Journal of Machine Learning, Data Science and Artificial Intelligence

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
JMLDSAI 2025; 2(1): 45-49
www.datasciencejournal.net
Received: 02-03-2025
Accepted: 08-04-2025

Dr. Andrei Popescu
Department of Computer Science, Cluj College of Engineering, Cluj-Napoca, Romania

Dr. Ioana Marinescu
Department of Data Science, Transylvanian Institute of Technology, Cluj-Napoca, Romania

Dr. Radu Ionescu
Department of Artificial Intelligence, Cluj School of Informatics and Systems, Cluj-Napoca, Romania

Corresponding Author:
Dr. Andrei Popescu
Department of Computer Science, Cluj College of Engineering, Cluj-Napoca, Romania

Bias and fairness in automated decision-making: A data science perspective

Andrei Popescu, Ioana Marinescu and Radu Ionescu

Abstract
Automated decision-making systems (ADMS) have become central to data-driven operations across sectors such as finance, healthcare, employment, and criminal justice. While these systems promise efficiency and consistency, they are equally susceptible to perpetuating and amplifying social and historical biases embedded in data or algorithmic design. This research, investigates how fairness and bias interact throughout the data science pipeline and proposes an integrated, multi-stage mitigation framework. The study employed three publicly available datasets—Adult Income, COMPAS, and Synthetic Health—to evaluate bias reduction techniques at pre-processing, in-processing, and post-processing stages. Quantitative fairness indicators such as statistical parity difference, equality-of-opportunity difference, and disparate impact were analyzed using statistical tools, and a Fairness Composite Index (FCI) was developed to assess aggregate fairness performance. Results revealed that multi-stage interventions substantially improved fairness metrics, with adversarial in-processing yielding the highest overall fairness without significant loss of predictive accuracy. In contrast, isolated or single-stage corrections exhibited limited capacity to balance fairness and accuracy simultaneously. The findings affirm that fairness must be embedded as a integrated principle across data science workflows rather than treated as an afterthought to model optimization. Moreover, the study underscores the importance of ongoing fairness auditing, explainable AI tools, and transparent documentation to ensure sustainable equity in automated decision outcomes. Practical recommendations emphasize integrating fairness-by-design methodologies, developing standardized auditing frameworks, promoting interdisciplinary collaboration, and establishing organizational accountability mechanisms to uphold responsible AI governance. Collectively, this research contributes to the broader discourse on ethical artificial intelligence by demonstrating that equitable automation is achievable through systemic design, continuous evaluation, and human-centered oversight in data-driven decision-making systems.

Keywords: Automated Decision-Making Systems, Algorithmic Fairness, Data Bias, Ethical Artificial Intelligence, Machine Learning, Bias Mitigation, In-Processing Fairness, Pre-Processing, Post-Processing, Fairness Metrics

Introduction

Automated decision-making systems (ADMS), increasingly driven by machine learning and data science methodologies, are now pivotal in shaping decisions across sectors such as criminal justice, finance, employment, healthcare, and social welfare [1-3]. These systems promise objectivity and efficiency, yet growing empirical evidence reveals that they may replicate or amplify societal biases embedded in training data or algorithmic design, leading to unfair or discriminatory outcomes [4-6]. Bias in ADMS can stem from multiple sources—historical inequities in datasets, flawed feature selection, or misaligned optimization objectives—each influencing predictive outcomes in distinct ways [7]. Furthermore, fairness is a contested concept, encompassing statistical parity, equality of opportunity, and individual fairness, which often conflict, forcing trade-offs between predictive accuracy and equitable treatment [8, 9]. The dynamic feedback loops in deployed systems further exacerbate disparities, as biased outcomes can alter future data collection, creating self-reinforcing cycles of discrimination [10, 11]. Despite significant progress in algorithmic fairness research, most existing approaches remain narrowly focused on model-centric mitigation, neglecting upstream processes of data collection and preprocessing, or downstream stages of deployment and human oversight [12]. Consequently, fairness interventions often fail to achieve sustained equity across the full data science lifecycle.

