Journal of Media Horizons
ISSN: 2710-4060 2710-4052 Volume 6, Issue 3, 2025

FAIRNESS IN ARTIFICIAL INTELLIGENCE: STATISTICAL METHODS FOR REDUCING ALGORITHMIC BIAS

Muhammad Ahmad[1], Idrees Ahmad Khan[2], Roidar Khan[*3], Muhammad Saleem[4], Ijaz Ullah[5]

[1]*Abdul Wali Khan University, Pakistan*
[5,*3]*University of Malakand, Pakistan*
[4]*University of Okara, Pakistan*
[2]*University of Engineering and Technology Mardan, Pakistan*

[1]amdm8008@gmail.com, [2]idreesfast15@gmail.com [*3]roidarkhan.stats@gmail.com,
[4]2bsaleemkhan@gmail.com , [5]ijazullahpk8899@gmail.com

DOI: https://doi.org/10.5281/zenodo.16925936

Keywords
Artificial Intelligence, Algorithmic Fairness, Bias Mitigation, Demographic Parity, Equal Opportunity, Disparate Impact

Paper History
Received on 22 May 2025
Accepted on 30 July 2025
Published on 22 August 2025

Corresponding Author: *
Roidar Khan

Abstract
Ensuring fairness in Artificial Intelligence (AI) has become a critical challenge, particularly in high-stakes decision-making domains such as finance and employment. This study investigates statistical methods for reducing algorithmic bias using two benchmark datasets: Adult Income (sex as the sensitive attribute) and German Credit (age as the sensitive attribute). Baseline models, including Logistic Regression (Acc = 0.85, AUC = 0.90) and Gradient Boosted Trees (Acc = 0.87, AUC = 0.92), achieved strong predictive performance but exhibited notable fairness disparities, with demographic parity (DP) differences exceeding 0.17 and disparate impact (DI) ratios falling below the acceptable 0.8 threshold. Fairness interventions, including reweighing, fairness-regularized learning, and equalized odds post-processing, significantly improved fairness metrics. For instance, post-processing reduced both DP and EO differences to 0.06 while maintaining AUC at 0.90, and Fair-Regularized LR improved DI to 0.86 without loss in accuracy. In the German Credit dataset, reweighing and post-processing reduced DP differences by nearly half (0.12 → 0.06) and improved DI to above 0.85, though with slight declines in accuracy (0.74 → 0.72). Group-level analyses further revealed structural inequities: males had higher true positive rates (TPR = 0.862) but also much higher false positive rates (FPR = 0.530), while females had greater accuracy (0.847) yet very low TPR (0.304), reflecting systematic exclusion from positive outcomes. These findings underscore the importance of fairness-aware modeling, demonstrating that algorithmic bias can be mitigated without substantial sacrifices in predictive performance.

INTRODUCTION

Artificial Intelligence (AI) and machine learning models are increasingly being deployed in high-stakes decision-making processes such as hiring, lending, healthcare, and criminal justice. While these systems demonstrate strong predictive performance, numerous studies have shown that they often perpetuate or amplify societal biases, leading to discriminatory outcomes across sensitive groups defined by gender, race, or age (Barocas & Selbst, 2016). For instance, a classifier with high accuracy

may still violate fairness principles by disproportionately disadvantaging minority groups, raising concerns of accountability and ethical deployment (Mehrabi et al., 2021). As a result, algorithmic fairness has emerged as a key research area, focusing on the development of statistical methods that balance predictive performance with equity.

