Fairness Improvement with Multiple Protected Attributes: How Far Are We?

Zhenpeng Chen
zp.chen@ucl.ac.uk
University College London
London, United Kingdom

Jie M. Zhang*
jie.zhang@kcl.ac.uk
King's College London
London, United Kingdom

Federica Sarro
f.sarro@ucl.ac.uk
University College London
London, United Kingdom

Mark Harman
mark.harman@ucl.ac.uk
University College London
London, United Kingdom

ABSTRACT

Existing research mostly improves the fairness of Machine Learning (ML) software regarding a single protected attribute at a time, but this is unrealistic given that many users have multiple protected attributes. This paper conducts an extensive study of fairness improvement regarding multiple protected attributes, covering 11 state-of-the-art fairness improvement methods. We analyze the effectiveness of these methods with different datasets, metrics, and ML models when considering multiple protected attributes. The results reveal that improving fairness for a single protected attribute can largely decrease fairness regarding unconsidered protected attributes. This decrease is observed in up to 88.3% of scenarios (57.5% on average). More surprisingly, we find little difference in accuracy loss when considering single and multiple protected attributes, indicating that accuracy can be maintained in the multiple-attribute paradigm. However, the effect on precision and recall when handling multiple protected attributes is about five times and eight times that of a single attribute. This has important implications for future fairness research: reporting only accuracy as the ML performance metric, which is currently common in the literature, is inadequate.

CCS CONCEPTS

• **Software and its engineering → Extra-functional properties**; • **Social and professional topics → User characteristics**; • **Computing methodologies → Machine learning**.

KEYWORDS

Fairness improvement, machine learning, protected attributes, intersectional fairness

ACM Reference Format:
Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2024. Fairness Improvement with Multiple Protected Attributes: How Far Are We?.

*Corresponding author

ICSE '24, April 14–20, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0217-4/24/04.
https://doi.org/10.1145/3597503.3639083

In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE '24), April 14–20, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3597503.3639083

1 INTRODUCTION

Machine Learning (ML) software is being increasingly applied to assist decision-making in social-critical scenarios. This has raised surging concerns on the fairness of such software [52]. Indeed, ML software frequently exhibits unfair behaviors related to protected attributes such as sex [6, 8] and race [10, 49]. Unfair behaviors may compromise the benefits of historically disadvantaged groups, and lead to consequences for Software Engineering (SE) if and when the software is found to contravene laws against discrimination [66].

Reducing software unfairness has become an ethical duty of software researchers and engineers [23, 26]. The SE community is endeavoring to address unfairness issues in ML software [23, 26]. In the SE domain, unfairness issues are also referred to as 'fairness bugs' [25]. SE researchers have been extensively exploring various techniques to fix fairness bugs and improve software fairness [16, 17, 23, 24, 35, 39, 40, 66].

In practice, software systems can have multiple protected attributes that need to be considered simultaneously [26]. From the humanities' perspective, unfair software systems built into society lead to systematic disadvantages along multiple intersecting attributes, such as sex, race, age, disability status, and so on [33]. From the SE perspective, these protected attributes pose multiple fairness requirements, some of which can be competing or conflicting, raising issues of negotiation, mediation, and conflict resolution for software engineers [32].

The intersection of these attributes creates different levels of privilege or disadvantage for various possible subgroups. For instance, black women may be vulnerable to both sexism and racism [29]. To cater for this, the literature measures intersectional fairness as the maximum disparity between subgroups that combine membership from different protected attributes [34, 68]. Intersectional fairness has been encoded in legal regulations [2]. It clearly has implications for software researchers and engineers, who must consider the fairness regarding multiple protected attributes simultaneously as multiple non-functional software requirements.

However, the current software fairness literature is lacking in this critical aspect. Existing fairness improvement research mostly focuses on singleton sets of protected attributes [16, 17, 24, 27, 35,

40, 50, 66]. Unfortunately, the implications of this prevalent practice remain unclear. We have yet to fully understand the potential impact on desirable fairness properties concerning other protected attributes when catering for fairness according to a single protected attribute. Moreover, considering the legal and ethical fairness requirements [2, 34, 59], there is an urgent demand to apply fairness improvement methods to deal with multiple protected attributes. Consequently, a comprehensive study on the effectiveness of these methods in such situations becomes imperative.

Furthermore, there is an important interplay between fairness and other functional SE requirements. Specifically, it is widely recognized that fairness improvement typically comes at the cost of ML performance (e.g., accuracy), known as the fairness-performance trade-off [15, 27, 28, 40, 63]. Based on the current literature, it remains unclear how existing fairness improvement methods would trade-off between fairness and performance when multiple protected attributes are considered.

To fill these gaps in the literature, we conduct an extensive study of fairness improvement regarding multiple protected attributes, with 11 state-of-the-art fairness improvement methods. We evaluate these methods on five widely-adopted datasets, which cover financial, social, and medical application domains, with widely-studied ML models, fairness metrics, and performance metrics. We investigate the effect of these methods on the fairness regarding unconsidered protected attributes. We also check the performance decrease when multiple protected attributes are considered. We analyze the effectiveness of these methods for intersectional fairness improvement and fairness-performance trade-offs. If our study reveals their effectiveness, no alternative approaches would be needed; otherwise, one can build on our study's results to seek improvements in current methods or to devise novel methods that could better tackle the problem at hand.

Our study reveals the following findings: **1)** Existing methods can largely decrease fairness regarding unconsidered protected attributes. This decrease happens in up to 88.3% of scenarios (57.5% on average), with a significantly large effect in up to 69.2% of scenarios (29.1% on average). **2)** There is a similar decrease in accuracy when considering single and multiple protected attributes, with a 0.3% difference in decrease rate. However, precision and recall are greatly affected, with the impact on precision and recall when dealing with multiple protected attributes being about five times and eight times that of a single attribute. **3)** According to a state-of-the-art benchmarking tool [40], existing methods outperform the fairness-performance trade-off baseline constructed by the tool in 9.5%~71.3% of cases (less than 50% on average) when dealing with multiple protected attributes. These methods even decrease both intersectional fairness and ML performance in 6.5%~42.6% of cases (18.6% on average).

Additionally, our results on the effectiveness of each studied method in addressing intersectional fairness and fairness-performance trade-offs offer references for software engineers when selecting fairness improvement methods. Furthermore, these results can serve as easy-to-access baselines for researchers to evaluate new fairness improvement methods.

In summary, this paper makes the following contributions:

- A rigorous empirical study on the impact of fairness improvement methods on fairness regarding unconsidered protected attributes.
- An extensive study of the effectiveness of state-of-the-art fairness improvement methods in enhancing intersectional fairness and achieving fairness-performance trade-offs when considering multiple protected attributes.
- A publicly-available package [11], containing all scripts and data in this study, to facilitate replication and extension.

2 PRELIMINARIES

We start with introducing the background knowledge of this study.

2.1 Protected Attributes and Fairness

Fairness has emerged as an important research topic in the SE research community, with a particular focus on fairness of ML software [67]. The ML software fairness literature primarily concentrates on ML classification that predicts class labels for individuals based on their personal features [18, 23, 26, 27, 56, 66]. These class labels can be categorized as favorable or unfavorable. For instance, in the context of credit scoring, a good credit is considered a favorable label, while a bad credit is deemed unfavorable.

During the classification, certain personal attributes need to be protected against discrimination. These attributes are referred to as protected attributes, also known as sensitive attributes. Common protected attributes include sex, race, age, religion, disability status, and national origin. In real-world applications, ML software often needs to consider multiple protected attributes simultaneously.

Based on the value of a protected attribute, individuals can be divided into a privileged group and an unprivileged group. In practice, the privileged group tends to be associated with favorable labels, while the unprivileged group is more likely to receive unfavorable labels. For example, in credit scoring tasks, race is often considered a protected attribute [12]. Due to potential biases favoring the white group in the credit scoring models, the white group may be viewed as privileged, while the non-white group may be considered unprivileged.

To address such biases, legal regulations and the fairness literature advocate for group fairness [52, 63], which requires ML software to treat privileged and unprivileged groups equally. Mathematical metrics have been developed to measure group fairness. We describe three metrics that have been widely adopted in the software fairness literature [16, 17, 26, 27, 40]:

- **SPD** (Statistical Parity Difference) calculates the disparity in favorable rates between the privileged and unprivileged groups.
- **AOD** (Average Odds Difference) captures the average discrepancy in false-positive rates and true-positive rates between the privileged and unprivileged groups.
- **EOD** (Equal Opportunity Difference) assesses the disparity in true-positive rates between the privileged and unprivileged groups.

