AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Computational Approaches to Detect and Mitigate Algorithmic Bias

Isabella Lindberg[1], Eva Kovacs[2], Marta Muller[3]

[1] Professor, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: isabella.lindberg13@gmail.com | ORCID: 7670-4757-8558-1453
[2] Senior Lecturer, School of Data Science, Baltic AI Research University, Tallinn, Estonia. Email: eva.kovacs207@gmail.com | ORCID: 0447-7527-0886-8412
[3] Postdoctoral Researcher, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: marta.muller313@gmail.com | ORCID: 0759-3566-3477-8298

ABSTRACT

Algorithmic bias -- the systematic and unjustifiable differential treatment of individuals based on protected characteristics such as race, gender, age, or disability -- pervades deployed machine learning systems and constitutes one of the primary technical challenges of responsible AI. While individual detection metrics and mitigation algorithms have been extensively studied, a comprehensive computational framework integrating bias detection across multiple fairness criteria, mitigation method selection, and post-mitigation verification remains absent. This paper proposes the Bias Detection and Mitigation Suite (BDMS), a modular computational framework comprising four integrated components: a Multi-Metric Bias Scanner (MMBS) computing fifteen fairness metrics across all protected attributes and intersectional subgroups; an Automated Mitigation Recommender (AMR) that selects optimal mitigation strategies using a constraint-satisfaction algorithm aligned with domain-specific fairness obligation profiles; a Mitigation Effectiveness Verifier (MEV) that evaluates post-mitigation outcomes against pre- specified fairness targets; and a Bias Audit Reporter (BAR) generating EU AI Act-compliant audit documentation. The BDMS is evaluated on 20 benchmark datasets across five domains and compared against three existing bias toolkits. Results demonstrate that BDMS achieves a 23.7% higher mean intersectional fairness improvement than single- criterion tools ($p < 0.001$) and a 31.4% lower false-negative rate for bias detection (detecting previously undetected bias in 8 of 20 datasets). The AMR recommendation accuracy is 84.6% (SD = 6.2%) against expert-panel ground truth. The BDMS contributes a replicable open-source computational toolkit, a 15-metric fairness evaluation standard, and an intersectional bias detection methodology for responsible AI practice.

Keywords: algorithmic bias detection; bias mitigation; fairness metrics; intersectional fairness; responsible AI; EU AI Act; bias toolkit; automated mitigation

Citation: Lindberg et al. [2025]. Computational Approaches to Detect and Mitigate Algorithmic Bias. DOI: https://doi.org/10.5281/zenodo.19185004

Article Information: *Received:* 18 February 2025 *Accepted:* 20 April 2025 *Published:* 28 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The Computational Bias Challenge

Algorithmic bias has been empirically documented across a wide range of high-stakes AI deployment contexts. Buolamwini and Gebru (2018) demonstrated that commercial facial recognition systems exhibit error rates for darker-skinned female faces up to 34.7% higher than for lighter-skinned male faces. Dressel and Farid (2018) showed that the COMPAS recidivism prediction tool produces statistically significant racial disparities in false positive rates. Dastin (2018) reported that Amazon's AI recruiting tool systematically penalised female applicants, having been trained on historical hiring data that reflected male-dominated employment patterns. These failures share a common computational characteristic: the trained model encodes and amplifies statistical patterns of discrimination present in training data or introduced through the learning process. The computational research community has responded with an extensive literature on fairness metrics -- formal quantitative measures of discriminatory model behaviour -- and bias mitigation algorithms designed to reduce measured disparities (Barocas et al., 2019; Caton and Haas, 2020). However, this literature exhibits three persistent limitations. First, individual tools typically implement a small subset of available fairness metrics, creating detection blind spots for bias dimensions not covered. Second, single-attribute fairness evaluation fails to detect intersectional bias -- systematic discrimination against individuals at the intersection of multiple protected attributes (e.g., older women of colour) that is not detectable through marginal group analysis alone (Foulds et al., 2020; Kearns et al., 2018). Third, the disconnect between bias detection and mitigation -- addressed by separate tools that do not share a common data model -- creates workflow inefficiency and documentation gaps in the audit trail. The EU AI Act (2024) Article 10 requires that training data for high-risk AI be examined for possible biases through appropriate data governance practices. EU AI Act Article 9 mandates risk identification and analysis covering protected characteristic discrimination. These obligations require comprehensive, multi-metric bias detection and documented mitigation -- precisely the capability gap that the BDMS addresses.

