A Comparative Study of Improved Disparate Impact Remover and Fair Adversarial Learning for Bias Mitigation

Junyu Zhou
University of Bristol
vb24220@bristol.ac.uk

Shuojingrui He
University of Bristol
eo24594@bristol.ac.uk

Jingyu Hu
University of Bristol
ym21669@bristol.ac.uk

Weiru Liu
University of Bristol
weiru.liu@bristol.ac.uk

***Abstract*—Machine learning models in high-stake domains can inherit and amplify historical biases from training data, deepening social inequality. We compare the Improved Disparate Impact Remover (IDIR) with adversarial learning (AL) for comprehensive bias mitigations. IDIR enhances bias detection using a dual-metric combination of KL divergence and Total Variation distance. KL divergence captures local distributional anomalies while TV distance measures macroscopic probability deviations, enabling robust identification of indirect bias across structural differences and population shifts. We apply controlled data repair with a tunable strength coefficient that selectively adjusts non-sensitive group distributions toward sensitive groups through targeted augmentation, preserving model discriminability while achieving effective mitigation. Adversarial learning addresses biases that emerge during training through an encoder-predictor-adversary architecture, which uses gradient reversal layers to learn representations that minimize the correlation between predictions and sensitive attributes. IDIR is a preprocessing technique that reduces historical biases before training, whilst in-processing AL maintains fairness constraints during optimization. Experiments on three datasets demonstrate consistent reductions in group-level disparities by both methods while maintaining predictive accuracy.**

***Index Terms*—Fairness, Bias Mitigation, Data Augmentation, Adversarial Learning**

I. Introduction

With the rapid development of machine learning technology in various fields, ML algorithm have played a key role in decision-making in medical care [1], [2], education [3], finance [4], [5] and other fields [6], [7]. However, there is a general bias problem in the existing training data, which may lead to the trained machine learning model aggravating existing social discrimination, resulting in a more unbalanced social structure [8], [9]. Existing bias mitigation methods can be categorized into three temporal stages: pre-processing methods that modify training data, in-processing techniques that incorporate fairness constraints during model training, and post-processing approaches that adjust model outputs. Compared with post-processing, pre-processing and in-processing methods address bias closer to its origins, offering more principled solutions with better theoretical guarantees.

Among pre-processing methods, the Disparate Impact Remover [10] addresses differential treatment by modifying feature distributions. However, existing methods face challenges in accurately detecting which features contain indirect bias and determining optimal repair strategies. Moreover, standard DIR does not explicitly specify how to identify indirectly biased features in a robust way, and overly aggressive repair can hurt utility; therefore, a principled detection rule and a tunable repair strength are required. We propose an Improved Disparate Impact Remover (IDIR) that enhances bias detection through a novel combination of Kullback-Leibler (KL) divergence and Total Variation (TV) distance metrics. While KL divergence captures sensitivity to local distributional anomalies and small probability events, TV distance measures macroscopic probability deviations between groups. This dual-metric approach enables more comprehensive detection of both structural differences and population shifts between sensitive and non-sensitive groups, significantly improving detection accuracy while maintaining robustness. Our data repair strategy introduces a controlled augmentation approach [11] that selectively adjusts non-sensitive group distributions toward sensitive group distributions through targeted sample generation. Rather than complete distributional alignment which could harm model discriminability, we apply a repair strength coefficient that controls the degree of distributional fusion, preserving model performance while achieving effective bias mitigation.

Considering that pre-processing alone may not capture all sources of bias that emerge during model training, we further consider a bias mitigation method based on adversarial learning (AL) during the in-process training stage. The adversarial learning model includes an Encoder–Predictor–Adversary framework [12] [13], and it predicts the target variable while minimizing the correlation between the target variable and sensitive attributes. AL can force the Encoder to learn from features that do not contain information about sensitive attributes, thereby alleviating the bias introduced into the model by the original data. By comparing bias mitigation at both the data preparation and model training stages, we provide a complementary analysis of both methods and find that pre-processing ensures that historical biases in the training data are reduced before model training begins, while in-processing maintains fairness constraints throughout the learning process.

