SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

Algorithmic Bias in Automated Decision-Making: A Statistical Study with Legal and Regulatory Implications

Amir Ahmad Dar [1,*], Shaik Afsar Jahan [1], Mohammad Shahfaraz Khan [2], Imran Azad [2], Murtaza M. Junaid Farooque [3], S. Sindhuja [4] and A. K. Abidha [5]

[1] Department of Statistics, Lovely Professional University, Jalandhar 144411, India

[2] College of Economics and Business Administration, University of Technology and Applied Sciences-Salalah, Salalah 211, Oman

[3] Department of MIS, College of Commerce and Business Administration, Dhofar University, Salalah 211, Oman

[4] Department of Mathematics, SRM Institute of Science and Technology, Chennai 600089, India

[5] Department Mathematics and Actuarial Science, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai 600048, India

[*] Correspondence: sagaramir200@gmail.com

Received: 19 January 2026; **Revised:** 23 February 2026; **Accepted:** 9 March 2026; **Published:** 9 April 2026

Abstract: The use of algorithmic decision systems is being expanded to high-risk areas like credit, recruiting, and distributing government resources. Despite the fact that these systems are usually claimed to be objective and efficient, there have been apprehensions about the likelihood of structural inequalities being perpetuated by the systems. This paper examines the effect of a fairness-aware pre-processing technique called reweighing on the performance of a predictive system in a controlled simulation environment. Using a synthetically created credit approval dataset with structural disadvantage embedded, we compare the performance of a logistic regression classifier with and without reweighing. Fairness is measured using demographic parity disparity (DPD), disparate impact ratio (DIR), and equalized odds difference (EO), along with predictive accuracy. In a single test scenario (seed = 42), reweighing does not improve all fairness metrics uniformly. However, when analyzed for robustness across 50 independent random seeds, we find modest average reductions in demographic parity disparity and equalized odds difference for reweighing, with little change in predictive accuracy. Threshold sensitivity analysis also shows that fairness metrics are sensitive to decision thresholds. These results show that fairness-aware pre-processing can lead to systematic improvements in expectation, although trade-offs across fairness metrics and performance remain context-dependent.

Keywords: Algorithmic Bias; Fairness-Aware Machine Learning; Demographic Parity; Disparate Impact; Equalized Odds; AI Governance

1. Introduction

Algorithmic decision-making systems are being used as a tool to assist with, or make, high-stakes decisions within areas such as credit assessment, employment, the distribution of social benefits, and the justice system [1,2]. Such systems are commonly offered as a more efficient, neutral, and sound alternative to human decision-making [3]. Despite this, a growing number of empirical studies have illustrated that algorithmic models are capable of replicating

existing inequalities when trained on a dataset that captures past inequalities [1,4,5]. This is particularly concerning in areas that consider equality a protected norm [6].

The central challenge in the governance of algorithmic decision-making can be seen as the mismatch between technical optimization objectives and legal norms [2,7]. Machine learning models are usually designed to optimize predictive performance, often in terms of accuracy or metrics derived thereof [8]. Legal frameworks, in turn, aim at the absence of unjustified discriminatory effects, whether they occur intentionally or indirectly [6,9]. It follows that very accurate models can still produce outcomes incompatible with anti-discrimination principles [10].

The following paper defines and makes a clear distinction between the concepts of bias, fairness and legality: Bias refers to the systematic statistical differences in model output among groups [11]. Fairness metrics refer to a group of quantitative diagnostics that may be applied to detect and quantify such differences [12,13]. The legislation, however, does not tend to take abstract concepts of fairness as freestanding commitments, but instead outlaws direct or indirect discriminatory practices which are not justified [9,14]. Therefore, this study is not supposed to perform doctrinal legal analysis or determine legal compliance. Instead, it employs statistical measures of fairness as analytical instruments to evaluate how the output of algorithms can be in conflict with legally cognizable issues of discrimination.

In order to examine such questions, this research employs a simulation study. It uses a synthetically created data set to simulate a stylized credit approval problem, in which structural adversity is introduced via differential distribution of features as well as success probabilities for different groups. The logistic regression classifier is trained as a baseline, which corresponds to a very common tool used for decision making that is easy to interpret in a regulated environment [15]. The performance of the models is assessed with regard to accuracy, as well as fairness metrics such as Demographic Parity Difference (DPD) and Disparate Impact Ratio (DIR), which are commonly cited in surveys on the impact of discrimination [16].

Then, this paper assesses the efficacy and cost of one preprocessing intervention based on reweighing aimed at reducing disparities at the group level [17,18]. The three main reasons why reweighing has been chosen are that it acts directly at the level of the data, it does not change any feature values, and it is also suitable for many auditability requirements in regulated decision-making. The resultant trade-off between fairness and accuracy is then explicitly factored into account, hyphenating the importance of benefits in group-level parity at the expense of non-trivial losses in predictive performance [19,20].

