A Large-Scale Empirical Study on Improving the Fairness of Image Classification Models

Junjie Yang
College of Intelligence and Computing
Tianjin University
Tianjin, China
jjyang@tju.edu.cn

Jiajun Jiang*
College of Intelligence and Computing
Tianjin University
Tianjin, China
jiangjiajun@tju.edu.cn

Zeyu Sun
Science & Technology on Integrated Information System
Laboratory, Institute of Software, Chinese Academy of
Sciences
Beijing, China
zeyu.zys@gmail.com

Junjie Chen
College of Intelligence and Computing
Tianjin University
Tianjin, China
junjiechen@tju.edu.cn

ABSTRACT

Fairness has been a critical issue that affects the adoption of deep learning models in real practice. To improve model fairness, many existing methods have been proposed and evaluated to be effective in their own contexts. However, there is still no systematic evaluation among them for a comprehensive comparison under the same context, which makes it hard to understand the performance distinction among them, hindering the research progress and practical adoption of them. To fill this gap, this paper endeavours to conduct the first large-scale empirical study to comprehensively compare the performance of existing state-of-the-art fairness improving techniques. Specifically, we target the widely-used application scenario of image classification, and utilized three different datasets and five commonly-used performance metrics to assess in total 13 methods from diverse categories. Our findings reveal substantial variations in the performance of each method across different datasets and sensitive attributes, indicating over-fitting on specific datasets by many existing methods. Furthermore, different fairness evaluation metrics, due to their distinct focuses, yield significantly different assessment results. Overall, we observe that pre-processing methods and in-processing methods outperform post-processing methods, with pre-processing methods exhibiting the best performance. Our empirical study offers comprehensive recommendations for enhancing fairness in deep learning models. We approach the problem from multiple dimensions, aiming to provide a uniform evaluation platform and inspire researchers to explore more effective fairness solutions via a set of implications.

---

*Corresponding author.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Software and its engineering** → *Software reliability*; • **General and reference** → **Empirical studies**.

KEYWORDS

Empirical study, Deep learning, Model fairness

1 INTRODUCTION

In recent years, Artificial Intelligence (AI) systems and Deep Learning (DL) models have attracted significant attention for their remarkable performance across diverse application domains, such as image processing [28, 68], machine translation [6, 21, 47, 48] and software engineering [39, 67, 92, 94]. However, the increasing adoption of DL models in real practice also poses new challenges to them. In particular, they have displayed troubling discriminatory tendencies in certain applications, such as elevated error rates, when confronted with specific groups or populations defined by attributes deemed *protected* or *sensitive* [4, 16, 20, 25, 54, 97], including factors like race, gender, and age, etc., causing ethical or even safety issues that will largely affect the reliability and usability of them in real practice. For example, AI judge [30, 36, 76, 91] may exhibit biased decision-making, leading to unequal treatment of individuals based on their sensitive attributes, such as race. This bias can result in disparities in sentencing or legal outcomes, raising concerns about fairness and justice within the legal system. Similarly, AI-powered hiring [33, 57] and loan systems [27, 84] tend to favor applicants of a particular gender or race. In other words, the manifestation of unfair behaviors in AI systems and DL models has raised profound ethical concerns, transforming fairness into an essential ethical and often legally mandated prerequisite for the

widespread adoption and utilization of DL models across various real-world contexts.

To improve the fairness of DL models, many approaches have been proposed in the last decades, particularly in the application domain of image classification, which has been widely adopted in diverse real-world applications. Typically, existing approaches can be classified into three categories according to the stage where de-biasing is applied, i.e., pre-processing, in-processing, and post-processing. Pre-processing techniques [8, 37, 62, 64, 70, 86, 93] endeavor to mitigate model biases by optimizing the training data, such as performing data transformation [93], augmentation [64], sampling [8, 37, 62, 70, 86] and so on. In contrast, in-processing techniques [35, 41, 45, 72, 81] dedicate to improve the model fairness by optimizing the model training process, such as updating the training loss [35, 72] or model structures [41, 45]. Different from the former two categories, post-processing techniques [46] aim to make the deployed models produce fairer outputs without changing the model structures and parameters, such as optimizing the given inputs [85, 96]. Since they do not change the original models, they can be flexibly applied to improve the fairness of deployed models in real practice.

Although these existing approaches have been evaluated to be effective in their own contexts, as will be presented in Section 2.2, they were often evaluated over different datasets and utilized inconsistent metrics for measuring the performance of the proposed methods, making it hard to obtain a uniform conclusion about the performance distinction among them over different application domains. In other words, there still lacks a comprehensive study to systematically compare the performance of different approaches under the same setup and thus provide in-depth analysis and accurate guidance for the better utilization of them in practice.

To fill this gap, we conducted the first extensive empirical study to comprehensively compare the performance of existing methods for improving the fairness of deep learning models. Specifically, we studied 13 state-of-the-art existing techniques for improving the fairness of deep learning models as the representatives, all of which are from the most recent research. To provide a comprehensive and fair comparison, we adopted three widely-used benchmarks that involve diverse image classification tasks related to different sensitive attributes. Moreover, we employed all five widely-used fairness metrics and two accuracy metrics in the result analysis for a consistent measurement, which is the most comprehensive study on model fairness as far as we are aware. According to the empirical results, we have summarized a set of findings and implications that can be valuable to guide future research and the utilization of them in real practice. We highlight partial key findings as follows:

(1) While the best-performing fairness improvement method does not exist, pre-processing and in-processing methods significantly outperform post-processing methods.
(2) Existing methods tend to be *insensitive* to different evaluation metrics, and thus a subset of the metrics can be representative in future studies.
(3) Existing methods are *sensitive* to different experimental datasets, and thus the conclusions drawn from a certain dataset may not be generalizable to others.

To sum up, our work makes the following major contributions:

- We conducted **the first large-scale empirical study** to comprehensively evaluate the performance of 13 state-of-the-art techniques that aim to improve DL models' fairness.
- We summarized **a set of findings and implications** by systematically analyzing and comparing the performance of different techniques under a uniform experimental setup.
- We re-implemented some of the state-of-the-art techniques and built **a uniform evaluation platform** of DL fairness techniques, which can facilitate the replication and comparison for future research in this research area. We have published all our experimental data.

2 BACKGROUND

Firstly, we would like to clarify the problem and concept of *fairness issues* in deep learning models. Furthermore, we will also conduct a succinct literature review of the most recent research on fairness improvement and identify the limitations of them, which motivates the necessity and significance of this study.

2.1 Fairness Issues

Over the years, both researchers and practitioners have introduced and investigated various fairness definitions [11, 19, 54]. These definitions typically fall into two broad categories: *individual fairness* and *group fairness*. Individual fairness requires that software produces similar predictive outcomes for individuals with similar characteristics, while group fairness mandates equitable treatment of different demographic groups. In particular, fairness is always closely related to the concept of *sensitive* or *protected* attributes, which represent characteristics demanding protection against discriminatory practices, such as gender and race. In particular, if the assigned value of a sensitive (or protected) attribute is considered advantageous or beneficial in a particular context, we call it a *privileged value*. For example, in a job application process [65, 82, 83], *male* would be a privileged value for the attribute of gender since man may tend to be preferred over women for certain roles. Therefore, given the sensitive attribute (e.g., gender), the input instances of the deep learning models can be typically divided into two distinct groups: a privileged group that associates instances with privileged values (e.g., male) and an unprivileged group that associates instances with unprivileged values (e.g., female). A *fair* DL model should produce similar or even equivalent prediction results over different groups.

In fact, software fairness issues (i.e., unfair software predictions) [24] have been a growing area of concern in both the Software Engineering (SE) and Deep Learning (DL) research communities. For example, they are typically referred to as *fairness defects* [7] or *fairness bugs* [20] in the SE community. In this study, we refer to the unfair predictions produced by DL models as *fairness issues* by following existing work [24], indicating a discordance between existing and expected fairness criteria. Many testing [2, 19, 26, 31, 59, 73, 75, 78, 79, 87] and repair studies [56, 71, 77] aim to automatically identify and fix such issues. In this study, we direct our attention specifically to the techniques that aim to resolve the fairness issues in image classification models since they have been already widely deployed in practice as explained in the

introduction. In addition, this study focuses on the *group fairness* of DL models.

2.2 Limitations in Existing Studies

In recent years, there has been a growing focus among researchers on fairness issues, leading to various of fairness improvement methods being proposed. In order to gain a deeper understanding of existing studies, we have conducted a succinct literature review of relevant papers published recently in prominent conferences and journals in the Software Engineering (SE) and Artificial Intelligence (AI) domains, including ICSE, ESEC/FSE, TOSEM, AIES, ICLR, ECCV, CVPR, NeurIPS, etc. Table 1 presents the details of these methods that aim to improve DL models' fairness.

In this table, besides the essential information of the method (e.g., method names and categories), we further classify the methods into *Tabular* and *Image* based on the type of dataset they are designed for. Additionally, we display the union sets of sensitive attributes and fairness metrics adopted by all methods within each category. Please note that there are also different classification methods [12]. In this paper we adopt the most widely-used one [4, 54, 97] for easy understanding. After a careful examination, we have the following two major observations regarding the limitations in existing studies.

Limitation 1 (Incomplete or Unrepresentative Datasets): Most of existing methods were evaluated on only one or two datasets, with a strong preference towards facial datasets. While face datasets offer advantages such as ease of acquisition and mature application scenarios, the real applications of such techniques are not limited to the classification based on the facial images, such as scenes or objects. However, only a few methods have been evaluated on a sufficiently diverse set of datasets and task scenarios, which is indeed inadequate. Furthermore, many methods use datasets that are rarely employed by other works, e.g., DiF [55] and Waterbirds [69], lacking the representation and making it harder to horizontally compare the performance of different techniques. In particular, studies in the SE community primarily focus on tabular datasets, which are too simple to assess the generalizability of the proposed method for a broader range of applications. In contrast, image data is more complex, and the models used are also more sophisticated. Fairness research on image data is more widespread and holds high research value. However, there still lacks a comprehensive evaluation of existing approaches over image datasets aiming to provide implications for future research.

