

Revisiting Pre-processing Group Fairness: A Modular Benchmarking Framework

Brodie Oldfield
brodie.oldfield@data61.csiro.au
CSIRO
Sydney, Australia

Ziqi Xu*
ziqu.xu@rmit.edu.au
RMIT University
Melbourne, Australia

Sevvandi Kandanaarachchi
sevvandi.kandanaarachchi@data61.csiro.au
CSIRO
Melbourne, Australia

Abstract

As machine learning systems become increasingly integrated into high-stakes decision-making processes, ensuring fairness in algorithmic outcomes has become a critical concern. Methods to mitigate bias typically fall into three categories: pre-processing, in-processing, and post-processing. While significant attention has been devoted to the latter two, pre-processing methods, which operate at the data level and offer advantages such as model-agnosticism and improved privacy compliance, have received comparatively less focus and lack standardised evaluation tools. In this work, we introduce FairPrep, an extensible and modular benchmarking framework designed to evaluate fairness-aware pre-processing techniques on tabular datasets. Built on the AIF360 platform, FairPrep allows seamless integration of datasets, fairness interventions, and predictive models. It features a batch-processing interface that enables efficient experimentation and automatic reporting of fairness and utility metrics. By offering standardised pipelines and supporting reproducible evaluations, FairPrep fills a critical gap in the fairness benchmarking landscape and provides a practical foundation for advancing data-level fairness research. The source code is available at <https://github.com/broldfield/FairPrep>.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **General and reference** → **Evaluation**.

Keywords

Machine Learning, Fairness, Pre-processing

ACM Reference Format:

Brodie Oldfield, Ziqi Xu, and Sevvandi Kandanaarachchi. 2025. Revisiting Pre-processing Group Fairness: A Modular Benchmarking Framework. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, November 10–14, 2025, Seoul, Republic of Korea. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3746252.3761650>

1 Introduction

Machine learning techniques are becoming increasingly prevalent, shaping how we interact with data and influencing decision-making

across diverse application areas. As these techniques are deployed in scenarios where decisions can significantly impact individual livelihoods, such as employment, finance, and healthcare, the question of fairness in automated decisions becomes essential. In particular, we must ask whether algorithms systematically treat individuals differently based on sensitive attributes such as race or gender [16].

Fairness-aware machine learning methods are commonly divided into three categories: pre-processing, in-processing, and post-processing. Pre-processing techniques modify the training data to reduce bias, such as through re-weighting, resampling, or transforming features to obscure sensitive attributes [6, 7, 12, 27]. In-processing methods embed fairness constraints into the training process, often via regularisation or adversarial objectives that penalise group disparities [1, 2, 14, 15, 24, 28]. Post-processing techniques adjust model outputs after training, for example by shifting decision thresholds to equalise outcomes across groups [8, 10, 13, 19]. These approaches differ in terms of model access, flexibility, and applicability, with pre-processing being especially appealing due to its simplicity and model-agnostic design.

Despite these established categories, prior research has mainly focused on in-processing and post-processing techniques, while pre-processing methods have received less attention. This discrepancy is largely due to the difficulty of standardising evaluation for pre-processing approaches, as many fairness benchmarks lack support for flexible and transparent data manipulation. Tools such as AIF360 [5], UST [26], and FairLearn [22] focus on model-level adjustments. By contrast, Fair-IRT [23] builds on broader research leveraging item response theory (IRT) for algorithm evaluation [18, 25]. Fair-IRT continues this trajectory by explicitly integrating fairness considerations, bridging both model- and data-level evaluation. In addition, inconsistencies in pre-processing, such as outlier removal strategies or methods for splitting training and testing sets, can cause substantial differences in reported performance. Even small changes in data preparation may yield significant variations in outcomes, raising concerns about the reliability and reproducibility of fairness evaluations [9].

To address the lack of standardised evaluation for pre-processing fairness methods, we introduce FairPrep, an extensible and modular framework for benchmarking fairness-aware data transformation techniques on tabular datasets. Built on top of the AIF360 foundation, FairPrep is designed with usability and research flexibility in mind. It enables users to seamlessly integrate new datasets, fairness techniques, and predictive models through a modular API. To facilitate large-scale and reproducible experiments, the framework includes a batch-processing interface that allows users to define experimental configurations and automatically compute both fairness and utility metrics. FairPrep contributes to the fairness research

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License. *CIKM '25, Seoul, Republic of Korea*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2040-6/2025/11

<https://doi.org/10.1145/3746252.3761650>

community by: (1) standardising pre-processing evaluation across datasets and metrics, (2) supporting easy extension and integration with new methods, and (3) streamlining the benchmarking pipeline for reproducible research. To the best of our knowledge, this is the first benchmarking study focused specifically on group fairness through pre-processing techniques.

2 The proposed FairPrep

This section presents FairPrep, a modular and extensible benchmarking framework for fairness-aware pre-processing. We introduce its two-stage architecture, describe the fairness metrics, outline the five benchmark datasets, and detail the four integrated pre-processing techniques.

