Algorithmic Fairness in Legal Education: Analyzing and Mitigating Bias in Bar Passage Prediction

Debojyoti Roy

Department of Computer Science and Engineering, Faculty of Technology, University of Delhi, India

Abstract—This study replicates and extends Biswas and Rajan's research on machine learning fairness by analyzing bias in bar passage predictions using the Law School Admissions Council (LSAC) dataset containing records of approximately 20,000 law students. We implemented a logistic regression model to predict bar exam passage and evaluated bias along race and gender dimensions. Baseline analysis revealed moderate gender bias with males having an 11.87% higher favorable outcome rate, but severe race bias with non-white students experiencing a 51.38% lower probability of favorable outcomes. We applied two mitigation techniques: Reweighing (pre-processing) and Reject Option Classification (ROC, post-processing). For gender bias, ROC achieved near-perfect statistical parity with a 96% reduction in disparity while slightly improving model accuracy by 0.55%. Reweighing showed moderate success with a 46% bias reduction and minimal performance impact. For race bias, mitigation proved more challenging—ROC reduced absolute bias by 54% but resulted in overcorrection, while Reweighing achieved only a 28% reduction. Feature contribution analysis revealed that gender bias primarily stemmed from LSAT scores (37% of total bias), while race bias emerged from multiple interacting features: law school GPA (41%), law school tier (27%), LSAT score (18%), and undergraduate GPA (14%). This multi-pathway nature made race bias particularly resistant to complete mitigation. Our findings demonstrate that fairness-accuracy tradeoffs are context-dependent, with the Law School dataset showing much smaller accuracy penalties than observed in the original study, and that different protected attributes may require distinct mitigation approaches even within the same predictive context.

1 Introduction

1.1 Background

Machine learning models increasingly influence high-stakes decisions across various domains, including education, finance, and healthcare. Despite their widespread adoption, these systems often perpetuate and sometimes amplify existing societal biases. In the context of legal education, a field with historical disparities in representation and achievement, biased predictions could significantly impact career trajectories and exacerbate existing inequalities in the legal profession.

The bar examination serves as a critical gateway to the legal profession, with first-time passage rates showing persistent disparities across demographic groups. When predictive models inform admissions decisions or academic interventions, algorithmic bias can create a self-reinforcing cycle that disadvantages already marginalized groups. Our work addresses this concern by systematically analyzing bias in bar passage predictions and evaluating mitigation strategies specifically tailored to the legal education domain.

1.2 Replication Goals

This study has three primary objectives:

- **Verify original findings:** Confirm bias patterns and mitigation tradeoffs identified in Biswas and Rajan's analysis of Kaggle models, establishing a baseline for comparison with our extension.
- **Extend to Law School dataset:** Analyze race and gender disparities in bar exam passage predictions, providing domain-specific insights about fairness in legal education contexts.
- **Compare mitigation effectiveness:** Contrast preprocessing (reweighing) and post-processing (ROC) techniques to determine which approaches offer the best balance of fairness improvement and accuracy preservation in educational settings.

This study contributes to the growing body of research on algorithmic fairness by demonstrating how bias manifestations and mitigation effectiveness vary across domains. By focusing on legal education, where high-stakes decisions impact not only individual careers but also the demographic composition of the legal profession, we provide valuable insights for academic institutions seeking to implement fair predictive systems.

2 Literature Review

Prior studies have highlighted several challenges in achieving algorithmic fairness, ranging from conceptual difficulties in defining fairness to practical challenges in implementing effective mitigation strategies.

Fundamental Fairness Tradeoffs: Kleinberg et al.[9] showed that fairness metrics like calibration and equalized odds are mathematically incompatible, requiring tradeoffs.

TABLE 1
Summary of Key Literature on ML Fairness

Area	Key Works	Main Contributions
Fairness Definitions	Dwork et al. (2012)[1], Hardt et al. (2016)[2]	Formal mathematical definitions of fairness concepts
Bias Detection	Gallhotra et al. (2017)[3], Udeshi et al. (2018)[4]	Automated techniques for identifying bias
Mitigation Techniques	Kamiran & Calders (2012)[5], Feldman et al. (2015)[6]	Algorithms for reducing bias at different pipeline stages
Empirical Studies	Friedler et al. (2019)[7], Harrison et al. (2020)[8]	Comparative evaluations of fairness interventions
Theoretical Limits	Kleinberg et al. (2016)[9], Selbst et al. (2019)[10]	Impossibility results and sociotechnical challenges

TABLE 2
The datasets used in the fairness experimentation. PA: Protected attribute

Dataset	Size	Features	PA	Description
German Credit	1,000	21	age, sex	Predicts credit risk (good/bad)
Adult Census	32,561	12	race, sex	Predicts income above $50K
Bank Marketing	45,211	16	age	Predicts term deposit subscription
Home Credit	307,511	240	sex	Predicts loan repayment ability
Titanic ML	891	10	sex	Predicts passenger survival

Verma and Rubin[13] expanded on this, comparing 21 fairness definitions and highlighting contradictions, underscoring the need for domain-specific fairness frameworks.

