Fairness without Sensitive Attributes via Knowledge Sharing

Hongliang Ni
The University of Queensland
Brisbane, Queensland, Australia

Lei Han
The University of Queensland
Brisbane, Queensland, Australia

Tong Chen
The University of Queensland
Brisbane, Queensland, Australia

Shazia Sadiq
The University of Queensland
Brisbane, Queensland, Australia

Gianluca Demartini
The University of Queensland
Brisbane, Queensland, Australia

ABSTRACT

While model fairness improvement has been explored previously, existing methods invariably rely on adjusting explicit sensitive attribute values in order to improve model fairness in downstream tasks. However, we observe a trend in which sensitive demographic information becomes inaccessible as public concerns around data privacy grow. In this paper, we propose a confidence-based hierarchical classifier structure called "Reckoner" for reliable fair model learning under the assumption of missing sensitive attributes. We first present results showing that if the dataset contains biased labels or other hidden biases, classifiers significantly increase the bias gap across different demographic groups in the subset with higher prediction confidence. Inspired by these findings, we devised a dual-model system in which a version of the model initialised with a high-confidence data subset learns from a version of the model initialised with a low-confidence data subset, enabling it to avoid biased predictions. Our experimental results show that Reckoner consistently outperforms state-of-the-art baselines in COMPAS dataset and New Adult dataset, considering both accuracy and fairness metrics.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

KEYWORDS

Fairness, Fairness in AI, Fairness without Sensitive Attributes

ACM Reference Format:
Hongliang Ni, Lei Han, Tong Chen, Shazia Sadiq, and Gianluca Demartini. 2024. Fairness without Sensitive Attributes via Knowledge Sharing. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3630106.3659014

1 INTRODUCTION

Automated models and algorithms have found wide application in various domains, including finance and justice, as tools to assist human decision-making processes [7, 11]. These applications collect information like age and education level in financial services, or misconduct incidents in policing, as well as sensitive data like race and gender from individuals, raising concerns about the ability of automation to deliver accurate and equitable judgments across diverse demographic groups [1]. Due to concerns about the misuse of private data, increasing regulatory restrictions have made it more challenging to access and make use of sensitive information in automated decision making [6, 37]. Approaches to improve fairness by not making use of sensitive data can be roughly divided into two categories: (1) focusing on ensuring that accuracy-related utility is equal across various demographic groups [24, 29, 38] and (2) focusing on limiting the impact of sensitivity-correlated proxies on predictions [16, 19, 22, 39, 41]. By relying on the correlated observed attributes, these methods have the potential benefit of mitigating bias.

However, we argue that fairness is still underachieved because of the unfair data being used. Most current approaches are limited by the presence of biased labels and other hidden bias concealed within the training data. For example, when faced with judges who are either overly strict or lenient, there may be unusual fluctuations in the number and severity of crimes committed by offenders, leading to biased labels [18]. In the supervised learning setting, the model, in an effort to minimize the loss with respect to the biased ground truth, may learn unnecessary biases [29]. With such data, attempts to obtain auxiliary information, about demographic groups to assist in tasks can unfortunately fail [29]. Moreover, approaches relying on selected proxy combinations are difficult to scale for sparse datasets and also difficult to be applied on unstructured data such as image and audio. [15, 41]

In this paper, we aim to answer the following research question: *How can we enhance the fairness of algorithmic predictions when we exclude sensitive information from the dataset?*. We begin by presenting a model confidence study. Confidence calibration is important for classification tasks [21]; however we do not focus on adjusting confidence but rather use the lens of confidence to explore model behaviour. We observe that as the model makes predictions with higher confidence scores, the bias gap between demographic groups increases as well. On the other hand, the analysis of feature value distributions on the COMPAS dataset reveals that, for subsets with higher confidence, the distribution pattern of selected features becomes more distinctive. Combining both analytical results, we argue that in a supervised learning setting, when the model tries to minimise the loss with respect to the training labels, it also learns the bias present in these biased labels. This results in subsets with easier-to-classify samples, yet with predictions which are less fair.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659014