The mitigation methods are evaluated across three datasets and three different models, and we assess the fairness improvement with five metrics. The experiments show that many

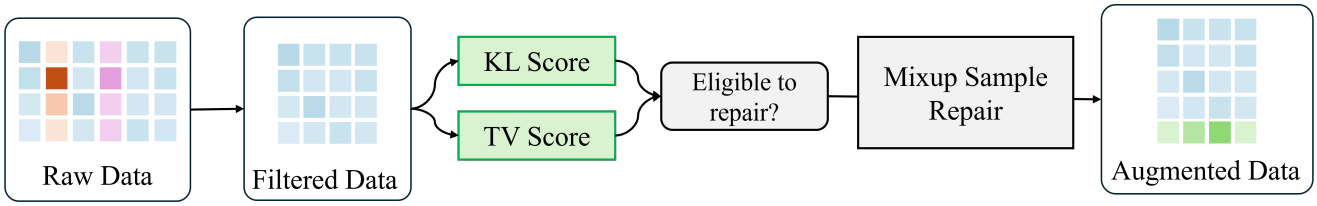


Fig. 1: The Pipeline of Improved Disparate Impact Remover

tabular datasets exhibit inherent bias, and both methods consistently reduce group-level disparities on biased datasets while preserving overall predictive performance. Across settings, we observe lower gaps in demographic parity and equalized odds, improved equal opportunity, better calibration across groups, and higher disparate-impact ratios, with negligible sacrifices in prediction accuracy. While extreme sparsity in minority groups can still limit improvements, the results indicate that IDIR and AL provide the effective and broadly applicable mitigation strategy for biased tabular data.

II. Related Work

Fairness improvement methods in machine learning have evolved from foundational bias frameworks into sophisticated intervention methods at different model learning stages. [14] established systematic frameworks for understanding bias in computer systems, empirical studies like Gender Shades project [15] demonstrated severe accuracy disparities across demographic groups in commercial systems. These findings have led to advocacy organizations like the Algorithmic Justice League[1] and Data for Black Lives[2] promoting algorithmic fairness in real-world applications. Methodologically, bias mitigation methods can be broadly categorized into three families: pre-processing, in-processing and post-processing stage approaches. Pre-processing methods aim to make the training data fairer prior to the training [16], [17]. The common strategies include re-weighting, resampling, fair learning and others. [14] introduced the Disparate Impact Remover, a geometric repair approach that edits feature values to reduce group disparities while preserving rank-ordering within groups. In-processing approaches integrate fairness constraints directly into model optimization, offering more targeted control over the fairness-accuracy trade-off. [18] introduced joint predictor–adversary training to maximize task accuracy while minimizing predictability of protected attributes; [13] generalized this to transferable fair representations; [19] used minimax training to make cross-group representations statistically indistinguishable while preserving utility.

III. Methods

This section presents the pre-processing method Improved Disparate Impact Remover, an in-processing method based on adversarial learning.

[1]https://www.ajl.org/
[2]https://d4bl.org/

Notations Given a dataset D for a classification task. Each data $d \in D$ is represented as $d = (X, Y, S)$, where X denotes the input features, Y represents the task label, and S denotes a set of sensitive features such as gender or race. For each sensitive feature f, we define $S(f_n) = 0$ as the n-th non-sensitive (majority) group and $S(f) = 1$ as the sensitive (minority) group. $\hat{Y}$ is the model's predicted label.

A. Improved Disparate Impact Remover (IDIR)

The proposed Improved Disparate Impact Remover(IDIR) is a data augmentation technique that combined Kullback-Leibler (KL) divergence and Total Variation (TV) distance to validate sensitive groups and also to mitigate bias through identifying indirect bias from non-sensitive features.

KL divergence measures the difference between two probability distributions, and for categorical distributions P and Q, it is formulated as $D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$, where P(i) and Q(i) represent the probabilities of the i-th class in the non-sensitive and sensitive groups, respectively. A larger KL divergence indicates greater distributional differences between the two groups, suggesting the feature requires repair. Total Variation distance captures the overall probability deviation between two distributions $D_{\mathrm{TV}}(P, Q) = \frac{1}{2} \sum_i |P(i) - Q(i)|$. KL divergence is sensitive to small probability events but cannot reflect population shifts, TV captures macroscopic differences between distributions. The combination of both metrics enables detection of both local structural anomalies and overall shifts, significantly improving mitigation solution while maintaining robustness against minor disturbances. In order to validate the sensitive group out of multiple groups of a feature, we do pair-wise KL calculations. The group that has the cumulated largest value is regarded as the sensitive group.