In this paper, three contributions have been made. First, it develops a clear and replicable simulation model on how indirect algorithmic discrimination is studied in stylized structural disadvantage conditions. Second, it empirically evaluates a bias mitigation policy based on reweighing with a set of outcomes of group-level fairness that is relevant in a legal and regulatory context. Finally, it conducts an interdisciplinary dialogue in which statistical data is incorporated within broader legal and governance frameworks by explicating the potential and the constraints of algorithm design in a fairness-conscious way [21,22].

2. Background and Related Literature

2.1. Algorithmic Bias and Fairness Metrics

The study of algorithmic bias has expanded quickly in computer science, statistics, and the social sciences, as an increasing number of people are worried about the social impact of automated decision-making [11,16,23]. Initial studies have demonstrated that machine learning systems that are trained using historical data can reproduce actual inequalities, even when the attributes that might cause such inequalities are not included in the models as protected [1,4]. This phenomenon, often referred to as indirect discrimination or disparate impact, arises through correlations between sensitive attributes and superficially neutral features.

Such disparities have been quantified using a variety of formal metrics of fairness proposed by scholars [12,13,24]. The most studied metrics are group-level metrics, including demographic parity, equalized odds, and disparate impact [12,25,26]. Demographic parity checks if different groups receive positive outcomes at comparable rates, whereas disparate impact compares the ratios of outcomes between advantaged and disadvantaged groups. These measures are highly relevant in regulated domains because they resonate with the legal doctrines that also focus on group-level effects rather than individual intent.

However, these metrics are not mutually compatible. The result of impossibility has revealed that it is rather

difficult to satisfy a set of multiple notions on fairness simultaneously when base rates vary from group to group [25]. This situation has engendered a consensus on the fact that fairness is more than a technical problem, which has been dealt with by treating the choice of fairness as a normative choice that has to be considered from the perspective of domain-specific risks, legal standards, and societal norms [21,22,27].

2.2. Legal/Regulatory View of Discrimination by Algorithms

In a legal context, the discussion on bias in algorithms is generally framed within a pre-existing set of legal foundations against discrimination, rather than as a set of mandates on bias in algorithms [6,9]. In most countries, such as the United States, the European Union, and India, a distinction is made within the law on discrimination between direct discrimination and Indirect Discrimination. Indirect Discrimination happens when a neutral practice affects a particular group of protected people in a disproportionate manner that is not justified.

In the United States, disparate impact analysis has long been applied in employment and credit contexts, notably under Title VII of the Civil Rights Act and the Equal Credit Opportunity Act [9]. The various European legal frameworks forbid indirect discrimination under EU equality law but emphasize proportionality and justification [14, 28]. In India, there is a guarantee of equality before the law in Article 14 of the Constitution, which the courts have interpreted to refer to direct and indirect discrimination [29].

The recent regulatory trends on the consideration of artificial intelligence, including the European AI Proposed Regulation, do not include any mandatory requirement concerning the requirement of fairness in the algorithmic system [30,31]. The regulatory instruments have largely been taken into account through the evaluation of risks, in terms of transparency, human judgment, and mitigating discrimination in the high-risk systems. Some scholars, especially jurists have claimed that the inputs of statistical measurements in the context of fairness are not adequate in legal terms, especially due to the situational nature of fairness [22,27].

2.3. Mitigation Techniques for Bias in Machine Learning

This has, in turn, prompted a sustained and vigorous program of research to develop algorithmic techniques for mitigating bias [17, 18, 24]. These approaches are typically grouped into three classes: pre-processing, in-processing, and post-processing methods.

Among pre-processing methods, re-weighting has gained popularity because of its simplicity [17]. The process involves assigning a certain weight to the training instances based on the category membership, thus balancing all the instances when training a model. This has been a significant advantage, especially when working in a regulated environment, where changing the original data might affect the traceability of the data.

The empirical research has demonstrated that the reweighing can significantly eliminate group-level differences in most areas although with a high likelihood to decrease predictive accuracy [19, 20]. This trade-off has been strongly documented and it constitutes a major dilemma in the machine learning consciousness of fairness: to lessen the likelihood of discrimination, there will need to be a trade-off between performance and a fixed extent. In more recent work, the idea of such trade-offs has been stressed to be considered in the context of domain-specific harms, legal requirements and institutional interests rather than as technical optimization problems [21,22].

2.4. Recent Interdisciplinary Developments

Much more recent work has also begun to view the concept of algorithmic fairness as an interdisciplinary project, including both theoretical frameworks in law and policy, as well as practical applications in particular domains [21–23]. On their side, legal scholars have highlighted algorithmic bias as a primary dilemma of AI governance by pointing to a needed complement to statistical methods in the shape of institutional protections, transparency standards and channels of contestations [22, 27]. In the meantime, empirical studies in other fields like finance, environmental policy, and engineering demonstrate the point that fairness-conscious modeling can be supportive to system robustness but does not eliminate the need to involve human intervention.

