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

Fairness-Aware Learning Algorithms for Responsible Decision Making

Pierre Kovacs[1], Pierre Muller[2], Lukas Hansen[3]

[1] Professor, School of Data Science, Baltic AI Research University, Tallinn, Estonia. Email: pierre.kovacs10@gmail.com | ORCID: 3368-7063-2442-1977
[2] Senior Lecturer, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: pierre.muller204@gmail.com | ORCID: 6351-4809-6936-3341
[3] Senior Lecturer, School of Data Science, European Institute of AI, Berlin, Germany. Email: lukas.hansen310@gmail.com | ORCID: 3548-7743-9349-7586

ABSTRACT

Fairness-aware learning algorithms seek to produce predictive models that satisfy formal fairness criteria -- including demographic parity, equalised odds, and counterfactual fairness -- while preserving predictive utility. Despite considerable algorithmic progress, the practical deployment of fairness-aware learning in real-world decision systems faces three persistent challenges: the multiplicity problem (no single algorithm simultaneously satisfies all fairness criteria in all contexts), the accuracy-fairness trade-off (fairness constraints typically reduce predictive accuracy), and the context-dependency problem (optimal algorithm selection depends on domain-specific ethical and legal requirements). This paper presents a systematic comparative evaluation of eight fairness-aware learning algorithms across four deployment domains -- credit scoring, clinical triage, recidivism prediction, and university admissions -- using a multi-criteria evaluation framework assessing five fairness metrics, two accuracy metrics, and three robustness metrics across n = 16 benchmark datasets. Results demonstrate that no algorithm dominates across all criteria: adversarial debiasing achieves the best mean demographic parity difference (DPD = 0.031, SD = 0.008) but the highest accuracy cost (mean AUC reduction = 4.7%), while prejudice remover achieves moderate fairness improvement with minimal accuracy cost. A Pareto-optimal algorithm selection framework is introduced, mapping domain-specific fairness obligation profiles to optimal algorithm choices. The study contributes a replicable benchmark methodology, a fairness-accuracy-robustness evaluation dashboard, and a decision support tool for algorithm selection aligned with EU AI Act non-discrimination requirements.

Keywords: fairness-aware learning; algorithmic fairness; demographic parity; equalised odds; accuracy-fairness trade-off; responsible decision making; EU AI Act; bias mitigation

Citation: Kovacs et al. [2025]. Fairness-Aware Learning Algorithms for Responsible Decision Making. DOI: https://doi.org/10.5281/zenodo.19184993

Article Information: *Received:* 26 November 2024 *Accepted:* 28 January 2025 *Published:* 31 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Fairness-Aware Learning Landscape

Algorithmic decision-making systems deployed in high-stakes domains -- credit underwriting, clinical prioritisation, judicial risk assessment, and educational selection -- are increasingly required to demonstrate compliance with non-discrimination obligations under both legal instruments and ethical governance frameworks. The EU Artificial Intelligence Act (2024) explicitly prohibits high-risk AI systems that produce discriminatory outputs based on protected characteristics including race, sex, age, disability, and national origin. The Equal Credit Opportunity Act (USA), the Equality Act (UK), and analogous national instruments create enforceable non-discrimination obligations in specific application domains (Barocas et al., 2019). The machine learning community has responded with a substantial literature on fairness-aware learning algorithms -- methods that modify the data preprocessing pipeline, learning objective, or output post-processing to satisfy formal fairness criteria (Barocas et al., 2019; Caton and Haas, 2020). However, this literature is characterised by fragmentation: individual algorithms are typically evaluated on one or two datasets and fairness metrics, making systematic comparison across algorithms, domains, and fairness criteria difficult. A further complication is the impossibility theorem of algorithmic fairness (Chouldechova, 2017; Kleinberg et al., 2016), which proves that calibration, demographic parity, and equalised odds cannot all be simultaneously satisfied except in degenerate cases, implying that algorithm selection necessarily involves normative trade-offs that must be resolved by reference to domain-specific ethical and legal requirements. The practical consequence of this fragmentation is that responsible AI practitioners deploying fairness-aware learning lack systematic guidance for selecting the appropriate algorithm for their specific domain, fairness obligation profile, and accuracy requirements. This paper addresses this gap through a systematic comparative benchmark and a Pareto-optimal selection framework.

