MirrorFair: Fixing Fairness Bugs in Machine Learning Software via Counterfactual Predictions

YING XIAO, Southern University of Science and Technology, China and King's College London, United Kingdom
JIE M. ZHANG*, King's College London, United Kingdom
YEPANG LIU *†, Southern University of Science and Technology, China
MOHAMMAD REZA MOUSAVI, King's College London, United Kingdom
SICEN LIU, Southern University of Science and Technology, China
DINGYUAN XUE, Southern University of Science and Technology, China

With the increasing utilization of Machine Learning (ML) software in critical domains such as employee hiring, college admission, and credit evaluation, ensuring fairness in the decision-making processes of underlying models has emerged as a paramount ethical concern. Nonetheless, existing methods for rectifying fairness issues can hardly strike a consistent trade-off between performance and fairness across diverse tasks and algorithms. Informed by the principles of counterfactual inference, this paper introduces MirrorFair, an innovative adaptive ensemble approach designed to mitigate fairness concerns. MirrorFair initially constructs a counterfactual dataset derived from the original data, training two distinct models—one on the original dataset and the other on the counterfactual dataset. Subsequently, MirrorFair adaptively combines these model predictions to generate fairer final decisions.

We conduct an extensive evaluation of MirrorFair and compare it with 15 existing methods across a diverse range of decision-making scenarios. Our findings reveal that MirrorFair outperforms all the baselines in every measurement (i.e., fairness improvement, performance preservation, and trade-off metrics). Specifically, in 93% of cases, MirrorFair surpasses the fairness and performance trade-off baseline proposed by the benchmarking tool Fairea, whereas the state-of-the-art method achieves this in only 88% of cases. Furthermore, MirrorFair consistently demonstrates its superiority across various tasks and algorithms, ranking first in balancing model performance and fairness in 83% of scenarios. To foster replicability and future research, we have made our code, data, and results openly accessible to the research community.

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

Additional Key Words and Phrases: Fairness Bugs, Bias Mitigation, Software Discrimination, Machine Learning

*Corresponding author.
†Yepang Liu is affiliated with the Research Institute of Trustworthy Autonomous Systems and the Department of Computer Science and Engineering of SUSTech.

Authors' addresses: Ying Xiao, Southern University of Science and Technology, Shenzhen, China and King's College London, London, United Kingdom, 12150075@mail.sustech.edu.cn; Jie M. Zhang, King's College London, London, United Kingdom, jie.zhang@kcl.ac.uk; Yepang Liu , Southern University of Science and Technology, Shenzhen, China, liuyp1@sustech.edu.cn; Mohammad Reza Mousavi, King's College London, London, United Kingdom, mohammad.mousavi@kcl.ac.uk; Sicen Liu, Southern University of Science and Technology, Shenzhen, China, 11910338@mail.sustech.edu.cn; Dingyuan Xue, Southern University of Science and Technology, Shenzhen, China, 11910213@mail.sustech.edu.cn.

ACM 2994-970X/2024/7-ART94
https://doi.org/10.1145/3660801

ACM Reference Format:
Ying Xiao, Jie M. Zhang, Yepang Liu, Mohammad Reza Mousavi, Sicen Liu, and Dingyuan Xue. 2024. MirrorFair: Fixing Fairness Bugs in Machine Learning Software via Counterfactual Predictions. *Proc. ACM Softw. Eng.* 1, FSE, Article 94 (July 2024), 23 pages. https://doi.org/10.1145/3660801

1 INTRODUCTION

Discrimination in machine learning software has been widely documented in various domains, such as finance [1, 17], healthcare [3, 16], and criminal justice [4, 19]. For instance, some auto-recruitment tools have been found to discriminate against women and minorities [16, 23], perpetuating existing biases and inequalities. This highlights the urgent need for fairness research. As software serves as the carrier of AI models and plays a significant role in the development of AI [7], addressing fairness bugs in ML software has become a pressing issue that requires attention from both Software Engineering and Machine Learning communities [19, 21, 25, 60, 61].

Previous research [34, 49, 60, 61] has shown that data bias can contribute significantly to fairness bugs in machine learning software. If the training data contains historical or other types of bias, the ML models may learn or even exacerbate such bias, resulting in unintended consequences such as discrimination against certain groups [49]. To address this issue, various fairness bug-fixing methods have been proposed by the ML and SE communities, including sample reweighting [35], feature value modification [14], label value modification [14], removing biased data points [17], and synthesizing minority group data points [16]. These methods attempt to modify the training data by re-balancing the distribution of sensitive attributes and labels to mitigate model bias. However, determining which data points to remove, synthesize, or mutate to improve fairness remains challenging. Furthermore, there is a typical trade-off between model performance and fairness[34], and many existing bias mitigation methods frequently lead to a considerable decline in performance. Additionally, a comprehensive empirical investigation [22] revealed that the effectiveness of existing methods varies considerably across different decision-making scenarios, influenced by tasks, datasets, models, and sensitive attributes.

To alleviate the limitations of existing bias-mitigating methods, we propose **MirrorFair**, a novel adaptive ensemble approach inspired by counterfactual inference [45, 46] and counterfactual fairness [39] to rectify fairness issues. In particular, MirrorFair constructs a mirror dataset by mutating sensitive attributes. It then trains two models from the original and mirror training sets, respectively, and makes a decision by adaptively ensembling the predictions from the two models (as shown in Figure 1). As an ensemble method, MirrorFair is novel in both what to ensemble and how to ensemble compared to the existing method MAAT [19]. Specifically, MirrorFair ensembles the original and Mirror models via counterfactual inference and adopts an adaptive ensemble strategy. On the contrary, MAAT ensembles models that are optimized for different objectives: fairness and ML performance, and simply ensembles by getting the average of two models' predictions.

To evaluate MirrorFair, we conduct a large-scale experiment using four different machine learning algorithms, including three classical algorithms and a deep neural network, on five benchmark datasets and 11 tasks. We report the effectiveness and applicability of MirrorFair across various scenarios, compare MirrorFair with 15 state-of-the-art baseline methods from ML and SE communities using 15 fairness-performance metric combinations (five performance metrics × three fairness metrics), and study the influence of ensemble strategies.

The results show that MirrorFair surpasses the Fairea Baseline in 93%/92% of scenarios with single/multiple sensitive attributes, while the state-of-the-art achieves only 88%/87%. Moreover, MirrorFair demonstrates broader superiority across various tasks and algorithms. In 83% of scenarios, it ranks first in model performance-fairness trade-off, while the state-of-the-art method achieves this in only 8% of cases in our study. To summarize, this paper makes the following contributions:

- We propose MirrorFair, a novel fairness improvement approach inspired by counterfactual fairness. MirrorFair demonstrates superior performance in mitigating bias related to single or multiple attributes across a wide range of tasks and algorithms.
- We empirically investigate the diverse impacts of flipping sensitive attribute values for all data instances across various decision-making scenarios. Additionally, we introduce an adaptive ensemble strategy to optimize the effectiveness of the ensemble method.
- We present a comprehensive evaluation of MirrorFair, comparing it with 14 state-of-the-art methods. Furthermore, we have made our code and experimental results publicly available, which facilitates the replication of our approach and the assessment of new bias-mitigation methods by fellow researchers [6].

In the remainder of this paper, we provide an overview of the background and related work in Section 2, followed by a description of the proposed MirrorFair approach in Section 3. Section 4 and Section 5 present the experimental evaluation and results, including the experimental design and a comparison with state-of-the-art bias-mitigating methods. We then discuss the findings and implications of the study in Section 6. Finally, we conclude the paper in Section 7.

2 BACKGROUND AND RELATED WORK

In this section, we introduce the background and previous research in this field.

2.1 Background

Fairness has emerged as a pressing concern in the Software Engineering (SE) domain, particularly as an increasing number of software applications incorporate machine learning models to enhance or fortify their functionality [17]. However, a significant body of research [60, 61] highlights that the datasets employed for training ML software often contain users' sensitive attributes. Sensitive attributes, synonymous with "sensitive features" and "protected attributes" in this paper, refer to individual characteristics such as gender, race, or age that may give rise to discrimination, unequal treatment, or limited opportunities [16, 19]. In machine learning classification tasks, models use provided features to make predictions or decisions [41]. However, when sensitive attributes are correlated with these predictions, these models can exhibit bias, leading to issues of group fairness where minority groups (e.g., based on gender, race, or nationality) are treated disparately [61]. Consequently, developing fair ML software represents not only a crucial ethical responsibility for engineers but also a vital prerequisite for achieving trustworthy ML software [22]. In this paper, we align with previous work [21] and focus on repairing group fairness issues.

2.2 Related Work

The growing use of machine learning software in domains such as education, healthcare, and finance has elevated fairness concerns in both computer science and social science [19, 49, 61]. In the software engineering community, there have been dedicated workshops or tracks at recent top conferences such as ICSE, ASE, and FSE to explore software fairness. Moreover, leading technology companies have established independent teams to seek solutions for providing fair and equitable AI services [9, 10, 40].

Source of bias: An imbalanced distribution of features and labels in training data can be a significant source of bias in machine learning [16, 21, 61]. Studies have shown that when certain groups are underrepresented in the training data, the resulting model may discriminate against those groups and produce unfair predictions [16]. Additionally, even if a sensitive attribute is removed from the training data, some non-sensitive attributes can act as proxies for sensitive attributes, leading to indirectly biased predictions [24, 30].

Fairness testing: Given that data and model bias are common in machine learning, there is a need for fairness testing to identify and address these issues [61]. Chen et al. [21] discuss the concept of fairness testing and clarify fairness bugs, which can lead to discriminatory outcomes. Zliobaite et al. [63] provide a comprehensive taxonomy and comparison of discrimination measurement and offer practical guidance on measuring indirect bias, which is particularly relevant for addressing discrimination caused by seemingly neutral rules or standards.

Fairness bug fixing: Recently, various fairness bug-fixing (also called bias-mitigation) methods have been proposed by different research communities. Zhang et al. [60] conduct an empirical study on five public benchmark datasets and four fairness metrics to examine the impact of feature sets and data sizes on model fairness. The study suggests that expanding the feature set can significantly improve fairness while increasing the amount of data has limited effect on mitigating model bias. Pessach and Zhang et al. [21, 49, 61] conducted comprehensive investigations of the working mechanisms of existing methods, which can be categorized into pre-processing, in-processing, and post-processing methods. Pre-processing methods focus on mitigating data bias for a fairer model, in-processing methods improve fairness during the training process, and post-processing methods correct the biased prediction by modifying the prediction result directly.