After identifying features with indirect bias using KL and D_{TV}, our data repair strategy is to adjust the distribution of non-sensitive groups to approximate that of sensitive groups through targeted sample augmentation. For example, if the probability distribution over Education Levels for the sensitive group (women) differs significantly from that of the non-sensitive group (men) by the values of D_{KL} and D_{TV}, a sample argumentation repair strategy will be applied to turn the probability distribution of men's group towards that of women's group. This approach selectively adds samples to the non-sensitive group to increase the proportion of high-quality instances while ensuring minimal sample modification—only one feature value is changed per augmented sample to preserve the target variable Y.

To maintain model discriminability and prevent accuracy degradation, complete distributional alignment is avoided. Inspired by the Mixup idea [20], [21], a repair strength coefficient α is introduced to control the fusion degree.

Given non-sensitive population distribution P and sensitive population distribution Q, the new distribution of the non-sensitive group for a specific categorical value is calculated as $P'(j) = \alpha \cdot P(j) + (1-\alpha) \cdot Q(j)$. Here j represents j-th non-sensitive group within a specific categorical feature, where $j \in \{1, \ldots, C-1\}$ and C is the total number of feature values. The repair process iteratively samples from the original non-sensitive population without replacement until the dataset distribution reaches the specified target degree, effectively completing the bias mitigation while preserving model performance.

B. Adversarial Learning (AL)

Adversarial learning (AL) [18] aims to learn a representation $Z = f_\theta(X)$ that maximally preserves task-relevant information while minimizing sensitive attribute information. This approach addresses systematic bias by decoupling the learned representation from protected attributes at the source level. The optimization objective can be formulated as: $\max_{\theta,\phi} \; I(Z;Y)$ subject to $\min_\theta \; I(Z;S)$, where $I(\cdot;\cdot)$ denotes mutual information between two variables. The adversarial learning framework consists of three interconnected modules that work in concert to achieve fair representation learning:

Encoder (f_θ) with parameter θ that maps X to a compressed representation Z, acting as an information bottleneck that selectively retains task-relevant information while filtering out sensitive attribute information.

Predictor (g_ϕ) with parameter ϕ that takes representation Z as input to generate task predictions $\hat{y} = g_\phi(Z)$, optimizing for high accuracy on the main task through binary cross-entropy loss minimization.

Adversary (h_ψ) with parameter ψ that attempts to predict sensitive attributes $\hat{s} = h_\psi(Z)$ from the learned representation, serving as a detector for sensitive information leakage. High adversary accuracy indicates insufficient protection of sensitive information, while low accuracy suggests successful bias mitigation.

The AL training process operates through two competing pathways. In the predictor pathway, the representation Z flows to the predictor to minimize task prediction loss: $\mathcal{L}_{pred} = -\mathbb{E}_{(X,Y)}\Big[Y \log(\hat{y}) + (1-Y)\log(1-\hat{y})\Big]$. Simultaneously, the adversary pathway processes Z through a gradient reversal layer (GRL) before attempting to predict sensitive attributes, with loss function:

$$\mathcal{L}_{adv} = -\mathbb{E}_{(X,S)}\Big[S \log(\hat{s}) + (1-S)\log(1-\hat{s})\Big]$$

The GRL reverses gradient signs during backpropagation, forcing the encoder to learn representations that confound the adversary's ability to predict sensitive attributes.

The complete objective function integrates both pathways through min-max optimization:

$$\min_{\theta,\phi} \; \mathcal{L}_{pred}\big(g_\phi(f_\theta(X)), Y\big) - \lambda \max_{\psi} \; \mathcal{L}_{adv}\big(h_\psi(f_\theta(X)), S\big)$$

where λ balances prediction performance against fairness constraints. To address potential training instability from simultaneous optimization, we employ an alternating strategy: the adversary is updated k times with fixed encoder and predictor to maximize $\mathcal{L}_{adv}$, followed by a single update of the encoder and predictor with fixed adversary to minimize $\mathcal{L}_{pred} - \lambda\mathcal{L}_{adv}$. This alternating approach ensures robust convergence while maintaining both task performance and fairness objectives.