Let A represent the protected attribute, with 1 denoting the privileged group and 0 denoting the unprivileged group. Let Y denote the actual label and $\hat{Y}$ denote the predicted label, where 1 is the favorable class and 0 is the unfavorable class. The calculation methods of these fairness metrics are shown in Table 1.

Table 1: Fairness metrics.

Metric	Definition
SPD	$P[\hat{Y}=1\|A=0]-P[\hat{Y}=1\|A=1]$
AOD	$\frac{1}{2}(P[\hat{Y}=1\|A=0,Y=0]-P[\hat{Y}=1\|A=1,Y=0]$ $+P[\hat{Y}=1\|A=0,Y=1]-P[\hat{Y}=1\|A=1,Y=1])$
EOD	$P[\hat{Y}=1\|A=0,Y=1]-P[\hat{Y}=1\|A=1,Y=1]$

Table 2: Intersectional fairness metrics.

Metric	Definition
SPD	$\max_{s\in S} P[\hat{Y}=1\|A=s]-\min_{s\in S} P[\hat{Y}=1\|A=s]$
AOD	$\frac{1}{2}[\max_{s\in S}(P[\hat{Y}=1\|A=s,Y=0]+P[\hat{Y}=1\|A=s,Y=1])$ $-\min_{s\in S}(P[\hat{Y}=1\|A=s,Y=0]+P[\hat{Y}=1\|A=s,Y=1])]$
EOD	$\max_{s\in S} P[\hat{Y}=1\|A=s,Y=1]-\min_{s\in S} P[\hat{Y}=1\|A=s,Y=1]$

2.2 Intersectional Fairness

To consider multiple protected attributes and their intersectionality, researchers divide a population into subgroups based on the combination of different protected attributes [33, 34, 68]. The intersectional fairness is measured as the maximum disparity between any two subgroups [34, 68]. For instance, considering two protected attributes Sex = {Male, Female} and Race = {White, Non-White}, the subgroup set S = {(Male, White), (Male, Non-White), (Female, White), (Female, Non-White)}. If the favorable rates for the four subgroups are 50%, 40%, 30%, and 20%, SPD is calculated as 50% − 20% = 30%.

Specifically, in the context of intersectional fairness, **SPD** measures the maximum difference between subgroups in obtaining favorable outcomes; **AOD** measures the maximum of the average of differences in false-positive rates and true-positive rates between subgroups; **EOD** measures the maximum difference between subgroups in true-positive rates.

Formally, we use A to denote the protected attributes and define S as the set of all possible combinations of the protected attributes. Let s be a subgroup, where $s \in S$. These intersectional fairness metrics are calculated as shown in Table 2.

Compared to single-attribute fairness, intersectional fairness can capture unfairness amplified in subgroups that combine membership from different unprivileged groups [34], especially if such subgroups are particularly underrepresented in historical platforms of opportunity, e.g., the (Female, Non-White) subgroup in the aforementioned example.

3 EXPERIMENTAL SETUP

In this section, we describe our research questions and experimental settings for the study.

3.1 Research Questions

RQ1: *How do existing fairness improvement methods affect the fairness regarding unconsidered protected attributes?* This RQ investigates the negative side effect of single-attribute fairness improvement by studying its impact on fairness regarding the unconsidered protected attributes.

RQ2: *What intersectional fairness do existing fairness improvement methods achieve when considering multiple protected attributes?* This RQ evaluates the effectiveness of state-of-the-art fairness improvement methods in improving intersectional fairness.

RQ3: *What fairness-performance trade-off do existing fairness improvement methods achieve when considering multiple protected attributes?* This RQ explores whether fairness improvement for multiple protected attributes can bring more decrease in ML performance and how state-of-the-art methods make the trade-off between intersectional fairness and ML performance.

RQ4: *How well do existing fairness improvement methods apply to different decision tasks, ML models, and fairness and performance metrics, when dealing with multiple protected attributes?* This RQ enriches the empirical knowledge of RQ2 and RQ3, and explores whether existing methods are widely applicable.

3.2 Datasets and Models

We use five real-world datasets for study: **Adult** [1], **Compas** [4], **Default** [5], **Mep15** [3], and **Mep16** [7]. A description of each dataset is presented in Table 3. These datasets have been widely adopted in the fairness literature [23, 26, 27, 56, 68]. They encompass tasks that involve individuals' personal information across diverse fairness-critical domains, such as finance, social, and medical. In line with previous fairness research [23, 26, 27, 56, 68], we select the two protected attributes provided by each dataset for our study.

For each dataset, we train four ML models, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Deep Neural Network (DNN), which have been extensively adopted in fairness literature [16, 23, 26, 40, 66, 68–70]. LR, SVM, and RF use default configurations from relevant studies [26, 27, 40, 66], while the DNN employs a fully-connected architecture with five hidden layers, containing 64, 32, 16, 8, and 4 units, respectively, which has been widely used in previous fairness research involving similar datasets [68–70].

3.3 Fairness Improvement Methods

We employ 11 state-of-the-art fairness improvement methods for study, covering pre-processing, in-processing, and post-processing methods. Pre-processing methods focus on reducing bias in training data to achieve a fairer model; in-processing methods optimize training algorithms to enhance fairness; post-processing methods modify ML model predictions to ensure fair outcomes [27, 38].

First, we use eight state-of-the-art methods proposed in the ML literature [68].

Pre-processing methods:

- **RW** (Reweighting) [41] employs differential weighting of training data for each combination of groups and labels to achieve fairness.
- **DIR** (Disparate Impact Remover) [31] adjusts feature values to enhance fairness while preserving the rank-ordering within groups.

Table 3: Datasets.

Name	#Samples	#Features	Protected attributes	Favorable label (Proportion)	Task
Adult	48,843	7	sex, race	income > 50k (23.9%)	Predicting if a person's income is greater than $50k
Compas	7,214	7	sex, race	no recidivism (54.9%)	Predicting if a criminal defendant will re-offend
Default	30,000	23	sex, age	default (22.1%)	Predicting if a customer will default on payment
Mep15	15,830	42	sex, race	utilizer (17.2%)	Predicting healthcare utilization of a person
Mep16	15,675	42	sex, race	utilizer (16.8%)	Predicting healthcare utilization of a person

In-processing methods:

- **META** (Meta Fair Classifier) [22] employs a meta-algorithm to optimize fairness regarding protected attributes.
- **ADV** (Adversarial Debiasing) [65] uses adversarial techniques to minimize the presence of protected attributes in predictions, while concurrently maximizing prediction accuracy.
- **PR** (Prejudice Remover) [45] incorporates discrimination-aware regularization to mitigate the influence of protected attributes.

Post-processing methods:

- **EOP** (Equalized Odds Processing) [37] uses linear programming to calculate probabilities for adjusting output labels, aiming to optimize equalized odds concerning protected attributes.
- **CEO** (Calibrated Equalized Odds) [57] optimizes the probabilities of modifying output labels based on calibrated classifier score outputs, with the objective of achieving equalized odds.
- **ROC** (Reject Option Classification) [43] assigns favorable outcomes to unprivileged instances and unfavorable outcomes to privileged instances near the decision boundary, particularly when there is high uncertainty.

Second, we use three state-of-the-art methods proposed in the SE literature, including Fair-SMOTE [23], MAAT [26], and FairMask [56].

- **Fair-SMOTE** [23] generates synthetic samples to achieve balanced distributions not only between different labels but also among various protected attributes within the training data. Additionally, it removes ambiguous samples from the training set.
- **MAAT** [26] combines individual models optimized for ML performance and fairness concerning each protected attribute, respectively. It ensures that both fairness and ML performance objectives are met.
- **FairMask** [56] trains extrapolation models to predict protected attributes based on other data features. Subsequently, it uses these extrapolation models to modify the protected attributes in test data, enabling fairer predictions.

We apply each fairness improvement method to the original models obtained in Section 3.2. We repeat each experiment 20 times. Each time we randomize the dataset by shuffling it and then divide it into 70% training data and 30% test data.