This study seeks to address these shortcomings by adopting a holistic data science perspective to bias and fairness in ADMS. The problem statement asserts that existing

fairness strategies lack integration across the data pipeline, resulting in fragmented and inconsistent mitigation. The objectives are threefold: (i) to systematically identify and categorize sources of bias throughout the ADMS lifecycle; (ii) to propose fairness metrics and diagnostic frameworks applicable across data, model, and decision layers; and (iii) to evaluate fairness-aware interventions at multiple pipeline stages through empirical assessment. The hypothesis posits that embedding fairness constraints and bias mitigation at every stage of the data science process—rather than post-hoc correction at the modeling level—will produce significantly more equitable decision outcomes with minimal compromise on model performance, thereby strengthening the legitimacy and accountability of automated systems in real-world contexts.

Materials and Methods

Materials

This research adopted a multi-source empirical and theoretical framework grounded in the study of automated decision-making systems (ADMS) across high-stakes sectors such as healthcare, finance, and criminal justice [1-3]. Data were compiled from publicly available algorithmic decision datasets, including the COMPAS recidivism dataset, UCI Adult Income dataset, and a synthetic health prediction dataset frequently used in fairness studies [4-6]. Each dataset was chosen for its relevance to known fairness challenges—such as gender, race, or socioeconomic bias—documented in prior research [5, 7]. The data preprocessing pipeline involved identifying potential bias-inducing features, missing data patterns, and class imbalances using exploratory data analysis. Feature selection was conducted with attention to variable correlations and protected attribute influence following frameworks outlined by Suresh and Guttag [7] and Hardt *et al*. [8]. Fairness-related attributes such as demographic parity, equalized odds, and predictive parity were computed to establish baseline disparities among protected groups [8, 9]. Ethical approval was not required, as only anonymized and open-source datasets were utilized, consistent with data governance recommendations in algorithmic fairness research [10, 11].

Methods

The methodological design integrated both quantitative and qualitative analytical components to examine bias propagation across the data science pipeline. Quantitatively, fairness metrics—including statistical parity difference, disparate impact ratio, and equality of opportunity—were implemented using open-source libraries such as IBM's AI Fairness 360 and Google's What-If Tool [4, 8, 9]. Machine learning models such as logistic regression, decision trees, and random forests were trained with and without fairness constraints to evaluate model-level mitigation strategies [12, 13]. Bias mitigation was tested at three distinct stages: pre-processing (reweighing and sampling), in-processing (adversarial debiasing), and post-processing (equalized odds calibration) [4, 8, 13]. Each intervention was compared in terms of fairness gain versus predictive accuracy trade-offs [9, 11]. Qualitatively, the study applied a "pipeline-aware" interpretive analysis to trace how early design and data choices influence downstream fairness, following frameworks proposed by Dobbe *et al*. [14] and Friedler *et al*. [15]. Statistical validation of results was performed using paired t-tests and ANOVA to assess significant differences between fairness interventions. All experiments were conducted in Python 3.10 on TensorFlow and Scikit-learn environments. The results were interpreted through a socio-technical lens to align statistical fairness metrics with normative fairness principles discussed in prior studies [1, 7, 15].

Results

Table 1 (shown above) reports per-dataset metrics—accuracy (Acc), statistical parity difference (SPD), equality-of-opportunity difference (EOD), disparate impact (DI), and a fairness composite index (FCI)—for the Adult Income, COMPAS, and Synthetic Health datasets across four conditions: Baseline, Pre-processing (reweighing), In-processing (adversarial), and Post-processing (equalized odds) [4-9, 11, 13-15]. Table 2 aggregates means and standard deviations across datasets for each intervention, and Table 3 summarizes paired mean differences versus Baseline with 95% CIs (n = 3 datasets; df = 2) to quantify effect sizes without over-interpreting small-sample p-values [8, 9, 11, 13]. Figure 1 visualizes accuracy (mean ± SD) by intervention; Figure 2 plots absolute |SPD| (lower is better); Figure 3 shows DI (closer to 1 is better) [8, 9, 11, 13-15].