In the SE community, Fairway [17] mitigates bias by removing biased data points through a pre-training-testing operation. Fair-SMOTE [16] counteracts the removal of data points by synthesizing new data points using SMOTE [18], a data augmentation technique. Chen et al. [19] introduced MAAT, a method that ensembles models optimized for fairness and ML performance by getting the average of the two models' predictions as the final prediction. Gohar et al. [29] investigated the use of voting and bagging in ensemble learning to promote fairer predictions. Peng et al. [48] proposed FairMask, which trains extrapolation models to predict a sensitive attribute value vector and replace the original sensitive attribute values to enhance fairness. As deep neural networks (DNNs) are increasingly deployed in the software industry, many DNNs repairing techniques are proposed to fix the fairness bugs in DNNs [11, 28, 40, 42, 44, 55]. Among these, CARE [55] emerges as a notable causality-based approach that adjusts the weights of neurons to fix fairness issues within neural networks. Moreover, given the variety of available bias-mitigation methods, Zhang et al. [62] devised a technique to assist researchers and developers in selecting the most suitable bias-mitigation strategy for their specific DNNs projects.

As fairness is just one of the critical properties of machine learning software, an ideal machine learning model should also be accurate, efficient, and privacy-friendly. Therefore, improving fairness should be balanced with other properties simultaneously. However, measuring the trade-off between fairness and other properties, such as performance, can be challenging [61]. To address this, Hort et al. propose the Fairea Baseline [33], which categorizes the effectiveness of trade-offs into five levels with fairness and performance changes after applying bias-mitigating methods. Chen et al. [22] utilize Fairea Baseline to conduct a large-scale empirical study on 17 existing bias-mitigating methods in eight benchmark classification tasks with multiple performance and fairness metrics. The results show that no existing method dominates others in all the scenarios (e.g., different tasks or machine learning algorithms). Therefore, researchers and practitioners need to carefully select the most suitable method based on their expertise and specific decision-making scenarios.

3 METHODOLOGY

In this section, we describe MirrorFair in detail.

3.1 Overview of MirrorFair

Drawing inspiration from both counterfactual inference and counterfactual fairness, we propose MirrorFair. MirrorFair adaptively counteracts the unfairness and generates fairer predictions by

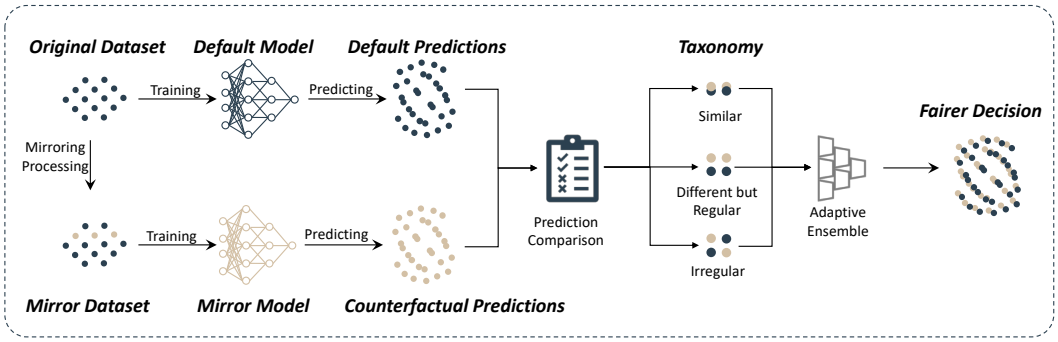

Fig. 1. Workflow of MirrorFair.

ensembling the original biased model and a mirror model trained from a pseudo mirror dataset. Figure 1 presents the workflow of MirrorFair. It begins with constructing a pseudo mirror dataset from the original training dataset and then training a mirror model from the mirror dataset. The mirror dataset is constructed by mutating (i.e., flipping) the protected attributes. We call the predictions from the original model *original predictions* and the predictions from the mirror model *counterfactual predictions*. MirrorFair then adaptively ensembles these two types of predictions and outputs the ensembled decisions.

3.2 Counterfactual Inference

Counterfactuals [45] are hypothetical scenarios proposed for events or situations that have already occurred. For example, "Bell had not invented the telephone" is a counterfactual assumption for a past event. Although not based on reality, contemplating these scenarios can help people better understand historical or real-world events.

Counterfactual inference [45, 46] is a reasoning method that involves inferring possibilities based on hypothetical scenarios. It often entails proposing a counterfactual scenario, such as "What would have happened if Bell had not invented the telephone?" and then exploring the potential outcomes and consequences by inferring from this scenario. Next, we introduce the steps of counterfactual inference. Given a known evidence W and causal model $M(U, V, F)$, where U is a set of latent background variables, V is a set of observable variables, F is a set of functions $\{f_1, \ldots, f_n\}$, counterfactual inference is the computation of probabilities $P(Y_{Z \leftarrow z}(U)|W = w)$, where W, Z and Y are subsets of V. Inference proceeds in three steps: (1) Abduction: for a given prior on U, compute the posterior distribution of U given the evidence $W = w$; (2) Action: Substitute the equations for Z with the interventional values z, resulting in the modified set of equations F_z; (3) Prediction: compute the implied distribution on the remaining elements of V using F_z and the posterior $P(U|W = w)$.

Counterfactual inference is commonly used to address causal inference problems, such as estimating the causal effect of an event, in fields such as statistics and machine learning to estimate the effect of a potential action [39]. Counterfactual fairness is a practice of counterfactual inference, which requires that the change in the sensitive attribute value should not change the original decision in a decision-making task [39]. Inspired by counterfactual inference and counterfactual fairness, we make a counterfactual hypothesis to mitigate bias by changing the contribution of sensitive attributes to model prediction. Instead of conducting all three steps of counterfactual inference, we only construct the counterfactuals, which we implement an experiment upon to explore the bias effect of sensitive attributes and improve the fairness of machine learning software. We describe notions of MirrorFair and how to construct counterfactuals in the following sections.

3.3 Mirroring Processing

The construction of counterfactuals involves modifying the value of the sensitive attribute while maintaining the values of other features. When dealing with multi-valued sensitive attributes such as those found in the "race" category, which includes values like "White", "Asian", and "Eskimo", determining which value to target for modification poses a challenge. To support fairness research, IBM introduced AIF360 [8], offering comprehensive APIs for accessing processed benchmark datasets and computing fairness metrics. AIF360 simplifies sensitive attributes into binary form; for instance, the "race" attribute in the Adult dataset is transformed to "White" and "non-White". Although this simplification might obscure distinctions between sub-groups, it considerably eases the complexity of modeling fairness issues and facilitates the construction of counterfactuals. Thus, in line with previous work [19, 22, 48], we use AIF360 to simplify multi-valued sensitive attributes into binary form for our approach implementation. This allows us to create counterfactuals — a virtual training dataset where the sensitive attribute is inverted from the original dataset, but all other feature values remain unchanged. The counterfactual is constructed by flipping the sensitive attribute values (e.g., changing "White" to "non-White") while keeping the values of other features constant. An illustration of mirroring an instance for the Adult-Race task is provided below.

$$x(White, f_1, f_2, f_3, ...) \longrightarrow x'(non\text{-}White, f_1, f_2, f_3, ...)$$

Where x denotes an original data instance; x' denotes such instance after mirroring processing; $White$ and $non\text{-}White$ denote two values of race attribute; f_1, f_2, f_3 denote values of the rest features. We refer to the operation of flipping the sensitive attribute to construct a new virtual training dataset as the **mirroring processing**, which resembles Plane Mirror Imaging (PMI) [27] that produces an upright, same-sized, and laterally inverted virtual image of an object. The new counterfactual training dataset is called the *mirror dataset*, the model trained on the mirror dataset is called the *mirror model*, and the predictions from mirror mode are called *mirror prediction* as well as *counterfactual prediction*.

3.4 Adaptive Ensemble Strategy

Prior work demonstrates data bias makes a significant contribution to the model discrimination [16, 17]. Zhang et al. propose the Causal Explanation Formula [59] and point out that the data bias consists of the counterfactual direct bias effect (Ctf-DE), counterfactual indirect bias effect (Ctf-IE) and counterfactual spurious bias effect (Ctf-SE) caused by sensitive attributes, mediators of sensitive attributes and the confounders respectively. The total variation, counterfactual direct bias effect, counterfactual indirect bias effect, and counterfactual spurious bias effect obey the following relationship:

$$TV_{x_0,x_1}(y) = DE_{x_0,x_1}(y|x_0) - IE_{x_0,x_1}(y|x_0) - SE_{x_0,x_1}(y|x_0) \tag{1}$$

where $TV_{x_0,x_1}(y)$ denotes the total variation measuring the demographic parity, which is a popular fairness metric adopted by many previous fairness research [19, 59]. $DE_{x_0,x_1}(y|x_0)$ denotes the causal effect difference between the sensitive attribute values x_0 and x_1; similarly,$IE_{x_0,x_1}(y|x_0)$ and $SE_{x_0,x_1}(y|x_0)$ denote the causal effect difference between the different mediator values and confounders values.

Chen et al. [21] demonstrate the effectiveness of enhancing fairness via modifying the sensitive attribute value distribution. The causal explanation formula enhances the reasonability and explainability of such techniques, and the formula further points out the efficacy of improving fairness by modifying sensitive attributes can be affected by the mediators and confounders while simply changing sensitive attributes can hardly keep high efficacy across all decision-making scenarios

(decision tasks × classifiers). In Section 5, we present the empirical investigation finding that "flipping" (mirroring processing) sensitive attribute values make different impacts on model prediction across various decision-making scenarios. To handle this situation, we propose a taxonomy to categorize different decision-making scenarios and design adaptive ensemble strategies to meet the requirement of enhancing fairness across different decision-making scenarios.