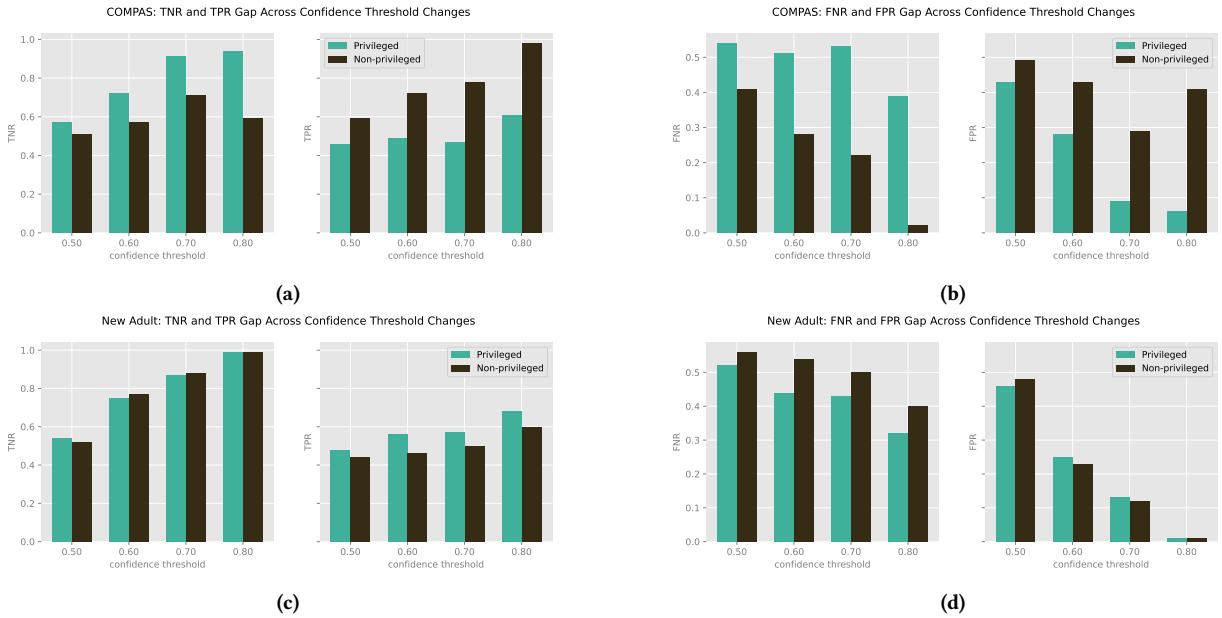


Figure 1: Observed measure factor gaps derived from the confusion matrix of the trained logistic regression classifier. (a) and (c) True Negative Rate (TNR) and True Positive Rate (TPR) gaps for two demographic groups across different confidence levels in COMPAS dataset and New Adult dataset, respectively. (b) and (d) False Negative Rate (FNR) and False Positive Rate (FPR) gaps for two demographic groups across different confidence levels in the COMPAS dataset and the New Adult dataset, respectively.

Inspired by these findings, we initially divide the original training set into two subsets based on confidence scores obtained from a simple linear classifier, and then initialise different classifiers with corresponding data subsets. After that, we introduce learnable noise into the original data, aiming to retain only the necessary information for prediction. In the next phase, one classifier acquires knowledge from the other classifier to learn fairness while also updating itself using the training labels to maintain high levels of effectiveness.

The main contributions of this paper are as follows:

- By analysing bias and the distribution of non-sensitive attributes across demographic groups in different model confidence intervals, we observe that as the model becomes more certain, it tends to make biased predictions more easily. This analysis reveals how biased labels or other hidden bias adversely affect fairness in machine learning, even when there are no sensitive attributes considered during the supervised learning process.
- We introduce a novel confidence-based classification framework, named Reckoner. This framework achieves improved group fairness while maintaining accurate classifications by utilising learnable noise and knowledge-sharing in a dual-model system architecture. This provides an effective approach to improve fairness without using sensitive attributes.
- We conduct extensive experiments to evaluate the effectiveness of the proposed framework as compared to other baselines in terms of fairness and predictive performance on datasets from which sensitive information is removed. We also present the results of an ablation study to understand the impact on effectiveness of the components in the proposed framework.

2 PROBLEM SETUP AND FAIRNESS METRICS

Problem definition. Our goal is to improve group fairness in prediction tasks in a non-sensitive attributes setting, where a set of labeled data $D = \{x_i, y_i\}_{i=1}^{N}$ is available for training. Each $x_i \in \mathbb{R}^{1\times m}$ is a m-dimensional data instance, and we use $F = \{f_i, \ldots, f_m\}$ to denote the m features. Sensitive attributes S are not used in training, i.e. $S \notin F$. Following the task settings on COMPAS dataset[31], New Adult dataset[17] and CelebA dataset[33], we focus on binary classification problems, i.e., $y_i \in \{0, 1\}$.

Group Fairness Metrics. We aim to reduce the difference in model predictions across various demographic groups. In this paper, we use two group fairness metrics for evaluation: Demographic Parity [12] and Equalised Odds [3]. Both fairness metrics are considered better when they have lower values.

- Demographic Parity measures the difference in favourable outcomes between privileged and non-privileged classes:

$$\Delta_{DP} = p(\hat{y} = 1|x_s = s_i) - p(\hat{y} = 1|x_s = s_j) \tag{1}$$

- Equalised Odds measures the difference in true positive rates and false positive rates, aiming for equality between privileged and non-privileged classes:

$$\begin{aligned}\Delta_{EOdds} =& \frac{1}{2}|p(\hat{y} = 1|x_s = s_i, y = 1) - p(\hat{y} = 1|x_s = s_j, y = 1)| + \\ & \frac{1}{2}|p(\hat{y} = 1|x_s = s_i, y = 0) - p(\hat{y} = 1|x_s = s_j, y = 0)|,\end{aligned} \tag{2}$$

	Previous Misconduct	Charge Degree	.....	Age	Race	Recidivist
Offender A	0	Felony	.....	23	White	True
Offender B	3	Felony	.....	22	Black	False
Offender C	2	Misdemeanor	.....	30	Black	True

(a)

	Education	Occupation	.....	Age	Race	Income Label
Individual A	Bachelors	Engineer	.....	23	White	False
Individual B	High School	Sales	.....	22	Black	False
Individual C	Master	Professor	.....	30	Black	True

(b)

Figure 2: (a) An example of the COMPAS dataset. In our experiment, the attribute 'Race' in the red box is not used. (b) An example of the New Adult dataset. In our experiment, the attribute 'Race' in the red box is not used.

where s_i and s_j are arbitrary sensitive attributes from sensitive attribute set S and $\hat{y}$ is prediction from the classifier. x_s specifies which demographic group data instance x belongs to. In addition to group fairness metrics, we employed the bias gap illustrated in Figure 1 in the confidence analysis in Section 3. Taking Figure 1 as an example, we analysed the bias gap of True Positive Rate (TPR) exhibited by two demographic groups: $\Delta_{TPR} = TPR_{s_i} - TPR_{s_j}$. On the other three measure factors, we applied a similar definition to obtain the corresponding bias gap results.