When conducting fairness improvement for multiple protected attributes, we simultaneously consider these attributes instead of applying a fairness improvement method independently for each attribute. It is because individually applying the method for each protected attribute cannot maintain fairness for previous considered attributes while also guaranteeing fairness for subsequently considered attributes. For example, let us consider a dataset with two protected attributes, and only one attribute is considered for fairness improvement at a time. Pre-processing methods may not preserve the optimized characteristics for the first considered protected attribute when optimizing data characteristics for the second attribute. For instance, if we use the RW method to assign different weights based on the second attribute, it can undermine its intended weights for the first attribute. In the case of in-processing methods, training models for one protected attribute results in models specific to that attribute. Therefore, in-processing methods can disregard fairness considerations for the first attribute, when optimizing for the second attribute. Additionally, concerning post-processing methods, modifying the output to optimize fairness for the second protected attribute may not ensure the preservation of fairness for the first attribute.

3.4 Measurement Metrics

We employ three fairness metrics and five ML performance metrics, resulting in a total of 15 fairness-performance measurements for study, as detailed in the following.

3.4.1 Fairness metrics. We use three fairness metrics introduced in Section 2, including **SPD**, **AOD**, and **EOD**, which have been widely adopted in the fairness literature [16, 17, 26, 27, 40]. We calculate the fairness metric values for individual attributes and intersectional fairness, as listed in Tables 1 and 2. We use absolute values for all fairness metrics, whereby these metrics indicate the highest fairness when they equal 0, and larger values indicate greater unfairness.

3.4.2 Performance metrics. We follow previous work [26, 27] to use a comprehensive set of five common ML performance metrics for study: **accuracy**, **precision**, **recall**, **F1-score**, and **MCC** (Matthews Correlation Coefficient). We provide the formal definitions of these metrics in Table 4, where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. For precision, recall, and F1-score, we report the macro-average values, as done in previous research [26], to enable comparisons of overall performance on the favorable and unfavorable classes. To achieve this, we average the precision, recall, and F1-score results obtained for the two classes. For each of the five metrics, a higher value indicates better ML performance.

3.4.3 Fairness-performance trade-off measurement. To assess the fairness-performance trade-off, we rely on **Fairea** [40], a state-of-the-art benchmarking tool that offers a unified trade-off baseline for comparing various fairness improvement methods.

Table 4: ML performance metrics.

Metric	Definition
Accuracy	$(TP+TN)/(TP+FP+TN+FN)$
Precision	$TP/(TP+FP)$
Recall	$TP/(TP+FN)$
F1-score	$2 \times Precision \times Recall/(Precision+Recall)$
MCC	$(TP \times TN - FP \times FN)/\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$

Fairea visualizes fairness and performance values using a two-dimensional coordinate system and establishes the trade-off baseline by connecting fairness-performance points of the original ML model and a set of mutated models. The mutated models are generated by gradually transforming the original model into models that produce only the majority class in the dataset. Throughout this process, fairness improves as the predictive performance becomes equally worse for privileged and unprivileged groups. Fairea uses these naive mutated models to establish the trade-off baseline, as it expects that fairness improvement methods should outperform them.

Fairea classifies the trade-off effectiveness of fairness improvement methods into five levels by comparing the fairness-performance trade-off achieved by these methods with the established baseline:

- The *win-win trade-off* level includes methods that increase both fairness and performance.
- The *good trade-off* level includes methods that increase fairness, decrease performance, and achieve a better trade-off than the baseline generated by Fairea.
- The *poor trade-off* level includes methods that increase fairness, decrease performance, and achieve a worse trade-off than the baseline.
- The *inverted trade-off* level includes methods that decrease fairness but increase performance.
- The *lose-lose trade-off* level includes methods that decrease both fairness and performance.

Different from the original paper of Fairea [40] that focuses on single-attribute tasks, our study extends the scope to multi-attribute tasks. We conduct a comprehensive evaluation by considering 15 fairness-performance measurements (i.e., the combination of three fairness metrics and five ML performance metrics).

For each combination of *(dataset, ML model, fairness-performance measurement)*, we establish a trade-off baseline. To achieve this, we first train the original model and then generate the mutated models based on it. The process is repeated 20 times. Following the recommendation of Fairea [40], we determine the baseline by averaging the results of these multiple runs.

3.5 Statistical Analysis

We use three statistical analysis methods in this study: Mann Whitney U-test [51], Cliff's δ [47], and Spearman's correlation coefficient ρ [55]. Since these methods do not assume normality of the data, they are suitable for our study, where we deal with diverse data that may not follow a normal distribution.

In RQ1 and RQ2, we use the Mann Whitney U-test [51] to assess whether fairness improvement methods significantly impact fairness. To establish statistical significance, we follow previous work [26, 27] to consider a p-value lower than 0.05. Specifically, when comparing two sets of fairness values using the test, we conclude that the two sets have statistically different fairness if the p-value of the test is lower than 0.05. Furthermore, to measure the effect size of the impact, we adopt the Cliff's δ [47], a commonly-used metric in the SE literature [14, 47, 62]. Consistent with the literature [14, 47, 62], we consider a change with an absolute value of δ greater than or equal to 0.428 as indicative of a large effect. Additionally, in RQ1, we use the Spearman's correlation coefficient ρ [55] to explore potential factors correlated with the impact on unconsidered protected attributes. The coefficient ρ ranges from -1 to 1, with 1 representing a perfect positive correlation, 0 indicating no correlation, and -1 representing a perfect negative correlation. A correlation is considered statistically significant only when the coefficient yields a p-value lower than 0.05 [27].

4 RESULTS

This section answers our RQs based on the experimental results. Due to the page limit, we primarily report statistical results in the paper and include the results of each fairness improvement method for each scenario in our repository [11].

4.1 RQ1: Impact on Unconsidered Protected Attributes

This RQ investigates how fairness improvement methods affect the fairness regarding unconsidered protected attributes when targeting a single protected attribute. Each dataset-protected attribute pair, as shown in Table 3, represents a single-attribute fairness improvement task. For instance, in the case of the Adult dataset, we have two tasks: Adult-Sex and Adult-Race. We apply existing methods to improve fairness for one task and then examine the influence on the fairness of the other task. Each application is repeated 20 times using four ML models and three fairness metrics (more details in Section 3). We treat each combination of *(task, ML model, fairness metric)* as a scenario and calculate the proportions of scenarios where existing methods reduce fairness regarding unconsidered protected attributes, based on the average value obtained from the 20 repeated runs.

Table 5 shows the results. The methods that we study decrease fairness regarding unconsidered protected attributes in up to 88.3% of the total scenarios (with an average of 57.5% across different methods). We further analyze the significance and effect size of the decrease by using Mann Whitney U-test and Cliff's δ, and find that such decrease has a significantly large effect in up to 69.2% of the scenarios (29.1% on average).

We take the three methods highlighted in Table 5 as examples to illustrate why they cause a large fairness decrease for unconsidered protected attributes. Fair-SMOTE aims to balance data for one protected attribute, which can lead to more severe data imbalance for other protected attributes, resulting in reduced fairness for those attributes. ROC targets predictions with high uncertainty and tends to assign favorable outcomes to the unprivileged members and unfavorable outcomes to the privileged. For example, if sex is the considered protected attribute and race is the unconsidered one, predictions for (Male, Non-White) and (Female, White) members tend to be uncertain because the two subgroups have both privileged and unprivileged properties. Therefore, improving fairness for

Table 5: (RQ1) Proportions of scenarios where existing methods reduce fairness regarding unconsidered protected attributes (the second column) and also have a significantly large effect (the third column). Significantly large reductions are highlighted in bold. The top three values in each column are shaded. The results indicate that existing methods decrease fairness regarding unconsidered protected attributes in up to 88.3% of scenarios (57.5% on average across different methods), with a significantly large effect observed in up to 69.2% of scenarios (29.1% on average).