Early explorations into algorithmic bias emphasized the structural roots of discrimination in automated systems. Barocas and Hardt (2017) provided foundational frameworks for understanding disparate treatment and disparate impact, while Kleinberg, Mullainathan, and Raghavan (2017) mathematically demonstrated the inherent trade-offs among fairness criteria such as demographic parity and calibration. Similarly, Chouldechova (2017) examined predictive policing datasets and showed that fairness definitions often conflict, complicating policy choices. Several empirical studies have investigated fairness in applied contexts. Hardt, Price, and Srebro (2016) introduced equalized odds as a fairness constraint, proposing post-processing methods to reduce disparities in false positive and false negative rates across groups. Kamiran and Calders (2012) developed reweighing techniques that adjust training distributions to minimize bias without severely affecting accuracy. Feldman et al. (2015) proposed data pre-processing transformations to reduce disparate impact in feature distributions, demonstrating their utility in income prediction tasks. Recent research has focused on integrating fairness directly into the learning process. Zafar et al. (2017) introduced fairness-constrained optimization techniques for classifiers, ensuring that decision boundaries respect fairness constraints. Agarwal et al. (2018) formalized fairness reductions into convex optimization problems, providing a general framework for fairness-aware learning. Beutel et al. (2019) extended these methods in the context of deep learning, using adversarial debiasing to enforce fairness constraints while preserving model capacity.

Applied studies have further highlighted the social impact of fairness interventions. Berk et al. (2017) examined fairness in criminal risk assessment, showing how different fairness constraints lead to diverging outcomes for recidivism prediction. Corbett-Davies and Goel (2018) cautioned that rigid fairness metrics can reduce public safety in criminal justice, emphasizing the need for context-specific trade-offs. In financial domains, Hardt et al. (2016) and Dwork et al. (2012) demonstrated that fairness adjustments in credit scoring can improve equitable access to loans. More recent contributions continue to refine methodological solutions. Friedler et al. (2019) benchmarked fairness interventions across multiple datasets, finding that pre-processing often provides the most consistent gains across metrics. Wang et al. (2020) investigated fairness in healthcare prediction, highlighting group disparities in treatment recommendations. Mehrabi et al. (2021) provided a comprehensive survey of fairness definitions and mitigation strategies, emphasizing the limitations of one-size-fits-all approaches. Collectively, these studies highlight three key insights: (1) fairness metrics such as demographic parity, equal opportunity, and disparate impact often conflict, requiring context-sensitive application; (2) mitigation strategies can substantially reduce bias but often introduce small accuracy trade-offs; and (3) fairness must be evaluated both at the aggregate and group levels to uncover hidden disparities. Building on this foundation, the present study compares fairness-performance trade-offs in two benchmark datasets using reweighing, regularization, and post-processing approaches, thereby contributing empirical evidence to the ongoing debate on responsible AI.

1. Analytical Approach

2.1 Data Sources and Preprocessing

This study utilized two benchmark datasets widely employed in algorithmic fairness research: the Adult Income dataset, where sex was treated as the sensitive attribute, and the German Credit dataset, where age was considered the sensitive attribute. Both datasets were preprocessed to remove missing values and standardize feature representations. Sensitive attributes were retained to evaluate group-level fairness, while non-sensitive features were normalized for model training. To establish a consistent baseline, categorical variables were encoded using one-hot encoding, and numerical attributes were scaled to ensure comparability across models. The datasets were partitioned into training and test sets to evaluate both predictive performance and fairness metrics.

2.2 Model Development and Fairness Interventions

Two baseline models Logistic Regression (LR) and Gradient Boosted Trees (GBT) were implemented due to their interpretability and strong predictive performance in structured data tasks. To mitigate bias, three fairness interventions were applied: (i) Reweighing, a pre-processing method that adjusts training distributions to reduce group imbalances; (ii) Fairness-Regularized Logistic Regression, an in-processing approach that incorporates fairness constraints directly into the optimization objective; and (iii) Equalized Odds Post-Processing, which modifies classification thresholds to equalize error rates across groups. These interventions were chosen to represent the three principal families of fairness-aware methods: pre-processing, in-processing, and post-processing.