3 ANALYSIS OF BIAS GAP ACROSS CONFIDENCE LEVELS

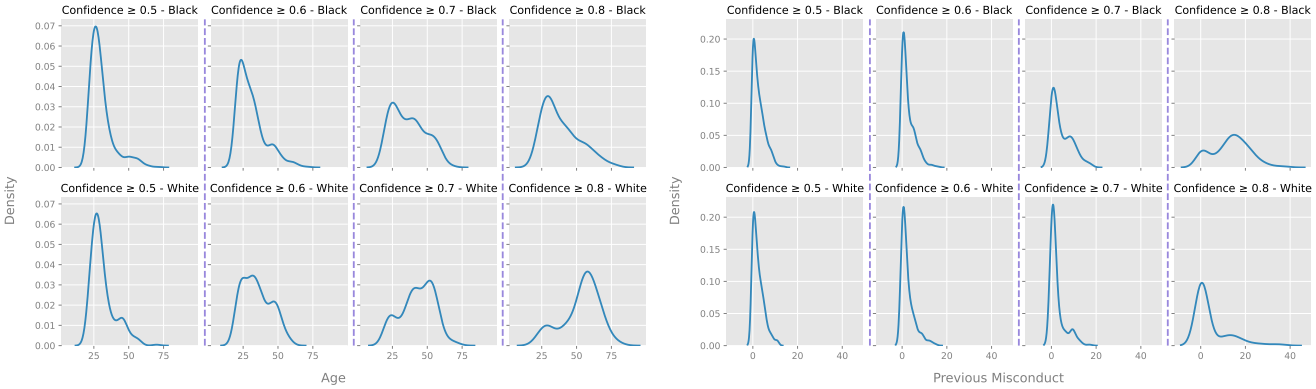
Given current confidence-based approaches, such as Out-of-Distribution (OOD) detection [26] and those for image classification [13, 14], we are interested in studying the behaviour of the classifier in subsets with varying levels of confidence. Viewing the data through the lens of confidence provides a more detailed picture of bias patterns across various datasets. In this analysis, we employ logistic regression on two real-world datasets: COMPAS [31] and New Adult [17]. Figure 2a provides a toy example using the COMPAS dataset, which is used for predicting offender recidivism. In our setup, sensitive information within the red dashed box is omitted from training. As a consequence, we partitioned the testing dataset into four disjoint subsets using confidence thresholds of 0.5, 0.6, 0.7, and 0.8, respectively. This threshold setting ensures a sufficient number of samples in each subset.

As shown in Figure 1, for COMPAS dataset, as the confidence threshold increases, the performance differences between demographic groups become more pronounced. Moreover, the gap in FNR and FPR values grows more significantly compared to TNR and TPR values. When the confidence is below 0.6, the gap between the two demographic groups is minimal across all four measures, with the largest bias gap being only 13%. When the confidence is above 0.8, we expect the classifier in this subset to accurately capture classification patterns. Ideally, there should be high TNR and TPR, the lowest FNR and FPR, and the smallest gap between the two demographic groups. However, the results show that although the first two criteria meet our expectations, the bias gap fluctuates between remarkable values between 35% and 37%. Additionally, the classifier has been found to perform better when identifying positive instances in the non-privileged group, as False Negative Rate (FNR) in the privileged group gradually decreases (from 0.54 to 0.39), while it sharply drops from 0.41 to 0.02 for the non-privileged group.

In the New Adult dataset, we can also observe the minimum bias gap in the low-confidence subsets in terms of TPR and FNR. Specifically, both the privileged group and the non-privileged group show an increasing trend in TNR and a sharp decreasing trend in FPR. This indicates that the performance of the classifier in identifying negative samples in both groups improves, and the performance gaps are small, achieving the ideal performance in the high-confidence subset mentioned above. In contrast, the model struggles in identifying samples where the salary exceeds the set income threshold. Plots of TPR and FNR values also show that bias gaps for these two metrics increase with increasing model confidence. The differences in bias gaps on the two datasets may be attributed to COMPAS being a dataset with biased labels [18]. The classifier may mistakenly treat bias as knowledge to learn when it tries to minimise the loss based on the available ground truth labels. Therefore, we observe a significant disparity in bias gap between the low-confidence subset (samples near the decision boundary, confidence < 0.6) and the high-confidence subsets (samples away from the decision boundary, confidence ≥ 0.6).

In summary, on both datasets, we observe that bias gaps tends to be relatively smaller in low-confidence subsets for all or some measure factors of classification results, and it increases as confidence levels rise. This indicates that models learn to make biased decision with high confidence from the biased labels present in the training data.

To further investigate why the model performs sub-optimally on COMPAS, we selected two non-sensitive attributes from the dataset to understand their distribution patterns across different confidence subsets. Figure 3a and Figure 3b show the distribution patterns of the 'Age' attribute and 'Previous Misconduct' in various subsets, partitioned based on confidence scores. We can observe that different racial backgrounds exhibit distinguishable distribution patterns within high-confidence subsets, while these differences are not observed in the low-confidence subset. Specifically, within high-confidence subsets, comprising approximately 65% of the testing data, there are varying tendencies of age distribution dispersion and right-skewness among different racial groups. However, we only observe a tendency toward right-skewness in age distribution of low-confidence subset. Similar patterns can be observed in the distribution of previous misconduct. We suspect that this is caused by biased labels and other hidden bias. The classifier, by making biased predictions toward the majority, is enabled to minimise the training loss based on predicting distinguishable patterns from most instances.



(a) (b)

Figure 3: (a) and (b) Distributions of the attribute 'Age' and 'Previous Misconduct' across different subsets of the testing set.

4 METHOD

Overview. As shown in Figure 4, our proposed method consists of two training stages, the Identification stage and Refinement stage. In the Identification stage (Sec. 4.2), we employ a simple linear classifier, such as logistic regression, to perform a binary classification task on the raw dataset under a supervised learning setting. Then the training data is split into two subsets based on a predefined confidence threshold: a high-confidence subset and a low-confidence subset. These subsets are then used to initialise their respective classifiers, both of which are three-layer multilayer perceptrons (MLPs). At the beginning of the Refinement stage (Sec. 4.3), we introduce learnable noise to the original dataset, generating noise-augmented data for training in this stage. Next, during the iterations, the *High-Confidence* (or "*High-Conf*") classifier produces pseudo-labels to train the *Low-Confidence* (or "*Low-Conf*") classifier, and the *Low-Conf* classifier updates its knowledge back to the *High-Conf* classifier. The *Low-Conf* classifier is trained for a limited number of iterations, for example, three epochs, before reverting to its initialised state. This approach ensures accurate and unbiased prediction while maintaining an efficient training process. Lastly, the *High-Conf* classifier uses ground truth and shared knowledge to update its parameters.