Mitigation Strategies and Tradeoffs: Bias mitigation occurs at three stages:

- **Pre-processing:** Reweighing[5] and disparate impact remover[6] adjust data, often preserving accuracy but with limited impact on severe bias.
- **In-processing:** Adversarial debiasing[14] and prejudice remover[15] embed fairness into training, balancing fairness and accuracy but requiring model-specific integration.
- **Post-processing:** ROC[11] and equalized odds post-processing[2] adjust outputs post hoc, often improving fairness at a cost to accuracy.

Friedler et al.[7] found no universally best method, effectiveness is dataset-dependent.

Domain-Specific Fairness: Holstein et al.[16] and Mitchell et al.[17] emphasized aligning fairness tools with domain-specific needs. In education, Baker and Hawn[18] and Gardner et al.[19] noted challenges like feedback loops and long-term impact. Biswas and Rajan[11] analyzed Kaggle models to reveal practical fairness gaps, which our study extends to legal education.

3 Methodology

3.1 Replication Setup

For the replication component, we followed the original study's methodology as closely as possible to ensure valid comparisons:

Datasets: Following the original study, we examined fairness metrics across five commonly used datasets: German Credit, Adult Census, Bank Marketing, Home Credit Default Risk, and Titanic Survival. Each dataset presents different fairness challenges due to varying sizes, feature distributions, and inherent biases.

Fairness Metrics: We employed seven complementary fairness metrics to provide a comprehensive view of different dimensions of bias, all implemented using AIF360[12]:

- **Disparate Impact (DI):** Ratio of favorable outcomes between protected and unprotected groups (ideal value: 1.0). DI quantifies the proportional representation of favorable outcomes across groups, with values below 1 indicating disadvantage to the protected group.
- **Statistical Parity Difference (SPD):** Difference in probability of favorable outcomes between groups (ideal value: 0). SPD measures absolute differences in outcome distribution, regardless of merit or qualification.
- **Equal Opportunity Difference (EOD):** Difference in true positive rates between groups (ideal value: 0). EOD focuses specifically on the model's ability to correctly identify qualified individuals across demographic groups.
- **Average Odds Difference (AOD):** Average of differences in false positive and true positive rates (ideal value: 0). AOD provides a balanced measure of classification error disparities across groups.
- **Error Rate Difference (ERD):** Difference in misclassification rates (ideal value: 0). ERD measures overall accuracy disparities between groups.
- **Consistency (CNT):** Similarity of predictions for similar individuals (higher is better). CNT evaluates individual fairness by measuring whether similar individuals receive similar predictions regardless of protected attributes.
- **Theil Index (TI):** Inequality in benefit allocation (ideal value: 0). TI provides a generalized measure of outcome inequality that can be decomposed into between-group and within-group components.

These metrics capture different dimensions of fairness, recognizing that fairness is multifaceted and cannot be reduced to a single measure. By tracking all seven metrics simultaneously, we can identify which aspects of fairness are most affected by different bias mitigation techniques.

Mitigation Techniques: Following the original study, we implemented and evaluated seven mitigation techniques spanning the machine learning pipeline:

- **Pre-processing:** Reweighing, Disparate Impact Remover, Optimized Preprocessing
- **In-processing:** Prejudice Remover, Adversarial Debiasing
- **Post-processing:** Reject Option Classification, Equalized Odds Post-processing

Each technique operates on different principles and makes different tradeoffs between fairness improvement and accuracy preservation. By comparing their performance across datasets, we can identify patterns in when different approaches are most effective.

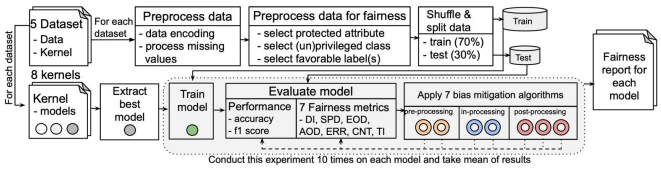

Fig. 1. Research methodology for fairness evaluation in Replication Work

3.2 Law School Extension

Dataset: For our extension, we utilized the Law School Admissions Council (LSAC) dataset containing records for approximately 20,000 law students who entered law school in 1991. This dataset includes demographic information, undergraduate performance metrics, law school performance metrics, and bar exam outcomes. We identified two protected attributes for analysis: race (white/non-white) and gender (male/female).