Algorithm 1 Pseudocode of Fair Adversarial Learning

Require: Labeled dataset $\mathcal{D} = \{(X_i, Y_i, S_i)\}_{i=1}^N$; encoder f_θ; predictor g_ϕ; adversary h_ψ; trade-off $\lambda > 0$; adversary steps k; learning rates $\eta_\theta, \eta_\phi, \eta_\psi$; epochs T; batch size B; binary cross-entropy BCE

Initialize parameters θ, ϕ, ψ.
function $\mathrm{GRL}(z, \lambda)$ ▷ Gradient Reversal Layer
  forward: return z
  backward: multiply the gradient w.r.t. z by $-\lambda$
end function
for epoch $= 1$ **to** T **do**
  for each mini-batch $\mathcal{B} = \{(X, Y, S)\} \subset \mathcal{D}$ of size B **do**
    Update adversary k **times with encoder/predictor frozen**
    for $t = 1$ **to** k **do**
      $Z \leftarrow f_\theta(X)$, $\hat{s} \leftarrow h_\psi(\mathrm{GRL}(Z, \lambda))$
      $\mathcal{L}_{adv} \leftarrow \frac{1}{B}\sum_i \mathrm{BCE}(\hat{s}_i, S_i)$
      gradient ascent on ψ: $\psi \leftarrow \psi + \eta_\psi \nabla_\psi \mathcal{L}_{adv}$
    end for
    Update encoder/predictor once with adversary frozen
    $Z \leftarrow f_\theta(X)$, $\hat{y} \leftarrow g_\phi(Z)$, $\hat{s} \leftarrow h_\psi(\mathrm{GRL}(Z, \lambda))$
    $\mathcal{L}_{pred} \leftarrow \frac{1}{B}\sum_i \mathrm{BCE}(\hat{y}_i, Y_i)$
    $\mathcal{L}_{adv} \leftarrow \frac{1}{B}\sum_i \mathrm{BCE}(\hat{s}_i, S_i)$
    combined objective: $\mathcal{L} \leftarrow \mathcal{L}_{pred} - \lambda\,\mathcal{L}_{adv}$
    gradient descent on θ, ϕ: $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}$, $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathcal{L}$
  end for
end for
Output: trained parameters $(\theta^\star, \phi^\star, \psi^\star)$; representation $Z = f_{\theta^\star}(X)$

C. The Comparison of IDIR and AL

To ensure minimal bias at the initial and intermediate stages while maximizing mitigation in the final results, each bias mitigation method is applied at its corresponding stage. IDIR adopts $\tau = 0.1$ as the standard threshold for both KL divergence and TV distance. A feature is repaired only if both $D_{KL} > \tau$ and $D_{TV} > \tau$, indicating perceptible distributional differences between sensitive and non-sensitive groups.

We set BCEWithLogitsLoss in the adversarial training process as the loss function. BCEWithLogitsLoss integrates the sigmoid activation and binary cross-entropy calculation into a single, numerically stable operation that directly processes model logits, avoiding the logarithm of zero issues that plague standard BCE implementations. This modification eliminates the need for manual sigmoid activation and clamping operations, resulting in more robust adversarial training.

IV. Experiments and Results

A. Experiment Settings

Datasets and Models. The experiments are under three selected datasets: Adult Income [22], Bank Marketing [23] and COMPAS [24]. We consider bias mitigation methods of IDIR and AL. All bias mitigation methods are applied on three machine learning models: logistic regression, random forest and SVM.

TABLE I: Statistics of Three Datasets

	Adult Income	Bank Marketing	COMPAS
#Features $\lvert X\rvert$	14	20	12
#Target Y	Income $>$ 50K	Subscribe a term deposit	Recidivism within 2 years
#Sens. S	Gender, Race	Age	Gender, Race

Evaluation Metrics. We evaluate predictive and fairness performance [25]–[27] using Accuracy, Demographic Parity difference (ΔDP), Average Odds difference (ΔAO), Equal Opportunity difference (ΔTPR), FPR difference (ΔFPR) and Max Equalized Odds difference (ΔEOmax). Higher accuracy indicates better predictive performance, while lower fairness metrics indicate fairer performance. Here we denote $SR_s = P(\hat{Y} = 1|S = s)$ as the selection rate, and TPR_s, FPR_s are the True Positive and False Positive Rates for group s, where $s \in \{0, 1\}$. Specifically, $\text{TPR}_{S=s} = P(\hat{Y} = 1 \mid Y = 1, S = s)$, $\text{FPR}_{S=s} = P(\hat{Y} = 1 \mid Y = 0, S = s)$. The calculation equations of these evaluation metrics are defined as below.