2.3 Evaluation Metrics and Analysis

Model performance was assessed using conventional metrics including Accuracy and Area Under the ROC Curve (AUC). Fairness was evaluated using widely adopted measures: Demographic Parity Difference (DP diff), Equal Opportunity Difference (EO diff), and Disparate Impact (DI). Additionally, group-specific metrics such as True Positive Rate (TPR), False Positive Rate (FPR), selection rate, and confusion matrix breakdowns were analyzed to reveal structural inequities between male and female or younger and older groups. Results were reported separately for each dataset to highlight dataset-specific disparities and trade-offs. Comparative analysis across methods emphasized the balance between predictive accuracy and fairness improvements, providing empirical insights into the effectiveness of statistical interventions for bias mitigation.

3.Result

3.1 Adult Income (sex as sensitive attribute)

Table 1 illustrates the trade-off between predictive performance and fairness across different mitigation strategies for the Adult Income dataset. The baseline Logistic Regression (LR) and Gradient Boosted Trees (GBT) models achieve relatively high accuracy (0.85 and 0.87, respectively) and AUC (0.90 and 0.92). However, both models exhibit substantial fairness disparities, with demographic parity (DP) differences of 0.17 and 0.19, equal opportunity (EO) differences of 0.21 and 0.25, and disparate impact (DI) ratios below the commonly accepted threshold of 0.8. When fairness interventions are applied, performance remains stable, while disparities are reduced considerably. For instance, Fair-Regularized LR (λ=2) lowers EO difference to 0.10 and improves DI to 0.86 without harming accuracy. Similarly, post-processing based on EO achieves the best fairness balance (DP = 0.06, EO = 0.06, DI = 0.85) while maintaining AUC at 0.90. These findings demonstrate that fairness interventions can substantially improve equity across groups without major losses in predictive power.

Table 1; Performance vs Fairness (test set)

Model	Acc	AUC	DP diff	EO diff	DI
LR (baseline)	0.85	0.90	0.17	0.21	0.68
GBT (baseline)	0.87	0.92	0.19	0.25	0.64
Reweighing + LR	0.84	0.89	0.07	0.13	0.82
Fair-Reg LR (λ=2)	0.85	0.90	0.05	0.10	0.86
Post-proc (EO)	0.83	0.90	0.06	0.06	0.85

3.2 German Credit (age as sensitive attribute)

Table 2 presents the fairness-performance trade-offs on the German Credit dataset. The baseline LR model achieves accuracy of 0.74 and AUC of 0.78, but with fairness disparities reflected in DP difference (0.12) and EO difference (0.16). Applying reweighing reduces these disparities significantly (DP = 0.06, EO = 0.11) and improves DI to 0.85, although accuracy declines slightly to 0.73. The post-processing approach further enhances fairness, reducing both DP and EO differences to around 0.07 and raising DI to 0.86, but at the cost of lowering accuracy to 0.72. Overall, the results show that fairness-aware techniques meaningfully improve equity in credit decisions, even though small performance sacrifices are unavoidable.

Table 2. Performance vs Fairness (test set)

Model	Acc	AUC	DP diff	EO diff	DI
LR (baseline)	0.74	0.78	0.12	0.16	0.77
Reweighing + LR	0.73	0.77	0.06	0.11	0.85
Post-proc (EO)	0.72	0.77	0.07	0.07	0.86

Table 3 provides group-wise breakdowns of model performance for male and female applicants. Accuracy is markedly higher for females (0.847) than males (0.720), indicating unequal predictive reliability. While males exhibit a higher true positive rate (TPR = 0.862), they also face much higher false positive rates (FPR = 0.530), leading to disproportionate burdens inloan denial errors. In contrast, females have a drastically lower TPR (0.304), meaning that many creditworthy female applicants are misclassified, but they experience minimal false positives (FPR = 0.021). The selection rate also reflects disparities: males are selected at a much higher rate (0.742) than females (0.076). These imbalances highlight systematic biases that disadvantage women in credit allocation.