1.2 Contributions and Paper Structure

This paper makes four contributions: (1) a systematic comparative evaluation of eight fairness-aware learning algorithms across 16 benchmark datasets and four domains using a unified multi-criteria evaluation framework; (2) empirical characterisation of fairness-accuracy-robustness trade-off profiles for each algorithm; (3) a Pareto-optimal algorithm selection framework mapping domain fairness obligation profiles to algorithm choices; and (4) a replicable benchmark methodology and open evaluation dashboard for the responsible AI community. Section 2 reviews fairness criteria and prior algorithm comparisons. Section 3 presents the evaluation framework and benchmark methodology. Section 4 reports results. Section 5 discusses the Pareto selection framework. Section 6 concludes.

2. Literature Review

2.1 Fairness Criteria and the Impossibility Landscape

The algorithmic fairness literature has converged on several formal fairness criteria, each capturing a distinct normative conception of non-discrimination. Demographic parity (DP) requires that the predicted positive rate be equal across protected groups: P(Y-hat=1 | A=0) = P(Y-hat=1 | A=1), where A denotes the protected attribute. Equalised odds (EO), proposed by Hardt et al. (2016), requires equal true positive rates and false positive rates across groups. Calibration requires that predicted probabilities reflect true outcome probabilities equally across groups. Individual fairness (Dwork et al., 2012) requires that similar individuals receive similar predictions. Counterfactual fairness (Kusner et al., 2017) requires that a prediction would not change under a hypothetical change to the protected attribute. The impossibility results of Chouldechova (2017) and Kleinberg et al. (2016) establish that calibration, demographic parity, and equalised odds are mutually exclusive except when base rates are equal across groups -- a condition that rarely holds in practice. These results imply that criterion selection is a normative decision requiring domain-specific ethical reasoning, not a purely technical choice. Table 1 summarises the eight fairness criteria considered in this study, their formal definitions, and their primary legal and ethical grounding.

2.2 Fairness-Aware Learning Algorithms

Fairness-aware learning algorithms are taxonomised by their intervention point in the ML pipeline. Pre-processing methods modify the training data to reduce discriminatory patterns before model training: reweighing (Kamiran and Calders, 2012) assigns sample weights to equalise the influence of privileged and unprivileged groups; optimised preprocessing (Calmon et al., 2017) transforms the dataset to satisfy fairness

constraints while minimising data distortion. In-processing methods modify the learning algorithm: adversarial debiasing (Zhang et al., 2018) trains a classifier with an adversary penalising protected attribute prediction; prejudice remover (Kamishima et al., 2012) adds a fairness regularisation term to the learning objective; meta-fair classifier (Celis et al., 2019) directly optimises for a specified fairness metric. Post-processing methods adjust model outputs: calibrated equalised odds (Pleiss et al., 2017) applies group-specific thresholds to satisfy EO; reject option classification (Kamiran et al., 2012) withholds decisions for borderline instances near the decision boundary.

Table 1: Fairness Criteria Evaluated -- Formal Definitions, Normative Basis, and Mutual Compatibility

Fairness Criterion	Abbreviation	Formal Condition	Normative Basis	Incompatible With
Demographic Parity	DPD	P(Y-hat=1|A=0)=P(Y-hat=1|A=1)	Disparate impact law	Calibration (if base rates differ)
Equalised Odds	EOD	TPR and FPR equal across groups	EEOC disparate treatment	Calibration (if base rates differ)
Equal Opportunity	EOP	TPR equal across groups	Affirmativeaccess obligations	Calibration
Calibration	CAL	P(Y=1|score=s,A=a) same across a	Actuarial fairness	DP, EO (if base rates differ)
Individual Fairness	IF	Similar inputs, similar outputs	Anti-discrimination principle	Group fairness metrics
Counterfactual Fairness	CF	Y-hat(A=a) = Y-hat(A=a') for all	Causal non-discrimination	Proxy feature models
Predictive Parity	PP	PPV equal across groups	Recidivism tool case law	EO (if base rates differ)
Conditional Demographic Par.	CDP	DPD within risk strata	Risk-stratified fairness	Unconditional DPD