This literature raises two critical issues to the fore that are specifically relevant to the present study. The first one is that the measures of fairness can be utilized to provide empirical determination of whether there is group-level bias, yet they do not offer a comprehensive view of the situation of decision-making that is normatively suitable in the legal context [22]. The second is that bias mitigation methods are to be assessed in regard to improvement, but also in regard to the implications of bias mitigation that go beyond quantitative improvement [21].

The current research draws on these observations to provide a transparent illustration of bias mitigation within a legally-informed framework.

3. Methodology

3.1. Experimental Framework

The effect of a pre-processing fairness intervention, reweighing, on predictive accuracy and several fairness metrics in a controlled supervised learning setting will be assessed. The goal is to examine behavior in a single run as well as robustness properties across multiple stochastic runs.

The experimental workflow includes:

1. Generation of synthetic data with a structural disadvantage explicitly embedded.
2. Use of logistic regression for supervised classification.
3. Use of the reweighing algorithm as a fairness mitigation technique [17].
4. Evaluation based on predictive and fairness metrics.
5. Robustness analysis over 50 different random seeds.
6. Analysis of sensitivity to the threshold.
7. Analysis of the weights assigned to instances by the model.

All experiments were carried out in Python. The implementation of the predictive model was done using scikit-learn [18], while the fairness mitigation technique was done using the AIF360 library [19].

3.2. Synthetic Data Generation

In order to retain complete control over the experiments, a synthetic dataset was created for each experiment run.

3.2.1. Sample Size and Protected Attribute

The size of the dataset produced by each run is as follows:

$$N = 1000$$

Each data point is also assigned a protected group attribute as follows:

- Group A (privileged): 60%.
- Group B (unprivileged): 40%.

3.2.2. Feature Construction

Each data point has three continuous features:

- Income: Normally distributed with a mean of 50,000 and a standard deviation of 15,000.
- Age: Normally distributed with a mean of 35 and a standard deviation of 10.
- Credit score: Normally distributed with a mean of 650 and a standard deviation of 50.

These features are generated independently.

3.2.3. Structural Disadvantage Mechanism

To model systemic bias, Group B is given a structural disadvantage at the feature level and outcome level.
Feature-level disadvantage:
The income of Group B is lowered by 5,000 units.
The credit score of Group B is lowered by 30 points.
Outcome-level disadvantage:
A negative shift of −0.5 is applied to the latent logit in the approval model for Group B.
The latent approval score is defined as:

$$z = 0.00002 \cdot income + 0.002 \cdot age + 0.005 \cdot credit\ score - 4$$

For Group B:

$$z = z - 0.5$$

Approval probability is obtained via the logistic function:

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

Binary outcomes are then sampled from a Bernoulli distribution with probability $P(Y = 1)$.

This setup imposes both:

- Observable disparity through differences in features, and
- Structural disparity through a direct group-level penalty on the latent decision function.

The data generation process is stylized to enable analysis of fairness interventions rather than modeling a particular real-world setting.

3.3. Train-Test Splitting and Preprocessing

For each experimental run:

- The dataset is split into 70% for training and 30% for testing.
- Stratified sampling is used based on the binary outcome label.
- The random seed used for splitting is set to be the same as the experimental seed.

The preprocessing of the features can be described as follows:

- The continuous features are normalized using z-score normalization.
- The scaler is trained on the training data alone.
- The test data is transformed using the statistics from the training data.

Crucially, membership in the protected groups is not used as a predictive feature in the classifier, such that any observed disparity is due to the correlation of features, not group membership.

3.4. Baseline Classification Model

The baseline classification model is logistic regression, trained with the following parameters:

- Maximum likelihood estimation,
- L2 regularization (default implementation),
- Maximum iterations set to 1,000.

The predictions are made using the default probability threshold of 0.5, except in the threshold sensitivity analysis.

The training data only contains the standardized feature inputs and not the sensitive attribute. The use of logistic regression is an indicator of a group of models that have been extensively applied in regulated decision-making processes [7], in which interpretability and auditability are more important than modeling complexity.

3.5. Fairness Mitigation via Reweighing

The algorithm of Fairness mitigation is a Reweighing algorithm of AIF360 [19].

Reweighing is a pre-processing technique that is used to weight the instance to balance the joint distribution of:

- Protected group membership, and
- Outcome label [17].

The process is applied in the following way:

- Group A is referred to as the privileged group.
- Group B is the group that is unprivileged.
- Calculate Instance weights only on the training data, i.e., reweighting.

- The logistic regression model is retrained using these weights using the sample weight argument.