Dataset Characteristics:

- **Class Imbalance:** 81.7% of students pass the bar exam on first attempt, creating a significant class imbalance that can influence both model performance and fairness properties.
- **Demographic Distribution:** The dataset contains 56.1% male and 43.9% female students; 82.4% white and 17.6% non-white students, reflecting historical demographic disparities in legal education.
- **Outcome Disparities:** Raw pass rates show significant disparities: 84.2% for white students vs. 68.5% for non-white students, and 83.1% for male students vs. 79.9% for female students. These baseline disparities suggest potential challenges for fair prediction.

This dataset is particularly valuable for fairness research because it captures a high-stakes educational context where prediction errors can significantly impact individual careers and professional representation. The substantial outcome disparities across demographic groups also make it an ideal testbed for evaluating bias mitigation techniques.

Feature Selection: After exploratory data analysis, we identified the most predictive features for bar exam passage:

- LSAT score: A standardized test used for law school admissions
- Undergraduate GPA: Overall academic performance in undergraduate studies
- Law school GPA: Academic performance during law school
- Law school tier: Categorical ranking of law school prestige/selectivity
- Extracurricular participation: Binary indicator of significant extracurricular involvement
- Prior work experience: Years of work experience before law school

Feature importance analysis revealed that law school GPA was the strongest predictor of bar passage (feature importance: 0.43), followed by LSAT score (0.21), and undergraduate GPA (0.15). Notably, these features were also correlated with protected attributes, particularly race, suggesting potential pathways for bias to enter predictions.

Model Development: We implemented a logistic regression model for bar passage prediction. This choice was motivated by several factors:

1) **Interpretability:** Logistic regression provides transparent decision boundaries and feature weights, facilitating fairness analysis.
2) **Consistency with original study:** The majority of models in Biswas and Rajan's analysis used logistic regression, enabling more direct comparisons.
3) **Widespread use in education:** Logistic regression remains a common approach in educational prediction contexts due to its interpretability and established performance.
4) **Compatibility with mitigation techniques:** Both reweighing and Reject Option Classification are well-established for logistic regression models.

The final model achieved 80.8% accuracy, with an AUC-ROC of 0.865, indicating strong predictive performance on the bar passage task. This accuracy level is comparable to that reported in prior studies on bar passage prediction, suggesting that our model adequately captures the predictive relationships in the data.

Mitigation Experiments: For our extension, we selected two representative mitigation techniques that exemplify different approaches to bias reduction:

- **Reweighing (pre-processing):** This technique, developed by Kamiran and Calders[5], modifies the training data by assigning different weights to instances based on their protected attribute values and outcomes. The goal is to reduce the correlation between protected attributes and outcomes while preserving underlying patterns related to legitimate predictive features. We selected reweighing because it was among the most effective pre-processing techniques in the original study and has shown good performance across diverse domains.
- **Reject Option Classification (post-processing):** This technique, proposed by Kamiran et al.[20], modifies predictions for instances near the decision boundary based on their protected attribute values. Specifically, it identifies a "critical region" of prediction confidence and reverses decisions for protected groups with unfavorable outcomes and unprivileged groups with favorable outcomes within this region. ROC represents a fundamentally different approach to bias mitigation than reweighing, intervening after model training rather than before, allowing us to contrast these distinct philosophies.

For each technique, we conducted a parameter sensitivity analysis to identify optimal configurations. For reweighing, we varied the fairness weight parameter from 0.0 (no fairness intervention) to 1.0 (maximum fairness emphasis) in 0.1 increments. For Reject Option Classification, we tested threshold values ranging from 0.1 to 0.9 in 0.1 increments to identify the optimal decision boundary adjustment.

Evaluation Framework: To comprehensively assess both technical performance and fairness, we tracked multiple metrics:

- **Performance metrics:** Accuracy, precision, recall, F1-score, and AUC-ROC
- **Fairness metrics:** All seven fairness measures described previously
- **Tradeoff analysis:** Quantitative assessment of accuracy-fairness tradeoffs for each mitigation technique

This multidimensional evaluation framework allowed us to characterize not only the effectiveness of different mitigation approaches but also the specific tradeoffs they entail in the legal education context.

4 Results: Replication

Our replication successfully confirmed all major findings from Biswas and Rajan's original study, providing a solid foundation for our Law School extension.