$$\text{Accuracy} = P(\hat{Y} = Y), \Delta\text{DP} = \text{SR}_{S=0} - \text{SR}_{S=1},$$
$$\Delta\text{FPR} = \text{FPR}_{S=0} - \text{FPR}_{S=1}, \Delta\text{TPR} = \text{TPR}_{S=0} - \text{TPR}_{S=1},$$
$$\Delta\text{EO} = \max\Big(\big|\text{TPR}_{S=0} - \text{TPR}_{S=1}\big|,\ \big|\text{FPR}_{S=0} - \text{FPR}_{S=1}\big|\Big).$$

B. Datasets Explorations

As Figure 2 shows, we first examine the biases present in the datasets. The Adult Income Dataset reveals that the male group has a higher probability of earning an annual income exceeding 50K compared to the female group. In the Bank Marketing Dataset, older demographics (over 60 years old) exhibit higher subscription response rates. The COMPAS Dataset shows that African Americans and Native Americans receive higher decile scores than other ethnic groups such as Caucasians. The inherent biases present in these datasets may cause models to perpetuate discriminatory patterns, such as predicting with higher probability that individuals of certain genders and ethnicities will have lower incomes and higher crime rates.

C. Bias Mitigation with IDIR and Adversarial Learning

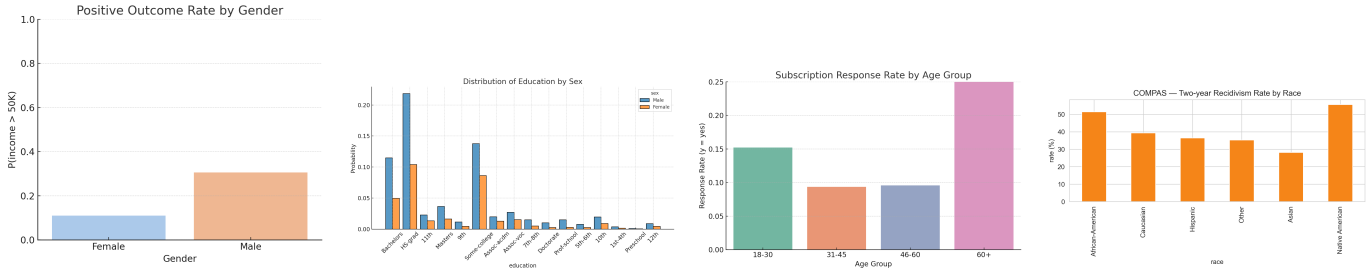
Figure 3 presents the age-group distribution within each job category in the repaired dataset. It can be seen that the repaired data has a more uniform distribution on each occupation category, the number of all age groups is basically the same, and the correlation between job category and sensitive attribute *age* is low, which alleviates the indirect bias. Now we apply the proposed methods to all datasets and Table II reports the fairness metrics before and after Improved Disparate Impact Remover and Adversarial Learning mitigation.

In Adult dataset, the experiments demonstrate improvements in both accuracy and fairness metrics as repair intensity increases. In the Bank dataset, models applying IDIR maintain stable predictive accuracy (> 0.91). In contrast, for the COMPAS dataset, the accuracy remains consistent while achieving meaningful fairness improvements. For gender bias mitigation, adversarial learning achieved substantial improvements across both datasets. In the COMPAS dataset, logistic regression demonstrated the largest improvement, with ΔDP decreasing from 0.3102 to 0.2674 and ΔTPR from 0.3412 to 0.2695. SVM achieved optimal performance with all fairness metrics approaching zero. Table 3 reports AL's performance in mitigating racial disparities for the COMPAS and Adult datasets, and results demonstrate that mitigation generalizes across sensitive features. Nevertheless, racial bias mitigation presented greater challenges. In the COMPAS dataset, performed best with ΔTPR decreasing from 0.8099 to 0.3798 and $\Delta EOdds$ diff from 0.7632 to 0.3798, while logistic regression maintained elevated bias levels with ΔDP diff remaining at 0.4253. For the Adult dataset, logistic regression and SVM achieved significant improvements, whereas Random Forest exhibited bias deterioration with ΔTPR increasing from 0.1011 to 0.1509. SVM consistently outperformed other models across both datasets and sensitive attributes. Compared with the COMPAS dataset, whose inherent historical bias remains resistant to adversarial methods, the Adult dataset's more uniform distribution enabled more effective bias mitigation.