Method	↓ unconsidered fairness	Significantly large effect
RW	49.2%	**2.5%**
DIR	54.2%	**1.7%**
META	81.7%	**69.2%**
ADV	64.2%	**40.0%**
PR	70.8%	**45.0%**
EOP	22.5%	**3.3%**
CEO	37.5%	**4.2%**
ROC	85.0%	**68.3%**
Fair-SMOTE	88.3%	**63.3%**
MAAT	29.2%	**4.2%**
FairMask	50.0%	**18.3%**
Average	57.5%	**29.1%**

sex can lead to more unfavorable outcomes for (Male, Non-White) and more favorable outcomes for (Female, White), causing further unfairness regarding race. META aims to improve fairness for the protected attribute during training, but this objective may conflict with fairness for other unconsidered protected attributes, resulting in reduced fairness for these attributes.

To gain further insight into the fairness reduction, we explore its potential reasons from the perspective of datasets. If the protected attributes in a dataset consistently have the same values (i.e., perfectly positively correlated), improving fairness for one protected attribute would be equivalent to doing so for the others. Drawing inspiration from this observation, we hypothesize that as the correlation between the considered and unconsidered protected attributes becomes more positive, fairness improvement methods will have a lesser adverse impact on the fairness concerning the unconsidered attributes.

To test this hypothesis, we assign 1 to denote the privileged group and 0 for the unprivileged group for each protected attribute. For each task, we calculate Spearman's correlation coefficient ρ to quantify the correlation between the considered and unconsidered protected attributes. Additionally, we determine the proportions of scenarios where existing methods reduce the fairness regarding the unconsidered attributes. Then, we measure the correlation between these proportions and the correlation of protected attributes.

Table 6 presents the results. Based on the results, we confirm our hypothesis with a correlation coefficient of $\rho = -0.640$ at a significance level of 0.05 (p-value < 0.05). We illustrate this negative correlation further using the Default dataset as an example. This dataset exhibits the most negative correlation ($\rho = -0.069$, p-value < 0.05) among all the datasets. Meanwhile, on this dataset,

Table 6: (RQ1) Correlation between considered and unconsidered protected attributes (second column) and proportions of scenarios where existing methods reduce fairness for unconsidered protected attributes (third column). * indicates a significant correlation with p-value < 0.05. We find that the more positive the correlation between the considered and unconsidered protected attributes, the less existing methods reduce fairness regarding the unconsidered protected attributes.

Task	Correlation between protected attributes	↓ unconsidered fairness
Adult-Sex	0.101*	44.7%
Adult-Race	0.101*	56.8%
Compas-Sex	0.068*	40.9%
Compas-Race	0.068*	33.3%
Default-Sex	-0.069*	84.1%
Default-Age	-0.069*	76.5%
Mep15-Sex	-0.015	74.2%
Mep15-Race	-0.015	64.4%
Mep16-Sex	-0.016*	48.5%
Mep16-Race	-0.016*	51.5%
Correlation between the last two columns		-0.640*

existing methods reduce fairness regarding unconsidered protected attributes in the highest proportion of scenarios.

> **Finding 1:** The fairness improvement methods that we study can lead to decreased fairness regarding unconsidered protected attributes to a large extent. Specifically, the decrease occurs in up to 88.3% of scenarios (on average 57.5%), with a significantly large effect in up to 69.2% of scenarios (on average 29.1%). Our correlation analysis suggests that the more positive the correlation between the considered and unconsidered protected attributes, the less existing methods reduce fairness regarding the unconsidered protected attributes.

4.2 RQ2: Intersectional Fairness Improvement

This RQ aims to evaluate the effectiveness of existing methods in improving intersectional fairness when dealing with multiple protected attributes. To this end, we use five datasets from Table 3, along with four ML models and three fairness metrics for each dataset. We consider each *(dataset, model, fairness metric)* combination as a scenario and calculate the proportions of scenarios where existing methods improve intersectional fairness based on the average results of 20 repeated runs. We also report the proportions of scenarios where the improvement has a significantly large effect by using Mann Whitney U-test and Cliff's δ.

Table 7 presents the results. The 11 methods studied improve intersectional fairness in a wide range of scenarios, ranging from 6.7% to 98.3%. In particular, MAAT, FairMask, and RW exhibit the most consistent improvements, achieving this in 98.3%, 93.3%, and 90.0% of scenarios, respectively. Furthermore, these three methods significantly improve intersectional fairness with a large effect in the most scenarios, accounting for 71.7%, 68.3%, and 68.3%, respectively.

Table 7: (RQ2) Proportions of scenarios where existing methods improve intersectional fairness (the second column) and also have a significantly large effect (the third column). The proportions of scenarios where such improvement has a significantly large effect are highlighted in bold. The top three values in each column are shaded. MAAT, FairMask, and RW improve intersectional fairness in the most scenarios.

Method	↑ intersectional fairness	Significantly large effect
RW	90.0%	**68.3%**
DIR	68.3%	**43.3%**
META	28.3%	**16.7%**
ADV	36.7%	**31.7%**
PR	75.0%	**53.3%**
EOP	85.0%	**51.7%**
CEO	61.7%	**43.3%**
ROC	6.7%	**0.0%**
Fair-SMOTE	56.7%	**38.3%**
MAAT	98.3%	**71.7%**
FairMask	93.3%	**68.3%**

The superiority of these methods can be attributed to their ability to mitigate data bias, preventing its amplification during training or decision-making. However, a common limitation of them is the need for access of training data. In situations where obtaining such access is infeasible, (e.g., due to privacy concerns), practitioners may prefer using post-processing methods that modify prediction outcomes to ensure fairness without requiring access to training data. Among the post-processing methods studied, EOP stands out, improving intersectional fairness in the most scenarios (85.0%), with a significantly large effect in 51.7% of cases.

We further measure the effectiveness of existing methods by calculating the absolute and relative changes in fairness metric values. Table 8 presents the results averaged over the five datasets and four models under study. Methods that lower fairness metric values to the largest extent contribute the most to improving intersectional fairness, as smaller fairness metric values indicate reduced unfairness. Notably, MAAT and FairMask, two state-of-the-art methods from the SE literature, demonstrate a general advantage in enhancing intersectional fairness across various fairness metrics. Specifically, they improve AOD fairness by 32.4% and 34.9%, respectively. Additionally, RW, PR, and EOP also yield favorable results in specific fairness metrics. Among the highlighted methods, EOP, as a post-processing method, is the only one that does not require access to training data. This makes EOP a suitable choice for scenarios where obtaining such access is infeasible.

Finding 2: The fairness improvement methods that we study improve intersectional fairness in 6.7%~98.3% of the scenarios. Notably, MAAT, FairMask, and RW achieve this goal in the most scenarios, accounting for 98.3%, 93.3%, and 90.0%, respectively; the improvement has a significantly large effect in 71.7%, 68.3%, and 68.3% of scenarios. For applications where obtaining access to training data is impossible (e.g., due to privacy concerns), EOP can be a better option, which improves intersectional fairness in 85.0% of scenarios, with a significantly large effect in 51.7% of scenarios.

Table 8: (RQ2) Absolute and relative changes (in parentheses) in intersectional fairness achieved by existing methods. The top three values in each column are highlighted. MAAT and FairMask demonstrate superiority in improving intersectional fairness across different fairness metrics.

Method	SPD		AOD		EOD	
RW	-0.030	(-20.0%)	-0.036	(-25.7%)	-0.040	(-22.8%)
DIR	-0.026	(-14.6%)	-0.024	(-7.6%)	-0.032	(-3.0%)
META	0.192	(236.4%)	0.139	(202.4%)	0.048	(66.0%)
ADV	-0.004	(-3.7%)	0.032	(48.6%)	0.046	(57.5%)
PR	-0.056	(-41.7%)	-0.031	(-13.1%)	-0.046	(-6.9%)
EOP	-0.038	(-20.6%)	-0.035	(-20.5%)	-0.037	(-17.2%)
CEO	-0.006	(-9.5%)	-0.003	(-8.4%)	-0.022	(-10.0%)
ROC	0.104	(80.2%)	0.085	(64.6%)	0.069	(39.8%)
Fair-SMOTE	-0.004	(15.2%)	-0.028	(-4.7%)	-0.038	(-14.9%)
MAAT	-0.041	(-29.3%)	-0.042	(-32.4%)	-0.054	(-29.2%)
FairMask	-0.034	(-22.2%)	-0.050	(-34.9%)	-0.067	(-32.0%)

4.3 RQ3: Fairness-performance Trade-off

This RQ aims to evaluate the fairness-performance trade-off achieved by existing methods when dealing with multiple protected attributes. We investigate this RQ by answering two sub-questions.