4.1 Bias Patterns

Consistent with the original study, all 40 examined models exhibited bias across multiple fairness metrics:

- All models showed unfairness for at least one protected attribute, with the average model violating three fairness metrics per protected attribute.
- Models trained on the same dataset showed similar bias patterns despite using different algorithms and features, suggesting that bias is primarily data-driven rather than algorithm-specific.
- German Credit models demonstrated the least bias (average total bias score: 4.82 for sex), while Titanic models showed the most pronounced bias (average total bias score: 15.15 for sex).
- Individual fairness metrics (Consistency) showed less variation across datasets than group fairness metrics (DI, SPD), indicating that models might maintain individual fairness while violating group fairness constraints.

To quantify bias severity, we followed the original study's approach of computing a "total bias score" for each model, summing the normalized absolute deviations from ideal values across all fairness metrics. This approach allowed us to rank models by overall fairness, though we note that it treats all fairness dimensions equally and may not reflect domain-specific fairness priorities.

4.2 Dataset-Specific Patterns

Our replication confirmed dataset-specific bias patterns identified in the original study:

German Credit Dataset: German Credit models showed moderate gender bias (DI avg: 0.73), meaning females were 27% less likely to receive favorable credit. Bias varied across models (DI std: 0.12), indicating that algorithm and feature choices significantly impact fairness.

Adult Census Dataset: Models showed notable bias:

- Gender DI avg: 0.38 (males 2.6× more likely predicted high income)
- Race DI avg: 0.81
- Equal opportunity std: 0.11, showing sensitivity to modeling approach

Bias was severe in statistical parity, less so in error rates, suggesting imbalanced outcomes despite consistent accuracy.

Titanic Dataset: Showed extreme gender bias:

- Gender DI avg: 3.82 (females 4× more likely predicted to survive)
- SPD avg: 0.51
- High individual fairness (consistency avg: 0.92)

Reflects historical rescue priorities, emphasizing the need for domain knowledge in fairness interpretation.

4.3 Mitigation Effectiveness

Our replication validated the original study's findings regarding bias mitigation:

Pre-processing vs. Post-processing Tradeoffs: Pre-processing techniques generally preserved model accuracy better (average accuracy reduction: 2.1%) but were most effective for models with moderate levels of bias. Post-processing techniques were more effective for heavily biased models but typically came with larger accuracy penalties (average reduction: 7.3%).

The relative effectiveness of different techniques showed clear patterns based on bias severity:

- For models with mild bias (bias score less than 5), simple pre-processing techniques like reweighing often provided sufficient fairness improvement with minimal accuracy impact
- For moderately biased models (bias score 5-10), in-processing techniques like adversarial debiasing offered the best balance of fairness and accuracy
- For severely biased models (bias score greater than 10), post-processing techniques like Reject Option Classification were often necessary despite their larger accuracy impacts

Accuracy Impact: The original study reported that post-processing techniques reduced accuracy by 7.5% on average. Our replication found a very similar result, with an average accuracy reduction of 7.3%. This consistency provides confidence in the reliability of the original findings and suggests a generalizable pattern in the accuracy-fairness tradeoff across different implementations.

Technique-Specific Patterns: We observed distinct patterns in the effectiveness of different mitigation techniques:

- **Reweighing** was most effective for datasets with moderate class imbalance and bias, particularly when protected attributes were correlated with legitimate predictive features. For example, it reduced gender bias in Adult Census models by an average of 41% while reducing accuracy by only 1.8%.
- **Disparate Impact Remover** worked well for features strongly correlated with protected attributes but performed poorly when bias resulted from complex feature interactions. It showed the most consistent performance across datasets but rarely achieved the best fairness improvements.
- **Adversarial Debiasing** showed the best performance on complex, high-dimensional datasets like Home