3.4.1 Taxonomy of Decision-making Scenarios. For clarity in this paper, we define the event of employing an algorithm to accomplish a decision-making task as a "decision-making scenario" (e.g., using LR for the Compas-Sex task). Given a default model M_{def} trained from the original dataset, a mirror model M_{mir} trained from the mirror dataset, a testing dataset D_{test}, and a decision-making scenario, the difference of two models on the given testing instance can be calculated as:

$$DIF_{d_{A=a}} = P_{def}(\hat{Y} = y | d_{A=a}) - P_{mir}(\hat{Y} = y | d_{A=a}) \tag{2}$$

where $\hat{Y}$ denotes the predictive output label (e.g., income), y denotes a label value (e.g., high income or low income); A denotes the sensitive attribute (e.g., sex); a denotes a sensitive attributes value (e.g., female or male); $DIF_{d_{A=a}}$ denotes the probability difference between the two models on the same test instance and same predictive label; $P_{def}(\hat{Y} = y | d_{A=a})$ denotes the probability of default model; $P_{mir}(\hat{Y} = y | d_{A=a})$ denotes the probability of mirror model.

Based on the characteristics of DIF near the decision boundary, we classify the decision-making tasks into three categories: mirror-insensitive, mirror-regular, and mirror-irregular. To elucidate our classification scheme, we employ the mathematical notion of a "neighborhood". A neighborhood of a point X is defined as the set $\mathcal{N}^{\delta}(X)$, comprising all points y for which $d(x, y) < \delta$, where δ denotes the radius of the neighborhood, i.e., $\mathcal{N}^{\delta}(X) := \{y \in X : d(x, y) < \delta\}$ [52]. With the concept of neighborhood, we categorize the three scenarios as follows:

$$S_{type} = \begin{cases} \text{mirror-insensitive} & \text{if } \forall d_{A=a} \in D_{test}, |DIF_{d_{A=a}}| \in \mathcal{N}^{\delta}(0) \\ \text{mirror-regular} & \text{if } \forall d_{A=a} \in D_{test}, |DIF_{d_{A=a}}| \in \mathcal{N}^{\delta}(c), c \neq 0 \\ \text{mirror-irregular} & \text{if } \exists d_{A=a} \in D_{test}, |DIF_{d_{A=a}}| \notin \mathcal{N}^{\delta}(c) \end{cases}$$

where S_{type} denotes decision-making scenario type, D_{test} denotes the testing dataset, $d_{A=a}$ denotes a testing instance, $\mathcal{N}(0)$ and $\mathcal{N}(c)$ denote the neighborhood of 0 and c, where c represents a constant value. In this paper, we set the radius δ to 0.05, a setting that distinctly differentiates the three types of scenarios. In the following, we introduce each decision-making scenario type in detail. Actual examples of each decision-making scenario and further analysis can be found in Section 5.4.1.

Mirror-insensitive scenario: This scenario happens when the default model makes almost identical predictions with the mirror model. This means that the mirroring processing has a tiny effect on the prediction of such scenarios. In other words, this type of decision-making is insensitive to mirroring processing. Therefore, we regard such a scenario as a mirror-insensitive scenario.

Mirror-regular scenario: This scenario refers to the case where the default model makes different predictions from the mirror model, but the absolute probability difference between the two models is regular and close to a constant value. That is, mirroring processing has a significant and regular effect on each prediction of such scenarios. Therefore, we regard such a scenario as a mirror-regular scenario.

Mirror-irregular scenario: This scenario refers to the case that the default model has different predictions from the mirror model, and the absolute probability difference between the two models is irregular for different test instances. The mirroring processing has a significant, irregular, and uncertain effect on the prediction of such scenarios. Therefore, we regard it as a mirror-irregular scenario.

3.4.2 Ensemble Strategies. To adaptively handle different decision-making scenarios, we propose two strategies **E-Mean** and **E-Max** to ensemble the prediction of the default model and mirror model to repair the fairness issue in machine learning software. Regarding mirror-regular scenarios, in which mirroring processing makes a significant and regular contribution to model predictions, we ensemble the two predictions via weighted averaging. The final output probability vector with the E-Mean strategy can be calculated by:

$$P_{final} = \left[\frac{P_{def}(Y=0) + P_{mir}(Y=0)}{2}, \frac{P_{def}(Y=1) + P_{mir}(Y=1)}{2} \right] \tag{3}$$

where $P_{def}(Y=0)$ denotes the output probability of class "0" by default model; $P_{def}(Y=1)$ denotes the output probability of class "1" by default model; $P_{mir}(Y=0)$ denotes the output probability of class "0" by mirror model; $P_{mir}(Y=0)$ denotes the output probability of class "1" by mirror model.

In terms of other decision-making scenarios, we repair the fairness issue in machine learning software by maximizing the favorable label probability of the unprivileged group near the decision boundary ($0.45 < P(Y=1) < 0.55$), because the previous work [36] points out most discrimination happens near the decision boundary. Maximizing the favorable label probability for the unprivileged instances within the boundary can calibrate some of the final predictions for unprivileged instances that deserve a favorable label, thereby improving fairness [19, 36]. Regarding the mirror-insensitive scenario, if the test instance belongs to an unprivileged group and the output class probability is near the decision boundary, we post-process the output to be a favorable label while the prediction out of the boundary still follows the E-Mean strategy. For the mirror-irregular scenario, if the test instance belongs to an unprivileged group and the output class probability is also near the decision boundary, the final output probability vector with the E-Max strategy can be calculated by:

$$P_{final} = \left[min(P_{def}(Y=0), P_{mir}(Y=0)), max(P_{def}(Y=1), P_{mir}(Y=1)) \right] \tag{4}$$

and the prediction out of the boundary still follows the weighted averaging ensemble strategy.

3.5 Protection of Multiple Sensitive Attributes

In machine learning prediction tasks, the feature set of the data can contain more than one sensitive feature that needs to be protected. For instance, there are sex and race sensitive features in both the Adult dataset and the Compas dataset. Covering multiple sensitive features protection is an important evaluation dimension for assessing a bias-mitigating approach. As an ensemble approach, MirrorFair combines the prediction of the default and mirror model to improve machine learning fairness. MirrorFair is uniquely advantageous for protecting multiple sensitive attributes due to its flexibility in combining strategies. In the case of protecting sex and race attributes in the Adult Census Income task, there are two deployment strategies that can improve fairness:

- Strategy 1 (MirrorMulti-S1): Select both sex and race as mirror features simultaneously and create a sex-race mirror training dataset by reversing both sex and race values in the original training dataset. Then, combine the predictions of both models as if protecting a single sensitive feature;
- Strategy 2 (MirrorMulti-S2): Select sex and race as the mirror features and produce separate sex mirror and race mirror training datasets. Then, combine the predictions from the default model, sex-mirror model, and race-mirror model to generate the final prediction. This strategy provides more flexibility in controlling the weight of each sensitive feature in the final prediction and allows for different weight settings for each feature to achieve a more fine-grained fairness level.

In Section 4, we evaluate the effectiveness of MirrorFair in balancing model performance and fairness in protecting single and multiple sensitive feature scenarios.

4 EXPERIMENTAL DESIGN

Here, we introduce our research questions and the experimental design for evaluating MirrorFair.

4.1 Research Questions

We evaluate MirrorFair by exploring the following research questions.

- **RQ1: Efficacy of MirrorFair:** To what extent can MirrorFair achieve mitigating model bias without losing too much performance? We conduct a comprehensive comparison between MirrorFair and existing bias-mitigating methods across different decision-making scenarios.
- **RQ2: Applicability and versatility:** To what extent can MirrorFair achieve the consistency in maintaining the efficacy? We design two experimental settings to explore the applicability and versatility of MirrorFair across different tasks and algorithms and compare MirrorFair with state-of-the-art methods.
- **RQ3: Effectiveness in mitigating multiple attributes biases:** To what extent can MirrorFair mitigate multiple sensitive attribute biases simultaneously? This research question compares the effectiveness of MirrorFair with that of existing methods in multiple sensitive attribute scenarios.
- **RQ4: Impact of mirroring processing and effectiveness of adaptive strategies:** What impact does the mirroring processing have on model predictions, and how effectively can adaptive ensemble strategies achieve? In this research question, we conduct an empirical investigation across various decision tasks and machine learning algorithms to explore the diverse effects of mirroring processing on model predictions. Subsequently, we present the results of adaptive ensemble strategies alongside those of fixed ensemble strategies to highlight the advantages of adaptive ensemble strategies.

4.2 Benchmark Datasets and Tasks

In order to ensure the reliability of the evaluation, we align our experimental setups with the recent empirical investigation [22] and adopt the same benchmarking dataset and machine learning algorithms to implement the experiments and comparison with existing bias mitigation approaches. The five public benchmark datasets come from diverse domains, including the Adult Income dataset [5] (a.k.a., Adult dataset), ProPublica Recidivism dataset [4] (a.k.a., Compas dataset), German Credit dataset [1] (a.k.a., German dataset), Bank Marketing dataset [2] (a.k.a., Bank dataset), and Medical Survey 2015 dataset [3] (a.k.a., Mep dataset). These datasets are commonly used in machine learning fairness studies due to their relevance to individual benefits and opportunities such as college admission and recruitment. Briefly, the Adult dataset contains information about individuals' demographic, social, and economic factors to predict whether their income is above or below a certain threshold. The Compas dataset contains information about individuals who were assessed for their likelihood of committing future crimes. The German dataset contains information about individuals' creditworthiness to predict whether they are likely to default on a loan. The Bank dataset contains information about individuals' financial attributes to predict whether they will subscribe to a term deposit. Finally, the Mep dataset contains information about individuals' health behaviors and outcomes.

Notably, literature [19, 22, 33] points out that 90% of the fairness research does not use more than three datasets, to address this, we adopt eleven fairness testing tasks in five datasets to evaluate the effectiveness of MirrorFair and existing methods comprehensively. Tasks 1-8 are single attribute tasks, where we mitigate a single sensitive attribute bias such as race or gender. Tasks 9-11 are multip-attribute tasks, where we mitigate multiple sensitive attribute biases simultaneously. Table 1 provides the details of each task and the corresponding dataset used. To comprehensively evaluate

Table 1. Benchmark datasets and tasks.