4.3.1 RQ3.1: Does the application of existing methods to improve fairness for multiple protected attributes lead to significantly greater performance reduction compared to improving fairness for a single attribute? It is well known that fairness improvement often comes at the expense of ML performance [15, 27, 28, 40, 63]. Intuitively, improving fairness for multiple protected attributes might result in a more substantial performance decrease than doing so for a single attribute. To explore this, we calculate the absolute and relative changes in the five performance metrics that we analyze when employing existing fairness improvement methods for one or multiple protected attributes. These changes are then averaged over the five datasets and four models used in our study.

Table 9 presents the results. Different from intuition, we observe a similar accuracy decrease when considering single and multiple protected attributes. Specifically, when considering two protected attributes, accuracy is further decreased by 0.3% (-2.1% vs. -2.4%) with an absolute change of 0.002 (-0.018 vs. -0.020), compared to considering a single protected attribute. This indicates that accuracy can be reasonably maintained in the multiple-attribute paradigm.

In contrast, precision and recall are greatly affected. The impact on precision and recall when dealing with two protected attributes is about five times (-2.3% vs. -11.3%) and eight times (1.3% vs. 11.0%) that of dealing with a single one. In particular, among all five metrics, only recall shows an overall improvement when multiple protected attributes are considered. This improvement can be attributed to existing methods enhancing the predictive power of ML models for underrepresented groups, resulting in improved recall across the population. However, since precision and recall often conflict with each other [19], the increase in recall may lead to a decrease in precision. Consequently, using the F1-score to compute the harmonic

Table 9: (RQ3.1) Absolute and relative changes (in parentheses) in ML performance when existing methods improve fairness for single or multiple protected attributes. On average, the accuracy decrease is similar when considering single or multiple protected attributes, with only a 0.3% difference in decrease rate. However, precision and recall show significant variations between the two scenarios.

Method	Accuracy		Precision		Recall		F1-score		MCC	
	single-attr	multi-attr	single-attr	multi-attr	single-attr	multi-attr	single-attr	multi-attr	single-attr	multi-attr
RW	-0.001 (-0.2%)	-0.001 (-0.2%)	0.001 (0.2%)	-0.087 (-11.4%)	-0.003 (-0.4%)	0.088 (13.7%)	-0.002 (-0.3%)	0.002 (0.2%)	-0.003 (-0.5%)	0.001 (0.2%)
DIR	-0.004 (-0.5%)	-0.008 (-0.9%)	-0.002 (-0.3%)	-0.107 (-14.1%)	-0.010 (-1.5%)	0.081 (12.6%)	-0.013 (-2.0%)	-0.024 (-3.6%)	-0.018 (-4.8%)	-0.036 (-9.5%)
META	-0.063 (-7.4%)	-0.079 (-9.4%)	-0.064 (-8.3%)	-0.039 (-5.0%)	0.056 (8.8%)	0.009 (1.6%)	0.000 (0.3%)	-0.011 (-1.4%)	0.001 (1.8%)	-0.020 (-3.8%)
ADV	0.002 (0.3%)	0.000 (0.0%)	0.005 (0.7%)	-0.091 (-11.9%)	0.005 (0.9%)	0.092 (14.4%)	0.007 (1.3%)	0.000 (0.3%)	0.010 (3.9%)	0.001 (1.8%)
PR	-0.003 (-0.4%)	-0.014 (-1.6%)	0.012 (1.7%)	-0.147 (-19.2%)	-0.021 (-3.2%)	0.122 (18.7%)	-0.023 (-3.4%)	-0.074 (-10.4%)	-0.023 (-5.2%)	-0.085 (-18.5%)
EOP	-0.013 (-1.7%)	-0.012 (-1.6%)	-0.018 (-2.4%)	-0.103 (-13.6%)	-0.018 (-2.7%)	0.072 (11.3%)	-0.020 (-2.8%)	-0.015 (-2.1%)	-0.037 (-9.1%)	-0.030 (-7.1%)
CEO	-0.010 (-1.3%)	-0.005 (-0.6%)	-0.008 (-1.0%)	-0.103 (-13.5%)	-0.027 (-4.0%)	0.081 (12.7%)	-0.033 (-4.6%)	-0.016 (-2.2%)	-0.045 (-10.5%)	-0.023 (-5.1%)
ROC	-0.056 (-6.7%)	-0.051 (-6.1%)	-0.069 (-9.0%)	-0.032 (-4.2%)	0.060 (9.4%)	0.023 (3.8%)	0.007 (1.3%)	0.010 (1.7%)	0.004 (1.9%)	0.004 (2.0%)
Fair-SMOTE	-0.044 (-5.4%)	-0.044 (-5.4%)	-0.061 (-7.9%)	-0.042 (-5.6%)	0.051 (7.9%)	0.028 (4.5%)	0.011 (1.8%)	0.009 (1.5%)	0.001 (0.5%)	-0.004 (-0.5%)
MAAT	0.000 (-0.1%)	-0.003 (-0.4%)	0.009 (1.2%)	-0.109 (-14.3%)	-0.009 (-1.3%)	0.095 (14.7%)	-0.007 (-1.0%)	-0.021 (-2.9%)	-0.005 (-0.9%)	-0.023 (-5.4%)
FairMask	-0.001 (-0.2%)	-0.003 (-0.4%)	0.000 (0.0%)	-0.090 (-11.9%)	0.001 (0.1%)	0.085 (13.3%)	0.001 (0.2%)	-0.002 (-0.2%)	0.000 (0.4%)	-0.005 (-0.9%)
Average	-0.018 (-2.1%)	-0.020 (-2.4%)	-0.018 (-2.3%)	-0.086 (-11.3%)	0.008 (1.3%)	0.071 (11.0%)	-0.007 (-0.8%)	-0.013 (-1.7%)	-0.011 (-2.0%)	-0.020 (-4.3%)

mean of precision and recall provides a more balanced measure [36] . In terms of F1-score, the performance decrease is twice as much (-0.8% vs. -1.7%) when considering two protected attributes. A similar decrease pattern is also observed for MCC.

The findings regarding performance decreases carry significant implications for the use of performance metrics in fairness research. As mentioned by previous studies [27], the majority of existing fairness research [20–22, 31, 40–42, 44, 64, 66] relies solely on accuracy as the performance metric. Our results demonstrate that by exclusively focusing on accuracy, researchers may overlook the significant impact on other performance metrics when dealing with multiple protected attributes. In real-world applications, metrics such as recall and precision are important [27, 53, 54]. Therefore, solely relying on accuracy may not provide engineers with a complete picture when selecting fairness improvement methods for such applications.

Furthermore, we find that different methods can exhibit distinct performance decrease patterns. For instance, we examine the accuracy decrease of two top-performing methods identified in RQ2 (i.e., FairMask and RW) when considering two protected attributes. FairMask, which improves fairness by modifying protected attribute information, experiences a doubled accuracy decrease (-0.2% vs. -0.4%) when dealing with two protected attributes. This is because FairMask needs to obfuscate more information to achieve fairness, resulting in a higher accuracy sacrifice. Compared to FairMask, RW adjusts only the weights of samples in training without modifying any attributes, avoiding introducing significant noise when dealing with more protected attributes. This characteristic enables RW to maintain a comparable accuracy when dealing with one or two protected attributes (-0.2% vs. -0.2%).

Finding 3: Different from intuition, we observe a similar accuracy decrease when considering single and multiple protected attributes (with a 0.3% difference in decrease rate), suggesting that accuracy can be maintained in the multiple-attribute paradigm. However, precision and recall are greatly affected, showing an impact around five times and eight times greater, respectively, when dealing with multiple protected attributes compared to a single attribute. Therefore, considering only change in accuracy (as most fairness studies do) cannot provide implications for real-world applications where precision or recall is crucial.