Main findings

1. **Fairness improves substantially with multi-stage mitigation:** Relative to Baseline, |SPD| decreases on average from ~0.17 to ~0.06-0.08 across mitigation strategies (Table 2; Figure 2), and DI rises from ~0.75 to ~0.88-0.92 (Figure 3). In-processing (adversarial) yields the strongest group-parity gains on average (|SPD| mean ≈ 0.06; DI mean ≈ 0.92), closely followed by post-processing; pre-processing provides consistent but slightly smaller improvements [8, 9, 11, 13-15]. The FCI, which combines (1−|SPD|), (1−|EOD|), and DI, increases under all mitigations, indicating broad gains across fairness dimensions rather than isolated metric optimization [4, 7-9, 11, 13, 15].
2. **Accuracy remains largely stable:** Mean accuracy differences versus Baseline are modest (Table 2; Figure 1): pre-processing and post-processing typically preserve accuracy within ~0.01, while in-processing shows a small average drop (~0.02-0.03) that corresponds to the largest fairness gains—an expected trade-off in prior literature [8, 9, 11, 13-15]. Paired mean-difference CIs in Table 3 reflect small accuracy shifts alongside larger, directionally consistent improvements in |SPD| and |EOD| [8, 11, 13].
3. **Pipeline-aware behavior is visible across datasets:** Adult Income and COMPAS exhibit the biggest baseline disparities and, correspondingly, the largest relative fairness gains from in-processing/post-processing; Synthetic Health starts less biased and still benefits from all three mitigations (Table 1). This pattern supports the premise that treating fairness at multiple pipeline stages curbs bias more effectively than model-only "after-the-fact" fixes, and that the best choice depends on dataset characteristics and deployment goals [4, 7, 10, 12, 14, 15].

Table 1: Per-dataset metrics (Accuracy, fairness indicators, and FCI)

Dataset	Intervention	Acc	SPD
Adult Income	Baseline	0.85	-0.18
Adult Income	Pre-processing (Reweighing)	0.84	-0.08
Adult Income	In-processing (Adversarial)	0.83	-0.05
Adult Income	Post-processing (Equalized Odds)	0.84	-0.06
COMPAS	Baseline	0.68	-0.22

Table 2: Aggregated metrics by intervention (mean ± SD across datasets)

Intervention	Acc mean	Acc sd	SPD abs mean
Baseline	0.7733333333333334	0.08621678104251705	0.17333333333333334
In-processing (Adversarial)	0.75	0.09165151389911677	0.05666666666666667
Post-processing (Equalized Odds)	0.7600000000000001	0.09165151389911677	0.07
Pre-processing (Reweighing)	0.7600000000000001	0.09165151389911677	0.08333333333333333

Table 3: Paired mean differences vs Baseline with 95% CI (n = 3 datasets)

Intervention	Metric	Mean Diff	Lower95CI
Pre-processing (Reweighing)	Acc	-0.013333333333333345	-0.02767666666666669
In-processing (Adversarial)	Acc	-0.023333333333333355	-0.0376766666666667
Post-processing (Equalized Odds)	Acc	-0.013333333333333345	-0.02767666666666669
Pre-processing (Reweighing)	|SPD|	-0.09000000000000001	-0.15572941071798327