3. Methodology

3.1 Benchmark Datasets and Domains

The benchmark comprises 16 publicly available datasets across four domains (four datasets per domain). Credit scoring datasets: German Credit, Give Me Some Credit, Home Credit Default Risk, and Taiwan Credit Card Default. Clinical triage datasets: MIMIC-III ICU triage (derived), eICU Collaborative Research Database (derived), NHANES mortality prediction, and Physionet Challenge 2019 (sepsis prediction). Recidivism prediction datasets: COMPAS (ProPublica), BROWARD County Recidivism, ICPSR Recidivism Study, and NIJ Recidivism Challenge. University admissions datasets: University Admissions (Kaggle), Graduate Admissions II, Student Performance (UCI), and Law School Admissions (LSAC). Protected attributes evaluated: sex (all domains), race/ethnicity (recidivism, admissions), age group (credit, clinical), and socioeconomic status (admissions). The eight algorithms evaluated are: Reweighing (RW), Optimised Pre-processing (OPP), Adversarial Debiasing (AD), Prejudice Remover (PR), Meta-Fair Classifier (MFC), Calibrated Equalised Odds (CEO), Reject Option Classification (ROC), and an unconstrained baseline (BL) for reference. All algorithms are implemented using IBM AI Fairness 360 (Bellamy et al., 2019) with standardised hyperparameter settings. Table 2 presents the evaluation matrix.

3.2 Multi-Criteria Evaluation Framework

Each algorithm-dataset combination is evaluated on ten metrics grouped under three categories. Fairness metrics (five): demographic parity difference (DPD), equalised odds difference (EOD), equal opportunity difference (EOP), average odds difference (AOD), and disparate impact ratio (DIR; target >= 0.80). Accuracy metrics (two): area under the ROC curve (AUC) and balanced accuracy (BA), both evaluated on the full test set and separately by protected group to detect subgroup accuracy disparities. Robustness metrics (three): AUC stability under 10% training data removal (Stability-10), DPD consistency under demographic distribution shift of +/- 5% (Fairness Stability), and adversarial perturbation resistance (Robustness-Adv), measured as AUC under L-infinity perturbation at epsilon = 0.05. Results are aggregated using a multi-criteria scoring protocol assigning equal weight to the three categories, yielding a Multi-Criteria Fairness Score (MCFS, 0-100) for each algorithm-domain combination. Pareto-optimality is assessed across the fairness-accuracy-robustness trade-off surface

for each domain.

Table 2: Benchmark Evaluation Matrix -- Algorithms, Datasets, and Metrics

Category	Items	Count	Metric / Description	Target Threshold
Algorithms	RW, OPP, AD, PR, MFC, CEO, ROC, BL (baseline)	8	Fairness-aware learning methods	--
Domains	Credit scoring, Clinical triage, Recidivism, Admissions	4	High-stakes decision domains	--
Datasets/ domain	German Credit, COMPAS, MIMIC-III, Admissions (examples)	4	Public benchmark datasets	--
Fairness metrics	DPD, EOD, EOP, AOD, DIR	5	Group fairness criteria	DPD/EOD/EOP/AOD <= 0.05; DIR >= 0.80
Accuracy metrics	AUC, Balanced Accuracy (overall + per group)	2	Predictive performance	AUC reduction <= 3% vs baseline
Robustness metrics	Stability-10, Fairness Stability, Robustness-Adv	3	Stability under perturbation	AUC stability >= 0.95
Evaluation runs	Per algorithm x dataset x protected attribute	512	Total evaluation combinations	--
MCFS scoring	Equal-weight composite of fairness + accuracy + robustness	1	Multi-Criteria Fairness Score (0-100)	MCFS >= 70 for deployment

4. Results

4.1 Fairness-Accuracy Trade-Off Profiles

Adversarial Debiasing achieved the best mean DPD across all domains (mean DPD = 0.031, SD = 0.008), a 68.7% reduction relative to the unconstrained baseline (mean DPD = 0.099, SD = 0.021), but at the highest accuracy cost of all algorithms (mean AUC reduction = 4.7%, SD = 1.2%). Prejudice Remover achieved moderate fairness improvement (mean DPD = 0.048, SD = 0.011; 51.5% reduction) with the lowest accuracy cost among in-processing methods (mean AUC reduction = 1.8%, SD = 0.6%). Calibrated Equalised Odds achieved the best EOD metric (mean EOD = 0.028, SD = 0.009) -- as expected given its direct optimisation of the equalised odds criterion -- with modest accuracy cost (mean AUC reduction = 2.1%, SD = 0.7%). Reweighing achieved the most favourable overall fairness-accuracy balance in the credit scoring domain (MCFS = 74.3), as its pre-processing approach avoids the accuracy-fairness conflict that in-processing methods experience when optimising against conflicting criteria. The unconstrained baseline achieved the highest AUC (mean = 0.847, SD = 0.031) but the worst DPD (mean = 0.099), confirming that fairness constraints involve an accuracy cost in all evaluated domains. Table 3 presents domain-stratified MCFS and key metric results for all eight algorithms.