The intervention only influences the effective weighting of training instances, whereas the values of the features and the test distribution remain unchanged.

3.6. Evaluation Metrics

All evaluation metrics are computed on the held-out test set.

3.6.1. Accuracy

Accuracy measures the proportion of correctly classified instances:

$$Accuracy = \frac{TP + TN}{Total}$$

3.6.2. Demographic Parity Disparity (DPD)

Demographic parity disparity measures the absolute difference in predicted approval rates between groups [12]:

$$DPD = \mid P(\hat{Y} = 1 \mid A) - P(\hat{Y} = 1 \mid B) \mid$$

Lower values indicate closer parity in positive prediction rates.

3.6.3. Disparate Impact Ratio (DIR)

Disparate impact ratio is defined as [12]:

$$DIR = \frac{P(\hat{Y} = 1 \mid B)}{P(\hat{Y} = 1 \mid A)}$$

A value of less than 0.80 is often mentioned in regulatory debates as a rule-of-thumb estimate of the possible disparate impact, but is not discussed in this context as a legal finding.

3.6.4. Equalized Odds Difference

Equalized odds difference captures disparities in error rates across groups [13]:

$$EO = \mid TPR_A - TPR_B \mid + \mid FPR_A - FPR_B \mid$$

where:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

This metric jointly captures disparities in both false positive and false negative rates.

3.7. Main Experimental Setup (Seed = 42)

The main experimental setup is performed under a fixed random seed of 42. This setup allows for:

- A reference scenario,
- Comparison between the baseline and re-weighted models,
- Visualization of the group-level predicted approval rates,
- Threshold sensitivity analysis.

The single seed setup is used as an anchor for interpretation before robustness analysis.

3.8. Robustness Analysis across 50 Random Seeds

In order to perform a robustness analysis, the whole process, from data generation to evaluation, is repeated 50 times with different random seeds.

For each seed, the following metrics are computed for both baseline and reweighed models:

- Accuracy,
- Demographic parity disparity (DPD),
- Equalized odds difference (EO).

Mean and standard deviation over seeds are calculated to assess:

- Robustness against stochastic noise,
- Fairness affects consistency,
- Predictive performance stability.

These calculations guarantee that the results are not contingent on a particular random draw.

3.9. Threshold Sensitivity Analysis

To analyze the sensitivity to operational decision policy, the predicted probabilities from the baseline classifier are compared for thresholds from 0.1 to 0.9 with a step size of 0.05.

For each threshold:

- Binary predictions are made,
- Accuracy is calculated,
- Demographic parity disparity is calculated.

The threshold sensitivity of the baseline classifier is analyzed to separate the impact of changes to the training distribution from the operational decision policy. This analysis will determine if the fairness disparities are dependent on the standard threshold of 0.5 or if they are robust to different decision boundaries.

3.10. Reweighing Weight Diagnostics

In order to measure the power and the efficacy of the intervention of the fairness, descriptive statistics of the learned weights are computed, including:

- Mean weight,
- Maximum weight,
- 95th percentile,
- Full distribution visualization.

These tests are aimed at making sure that the reweighing does not create extreme leverage or numerical instability, and that the intervention is a moderate redistribution and not a redistribution that is hostile.

4. Results

4.1. Main Experimental Results (Seed = 42)

To have a deterministic setting, a significant experiment is conducted with a fixed random seed (seed = 42). This single-seed test allows the head-to-head comparison of the baseline classifier and the reweighed model on the same data realization and split (**Table 1**).

Table 1. Main Experimental Results (Seed 42).

Model	Accuracy	DPD	DIR	EO Difference
Baseline	0.6200	0.1624	0.7649	0.2440
Reweighed	0.6267	0.1745	0.7582	0.2687

In the case of the baseline model, the test accuracy is 0.6200 indicating a moderate degree of predictive accuracy on the synthetic approval decision. Nonetheless, the difference in the demographic parity (DPD) is 0.1624

indicating that there is no trivial difference in the predicted approval rates between the privileged and unprivileged subpopulations. The disparate impact ratio (DIR) stands at 0.7649 which falls short of the traditional 0.80 cutoff value that is addressed in the context of disparate impact. Equalized odds difference is 0.2440 indicating a joint difference of both in true positive and false positive rates.

Following the reweighing intervention implementation:

- Accuracy: Accuracy has a slight improvement to 0.6267.
- DPD: DPD is on the rise to 0.1745.
- DIR: DIR declines to a slight extent to 0.7582.

There is also an increase in the equalized odds difference to 0.2687. Therefore, in the current deterministic environment, the reweighing intervention fails to bridge the gap of disparity but instead fails to bridge the parity disparity and error rate disparity slightly. This small gain in accuracy requires that mitigation of fairness is not associated with a performance cost but also, the fairness change is not all good in the specific environment. This fact proves the importance of robustness analysis in a single experimental study.