Task	Protected attribute(s)	Dataset	Size	Favourable label	Majority label
1. Adult-sex	Sex	Adult	45,222	1 (income > 50k)	0 (75.2%)
2. Adult-race	Race	Adult	45,222	1 (income > 50k)	0 (75.2%)
3. Compas-sex	Sex	Compas	6,167	0 (no recidivism)	0 (54.5%)
4. Compas-race	Race	Compas	6,167	0 (no recidivism)	0 (54.5%)
5. German-sex	Sex	German	1,000	1 (good credit)	1 (70.0%)
6. German-age	Age	German	1,000	1 (good credit)	1 (70.0%)
7. Bank-age	Age	Bank	30,488	1 (subscriber)	0 (87.3%)
8. Mep-race	Race	Mep	15,830	1 (utilizer)	0 (82.8%)
9. Adult-sex-race	Sex, Race	Adult	45,222	1 (income > 50k)	0 (75.2%)
10. Compas-sex-race	Sex, Race	Compas	6,167	0 (no recidivism)	0 (54.5%)
11. German-sex-age	Sex, Age	German	1,000	1 (good credit)	1 (70.0%)

the effectiveness of the proposed methods, we adopt four classifiers, including logistic regression (LR) [38], random forest (RF) [12], support vector machine (SVM) [32], and deep learning classifiers (DNN) [41].

4.3 Metrics and Measurements

We follow previous research [19, 20, 22] using the most popular five performance metrics and three fairness to measure the effectiveness of MirrorFair and existing methods by the state-of-the-art fairness and performance measurement tool Fairea. Next, We describe the performance metrics, fairness metrics, and Fairea we leveraged in this paper.

4.3.1 Performance Metrics. Precision and recall are significant metrics to evaluate machine learning classifiers' performance. Precision reflects the ability of the classifier to predict sample classes correctly; recall reflects the ability of the classifier to find out each class completely. They are calculated as follows:

$$Precision@c = Pr[Y = c|\hat{Y} = c] = \frac{TP}{TP + FP} \tag{5}$$

$$Recall@c = Pr[\hat{Y} = c|Y = c] = \frac{TP}{TP + FN} \tag{6}$$

where $Precision@c$ denotes the precision on class c; $Recall@c$ denotes the recall on class c; True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) denote numbers of favorable samples predicted as favorable, unfavorable samples predicted as unfavorable, unfavorable samples predicted as favorable, and favorable samples predicted as unfavorable, respectively. As both precision and recall reflect a single dimension of the model's performance, F1-score and accuracy are widely adopted to evaluate the overall performance. They are calculated as follows:

$$F1@c = \frac{2 \times Precision@c \times Recall@c}{Precision@c + Recall@c} \tag{7}$$

$$Acc = Pr[\hat{Y} = Y] = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

where $F1@c$ denotes the F1-score on class c. Additionally, as a model can easily obtain high Acc in extremely biased datasets by making all predictions as majority label, Rodriguez et al. [53] propose the Matthews Correlation Coefficient (MCC) metric, which is calculated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

In terms of precision, recall, and F1-score, We follow previous research using Macro-Precison, Macro-Recall, and Macro-F1-score to represent such metrics among all classes.

4.3.2 Fairness Metrics. There are various fairness metrics, according to different definitions. We follow previous research [20, 22] leveraging the Statistical Parity Difference (SPD), Average Odds Difference (AOD), and Equal Opportunity Difference (EOD) to measure the fairness of the machine learning models, which are mitigated bias by MirrorFair or existing methods.

SPD reflects the difference between privileged and unprivileged groups predicted as good labels by the models, which is calculated as follows:

$$SPD = Pr[\hat{Y} = 1|A = 0] - Pr[\hat{Y} = 1|A = 1] \tag{10}$$

AOD reflects the difference between the false positive rate and the true positive rate of the privileged and unprivileged groups, it is calculated as follows:

$$\begin{aligned} AOD = \frac{1}{2}(&|Pr[\hat{Y} = 1|A = 0, Y = 0] - Pr[\hat{Y} = 1|A = 1, Y = 0]| \\ &+|Pr[\hat{Y} = 1|A = 0, Y = 1] - Pr[\hat{Y} = 1|A = 1, Y = 1]|) \end{aligned} \tag{11}$$

EOD reflects the difference between the privileged group and the unprivileged group in terms of good labels predicted by the models, it is calculated as follows:

$$EOD = Pr[\hat{Y} = 1|A = 0, Y = 1] - Pr[\hat{Y} = 1|A = 1, Y = 1] \tag{12}$$

where $\hat{Y}$ denotes the prediction of models, Y denotes the model label, A demotes attribute, and Pr denotes the proportion respectively.

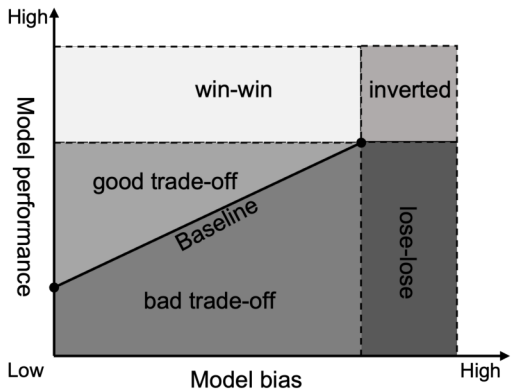

Fig. 2. The performance and fairness trade-off effectiveness regions categorized by the Fairea baseline.

4.3.3 Fairea Baseline. Due to the variety of performance and fairness metrics in machine learning models, evaluating a bias-mitigating method based on a single metric can be one-sided. To address this issue, Hort et al. propose Fairea, a comprehensive trade-off baseline that balances both model performance and fairness. As shown in Figure 2, Fairea Baseline categorizes the performance-bias Cartesian coordinate system into five regions, including "win-win" (increasing both fairness and performance), "lose-lose" (decreasing both fairness and performance), and "inverted" (increasing performance but decreasing fairness). Fairea Baseline further divides cases that increase fairness but decrease performance into "good trade-off" and "bad trade-off." As recent studies [19, 22] have adopted Fairea to evaluate the effectiveness of bias-mitigating methods, we align with them and use Fairea with the same 15 (each fairness metric combining any performance metrics can generate a trade-off baseline, three fairness × five performance metrics = 15 baseline) trade-off baselines to evaluate the effectiveness of MirrorFair and existing methods.

Table 2. Existing methods and description.

Method	Type	Venue/Journal	Description
Optimized Pre-processing (OP) [14]	Pre-processing	NeurIPS	Modify data features and labels.
Learning Fair Representation (LFR) [57]	Pre-processing	ICML	Obfuscating information about sensitive attributes
Reweighting (RW) [35]	Pre-processing	KAIS	Set different weights for samples in different groups.
Disparate Impact Remover (DIR) [26]	Pre-processing	SIGKDD	Modify data feature values.
Fairway [17]	Pre-processing	ESEC/FSE	Remove ambiguous data points.
Fair-SMOTE [16]	Pre-processing	ESEC/FSE	Remove ambiguous data points and synthesize new data points.
FairMask [48]	Pre-processing	TSE	Replace the sensitive attribute vector of testing data.
MAAT [19]	Pre-Post-processing	ESEC/FSE	Ensemble prediction of fairness model and performance model.
Prejudice Remover (PR) [37]	In-processing	ECML-PKDD	Add a fair regularization term to the learning objective.
Adversarial Debiasing (AD) [58]	In-processing	AAAI	Reduce the contribution of protected attributes to prediction.
Meta Fair Classifier (MFC) [15]	In-processing	FAT	Optimize classifier with fairness metrics.
CARE [50, 55]	Post-processing	ICSE	Using causality analysis to modify neurons weights.
Reject Option Classification (ROC) [36]	Post-processing	ICDM	Modify prediction near the threshold.
Equalized Odds Post-processing (EOP) [31]	Post-processing	NeurIPS	Modify predictions to make the Odds Difference equal.
Calibrated Equalized Odds Post-processing (CEO) [51]	Post-processing	NeurIPS	Modify predictions with calibrated probability.

4.4 Baseline Methods

To ensure the reliability of our experiments, we have carefully selected a set of state-of-the-art methods from different communities as benchmark methods to compare with MirrorFair. Our selection includes ten bias mitigation methods integrated into the AIF 360 toolbox [8], as well as five advanced methods [16, 17, 19, 48, 55] proposed in software engineering. This large-scale comparison exceeds the scale of some previous empirical studies [22]. Table 2 provides the names, sources, types, and brief descriptions of each selected baseline method. We compare different methods by checking their frequency of improving fairness, maintaining performance, and surpassing the Fairea trade-off baseline.

4.5 Experimental Design

In this section, we provide the experimental details for replicating our work. To mitigate personal bias, we used AIF360 [8], Scikit-learn [13, 47], and TensorFlow Keras [56] for implementing existing methods, machine learning algorithms, and evaluation metrics. AIF360 is a fairness research and testing tool developed by IBM that includes current state-of-the-art bias-mitigating methods and all fairness metrics we leveraged. We performed large-scale experiments across various scenarios, each replicated 50 times and taking the average as the final result to minimize random errors. We use different random seeds to split the dataset into 70% training data and 30% testing data each time in the experiment without involving cross-validation. All of these experimental and model parameter settings align with previous research [19, 22], ensuring the soundness of experiments and fairness of the comparison among MirrorFair and existing methods. All experiments were carried out on a Linux server running Ubuntu 20.04 focal with 256 GB RAM, 3.7 GHz Intel Xeon Gold 6238, Python 3.8.16, and TensorFlow 2.11.1.

5 RESULTS

This section presents the experimental results to answer our four research questions.

5.1 RQ1: Efficacy of MirrorFair

This RQ aims to assess the efficacy of MirrorFair and the existing methods in mitigating a single sensitive attribute bias. To achieve this, we design two evaluation settings to explore the impact of MirrorFair on model performance and fairness, as well as the priority of MirrorFair against the 15 existing methods. The evaluation is based on the results of 50 repetitions of each bias-mitigating method in different decision-making scenarios (tasks × classification algorithms).

5.1.1 Impact of MirrorFair on Model Performance and Fairness. Table 3 presents the detailed performance metrics (accuracy, recall, precision, and f1-score) and fairness metrics (SPD, AOD,

Table 3. (RQ1) Detailed performance and fairness metrics of the models before and after applying MirrorFair. "Default" means not applying any bias-mitigating method, and "MirrorFair" means applying MirrorFair to mitigate bias. "(+)" means higher value of the metrics is better and "(-)" means lower value of the metrics is better. Each metric value is the average of 50 times repetitions.