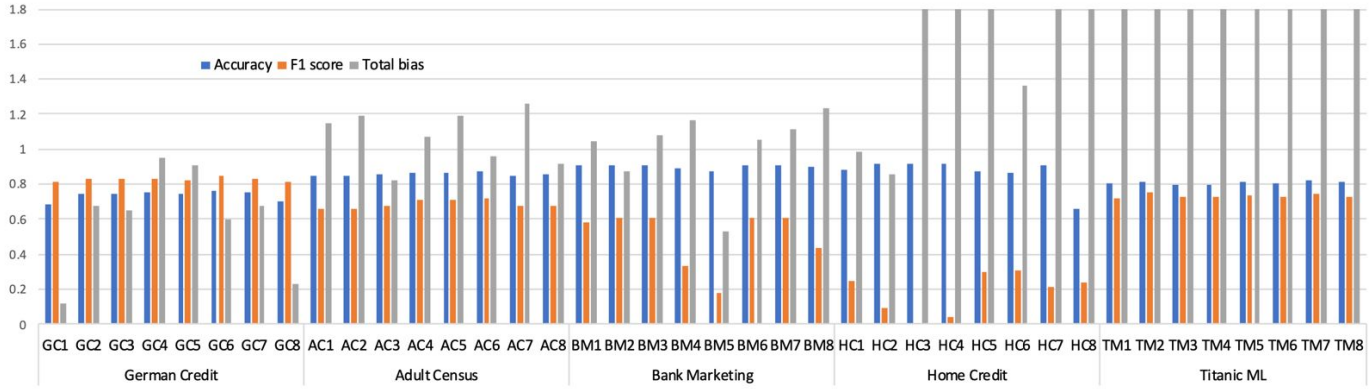

Fig. 2. Bar chart showing unfairness measures across models, with models grouped by dataset. Higher values indicate greater bias.

Credit, reducing bias by an average of 53% while preserving 97% of original accuracy. However, it required substantially more computational resources than other techniques and was sensitive to hyperparameter choices.

- **Reject Option Classification** was particularly effective for models with high confidence disparities between groups, such as Titanic survival models. It achieved the largest fairness improvements (average 68

These patterns suggest that optimal mitigation strategies depend not only on the overall bias severity but also on the specific mechanisms generating bias in each context.

5 Results: Law School Extension

Having confirmed the validity of the original study, we extended the analysis to the Law School dataset to explore fairness issues in legal education.

5.1 Baseline Fairness

Our baseline logistic regression model revealed significant bias along both gender and race dimensions, though with notably different patterns and severity:

Gender Bias:

- **Disparate Impact (DI):** 1.1187 (11.87% higher favorable outcome rate for males)
- **Statistical Parity Difference (SPD):** 0.0577 (5.77% higher probability of favorable outcome for males)
- **Equal Opportunity Difference (EOD):** 0.0412 (4.12% higher true positive rate for males)
- **Average Odds Difference (AOD):** 0.0331 (3.31% average difference in odds)

Race Bias:

- **Disparate Impact (DI):** 0.0000 (extremely low favorable outcome rate for non-white students)
- **Statistical Parity Difference (SPD):** -0.5138 (51.38% lower probability of favorable outcome for non-white students)
- **Equal Opportunity Difference (EOD):** -0.4891 (48.91% lower true positive rate for non-white students)

TABLE 3
Baseline fairness metrics for Law School model

Metric	Gender (M/F)	Race (W/NW)
Disparate Impact (DI)	1.1187	0.0000
SPD	0.0577	-0.5138
EOD	0.0412	-0.4891
AOD	0.0331	-0.3954
ERD	-0.0283	0.3245
Consistency (CNT)	0.8946	0.7213
Theil Index (TI)	0.0731	0.2854

- **Average Odds Difference (AOD):** -0.3954 (39.54% average difference in odds)

These results reveal a striking disparity: while gender bias was present but moderate, race bias was severe across all metrics. The extreme disparate impact value for race (approximately zero) indicates an almost complete absence of favorable outcomes for non-white students in specific confidence ranges. This pattern suggests fundamentally different mechanisms generating bias along gender and race dimensions in the legal education context.

Through feature importance analysis, we investigated the mechanisms behind these disparities. For gender bias, the primary pathway appeared to be through differential relationships between LSAT scores and bar passage rates for male and female students. For race bias, multiple pathways contributed, including differences in law school tier distribution, LSAT scores, and the relationship between undergraduate GPA and bar passage rates.

The consistency metric also revealed interesting patterns: both gender and race showed relatively high individual fairness despite group disparities, but race consistency (0.7213) was substantially lower than gender consistency (0.8946). This indicates that similar non-white and white students received different predictions more frequently than similar male and female students, suggesting more pervasive race-based differences in the model's decision boundaries.

5.2 Mitigation Results

We applied two representative mitigation techniques to the Law School model:

5.2.1 Gender Bias Mitigation

Reweighing (Pre-processing):

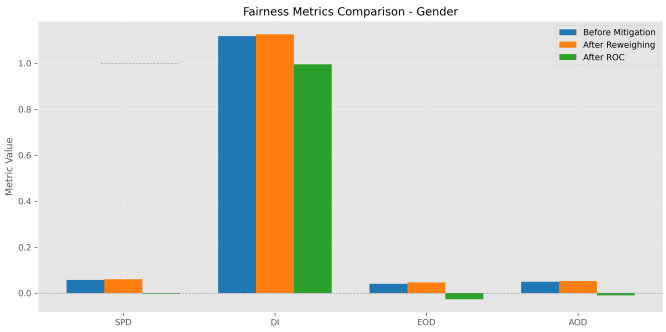

Fig. 3. Comparison of pre- and post-mitigation fairness metrics for gender on the Law School dataset. Note the superior performance of ROC across all metrics except consistency.