4.2. Predicted Approval Rates by Group

Figure 1 shows the estimated level of approval of protected groups in the baseline and reweighed models (seed = 42 configuration).

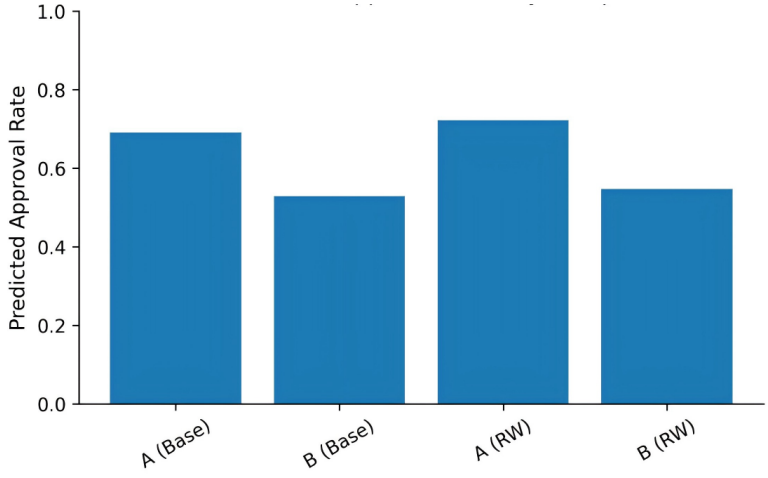

Figure 1. Predicted Approval Rates by Group (Baseline vs. Reweighed).

Under the baseline model:

- Group A has a predicted approval rate of about 0.66.
- Group B has a predicted approval rate of about 0.50.

The absolute difference corresponds to the demographic parity disparity (DPD = 0.1624) reported in **Table 1**. Applying reweighing:

- The predicted approval rate for Group A increases to about 0.69.
- The predicted approval rate for Group B increases slightly to about 0.51.
- The absolute difference obtained is 0.1745, as reported in **Table 1**.

Thus, in the single seed context, reweighing fails to bridge the demographic parity gap. Instead, it causes the separation between the estimated approval rates to be slightly more.

This proves that the mitigation strategies aimed at fairness do not have to be monotonic in ensuring that all metrics are enhanced in one environment. The reweighing effect depends on both the distribution of the synthetic data and the sampling process as well as the learned decision boundary.

4.3. Model Accuracy Comparison

Figure 2 shows the comparison of overall predictive accuracy between the baseline and reweighed models.

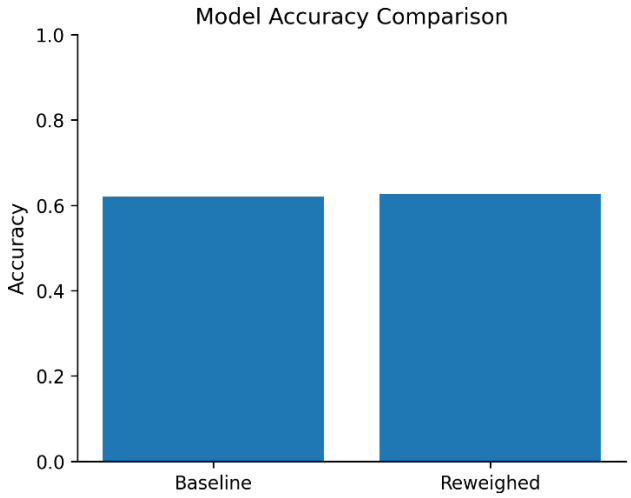

Figure 2. Model Accuracy Comparison (Seed 42).

Baseline model has an accuracy of 0.6200 compared to the reweighed model which has an accuracy of 0.6267. The difference is minute but it is consistent with the numerical difference presented in **Table 1**.

This result demonstrates that instance weights through reweighing do not have any negative effect on the model accuracy. However, the observation that an increased accuracy does not imply better fairness results is also noted.

4.4. Threshold Sensitivity Analysis

To investigate if fairness gaps are contingent on the default decision threshold of 0.5, the predicted probabilities of the baseline classifier were assessed for different thresholds between 0.1 and 0.9.

The solid line in **Figure 3** corresponds to accuracy, and the dashed line corresponds to demographic parity disparity (DPD).