Task	Method	LR							SVM						
		Accuracy (+)	Recall (+)	Precision (+)	F1-Score (+)	SPD (-)	AOD (-)	EOD (-)	Accuracy (+)	Recall (+)	Precision (+)	F1-Score (+)	SPD (-)	AOD (-)	EOD (-)
Adult-Sex	Default	0.85	0.76	0.80	0.78	0.19	0.10	0.12	0.85	0.76	0.81	0.78	0.18	0.08	0.09
	MirrorFair	0.84	0.74	0.81	0.76	**0.11**	**0.04**	**0.05**	0.84	0.74	0.81	0.76	**0.11**	**0.05**	**0.07**
Adult-Race	Default	0.85	0.76	0.80	0.78	0.10	0.06	0.09	0.85	0.76	0.81	0.78	0.10	0.05	0.07
	MirrorFair	0.85	0.76	0.80	0.78	**0.07**	**0.02**	**0.02**	0.85	0.76	0.81	0.78	**0.07**	**0.02**	**0.02**
Compas-Sex	Default	0.67	0.66	0.67	0.66	0.28	0.25	0.20	0.66	0.66	0.66	0.66	0.26	0.24	0.18
	MirrorFair	0.67	0.65	0.67	0.65	**0.12**	**0.10**	**0.06**	0.66	0.65	0.67	0.64	**0.11**	**0.08**	**0.04**
Compas-Race	Default	0.67	0.66	0.67	0.66	0.18	0.16	0.11	0.66	0.66	0.66	0.66	0.17	0.15	0.10
	MirrorFair	0.66	0.65	0.67	0.65	**0.06**	**0.05**	**0.02**	0.66	0.64	0.67	0.64	**0.05**	**0.04**	**0.02**
German-Sex	Default	0.75	0.67	0.70	0.68	0.11	0.10	0.07	0.75	0.67	0.70	0.68	0.11	0.10	0.07
	MirrorFair	0.74	0.65	0.69	0.66	**0.05**	**0.08**	**0.04**	0.74	0.63	0.70	0.64	**0.05**	**0.08**	**0.04**
German-Age	Default	0.75	0.67	0.70	0.68	0.21	0.17	0.16	0.75	0.67	0.70	0.68	0.20	0.17	0.16
	MirrorFair	0.74	0.67	0.70	0.68	**0.05**	**0.07**	**0.05**	0.75	0.64	0.71	0.65	**0.05**	**0.10**	**0.05**
Bank-Age	Default	0.90	0.68	0.79	0.72	0.09	0.08	0.13	0.90	0.67	0.79	0.71	0.07	0.05	0.08
	MirrorFair	0.90	0.69	0.79	0.73	**0.05**	**0.03**	**0.04**	0.90	0.70	0.79	0.73	**0.05**	**0.03**	**0.04**
Mep-Race	Default	0.86	0.68	0.78	0.71	0.12	0.11	0.18	0.86	0.67	0.78	0.70	0.10	0.08	0.12
	MirrorFair	0.86	0.67	0.78	0.71	**0.08**	**0.05**	**0.07**	0.86	0.67	0.78	0.70	**0.07**	**0.03**	**0.05**

Task	Method	RF							DNN						
		Accuracy (+)	Recall (+)	Precision (+)	F1-Score (+)	SPD (-)	AOD (-)	EOD (-)	Accuracy (+)	Recall (+)	Precision (+)	F1-Score (+)	SPD (-)	AOD (-)	EOD (-)
Adult-Sex	Default	0.84	0.77	0.79	0.78	0.19	0.08	0.08	0.85	0.77	0.80	0.78	0.18	0.08	0.08
	MirrorFair	0.84	0.77	0.79	0.78	**0.16**	**0.04**	**0.02**	0.85	0.75	0.81	0.77	**0.13**	**0.04**	**0.04**
Adult-Race	Default	0.84	0.77	0.79	0.78	0.10	0.05	0.04	0.85	0.77	0.80	0.78	0.09	0.04	0.05
	MirrorFair	0.85	0.77	0.80	0.79	**0.06**	**0.03**	**0.04**	0.85	0.77	0.80	0.78	**0.07**	**0.02**	**0.02**
Compas-Sex	Default	0.65	0.64	0.64	0.64	0.17	0.14	0.12	0.65	0.65	0.65	0.65	0.19	0.16	0.13
	MirrorFair	0.66	0.64	0.66	0.64	**0.03**	**0.03**	**0.03**	0.66	0.65	0.65	0.65	**0.14**	**0.11**	**0.09**
Compas-Race	Default	0.65	0.64	0.64	0.64	0.14	0.12	0.09	0.65	0.65	0.65	0.65	0.16	0.14	0.10
	MirrorFair	0.65	0.64	0.65	0.64	**0.04**	**0.02**	**0.02**	0.65	0.64	0.65	0.64	**0.06**	**0.04**	**0.03**
German-Sex	Default	0.76	0.66	0.73	0.67	0.07	0.07	0.04	0.73	0.66	0.68	0.67	0.09	0.09	0.06
	MirrorFair	0.76	0.64	0.73	0.66	**0.04**	**0.05**	**0.03**	0.74	0.65	0.70	0.66	**0.07**	**0.07**	**0.05**
German-Age	Default	0.76	0.66	0.73	0.67	0.13	0.11	0.07	0.73	0.66	0.68	0.67	0.19	0.16	0.15
	MirrorFair	0.76	0.65	0.73	0.66	**0.05**	**0.07**	**0.04**	0.74	0.65	0.69	0.66	**0.08**	**0.08**	**0.06**
Bank-Age	Default	0.90	0.72	0.79	0.75	0.08	0.05	0.06	0.90	0.75	0.77	0.76	0.09	0.06	0.07
	MirrorFair	0.90	0.73	0.79	0.75	**0.06**	**0.04**	**0.05**	0.90	0.77	0.78	0.77	0.09	**0.05**	**0.06**
Mep-Race	Default	0.86	0.67	0.76	0.70	0.09	0.07	0.09	0.85	0.67	0.74	0.69	0.11	0.09	0.13
	MirrorFair	0.86	0.68	0.75	0.71	**0.06**	**0.02**	**0.02**	0.86	0.68	0.75	0.70	**0.08**	**0.05**	**0.06**

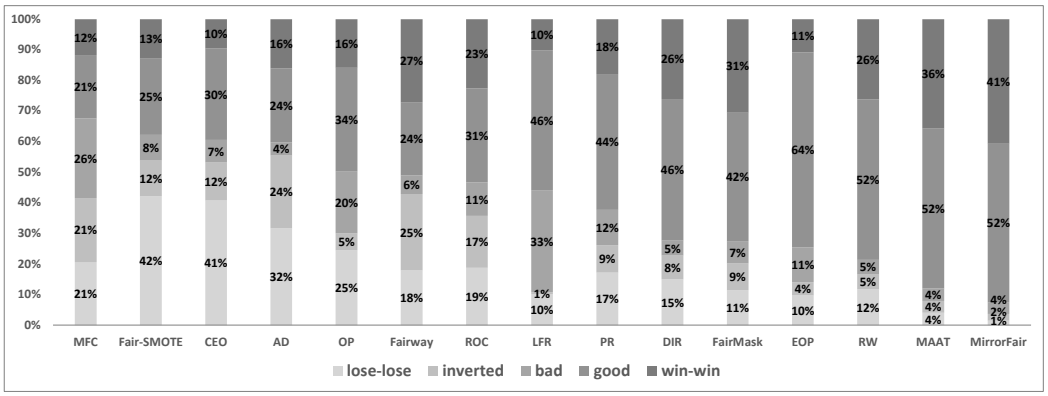

Fig. 3. RQ1: Trade-off Effectiveness distributions in mitigating single sensitive attribute bias of different bias-mitigating methods. Overall, MirrorFair achieves the best trade-off, with 93% of the mitigation cases falling in the "good" or "win-win" trade-off regions.

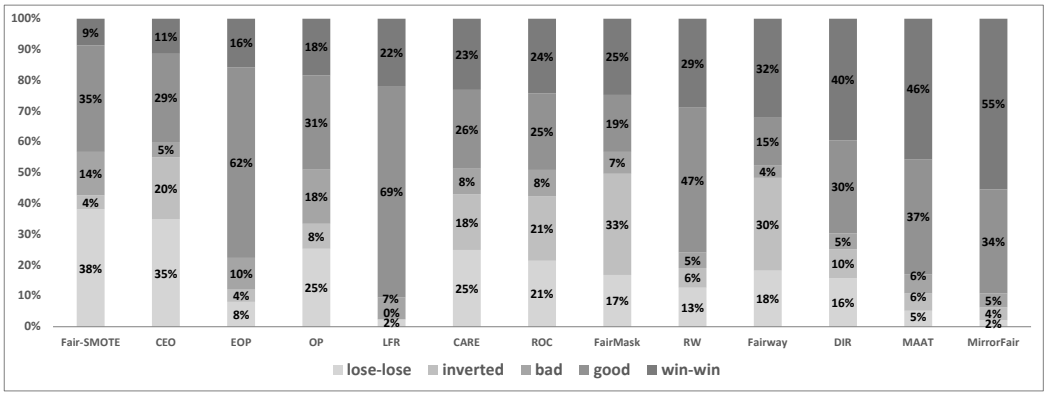

Fig. 4. RQ1: Trade-off Effectiveness distributions in mitigating single sensitive attribute bias of different bias-mitigating methods for DNN only.

and EOD) values of default models and MirrorFair models in each decision-making scenario (task × algorithms). We can see that in most cases, after applying MirrorFair, the overall prediction performance is maintained, and the bias is evidently reduced. In some scenarios, such as Mep-Race-DNN, MirrorFair improves all four performance metrics and reduces all three bias metrics. The cases that enhance fairness but reduce performance are rare, and the most significant drop in model performance occurs in the German-Sex-SVM scenario. When significantly enhancing model fairness, MirrorFair increases model precision by one percent but simultaneously decreases model recall and f1-score by three percent.

5.1.2 Superiority of MirrorFair against Existing Bias-mitigating Methods. We investigate the superiority of MirrorFair from three aspects: the superiority in fairness-performance trade-offs, the superiority in overall fairness and performance improvement, and the superiority in significant fairness improvement.