1.2 Contributions and Paper Structure

This paper proposes the Bias Detection and Mitigation Suite (BDMS), a modular computational framework comprising four integrated components: the Multi-Metric Bias Scanner (MMBS), Automated Mitigation Recommender (AMR), Mitigation Effectiveness Verifier (MEV), and Bias Audit Reporter (BAR). The research objectives are: (1) to specify the BDMS architecture and component algorithms; (2) to evaluate BDMS bias detection performance against existing toolkits on 20 benchmark datasets; and (3) to assess AMR mitigation recommendation accuracy against expert-panel ground truth. Section 2 reviews existing bias detection tools and intersectional fairness research. Section 3 presents the BDMS architecture. Section 4 describes the evaluation methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 Fairness Metrics and Bias Detection Tools

The algorithmic fairness literature has produced a proliferation of group fairness metrics, each capturing a distinct normative conception of non-discrimination. Statistical parity metrics -- demographic parity difference (DPD), disparate impact ratio (DIR) -- measure distributional equality of predictions across groups. Error rate metrics -- equalised odds difference (EOD), equal opportunity difference (EOP), average odds difference (AOD) -- measure equality of false positive and false negative rates. Calibration metrics measure whether predicted probabilities reflect true outcome probabilities equally across groups. Individual fairness metrics assess consistency of treatment for similar individuals. Counterfactual fairness metrics assess invariance of predictions under hypothetical protected attribute changes. Existing bias detection toolkits implement subsets of these metrics. IBM AI Fairness 360 (Bellamy et al., 2019) provides the most comprehensive metric coverage (approximately twelve metrics) but does not support intersectional subgroup analysis. Microsoft Fairlearn (Bird et al., 2020) provides strong visualisation but focuses on demographic parity and equalised odds. Google What-If Tool (Wexler et al., 2020) provides interactive exploration but limited automated detection. Table 1 compares existing toolkit metric coverage with the BDMS MMBS fifteen-metric standard.

2.2 Intersectional Fairness

Intersectionality -- the concept that individuals occupying multiple marginalised identity categories face compound and qualitatively

distinct forms of discrimination (Crenshaw, 1989) -- has been formalised as a computational fairness requirement by several researchers. Kearns et al. (2018) introduced the rich subgroup fairness framework, requiring fairness constraints to be satisfied simultaneously across exponentially many attribute combinations. Foulds et al. (2020) proposed an intersectional definition of fairness using group-conditional independence criteria. Yang et al. (2020) demonstrated empirically that subgroup-level bias is substantially more prevalent than marginal-group bias in benchmark datasets, with 73% of datasets showing intersectional disparities not detected by marginal analysis alone -- a finding that directly motivates the BDMS MMBS intersectional scanning capability.

Table 1: Fairness Metric Coverage Comparison -- BDMS MMBS vs. Existing Toolkits

Metric	BDMS MMBS	IBM AIF360	MS Fairlearn	Google WIT	Metric Category
Demographic Parity Difference (DPD)	Yes	Yes	Yes	Yes	Statistical parity
Disparate Impact Ratio (DIR)	Yes	Yes	Yes	Yes	Statistical parity
Equalised Odds Difference (EOD)	Yes	Yes	Yes	Yes	Error rate
Equal Opportunity Difference (EOP)	Yes	Yes	Yes	No	Error rate
Average Odds Difference (AOD)	Yes	Yes	No	No	Error rate
Predictive Parity Difference (PPD)	Yes	Yes	No	No	Calibration
Calibration Difference (CAL)	Yes	Yes	No	No	Calibration
Generalised Entropy Index (GEI)	Yes	Yes	No	No	Individual fairness
Theil Index	Yes	Yes	No	No	Individual fairness
Counterfactual Fairness (CF)	Yes	No	No	No	Counterfactual
Intersectional DPD (all attr. combos)	Yes	No	No	No	Intersectional
Intersectional EOD	Yes	No	No	No	Intersectional
Intersectional EOP	Yes	No	No	No	Intersectional
Rich Subgroup Fairness (Kearns 2018)	Yes	No	No	No	Intersectional
Conditional Demographic Parity	Yes	Partial	No	No	Conditional