These findings indicate that IDIR is superior for smaller datasets or those with high-cardinality categorical variables, as targeted augmentation provides interpretable bias reduction without requiring massive training overhead. With small sample sizes, a limited number of IDIR-generated samples can effectively alleviate category-specific bias, while AL suffers from overfitting due to insufficient samples for stable convergence of both the adversary and predictor, making the debiasing process inefficient. However, as sample size increases, IDIR-generated samples grow exponentially and may exceed the original dataset volume, whereas AL becomes more effective with abundant training data, achieving better debiasing through stable optimization of competing pathways.

V. Discussion

The experimental results show that bias mitigation effectiveness varies across datasets, models, and sensitive attributes. The Adult Income dataset, with balanced feature distributions and moderate historical bias, benefits consistently from both



(a) Positive Outcome Rate by Gender

(b) Distribution of Education by Gender

(c) Subscription Response Rate by Age Group

(d) Distribution of Two-year Recidivism Rate by Race

Fig. 2: The imbalanced distributions in datasets.

TABLE II: Models Performance under Improved Disparate Impact Remover and Adversarial Learning. Gender is used as the sensitive feature in COMPAS and Adult datasets, and Age is set as the sensitive feature in Bank dataset.

		COMPAS Dataset					Adult Income Datset					Bank Marketing Dataset				
	Models	Acc.	ΔDP	ΔTPR	ΔFPR	$\Delta EO_{\max}$	Acc.	ΔDP	ΔTPR	ΔFPR	$\Delta EO_{\max}$	Acc.	ΔDP	ΔTPR	ΔFPR	$\Delta EO_{\max}$
$\alpha = 1$	Logistic Regression	0.6951	0.3015	0.2928	0.2194	0.2928	0.8258	0.1755	0.2856	0.0727	0.2856	0.9104	0.1293	0.0523	0.0784	0.0784
	Random Forest	0.6748	0.2417	0.195	0.1989	0.1989	0.8584	0.1811	0.1100	0.0691	0.1100	0.9183	0.1043	0.0043	0.0381	0.0381
	SVM	0.6894	0.3485	0.3515	0.2619	0.3515	0.8145	0.0805	0.1365	0.0165	0.1365	0.9112	0.0737	0.0539	0.0429	0.0539
$\alpha = 0.5$	Logistic Regression	0.6842	0.2515	0.2181	0.1965	0.2181	0.8336	0.1726	0.2473	0.0766	0.2473	0.9105	0.1225	0.1082	0.0626	0.1082
	Random Forest	0.7149	0.2000	0.2304	0.077	0.2304	0.8754	0.1797	0.0268	0.0793	0.0793	0.9209	0.1284	0.2600	0.0082	0.2600
	SVM	0.6897	0.2765	0.2572	0.2087	0.2572	0.8212	0.058	0.0578	0.0082	0.0578	0.9097	0.0988	0.1263	0.0313	0.1263
$\alpha = 0$	Logistic Regression	0.6834	0.2378	0.2336	0.1559	0.2336	0.8492	0.1694	0.2237	0.0768	0.2237	0.9072	0.1059	0.1024	0.058	0.1024
	Random Forest	0.7359	0.1851	0.2546	0.0192	0.2546	0.8960	0.1712	0.0662	0.0774	0.0774	0.9172	0.1045	0.2965	0.0067	0.2965
	SVM	0.6923	0.2405	0.2223	0.165	0.2223	0.8387	0.0511	0.0285	0.0062	0.0285	0.9081	0.0812	0.1138	0.0263	0.1138
AL	Logistic Regression	0.6765	0.2674	0.2695	0.2077	0.2695	0.8695	0.1315	0.0462	0.0421	0.0496	0.9092	0.0824	0.0104	0.0436	0.0436
	Random Forest	0.6514	0.1365	0.0972	0.1165	0.1165	0.8703	0.1276	0.0346	0.0537	0.0357	0.9074	0.0728	0.0363	0.0347	0.0363
	SVM	0.6687	0.1612	0.1569	0.1145	0.1569	0.8864	0.1367	0.0157	0.0369	0.0219	0.9075	0.1015	0.0621	0.0498	0.0621