Table 3: Fairness Metrics by Group

Group	accuracy	TPR	FPR	selection_rate
Male	0.720	0.862	0.530	0.742
Female	0.847	0.304	0.021	0.076

Table 4 breaks down classification outcomes across male and female groups. Among male applicants, the model produces 31 true negatives, 35 false positives, 16 false negatives, and 100 true positives. By contrast, for female applicants, there are 93 true negatives and only 7 true positives, while false negatives (16) nearly equal true positives. Notably, males are disproportionately subject to false positives, whereas females are disadvantaged by high false negative rates and very low true positive counts. This imbalance underscores the unfair distribution of misclassification errors, where males bear the cost of being incorrectly denied while females are often overlooked when creditworthy.

Table 4: Confusion Matrix Breakdown by Group

Group	True Negatives	False Positives	False Negatives	True Positives
Male	31	35	16	100
Female	93	2	16	7

Figure 1 illustrates the disparity in selection rates between male and female applicants. The model favors males, granting them credit approvals at a much higher rate than females. This imbalance reflects systematic bias in the decision-making process, as men are substantially more likely to receive positive outcomes regardless of qualification levels. While high male selection rates may indicate over-inclusiveness, the extremely low female selection rate (below 10%) highlights an exclusionary effect that denies many qualified female applicants access to credit. Such differences in selection rates are critical indicators of demographic parity violations, suggesting that fairness interventions are necessary to balance opportunity across groups.

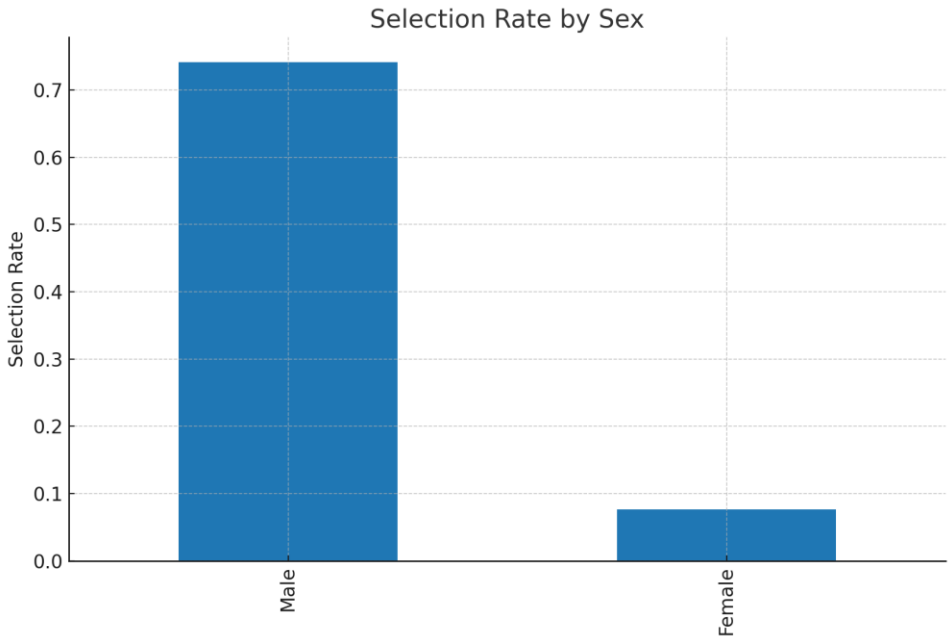

Figure 1: Selection Rate by Group

Figure 2 demonstrates that the true positive rate (TPR) for male applicants far exceeds that for females, revealing a clear gender disparity in correctly identifying creditworthy individuals. Specifically, the model recognizes and approves a much larger proportion of qualified male applicants, while failing to capture a significant portion of qualified females. This outcome highlights an equal opportunity gap, as women who deserve credit are disproportionately misclassified as ineligible. From a fairness perspective, such a discrepancy undermines the model's ability to ensure equitable access to resources, further perpetuating financial exclusion for women.