Superiority in fairness-performance trade-offs: Following the state-of-the-art method [19] and recent empirical investigation [22], we use the Fairea benchmark to assess MirrorFair and other existing methods. Fairea quantifies the effectiveness of trade-offs by calculating the proportion of cases where both performance and fairness are increased or decreased among the experimental cases [33]. Figure 3 shows the effectiveness distribution ("win-win", "good", "bad", "inverted" and "lose-lose") of each bias mitigating method, which implies the proportion of cases with improved fairness ("win-win" + "good" + "bad"), decreased performance ("lose-lose" + "good" + "bad"), and surpassed the Fairea baseline ("win-win" + "good"). The comparison is based on 1,600 cases for each bias-mitigating method.

Figure 3 demonstrates that MirrorFair achieves the highest "win-win" proportion, simultaneously enhancing model performance and fairness in 41% of all the cases, while the state-of-the-art method MAAT only realizes a 36% "win-win" proportion of cases. According to Fairea [19, 21, 33], MirrorFair surpasses the Fairea trade-off baseline in 93% of cases, which is five percent points higher than the state-of-the-art method MAAT and much far leads the other. Consequently, MirrorFair outperforms all the existing methods that we studied in achieving good performance and fairness trade-off.

We also compare MirrorFair with the state-of-the-art DNN fairness repairing technique CARE [50, 55], which is specially designed for repairing the fairness bugs in deep neural networks (DNNs). Figure 4 displays the comparison results, together with other baseline methods. It turns out MirrorFair still has the best effectiveness on DNNs only when taking CARE into consideration.

Table 4. (RQ1) Impact of MirrorFair and state-of-the-art methods on model performance and bias. MirrorFair achieves the highest bias reduction with the lowest performance compromise.

Method	Accuracy	Recall	Precision	F1-Score	Overall Performance	SPD	AOD	EOD	Overall Bias
FairMask	-0.23%	-0.94%	-1.10%	-1.51%	-1.48%	-15.53%	-19.18%	-20.24%	-18.32%
DIR	-0.39%	-0.75%	-0.37%	-0.94%	-0.61%	-18.94%	-17.54%	-18.95%	-18.47%
RW	-0.32%	-0.78%	-0.37%	-0.71%	-0.54%	-49.24%	-37.32%	-26.25%	-37.60%
MAAT	0.03%	-1.33%	0.83%	-1.01%	-0.37%	-37.02%	-43.44%	-42.51%	-40.99%
EOP	-2.39%	-3.00%	-3.55%	-3.61%	-3.14%	-44.54%	-47.42%	-44.33%	-45.43%
MirrorFair	0.03%	-0.46%	0.47%	-0.54%	**-0.12%**	-44.69%	-51.93%	-55.43%	**-50.68%**

Table 5. (RQ1) Proportions of cases where each method significantly improves fairness compared with other methods. Compared with the default model, MirrorFair significantly improves fairness in 99% of the cases, and compared with MAAT, MirrorFair significantly improves fairness in 44% of the cases.

Method	VS Default	VS FairMask	VS DIR	VS EOP	VS RW	VS MAAT	VS MirrorFair
Default	0%	9%	13%	4%	7%	0%	0%
FairMask	61%	0%	24%	20%	18%	7%	0%
DIR	71%	55%	0%	22%	27%	20%	10%
EOP	77%	65%	54%	0%	29%	44%	42%
RW	80%	54%	44%	21%	0%	40%	34%
MAAT	92%	65%	51%	31%	38%	0%	16%
MirrorFair	99%	92%	67%	34%	45%	44%	0%

Superiority in overall fairness and performance improvement: Table 4 presents the overall impact of the five state-of-the-art methods that perform well in trade-off model fairness and performance based on Fairea (according to Figure 3) to explore to what extent MirrorFair and existing methods enhance model fairness as well as the side effect on model performance. In general, MirrorFair reduces the most bias with minimal impact on performance.

In particular, regarding enhancing fairness, MirrorFair reduces the SPD bias by 44.69%, AOD bias by 51.93%, EOD bias by 55.43%, and reduces overall bias by 50.68%. MAAT, the state-of-the-art method in trade-off model performance and fairness, reduces the overall bias by 40.99%; EOP, the state-of-the-art in mitigating bias, reduces the overall bias by 45.43%. As to the impact on performance, MirrorFair and MAAT perform best in maintaining model prediction performance, and they all slightly increase the accuracy and precision with slight decreases in recall and f1-score, but MirrorFair outperforms MAAT with a 0.25% advantage in maintaining the overall performance.
Significance of difference in fairness between different methods: In the prior evaluation, we demonstrated its effectiveness by comparing each method with default models. Here, we compare different methods with each other via the Mann-Whitney U-Test [19, 43], a non-parametric statistical method, to explore the significant difference in fairness between different methods. Aligning with previous work [19], we consider the change in fairness to be statistically significant only if the resulting p-value from the test is less than 0.05, whilst we calculate the proportion of significantly improving fairness cases. Table 5 presents 49 (seven baselines × seven methods) comparison results of seven methods, including Default and top six methods in trade-off fairness and performance based on the result in Figure 3. The value in the cross of the "VS Default" column and "MirrorFair" row means that 99% of MirrorFair cases are significantly fairer than Default cases. The value in the cross of the "VS MAAT" column and "MirrorFair" row means that 44% of MirrorFair cases are significantly fairer than MAAT cases.

Answer to RQ1:

MirrorFair surpasses the Fairea trade-off baseline by 93%, outperforming the state-of-the-art method by 5 percent points. MirrorFair also shows the highest efficacy in both enhancing fairness and maintaining performance. In particular, regarding the extent of mitigating bias, MirrorFair reduces overall bias by 50% with the minimum effect on performance, while the state-of-the-art methods only reduce it by 45% (EOP) and 41% (MAAT).

5.2 RQ2: Applicability and Versatility

We showcase the superior overall efficacy of MirrorFair in bias mitigation while preserving performance, surpassing existing methods across 32 decision-making scenarios (comprising eight decision-making tasks × four algorithms) to answer RQ2. The efficacy range, computed as the disparity between the highest and lowest efficacy values, serves as a metric for assessing efficacy stability. In this research question, we delve into the applicability and versatility of MirrorFair in comparison to state-of-the-art methods across various tasks and algorithms, leveraging both overall efficacy and efficacy range as key indicators.

Table 6 presents the proportion of cases where MirrorFair and other methods surpass the Fairea baseline with different tasks and algorithms. Concerning algorithms, MirrorFair outperforms all the existing methods in our study in terms of surpassing the Fairea trade-off baseline. Specifically, MirrorFair achieves the highest efficacy in the LR algorithm (95.38%) and the lowest efficacy in the RF algorithm (89.23%), whereas the state-of-the-art method MAAT achieves the highest efficacy in the LR algorithm (93.35%) and the lowest efficacy in the DNN algorithm (82.97%).

Regarding decision-making tasks, MirrorFair achieves the best trade-off efficacy in six out of the eight tasks. It is slightly surpassed by the DIR method in the Compas-Sex task (outperforming MirrorFair by two percentage points) and by MAAT in the Mep-Race task (outperforming MirrorFair by one percentage point). However, MirrorFair achieves its lowest efficacy in the German-Sex task (83.3%), which is significantly better than the state-of-the-art method MAAT, which achieves its lowest efficacy in the German-Sex (71.70%) and German-Age (73.07%) tasks.

Table 6. (RQ2) Surpassing Fairea baseline proportion across different algorithms and tasks.

Method	Algoritm				Task							
	LR	RF	SVM	DNN	Adult-Sex	Adult-Race	Compas-Sex	Compas-Race	German-Sex	German-Age	Bank-Age	Mep-Race
FairMask	76.52%	65.08%	75.97%	43.27%	63.80%	66.73%	78.53%	42.93%	70.33%	66.97%	57.77%	74.60%
DIR	77.02%	70.10%	72.58%	69.70%	54.23%	89.10%	**100.00%**	97.20%	64.87%	76.23%	7.43%	89.73%
RW	89.52%	65.80%	83.22%	75.95%	51.40%	73.17%	96.20%	90.37%	72.30%	77.27%	78.17%	90.10%
EOP	84.15%	53.05%	83.45%	77.70%	79.13%	72.50%	95.20%	90.57%	68.47%	52.50%	61.90%	76.43%
MAAT	93.35%	85.53%	90.23%	82.97%	90.53%	93.40%	98.77%	97.00%	71.70%	73.07%	82.63%	**97.07%**
MirrorFair	**95.38%**	**89.23%**	**95.10%**	**90.35%**	**96.60%**	**96.43%**	98.13%	**99.93%**	**83.03%**	**84.03%**	**85.90%**	96.07%

Answer to RQ2:

MirrorFair demonstrates higher efficacy and a narrower efficacy range compared to the state-of-the-art method, showcasing its superior applicability and versatility across different algorithms and tasks.

5.3 RQ3: Effectiveness in Mitigating Multiple Attribute Biases

Datasets may contain multiple sensitive attributes that require protection. Among the 14 existing methods we evaluated, only some are capable of handling multiple sensitive attribute protection,

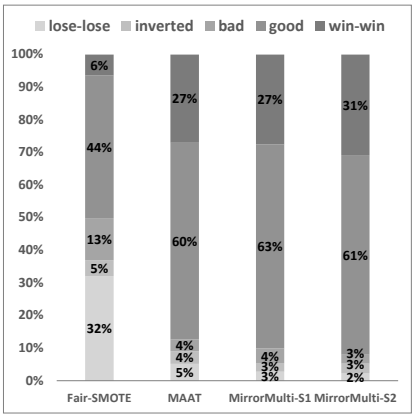

Fig. 5. RQ3: The efficacy level distribution in mitigating multiple attribute biases.

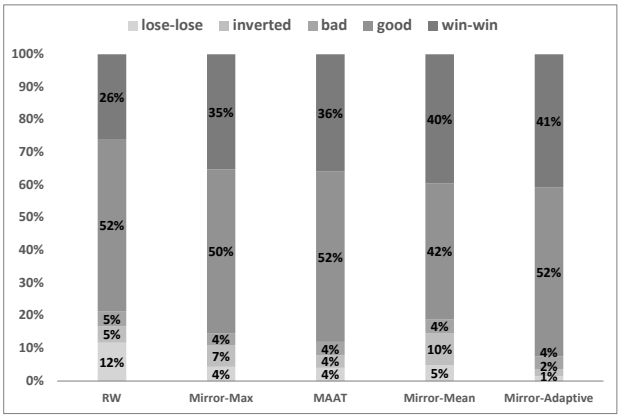

Fig. 6. RQ4: The efficacy level distribution of different ensemble strategies of MirrorFair and the existing methods.

as stated in their original papers. To answer this research question, we compared MirrorMulti-S1 and MirrorMulti-S2, the two variants of MirrorFair introduced in Section 3, with state-of-the-art methods in balancing model performance and fairness when mitigating biases associated with multiple sensitive attributes. We conducted these comparisons using the Adult and Compas datasets.