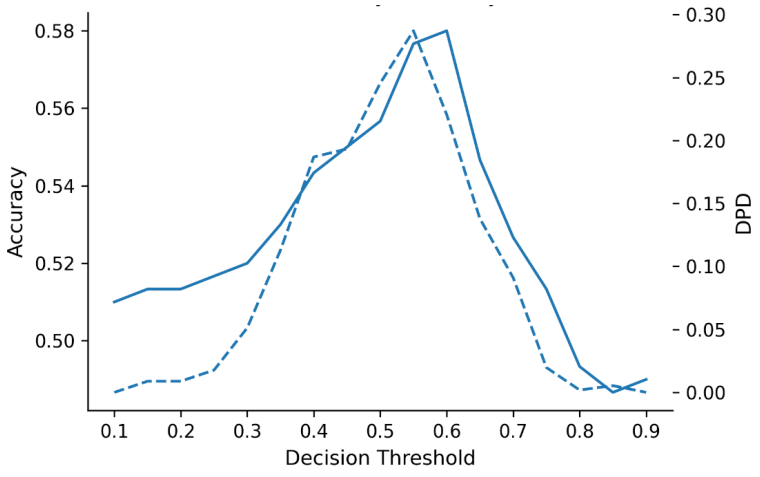

Figure 3. Threshold Sensitivity: Accuracy and Demographic Parity Disparity.

The plot shows several key behaviors:

1. Accuracy Behavior: Accuracy improves as the threshold gets closer to the mid-range and reaches its peak around thresholds of 0.55 and 0.60, where it reaches its local maximum in the mid-range (approximately between 0.55 and 0.60). On the other hand, accuracy gradually decreases towards both ends of the lower and higher thresholds.
2. Fairness Behavior (DPD): DPD is not a linear function of the threshold. It improves as the threshold gets closer to the mid-range and reaches its peak around the points where accuracy is highest. On the other hand, at the extreme points (0.1 and 0.9), DPD drops significantly.
3. Performance-Fairness Interaction: The region where accuracy is highest does not correspond to the region where DPD is lowest. In fact, fairness disparity is higher in the region where the predictive performance is highest.

These results suggest that fairness gaps are not merely a result of the default threshold of 0.5. Although the threshold can be tuned to close the gap in extreme decision boundaries, this is often at the expense of predictive accuracy. Therefore, decision policy selection is a factor that interacts with structural bias but does not mitigate it.

4.5. Robustness Analysis across 50 Random Seeds

To ensure that results for a single seed are robust to random fluctuations, the entire experimental pipeline was run for 50 different random seeds (**Table 2**). For each seed, the values of accuracy, demographic parity disparity, and equalized odds difference were calculated for both the baseline and reweighed models.

Table 2. Mean Results across 50 Seeds.

Metric	Baseline	Reweighed
Accuracy	0.5907	0.5893
DPD	0.2572	0.2400
EO Diff	0.4642	0.4306

Over 50 seeds, several patterns stand out (**Table 3**):

- Accuracy is virtually unaffected (0.5907 vs. 0.5893), suggesting that reweighing does not tend to worsen accuracy.
- DPD reduces on average from 0.2572 to 0.2400 with reweighing.
- Equalized odds difference reduces as well, from 0.4642 to 0.4306.

Table 3. Standard Deviation across 50 Seeds.

Metric	Baseline	Reweighed
Accuracy	0.0249	0.0266
DPD	0.0555	0.0583
EO Diff	0.1207	0.1272

This is unlike the single seed set, the average of the multi-seed set exhibits a distinct, but insignificant fairness enhancement. This is used to show that one observation cannot be reflective of the trend. The small standard deviations suggest that the patterns of fairness observed do not change significantly across stochastic realizations, though no formal hypothesis testing is done.

4.6. Distribution of Reweighing Instance Weights

The learning instance weights were subjected to descriptive statistics to determine how strong the fairness effect was as shown in **Figure 4**.

The weight diagnostics indicate that:

- Mean weight = 1.0.
- Maximum weight ≈ 1.2986.
- 95th percentile ≈ 1.2986.

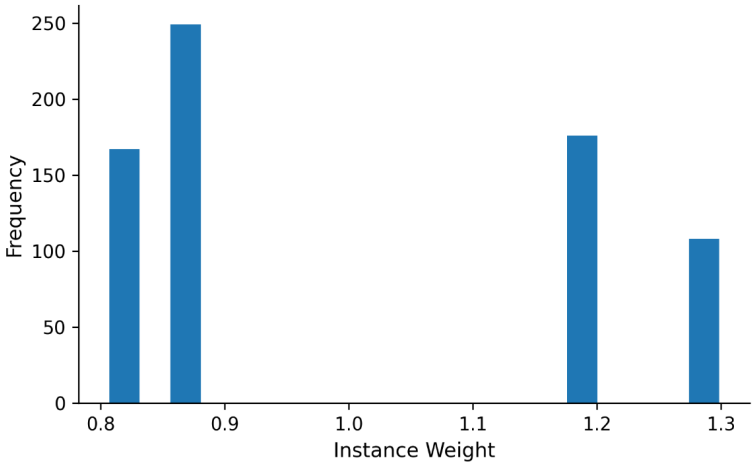

Figure 4. Distribution of Reweighing Instance Weights.

The histogram shows that the weights are constrained by discrete points giving the combinations of group label with outcome values. Most importantly, there are no extreme and massively skewed weights.