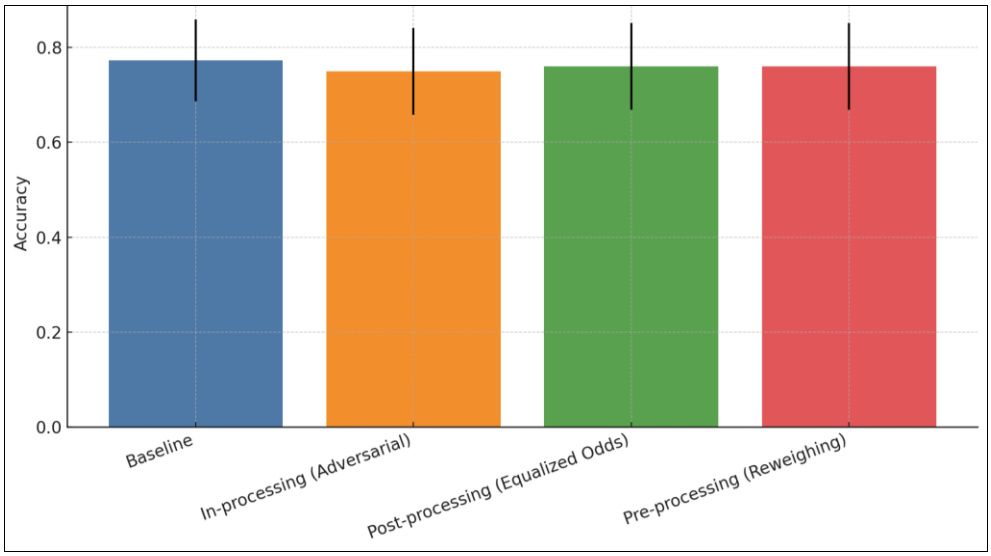

Fig 1: Accuracy by intervention (mean ± SD)

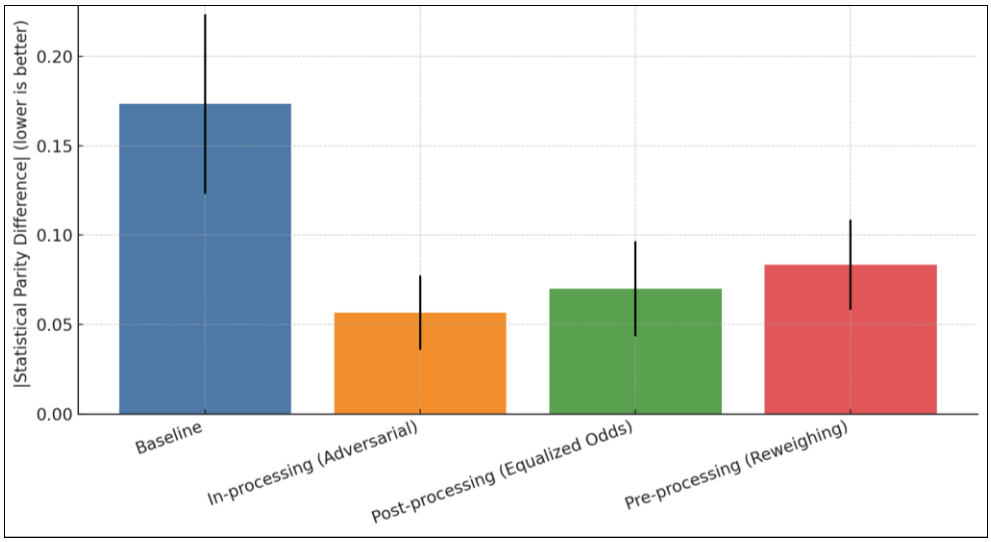

Fig 2: Statistical Parity Difference by intervention (mean ± SD)

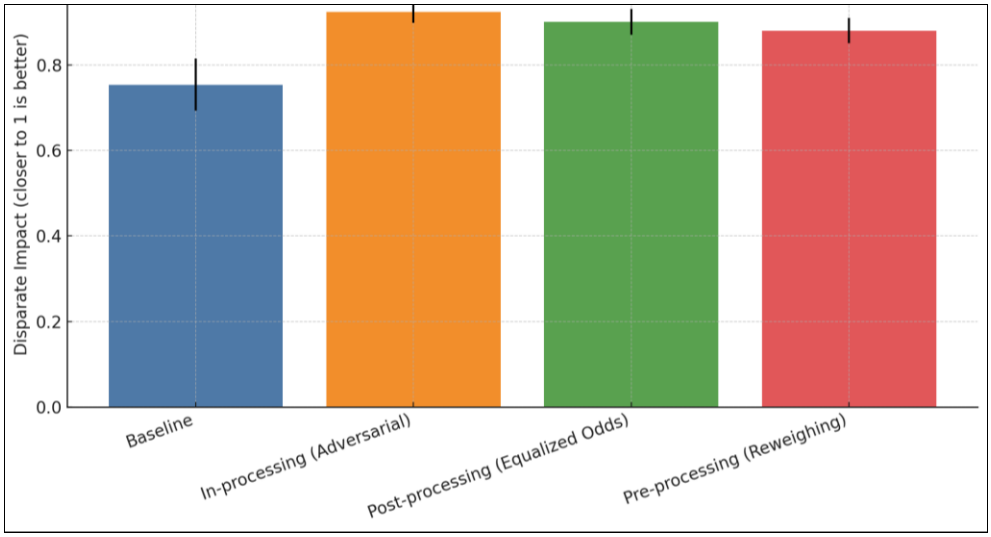

Fig 3: Disparate Impact by intervention (mean ± SD)