Figure 5 presents the trade-off distribution of mitigating multiple sensitive attributes. We observed that MirrorMulti-S2 achieves 30.86% "win-win" cases and 61.13% "good" cases. MirrorMulti-S1 achieves 27.50% "win-win" cases and 62.60% "good" cases while the state-of-the-art method MAAT achieves 27.06% "win-win" cases and 60.35% "good" cases. It means that both MirrorMulti-S1 and MirrorMulti-S2 outperformed MAAT and Fair-SMOTE in both achieving "win-win" cases and surpassing the Fairea baseline.

Answer to RQ3:

The two variants of MirrorFair (MirrorMulti-S1 and MirrorMulti-S2) are both more effective in protecting multiple sensitive attributes than the existing methods. In particular, MirrorMulti-S2 surpasses the Fairea baseline in 91.99% of cases, which is 4.58 percentage points higher than the state-of-the-art method.

5.4 RQ4: Impact of Mirroring Processing and Effectiveness of Adaptive Strategies

In this research question, we investigate the impact of mirroring processing in Section 5.4.1 via investigating the probability difference between default predictions and counterfactual predictions (mirror predictions). Following this, we present the comparative results between different MirrorFair ensemble strategies and existing methods in Section 5.4.2.

5.4.1 Impact of Mirroring Processing on Decision-making Scenarios. We conducted empirical investigations across eight decision-making tasks using four distinct ML algorithms. Due to space constraints, we provide part of the *DIF* value in Table 7, illustrating potential patterns in mirroring processing effects. Comprehensive statistical results for all experiments are presented in Table 8.

In Table 7, cases in dark grey (i.e., "Compas-Race-LR" and "Compas-Race-SVM") are mirror-insensitive scenarios, as the *DIF* values (probability differences between the default and the mirror model) are close to zero, suggesting that mirroring processing has minimal impact on these particular

Table 7. Probability difference between the default and mirror models (Random sampling 5 testing instances). "Adult-Sex-LR" means mitigating "Sex" bias in "Adult" using "LR" algorithms; $P_{def}(1,0)$ denotes the prediction probability of output $Y = 1$ on the condition of sensitive attribute $A = 0$ of the default model; $P_{mir}(1,0)$ denotes the prediction probability of output $Y = 1$ on the condition of sensitive attribute $A = 0$ of the mirror model; DIF denotes the probability difference between the default and mirror model.

Adult-Sex-LR			**Adult-Sex-SVM**			**Adult-Sex-RF**			**Adult-Sex-DNN**		
$P_{def}(1,0)$	$P_{mir}(1,0)$	**DIF**	$P_{def}(1,0)$	$P_{mir}(1,0)$	**DIF**	$P_{def}(1,0)$	$P_{mir}(1,0)$	**DIF**	$P_{def}(1,0)$	$P_{mir}(1,0)$	**DIF**
0.34	0.50	-0.17	0.36	0.50	-0.14	0.60	0.42	0.17	0.30	0.52	-0.22
0.44	0.62	-0.17	0.48	0.68	-0.20	0.16	0.72	-0.56	0.31	0.63	-0.31
0.36	0.53	-0.17	0.46	0.66	-0.20	0.52	0.43	0.09	0.46	0.58	-0.12
0.45	0.60	-0.16	0.48	0.68	-0.20	0.10	0.67	-0.57	0.41	0.75	-0.35
0.39	0.56	-0.17	0.43	0.61	-0.18	0.28	0.76	-0.49	0.47	0.79	-0.32
$P_{def}(1,1)$	$P_{mir}(1,1)$	**DIF**	$P_{def}(1,1)$	$P_{mir}(1,1)$	**DIF**	$P_{def}(1,1)$	$P_{mir}(1,1)$	**DIF**	$P_{def}(1,1)$	$P_{mir}(1,1)$	**DIF**
0.59	0.42	0.17	0.62	0.44	0.18	0.38	0.50	-0.12	0.52	0.43	0.09
0.64	0.47	0.17	0.56	0.35	0.21	0.55	0.46	0.10	0.54	0.20	0.35
0.55	0.38	0.17	0.52	0.30	0.22	0.53	0.38	0.15	0.72	0.35	0.37
0.55	0.38	0.17	0.52	0.30	0.22	0.47	0.53	-0.06	0.70	0.43	0.27
0.59	0.41	0.17	0.61	0.43	0.18	0.56	0.47	0.09	0.56	0.24	0.31
Compas-Race-LR			**Compas-Race-SVM**			**Compas-Race-RF**			**Compas-Race-DNN**		
$P_{def}(1,0)$	$P_{mir}(1,0)$	**DIF**	$P_{def}(1,0)$	$P_{mir}(1,0)$	**DIF**	$P_{def}(1,0)$	$P_{mir}(1,0)$	**DIF**	$P_{def}(1,0)$	$P_{mir}(1,0)$	**DIF**
0.50	0.50	0.00	0.55	0.54	0.01	0.29	0.52	-0.23	0.62	0.42	0.20
0.50	0.50	0.00	0.50	0.50	0.00	0.55	0.45	0.10	0.69	0.32	0.37
0.50	0.50	0.00	0.48	0.47	0.01	0.97	0.16	0.81	0.43	0.53	-0.10
0.50	0.50	0.00	0.47	0.45	0.01	0.42	0.65	-0.23	0.38	0.52	-0.14
0.50	0.50	0.00	0.49	0.48	0.01	0.27	0.87	-0.60	0.33	0.64	-0.32
$P_{def}(1,1)$	$P_{mir}(1,1)$	**DIF**	$P_{def}(1,1)$	$P_{mir}(1,1)$	**DIF**	$P_{def}(1,1)$	$P_{mir}(1,1)$	**DIF**	$P_{def}(1,1)$	$P_{mir}(1,1)$	**DIF**
0.45	0.45	0.00	0.47	0.47	0.01	0.50	0.98	-0.48	0.50	0.61	-0.11
0.51	0.51	0.00	0.51	0.52	0.00	0.45	0.99	-0.54	0.28	0.54	-0.26
0.49	0.49	0.00	0.49	0.49	0.00	0.48	0.61	-0.13	0.67	0.42	0.24
0.53	0.52	0.00	0.52	0.53	0.00	0.45	0.58	-0.13	0.37	0.61	-0.24
0.50	0.50	0.00	0.51	0.51	0.00	0.32	0.69	-0.37	0.46	0.66	-0.19

Table 8. Taxonomy results of decision-making scenarios using $\delta = 0.05$. $DIF \in \mathcal{N}^{\delta}(c)$ means that the probability difference DIF between two model predictions near the decision boundary, belonging to the neighborhood of a constant c. Where $\mathcal{N}$ denotes the neighborhood of constant c, δ denotes the width of the neighborhood. $Avg.DIF$ denotes the average of DIF.

Decision Task	LR			RF			SVM			DNN		
	$DIF \in \mathcal{N}^{\delta}(c)$	$Mean_{DIF}$	Type	$DIF \in \mathcal{N}^{\delta}(c)$	$Mean_{DIF}$	Type	$DIF \in \mathcal{N}^{\delta}(c)$	$Mean_{DIF}$	Type	$DIF \in \mathcal{N}^{\delta}(c)$	$Mean_{DIF}$	Type
Adult-Sex	✓	0.17	Regular	×	-	Irregular	✓	0.18	Regular	×	-	Irregular
Adult-Race	✓	0.03	Regular	×	-	Irregular	✓	0.04	Regular	×	-	Irregular
Compas-Sex	✓	0.08	Regular	×	-	Irregular	✓	0.06	Regular	×	-	Irregular
Compas-Race	✓	0.00	Insensitive	×	-	Irregular	✓	0.01	Insensitive	×	-	Irregular
German-Sex	✓	0.02	Regular	×	-	Irregular	✓	0.02	Regular	×	-	Irregular
German-Age	✓	0.02	Regular	×	-	Irregular	✓	0.01	Insensitive	×	-	Irregular
Bank-Age	✓	0.08	Regular	×	-	Irregular	✓	0.13	Regular	×	-	Irregular
Mep-Race	✓	0.11	Regular	×	-	Irregular	✓	0.09	Regular	×	-	Irregular

decision-making scenarios. Cases in light grey (i.e., "Adult-Sex-LR" and "Adult-Sex-SVM") are mirror-regular scenarios (as introduced in Section 3.4.1), as the DIF values are close to a constant value. For example, for "Adult-Sex-LR", all the DIF values for $A = 1$ are around 0.17. Cases in white are mirror-irregular scenarios, as we observe no pattern in the DIF values.

A comprehensive breakdown of these taxonomy results for all the decision-making scenarios under examination is presented in Table 8. We observe that our mirroring processing has an irregular effect on Random Forest (RF) and DNN, which might be because Random Forest and DNNs are more complex models capable of capturing non-linear relationships between features and the target variable. The interactions between features of these models are flexible and diverse. On the contrary, Logistic Regression (LR) and SVM are more straightforward in how they use features to make predictions, typically relying on linear boundaries. Changes in sensitive features, therefore, have more predictable effects due to the linear nature of these models.

5.4.2 Effectiveness of Adaptive Strategies. Figure 6 showcases the effectiveness of various ensemble strategies within MirrorFair, alongside the performance of the state-of-the-art non-ensemble method RW and the leading ensemble approach MAAT. Specifically, Mirror-Mean and Mirror-Max employ only the E-Mean and E-Max strategies, respectively, as delineated in Section 3.4.2, to ensemble the predictions from the mirror model and the default model. Conversely, Mirror-Adaptive dynamically chooses the most suitable ensemble strategy—either E-Mean or E-Max—tailored to the specific decision-making scenario at hand.