TABLE III: Model Performance under Adversarial Learning with Race as the Sensitive Feature

		COMPAS Dataset				Adult Income Dataset			
		ΔDP	ΔTPR	ΔFPR	ΔEOdds(max)	ΔDP	ΔTPR	ΔFPR	ΔEOdds(max)
Before	Logistic Regression	0.5302	0.7500	0.3453	0.7500	0.2557	0.2936	0.1510	0.2936
	Random Forest	0.3214	0.6628	0.3867	0.6628	0.1850	0.1011	0.0689	0.1011
	SVM	0.6308	0.8099	0.4478	0.7632	0.1755	0.1789	0.0802	0.1789
After	Logistic Regression	0.4253	0.5862	0.3008	0.6003	0.2007	0.2308	0.1057	0.2246
	Random Forest	0.2924	0.4648	0.3509	0.4124	0.1389	0.1209	0.0605	0.1509
	SVM	0.3609	0.3798	0.2918	0.3798	0.1107	0.1004	0.0578	0.1053

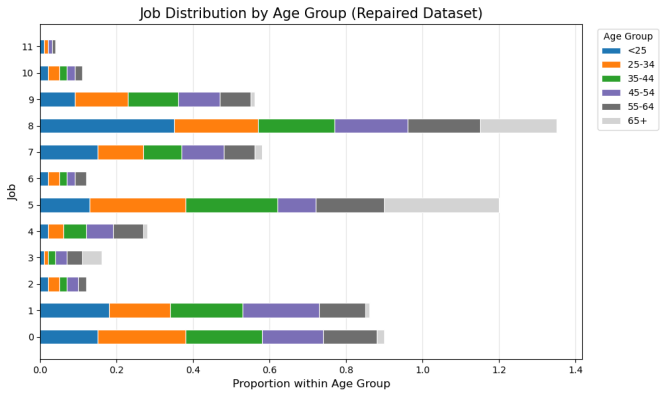


Fig. 3: Job Distribution by Age Group after IDIR

IDIR and adversarial learning. The COMPAS dataset, with stronger historical bias and sparser minority groups, poses greater challenges, and residual disparities remain after intervention. IDIR reduces historical and structural distributional bias by repairing the indirectly biased features prior to training, while adversarial learning suppresses emergent bias introduced during model optimization. When used in isolation, IDIR cannot address bias arising in the training stage, and Adversarial Learning can suffer from instability when trained on heavily biased data. Future work could sequentially combine these two stages to allow adversarial learning to operate on a less biased data distribution, potentially leading to more stable convergence and improved fairness–accuracy trade-offs across datasets and models.

Method selection depends on dataset characteristics. IDIR is effective when sample sizes are limited and categorical variables are numerous, as targeted augmentation can reduce indirect bias while maintaining interpretability. However, when datasets are large, augmented samples may exceed the original data scale, and adversarial learning becomes more suitable given sufficient samples for stable optimization. Model choice also affects outcomes. Support Vector Machines achieved stronger fairness improvements with minimal accuracy loss. Logistic Regression showed limited ability to suppress complex bias patterns, especially for racial attributes. Random forests exhibited mixed results, sometimes with fairness deteri-

oration, suggesting unfavorable interactions between ensemble decision boundaries and certain mitigation strategies. Despite their effectiveness, hyperparameters such as the repair strength coefficient α and the adversarial trade-off parameter λ require careful tuning, and sparse minority groups further constrain both data-level repair and adversarial learning.

VI. Conclusion

This paper compares two improved bias-mitigation methods, improved disparate impact remover (IDIR) and adversarial learning (AL). Both methods demonstrate effective debiasing performance. IDIR is a dual-metric detector and repair framework that combines KL divergence and total variation (TV) distance with controlled data repair. AL applies an encoder–predictor–adversary architecture to reduce dependence between predictions and sensitive features. Future work can extend IDIR and AL to intersecting sensitive features [28], [29] and develop adaptive repair strategies [30], [31] that automatically adjust to dataset-specific bias patterns.

VII. Acknowledgement

Weiru Liu is partially funded by ESRC Centre for Sociodigital Futures (ES/W002639/1) and Jingyu Hu is funded by Doctoral Training Partnership Studentship of Engineering and Physical Sciences Research Council (EPSRC-DTP, EP/W524414/1/2894964).