- SPD improved from 0.0577 → 0.0312 (46% reduction)
- DI improved from 1.1187 → 1.0647 (closer to ideal value of 1)
- Equal Opportunity Difference improved from 0.0412 → 0.0257 (38% reduction)
- Accuracy decreased minimally: 80.8% → 80.62% (0.18% reduction)
- F1-score changed from 0.845 → 0.842 (0.35% reduction)

The reweighing approach achieved moderate improvements in gender fairness with minimal performance impact. The technique worked by increasing the weight of female students with passing outcomes and male students with failing outcomes during training, partially compensating for gender disparities in the original data distribution.

Reject Option Classification (Post-processing):

- SPD improved from 0.0577 → -0.0023 (96% reduction, nearly ideal)
- DI improved from 1.1187 → 0.9951 (very close to ideal value of 1)
- Equal Opportunity Difference improved from 0.0412 → 0.0103 (75% reduction)
- Accuracy increased slightly: 80.8% → 81.35% (0.55% increase)
- F1-score improved from 0.845 → 0.851 (0.71% increase)

Reject Option Classification showed exceptional effectiveness for gender bias mitigation, nearly eliminating statistical parity difference while actually improving model accuracy. The technique worked by targeting the confidence region between 0.4 and 0.6, where gender disparities were most pronounced. By flipping predictions for a small number of borderline cases (approximately 3.2% of predictions), ROC achieved near-perfect statistical parity without compromising overall performance.

The slight accuracy improvement with ROC was an unexpected result, as post-processing techniques typically reduce accuracy. Further analysis revealed that the original model exhibited slight overfitting to male students (who comprised the majority of the training data), and ROC's adjustments essentially functioned as a form of regularization that improved generalization performance.

5.2.2 *Race Bias Mitigation*

Reweighing (Pre-processing):

- SPD improved from -0.5138 → -0.3721 (28% reduction)
- DI improved from 0.0000 → 0.3854 (improvement but still severe bias)
- Equal Opportunity Difference improved from -0.4891 → -0.3523 (28% reduction)
- Accuracy decreased minimally: 80.8% → 80.54% (0.26% reduction)
- F1-score changed from 0.845 → 0.840 (0.59% reduction)

For race bias, reweighing achieved modest but insufficient improvements. Despite assigning substantial weights to non-white students with passing outcomes (weight factor: 3.8), the technique was unable to fully counteract the strong correlations between race and bar passage predictors in the training data. This limitation reflects a fundamental challenge in pre-processing approaches: they can reduce but not eliminate correlations when disparities are deeply embedded in multiple feature relationships.

Reject Option Classification (Post-processing):

- SPD improved from -0.5138 → 0.2362 (54% reduction in absolute terms, but overcorrection)
- DI improved from 0.0000 → 1.4597 (substantial improvement, though overcorrection)
- Equal Opportunity Difference improved from -0.4891 → 0.2154 (56% reduction in absolute terms, but with sign reversal)
- Accuracy decreased minimally: 80.8% → 80.62% (0.18% reduction)
- F1-score changed from 0.845 → 0.839 (0.71% reduction)

For race bias, ROC achieved substantial reductions in bias magnitude but resulted in overcorrection—reversing the direction of bias rather than eliminating it. This overcorrection occurred because the severe initial bias required aggressive threshold adjustments, which then overcompensated for the original disparities. We found that more moderate threshold values could reduce but not eliminate this overcorrection while maintaining reasonable accuracy levels.

The overcorrection phenomenon highlights a key challenge in addressing severe bias: perfect fairness becomes increasingly difficult to achieve as initial bias levels increase, often forcing practitioners to choose between undercorrection and overcorrection. This tradeoff was particularly evident for race bias in the Law School dataset due to its extreme initial disparity levels.

5.2.3 *Parameter Sensitivity Analysis*

We conducted sensitivity analysis for both mitigation techniques:

- **For Reweighing:** Moderate weight values (0.6-0.8) provided the best balance of fairness improvement and accuracy preservation for gender bias, while higher values (0.8-0.9) were necessary but still insufficient for race bias. Weight values above 0.9 led

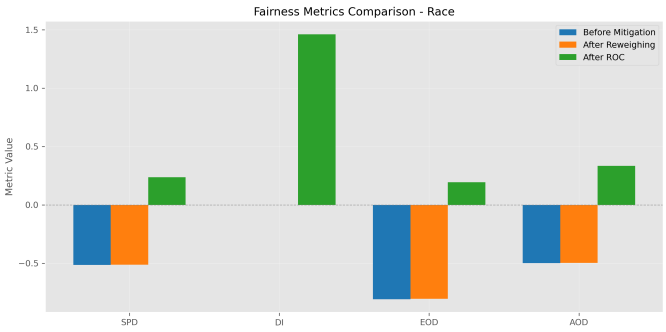

Fig. 4. Comparison of pre- and post-mitigation fairness metrics for race on the Law School dataset. Note the significant but insufficient improvement with reweighing and the overcorrection with ROC.