Limitation 2 (Inconsistent Fairness Evaluation Metrics): Most of the existing studies only adopted one or two metrics, which may not offer a complete assessment of model fairness. Moreover, some works introduce unique fairness metrics [46, 50] tailored to their specific task scenarios, but these metrics are rarely cited or adopted by other research, raising questions about the reliability and generalizability of these metrics. In particular, the different metrics adopted by existing studies further hinder the comparison of different techniques.

The limitations above pose an emergent need to conduct a large-scale empirical study for comprehensively comparing their performance under a uniform context, which motivates this work. In particular, to address these limitations, we have replicated the experimental results of many existing approaches over their original datasets and then further evaluated their effectiveness over three different datasets involving both facial and non-facial images (i.e., objects). Furthermore, we have employed all the widely-used performance metrics (i.e., five fairness metrics and two accuracy metrics) for systematically demonstrating the effectiveness of them. In this way, we can better understand the performance of existing techniques and derive valuable insights for bias mitigation in deep learning models. Ultimately, our aspiration is to provide meaningful guidance that will facilitate more effective explorations of fairness issues in the realm of image fairness.

3 STUDIED METHODS IN THIS STUDY

As summarized in Table 1, many novel methods have been published in a wide range of conferences and journals, aiming to enhance the fairness of neural networks. The latest research [18] has explored the performance of existing methods that focus on tabular datasets, while those methods targeting image datasets have not been systematically investigated yet. Therefore, our study will primarily focus on fairness improving methods for image datasets. Specifically, we follow the selection criteria below to select the methods ultimately studied in our study.

Criterion 1 (Open-source or Reproducible): To ensure the correctness and reliability of our study, we have tried to reproduce the results of the selected methods before conducting our experiment. Then, we selected the methods that are open-source or can be replicated via our re-implementation according to the reported configurations, while ignored those that cannot replicate. In particular, we appreciate the corresponding authors that assisted us in re-implementing and reproducing their results.

Criterion 2 (Covering Different Categories): To ensure the comprehensiveness of our study, we have selected methods for investigation in all three categories: pre-processing, in-processing, and post-processing, ensuring coverage across all categories.

Criterion 3 (Latest Studies): To ensure the effectiveness and representativeness of our study, we prefer to select the methods that were most recently proposed since they usually can achieve much better performance than the previous ones.

As a result, 13 state-of-the-art methods (including two variants of FR [96]) were included, involving all the three categories, i.e., pre-processing, in-processing, and post-processing. We also emphasized these methods in **bold** in Table 1 and introduced them as follows.

3.1 Pre-processing

Pre-processing methods, as a general practice, are primarily concerned with reducing discrimination and bias within the training data, with the ultimate objective of improving the fairness of the resulting trained model. These techniques involve data transformations and augmentations designed to mitigate inherent forms of discrimination. Pre-processing strategies are employed when the algorithm has the capacity to modify the training data to mitigate bias. From the plethora of available pre-processing methods [8, 9, 37, 40, 62, 63, 70, 86, 93], we specifically concentrate on the following four approaches as the representative based on the selection criteria.

Table 1: Summary of fairness improving methods.

Category	Data Type	Methods	Sensitive Attributes	Fairness Metrics
Pre-processing	Tabular	OP [10], Fair-SMOTE [14], RW [42], DIR [23], LTDD [49], FairPreprocessing [5], Fairway [15]	gender, age, race, black	AAOD, EOD, BER, SPD, DI, FPR, ERD
	Image	FairHSIC [63], **US** [70], **OS** [86], **UW** [62], GAN-debiasing [64], CGL [40], **BM** [62]	gender, background, age, race, color/grayscale	EOD, DEO, BA, BC, KL
In-processing	Tabular	META [13], AD [95], GR [1], MAAT [20], PR [44], FMT [38], CARE [77], Parfait-ML [80]	gender, age, race	NMI, PI/MI, NPI, UEI, CVS, SPD, AAOD, EOD
	Image	**Adv** [86], **DI** [86], **BC+BB** [35], **MFD** [41], FairBatch [66], FSCL [58], FairGRAPE [50], **FDR** [53], **FLAC** [72]	background color, color (bright/dark), age, race, texture bias, gender, color/grayscale	EOD, BC, WA, DEO, SPD, AED
Post-processing	Tabular	EO [32], CEO [61], ROC [43], Fax-AI [29]	gender, age, race	DEO, EOD
	Image	**FairReprogram** [96], Multiaccuracy [46], **FAAP** [85]	gender	DEO, SPD

- **Undersampling (US)** [8, 37, 70] aims at achieving class balance by reducing samples from the majority class. It has also been extended to tackle bias by selectively dropping samples from specific subgroups, which include samples that share common class and bias attributes. This approach effectively balances the sizes of these subgroups. However, a notable challenge arises as the model's exposure to samples becomes constrained by the size of the smallest subgroup.
- **Oversampling (OS)** [86] is employed to rectify class imbalance by duplicating samples, thereby equalizing class sizes. This approach has been adapted to effectively address bias by ensuring balance within sensitive subgroups. In practice, we ascertain the size of the largest subgroup and then replicate samples in other subgroups proportionally. However, it's important to exercise caution, as excessive sample duplication can result in overfitting, especially when dealing with highly parameterized models like deep neural networks.
- **Upweighting (UW)** [9, 62] balances the impact of different samples on the loss function by scaling the loss value with the inverse of the sample's class frequency. Therefore, this technique can mitigate bias across different subgroups rather than individual classes. We classify it into the pre-processing category since it depends on the statistics of the training data. Nevertheless, Upweighting possesses a notable limitation—it can exhibit instability when used in conjunction with stochastic gradient descent [3].
- **Bias Mimicking (BM)** [62] was proposed to address dataset bias in Visual Recognition tasks. It achieves statistical independence between bias and class labels by emulating bias distributions across classes. This method involves subsampling the dataset into diverse distributions per class, striking a balance between simplicity and performance enhancement, effectively bridging the gap between sampling and non-sampling methods.

3.2 In-processing

In-processing methods endeavor to modify state-of-the-art learning algorithms during the model training process to eliminate discrimination. This is primarily achieved through the integration of adjustments into the objective function or the imposition of constraints on the learning process, provided that the model's learning procedure is adaptable to such modifications. These techniques employ various means to adjust the model and mitigate bias in its predictions. In our study, we employ the following six methods.

- **Adversarial Training (Adv)** with uniform confusion [62, 81] introduces an adversarial loss to induce a randomized feature representation of target labels and bias groups within the model. It introduces an innovative CNN architecture specifically crafted to leverage unlabeled and sparsely labeled data in the target domain. Additionally, the architecture incorporates domain confusion and softmax cross-entropy losses for model training.
- **Domain Independent Training (DI)** [62, 86] involves the utilization of extra prediction heads for each bias group. In response to the issue of discriminative models learning unnecessary domain-specific class boundaries, DI promotes the adoption of distinct domain-specific classifiers. Nonetheless, training separate classifiers diminishes data exposure. To address this concern, DI proposes a shared feature representation alongside an ensemble of classifiers. It recommends direct reasoning on class boundaries of domains, effectively eradicating class-domain correlations.
- **Bias-Contrastive and Bias-Balanced Learning (BC+BB)** [35, 62] introduces an innovative method to mitigate bias by combining two distinct losses. The Bias-Contrastive (BC) loss incorporates contrastive learning, leveraging bias labels to effectively mitigate bias. Furthermore, the Bias-Balanced (BB) regression loss enhances debiasing performance by optimizing the model to achieve a uniform target-bias correlation distribution.
- **FLAC** [72] mitigates bias in DL models by reducing mutual information between model-extracted features and a protected attribute. It utilizes a sampling strategy to highlight underrepresented data and transforms fair representation learning into a probability matching task using representations obtained from a bias-capturing classifier. This approach disentangles the target representation from bias and protected attributes, ultimately leading to fairer outcomes.

- **MMD-based Fair Distillation (MFD)** [41] combines feature distillation with maximum mean discrepancy (MMD) to enhance both prediction accuracy and fairness simultaneously. The theoretical foundation demonstrates that its MMD-based regularization fosters fairness by aligning group-conditioned features of the student model across sensitive groups and aligning them with the teacher model's group-averaged features, resulting in improved accuracy as well.
- **Fair Deep Feature Reweighting (FDR)** [53] is a simple and straightforward method that improves the model fairness by fine-tuning, where the fairness metrics are incorporated when computing the loss for model updating. In particular, only the last-layer will be updated during the tuning process.

3.3 Post-processing

Post-processing methods are employed to improve fairness in deep learning software by modifying prediction outcomes. These methods are applied after the model training process. When dealing with models as black boxes without the capability to modify training data or algorithms, post-processing becomes essential and easy to use in practice. In this study, we adopted three methods in this category, including two variants of the FairReprogram.

- **FairReprogram (FR)** [96] leverages model reprogramming techniques to address scenarios where it's not feasible to modify existing models. Instead, it appends a set of perturbations known as the "fairness trigger" to the input data, optimizing it in a min-max framework to align with fairness criteria. In particular, FR applies two methodologies to append the trigger, corresponding to its 2 variants, named **FR-P** (appending the trigger like a patch to the original image) and **FR-B** (appending the trigger at the border of the original image). We applied both of them in our study. Specifically, they perturb the inputs by adding a constant global vector/feature, which is tailored to improve fairness while keeping the deep learning model unchanged. FairReprogram is a versatile framework applicable to various tasks and domains, making it a generic solution.
- **Fairness-Aware Adversarial Perturbation (FAAP)** [85] mitigates unfairness in deployed deep models by updating the given input without modifying model parameters or structures. Specifically, it achieves this by learning to perturb input data, effectively blinding deployed models on fairness-related attributes such as gender or ethnicity. FAAP employs a discriminator-generator adversarial framework to ensure that protected attributes are not correlated with model predictions, making it flexible and practical to addressing fairness bugs in real-world AI systems.