References

[1] Mohsen Khosravi, Zahra Zare, Seyyed Morteza Mojtabaeian, and Reyhane Izadi, "Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews," *Health services research and managerial epidemiology*, vol. 11, pp. 23333928241234863, 2024.

[2] Yu Huang, Jingchuan Guo, Wei-Han Chen, Hsin-Yueh Lin, Huilin Tang, Fei Wang, Hua Xu, and Jiang Bian, "A scoping review of fair machine learning techniques when using real-world data," *Journal of Biomedical Informatics*, vol. 151, pp. 104622, 2024.

[3] René F. Kizilcec and Hansol Lee, "Algorithmic fairness in education," in *The Ethics of Artificial Intelligence in Education*, Wayne Holmes and Kaska Porayska-Pomsta, Eds., pp. 174–202. Routledge, London, 2022.

[4] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1083–1094, 2022.

[5] Ebikella Mienye, Nobert Jere, George Obaido, Ibomoiye Domor Mienye, and Kehinde Aruleba, "Deep learning in finance: A survey of applications and techniques," 2024.

[6] Thyago P Carvalho, Fabrízzio AAMN Soares, Roberto Vita, Roberto da P Francisco, João P Basto, and Symone GS Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computers & Industrial Engineering*, vol. 137, pp. 106024, 2019.

[7] Dana Pessach and Erez Shmueli, "A review on fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[8] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[9] Alessandro Fabris, Nina Baranowska, Matthew J Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J Biega, "Fairness and bias in algorithmic hiring: A multidisciplinary survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 1, pp. 1–54, 2025.

[10] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 259–268, ACM.

[11] Faisal Kamiran and Toon Calders, "Classification with no discrimination by preferential sampling," in *Benelearn 2010: Annual Machine Learning Conference of Belgium and The Netherlands*, 2010, pp. 1–6.

[12] Harrison Edwards and Amos Storkey, "Censoring representations with an adversary," in *International Conference on Learning Representations (ICLR)*, 2016, Workshop Track; arXiv:1511.05897.

[13] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel, "Learning adversarially fair and transferable representations," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 3384–3393, PMLR.

[14] Batya Friedman and Helen Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1988.

[15] Joy Buolamwini and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* 2018)*. 2018, vol. 81 of *Proceedings of Machine Learning Research*, pp. 77–91, PMLR.

[16] Faisal Kamiran and Toon Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

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

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

[19] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang, "Learning fair representations via an adversarial framework," *arXiv preprint arXiv:1904.13341*, 2019.

[20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[21] Jingyu Hu, Jun Hong, Mengnan Du, and Weiru Liu, "Proximix: Enhancing fairness with proximity samples in subgroups," *arXiv preprint arXiv:2410.01145*, 2024.

[22] Dheeru Dua and Casey Graff, "Uci machine learning repository," http://archive.ics.uci.edu/ml, 2019, Adult Data Set, originally contributed by Ronny Kohavi and Barry Becker.

[23] Sérgio Moro, Paulo Cortez, and Paulo Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[24] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.," *ProPublica*, May 2016.

[25] Moritz Hardt, Eric Price, and Nati Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, vol. 29, pp. 3315–3323.

[26] Sahil Verma and Julia Rubin, "Fairness definitions explained," *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7, 2018.

[27] U.S. Supreme Court, "Griggs v. duke power co., 401 u.s. 424," https://supreme.justia.com/cases/federal/us/401/424/, 1971, Landmark case that established the doctrine of disparate impact under U.S. employment law.

[28] Usman Gohar, Yang Lu, Peter Li, and Imran Razzak, "A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges," *arXiv preprint arXiv:2305.06969*, 2023.

[29] Hongbo Bo, Jingyu Hu, Debbie Watson, and Weiru Liu, "Failing on bias mitigation: Investigating why predictive models struggle with government data," *arXiv preprint arXiv:2601.17054*, 2026.

[30] Junyi Chai and Xiaoqian Wang, "Fairness with adaptive weights," in *International conference on machine learning*. PMLR, 2022, pp. 2853–2866.

[31] Jingyu Hu, Hongbo Bo, Jun Hong, Xiaowei Liu, and Weiru Liu, "Mitigating degree bias adaptively with hard-to-learn nodes in graph contrastive learning," *arXiv preprint arXiv:2506.05214*, 2025.