4.1 Motivation

Biased labels and other hidden bias concealed within the dataset results in the classifier learning misleading classification patterns and feature distributions, leading to 1) model-learned parameters that do not enable fair predictions, and 2) limited auxiliary information that non-sensitive attributes can provide. The motivation behind our proposed framework stems from the analytical findings in Section 3. We observed the smallest bias gap in the results of the classifier on low-confidence subsets. This insight can be utilised to introduce regularisation to the classifier's learning based on the knowledge acquired from low-confidence subsets. We also introduced learnable noise, to obtain more reliable auxiliary information. To improve fairness in the settings of missing sensitive attributes, we design a novel framework, named Reckoner, which seamlessly integrates learnable noise and a knowledge-sharing mechanism between dual models. We have demonstrated the necessity of combining these two components in our ablation study (see Section 5.2).

4.2 Identification stage

In this stage, we perform a simple confidence-based sample split on the training data to obtain high-confidence samples and low-confidence samples. Specifically, we train a logistic regression classifier on the original training set and we split the data using confidence threshold (in this case, it is set to 0.6 following the results in [30]). Our hypothesis is that the model trained on the low-confidence subset is more inclined towards classifications with smaller bias gap, even if its predictive accuracy is relatively modest. However, by integrating the knowledge derived from the model trained on the high-confidence subset, it is possible to enhance prediction effectiveness while maintaining good fairness.

4.3 Refinement stage

4.3.1 Learnable noise. In the initial phase of this stage, learnable noise is introduced into the training set. Intuitively, learnable noise has a similar effect to L1 regularisation. L1 regularisation is employed to encourage sparsity in the model parameters, resulting in a model that utilises only a subset of the initial features. This is beneficial for reducing bias gaps in model predictions because, even if the dataset does not contain sensitive information, the associated non-sensitive information may still be biased. Therefore, employing feature selection methods becomes necessary. However, we do not directly use L1 regularisation here because it is constrained by the model structure and feature scale thus being less generalisable. L1 regularisation imposes a penalty on the weight of each feature, which is the sum of the absolute values of the weights. Consequently, changes to the model structure or adjustments to the feature scale can change the weights, affecting the strength of L1 regularisation. On the other hand, adding learnable noise to input data can be considered as disrupting the original inputs to increase the sparsity. Through supervised learning, the model is encouraged to select features that have higher cross-entropy with the ground truth. Moreover, learnable noise is robust to changes in model structure or feature scale, allowing us to leverage this

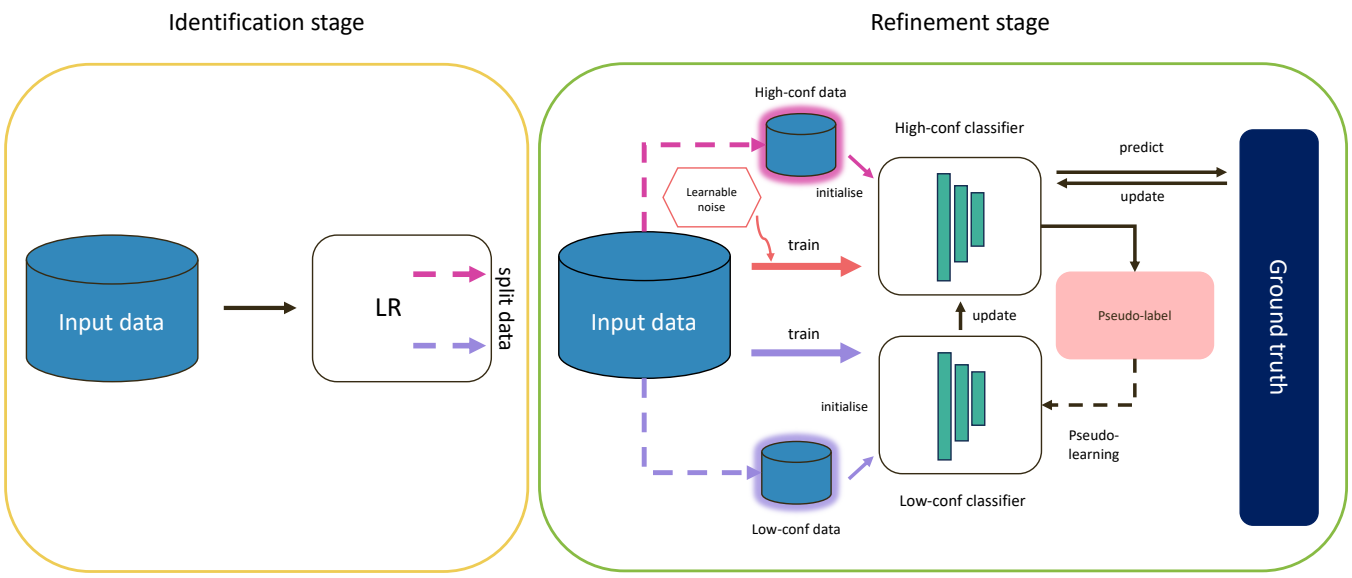


Figure 4: Overview of *Reckoner*. Reckoner consists of two stages. *Identification stage*: we first train a logistic regression classifier on the raw data, and then split the data based on confidence scores. In *Refinement stage*, we introduce learnable noise into the original dataset. We employ two classifiers, one for low-confidence instances and another for high-confidence ones. The *Low-Conf* classifier uses pseudo-labels produced by the *High-Conf* classifier for limited training times and restores for each new data. Knowledge acquired during this process is then shared with the *High-Conf* classifier, which incorporates ground truth data to refine its model weights.

property to design a more flexible framework where the classifier can be any model suitable for downstream tasks. To be more specific, we add a noise wrapper to vectors of the same dimensions as the input (denoted as η). The noise wrapper is a simple two-layer MLP that is subsequently applied to modify the input. The resulting modified input is referred to as a noise-augmented input and can be represented as follows:

$$\tilde{x}_i = x_i + tanh(g_\omega(\eta)), \tag{3}$$

where ω is the set of parameters in the noise wrapper g, and $tanh(\cdot)$ helps constrain the range of noise values to [-1, 1]. In the rest of the refinement stage, $\tilde{x}_i$ is the new input we use for the *High-Conf* classifier.