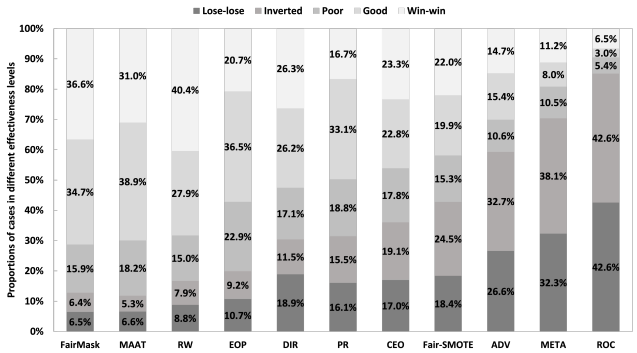


Figure 1: (RQ3.2) Effectiveness level distributions of existing methods in fairness-performance trade-off when dealing with multiple protected attributes. FairMask, MAAT, and RW achieve the best trade-off, with 71.3%, 69.9%, and 68.3% of cases falling into the win-win or good trade-off, respectively.

4.3.2 RQ3.2: Which trade-off effectiveness levels do existing fairness improvement methods fall into according to Fairea? In this RQ, we use Fairea [40], a state-of-the-art benchmarking tool described in Section 3.4.3, to evaluate the effectiveness of existing methods in achieving the trade-off between intersectional fairness and ML performance when dealing with multiple protected attributes. For each of the five datasets, we use four ML models and 15 fairness-performance measurements. We apply each fairness improvement method to the $5 \times 4 \times 15 = 300$ *(dataset, model, measurement)* combinations. We repeat the experiments 20 times and treat each single run as an individual case. As a result, we have $300 \times 20 = 6,000$ cases for each method. We use Fairea to classify the trade-offs achieved by each method in these cases into different effectiveness levels, and then calculate the distribution of the effectiveness levels.

We illustrate the results in Figure 1 and present the methods in descending order by the proportion of cases where each method

beats the trade-off baseline constructed by Fairea (i.e., achieving win-win or good trade-off). These methods surpass the trade-off baseline in 9.5%∼71.3% of cases, with an average of less than 50% of cases (46.9%). They also achieve a lose-lose trade-off (i.e., decrease both intersectional fairness and performance) in 6.5%∼42.6% of cases (18.6% on average).

Among the 11 methods under study, FairMask, MAAT, and RW achieve the best trade-off effectiveness. They beat the trade-off baseline constructed by Fairea in 71.3%, 69.9%, and 68.3% of the evaluated cases, respectively. In particular, they improve both intersectional fairness and performance (i.e., win-win trade-off) in 36.6%, 31.0%, and 40.4% of cases. Nevertheless, they still suffer from a lose-lose trade-off (i.e., decreasing both intersectional fairness and ML performance) in 6.5%, 6.6%, and 8.8% of cases.

> **Finding 4:** The state-of-the-art fairness improvement methods that we study beat the fairness-performance trade-off baseline constructed by Fairea in 9.5%∼71.3% of cases (less than 50% on average) when dealing with multiple protected attributes. They also lead to a decrease of both intersectional fairness and performance in 6.5%∼42.6% of cases (18.6% on average). Among these methods, FairMask, MAAT, and RW are the most effective, surpassing the trade-off baseline in 71.3%, 69.9%, and 68.3% of the evaluated cases, respectively.

4.4 RQ4: Applicability

This RQ aims to explore whether existing fairness improvement methods are widely applicable to different datasets, models, and fairness-performance measurements. Specifically, we analyze the effectiveness of these methods in improving intersectional fairness and achieving the trade-off between intersectional fairness and performance. For the effectiveness in fairness improvement, we calculate the proportions of scenarios where existing methods improve intersectional fairness for each dataset, model, and fairness measurement, respectively. For example, for each dataset, we have $4 \times 3 = 12$ *(model, fairness metric)* combinations, and compute the proportion of the 12 scenarios in which each method improves intersectional fairness. For the effectiveness in the fairness-performance trade-off, we use the proportion of cases that surpass the trade-off baseline constructed by Fairea as the indicator [26], and calculate the proportions achieved by each method for each dataset, model, and fairness-performance measurement.

We find that for each dataset, model, and measurement, at least one of the top three methods identified in RQ2 and RQ3 (i.e., RW, MAAT, and FairMask) can achieve the best intersectional fairness improvement and the best fairness-performance trade-off. Due to the page limit, we show only the results of these three methods in Figure 2, and the results for all methods can be found in our repository [11].

As shown in Figure 2(a), for each dataset and each model, at least one of the methods RW, MAAT, and FairMask can improve intersectional fairness in 100% of scenarios. However, regarding the fairness measurements, all three methods cannot do so for AOD. It is reasonable since the AOD fairness is more complex and difficult to satisfy than SPD and EOD, as demonstrated in previous work [27].

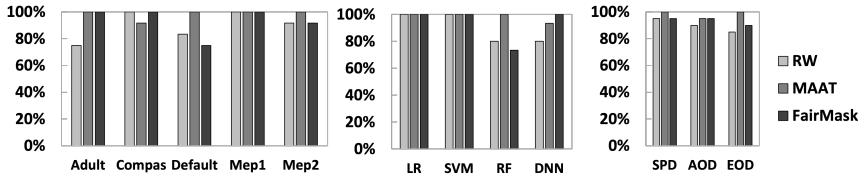
Figure 2(b) reveals that these methods tend to achieve worse fairness-performance trade-offs on imbalanced datasets compared to balanced datasets. Specifically, from Table 3, we find that the majority class in the Adult, Compas, Default, Mep1, and Mep2 datasets accounts for 76.1%, 54.9%, 77.9%, 82.8%, and 83.2%, respectively. Among these datasets, Compas, being the most balanced, exhibits the best fairness-performance trade-off results. This observation is expected since the classification on balanced datasets is generally considered easier than on imbalanced ones [58], making it relatively easier for existing methods to retain performance while improving fairness on such datasets. Our findings indicate that achieving a good trade-off between fairness and precision is more challenging for existing methods compared to the trade-off between fairness and other performance metrics. This observation aligns with the results from RQ3.1, where existing fairness improvement methods are shown to cause the most significant decrease in precision among all the performance metrics.

> **Finding 5:** It is challenging for fairness improvement methods to achieve good fairness-performance trade-offs for imbalanced datasets and applications where precision matters when dealing with multiple protected attributes. RW, MAAT, and FairMask achieve the best intersectional fairness improvement and fairness-performance trade-off results for each dataset, model, and measurement that we study.

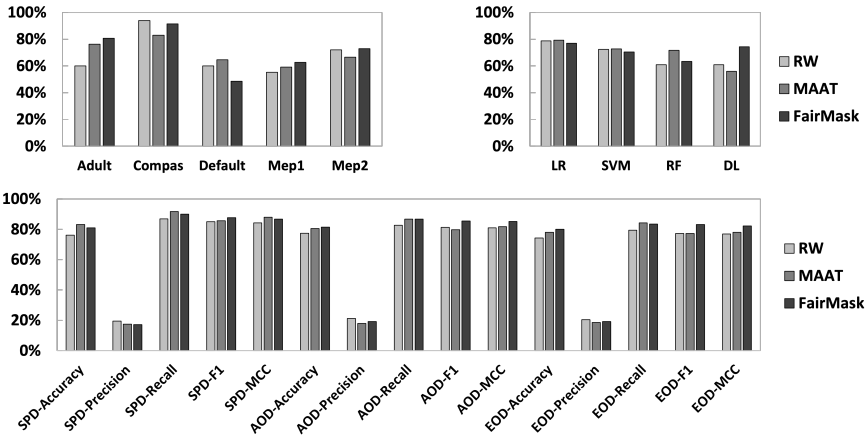
5 IMPLICATIONS

Implications for software engineers: 1) There is a substantial risk of inadvertently exacerbating unfairness for unconsidered protected attributes and violating anti-discrimination laws when software engineers focus on certain protected attributes. This is due to the presence of a noteworthy trade-off between fairness across different protected attributes observed in our study. If the trade-off comes simply because the data is skewed thus creating 'artificial contention' between protected attributes, it can be corrected by software engineers, as a type of fairness bug. Otherwise, if it is inherent to the problem that there is a trade-off between the fairness regarding different protected attributes, the competing fairness requirements raise issues of negotiation, mediation, and conflict resolution for engineers. **2)** We have compared 11 state-of-the-art fairness improvement methods when dealing with multiple protected attributes based on several different metrics. The results offer valuable insights and references for software engineers when they select fairness improvement methods that address multiple protected attributes in line with their specific objectives, thereby mitigating legal risks associated with software discrimination. For example, the results of RQ2 reveal that when faced with limited access to training data, the EOP method emerges as a viable choice for improving intersectional fairness. Conversely, MAAT can be a suitable option while having access to training data.