2.1 System Architecture

FairPrep operates in two stages: a pre-processing stage and a benchmarking stage. These are executed independently via command-line interfaces, allowing modular and flexible usage. In the pre-processing stage, the user specifies a dataset, a sensitive attribute, and a fairness pre-processing method. The technique is applied to transform the dataset and reduce bias with respect to the sensitive attribute. Fairness metrics are computed for both the original and processed datasets to assess the effect of the transformation. Both versions are cached for reuse in benchmarking. In the benchmarking stage, a predictive model is trained on both datasets using a holdout validation scheme (e.g., 70% training, 15% validation, 15% testing). Performance and fairness metrics are computed from the predictions to evaluate trade-offs. To support fine-grained analysis, thresholds are swept from 0.01 to 0.99 in increments of 0.01. At each threshold, metrics are recorded, and the optimal threshold balancing fairness and accuracy is identified and visualised.

To enable scalable experiments, FairPrep supports YAML-based batch execution. Users may define fixed jobs or provide lists of inputs to automatically generate all valid combinations. This supports reproducible and extensible benchmarking across datasets, models, and fairness methods.

2.2 Group Fairness Metrics

FairPrep adopts a set of comprehensive metrics to assess data-level fairness before training and model-level fairness and utility after training. These metrics are grouped by stage: pre-processing or benchmarking.

2.2.1 Metrics for Pre-processing Evaluation. In the pre-processing stage, FairPrep evaluates the effectiveness of fairness interventions using several group-level and distributional metrics. These include: (1) *base rate*, which measures the overall proportion of positive labels, and its conditional variant $\Pr(Y = 1 \mid S = s)$; (2) *consistency*, which assesses individual-level fairness by measuring the agreement between a sample's label and those of its neighbors in the feature space; (3) *disparate impact*, and (4) *statistical parity*, which capture group-level fairness by quantifying outcome ratios and differences across sensitive groups; (5) *number of positives* and (6) *number of negatives*, which are computed as the total counts of positive and negative labels, respectively, providing a basic view of label distributions; and (7) *empirical difference*, which quantifies the disparity of a selected metric between privileged and unprivileged

groups. Together, these metrics offer a comprehensive assessment of data-level bias prior to model training.

2.2.2 Metrics for Benchmarking Evaluation. In the benchmarking stage, FairPrep evaluates the predictive models trained on both original and pre-processed datasets using a combination of performance and fairness metrics. Utility is assessed using (8) *balanced accuracy*, defined as the average of the true positive rate (TPR) and true negative rate (TNR), which is particularly useful in imbalanced classification tasks. Fairness is evaluated using several group-based metrics, including (9) *equal opportunity*, which measures the difference in TPR between underprivileged and privileged groups, and (10) *equal odds*, which averages disparities in both TPR and false positive rate (FPR) across groups. For consistency with the pre-processing evaluation, the stage also computes *disparate impact* and *statistical parity difference*. Additionally, (11) *the theil index*, a generalised entropy-based metric with parameter $\alpha = 1$ [21], is included to capture both individual- and group-level fairness violations. Together, these benchmarking metrics provide a comprehensive evaluation of how fairness-aware pre-processing techniques affect downstream model behaviour.

2.3 Datasets

We evaluate FairPrep on 5 commonly used fairness-related datasets, briefly described below: (1) the *Adult Census* dataset [4] contains demographic and financial information of individuals from the U.S. Census Bureau, with the target variable indicating whether an individual earns over \$50K annually; the sensitive attribute is typically gender or race; (2) the *Bank Marketing* dataset [17] includes client information and marketing interaction data collected by a Portuguese banking institution; the target variable denotes whether a client subscribes to a term deposit, with age or marital status used as the sensitive attribute; (3) the *ProPublica COMPAS* dataset [20] comprises criminal history and demographic attributes of defendants; the prediction task focuses on recidivism within two years, and race is commonly used as the sensitive attribute; (4) the *German Credit* dataset [11] contains features from credit applicants at a German bank, with the target being a binary indicator of credit risk (good or bad); gender or age is often used as the sensitive attribute; and (5) the *MEPS Panel 21* dataset [3] is drawn from the Medical Expenditure Panel Survey, which includes data on demographics, healthcare usage, costs, and quality of care collected from individuals, providers, and employers in the United States. In this dataset, the target variable is *utilization*, a score reflecting the appropriateness and frequency of care received. We binarize this score such that values less than 10 are labelled as 0 (low utilisation), and values greater than or equal to 10 are labelled as 1 (high utilisation), with race used as the sensitive attribute.

2.4 Fairness Pre-processing Methods

The proposed FairPrep involve four widely studied fairness pre-processing methods: Reweighting, Learned Fair Representations, Disparate Impact Remover, and Optimised Pre-processing. We consider a dataset $D = (S, X, Y)$, with protected attribute S , features X , and binary label Y in the following discussion.

Table 1: Results of fairness metrics in the pre-processing stage for 3 datasets and 4 methods.