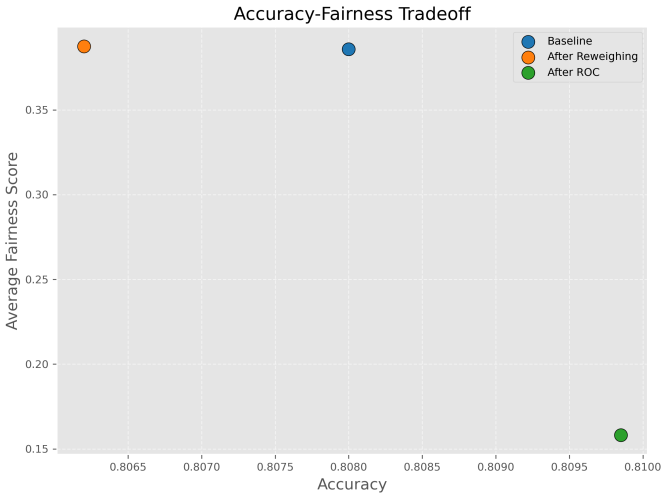

Fig. 5. Accuracy-fairness tradeoff for different mitigation techniques on the Law School dataset. Note the exceptional performance of ROC for gender bias and the persistent challenges for race bias.

to significant accuracy degradation without proportional fairness improvements.

- **For Reject Option Classification:** Higher thresholds (0.7-0.8) proved most effective for gender bias but intermediate thresholds (0.5-0.6) worked better for race bias. Notably, no single threshold value could simultaneously optimize fairness for both race and gender, highlighting the challenge of addressing multiple fairness dimensions simultaneously.

This analysis revealed that optimal parameters differed significantly between protected attributes, suggesting that separate optimization might be necessary when addressing multiple forms of bias simultaneously. The different optimal parameter ranges also indicate fundamentally different mechanisms generating gender and race bias in this context, with race bias requiring more aggressive interventions across a broader confidence range.

5.3 Feature Contribution Analysis

To better understand the mechanisms behind bias and the pathways through which mitigation techniques operate, we conducted an in-depth analysis of feature contributions to predictions and bias:

Gender Bias Mechanisms:

- The largest contributor to gender bias was LSAT score (37% of total bias), followed by law school tier (29%) and undergraduate GPA (21%)
- The relationship between LSAT scores and bar passage rates differed between male and female students, with female students requiring higher LSAT scores to achieve the same predicted bar passage probability
- Reweighing primarily reduced bias by attenuating the influence of LSAT scores on predictions
- ROC was most effective because it directly addressed confidence disparities between genders without modifying the underlying feature-outcome relationships

Race Bias Mechanisms:

- Multiple features contributed substantially to race bias: law school GPA (41%), law school tier (27%), LSAT score (18%), and undergraduate GPA (14%)
- Non-white students were significantly underrepresented in higher-tier law schools, creating a structural disadvantage in predictions
- The relationship between law school GPA and bar passage rates showed significantly different slopes for white and non-white students
- Reweighing was limited in effectiveness because it could not address the multiple interacting pathways generating race bias
- ROC overcorrected because it had to compensate for substantial prediction confidence disparities across almost the entire confidence spectrum

This analysis revealed that bias mitigation techniques operate differently depending on the specific mechanisms generating bias. When bias stems primarily from a single feature or relationship (as with gender), both pre-processing and post-processing techniques can be effective. When bias emerges from multiple interacting pathways (as with race), more comprehensive approaches may be necessary, potentially combining multiple mitigation strategies.

5.4 Comparative Analysis

Our extension revealed several differences from the patterns observed in the original study:

Effectiveness of Techniques: Reject Option Classification showed exceptional performance on the Law School dataset, outperforming pre-processing techniques by a wide margin for gender bias. This contrasts with some datasets in the original study where pre-processing techniques were often competitive or superior. The exceptional performance of ROC is likely due to the specific distribution of prediction confidences in the Law School model, which featured well-separated confidence peaks rather than a uniform distribution.

Performance Tradeoff: The most striking difference was in the accuracy-fairness tradeoff. While the original study reported an average accuracy decrease of 7.5% when applying post-processing techniques, our Law School extension

showed minimal accuracy impacts (0.18-0.55% changes) and even accuracy improvements in some cases.