Implications for policy makers: Despite many laws and regulations seeking to protect multiple attributes simultaneously [2, 34], our findings reveal that fairness objectives for protected attributes such as sex and race may compete with each other. As a result, expecting software systems to perfectly satisfy these competing



(a) Proportions of scenarios where existing methods improve intersectional fairness, organized by different datasets, models, and measurements



(b) Proportions of cases beating the trade-off baseline, organized by different datasets, models, and measurements

Figure 2: (RQ4) Effectiveness in intersectional fairness improvement and fairness-performance trade-off of the best three methods identified in this study (i.e., RW, MAAT, and FairMask) across various datasets, models, and measurements. We observe that it is challenging for these methods to achieve a good fairness-performance trade-off for imbalanced datasets and precision-critical applications.

fairness objectives under a single law or regulation can be unrealistic. To achieve a balanced approach towards fairness in software systems, policy makers and legislative bodies should carefully consider these competing fairness considerations when formulating laws and regulations.

Implications for researchers: 1) There is a potential risk associated with the common research practice of focusing on one protected attribute at a time, as fairness improvement methods can significantly impact fairness regarding unconsidered protected attributes (RQ1). This emphasizes the importance of considering multiple protected attributes, not only in real-world applications, but also as a crucial objective in research. Researchers should be mindful of the potential consequences of neglecting the impact on unconsidered protected attributes and strive to broaden the scope of their investigations to encompass multiple dimensions of fairness. **2)** Considering the well-known fairness-performance trade-off and the trade-off between fairness regarding different protected attributes observed in our study (RQ1), researchers have the opportunity to develop multi-objective optimization techniques that address both these trade-offs simultaneously. **3)** Researchers can prioritize proposing post-processing fairness improvement techniques for tackling multiple protected attributes. This focus is driven by the finding that RW, MAAT, and FairMask are the most effective methods for enhancing intersectional fairness (RQ2), but they all require access to training data, posing challenges in real-world fairness-related applications due to concerns about releasing sensitive personal information. In contrast, EOP, the top-performing post-processing method that does not require such access, achieves intersectional fairness improvement in 18.3% fewer scenarios (RQ2). **4)** Researchers should include precision and recall in their evaluations when dealing with multiple protected attributes, moving beyond sole reliance on accuracy, as commonly observed in existing fairness research [20–22, 31, 40–42, 44, 64, 66]. It is because fairness improvements can have a significant impact on precision and recall when considering multiple protected attributes (RQ3). Precision and Recall's wide adoption in real-world applications further emphasizes the importance [27, 53, 54]. **5)** Researchers can design novel methods specifically tailored to optimize the fairness-performance trade-off for imbalanced datasets and precision-critical applications, because existing methods may not suffice under such circumstances (RQ4). This is important especially considering that these circumstances are common in the real-world applications [46, 48].

6 THREATS TO VALIDITY

Datasets: Due to the lack of public availability of datasets across all domains with fairness issues, we use five widely-adopted datasets

that cover common domains frequently explored in the fairness literature. However, it is important to note that these widely-adopted datasets can have potential limitations [30], which may affect the validity of our findings. In addition, regarding protected attributes, we consider only sex, race, and age, which are the most widely-studied ones in the fairness literature [60]. In the future, one could replicate this study with more datasets and more protected attributes.

ML models: To mitigate potential concerns regarding the selection of ML models, we have carefully chosen representative models for our study. Our selection includes both traditional ML models such as LR, RF, and SVM, as well as DNN. LR, RF, and SVM have been widely adopted in decision-making scenarios of social significance where fairness is a critical factor, as supported by existing research [26] and a recent official report from the UK government [9]. Moreover, DNN is increasingly adopted in the fairness literature due to their expanding applications in decision-making contexts [27, 61, 68, 70].

Fairness improvement methods: In recent years, the significance of fairness has gained considerable attention, resulting in an increasing number of fairness improvement methods. Given the extensive range of methods available, it is challenging to incorporate all of them in our study. To address this limitation, we choose 11 representative methods that have been recognized as state-of-the-art in the literature [23, 26, 56, 68]. While we have considered a wide range of fairness improvement methods that can be applied to different phases of the machine learning pipeline, we acknowledge that, in practice, they are not always applicable, given the constraints of the data sources and the application domain.

Evaluation metrics: Fairness metrics have been increasingly emerging in the literature. It is impractical to incorporate all of these metrics in our study. To address this limitation, we have followed previous studies [26] to use three fairness metrics that have gained significant adoption in the literature. Similarly, for performance evaluation, we have used the most widely-adopted metrics for ML classification [26]. We have employed a comprehensive set of 15 fairness-performance measurements, which is the most extensive range used in the literature.

7 RELATED WORK

Researchers have made significant efforts to address unfairness issues in ML software by proposing various fairness improvement methods. For instance, IBM has launched the AIF360 toolkit that integrates cutting-edge fairness improvement methods [13], such as Reweighting [41], Prejudice Remover [45], and Equalized Odds Processing [37]. These methods can be categorized into pre-, in-, and post-processing methods, which respectively optimize training data, the learning process, and decision outputs to improve fairness [27]. While a plethora of fairness improvement methods have been proposed, the majority of them primarily concentrate on addressing individual protected attributes, as emphasized in recent work [23, 26, 34, 56].

With the increasing number of fairness improvement methods, previous studies have aimed to empirically evaluate and compare existing methods. For instance, Biswas and Rajan [16] assessed seven fairness improvement methods using ML models gathered from a crowd-sourced platform, analyzing the resulting fairness outcomes and their impact on performance. Hort et al. [40] introduced Fairea, a benchmarking tool that provides a unified baseline for evaluating the fairness-performance trade-off obtained by different methods. Chen et al. [27] used Fairea to conduct a comprehensive empirical study of state-of-the-art fairness improvement methods. However, all these evaluations are limited to tasks involving a single protected attribute at a time.

Recent SE studies have presented methods capable of handling multiple protected attributes simultaneously [23, 26, 56]. However, the systematic comparison of these methods remains understudied. Specifically, when evaluating their method for dealing with multiple protected attributes, Chakraborty et al. [23] did not employ any method for comparison; Chen et al. [26] and Peng et al. [56] compared the proposed methods with only the one proposed by Chakraborty et al. [23]. Additionally, the effectiveness of these methods in improving intersectional fairness was not evaluated in previous work. Recently, Zhang and Sun [68] adapted fairness improvement methods previously proposed in the ML community so that they can handle multiple protected attributes. However, they did not compare these methods with the recent ones proposed by the SE community [23, 26, 56], and they used SPD as the only group fairness metric and accuracy as the only performance metric for evaluation. In this paper, we systematically study the effectiveness of 11 state-of-the-art fairness improvement methods (covering methods from both ML and SE communities) in improving intersectional fairness with multiple widely-adopted fairness metrics. We also investigate the fairness-performance trade-off achieved by these methods in the context of multiple protected attributes using 15 fairness-performance measurements.

8 CONCLUSION

This paper presents an extensive study of fairness improvement with multiple protected attributes. We systematically study 11 state-of-the-art fairness improvement methods from the literature, on widely-adopted benchmark datasets, ML models, performance metrics, and fairness metrics. We uncover the potential trade-off between fairness regarding different protected attributes and find that the correlation between the attributes can be a possible reason. We also explore the influence on performance when improving fairness for multiple protected attributes. Moreover, we benchmark existing methods and compare their effectiveness in improving intersectional fairness and achieving the trade-off between intersectional fairness and performance. The results provide actionable implications for researchers, software engineers, and policy makers.

9 DATA AVAILABILITY

We have made the code and data used in this paper publicly accessible [11].

ACKNOWLEDGMENTS

Zhenpeng Chen, Federica Sarro, and Mark Harman are supported by the ERC Advanced Grant No.741278 (EPIC: Evolutionary Program Improvement Collaborators). Jie M. Zhang is supported by the UKRI Trustworthy Autonomous Systems Node in Verifiability, with Grant Award Reference EP/V026801/2.

REFERENCES

[1] 1996. The Adult Census Income dataset. https://archive.ics.uci.edu/ml/datasets/adult.