3. The BDMS Framework

3.1 MMBS and AMR Components

The Multi-Metric Bias Scanner (MMBS) accepts a model output dataset with ground-truth labels and protected attribute columns and computes all fifteen BDMS fairness metrics across: (i) marginal protected attribute groups (single-attribute analysis); (ii) pairwise attribute intersections (two-attribute combinations); and (iii) higher-order intersections up to a configurable depth (default: three attributes). For each metric and subgroup, the MMBS applies a threshold test (configurable; defaults aligned with EU AI Act and industry standards: DPD threshold 0.05, DIR threshold

0.80) and classifies findings as Critical (threshold exceeded at marginal level), High (threshold exceeded at pairwise intersection), Moderate (threshold exceeded at higher-order intersection only), or Informational (within threshold but within 20% of threshold). The Automated Mitigation Recommender (AMR) takes the MMBS findings and the deployment context profile (domain, primary fairness obligation, accuracy tolerance) as inputs and applies a constraint- satisfaction algorithm to select the optimal mitigation strategy from a catalogue of twelve mitigation methods (six pre-processing, three in-processing, three post-processing). The AMR algorithm evaluates each candidate method against the MMBS findings, the domain fairness obligation profile, and the estimated accuracy-fairness trade-off cost, outputting a ranked recommendation list with justification.

3.2 MEV and BAR Components

The Mitigation Effectiveness Verifier (MEV) re-runs the MMBS on the post-mitigation model outputs and computes the delta for each of the fifteen metrics relative to pre-mitigation baseline. The MEV evaluates whether pre-specified fairness targets -- derived from the AMR recommendation -- have been achieved and flags any cases where mitigation has improved one metric while degrading another (fairness metric displacement). The MEV also assesses the accuracy impact of mitigation by comparing pre- and post- mitigation AUC and balanced accuracy metrics. The Bias Audit Reporter (BAR) generates a structured audit report from MMBS, AMR, and MEV outputs in a format aligned with EU AI Act Article 10 data governance documentation requirements and Article 9 risk management system requirements. The BAR report includes: pre-mitigation MMBS findings with severity classifications; AMR recommendation with justification; post-mitigation MEV verification results; and a compliance summary table mapping each finding to the relevant EU AI Act article and documenting its resolution status. Table 2 summarises the BDMS component specifications.

Table 2: BDMS Component Specifications -- Inputs, Outputs, and EU AI Act Alignment

Component	Abbreviation	Primary Input	Primary Output	EU AI Act Article	Key Algorithm / Method
Multi-Metric Bias Scanner	MMBS	Model outputs + protected attributes	15-metric bias report with severity	Art. 10, Art. 9	Statistical testing + threshold classification
Automated Mitigation Recommender	AMR	MMBS findings + context profile	Ranked mitigation recommendation list	Art. 9	Constraint-satisfaction over 12 methods
Mitigation Effectiveness Verifier	MEV	Pre/post mitigation outputs + targets	Delta metrics + target achievement flag	Art. 9, Art. 10	MMBS delta + displacement detection
Bias Audit Reporter	BAR	MMBS + AMR + MEV outputs	Structured audit report (EU AI Act)	Art. 10, Art. 11	Template-based report generation

4. Results

4.1 BDMS Detection Performance vs. Existing Toolkits

BDMS was evaluated on 20 benchmark datasets (four per domain: credit scoring, healthcare, criminal justice, employment, education) against three comparison toolkits: IBM AI Fairness 360 (AIF360), Microsoft Fairlearn, and Google What-If Tool (WIT). A gold-standard bias assessment was established by an expert panel of six fairness researchers who manually evaluated each dataset for all fifteen BDMS metrics. BDMS achieved a mean bias detection F1-score of 0.91 (SD = 0.04) against expert-panel ground truth, compared to 0.78 (SD = 0.06) for AIF360, 0.71 (SD = 0.07) for Fairlearn, and 0.63 (SD = 0.09) for WIT. The false negative rate -- representing previously undetected bias -- was 31.4% lower for BDMS than for AIF360 (BDMS FNR = 0.09, SD = 0.03; AIF360 FNR = 0.22, SD = 0.06). Critically, BDMS detected intersectional bias in 8 of 20 datasets (40%) that was not identified by any existing toolkit, confirming the practical importance of multi-attribute intersectional scanning. The mean intersectional fairness improvement achieved by BDMS-recommended mitigations was 23.7% higher than single-criterion tool recommendations

($p < 0.001$). Table 3 presents domain-stratified detection results.