4.3.2 Dual-model and knowledge sharing. The lower predictive performance prevents us from relying on the *Low-Conf* classifier to perform classification tasks, but it can guide the *High-Conf* classifier to make fairer classifications. Within the Reckoner framework, the *Low-Conf* classifier relies on the pseudo-labels produced by the *High-Conf* classifier for supervised learning to update its own parameters. Since ground truth data is not used in this phase, the learning process can be referred to as *pseudo-learning*. The reason we do not use ground truth here is that some datasets with hidden bias contain a considerable amount of biased labels which may be propagated during model learning. If these labels were used during the learning phase of the *Low-Conf* classifier, it might lead the model to make biased predictions, thereby limiting its regularisation effect on the *High-Conf* classifier. We use binary cross entropy as the supervised loss of *pseudo-learning*:

$$\mathcal{L}_{\mathrm{L}} = BCE(f_{\Theta_{\mathrm{L}}}(x), \tilde{y}), \tag{4}$$

where $f_{\Theta_{\mathrm{L}}}$ is *Low-Conf* classifier and $\tilde{y}$ is the pseudo-label produced by *High-Conf* classifier. The subscripts "L" refers to the *Low-Conf* classifier. Note that during this phase, the training iterations of the *Low-Conf* classifier are limited (set to only 3 times in our experiments) for training efficiency. Furthermore, once these iterations end, the *Low-Conf* classifier has a rollback operation, reverting its parameters to their initialised values. This design maintains the effectiveness of *Low-Conf* classifier in providing fairness guidance to the parameters of *High-Conf* classifier, and avoids learning the bias inherent in the dataset.

On the other hand, we rely on the *High-Conf* classifier, which offers higher accuracy, to perform classification tasks. However, as revealed by the analysis in Section 3, we are aware of its poor performance in terms of decreasing bias. Previous work shows that leveraging the strengths of both models is a common strategy for various research problems, such as, for example, the high-pass and low-pass filters in graph neural networks [5], and efforts to average model weights for improved image classification [36]. The most promising improvement on fairness for the *High-Conf* classifier lies in integrating the knowledge from the *Low-Conf* classifier, and its parameter update mechanism can be expressed as follows:

$$\Theta_{\mathrm{H}} \leftarrow \alpha\Theta_{\mathrm{H}} + (1-\alpha)\Theta_{\mathrm{L}}, \tag{5}$$

where α controls the proportion of the knowledge of *High-Conf* classifier. The subscripts "H" and "L" refer to the *High-Conf* classifier and the *Low-Conf* classifier, respectively. In order to enhance

predictive accuracy, we use ground truths and employ the backpropagation algorithm to update the *High-Conf* classifier. By integrating the knowledge from the *Low-Conf* classifier, the final update mechanism can be formulated as follows:

$$\theta_i^{\mathrm{H}} \leftarrow \hat{\theta}_{i-1}^{\mathrm{H}} - \gamma \frac{\partial \mathcal{L}_{\mathrm{H}}}{\partial \hat{\theta}_{i-1}^{\mathrm{H}}}, \quad \hat{\theta}_{i-1}^{\mathrm{H}} \leftarrow \alpha \theta_{i-1}^{\mathrm{H}} + (1-\alpha)\theta_k^{\mathrm{L}}, \tag{6}$$

where θ_i^{H} is the weight of *High-Conf* classifier at i-th iteration, $\hat{\theta}_{i-1}^{\mathrm{H}}$ is the temporary weight integrating both *High-Conf* classifier and *Low-Conf* classifier knowledge controlled by α, and k is the iteration number when the *Low-Conf* classifier achieves the best performance during *pseudo-learning*. We also employ binary cross entropy as the supervised loss of classification task:

$$\mathcal{L}_{\mathrm{H}} = BCE(f_{\Theta_{\mathrm{H}}}(x), y). \tag{7}$$

Intuitively, the *pseudo-learning* applied to the *High-Conf* classifier can be interpreted as shifting the decision boundary closer to the feature space of the the major samples in high-confidence subsets, with the hyperparameter α controlling stability. Hence, the model will not misclassify similar instances based on distribution patterns of the majority. Another component of the framework, learnable noise, offers more auxiliary information for demographic groups, ensuring both accuracy and enhanced fairness. We will discuss the contribution of each of these two components to prediction fairness when discussing the results of the ablation study (see Section 5.2).

5 EXPERIMENTAL EVALUATION

Datasets. We validate our model on three benchmark datasets:(1) **New Adult**: as introduced by [17], it comprises of 49,531 samples, each associated with 14 attributes. The primary objective is to predict whether an individual's income exceeds 50k. In our experiments, we convert income into binary labels, and we set 'race' as the sensitive attribute and exclude it from our experiments. (2) **COMPAS**: COMPAS [31] comprises 7,215 data samples, each associated with 11 attributes. Following previous work on fairness without sensitive attributes [8], we have filtered this dataset to include only African American and Caucasian offenders, hence we use a modified dataset containing 6,150 samples. The primary objective is to predict whether an offender will commit another offense within two years. We set 'race' as the sensitive attribute and exclude it from our experiments. (3) **CelebA**: The CelebA dataset [33] comprises 202,599 image samples with resolution 178*218, each associated with 40 attributes. The primary objective is to predict the attractiveness of each image. We set 'gender' as the sensitive attribute and exclude it from our experiments.