[2] 2003. U.S. Equal Employment Opportunity Commission. https://www.eeoc.gov/initiatives/e-race/significant-eeoc-racecolor-casescovering-private-and-federal-sectors#intersectional.

[3] 2015. The Mep15 dataset. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181.

[4] 2016. The Compas dataset. https://github.com/propublica/compas-analysis.

[5] 2016. The Default dataset. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

[6] 2016. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] 2016. The Mep16 dataset. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192.

[8] 2018. Study finds gender and skin-type bias in commercial artificial-intelligence systems. https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212.

[9] 2020. Review into bias in algorithmic decision-making. https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making.

[10] 2021. When good algorithms go sexist: Why and how to advance AI gender equity. https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity.

[11] 2023. Replication package. https://doi.org/10.6084/m9.figshare.24943590.v1.

[12] Galina Andreeva, Jake Ansell, and Jonathan Crook. 2004. Impact of anti-discrimination laws on credit scoring. *Journal of Financial Services Marketing* 9 (2004), 22–33.

[13] 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 T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4:1–4:15.

[14] Kwabena Ebo Bennin, Jacky Keung, Akito Monden, Passakorn Phannachitta, and Solomon Mensah. 2017. The significant effects of data sampling approaches on software defect prioritization and classification. In *Proceedings of the 2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2017*. 364–373.

[15] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[16] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2020*. 642–653.

[17] Sumon Biswas and Hridesh Rajan. 2021. Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2021*. 981–993.

[18] Sumon Biswas and Hridesh Rajan. 2023. Fairify: Fairness verification of neural networks. In *Proceedings of the 45th IEEE/ACM International Conference on Software Engineering, ICSE 2023*. 1546–1558.

[19] Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American Society for Information Science* 45, 1 (1994), 12–19.

[20] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *Proceedings of the 2009 IEEE International Conference on Data Mining*. 13–18.

[21] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[22] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm wit provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*. 319–328.

[23] 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 European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2021*. 429–440.

[24] 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 European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2020*. 654–665.

[25] Zhenpeng Chen, Jie M. Zhang, Max Hort, Federica Sarro, and Mark Harman. 2022. Fairness testing: A comprehensive survey and analysis of trends. *CoRR* abs/2207.10223 (2022).

[26] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2022. MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022*. 1122–1134.

[27] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology* 32, 4 (2023), 106:1–106:30.

[28] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017*. 797–806.

[29] Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, deminist theory and antiracist politics. *Feminist Legal Theories* (1989), 139–167.

[30] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems* 34 (2021), 6478–6490.

[31] 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.

[32] Anthony Finkelstein, Mark Harman, S. Afshin Mansouri, Jian Ren, and Yuanyuan Zhang. 2008. "Fairness analysis" in requirements assignments. In *Proceedings of the 16th IEEE International Requirements Engineering Conference, RE 2008*. 115–124.

[33] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *Proceedings of the 36th IEEE International Conference on Data Engineering, ICDE 2020*. 1918–1921.

[34] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. In *Proceedings of the Artificial Intelligence Diversity, Belonging, Equity, and Inclusion, AIDBEI 2021*. 22–34.

[35] Usman Gohar, Sumon Biswas, and Hridesh Rajan. 2023. Towards understanding fairness and its composition in ensemble machine learning. In *Proceedings of the 45th IEEE/ACM International Conference on Software Engineering, ICSE 2023*. 1533–1545.

[36] Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Proceedings of the 27th European Conference on IR Research, ECIR 2005*. 345–359.

[37] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2016, NIPS 2016*. 3315–3323.

[38] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2023. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing* (2023).

[39] Max Hort and Federica Sarro. 2021. Did you do your homework? Raising awareness on software fairness and discrimination. In *Proceedings of the 36th IEEE/ACM International Conference on Automated Software Engineering, ASE 2021*. 1322–1326.

[40] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2021*. 994–1006.

[41] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2011), 1–33.

[42] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *Proceedings of the 10th IEEE International Conference on Data Mining, ICDM 2010*. 869–874.

[43] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *Proceedings of the 12th IEEE International Conference on Data Mining, ICDM 2012*. 924–929.

[44] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Science* 425 (2018), 18–33.

[45] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML/PKDD 2012*. 35–50.

[46] Fumihiro Kanei, Daiki Chiba, Kunio Hato, and Mitsuaki Akiyama. 2019. Precise and robust detection of advertising fraud. In *Proceedings of the 2019 IEEE 43rd Annual Computer Software and Applications Conference, COMPSAC 2019*. 776–785.

[47] Barbara A. Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart M. Charters, Shirley Gibbs, and Amnart Pohthong. 2017. Robust

statistical methods for empirical software engineering. *Empirical Software Engineering* 22, 2 (2017), 579–630.

[48] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 1 (2006), 25–36.

[49] Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. 2023. Dark-skin individuals are at more risk on the street: Unmasking fairness issues of autonomous driving systems. *CoRR* abs/2308.02935 (2023).

[50] Yanhui Li, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou, and Baowen Xu. 2022. Training data debugging for the fairness of machine learning software. In *Proceedings of the 44th IEEE/ACM 44th International Conference on Software Engineering, ICSE 2022*. 2215–2227.

[51] 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.

[52] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (2021), 115:1–115:35.

[53] Tim Menzies, Alex Dekhtyar, Justin Distefano, and Jeremy Greenwald. 2007. Problems with precision: A response to "comments on 'data mining static code attributes to learn defect predictors'". *IEEE Transactions on Software Engineering* 33, 9 (2007), 637–640.

[54] Rebecca Moussa and Federica Sarro. 2022. On the use of evaluation measures for defect prediction studies. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2022*. 101–113.

[55] Jerome L Myers, Arnold D Well, and Robert F Lorch Jr. 2013. *Research design and statistical analysis*. Routledge.

[56] Kewen Peng, Joymallya Chakraborty, and Tim Menzies. 2023. FairMask: Better Fairness via Model-Based Rebalancing of Protected Attributes. *IEEE Transactions on Software Engineering* 49, 4 (2023), 2426–2439.

[57] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2017, NIPS 2017*. 5680–5689.

[58] D Ramyachitra and Parasuraman Manikandan. 2014. Imbalanced dataset classification and solutions: A review. *International Journal of Computing and Business Research* 5, 4 (2014), 1–29.

[59] Federica Sarro. 2023. Search-based software engineering in the era of modern software systems. In *Proceedings of the 31st IEEE International Requirements Engineering Conference, RE 2023*.

[60] Ezekiel Soremekun, Mike Papadakis, Maxime Cordy, and Yves Le Traon. 2022. Software fairness: An analysis and survey. *arXiv preprint arXiv:2205.08809* (2022).

[61] Guanhong Tao, Weisong Sun, Tingxu Han, Chunrong Fang, and Xiangyu Zhang. 2022. RULER: Discriminative and iterative adversarial training for deep neural network fairness. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022*. 1173–1184.

[62] András Vargha and Harold D Delaney. 2000. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics* 25, 2 (2000), 101–132.

[63] Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: A trade-off revisited. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*. 8780–8789.

[64] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*. 962–970.

[65] 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, AIES 2018*. 335–340.

[66] Jie M. Zhang and Mark Harman. 2021. Ignorance and prejudice in software fairness. In *Proceedings of the 43rd IEEE/ACM International Conference on Software Engineering, ICSE 2021*. 1436–1447.

[67] Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. 2022. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* 48, 2 (2022), 1–36.

[68] Mengdi Zhang and Jun Sun. 2022. Adaptive fairness improvement based on causality analysis. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022*. 6–17.

[69] Peixin Zhang, Jingyi Wang, Jun Sun, Guoliang Dong, Xinyu Wang, Xingen Wang, Jin Song Dong, and Ting Dai. 2020. White-box fairness testing through adversarial sampling. In *Proceedings of the 42nd International Conference on Software Engineering, ICSE 2020*. 949–960.

[70] Haibin Zheng, Zhiqing Chen, Tianyu Du, Xuhong Zhang, Yao Cheng, Shouling Ji, Jingyi Wang, Yue Yu, and Jinyin Chen. 2022. NeuronFair: Interpretable white-box fairness testing through biased neuron identification. In *Proceedings of the 44th IEEE/ACM 44th International Conference on Software Engineering, ICSE 2022*. 1519–1531.