4 EXPERIMENTAL SETUP

4.1 Dataset Selection

As aforementioned, existing studies tend to employ different datasets to evaluate the performance of the proposed method. In this study, we aim to provide a comprehensive and uniform comparison among existing techniques. Specifically, we selected the datasets with regard to the following criteria.

Criterion 1 (Different Image Types): We selected the datasets that associate to different image types (i.e., facial and non-facial),

involve different sensitive attributes (e.g., age, race and gender), and are used for different classification tasks.

Criterion 2 (Universal and Adaptable): To make the adaptation of those studied methods easy to our experiment, we preferred the datasets that were widely used by existing studies.

As a result, we selected three datasets (i.e., CelebA, UTKFace, and CIFAR-10S) for conducting the empirical study. The summary of the datasets is shown in Table 2, which presents the sensitive attribute considered in the study, the labels of the classification tasks, and the studied methods that previously used the same settings. In addition, we also present the number of instances for model training and testing per each dataset by following existing studies.

Table 2: Summary of adopted benchmark datasets.

Dataset	Sensi. Attr.	Task Label	Used By	#Train	#Valid	#Test
CelebA	Gender	BlondHair	All	93,141	19,867	19,962
UTKFace	Age	Gender	UW,BM,FLAC	10,521	2,220	2,221
	Race		UW,BM,BC+BB,MFD,FLAC	10,744	2,370	2,370
CIFAR-10S	Color	Objects	OS,UW,BM,Adv,DI,MFD,	50,000	2,000	18,000

Specifically, CelebA and UTKFace datasets represent human-centric datasets that involve sensitive attributes (e.g., age and race) reflected in facial images, while CIFAR-10S is a more general dataset in fairness study, which introduces biases related to image colors of objects (non-facial images). They were previously used for evaluation by partial/all methods under different conditions, e.g., using different metrics. From the table we can also find that previous studies tend to employ different datasets, hindering their comprehensive comparison. In this study, we aim to provide an extensive study through a systematic evaluation by employing all of them, which vary in multiple perspectives.

4.2 Measurement Metrics

4.2.1 Fairness Metrics. Numerous prior studies have predominantly assessed fairness using a restricted set of fairness metrics. Although certain approaches have introduced their unique fairness metrics, these metrics have not achieved widespread adoption and lack broad applicability. To address these limitations, we employ the widely-used five metrics while excluding the self-defined metrics for measuring the fairness of models, i.e., **SPD** [4, 5, 20, 54], **DEO** [41, 54, 96], **EOD** [4, 5, 20, 54], **AAOD** [4, 5, 20], and **AED** [4, 5, 53], following previous work. The reason we chose these five metrics is twofold: (1) They have been integrated into the well-known *AIF360* and *Fairlearn* toolkits, indicating the importance and representativeness of them. (2) They have been well recognized and adopted by at least three previous studies shown in Table 1.

Before delving into a detailed explanation of each metric, we will provide some definitions for the symbols used. Formally, let A be a sensitive (or protected) attribute, and $A = 1$ represents the instance that belongs to the privileged group while $A = 0$ to the unprivileged group. We use Y and $\hat{Y}$ to denote the expected and actual prediction labels, respectively, with 1 as the favorable label and 0 as the unfavorable label. We use P to represent the probability of model prediction. Then, the definitions of the above metrics are defined as follows.

- **SPD** (Statistical Parity Difference): the difference of probabilities that the (un)privileged groups receive favorable outcomes:

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

- **DEO** (Equalized Odds Difference): the max of differences about true/false positive rates for unprivileged and privileged groups:

$$DEO = max\{|P[\hat{Y} = 1|A = 0, Y = 0] - P[\hat{Y} = 1|A = 1, Y = 0]|,$$
$$|P[\hat{Y} = 1|A = 0, Y = 1] - P[\hat{Y} = 1|A = 1, Y = 1]|\}. \tag{2}$$

- **EOD** (Equal Opportunity Difference): the true positive rate difference between unprivileged and privileged groups:

$$EOD = P[\hat{Y} = 1|A = 0, Y = 1] - P[\hat{Y} = 1|A = 1, Y = 1]. \tag{3}$$

- **AAOD** (Average Absolute Odds Difference): the average of absolute difference of true/false positive rates between unprivileged and privileged groups:

$$AAOD = \frac{1}{2}(|P[\hat{Y} = 1|A = 0, Y = 0] - P[\hat{Y} = 1|A = 1, Y = 0]|$$
$$+|P[\hat{Y} = 1|A = 0, Y = 1] - P[\hat{Y} = 1|A = 1, Y = 1]|). \tag{4}$$

- **AED** (Accuracy Equality Difference): the difference of model misclassification rates across difference sensitive groups:

$$AED = P[\hat{Y} \neq Y|A = 0] - P[\hat{Y} \neq Y|A = 1]. \tag{5}$$

Considering that the classification task for CIFAR-10S involves multi-class classification, whereas the classic fairness metrics mentioned earlier are primarily designed for binary classification tasks, we employ the **"ovr (one-vs-rest)"** strategy, inspired by [41, 96] and *sklearn* to decompose the N-ary classification problem into N binary classification subproblems. In our experiments, we initiate the process by computing the bias score for each class, following the same procedure as described for binary classification. Specifically, we calculate the fairness metrics for each class against the rest and then average the results across different classes to derive the multi-class versions of these fairness metrics.

According to the definitions of fairness metrics, a smaller metric value indicates a fairer model. 0 denotes absolute fairness.

4.2.2 Performance Metrics. To evaluate the quality of a model's predictions, we use traditional classification metrics, including **Accuracy** and **Balanced Accuracy**, which are defined by formulas 6 and 7. In particular, the latter can mitigate the effect of imbalanced dataset to the final result. In the formulas, *TP/TN* denotes the number of instances that are correctly classified as positive/negative, while *FP/FN* denotes the number of instances incorrectly classified as positive/negative.

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \tag{6}$$

$$Balanced\text{-}Accuracy = (TP/(TP + FN) + TN/(TN + FP))/2 \tag{7}$$

These two metrics are applicable to both binary and multi-class problems. For all metrics, larger values indicate better performance. The values of performance metrics are between 0 and 1.

4.3 Implementation Details

For all experiments, we use ResNet-18 [34] as the backbone architecture by following existing studies [41, 62, 72, 85, 96]. Moreover, we used the best default configurations of the studied techniques provided in the respective papers. Specifically, we firstly preprocessed the selected three datasets to ensure uniform data partitioning for

consistency. Then, for each studied method, we utilized the optimal default configuration hyper-parameters provided in their respective papers to ensure the reliability of the experiments. However, since some methods originally did not utilize all of the studied datasets used in this paper, we adhered to their principles and adapted them to the new datasets , fine-tuning them for optimal results. In summary, to effectively and fairly compare the experimental results of each method, we standardized various factors such as data partitioning, model backbone, experiment repetition, experimental environment and so on. In particular, to mitigate the effect of randomness, we have also repeated all our experiments **10 times** to ensure the reliability and stability of the results, which took about **40 days** apart from the costs of tool implementation and configuration. It's worth noting that since some approaches depend on a well-trained model (i.e., post-processing) while some approaches train the models from scratch (e.g., pre-processing), we compare the performance of the final optimized models directly. For methods that depend on existing well-trained models, we first conducted a grid search process to obtain the best-performing models as their inputs. We have published all the implementation details in our open-source repository.

Both our performance and fairness metrics are implemented based on *sklearn* [60] and *Fairlearn* [88] libraries. Particularly, we have implemented them in an easy-to-use framework for future studies. It is available in our open-source repository. All experiments are implemented with Python 3.8.18 and PyTorch 1.10.1, and executed on an Ubuntu 20.04.6 LTS with 2.90GHz Intel(R) Xeon(R) Gold 6326 CPUs and eight NVIDIA GeForce RTX 3090 GPUs.

4.4 Research Questions

Our empirical study focuses on the following research questions.

- **RQ1 (Overall effectiveness of fairness improving methods):** *How do different fairness improvement methods perform in terms of accuracy and fairness?*
- **RQ2 (Influence of evaluation metrics):** *How do different accuracy and fairness metrics affect the evaluation results of DL models?*
- **RQ3 (Influence of datasets):** *How do different dataset settings (dataset characteristics, variations in sensitive attributes and so on) impact fairness improvements?*
- **RQ4 (Efficiency of fairness improving methods):** *What are the time costs of different fairness improvement methods?*

5 EVALUATION RESULTS

5.1 RQ1: Overall Effectiveness

This research question assesses the overall performance of studied methods in terms of both the fairness and performance metrics. Our analysis takes both fairness and accuracy into consideration since an excellent fairness improving method should achieve good fairness scores without excessively sacrificing the accuracy, striking a balance or "trade-off" between the two. Its objective is to see whether the best-performing method exists, and thereby offering valuable choices and insights for software engineering researchers engaged in addressing fairness-related issues.

As introduced in Sections 3 and 4, we have systematically studied 13 fairness improving methods over 3 diverse datasets. The experimental results are presented in Table 3 and Figure 1. According to

the experimental results, the performance of different approaches varies greatly over different datasets and even over different evaluation metrics. In particular, the best-performing method does not exist. For example, although the method DI performs better over the UTKFace-race and CIFAR-10S datasets than all the other methods, its performance over the other two datasets suffer a large decline compared with many others, e.g., US and FDR. However, we can observe that the methods in pre-processing and in-processing categories are more effective to improve model fairness than those in the post-processing category, which yield less satisfactory results.