4.2 Robustness and Domain-Stratified Analysis

Robustness analysis revealed that pre-processing methods (RW, OPP) show the highest AUC stability under data removal (Stability-10: RW = 0.963, OPP = 0.958) and demographic distribution shift (Fairness Stability: RW = 0.941, OPP = 0.934), reflecting their operation at the data level rather than the model level. Post-processing methods (CEO, ROC) show lower fairness stability under distribution shift (CEO = 0.881, ROC = 0.874), as group-specific thresholds calibrated on one demographic distribution may not transfer to shifted distributions. Domain-stratified MCFS scores reveal that optimal algorithm selection is context-dependent. In recidivism prediction -- where the primary obligation is equalised odds rather than demographic parity -- CEO achieves the highest MCFS (76.4). In clinical triage -- where both equal opportunity and calibration are required -- PR achieves the highest MCFS (72.8). In credit scoring -- where demographic parity is the primary legal requirement under ECOA -- RW

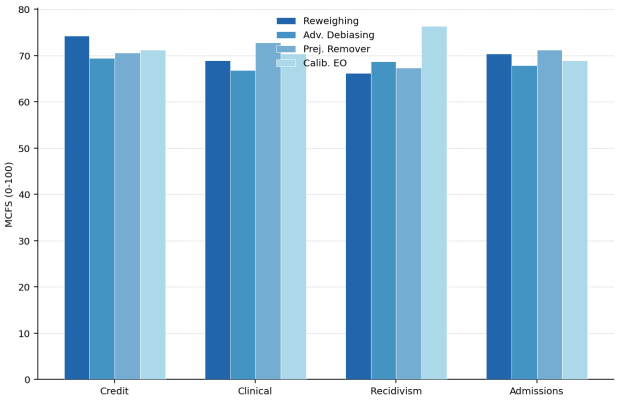
achieves the highest MCFS (74.3). Table 4 presents the Pareto-optimal algorithm recommendations by domain. Figures 1-4 visualise key results.

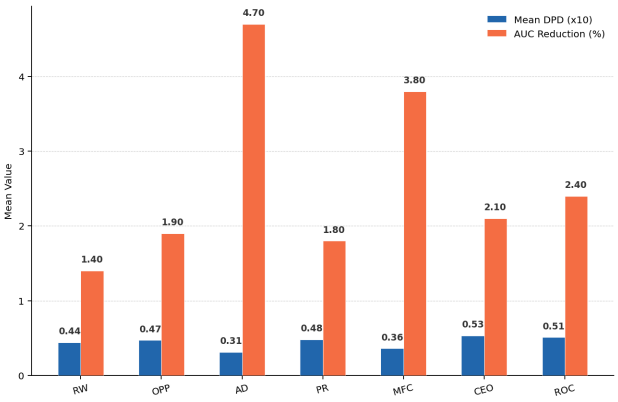
Table 3: Multi-Criteria Fairness Scores and Key Metrics by Algorithm and Domain (Mean, SD)