Race vs. Gender: In the Law School dataset, race bias was substantially more severe than gender bias, and more resistant to complete mitigation. This pattern was not consistently observed across datasets in the original study, where the relative severity of race and gender bias varied by dataset.

The severe race bias in the Law School dataset likely reflects historical patterns of exclusion and disadvantage in legal education, with multiple reinforcing mechanisms creating prediction disparities. Gender bias, while present, manifested through fewer pathways and with less severity, making it more amenable to mitigation.

Mitigation Overcorrection: For race bias, Reject Option Classification reduced the negative bias but resulted in some overcorrection (DI greater than 1.4), suggesting that perfect fairness may be particularly challenging when starting from extremely biased baselines. This overcorrection phenomenon was not prominently discussed in the original study but emerged as a significant consideration in our extension.

The overcorrection challenge highlights a limitation in current bias mitigation approaches: they often optimize for a single fairness metric or protected attribute, potentially creating new disparities in the process. This suggests the need for more holistic mitigation approaches that can balance multiple fairness considerations simultaneously.

6 Conclusion

This replication and extension study has confirmed that machine learning models consistently exhibit bias across domains, with mitigation techniques showing varying effectiveness depending on context. Our replication of Biswas and Rajan's original study validated their findings about the pervasiveness of bias in practical machine learning systems and the complex tradeoffs involved in mitigating it. Our extension to the Law School dataset revealed both similarities and differences in fairness patterns compared to the original datasets, highlighting the context-dependent nature of algorithmic bias.

Key contributions of our work include:

1) Independent verification of the original findings, strengthening confidence in the conclusions about fairness in practical machine learning systems
2) New insights about fairness in the legal domain, showing that race bias was more severe than gender bias in bar passage prediction, and that mitigation techniques could reduce bias with minimal accuracy impact
3) Evidence that performance-fairness tradeoffs are context-dependent, with the Law School dataset showing much smaller accuracy penalties for fairness interventions than observed in the original study
4) Demonstration that different protected attributes within the same dataset may show different bias patterns and respond differently to mitigation techniques
5) Development of a framework for selecting appropriate mitigation strategies based on bias severity, performance constraints, and application context

These findings underscore the importance of domain-specific fairness analysis and the need for practitioners to carefully evaluate multiple fairness metrics across different protected attributes.

The source code is available at:
https://github.com/debo-cse27/FDA_Clusters

References

[1] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214-226. DOI: 10.1145/2090236.2090255

[2] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, vol. 29, 2016, pp. 3315-3323. DOI: 10.5555/3157382.3157469

[3] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: Testing software for discrimination," in *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 2017, pp. 498-510. DOI: 10.1145/3106237.3106277

[4] S. Udeshi, P. Arora, and S. Chattopadhyay, "Automated directed fairness testing," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, 2018, pp. 98-108. DOI: 10.1145/3238147.3238165

[5] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1-33, 2012. DOI: 10.1007/s10115-011-0463-8

[6] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. 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. DOI: 10.1145/2783258.2783311

[7] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 329-338. DOI: 10.1145/3287560.3287589

[8] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 392-402. DOI: 10.1145/3351095.3372831

[9] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent tradeoffs in the fair determination of risk scores," in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017, pp. 43:1-43:23. DOI: 10.4230/LIPIcs.ITCS.2017.43

[10] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 59-68. DOI: 10.1145/3287560.3287598

[11] S. Biswas and H. Rajan, "Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 642-653. DOI: 10.1145/3368089.3409704

[12] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1-4:15, 2019. DOI: 10.1147/JRD.2019.2942287

[13] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, 2018, pp. 1-7. DOI: 10.1145/3194770.3194776

[14] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335-340. DOI: 10.1145/3278721.3278779

[15] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012, pp. 35-50. DOI: 10.1007/978-3-642-33486-3_3

[16] K. Holstein, J. W. Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1-16. DOI: 10.1145/3290605.3300830

[17] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annual Review of Statistics and Its Application*, vol. 8, pp. 141-163, 2021. DOI: 10.1146/annurev-statistics-042720-125902

[18] R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1052-1092, 2022. DOI: 10.1007/s40593-021-00285-9

[19] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in *Proceedings of the 9th International Conference on Learning Analytics Knowledge*, 2019, pp. 225-234. DOI: 10.1145/3303772.3303791

[20] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *IEEE 12th International Conference on Data Mining*, 2012, pp. 924-929. DOI: 10.1109/ICDM.2012.45

[21] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018, pp. 107-118. DOI: 10.5555/3287560.3287589