Furthermore, pre-processing and in-processing methods also tend to achieve higher model accuracy together with improving model fairness. This aligns with our intuition, as the sources of bias in fairness mainly originate from the dataset and the training procedure, making direct improvements in these aspects should be the most straightforward and effective approach. Nevertheless, post-processing methods remain essential, particularly with the increasing prevalence of large-scale models in industry. When dealing with many already deployed industrial-scale models, improving their fairness may only be achievable through fine-tuning specific layers or enhancing the entire model as a black box. Existing post-processing methods often employ adversarial perturbations on input images or introduce fairness triggers to obscure the model's identification of sensitive attributes so as to make fair predictions. However, we should acknowledge that their effectiveness still has much room for further improvement. More effective post-processing methods are urgent to be proposed.

Finding 1: While the best-performing fairness improvement method does not exist, pre-processing and in-processing methods have presented much better effectiveness than the post-processing methods.

Considering the methods belonging the same categories, their effectiveness is very close although different methods may perform slightly better in some cases than the others for the pre-processing and in-processing methods regarding improving model fairness. For example, US slightly outperforms other pre-processing methods on the CelebA dataset, but slightly worse than BM on the CIFAR-10S dataset. Similarly, while FLAC and FDR achieve better model fairness than the other in-processing methods on the CelebA and UTKFace-age, they perform relatively poor than DI on the other two datasets. In contrast, FAAP in the post-processing category stands out as it achieve better than the other two methods (i.e., FR-B and FR-P) over almost all the datasets. However, when taking the model accuracy into consideration, the conclusion will be different. For example, BM in the pre-processing category achieves the best accuracy over almost all the datasets. In fact, BM achieves the best model accuracy than almost all the other methods regardless of their categories. Recall the fairness performance of BM in Table 3, we can reasonably conclude that BM achieves the best performance for balancing the model fairness and accuracy. It is important to note that BM actually combines pre-processing and in-processing techniques as it first samples the original dataset to create different distributions for each class and then it employs a specialized training and inference process to achieve the desired trade-off between accuracy and fairness by considering the distributions. Furthermore, from Table 3 we can also observe that the performance of

different methods over different datasets vary greatly, most of them tend to be affected. We leave a more comprehensive analysis on this point to Section 5.3.

Finding 2: BM achieves the best performance for balancing the model fairness and accuracy by combining the strength of both pre-processing and in-processing techniques.

For further analysis of the reasons for the different effectiveness of methods belonging to the same category, we first take the pre-processing category as an example. We believe that the main reason is due to their underlying principles and their suitability for addressing class imbalance and bias in the data. For example, US works well when the class imbalance is significant, as it directly addresses this issue by reducing the majority class. Additionally, by selectively dropping samples from specific subgroups, it can effectively mitigate bias by balancing the sizes of these subgroups. However, OS may result in over-fitting, especially with complex models like deep neural networks, when samples are excessively duplicated. This can lead to poorer generalization on the test set, which might explain its lower performance. To our surprise, as a newly proposed method, BM might not always outperform other techniques, such as US, in terms of fairness. The possible reason is that BM might introduce additional complexity to the pre-processing pipeline by emulating bias distributions across classes. This added complexity may lead to over-fitting or decreased generalization performance. In this case, a simpler method like US might suffice and even outperform BM on fairness. However, as aforementioned, when considering both model fairness and accuracy, BM should still be the best.

Finding 3: For pre-processing methods, over-fitting often proves to be a key factor affecting the performance of various methods.
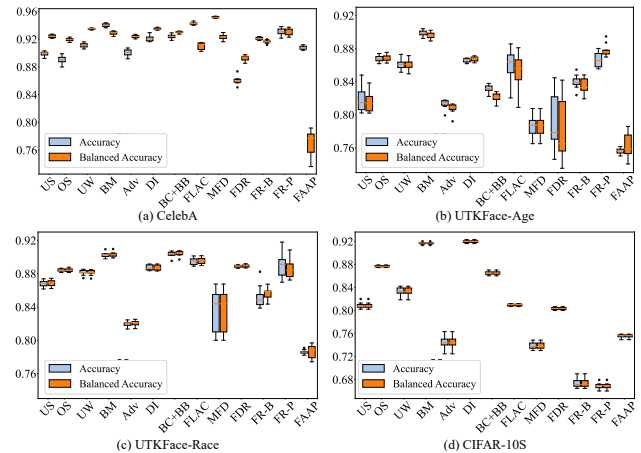
Figure 1: Value distribution regarding accuracy metrics after applying different approaches on different datasets.

Similarly, for in-processing methods, we observe that although we cannot determine which method performs the best, the Adv method exhibits the poorest fairness performance across all datasets, especially on the UTKFace-age as shown in Table 3. The core idea behind Adv is *fairness through blindness* [86], which means attempting to prevent a model from explicitly encoding information

Table 3: Result comparison among different studied methods. The values in each cell denote the mean and standard deviation of the fairness metrics obtained from multiple experiments for the current method and dataset setting. We use different colors to highlight the values of different metrics, the darker of the color, the larger of the value, and the worse of the performance.

Dataset	Metric	Pre-processing				In-processing						Post-processing		
		US	OS	UW	BM	Adv	DI	BC+BB	FLAC	MFD	FDR	FR-B	FR-P	FAAP
CelebA	SPD	$0.186_{\pm0.004}$	$0.338_{\pm0.007}$	$0.258_{\pm0.004}$	$0.208_{\pm0.003}$	$0.330_{\pm0.007}$	$0.230_{\pm0.006}$	$0.272_{\pm0.006}$	$0.170_{\pm0.004}$	$0.210_{\pm0.003}$	$0.139_{\pm0.010}$	$0.289_{\pm0.004}$	$0.265_{\pm0.005}$	$0.185_{\pm0.005}$
	DEO	$0.048_{\pm0.007}$	$0.438_{\pm0.029}$	$0.183_{\pm0.023}$	$0.247_{\pm0.016}$	$0.495_{\pm0.027}$	$0.207_{\pm0.023}$	$0.401_{\pm0.025}$	$0.160_{\pm0.023}$	$0.399_{\pm0.026}$	$0.127_{\pm0.026}$	$0.576_{\pm0.011}$	$0.559_{\pm0.001}$	$0.524_{\pm0.002}$
	EOD	$0.047_{\pm0.008}$	$0.438_{\pm0.029}$	$0.183_{\pm0.023}$	$0.247_{\pm0.016}$	$0.495_{\pm0.027}$	$0.207_{\pm0.023}$	$0.401_{\pm0.025}$	$0.160_{\pm0.023}$	$0.399_{\pm0.026}$	$0.127_{\pm0.026}$	$0.576_{\pm0.011}$	$0.559_{\pm0.001}$	$0.524_{\pm0.002}$
	AAOD	$0.041_{\pm0.004}$	$0.315_{\pm0.014}$	$0.144_{\pm0.011}$	$0.149_{\pm0.008}$	$0.337_{\pm0.012}$	$0.141_{\pm0.012}$	$0.258_{\pm0.013}$	$0.090_{\pm0.012}$	$0.224_{\pm0.013}$	$0.068_{\pm0.012}$	$0.358_{\pm0.007}$	$0.327_{\pm0.006}$	$0.299_{\pm0.004}$
	AED	$0.018_{\pm0.003}$	$0.142_{\pm0.007}$	$0.077_{\pm0.004}$	$0.042_{\pm0.002}$	$0.132_{\pm0.007}$	$0.053_{\pm0.004}$	$0.086_{\pm0.004}$	$0.028_{\pm0.002}$	$0.044_{\pm0.002}$	$0.026_{\pm0.008}$	$0.103_{\pm0.001}$	$0.073_{\pm0.001}$	$0.114_{\pm0.001}$
UTKFace Age	SPD	$0.109_{\pm0.018}$	$0.458_{\pm0.016}$	$0.228_{\pm0.037}$	$0.098_{\pm0.036}$	$0.596_{\pm0.011}$	$0.492_{\pm0.022}$	$0.145_{\pm0.020}$	$0.081_{\pm0.035}$	$0.052_{\pm0.029}$	$0.131_{\pm0.071}$	$0.520_{\pm0.004}$	$0.499_{\pm0.004}$	$0.475_{\pm0.005}$
	DEO	$0.256_{\pm0.027}$	$0.755_{\pm0.017}$	$0.432_{\pm0.050}$	$0.315_{\pm0.031}$	$0.863_{\pm0.015}$	$0.810_{\pm0.040}$	$0.267_{\pm0.032}$	$0.269_{\pm0.035}$	$0.265_{\pm0.031}$	$0.269_{\pm0.091}$	$0.811_{\pm0.009}$	$0.806_{\pm0.003}$	$0.734_{\pm0.003}$
	EOD	$0.127_{\pm0.026}$	$0.073_{\pm0.019}$	$0.069_{\pm0.024}$	$0.213_{\pm0.049}$	$0.256_{\pm0.016}$	$0.086_{\pm0.016}$	$0.056_{\pm0.022}$	$0.190_{\pm0.039}$	$0.185_{\pm0.055}$	$0.075_{\pm0.056}$	$0.155_{\pm0.004}$	$0.075_{\pm0.002}$	$0.153_{\pm0.003}$
	AAOD	$0.196_{\pm0.015}$	$0.414_{\pm0.016}$	$0.251_{\pm0.020}$	$0.263_{\pm0.020}$	$0.560_{\pm0.011}$	$0.448_{\pm0.022}$	$0.161_{\pm0.019}$	$0.228_{\pm0.018}$	$0.213_{\pm0.022}$	$0.171_{\pm0.039}$	$0.486_{\pm0.003}$	$0.435_{\pm0.004}$	$0.442_{\pm0.005}$
	AED	$0.196_{\pm0.016}$	$0.335_{\pm0.008}$	$0.247_{\pm0.019}$	$0.266_{\pm0.020}$	$0.306_{\pm0.010}$	$0.354_{\pm0.020}$	$0.177_{\pm0.018}$	$0.239_{\pm0.017}$	$0.213_{\pm0.022}$	$0.185_{\pm0.035}$	$0.326_{\pm0.003}$	$0.404_{\pm0.001}$	$0.277_{\pm0.001}$
UTKFace Race	SPD	$0.031_{\pm0.007}$	$0.079_{\pm0.005}$	$0.045_{\pm0.007}$	$0.023_{\pm0.009}$	$0.240_{\pm0.012}$	$0.029_{\pm0.008}$	$0.052_{\pm0.006}$	$0.045_{\pm0.012}$	$0.036_{\pm0.014}$	$0.066_{\pm0.007}$	$0.200_{\pm0.006}$	$0.168_{\pm0.006}$	$0.153_{\pm0.005}$
	DEO	$0.021_{\pm0.007}$	$0.068_{\pm0.009}$	$0.029_{\pm0.009}$	$0.024_{\pm0.010}$	$0.234_{\pm0.015}$	$0.016_{\pm0.007}$	$0.040_{\pm0.007}$	$0.041_{\pm0.019}$	$0.049_{\pm0.011}$	$0.069_{\pm0.012}$	$0.179_{\pm0.003}$	$0.150_{\pm0.002}$	$0.166_{\pm0.003}$
	EOD	$0.011_{\pm0.009}$	$0.055_{\pm0.015}$	$0.023_{\pm0.012}$	$0.018_{\pm0.010}$	$0.228_{\pm0.021}$	$0.008_{\pm0.006}$	$0.016_{\pm0.005}$	$0.011_{\pm0.002}$	$0.037_{\pm0.016}$	$0.016_{\pm0.007}$	$0.179_{\pm0.003}$	$0.130_{\pm0.002}$	$0.090_{\pm0.003}$
	AAOD	$0.014_{\pm0.005}$	$0.057_{\pm0.005}$	$0.023_{\pm0.007}$	$0.016_{\pm0.009}$	$0.222_{\pm0.012}$	$0.010_{\pm0.005}$	$0.028_{\pm0.006}$	$0.026_{\pm0.010}$	$0.038_{\pm0.011}$	$0.042_{\pm0.007}$	$0.177_{\pm0.004}$	$0.139_{\pm0.003}$	$0.128_{\pm0.006}$
	AED	$0.010_{\pm0.006}$	$0.012_{\pm0.006}$	$0.006_{\pm0.006}$	$0.015_{\pm0.009}$	$0.009_{\pm0.008}$	$0.007_{\pm0.004}$	$0.012_{\pm0.003}$	$0.019_{\pm0.011}$	$0.014_{\pm0.007}$	$0.028_{\pm0.007}$	$0.007_{\pm0.001}$	$0.016_{\pm0.001}$	$0.048_{\pm0.002}$
CIFAR-10S	SPD	$0.003_{\pm0.001}$	$0.026_{\pm0.004}$	$0.033_{\pm0.002}$	$0.002_{\pm0.001}$	$0.069_{\pm0.004}$	$0.002_{\pm0.001}$	$0.032_{\pm0.001}$	$0.029_{\pm0.001}$	$0.045_{\pm0.017}$	$0.016_{\pm0.002}$	$0.069_{\pm0.001}$	$0.069_{\pm0.001}$	$0.015_{\pm0.005}$
	DEO	$0.015_{\pm0.001}$	$0.108_{\pm0.002}$	$0.131_{\pm0.009}$	$0.012_{\pm0.001}$	$0.282_{\pm0.019}$	$0.010_{\pm0.001}$	$0.128_{\pm0.006}$	$0.120_{\pm0.004}$	$0.174_{\pm0.064}$	$0.074_{\pm0.008}$	$0.249_{\pm0.002}$	$0.250_{\pm0.004}$	$0.082_{\pm0.003}$
	EOD	$0.015_{\pm0.001}$	$0.108_{\pm0.002}$	$0.131_{\pm0.009}$	$0.012_{\pm0.001}$	$0.282_{\pm0.019}$	$0.010_{\pm0.001}$	$0.128_{\pm0.006}$	$0.120_{\pm0.004}$	$0.174_{\pm0.064}$	$0.074_{\pm0.008}$	$0.249_{\pm0.002}$	$0.250_{\pm0.004}$	$0.082_{\pm0.003}$
	AAOD	$0.008_{\pm0.001}$	$0.063_{\pm0.001}$	$0.076_{\pm0.005}$	$0.007_{\pm0.001}$	$0.163_{\pm0.011}$	$0.006_{\pm0.001}$	$0.074_{\pm0.004}$	$0.069_{\pm0.003}$	$0.102_{\pm0.038}$	$0.042_{\pm0.005}$	$0.149_{\pm0.003}$	$0.150_{\pm0.003}$	$0.048_{\pm0.003}$
	AED	$0.002_{\pm0.001}$	$0.005_{\pm0.001}$	$0.007_{\pm0.001}$	$0.001_{\pm0.001}$	$0.015_{\pm0.001}$	$0.001_{\pm0.001}$	$0.006_{\pm0.001}$	$0.006_{\pm0.001}$	$0.011_{\pm0.005}$	$0.004_{\pm0.001}$	$0.019_{\pm0.001}$	$0.019_{\pm0.001}$	$0.005_{\pm0.001}$