Baselines. We compare our method with four related methods for comparisons: (1) Distributed Robust Optimisation (DRO) [24]: The primary objective of this method is to enhance Rawlsian Max-Min Fairness [35]. It specifically focuses on mitigating the prioritisation of benefits for the majority group that may arise from employing empirical risk minimisation. The approach establishes upper and lower bounds for the objective function of each group based on different group proportions, ensuring non-discrimination by the algorithm. (2) ARL [29]: This approach also aims at optimising Rawlsian Max-Min Fairness and uses adversary learning to optimise worst-case performance by prioritising instances with higher loss. (3) FairRF [41]: This approach identifies proxy features strongly correlated with sensitive attributes and minimises the correlation by re-weighting to achieve fairness. (4) [8]: This approach applies knowledge distillation requiring one model to produce soft labels, and uses them to train a second model to obtain a better decision boundary. It has two variants: either with labels determined by softmax function or linear function.

Experimental Setting. We employ feature hashing on categorical features if the dataset we use contains categorical features. For the proposed framework, we apply logistic regression to train a simple binary classifier and follow [30] by setting 0.6 as the confidence threshold for data splitting in the identification stage. In the refinement stage, we use each confidence-based subset and use 10% of the total model training iterations to initialise both the *High-Conf* classifier and the *Low-Conf* classifier. In the *pseudo-learning* phase, the *Low-Conf* classifier is trained three times. In the training for the whole proposed framework, we use Adam to be the optimiser and binary cross entropy for classification loss. We employ Resnet-50 [25] as the backbone of the classifiers for the CelebA dataset. For evaluation, we use Equalised Odds [3] and Demographic Parity [12] as fairness metrics and report accuracy for classification. Both fairness metrics are considered better when they have lower values.

5.1 Results

Tables 1 - 3 show the results comparing our models to other baselines. Results for both variants of [8] and of FairRF [41] are from [8] using the same datasets and same train-valid-test split. Our model outperforms the selected baselines. In the COMPAS dataset, we can observe that Reckoner achieves the best result in Equalised Odds with a relative improvement of about 2.84% over the best baseline, and also secures the second-best effectiveness in terms of accuracy. Compared to [8] with the optimal Demographic Parity, although our method exhibits a marginal difference of 1.2%, we obtain improvements in terms of accuracy and Equalised Odds, with improvements of 1.45% and 3.85%, respectively. In comparison to the highest accuracy achieved by ARL [29], Reckoner exhibits only a 0.4% gap in accuracy. However, it holds a significant edge in fairness, with improvements of 5.54% for Equalised Odds and 4.65% for Demographic Parity. In the New Adult dataset, Reckoner exhibits improvements in fairness compared to all the baselines. In comparison to the best-performing baselines in terms of fairness, it achieves a 4.51% improvement in Equalised Odds and a 0.43% improvement in Demographic Parity. At the same time, Reckoner achieves the second-highest position in terms of prediction accuracy, with a marginal 0.02% gap compared to the most accurate baseline. In the CelebA dataset [33], on the other hand, we obtain a third position in accuracy and Equalised Odds, but achieve the best Demographic Parity.

Note that we differentiate our study from those that emphasise equal accuracy-related utility across different demographic groups (such as the percentage of matched binary labels and user satisfaction [24] or AUC [29]). This branch of methods may not be very competitive in fairness performance with evaluations such as Equalised Odds and Demographic Parity, as these metrics are designed for Group Fairness methods. However, as mentioned before,

Table 1: Results on the COMPAS dataset.

Metrics(%) Methods	Accuracy	Equalised Odds	Demographic Parity
DRO [24]	64.88 ± 0.34%	23.11 ± 1.80%	25.32 ± 1.22%
ARL [29]	65.32 ± 0.70%	23.01 ± 1.21%	25.37 ± 1.01%
FairRF [41]	63.26 ± 0.83%	25.67 ± 2.63%	21.47 ± 1.76%
Chai's work [8](softmax label)	63.47 ± 0.44%	21.32 ± 1.97%	19.52 ± 2.46%
Chai's work [8](linear label)	63.34 ± 0.46%	20.31 ± 2.62%	20.27 ± 2.34%
Reckoner	64.92 ± 0.63%	17.47 ± 0.87%	20.72 ± 0.97%
Reckoner (w/o noise)	64.95 ± 0.51%	17.91 ± 1.32%	21.21 ± 1.33%
Reckoner (w/o *pseudo-learning*)	64.38 ± 0.83%	17.98 ± 1.34%	21.18 ± 1.46%

Table 2: Results on the New Adult dataset.

Metrics(%) Methods	Accuracy	Equalised Odds	Demographic Parity
DRO [24]	85.15 ± 0.93%	11.56 ± 2.10%	12.23 ± 1.41%
ARL [29]	85.37 ± 1.91%	11.79 ± 1.77%	13.05 ± 1.57%
FairRF [41]	83.74 ± 0.86%	11.23 ± 1.42%	11.37 ± 1.46%
Chai's work [8](softmax label)	84.63 ± 0.47%	10.34 ± 1.22%	10.63 ± 1.34%
Chai's work [8](linear label)	84.27 ± 0.31%	10.57 ± 1.64%	10.21 ± 1.52%
Reckoner	85.35 ± 0.07%	5.83 ± 0.51%	9.78 ± 0.17%
Reckoner (w/o noise)	85.36 ± 0.09%	5.82 ± 0.27%	9.98 ± 0.19%
Reckoner (w/o *pseudo-learning*)	85.53 ± 0.13%	4.82 ± 0.41%	9.11 ± 0.18%

Table 3: Results on the CelebA dataset.

Metrics(%) Methods	Accuracy	Equalised Odds	Demographic Parity
DRO [24]	77.12 ± 0.58%	17.22 ± 1.69%	19.04 ± 1.51%
ARL [29]	78.91 ± 0.41%	17.53 ± 1.72%	19.46 ± 1.96%
Chai's work [8](softmax label)	80.87 ± 0.14%	11.43 ± 1.25%	15.27 ± 1.71%
Chai's work [8](linear label)	80.76 ± 0.73%	10.62 ± 1.10%	14.47 ± 1.64%
Reckoner	79.47 ± 0.25%	11.58 ± 0.49%	13.92 ± 1.27%
Reckoner (w/o noise)	79.86 ± 0.09%	11.96 ± 0.67%	14.23 ± 0.85%
Reckoner (w/o *pseudo-learning*)	77.99 ± 0.11%	11.01 ± 0.74%	13.19 ± 0.94%

some datasets may contain biased labels, and focusing on increasing classification performance for the minority may lead to an increase in the bias gap, which is what we aim to avoid. Therefore, our study underscores the significance of group fairness for a more comprehensive improvement, steering away from inadvertently perpetuating bias in labelled datasets.