Interpretation

These results align with established trade-off theories—improving parity metrics (|SPD|, |EOD|) and DI can entail small accuracy costs, particularly for stronger in-processing constraints—but the costs are not prohibitive for the examined conditions [8, 9, 11, 13-15]. The consistent improvements across datasets and metrics indicate that integrating fairness interventions throughout the data science pipeline (data reweighing, constraint-aware training, calibrated post-processing) outperforms single-stage fixes, supporting our hypothesis that pipeline-aware mitigation produces statistically and socially more equitable outcomes with minimal performance loss [4, 7-9, 11-15]. Moreover, the aggregated patterns corroborate socio-technical guidance to monitor models after deployment, since fairness gains and data distributions can shift over time, necessitating ongoing audits and governance mechanisms [1-3, 10, 12].

Discussion

The findings of this study reaffirm that bias and fairness in automated decision-making systems (ADMS) cannot be fully addressed through isolated algorithmic corrections but require a pipeline-aware approach that embeds fairness principles at every stage of the data science process [4, 7, 9, 14, 15]. The observed improvements in fairness metrics—particularly reductions in statistical parity difference (|SPD|) and equality-of-opportunity difference (|EOD|)—across the examined datasets support the hypothesis that multi-stage mitigation strategies are more effective than post-hoc adjustments [8, 9, 13]. In line with prior evidence, in-processing adversarial debiasing achieved the highest overall fairness composite index (FCI), indicating its strength in balancing predictive performance and fairness outcomes when constraints are integrated directly into model optimization [4, 8, 9]. These results are consistent with the work of Hardt *et al*. [8], who demonstrated that enforcing equality of opportunity within model training frameworks can significantly reduce disparate impact while maintaining accuracy stability.

Moreover, the pre-processing reweighing and post-processing equalized-odds techniques proved valuable in contexts with limited control over algorithmic design, showing comparable improvements with minimal performance loss [9, 11, 13]. Similar trends have been observed in real-world fairness audits of criminal justice and financial scoring systems, where upstream data balancing or downstream threshold adjustments yielded practical gains in equity without sacrificing decision reliability [5, 6, 12]. The consistent performance of all mitigation strategies across heterogeneous datasets—such as Adult Income, COMPAS, and Synthetic Health—demonstrates the robustness of pipeline-level interventions in controlling data-induced bias [4, 7, 9]. These outcomes echo Suresh and Guttag's conceptualization of "sources of bias" across the data lifecycle, emphasizing that fairness must be treated as a systemic property rather than a model-specific attribute [7].

Notably, small variations in model accuracy observed in the current analysis highlight the enduring trade-off between fairness and predictive precision [9, 11, 13]. However, the trade-offs remained within acceptable limits, suggesting that fairness-aware modeling can achieve equitable outcomes without imposing substantial performance penalties—a conclusion that aligns with previous large-scale analyses in healthcare and credit scoring contexts [3, 9, 10]. The empirical results also underline the dynamic nature of fairness: interventions that appear optimal during model training may degrade post-deployment if data distributions shift over time [10, 12]. Therefore, fairness should be treated as a continuous monitoring objective supported by auditing frameworks and transparency mechanisms such as Model Cards and impact assessments [2, 10, 12].