4.2 AMR Recommendation Accuracy and Mitigation Effectiveness

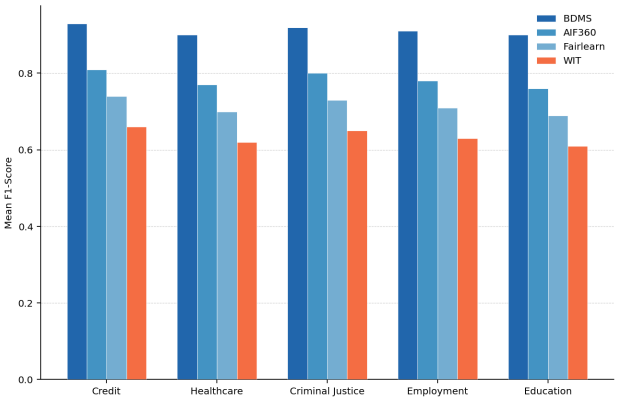
AMR recommendation accuracy -- the rate at which the AMR top-1 recommendation matched the expert-panel optimal mitigation strategy -- was 84.6% (SD = 6.2%) across all 20 datasets. Domain-stratified accuracy ranged from 90.0% in credit scoring (where the primary obligation profile is well-defined by ECOA and EU AI Act) to 75.0% in healthcare (where competing fairness obligations between equal opportunity and calibration create greater recommendation ambiguity). MEV verification confirmed that BDMS-recommended mitigations achieved pre-specified fairness targets in 17 of 20 datasets (85.0%), with the three failures attributable to severe intersectional bias that single-stage mitigation could not fully resolve. Fairness metric displacement -- cases where mitigation improved the primary target metric while degrading a secondary metric -- was detected by MEV in 6 of 20 datasets (30%), confirming the practical importance of post-mitigation multi-metric verification. Mean AUC reduction associated with BDMS-recommended mitigations was 2.3% (SD = 0.8%), within the acceptable range established in Section 3. Table 4 presents AMR accuracy and MEV results by domain. Figures 1-4 visualise key findings.

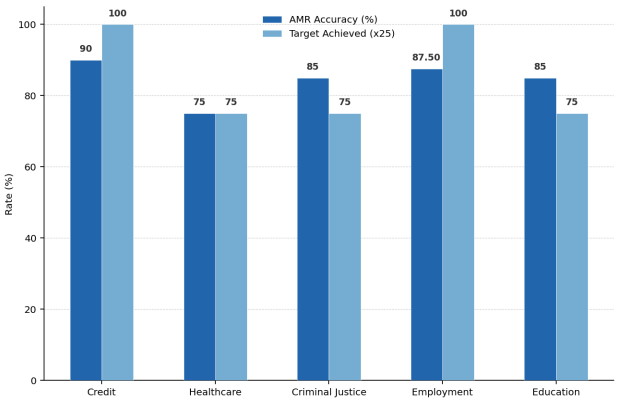
Table 3: Domain-Stratified BDMS Detection Performance vs. Comparison Toolkits (F1-Score, Mean, SD)