5.2 Ablation Study

In our ablation study, we look at the effectiveness of the two components in the proposed framework, Reckoner, and the necessity of combining them. Tables 1 - 3 show prediction accuracy and fairness measurements for all variants. In general, Reckoner and its variants achieve superior performance on the COMPAS and NewAdult datasets. In the CelebA dataset, Reckoner and its variants exhibit better performance only in Demographic Parity compared all the baselines. Based on these results, we are interested in understanding the differences in performance among each variant across different datasets.

Effect of the Learnable Noise. Our model without learnable noise trains both classifiers using original inputs in the refinement stage. It achieves a slight advantage in accuracy on COMPAS and CelebA datasets, with a minor gap in fairness results. However, in the New Adult dataset there is no particularly noticeable difference in performance across the three metrics compared with the full Reckoner. It is worth noting that we argue that COMPAS and CelebA datasets contain biased labels, as their ground truth is assigned by officials of the legal department and human annotators. Due to the absence of learnable noise, the classifier in the *pseudo-learning* phase tends to push the decision boundary significantly into the feature space of the majority, enhancing discrimination against the

minority and leading to an increase in accuracy. However, when the dataset labels are considered biased, this can also result in a decrease in fairness values. This is why in the New Adult dataset, whether the proposed framework introduces learnable noise or not, the performance difference is not significant. However, in datasets suspected of having biased labels there is a larger fluctuation in fairness among these two variants of the proposed approach.

Effect of the *pseudo-learning*. In addition to the learnable noise, we also evaluate the effectiveness of *pseudo-learning*. Our model without *pseudo-learning* employs a three-layer MLP as the main classifier, takes noise-augmented information as input, and does not involve the identification stage. Similar to our previous speculation, its poor performance on datasets with biased labels is also related to the misguidance caused by supervised learning with problematic ground truth. Without the regularisation provided by *pseudo-learning*, the classifier mistakenly treats bias as knowledge, and learnable noise reinforces this by selecting features that maximise the cross-entropy with the ground truth. Additionally, experimental results seem to suggest overfitting issues with this variant on these datasets. As compared to the results on New Adult, where our method outperforms five baselines and other variants, both in terms of predictive accuracy and fairness, these results suggest that the bias in labels will greatly impact algorithmic fairness. Perhaps in a dataset with unbiased labels, improving predictive accuracy could effectively reduce the bias gap between different demographic groups.

6 RELATED WORK

Group Fairness. In contrast to approaches that emphasise the equitable treatment of similar individuals in pursuit of individual fairness, our work focuses on group fairness, manifesting in the differential treatment of distinct demographic groups. Some foundational methods to group fairness include incorporating fairness regularisation into the objective function or converting it into a constrained optimisation problem. Kamishima et al. [28] introduced a method for reducing mutual information between sensitive groups and targets by quantifying the mutual distribution between them. This approach aims to diminish the dependency between sensitive groups and targets. A similar concept is also adopted by Beutel et al. [4], where fairness is achieved by minimising the absolute correlation between these two entities. In contrast to the aforementioned methods, Hardt et al. [23] propose the use of the Equalised Odds fairness metric, which underscores the equalisation of true positive and false positive rates across different demographic groups. It transforms the general loss function into an optimisation problem subject to fairness constraints, ensuring that the revised unbiased predictions closely approximate the original predictions. Similarly, Zafar et al. [40] achieve fair classification by adding tractable constraints at the decision boundary. However, as the desire for both algorithmic fairness and privacy grows, we observe the requirement of avoiding the use of sensitive attributes in machine learning model training, leading to legislative restrictions on such practices like, e.g., the General Data Protection Regulation (GDPR) [37]. To manage such requirements, some approaches have been designed under the assumption that sensitive attributes are either difficult to obtain or prohibited from use, like we do in our work.

Fairness Without Sensitive Attributes. As public concerns about privacy are on the rise, an increasing amount of research on group fairness is turning its attention to "imperfect" data, such as missing protected class labels [2, 10, 32] or noisy sensitive attributes [19, 34]. To deal with fairness in this setting, the main idea of some studies is leveraging the correlation between sensitive and non-sensitive attributes to mitigate bias. Representative work includes the use of proxy features [22], in which a proxy group is obtained from clustering the data and is used to replace actual sensitive attributes during training. A well-known example is using 'zip code' instead of 'race' as this can have similar effects on individual splits since the two attributes are highly correlated [15]. Similarly, Yan et al. [39], Zhao et al. [41] explore features which have strong correlation with sensitive attributes to learn fair classifiers by using them for training and for regularisation in learning. However, this approach needs a careful selection of proxy attributes and even of fairness metrics. To address underlying issues, Zhu et al. [42] estimate fairness using only weak proxies. Through estimating the transition probabilities between sensitive group target values, it uses auxiliary models to calibrate the fairness metrics. Another family of approaches [24] addresses fairness without sensitive attributes via distributionally robust optimisation (DRO). The main idea is that the fairness of the algorithm is related to the quantity of individuals in different demographic groups. If empirical risk minimisation is employed to optimise the algorithm, it may lead to a prioritisation of benefits for the privileged group, as the privileged group constitutes the majority. This could result in the non-privileged group gradually avoiding the algorithm due to a poor experience, creating a vicious cycle that ultimately causes the algorithm to increasingly overlook the non-privileged group. This approach utilises different group proportions to design upper and lower bounds for the objective function of each group, ensuring that no group is discriminated against by the algorithm. Recently, Jung et al. [27], targeting group fairness, extended DRO with fairness constraints in the resulting objective function using a re-weighting based learning method. Beside the aforementioned methods, others have recently utilised various techniques to address unfairness without knowledge of demographics. For example, Lahoti et al. [29] adversarially reweigh the samples to achieve a Rawlsian Max-Min fairness and learn the classifier. However, these methods can be easily influenced by outliers. Others tackle the problem through knowledge distillation [8], reweighing-based contrastive learning [9], and causal variational autoencoders [20]. However, these methods need the prior identification of proxies to harness their interactions with sensitive attributes, such as correlation and causality, in order to achieve fairness. Our approach avoids the need for such analysis. Instead, it leverages learnable noise applied to all data, forcing the data to retain only essential information for better predictions. Additionally, it employs a dual-model knowledge-sharing mechanism to acquire fairness-related knowledge, thereby improving predictive fairness. Hence, our proposed framework exhibits greater generalisability, particularly when dealing with data where proxy identification is challenging, such as images and audio.