The boundedness of the weight distribution indicates that:

- There is no numerical instability,
- There is no extreme leverage effect,
- There is a moderate redistribution of influence for the group-label combinations.

General Implication of Findings

The findings give three valuable lessons:

- The non-monotonic dynamics of fairness may occur in individual runs, as in the case of the seed 42 setting.
- In repeated stochastic runs, reweighing causes small, but systematic decreases in the gaps in fairness, with no significant decrease in accuracy.
- Whereas the threshold selection can influence accuracy and parity, equality gaps do not change in a large interval of operational thresholds.

5. Discussion

The effect of a pre-processing fairness intervention, reweighing, on predictive accuracy and a set of measures of fairness in a controlled experimental setting was examined in this paper. The findings indicate that there is a complicated interplay of mitigation, statistical variation, and decision policy.

In the single seed case (seed = 42), there was a minor increase in predictive accuracy with reweighing but it did not achieve fairness measures consistently. The disparity in demographic parity and equalized odds difference marginally improved in this arrangement but the disparate impact ratio was below the 0.80 mark. It can be seen in this example that the effects of fairness mitigation are not always monotonic at the individual level, and may depend on the stochastic variation.

The strength analysis provided by multi-seed however leads to a more stable perspective. Reweighing on 50 different runs led to a small decrease in the average demographic parity disparity (0.2572–0.2400) and equalized odds difference (0.4642–0.4306), but not a change in predictive accuracy. The comparatively tiny standard deviations indicate that the trends of fairness do not vary among stochastic cases, but no statistical conclusion is drawn.

The threshold sensitivity analysis also demonstrates that the outcomes of fairness do not only rely on the model but also on the decision policy. The difference between the demographic parity is a non-linear relationship on the thresholds and is the greatest at the medium thresholds and declines at the higher/lower thresholds. This once again highlights the importance of distinguishing model bias and policy bias. The weight diagnostics show that the re-weighing process has moderate changes (max weight ≈ 1.30) and does not experience severe leverage effects and numerical instability. The intervention relies on a controlled redistribution as opposed to drastic rescaling.

In summary, the results emphasize three important insights:

- Fairness mitigation effects are not necessarily uniform for individual instances.
- Multi-seed testing is a crucial requirement for drawing valid conclusions.
- Fairness is a multi-dimensional concept that is sensitive to policies.

6. Limitations

The data generation mechanism in this study is synthetic, which allows for experimental control and replicability but does not faithfully represent the complexity of socio-technical systems in the real world. The structural disadvantage mechanism is thus more illustrative than descriptive.

The predictive model considered in this study is restricted to logistic regression. While this is deliberate to reflect the nature of interpretable models in regulated settings, the fairness-accuracy trade-off could be different for more complex models.

Though robustness is evaluated over 50 random seeds, other structural settings and levels of disadvantage are not considered. The findings thus establish robustness over stochastic variation rather than structural variation.

Lastly, fairness evaluation in this study is restricted to group-level statistical metrics. These metrics reflect distributional and error rate disparities but do not subsume individual-level, causal, or more general notions of fairness.

7. Conclusion

This work examined the relationship between predictive accuracy and fairness-aware pre-processing in a controlled simulated decision-making task. We assessed logistic regression models on a synthetically created credit approval task with structural disadvantage injected into the data using the reweighing fairness mitigation method on a dataset with 50 independent random seeds.

Single-seed experiments show that fairness metrics can differ between individual runs, but robustness analysis on 50 independent random seeds shows small but consistent improvements in demographic parity disparity and equalized odds difference using reweighing, with little effect on average predictive accuracy. Threshold sensitivity analysis also shows that fairness metrics are sensitive to decision thresholds, highlighting the complex nature of fairness evaluation in algorithmic decision-making systems.

The results of this work suggest that fairness-aware pre-processing can lead to systematic improvements in group fairness metrics in expectation, but that trade-offs between fairness metrics and predictive accuracy are still context-dependent. Future studies may apply this framework to other model classes, other mitigation techniques, and real-world datasets to further test external validity.

Author Contributions

A.A.D. conceptualized the study, supervised the research, and drafted the manuscript. S.A.J. contributed to data analysis and manuscript writing. M.S.K. and I.A. were involved in data collection, methodology, and initial drafting. M.M.J.F. assisted in formal analysis and validation. S.S. contributed to visualization and editing. A.K.A. supported review, proofreading, and final revisions. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data used in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper. The authors have no financial, personal, institutional, or other relationships that could inappropriately influence or bias the work reported in this manuscript.

AI Use Statement

The authors used ChatGPT3 solely for grammar checking, sentence structure refinement, and improving the readability of the English text in this manuscript. The authors take full responsibility for all academic content, including all ideas, data, analyses, and conclusions presented herein. The use of AI was thoroughly reviewed and supervised by the authors.