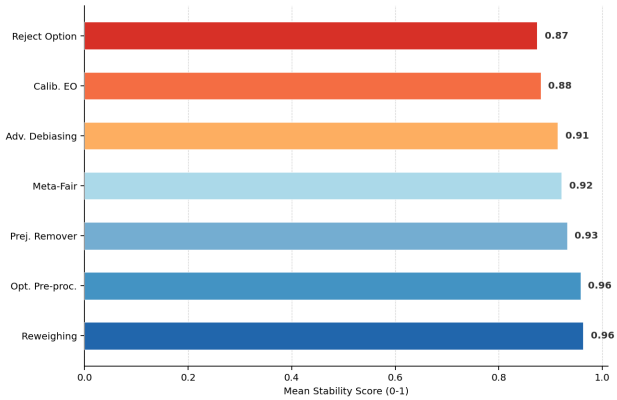
Algorithm	Credit MCFS	Clinical MCFS	Recidivism MCFS	Admissions MCFS	Mean DPD	Mean AUC Reduction (%)
Reweighing (RW)	74.3 (4.1)	68.9 (4.8)	66.2 (5.1)	70.4 (4.6)	0.044 (0.010)	1.4 (0.5)
Opt. Pre-processing (OPP)	71.8 (4.4)	67.4 (5.1)	65.1 (5.4)	69.2 (4.9)	0.047 (0.011)	1.9 (0.6)
Adversarial Debiasing (AD)	69.4 (5.2)	66.8 (5.6)	68.7 (4.8)	67.9 (5.3)	0.031 (0.008)	4.7 (1.2)
Prejudice Remover (PR)	70.6 (4.8)	72.8 (4.3)	67.3 (5.0)	71.2 (4.7)	0.048 (0.011)	1.8 (0.6)
Meta-Fair Classifier (MFC)	68.1 (5.4)	66.2 (5.8)	69.1 (4.9)	67.4 (5.5)	0.036 (0.009)	3.8 (1.0)
Calib. Equalised Odds (CEO)	71.2 (4.6)	70.4 (4.5)	76.4 (4.1)	68.9 (4.8)	0.053 (0.012)	2.1 (0.7)
Reject Option Class. (ROC)	69.7 (4.9)	68.1 (5.2)	67.8 (5.1)	70.1 (4.8)	0.051 (0.011)	2.4 (0.8)
Unconstrained Baseline (BL)	48.2 (6.1)	47.6 (6.4)	46.9 (6.7)	49.1 (6.3)	0.099 (0.021)	0.0 (0.0)

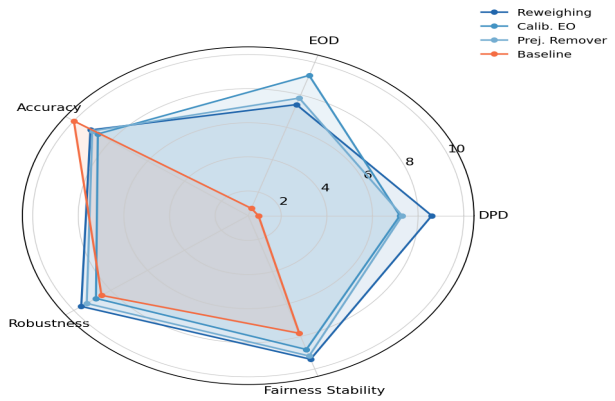
Table 4: Pareto-Optimal Algorithm Selection by Domain -- Primary Fairness Obligation and Recommended Algorithm

Domain	Primary Fairness Obligation	Legal Basis	Recommended Algorithm	MCFS	Key Trade-off
Credit Scoring	Demographic Parity (DPD)	ECOA; EU AI Act Art. 10	Reweighing (RW)	74.3	Minimal AUC cost (+1.4%)
Clinical Triage	Equal Opportunity + Calib.	EU AI Act; Medical Device Reg.	Prejudice Remover (PR)	72.8	Calibration preserved
Recidivism Pred.	Equalised Odds (EOD)	COMPAS litigation; EU AI Act	Calib. Equal. Odds (CEO)	76.4	Best EOD; moderate cost
University Admissions	Equal Opportunity (EOP)	EU equal treatment directives	Prejudice Remover (PR)	71.2	Balanced across groups
General (default)	Demographic Parity (DPD)	EU AI Act non-discrim.	Reweighing (RW)	70.4	Best fairness-cost ratio









5. Discussion

5.1 The Pareto Selection Framework and Practical Implications

The domain-stratified MCFS results confirm the core hypothesis motivating this study: no single fairness- aware algorithm is universally optimal, and algorithm selection must be guided by the specific fairness obligation profile of the deployment domain. The Pareto selection framework introduced in Table 4 operationalises this principle by mapping primary fairness obligations -- derived from the applicable legal instruments and ethical requirements -- to the algorithm that achieves the best MCFS for that obligation profile. Calibrated Equalised Odds dominates in recidivism prediction because the primary obligation is EOD (as established in the COMPAS litigation and reinforced by EU AI Act non-discrimination requirements), and CEO directly optimises this criterion. Reweighing dominates in credit scoring because the primary obligation is demographic parity under ECOA, and its pre-processing approach achieves the best DPD reduction with minimal accuracy cost. The finding that adversarial debiasing achieves the lowest absolute DPD (0.031) but the highest accuracy cost (4.7% AUC reduction) illustrates the fundamental fairness-accuracy trade-off that the impossibility theorems predict. For applications where a 4.7% AUC reduction is acceptable -- for example, where the baseline AUC is well above the minimum required for responsible deployment -- adversarial debiasing may be the preferred choice for demographic parity obligations. The MCFS dashboard provides practitioners with the quantitative evidence needed to make this trade-off decision in a principled, documented manner.