7 CONCLUSIONS

In this paper, we present a novel framework for classification tasks that improves fairness without using sensitive attributes. Through an analysis of the bias gap and the distribution of selected non-sensitive attributes across different confidence subsets with respect to different demographic groups, we gain insights into how biased labels and other hidden bias harm fairness in predictions and mislead the classifier. Our proposed framework integrates (1) learnable noise and (2) a dual-model system, enabling a knowledge-sharing framework for fair predictions for different demographic groups. Our experimental results show the superiority of the proposed method, which can make accurate and fair predictions, as compared to state of the art baseline methods. Our ablation study also confirms the benefits of the two main components in our proposed solution. The code of our method is publicly available at https://github.com/uewopq88/Reckoner-Fairness. It is important to acknowledge that the critical issue of the intersectionality of multiple sensitive attributes in automated decision-making systems warrants further investigation in our future research. While our proposed framework is able to perform classification tasks even with multiple sensitive attributes, this paper does not discuss intersectional fairness due to challenges such as data scarcity at intersections of minority groups and a lack of proper group fairness metrics. Future studies will focus on enhancing intersectional fairness in classification and extending current group fairness metrics.

ACKNOWLEDGMENTS

This work is partially supported by the Australian Research Council (ARC) Training Centre for Information Resilience (Grant No. IC200100022).

REFERENCES

[1] McKane Andrus and Sarah Villeneuve. 2022. Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1709–1721.

[2] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. 2021. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 206–214.

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

[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[5] Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. 2021. Beyond low-frequency information in graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3950–3957.

[6] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 492–500.

[7] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior* 36, 1 (2009), 21–40. https://doi.org/10.1177/0093854808326545

[8] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. 2022. Fairness without Demographics through Knowledge Distillation. In *Advances in Neural Information Processing Systems*, Vol. 35. 19152–19164. https://proceedings.neurips.cc/paper_files/paper/2022/file/79dc391a2c1067e9ac2b764e31a60377-Paper-Conference.pdf

[9] Junyi Chai and Xiaoqian Wang. 2022. Self-supervised fair representation learning without demographics. *Advances in Neural Information Processing Systems* 35, 27100–27113.

[10] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348.

[11] Consumer Financial Protection Bureau 2022. *CFPB Targets Unfair Discrimination in Consumer Finance*. Consumer Financial Protection Bureau, Washington, DC, USA. https://www.consumerfinance.gov/about-us/newsroom/cfpb-targets-unfair-discrimination-in-consumer-finance/

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[13] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems* 32 (2019).

[14] Peng Cui, Yang Yue, Zhijie Deng, and Jun Zhu. 2022. Confidence-based Reliable Learning under Dual Noises. *Advances in Neural Information Processing Systems* 35 (2022), 35116–35129.

[15] 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).

[16] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2022. Multiaccurate proxies for downstream fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1207–1239.

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

[18] Laurel Eckhouse. 2017. *Big data may be reinforcing racial bias in the criminal justice system*. Report. The Washington Post.

[19] Azin Ghazimatin, Matthaus Kleindessner, Chris Russell, Ziawasch Abedjan, and Jacek Golebiowski. 2022. Measuring fairness of rankings under noisy sensitive information. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2263–2279.

[20] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2022. Fairness without the Sensitive Attribute via Causal Variational Autoencoder. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligenceg*. 696–702.

[21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[22] Maya Gupta, Andrew Cotter, Mahdi M Fard, and Serena Wang. 2018. Proxy Fairness. *arXiv preprint arXiv:1806.11212* (2018).

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

[24] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning*. 1929–1938.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[26] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).

[27] Sangwon Jung, Taeeon Park, Sanghyuk Chun, and Taesup Moon. 2023. Reweighting Based Group Fairness Regularization via Classwise Robust Optimization. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=Q-WfHzmiG9m

[28] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[29] 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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 728–740.

[30] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. AAAI Press, 2124–2132.

[31] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. Compas analysis. https://github.com/propublica/compas-analysis

[32] David Liu, Virginie Do, Nicolas Usunier, and Maximilian Nickel. 2023. Group fairness without demographics using social networks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1432–1449.

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

[34] Anay Mehrotra and L Elisa Celis. 2021. Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM conference on fairness,*

accountability, and transparency. 237–248.
[35] John Rawls. 2001. *Justice as fairness: A restatement*. Harvard University Press.
[36] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 30 (2017).
[37] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.
[38] Jiaheng Wei, Harikrishna Narasimhan, Ehsan Amid, Wen-Sheng Chu, Yang Liu, and Abhishek Kumar. 2023. Distributionally robust post-hoc classifiers under prior shifts. *arXiv preprint arXiv:2309.08825* (2023).
[39] Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. 1715–1724.
[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.
[41] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2022. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1433–1442.
[42] Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. 2023. Weak Proxies are Sufficient and Preferable for Fairness with Missing Sensitive Attributes. *arXiv preprint arXiv:2210.03175* (2023).