References

1. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; et al. Discrimination in the Age of Algorithms. *J. Leg. Anal.* **2018**, *10*, 113–174.
2. Ebers, M. Automating due process—The promise and challenges of AI-based techniques in consumer online dispute resolution. In *Frontiers in Civil Justice*; Edward Elgar Publishing: Cheltenham, UK, 2022; pp. 142–168.
3. Corbett-Davies, S.; Gaebler, J.D.; Nilforoshan, H.; et al. The measure and mismeasure of fairness. *J. Mach. Learn. Res.* **2023**, *24*, 1–117.
4. Barocas, S.; Selbst, A.D. Excerpt from Big Data's Disparate Impact. In *Ethics of Data and Analytics*; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 303–318.
5. Chen, Y.C.; Ahn, M.J.; Wang, Y.F. Artificial intelligence and public values: Value impacts and governance in the public sector. *Sustainability* **2023**, *15*, 4796.
6. Burda, P.; Van Otterloo, S. Fairness definitions explained and illustrated with examples. *Comput. Soc. Res. J.* **2025**, *2*, 1–23.
7. Yan, C.; Zhang, X.; Shen, J. Credit score classification using advanced machine learning: A comprehensive approach. *J. Softw. Eng. Appl.* **2025**, *18*, 98–112.
8. Popat, A.K.; Amemiya, J.; Heyman, G.D.; et al. Hiding discrimination in plain sight: The development of reasoning about disparate impact policies. *J. Exp. Psychol. Gen.* **2025**, *155*, 116–132.
9. Pessach, D.; Shmueli, E. A review on fairness in machine learning. *ACM Comput. Surv.* **2022**, *55*, 1–44.
10. Slussareff, M. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Crown: New York, NY, USA, 2022.
11. Angwin, J.; Larson, J.; Mattu, S.; et al. Machine bias. In *Ethics of Data and Analytics*; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 254–264.
12. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
13. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference, Berkeley, CA, USA, 9–11 January 2017.
14. Ellis, E.; Watson, P. *EU Anti-Discrimination Law*; Oxford University Press: Oxford, UK, 2012.
15. Van Iddekinge, C.H.; Lievens, F.; Sackett, P.R. Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. *Pers. Psychol.* **2023**, *76*, 651–686.
16. The Council of the European Union. *Council Directive 2000/43/EC of 29 June 2000 Implementing the Principle of Equal Treatment between Persons Irrespective of Racial or Ethnic Origin*; The Council of the European Union: Brussels, Belgium, 2000.
17. Bhatia, G. India: A Constitution in Search of an Identity. *SSRN* **2022**. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4075049
18. Lognoul, M. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act–AI Act). *Rev. Droit Technol. Inf.* **2025**, 145–189.
19. De Troya, Í.; Kernahan, J.; Doorn, N.; et al. Misabstraction in Sociotechnical Systems. In Proceedings of the

2025 ACM Conference on Fairness, Accountability, and Transparency, Athens, Greece, 23–26 June 2025; pp. 1829–1842.
20. Kamiran, F.; Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **2012**, *33*, 1–33.
21. Raghavan, M.; Barocas, S.; Kleinberg, J.; et al. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 469–481.
22. Abowd, J.M.; Hawes, M.B. 21st century statistical disclosure limitation: Motivations and challenges. In *Handbook of Sharing Confidential Data*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2024; pp. 24–36.
23. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
24. Loganathan, M.; Sharifzadeh, H.; Keivanmarz, A. Towards Improving Fairness in AI Systems: A Framework for Bias Mitigation. In Proceedings of the 2025 IEEE Region 10 Symposium (TENSYMP), Christchurch, New Zealand, 7–9 July 2025.
25. Wachter, S.; Mittelstadt, B.; Russell, C. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Comput. Law Secur. Rev.* **2021**, *41*, 105567.
26. Zafar, M.B.; Valera, I.; Rogriguez, M.G.; et al. Fairness constraints: Mechanisms for fair classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, April 2017; pp. 962–970.
27. Binns, R. On the apparent conflict between individual and group fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 514–524.
28. Kour, M.; Schutte, D.P. *Artificial Intelligence and Accounting: Ethical, Legal, and Social Implications*; Routledge: London, UK, 2025.
29. Caton, S.; Haas, C. Fairness in machine learning: A survey. *ACM Comput. Surv.* **2024**, *56*, 1–38.
30. Schäferling, S. Automated Decision-Making and the Law. In *Governmental Automated Decision-Making and Human Rights: Reconciling Law and Intelligent Systems*; Springer Nature: Cham, Switzerland, 2023; pp. 23–90.
31. European Commission. *White Paper on Artificial Intelligence—A European Approach to Excellence and Trust*; European Commission: Brussels, Belgium, 2020.