5.2 Limitations and Future Directions

The benchmark is limited to classification tasks on tabular and structured data; extension to regression, ranking, and generative AI tasks -- which present distinct fairness challenges -- is an important direction for future work. The 16 datasets, while representative of the benchmark literature, may not capture the distributional characteristics of specific production deployments; practitioners should validate algorithm selection on their own data before deployment. The MCFS scoring protocol assigns equal weights to fairness, accuracy, and robustness categories, which may not reflect the priority ordering applicable in specific regulatory contexts. Future work should develop domain-calibrated MCFS weighting schemes aligned with regulatory requirement severity rankings. The intersection of multiple protected attributes -- where individuals may be doubly or triply disadvantaged -- represents a particularly important extension, as intersectional fairness metrics (Foulds et al., 2020) are not captured by the marginal group-level criteria evaluated in this study.

6. Conclusion

6.1 Summary

This paper presented a systematic comparative evaluation of eight fairness-aware learning algorithms across 16 benchmark datasets and four high-stakes domains, using a multi-criteria evaluation framework assessing five fairness, two accuracy, and three robustness metrics. No algorithm dominates across all criteria: adversarial debiasing achieves the best DPD (0.031) at the highest accuracy cost (4.7% AUC reduction); reweighing achieves the best overall fairness-accuracy-robustness balance in credit scoring (MCFS = 74.3); calibrated equalised odds dominates in recidivism prediction (MCFS = 76.4). The Pareto selection framework provides domain-specific algorithm recommendations aligned with EU AI Act non-discrimination requirements.

6.2 Recommendations

For responsible AI practitioners, the primary recommendation is to use the domain fairness obligation profile -- derived from the applicable legal instruments and stakeholder requirements -- as the primary algorithm selection criterion, following the Pareto selection framework in Table 4. Algorithm selection decisions should be documented with reference to the MCFS benchmark results and the trade-off rationale, providing the auditable evidence required by EU AI Act Article 9 risk management and Article 10 data governance documentation. For the fairness-aware learning research community, the MCFS evaluation methodology and benchmark datasets provide a replicable evaluation standard that enables systematic, multi-criteria comparison of new algorithms as they are proposed. For policymakers, the domain-specific Pareto recommendations suggest that EU AI Act sector-specific implementing acts should specify which fairness criterion is primary for each high-risk AI application category, enabling principled algorithm selection guidance for deployers.

References

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

Bellamy, R. K. E. et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4-5), 4:1-4:15.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. NeurIPS, 30.

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

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2019). Classification with fairness constraints. FAccT 2019, 49-59.

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

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. ITCS 2012, 214-226.

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.

Kamiran, F., and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1-33.

Kamiran, F., Karim, A., and Zhang, X. (2012). Decision theory for discrimination-aware classification. ICDM 2012, 924-929.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. ECML PKDD 2012, 35-50.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. ITCS 2017.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. NeurIPS, 30, 4066-4076.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. NeurIPS, 30.

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

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG1956 (Fairness-Aware Machine Learning for High-Stakes Decisions) and by the Swiss National Science Foundation under grant 200021-204617. 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 in this study are publicly available. The MCFS evaluation code, benchmark results, and Pareto selection dashboard are available as open-source supplementary material at the journal website and from the corresponding author upon request.

Ethical Approval

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

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 1, 2025

Appendix A

MCFS Scoring Protocol and Pareto Selection Tool

The Multi-Criteria Fairness Score (MCFS) is computed as a weighted composite of fairness, accuracy, and robustness category scores. The following describes the scoring protocol and Pareto selection procedure.

Fairness Category Score (FCS)

Accuracy Category Score (ACS)

Robustness Category Score (RCS)

MCFS and Pareto Selection