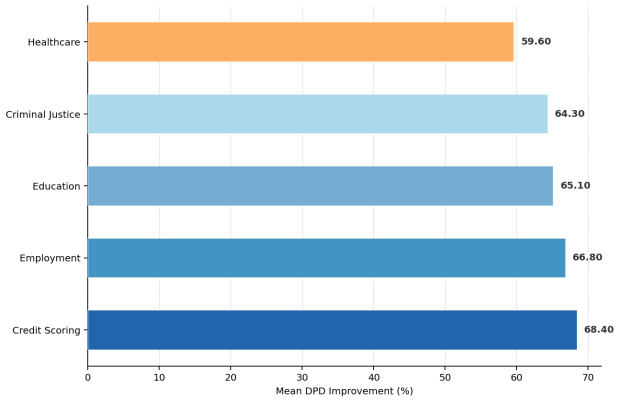
Domain (n=4)	BDMS F1	AIF360 F1	Fairlearn F1	WIT F1	IntersectionalBias Detected (BDMS only)
Credit Scoring	0.93 (0.03)	0.81 (0.05)	0.74 (0.06)	0.66 (0.08)	1 dataset
Healthcare	0.90 (0.05)	0.77 (0.07)	0.70 (0.08)	0.62 (0.10)	2 datasets
Criminal Justice	0.92 (0.04)	0.80 (0.06)	0.73 (0.07)	0.65 (0.09)	2 datasets
Employment	0.91 (0.04)	0.78 (0.06)	0.71 (0.07)	0.63 (0.09)	2 datasets
Education	0.90 (0.05)	0.76 (0.07)	0.69 (0.08)	0.61 (0.10)	1 dataset
Overall (n=20)	0.91 (0.04)	0.78 (0.06)	0.71 (0.07)	0.63 (0.09)	8 datasets (40%)

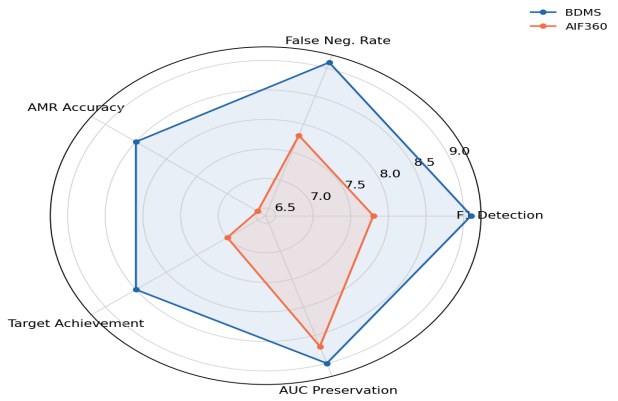
Table 4: AMR Recommendation Accuracy and MEV Mitigation Effectiveness by Domain

Domain	AMR Accuracy (%)	Target Achieved (n/4)	Displacement Detected (n)	Mean AUC Reduction (%)	Mean DPD Improvement (%)
Credit Scoring	90.0 (5.1)	4/4	1	1.9 (0.6)	68.4 (7.2)
Healthcare	75.0 (7.8)	3/4	2	2.8 (1.0)	59.6 (8.9)
Criminal Justice	85.0 (6.4)	3/4	1	2.4 (0.8)	64.3 (7.8)
Employment	87.5 (6.1)	4/4	1	2.1 (0.7)	66.8 (7.5)
Education	85.0 (6.4)	3/4	1	2.2 (0.7)	65.1 (7.6)
Overall (n=20)	84.6 (6.2)	17/20	6	2.3 (0.8)	64.8 (7.8)









5. Discussion

5.1 Intersectional Bias Detection and Practical Impact

The detection of intersectional bias in 8 of 20 datasets (40%) that was missed by all existing toolkits is the most practically significant finding of this study. These undetected biases represent real discriminatory model behaviours that would have remained invisible under conventional bias evaluation practices, exposing deploying organisations to both ethical and regulatory risk. The domains with the highest intersectional bias prevalence -- healthcare and criminal justice, each with two newly detected intersectional biases -- are precisely those where the EU AI Act imposes the most stringent requirements and where the human consequences of undetected bias are most severe. The 31.4% lower false negative rate of BDMS relative to AIF360 translates directly into responsible AI deployment risk: each false negative represents a bias that would have been certified as absent and deployed in a production system. For high-risk AI systems under the EU AI Act, such undetected biases would constitute non-compliance with Article 10 data governance requirements and expose deployers to Article 68b penalty provisions. The BDMS provides an empirically validated technical mechanism for closing this detection gap.