about sensitive attributes through adversarial training. However, as we have observed, this practice of ignoring sensitive attribute encoding has several issues. Firstly, due to the phenomenon of redundant encoding [22, 32], even if there is no particular sensitive attribute in the classifier's feature representation, combinations of other attributes can be used as a proxy. For instance, considering a real-world task where a bank assesses a loan application without considering the applicant's gender. If the applicant's occupation is *nurse*, we can still infer with a high probability that the applicant is likely female based on the *nurse* attribute. Secondly, this practice of discarding sensitive attribute encoding may harm accuracy, and the results in Figure 1 support this hypothesis, with Adv's lowest accuracy across all datasets among in-processing methods. The reason may be that internal connections between different attributes likely have a significant impact on accuracy.

Finding 4: Sensitive attributes play an important role in model predictions. Simply improving fairness by removing the encoding of sensitive attributes may not be an effective approach.

Finally, for post-processing methods, we observe that all the studied methods have shown unstable performance. The reason may be that these methods share a common optimization objective, which is the min-max adversarial training framework. In this framework, the discriminator attempts to improve its predictions of sensitive attributes, while the fairness trigger (in FR-B and FR-P) or the generator (in FAAP) modifies input images to reduce the information about protected attributes in the latent representation of a deployed black-box model, thus attempting to degrade predictions. These methods primarily aim to mitigate model bias by hiding the information of sensitive attributes during the feature extraction process, so that the model does not associate predictions with sensitive attributes. However, this approach has its drawbacks. Hiding information about sensitive attributes may also hide features closely related to those attributes, ultimately affecting both accuracy and fairness performance. Supposing an AI judger makes

predictions based on human faces, if we attempt to hide racial encoding information in facial images, we may unintentionally affect information related to skin color, which is closely linked to race. The loss of this information can potentially impact the performance of other prediction tasks.

Finding 5: For post-processing methods, hiding sensitive information encoding may not be an ideal choice, and more effective approaches are still in urgent need.

5.2 RQ2: Influence of evaluation metrics

In this RQ, we investigate the performance of existing methods over different metrics, aiming to inspire researchers to choose appropriate metrics for evaluating DL models in future studies.
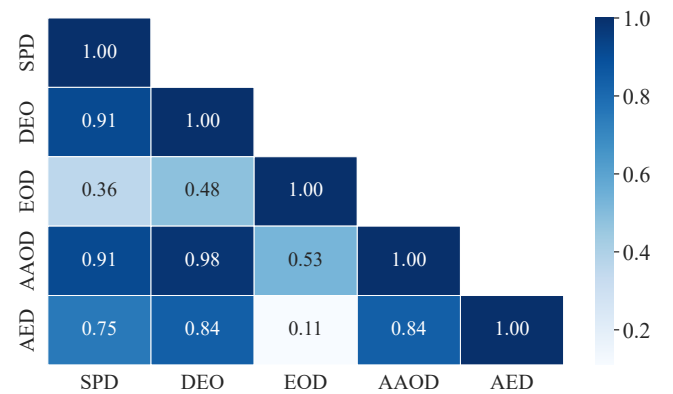
Figure 2: Pearson correlation across different metrics.

Regarding fairness metrics, we find that if a method performs well on one metric, it tends to perform well on other metrics as well, and vice versa. From Table 3, we can also observe that the value distributions of each metric are highly similar within each dataset.

Although different fairness metrics evaluate the performance of DL models from different aspects, when designing fairness improving methods, common strategies or techniques are often employed, including using fairness metrics as optimization objectives and balancing sampling, to ensure good performance across different metrics. To more clearly understand the relations between different fairness metrics, we have calculated the Pearson Correlation coefficient [89] across different metrics over all datasets. The result shown in Figure 2 further confirms our hypothesis that all fairness metrics are positively correlated with each other.

> **Finding 6:** Existing fairness improvement approaches tend to be *insensitive* to the fairness metrics. That is, if a method performs well regarding one metric, it also tends to perform well regarding the other metrics.

Moreover, we conducted a further in-depth investigation into the correlations between different fairness metrics and aimed to provide valuable guidance for future studies. From Figure 2 we can observe that AAOD has the highest overall correlation as its correlation coefficients with the other four metrics never fall below 0.5. We analyze the reasons based on the definitions of each fairness metric (see Section 4.2.1). The goal of SPD is to ensure a DL model's predictions are independent of membership in a sensitive group. It is an easy metric but it only considers a single binary protected attribute, which is sometimes insufficient. For DEO, EOD and AAOD, the goal of these metrics is to ensure a DL model performs equally well for different groups. These metrics are stricter than SPD because they require consideration of both the sensitive attribute and the true labels, involving the measurement of the difference between the true positive rate and false positive rate for the *unprivileged* and *privileged* groups. AAOD, as an arithmetic average of the difference in false positive rates and the difference in true positive rates, effectively balances the effects of both difference metrics. This is supported by the experimental results, where AAOD generally exhibits higher differentiation compared to DEO and EOD. Therefore, AAOD appears to be a robust fairness evaluation metric compared to DEO and EOD, especially in scenarios where DEO and EOD show substantial discrepancies. AED measures the difference in misclassification rates of the model on different sensitive attribute groups. The major drawback of this metric is that its values are generally small, particularly on the CIFAR-10S dataset, resulting in lower differentiability compared to SPD and AAOD. Therefore, it is not considered an excellent fairness evaluation metric.