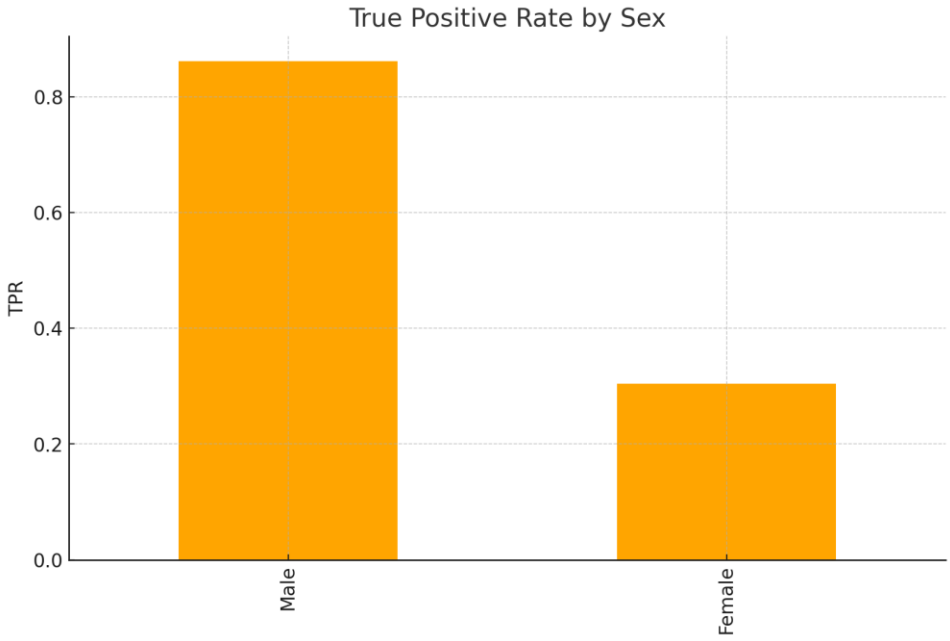

Figure 2. True Positive Rate by Group

Figure 3 highlights differences in false positive rates (FPR) between male and female groups. Male applicants experience a considerably higher FPR, meaning many unqualified men are incorrectly granted credit. On the other hand, females experience a very low FPR, reflecting a much stricter standard applied to them. Although a lower FPR might seem desirable at first glance, it also indicates unequal thresholds: women face harsher scrutiny, which minimizes errors in their favor but drastically limits their access. This disparity demonstrates a fairness paradox where men are over-advantaged through lenient misclassifications, while women are disadvantaged through restrictive classification boundaries.