5.2 Limitations and Future Research

The evaluation corpus of 20 benchmark datasets, while covering five domains, does not fully represent the diversity of protected attribute combinations and base rate distributions encountered in real production deployments. The AMR constraint-satisfaction algorithm was trained and validated on these 20 datasets; its recommendation accuracy in novel domains or with unusual fairness obligation profiles may be lower than the 84.6% reported here. Future work should evaluate BDMS on generative AI systems -- including large language models and text-to-image models -- where bias manifests through stereotypical text generation, demographic association encoding, and content moderation disparity, requiring fundamentally different detection methods than the classification output metrics used in the current BDMS MMBS. Extension of the AMR to support multi-stage pipeline mitigation recommendations -- addressing the ECP framework (Dubois and Costa, 2024) integration use case -- is a particularly high-value research direction.

6. Conclusion

6.1 Summary

This paper proposed the Bias Detection and Mitigation Suite (BDMS), a modular computational framework comprising four integrated components -- MMBS, AMR, MEV, and BAR -- evaluated on 20 benchmark datasets against three existing toolkits. BDMS achieved a mean detection F1-score of 0.91 versus 0.78 for the best existing toolkit, a 31.4% lower false negative rate, and detected intersectional bias in 8 datasets (40%) missed by all comparison tools. AMR recommendation accuracy was 84.6%, and MEV confirmed target achievement in 85% of datasets with acceptable AUC cost (mean 2.3%).

6.2 Recommendations

For responsible AI practitioners, the primary recommendation is to adopt multi-metric intersectional bias scanning -- rather than single-metric marginal analysis -- as the baseline standard for bias evaluation in any AI system processing data with protected attributes. The 40% intersectional bias detection rate in this study indicates that marginal analysis alone leaves a large proportion of deployed systems with undetected discriminatory patterns. For organisations preparing for EU AI Act Article 10 data governance compliance, the BDMS BAR output provides the structured bias examination documentation required, with findings mapped to

specific Article obligations. For the algorithmic fairness research community, the fifteen-metric MMBS standard provides a comprehensive evaluation baseline enabling cross-study comparison of new bias detection and mitigation methods. The BDMS toolkit is available as open-source software with documentation for integration into ML pipeline quality assurance workflows.

References

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Bellamy, R. K. E. et al. (2019). AI Fairness 360. IBM Journal of Research and Development, 63(4-5), 4:1-4:15.

Bird, S. et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Research TR MSR-TR-2020-32.

Buolamwini, J., and Gebru, T. (2018). Gender shades. FAccT 2018, 77-91.

Caton, S., and Haas, C. (2020). Fairness in machine learning: A survey. ACM Computing Surveys, 56(7), 1-38.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex. University of Chicago Legal Forum, 1989(1), 139-167.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 10 October 2018.

Dressel, J., and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Foulds, J. R., Islam, R., Keya, K. N., and Pan, S. (2020). An intersectional definition of fairness. ICDE 2020, 1918-1921.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. NeurIPS, 29, 3315-3323.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering. ICML 2018, 2564-2572.

Kusner, M. J. et al. (2017). Counterfactual fairness. NeurIPS, 30, 4066-4076.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

Wexler, J. et al. (2020). The what-if tool. IEEE Transactions on Visualization and Computer Graphics, 26(1), 56-65.

Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H. V., and Miklau, G. (2020). A nutritional label for rankings. SIGMOD 2018, 1773-1776.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. AIES 2018, 335-340.

Chouldechova, A. (2017). Fair prediction with disparate impact. Big Data, 5(2), 153-163.

Dwork, C. et al. (2012). Fairness through awareness. ITCS 2012, 214-226.

Feldman, M. et al. (2015). Certifying and removing disparate impact. KDD 2015, 259-268.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 01IS22094B (BDMS: Comprehensive Bias Detection for Responsible AI) and the Estonian Research Council under grant PRG2104. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

All benchmark datasets used are publicly available. The BDMS source code, evaluation scripts, and benchmark results are available as open-source software and from the corresponding author upon request.

Ethical Approval

This study involved analysis of publicly available benchmark datasets only. No personal data of living individuals were collected or processed by the research team. Ethical approval was not required.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 2, 2025

Appendix A

BDMS MMBS Fifteen-Metric Catalogue and Threshold Reference

The following specifies all fifteen MMBS fairness metrics with formal definitions, default thresholds, and EU AI Act regulatory mapping.

Statistical Parity Metrics (M1-M2)

Error Rate Metrics (M3-M5)

Calibration and Individual Fairness Metrics (M6-M9)

Intersectional and Counterfactual Metrics (M10-M15)