We observe that all ensemble strategies outperform the state-of-the-art non-ensemble method RW in both achieving "win-win" scenarios and surpassing the Fairea baseline. Among the ensemble methods, Mirror-Adaptive stands out with 41% "win-win" cases and a 93% proportion in surpassing the Fairea baseline. MAAT surpasses the Fairea baseline in 88% of cases, outperforming Mirror-Max (85%), and Mirror-Mean (82%). MAAT treats different decision-making scenarios equally, and thus has lower efficacy than MirrorFair. In summary, the comparison results highlight the advantages of MirrorFair in mitigating model bias.

Answer to RQ4:

Mirroring processing has varying impacts on different decision-making scenarios. MirrorFair can discern these differences and categorize them into distinct types. The comparison results, as shown in Figure 6, underscore the substantial advantages of adaptive ensemble strategies over fixed ensemble strategies in mitigating bias across diverse scenarios.

6 DISCUSSION

Here, we highlight the novelty of MirrorFair and discuss the threats to validity and future work.

6.1 Novelty and Superiority of MirrorFair Over Existing Methods

The results demonstrate that MirrorFair outperforms the state-of-the-art methods in mitigating bias, maintaining performance, and keeping consistency across various tasks and algorithms. Multiple models can describe the data from different perspectives and levels, and ensemble methods allow multiple models to complement each other and obtain better results [54]. As an ensemble approach, MirrorFair can learn more information from the training dataset and directionally and purposefully decrease the contribution of sensitive attributes on the model prediction to mitigate model bias better. The ensemble design makes MirrorFair much further ahead of the non-ensemble methods. As for ensemble techniques, there are two crucial aspects: what to ensemble and how to ensemble. MirrorFair is novel in both aspects compared to the existing ensemble method MAAT. In particular, regarding what to ensemble, MirrorFair ensembles the original model and the Mirror Model via counterfactual inference, while MAAT ensembles models that are optimized for different objectives: fairness and ML performance. As for how to ensemble, MirrorFair adopts an adaptive strategy that chooses optimal ensemble strategies according to different tasks, while MAAT simply gets the average of two models' predictions as the final prediction.

6.2 Threats to Validity

6.2.1 Internal Threats. The selection of measurement and metrics could potentially introduce threats to the validity of experimental results. To mitigate this concern, we have aligned our approach with prior research [19] and the latest empirical investigation [22], employing the state-of-the-art benchmarking tool Fairea [33] in conjunction with multiple performance and fairness metrics to assess the effectiveness of MirrorFair.

6.2.2 External Threats. The choice of benchmark datasets, algorithms, and existing methods has implications for the external validity of this study. To address this issue, we have adhered to established experimental settings from prior research and conducted our experiments using five widely recognized benchmark datasets and four commonly employed machine learning algorithms, including deep neural networks. This comprehensive approach allows us to make meaningful comparisons with 14 existing methods[19, 22, 48] from both the SE and AI communities.

6.3 Limitations of MirrorFair and Future Improvements

6.3.1 Limitation of MirrorFair. Typical fairness metrics (e.g., SPD, AOD, EOD) support only binary sensitive attributes in AIF360. Therefore, similar to existing methods [16, 17, 19, 48], MirrorFair converts sensitive attributes into binary categories, such as classifying the race attribute in the Adult dataset as "White" and "non-White". This simplification enables researchers to concentrate on addressing the most pronounced biases and reduces the complexity of modeling fairness issues. Nevertheless, this practice oversimplifies human characteristics and may obscure the distinctions within sub-groups of the binary categories (e.g., "Asian" and "Eskimo" within "non-White"), potentially introducing new biases.

6.3.2 Future Improvements. In our future work, we will introduce novel fairness metrics that calculate the overall standard deviation of fairness scores across groups, allowing for the measurement of more fine-grained biases with non-binary sensitive attributes. Subsequently, we will leverage these new metrics to perform a comprehensive empirical analysis of the effectiveness of existing methods in mitigating fine-grained biases and to develop innovative bias-mitigation approaches that do not rely on binary conversion of sensitive attributes. Additionally, we intend to expand the capabilities of the MirrorFair approach to encompass regression and generation tasks.

7 CONCLUSIONS

This paper introduces MirrorFair, a novel adaptive ensemble approach for mitigating bias in machine learning software through ensembling mirror predictions generated by counterfactual inference. We have demonstrated that MirrorFair surpasses the state-of-the-art methods in both the ML and SE communities by striking a balance between model performance and fairness across various decision tasks and ML algorithms. Our results underscore the efficacy of leveraging causal analysis and counterfactual inference to mitigate bias and enhance fairness in machine learning software. In conclusion, the successful implementation of MirrorFair enriches the array of bias-mitigating methods and underscores the potential of integrating counterfactual inference in future research.

8 DATA AVAILABILITY

To facilitate future work and replication of our approach, we make the replication package of MirrorFair, including code, results, and supplementary materials available on the website [6].

ACKNOWLEDGMENTS

Yepang Liu is partially supported by the National Natural Science Foundation of China (Grant No. 61932021) and the National Key Research and Development Program of China (Grant No. 2019YFE0198100). Mohammad Reza Mousavi has been partially supported by the UKRI Trustworthy Autonomous Systems Node in Verifiability, Grant Award Reference EP/V026801/2 and the EPSRC grant on Verified Simulation for Large Quantum Systems (VSL-Q), Grant Award Reference EP/Y005244/1, and the EPSRC project on Robust and Reliable Quantum Computing (RoaRQ), Investigation 009, grant reference EP/W032635/1. Also, the King's Quantum grants provided by King's College London are gratefully acknowledged.

REFERENCES

[1] 1994. The German dataset. https://archive.ics.uci.edu/ml/datasets/statlog(german+credit+data)

[2] 2014. The Bank dataset. https://archive.ics.uci.edu/ml/datasets/bank+marketing

[3] 2015. The Mep dataset. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp

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

[5] 2017. The Adult Census Income dataset. https://archive.ics.uci.edu/ml/datasets/adult

[6] 2024. Replication package for MirrorFair. https://github.com/XY-Showing/FSE2024-MirrorFair.

[7] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.

[8] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 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.

[9] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220.

[10] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

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

[12] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[13] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[14] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[15] 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.

[16] 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*. 429–440.

[17] 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.

[18] 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.

[19] Zhenpeng Chen, Jie Zhang, Federica Sarro, and Mark Harman. 2022. MAAT: A Novel Ensemble Approach to Addressing Fairness and Performance Bugs for Machine Learning Software. In *The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*.

[20] Zhenpeng Chen, Jie Zhang, Federica Sarro, and Mark Harman. 2023. Fairness Improvement with Multiple Protected Attributes: How Far Are We?. In *46th International Conference on Software Engineering (ICSE 2024)*. ACM.

[21] Zhenpeng Chen, Jie M Zhang, Max Hort, Federica Sarro, and Mark Harman. 2022. Fairness Testing: A Comprehensive Survey and Analysis of Trends. *arXiv preprint arXiv:2207.10223* (2022).

[22] 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* (2023).

[23] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women — reuters.com. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [Accessed 13-Apr-2023].

[24] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120* (2017).

[25] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[26] 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.

[27] Grant R Fowles. 1989. *Introduction to modern optics*. Courier Corporation.

[28] Xuanqi Gao, Juan Zhai, Shiqing Ma, Chao Shen, Yufei Chen, and Qian Wang. 2022. FairNeuron: improving deep neural network fairness with adversary games on selective neurons. In *Proceedings of the 44th International Conference on Software Engineering*. 921–933.

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

[30] Google. [n. d.]. Machine Learning Glossary: Fairness | Google Developers — developers.google.com. https://developers.google.com/machine-learning/glossary/fairness [Accessed 26-Mar-2023].

[31] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[32] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[33] 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 Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 994–1006.

[34] Zhenlan Ji, Pingchuan Ma, Shuai Wang, and Yanhui Li. 2023. Causality-Aided Trade-off Analysis for Machine Learning Fairness. *arXiv preprint arXiv:2305.13057* (2023).

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

[36] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*. IEEE, 924–929.

[37] 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: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 35–50.

[38] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.

[39] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[40] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems* 33 (2020), 728–740.

[41] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[42] Tianlin Li, Xiaofei Xie, Jian Wang, Qing Guo, Aishan Liu, Lei Ma, and Yang Liu. 2023. Faire: Repairing Fairness of Neural Networks via Neuron Condition Synthesis. *ACM Transactions on Software Engineering and Methodology* (2023).

[43] 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.

[44] Verya Monjezi, Ashutosh Trivedi, Gang Tan, and Saeid Tizpaz-Niari. 2023. Information-Theoretic Testing and Debugging of Fairness Defects in Deep Neural Networks. *arXiv preprint arXiv:2304.04199* (2023).

[45] Judea Pearl. 2009. *Causality*. Cambridge university press.

[46] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[48] Kewen Peng, Joymallya Chakraborty, and Tim Menzies. 2022. FairMask: Better Fairness via Model-based Rebalancing of Protected Attributes. *IEEE Transactions on Software Engineering* (2022).

[49] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[50] Long H Pham, Jiaying Li, and Jun Sun. 2020. SOCRATES: Towards a Unified Platform for Neural Network Analysis. *arXiv preprint arXiv:2007.11206* (2020).

[51] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).

[52] Alexander S. Poznyak. 2008. Chapter 14 - Sets, Functions and Metric Spaces. In *Advanced Mathematical Tools for Automatic Control Engineers: Deterministic Techniques*, Alexander S. Poznyak (Ed.). Elsevier, Oxford, 251–274. https://doi.org/10.1016/B978-008044674-5.50017-1

[53] Daniel Rodriguez, Israel Herraiz, Rachel Harrison, Javier Dolado, and José C Riquelme. 2014. Preliminary comparison of techniques for dealing with imbalance in software defect prediction. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. 1–10.
[54] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1249.
[55] Bing Sun, Jun Sun, Long H Pham, and Jie Shi. 2022. Causality-based neural network repair. In *Proceedings of the 44th International Conference on Software Engineering*. 338–349.
[56] Developers TensorFlow. 2018. TensorFlow. *Site oficial* (2018).
[57] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.
[58] 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.
[59] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[60] Jie M Zhang and Mark Harman. 2021. "Ignorance and Prejudice" in Software Fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1436–1447.
[61] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* 48, 1 (2020), 1–36.
[62] 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*. 6–17.
[63] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148* (2015).

Received 2023-09-28; accepted 2024-04-16