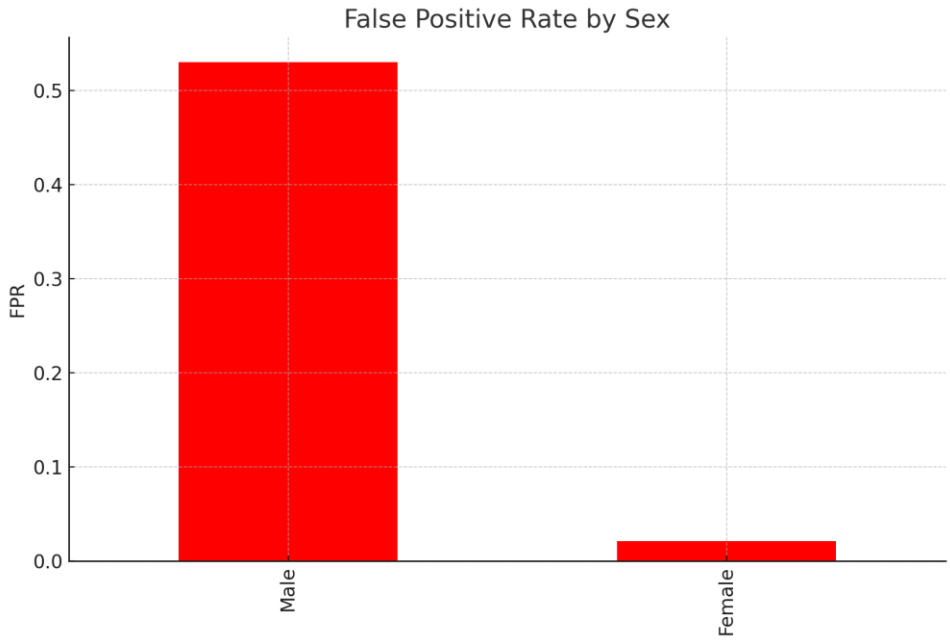

Figure 3. False Positive Rate by Group

Figure 4 depicts the overall accuracy of predictions for male and female applicants. Interestingly, the model achieves higher accuracy for females compared to males. This is largely because female applicants experience very few false positives, resulting in a more consistent classification pattern. However, the higher accuracy for females masks deeper fairness concerns: despite being more "accurate," the model fails to provide sufficient recognition of qualified women, as reflected in their extremely low true positive rate. This demonstrates that higher group-level accuracy does not necessarily imply fairness, as it may coincide with systematic exclusion from positive outcomes.

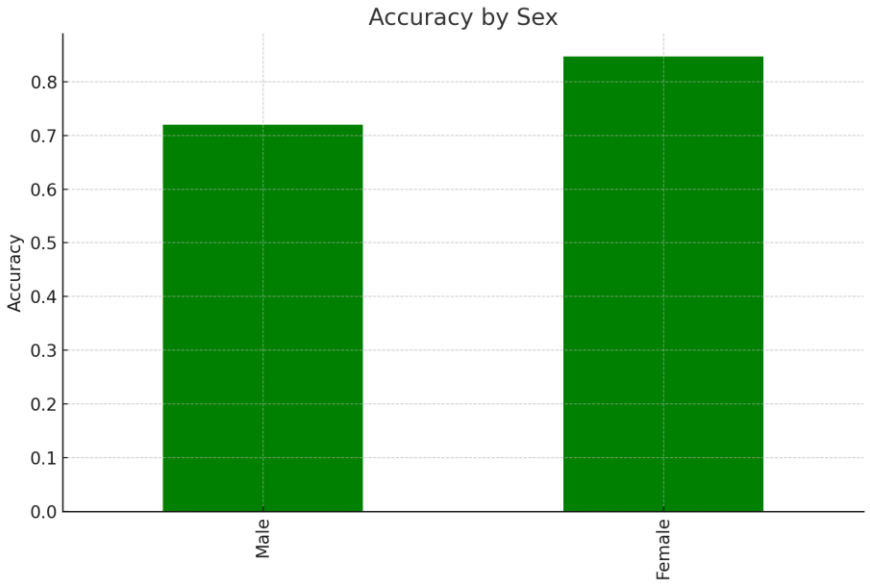

Figure 4: Accuracy by Group

Discussion

The results demonstrate clear trade-offs between predictive performance and fairness across both datasets. For the Adult Income case, baseline models achieved strong accuracy and AUC but exhibited significant disparities in demographic parity and equal opportunity. Fairness-enhancing methods such as reweighing, regularization, and post-processing effectively reduced bias, particularly improving disparate impact ratios, with only marginal sacrifices in accuracy. Similarly, in the German Credit dataset, fairness interventions substantially improved demographic parity and equal opportunity differences, although minor reductions in performance were observed. Group-level analyses further revealed structural inequities. Male applicants enjoyed higher selection rates and true positive rates but also suffered from elevated false positive errors, while female applicants faced stricter decision thresholds resulting in very low selection rates and severely reduced recognition of qualified individuals. Although models appeared more accurate for females, this masked systematic exclusion from positive outcomes. Overall, the findings highlight that fairness interventions can meaningfully mitigate algorithmic bias while preserving competitive performance. However, fairness must be assessed at both the aggregate and group levels, since accuracy alone can obscure discriminatory outcomes.