> **Finding 7:** AAOD emerges as the most promising and representative metric for evaluating model fairness.

When considering the two performance metrics, i.e., *accuracy* and *balanced accuracy*, they are very close to each other according to the results shown in Figure 1. In particular, on UTKFace and CIFAR-10S, accuracy and balanced accuracy results are almost equal. This indicates that the distribution of sensitive attributes on these datasets is balanced, in which case the two metrics are equivalent. However, on the CelebA dataset, we find that there is a significant difference between these two accuracy metrics, especially evident in the case of the FAAP method. It suggests that the FAAP method is highly sensitive to imbalanced class distributions in CelebA. We have mentioned earlier that FAAP, as a post-processing method,

heavily relies on modifying input data to improve fairness. However, the imbalanced data distribution may make the model more biased towards the majority class, thereby affecting its normal predictions. After class balancing, we observe a significant decrease in accuracy. From Figure 1, we also find that while FAAP exhibits substantial differences, there are still many methods that show consistent accuracy on the CelebA dataset, such as BM, BC+BB, and FR-P. Nevertheless, it is better to consider the two metrics together for a systematic and comprehensive analysis of the results.

> **Finding 8:** For an imbalanced dataset, accuracy and balanced accuracy tend to produce diverse results, and thus both of them should be used to comprehensively measure the performance of fairness improvement methods.

5.3 RQ3: Influence of datasets

Deep learning models tend to yield varying results on different datasets. An excellent fairness improvement method should perform competitively on different datasets. However, existing studies only use limited datasets. Therefore, this RQ aims to comprehensively analyze the performance of different fairness improving methods on various datasets.

For each method, we find that both accuracy and fairness vary significantly across different datasets according to the results shown in Table 3 and Figures 1. Regarding fairness, the results from Table 3 indicate that the overall fairness scores on CelebA and UTKFace-age are darker than the other two datasets. This suggests that for each method, fairness performance on CelebA and UTKFace-age is inferior to that on the other datasets. On the other hand, when considering accuracy, we have also observed that methods on the CelebA dataset have significantly higher accuracy compared to the other datasets, while CIFAR-10S exhibits the lowest overall accuracy. Particularly, we notice that the FR-B and FR-P methods, which have decent accuracy on other datasets, perform catastrophically poorly with accuracy below 0.70 on CIFAR-10S. Our analysis suggests that this should be due to the overreliance of FR methods on input data. CIFAR-10S, being an artificial dataset, has sensitive attributes that are artificially constructed (color or grayscale), and FR methods may not make appropriate modifications to such sensitive attributes during image alteration, thus disrupting the model's normal accuracy. Similarly, we find that what may be a superior method on one dataset doesn't necessarily hold when applied to a different dataset. For instance, by examining the table, we observe that on UTKFace-age, BC+BB exhibits significantly better fairness than DI, but conversely, on the other three datasets, DI shows substantial advantages. This indicates that existing methods are sensitive to the dataset they are applied to, and there is no best-performing method on all datasets.

> **Finding 9:** Existing methods are *sensitive* to different datasets in the evaluation, which are supported by two observations: 1) For a certain method, it tends to achieve different fairness and accuracy over different datasets; 2) For two different methods, one may outperform the other on certain dataset A, but underperform the other on another dataset B. Therefore, the conclusions from one certain dataset are hard to generalize to others.

There is another point worth our attention. In the UTKFace dataset, age and race, two sensitive attributes, yield markedly different results. When age is treated as the sensitive attribute, the overall accuracy and fairness of various methods are lower compared to when race is the sensitive attribute. This difference is particularly pronounced in terms of fairness. By observing the table, we can see that, even within the UTKFace dataset, the colors for age are noticeably darker than those for race, indicating the differences in fairness across various methods due to different sensitive attributes. Therefore, the choice of sensitive attributes is crucial in fairness tasks since different methods are suited for different ones.

5.4 RQ4: Efficiency of studied methods

The first three RQs explore the impact of different factors. In practical applications, the time efficiency of each method is also an important consideration for method selection. Therefore, our final RQ primarily investigates the time cost of different methods. Since the optimal experimental hyperparameters vary for different methods (such as batch size and epochs), we decide to measure the efficiency of each method by averaging the time cost for each epoch. It's important to note that all our experimental results are the averages obtained after conducting multiple experiments.

Table 4 presents the time cost per epoch of each fairness improving method. In the table, we highlight the highest efficiency in gray color regarding the complete training time, while **underline** the shortest time regarding a single epoch. Please note that analyzing both types (complete *vs* single) of time cost have their values. The complete training time reflects the overall complexity of methods under their optimal parameter settings, while the time per epoch indicates the intrinsic computation complexity of the optimizing process of the method's effectiveness in practical use.

From the table, we can observe that FDR is much more efficient than other approaches considering a single training epoch. The reason is that FDR improves model fairness only through last-layer fine-tuning by incorporating fairness constraints and data reweighting [53]. These operations only require updating the parameters of the last layer, resulting in very short runtime and high efficiency. In comparison, other in-processing methods incur significantly higher time overheads. This is primarily because these methods either introduce new loss functions and optimization objectives, design novel model modules, or employ techniques like distillation, all of which involve learning a large number of parameters during the training process, making it time-consuming. In contrast, considering the complete training time, there is no individual method outperforms all the others. For example, OS is the most efficient on UTKFace while FAAP is the most efficient on CIFAR-10S. As previously analyzed, BM emerges as the most effective method, but it is not the most efficient one over every studied datasets. Similarly, FDR, which exhibits overwhelming efficiency and good fairness improvement among the in-processing methods, comes at the cost of significant accuracy sacrifices. Therefore, end users should to choose the most suitable method based on the specific task's varying emphasis on effectiveness and efficiency.

Finding 10: An excellent fairness improving method should strike a balance between effectiveness and efficiency, ensuring

Table 4: Average time cost per epoch of each method for improving model fairness (in seconds).

Category	Method	CelebA		UTKFace Age		UTKFace Race		CIFAR-10S	
		t/epoch	epochs	t/epoch	epochs	t/epoch	epochs	t/epoch	epochs
Pre-proc.	US	81.4	170	11.5	400	15.6	120	14.8	2000
	OS	590.3	4	**16.6**	**7**	**13.7**	**10**	65.0	100
	UW	252.9	10	8.5	20	8.9	20	71.5	200
	BM	528.2	10	14.3	20	15.5	20	72.2	200
In-proc.	Adv	242.5	10	13.3	20	14.1	20	72.6	200
	DI	242.3	10	13.3	20	13.5	20	68.1	200
	BC+BB	564.3	10	23.3	20	23.5	20	67.9	200
	FLAC	277.2	10	14.5	20	13.4	20	40.5	200
	MFD	523.8	50	70.6	100	48.7	100	106.4	50
	FDR	**2.2**	**1000**	**0.6**	1500	**0.8**	1500	**1.7**	1000
Post-proc.	FR-B	379.8	20	41.8	20	42.7	20	76.7	20
	FR-P	586.3	20	46.8	20	43.4	20	74.0	20
	FAAP	288.1	50	17.1	50	17.5	50	**14.5**	**50**

that it not only improves fairness but also maintains acceptable levels of accuracy without consuming excessive computational resources.

6 IMPLICATIONS

Based on our result analysis, we have summarized several implications that can facilitate future research in this research area.

Combining the strength of different methods. According to our empirical results (i.e. Finding 1), there is no best-performing methods in all application scenarios and different methods are conceptually complementary. For example, pre-processing methods focus more on transforming or augmenting training data, while in-processing methods primarily optimize the design of objective functions and model structures. Our experimental results also confirm that. Therefore, it should be promising to combine the strength of individual methods since they improve the models from different perspectives, e.g., training data and training process. In particular, directly combining the results of different methods can be effective. However, investigating the deep combination of different methods by considering their core novelty still needs more exploration.

Developing more effective post-processing methods. With the rapid development and inspiring performance of large language models (LLMs), they have been adopted in various applications. However, re-training LLMs can be a hard or even impossible task for end users due to the high requirement on computing resources and the reliance on training data. Besides, for deployed models, post-processing methods are intuitively the best choice for improving model fairness. Nevertheless, according to Finding 5, the results shown in Table 3 demonstrate that existing post-processing methods are not satisfactory and still have much room for further improvement compared with the other two categories. Therefore, more studies should focus on the post-processing methods.

Understanding the source of unfairness. Based on the introduction of existing approaches, almost all of them aim to improve the model fairness in somehow a blind way without understanding the core reasons of the unfairness in DL models. Specifically, pre-processing methods assume that the training data contain bias while in-processing methods deem the bias is involved during the training process. However, even adopting the same training data and the same training process (see Section 5.3), the model performance regarding different attributes can also be different. Therefore, understanding the source of unfairness in the DL models may help

to better overcome this issue. In particular, model interpretation techniques [17, 74] can be incorporated to facilitate this task.

Incorporating the correlations among different attributes deeply. Indeed, sensitive attributes play a crucial role in model predictions regarding result fairness. However, according to Finding 4, simply removing the sensitive attributes does not necessarily improve model fairness, and incorporating the correlations between the sensitive attributes and the others is also important. In fact, the existing study [49] has already considered such information for improving model fairness. Nevertheless, it was designed for the structured (i.e., tabular) data, which should be much simpler than the images since the image features are not clearly defined. In other words, how to effectively identify the implicit correlations between different image features and incorporate them into the process of fairness improvement still requires more in-depth exploration.

7 THREATS TO VALIDITY

The *external* threats to validity mainly lie in the data selection in our study. In order to comprehensively evaluate the performance of existing methods, we have employed three widely-used datasets involving different sensitive attributes and classification tasks. In addition, according to our findings that different methods tend to have different effectiveness. Although our results cannot generalize to other datasets, the findings should not be affected.