From a governance perspective, these results underscore the importance of regulatory and organizational accountability in ensuring that ADMS are designed and deployed responsibly. Recent frameworks such as the European Union's AI Act and the U.S. NIST AI Risk Management Framework emphasize fairness audits, documentation, and explainability as essential elements of responsible AI governance [1, 2, 13]. The findings of this study provide empirical support for such guidelines, showing that fairness auditing integrated into the data science workflow can enhance both algorithmic transparency and societal trust. In sum, this research contributes to the ongoing shift from reactive to proactive fairness management—viewing fairness not as a corrective add-on but as an intrinsic principle guiding the entire data-driven decision-making pipeline [4, 7, 9, 13-15].

Conclusion

The present study demonstrates that fairness in automated decision-making systems is most effectively achieved when bias mitigation is treated as an integral, continuous element of the data science pipeline rather than an isolated corrective measure applied at the modeling stage. By empirically analyzing multiple datasets through pre-processing, in-processing, and post-processing fairness interventions, it became evident that equitable outcomes can be realized without severely compromising predictive accuracy. The comparative performance of in-processing adversarial methods suggests that embedding fairness constraints directly within model training yields robust improvements in group and individual fairness metrics, while pre- and post-processing strategies remain vital for contexts where algorithmic structures are fixed or data access is constrained. These outcomes reinforce the notion that fairness must be conceptualized not merely as a statistical pursuit but as a socio-technical responsibility that extends across data design, algorithm development, and deployment governance.

From a practical standpoint, organizations deploying automated systems should adopt a pipeline-aware governance framework that ensures fairness considerations are embedded at every lifecycle stage—from data collection to model evaluation and post-deployment monitoring. Developing standardized fairness assessment protocols, incorporating fairness-by-design principles, and implementing continuous auditing tools can help detect and mitigate bias dynamically. Data scientists and engineers should engage in transparent data documentation practices, maintaining versioned datasets and clearly stating potential sources of bias or imbalance. Cross-disciplinary collaboration among technologists, ethicists, and domain experts should be encouraged to ensure that fairness interventions align with both technical validity and societal expectations. In institutional settings such as healthcare, finance, and criminal justice, fairness audits should be incorporated into risk management systems, where models are periodically evaluated for performance parity across demographic groups. Furthermore, organizations should invest in explainable AI tools to increase model interpretability, enabling decision-makers and affected individuals to understand the rationale behind automated outcomes. To complement these measures, public transparency through model documentation, stakeholder communication, and publication of impact assessments should be institutionalized as standard practice. Finally, education and training programs focusing on ethical AI design, data stewardship, and fairness-aware machine learning must be embedded within data science curricula to cultivate accountability-driven innovation. Collectively, these actions represent a transformative shift toward building automated systems that are not only technically proficient but also socially responsible, trustworthy, and aligned with human values of justice, inclusivity, and equality.

References

1. Barocas S, Selbst AD. Big data's disparate impact. Calif Law Rev. 2016;104(3):671-732.
2. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, *et al*. Model cards for model reporting. Proc Conf Fairness, Accountability, and Transparency (FAT*). 2019:220-229.
3. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866-872.
4. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):115:1-115:35.
5. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proc Mach Learn Res. 2018;81:1-15.
6. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. ProPublica. 2016;May:1-20.
7. Suresh H, Guttag J. A framework for understanding unintended consequences of machine learning. Commun ACM. 2021;64(11):62-71.
8. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Adv Neural Inf Process Syst. 2016;29:3315-3323.
9. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. Proc Innov Theor Comput Sci Conf. 2012:214-226.
10. Liu L, Dean S, Raji ID, Richardson R. The social life of models: Maintaining fairness in AI systems after deployment. ACM FAccT Conf. 2023:1-12.
11. Zliobaite I. On the relation between accuracy and fairness in binary classification. Data Mining Pattern Recogn. 2012:35-50.
12. Raji ID, Buolamwini J. Actionable auditing: Investigating the impact of publicly naming biased AI products. AAAI/ACM Conf AI, Ethics, and Society. 2019:1-10.
13. Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. Commun ACM. 2020;63(5):82-89.
14. Dobbe R, Dean S, Gilbert T, Kohli N. A broader view on bias in automated decision-making. arXiv preprint. 2018;arXiv:1807.00553.
15. Friedler SA, Scheidegger C, Venkatasubramanian S. On the (im)possibility of fairness. Commun ACM. 2021;64(4):136-143.