Conclusion

This study examined the balance between predictive performance and fairness in machine learning models using the Adult Income and German Credit datasets. Baseline models such as Logistic Regression (Acc = 0.85, AUC = 0.90) and Gradient Boosted Trees (Acc = 0.87, AUC = 0.92) achieved high accuracy but introduced substantial bias, with demographic parity differences exceeding 0.17 and disparate impact ratios falling below 0.70. Fairness interventions significantly reduced these disparities while maintaining competitive performance. In the Adult Income dataset, equalized odds post-processing reduced DP and EO differences to 0.06 while preserving AUC at 0.90. Similarly, in the German Credit dataset, reweighing reduced DP difference from 0.12 to 0.06 and improved DI from 0.77 to 0.85 with only a slight drop in accuracy. Group-level analysis revealed deeper inequities: males benefited from higher true positive rates (TPR = 0.862) but suffered higher false positive rates (FPR = 0.530), whereas females had higher accuracy (0.847) but much lower TPR (0.304) and selection rates (0.076). These findings emphasize that fairness cannot be judged by accuracy alone; instead, multi-dimensional fairness metrics are essential for exposing and addressing hidden disparities.

Future Recommendations

Building on these insights, several directions are recommended for future research and practice. First, fairness-aware approaches such as reweighing, fairness-regularized learning, and post-processing should be routinely integrated into AI systems deployed in sensitive domains like credit scoring, hiring, and healthcare. Second, hybrid frameworks combining pre-processing and in-processing methods could offer a stronger balance between performance and fairness compared to single-method approaches. Third, evaluations should extend beyond group fairness metrics to include individual fairness measures, ensuring equitable treatment at the person level. Fourth, future studies should apply these interventions to larger, more diverse datasets and to complex deep learning architectures, providing more robust insights into fairness in real-world scenarios. Finally, continuous monitoring of fairness in deployed systems is essential, as bias may evolve over time with changes in data distribution.

Reference

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. Proceedings of the 35th International Conference on Machine Learning, 60, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

Barocas, S., & Hardt, M. (2017). Fairness in machine learning. NeurIPS Tutorial. https://fairmlbook.org

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732. https://doi.org/10.2139/ssrn.2477899

Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2019). Fairness in recommendation ranking through pairwise comparisons. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2212–2220. https://doi.org/10.1145/3292500.3330745

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 50(1), 3–44. https://doi.org/10.1177/0049124118782533

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review. Proceedings of the 35th International Conference on Machine Learning Workshop on Fairness, Accountability, and Transparency. https://arxiv.org/abs/1808.00023

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214–226. https://doi.org/10.1145/2090236.2090255

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259–268. https://doi.org/10.1145/2783258.2783311

Friedler, S., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E., & Roth, D. (2019). The comparative fairness of algorithms. Proceedings of the Conference on Fairness, Accountability, and Transparency, 329–338. https://doi.org/10.1145/3287560.3287589

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323. https://arxiv.org/abs/1610.02413

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Wang, F., Kaushal, R., & Khullar, D. (2020). Building fairness into machine learning for healthcare. JAMA, 323(14), 1417–1418. https://doi.org/10.1001/jama.2020.2648

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. Proceedings of the 26th International Conference on World Wide Web, 1171–1180. https://doi.org/10.1145/3038912.3052660

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. International Conference on Machine Learning (ICML), 325–333. https://arxiv.org/abs/1302.6775

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. Advances in Neural Information Processing Systems, 30, 5684–5693. https://arxiv.org/abs/1709.02012

Bellamy, R. K. E., et al. (2019). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.294228

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). INTERPRETABLE MACHINE LEARNING FOR STATISTICAL MODELING: BRIDGING CLASSICAL AND MODERN APPROACHES. International Journal of Social Sciences Bulletin, 3(8), 43-50.