The *internal* threat to validity mainly lies in the implementations of the studies methods. To mitigate this threat, we mainly employed the open-source implementations of the corresponding papers if available, while for other methods, we have double-checked their performance with the corresponding authors to ensure they are correctly implemented and configured. Furthermore, to ease the replication and promote future research, we have made all our experimental results and implementations open-source.

8 RELATED WORK

8.1 Fairness Improving Methods

Pre-processing Methods Pre-processing methods [54, 63, 64, 93, 98] aim to remove underlying discrimination by calibrating training data to eliminate spurious correlations and training fairer models on the modified data. Ramaswamy *et al.* [64] proposed a GAN-based data augmentation method to balance the training data. Yao *et al.* [93] proposed the methods to maintain useful information while filtering out bias information. Quadrianto *et al.* [63] learned a mapping from an input domain to a fair target domain to mitigate gender bias. Zhang *et al.* [98] employed adversarial examples to balance the training data for visual debiasing.

In-processing Methods In-processing techniques [45, 50, 52, 54, 58, 66, 90] focus on modifying and changing learning algorithms to remove discrimination during training. Lokhande *et al.* [52] offered a simplified approach by treating fairness measures as constraints on the model's output, which is incorporated through an augmented Lagrangian framework. Kehrenberg *et al.* [45] proposed NIFR model to learn invariant representations to improve algorithmic fairness. Roh *et al.* [66] introduced FairBatch, a bilevel optimization-based batch selection algorithm that adaptively chooses minibatch sizes to improve model fairness without requiring modifications to model training. Xu *et al.* [90] proposed a false positive rate penalty loss to

mitigate bias in face recognition by increasing the consistency of instance false positive rate. Lin *et al.* [50] proposed FairGRAPE, a pruning method that minimizes the disproportionate impacts by calculating per-group weight importance and preserving relative between-group importance during network edge pruning. Park *et al.* [58] proposed Fair Supervised Contrastive Loss for fair visual representation learning. This line of work aims at getting a fairer model by explicitly changing the training procedure.

Post-processing Methods Post-processing work [46, 51, 54, 85] focused on adjusting model predictions based on specific fairness criteria after the model training, which often use a holdout set which is not involved during the training procedure. Lohia *et al.* [51] introduced a technique that detects fairness bugs from outputs and makes appropriate adjustments. Kim *et al.* [46] introduced a technique that creates a new classifier with equal accuracy across different protected attributes. Wang *et al.* [85] presented an approach that learns to perturb input data, rendering the models incapable of recognizing fairness-related features.

8.2 Empirical Studies of Model Fairness

Empirical studies are essential for evaluating the effectiveness and efficiency of fairness improvement methods. Additionally, it can inspire researchers to pursue further innovations in fairness improvement. Most previous studies [4, 16, 25, 97] on fairness have focused on numerical or tabular inputs. However, fairness issues in image data are more complex. Wang *et al.* [86] conducted an analysis of computer vision models, specifically targeting adversarial training approaches, while Chen *et al.* [18] studied fairness improvement methods that work with tabular inputs. In contrast, our research delves into 13 approaches working with image inputs from three major categories. Additionally, we use three distinct datasets and encompass all commonly-used performance metrics.

9 CONCLUSION

Deep Learning (DL) models have been widely-adopted in many applications, particularly in ethical-sensitive domains. As a result, ensuring the fairness of DL models has been a emerging research problem. Although many fairness improving methods have been proposed in recent years, there still lacks a systematic empirical study to comprehensively compare the performance of different approaches. To fill this gap, this paper conducted the first large-scale empirical study, where we comprehensively analyzed the performance of 13 state-of-the-art methods from multiple aspects, e.g., metrics, datasets, and tasks. Based on the results, we have summarized a set of findings and implications to promote future studies in this research area.

DATA AVAILABILITY

For the sake of Open Science, we make the replication package with source code and the experimental results publicly available at: **https://github.com/junjie1003/DL-Fairness-Study**.

ACKNOWLEDGMENTS

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *ICML 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[2] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *ESEC/FSE 2019*, Marlon Dumas, Dietmar Pfahl, Sven Apel, and Alessandra Russo (Eds.). ACM, 625–635.

[3] Jing An, Lexing Ying, and Yuhua Zhu. 2021. Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients. In *ICLR 2021*. OpenReview.net.

[4] 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 *ESEC/FSE 2020*, Prem Devanbu, Myra B. Cohen, and Thomas Zimmermann (Eds.). ACM, 642–653.

[5] Sumon Biswas and Hridesh Rajan. 2021. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *ESEC/FSE 2021*, Diomidis Spinellis, Georgios Gousios, Marsha Chechik, and Massimiliano Di Penta (Eds.). ACM, 981–993.

[6] Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Comput. Linguistics* 19, 2 (1993), 263–311.

[7] Yuriy Brun and Alexandra Meliou. 2018. Software fairness. In *ESEC/FSE 2018*, Gary T. Leavens, Alessandro Garcia, and Corina S. Pasareanu (Eds.). ACM, 754–759.

[8] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106 (2018), 249–259.

[9] Jonathon Byrd and Zachary Chase Lipton. 2019. What is the Effect of Importance Weighting in Deep Learning?. In *ICML 2019 (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 872–881.

[10] Flávio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *NeurIPS 2017*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 3992–4001.

[11] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, and Daniele Regoli. 2021. The zoo of Fairness metrics in Machine Learning. *CoRR* abs/2106.00467 (2021).

[12] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *Comput. Surveys* (2020).

[13] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *FAccT 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, 319–328.

[14] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in machine learning software: why? how? what to do?. In *ESEC/FSE 2021*, Diomidis Spinellis, Georgios Gousios, Marsha Chechik, and Massimiliano Di Penta (Eds.). ACM, 429–440.

[15] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: a way to build fair ML software. In *ESEC/FSE 2020*, Prem Devanbu, Myra B. Cohen, and Thomas Zimmermann (Eds.). ACM, 654–665.

[16] Joymallya Chakraborty, Tianpei Xia, Fahmid M. Fahid, and Tim Menzies. 2019. Software Engineering for Fairness: A Case Study with Hyperparameter Optimization. *CoRR* abs/1905.05786 (2019).

[17] Runjin Chen, Hao Chen, Ge Huang, Jie Ren, and Quanshi Zhang. 2019. Explaining Neural Networks Semantically and Quantitatively. In *ICCV 2019*. IEEE, 9186–9195.

[18] Zhenpeng Chen, Jie Zhang, Federica Sarro, and Mark Harman. 2024. Fairness Improvement with Multiple Protected Attributes: How Far Are We?. In *ICSE 2024*. ACM.

[19] 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).

[20] 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 *ESEC/FSE 2022*, Abhik Roychoudhury, Cristian Cadar, and Miryung Kim (Eds.). ACM, 1122–1134.

[21] David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *ACL 2005*, Kevin Knight, Hwee Tou Ng, and Kemal Oflazer (Eds.). The Association for Computer Linguistics, 263–270.

[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *ITCS 2012*, Shafi Goldwasser (Ed.). ACM, 214–226.

[23] 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, Sydney, NSW, Australia, August 10-13, 2015*, Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams (Eds.). ACM, 259–268.

[24] Anthony Finkelstein, Mark Harman, S. Afshin Mansouri, Jian Ren, and Yuanyuan Zhang. 2008. "Fairness Analysis" in Requirements Assignments. In *RE 2008*. IEEE Computer Society, 115–124.

[25] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *FAccT 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, 329–338.

[26] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *ESEC/FSE 2017*, Eric Bodden, Wilhelm Schäfer, Arie van Deursen, and Andrea Zisman (Eds.). ACM, 498–510.

[27] Jorge Galindo and Pablo Tamayo. 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational economics* 15 (2000), 107–143.

[28] Rafael C. González and Richard E. Woods. 2008. *Digital image processing, 3rd Edition.* Pearson Education.

[29] Przemyslaw A. Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying Fairness and Explainability in Supervised Learning. In *FAccT 2022*. ACM, 1905–1916.

[30] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[31] Huizhong Guo, Jinfeng Li, Jingyi Wang, Xiangyu Liu, Dongxia Wang, Zehong Hu, Rong Zhang, and Hui Xue. 2023. FairRec: Fairness Testing for Deep Recommender Systems. In *ISSTA 2023*, René Just and Gordon Fraser (Eds.). ACM, 310–321.

[32] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS 2016*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3315–3323.

[33] Mercedeh Harirchian, Fereshteh Amin, Saeed Rouhani, Aref Aligholipour, and Vahid Amiri Lord. 2022. AI-enabled exploration of Instagram profiles predicts soft skills and personality traits to empower hiring decisions. *arXiv preprint arXiv:2212.07069* (2022).

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016*. IEEE Computer Society, 770–778.

[35] Youngkyu Hong and Eunho Yang. 2021. Unbiased Classification through Bias-Contrastive and Bias-Balanced Learning. In *NeurIPS 2021*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 26449–26461.

[36] Aziz Z Huq. 2018. Racial equity in algorithmic criminal justice. *Duke LJ* 68 (2018), 1043.

[37] Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intell. Data Anal.* 6, 5 (2002), 429–449.

[38] Jiajun Jiang, Junjie Yang, Yingyi Zhang, Zan Wang, Hanmo You, and Junjie Chen. 2023. A Post-Training Framework for Improving the Performance of Deep Learning Models via Model Transformation. *ACM Transactions on Software Engineering and Methodology* (2023).

[39] Frederick P. Brooks Jr. 1987. No Silver Bullet - Essence and Accidents of Software Engineering. *Computer* 20, 4 (1987), 10–19.

[40] Sangwon Jung, Sanghyuk Chun, and Taesup Moon. 2022. Learning Fair Classifiers with Partially Annotated Group Labels. In *CVPR 2022*. IEEE, 10338–10347.

[41] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. 2021. Fair Feature Distillation for Visual Recognition. In *CVPR 2021*. Computer Vision Foundation / IEEE, 12115–12124.

[42] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33, 1 (2011), 1–33.

[43] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In *ICDM 2012*, Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu (Eds.). IEEE Computer Society, 924–929.

[44] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *ECML PKDD 2012 (Lecture Notes in Computer Science, Vol. 7524)*, Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.). Springer, 35–50.

[45] Thomas Kehrenberg, Myles Bartlett, Oliver Thomas, and Novi Quadrianto. 2020. Null-Sampling for Interpretable and Fair Representations. In *ECCV 2020 (Lecture Notes in Computer Science, Vol. 12371)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 565–580.

[46] Michael P. Kim, Amirata Ghorbani, and James Y. Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *AIES 2019*, Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor (Eds.). ACM, 247–254.

[47] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MTSummit 2005*. 79–86.

[48] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007.

Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007*, John Carroll, Antal van den Bosch, and Annie Zaenen (Eds.). The Association for Computational Linguistics.

[49] 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 *ICSE 2022*. ACM, 2215–2227.

[50] Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. 2022. FairGRAPE: Fairness-Aware GRAdient Pruning mEthod for Face Attribute Classification. In *ECCV 2022 (Lecture Notes in Computer Science, Vol. 13673)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 414–432.

[51] Pranay Kr. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2019. Bias Mitigation Post-processing for Individual and Group Fairness. In *ICASSP 2019*. IEEE, 2847–2851.

[52] Vishnu Suresh Lokhande, Aditya Kumar Akash, Sathya N. Ravi, and Vikas Singh. 2020. FairALM: Augmented Lagrangian Method for Training Fair Models with Little Regret. In *ECCV 2020 (Lecture Notes in Computer Science, Vol. 12357)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 365–381.

[53] Yuzhen Mao, Zhun Deng, Huaxiu Yao, Ting Ye, Kenji Kawaguchi, and James Zou. 2023. Last-Layer Fairness Fine-tuning is Simple and Effective for Neural Networks. *CoRR* abs/2304.03935 (2023).

[54] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2022), 115:1–115:35.

[55] Michele Merler, Nalini K. Ratha, Rogério Schmidt Feris, and John R. Smith. 2019. Diversity in Faces. *CoRR* abs/1901.10436 (2019).

[56] Giang Nguyen, Sumon Biswas, and Hridesh Rajan. 2023. Fix Fairness, Don't Ruin Accuracy: Performance Aware Fairness Repair using AutoML. In *ESEC/FSE 2023*, Satish Chandra, Kelly Blincoe, and Paolo Tonella (Eds.). ACM, 502–514.

[57] Girish Keshav Palshikar, Rajiv Srivastava, Mahek Shah, and Sachin Pawar. 2018. Automatic Shortlisting of Candidates in Recruitment.. In *ProfS/KG4IR/Data: Search@ SIGIR*. 5–11.

[58] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2022. Fair Contrastive Learning for Facial Attribute Classification. In *CVPR 2022*. IEEE, 10379–10388.

[59] Ankita Ramjibhai Patel, Jaganmohan Chandrasekaran, Yu Lei, Raghu N. Kacker, and D. Richard Kuhn. 2022. A Combinatorial Approach to Fairness Testing of Machine Learning Models. In *ICST Workshops 2022*. IEEE, 94–101.

[60] 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.

[61] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *NeurIPS 2017*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5680–5689.

[62] Maan Qraitem, Kate Saenko, and Bryan A. Plummer. 2022. Bias Mimicking: A Simple Sampling Approach for Bias Mitigation. *CoRR* abs/2209.15605 (2022).

[63] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. 2019. Discovering Fair Representations in the Data Domain. In *CVPR 2019*. Computer Vision Foundation / IEEE, 8227–8236.

[64] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. 2021. Fair Attribute Classification Through Latent Space De-Biasing. In *CVPR 2021*. Computer Vision Foundation / IEEE, 9301–9310.

[65] Harriet Rodney, Katarina Valaskova, and Pavol Durana. 2019. The artificial intelligence recruitment process: How technological advancements have reshaped job application and selection practices. *Psychosociological Issues in Human Resource Management* 7, 1 (2019), 42–47.

[66] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. FairBatch: Batch Selection for Model Fairness. In *ICLR 2021*. OpenReview.net.

[67] W. W. Royce. 1987. Managing the Development of Large Software Systems: Concepts and Techniques. In *ICSE 1987*, William E. Riddle, Robert M. Balzer, and Kouichi Kishida (Eds.). ACM Press, 328–339.

[68] Kari Saarinen. 1994. Image processing, analysis and machine vision : by Milan Sonka, Vaclav Hlavac and Roger Boyle. Publisher: Chapman & Hall, 2-6 Boundary Row, London SE1 8HN, UK, 1993, xix+555 pp., ISBN 0-412-45570-6. *Signal Process.* 35, 1 (1994), 102–104.

[69] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *CoRR* abs/1911.08731 (2019).

[70] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. *CoRR* abs/2005.04345 (2020).

[71] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *SIGMOD 2019*, Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska (Eds.). ACM, 793–810.

[72] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. 2023. FLAC: Fairness-Aware Representation Learning by Suppressing Attribute-Class Associations. *CoRR* abs/2304.14252 (2023).

[73] Qingchao Shen, Junjie Chen, Jie M. Zhang, Haoyu Wang, Shuang Liu, and Menghan Tian. 2022. Natural Test Generation for Precise Testing of Question Answering Software. In *ASE 2022*. ACM, 71:1–71:12.

[74] Wen Shen, Zhihua Wei, Shikun Huang, Binbin Zhang, Jiaqi Fan, Ping Zhao, and Quanshi Zhang. 2021. Interpretable Compositional Convolutional Neural Networks. In *IJCAI 2021*, Zhi-Hua Zhou (Ed.). ijcai.org, 2971–2978.

[75] Ezekiel O. Soremekun, Sakshi Udeshi, and Sudipta Chattopadhyay. 2022. Astraea: Grammar-Based Fairness Testing. *IEEE Trans. Software Eng.* 48, 12 (2022), 5188–5211.

[76] Maya Steinitz. 2014. The case for an international court of Civil Justice. *Stan. L. Rev. Online* 67 (2014), 75.

[77] Bing Sun, Jun Sun, Long H. Pham, and Tie Shi. 2022. Causality-Based Neural Network Repair. In *ICSE 2022*. ACM, 338–349.

[78] Zeyu Sun, Jie M Zhang, Mark Harman, Mike Papadakis, and Lu Zhang. 2020. Automatic testing and improvement of machine translation. In *ICSE 2020*. 974–985.

[79] Zeyu Sun, Jie M Zhang, Yingfei Xiong, Mark Harman, Mike Papadakis, and Lu Zhang. 2022. Improving machine translation systems via isotopic replacement. In *ICSE 2022*. 1181–1192.

[80] Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. 2022. Fairness-aware Configuration of Machine Learning Libraries. In *ICSE 2022*. ACM, 909–920.

[81] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous Deep Transfer Across Domains and Tasks. In *ICCV 2015*. IEEE Computer Society, 4068–4076.

[82] Patrick van Esch, J Stewart Black, and Denni Arli. 2021. Job candidates' reactions to AI-enabled job application processes. *AI and Ethics* 1 (2021), 119–130.

[83] Patrick Van Esch, J Stewart Black, and Joseph Ferolie. 2019. Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior* 90 (2019), 215–222.

[84] Gang Wang, Jinxing Hao, Jian Ma, and Hongbing Jiang. 2011. A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications* 38, 1 (2011), 223–230.

[85] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. 2022. Fairness-aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models. In *CVPR 2022*. IEEE, 10369–10378.

[86] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. In *CVPR 2020*. Computer Vision Foundation / IEEE, 8916–8925.

[87] Zan Wang, Ming Yan, Junjie Chen, Shuang Liu, and Dongdi Zhang. 2020. Deep learning library testing via effective model generation. In *ESEC/FSE 2020*, Prem Devanbu, Myra B. Cohen, and Thomas Zimmermann (Eds.). ACM, 788–799.

[88] Hilde J. P. Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *CoRR* abs/2303.16626 (2023).

[89] Wikipedia contributors. 2023. Pearson correlation coefficient — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=1177721736.

[90] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. 2021. Consistent Instance False Positive Improves Fairness in Face Recognition. In *CVPR 2021*. Computer Vision Foundation / IEEE, 578–586.

[91] Zichun Xu. 2022. Human Judges in the era of artificial intelligence: challenges and opportunities. *Applied Artificial Intelligence* 36, 1 (2022), 2013652.

[92] Lin Yang, Junjie Chen, Hanmo You, Jiachen Han, Jiajun Jiang, Zhe Sun, Xinqi Lin, Fang Liang, and Yuning Kang. 2023. Can Code Representation Boost IR-Based Test Case Prioritization?. In *ISSRE 2023*. 240–251. https://doi.org/10.1109/ISSRE59848.2023.00077

[93] Ruichen Yao, Ziteng Cui, Xiaoxiao Li, and Lin Gu. 2022. Improving Fairness in Image Classification via Sketching. *CoRR* abs/2211.00168 (2022).

[94] Hanmo You, Zan Wang, Junjie Chen, Shuang Liu, and Shuochuan Li. 2023. Regression Fuzzing for Deep Learning Systems. In *ICSE 2023*. IEEE, 82–94.

[95] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AIES 2018*, Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi (Eds.). ACM, 335–340.

[96] Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. 2022. Fairness Reprogramming. In *NeurIPS*.

[97] Mengdi Zhang and Jun Sun. 2022. Adaptive fairness improvement based on causality analysis. In *ESEC/FSE 2022*, Abhik Roychoudhury, Cristian Cadar, and Miryung Kim (Eds.). ACM, 6–17.

[98] Yi Zhang and Jitao Sang. 2020. Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing. In *MM 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 4346–4354.