	Adult Census					ProPublica COMPAS					German Credit				
Metrics	Orig.	RW	LFR	DIR	OPP	Orig.	RW	LFR	DIR	OPP	Orig.	RW	LFR	DIR	OPP
(1) base rate	0.239	0.239	0.095	0.239	0.252	0.530	0.530	0.031	0.530	0.529	0.700	0.700	1	0.700	0.694
(2) consistency	0.719	0.719	1	0.719	0.700	0.595	0.595	1	0.595	0.610	0.661	0.661	1	0.661	0.644
(3) disparate impact	0.360	1.000	0.601	0.360	0.816	0.788	1.000	0.786	0.788	0.906	0.897	1.000	1	0.897	0.927
(4) statistical parity	-0.195	0.000	-0.043	-0.195	-0.050	-0.135	0.000	-0.008	-0.135	-0.054	-0.075	0.000	0	-0.075	-0.052
(5) Num. positives	11687	11687	4664	11687	12330	2795	2795	161	2795	2793	700	700	1000	700	694
(6) Num. negatives	37155	37155	44178	37155	36512	2483	2483	5117	2483	2485	300	300	0	300	306
(7) empirical difference	1.022	0.000	0.509	1.022	0.205	0.317	0.000	0.249	0.317	0.118	0.239	0.001	0.798	0.239	0.166

2.4.1 Reweighting (RW) [12]. RW assigns weights to samples based on the observed and expected probabilities of each (group, label) pair, assuming independence between group membership and outcome. For a dataset $D = (S, X, Y)$, the weight for a given pair (s, y) is computed as:

$$W(s, y) = \frac{Pr(S = s) \times Pr(Y = y)}{Pr(S = s \wedge Y = y)}.$$

This adjusts the training distribution to reduce the disparity between groups. For example, if the observed success rate of a group is lower than expected, its positive instances will be up-weighted accordingly. These weights are passed to the classifier and used to reweight the loss during training.

2.4.2 Learned Fair Representations (LFR) [27]. LFR learns a latent representation of the input data that retains task-relevant information while obfuscating sensitive attribute information [27]. Given a dataset $D = (S, X, Y)$, the goal is to find a prototype set Z such that: (1) the mapping from X to Z satisfies statistical parity; (2) Z retains as much information as possible unrelated to S ; and (3) the composed mapping from X to Z and then to Y approximates the original classifier $f : X \rightarrow Y$.

2.4.3 Disparate Impact Remover (DIR) [7]. DIR edits feature values to improve group fairness while preserving the rank-ordering of features within groups [7]. The concept of disparate impact has legal origins and is defined as:

Definition 2.1. Given a dataset $D = (S, X, Y)$, with protected attribute S , features X , and binary label Y , disparate impact exists if

$$\frac{Pr(Y = 1 \mid S = 0)}{Pr(Y = 1 \mid S = 1)} \leq \tau = 0.8,$$

for positive outcomes $Y = 1$ and majority group $S = 1$.

DIR produces a repaired version of X , denoted \tilde{X} , such that the marginal distributions maintain the same cumulative probabilities. Formally, if $F_s : X_s \rightarrow [0, 1]$ is the cumulative distribution function for group $S = s$, then the repaired features \tilde{x} satisfy $F_s(x) = F_s(\tilde{x})$.

2.4.4 Optimised Pre-processing (OPP) [6]. OPP learns a probabilistic transformation of both features and labels to minimise discrimination while preserving individual utility and limiting distortion [6]. Given a dataset $D = (S, X, Y)$, the transformed dataset $\hat{D} = (S, \hat{X}, \hat{Y})$ is constructed such that the joint distribution (\hat{X}, \hat{Y}) is close to the original (X, Y) , while ensuring that $Pr(\hat{Y} \mid S)$ is close to a target distribution and that the changes to individual data points are small.

3 Experiments

In this section, we present empirical evaluations of FairPrep to demonstrate its flexibility, modular design, and practical utility in benchmarking fairness-aware pre-processing techniques. Applied to multiple benchmark datasets, the results show how different techniques influence group fairness and confirm FairPrep’s ability to support standardised, reproducible evaluation.

3.1 Experimental Setup

To showcase the practicality and extensibility of *FairPrep*, we conduct experiments using five commonly studied tabular datasets and four representative fairness pre-processing techniques. These datasets are processed using our pre-processing interface and evaluated via predictive models. It is important to note that FairPrep is designed as a modular and extensible framework. All components, including datasets, fairness pre-processing techniques, and predictive models, are exposed via standardised interfaces. This allows users to easily plug in their own datasets, implement custom pre-processing methods, or evaluate new classifiers without modifying the core framework. The experiments presented here serve as instantiations of this general pipeline.

Our evaluation follows a two-stage pipeline. In the pre-processing stage, the framework computes pre-processing metrics on both the original and transformed datasets to assess the effect of bias mitigation before model training. These results are summarised in Table 1. In the benchmarking stage, each model is trained on the original and pre-processed datasets using a holdout validation scheme, and evaluated on both fairness and performance metrics. The resulting metrics are used to generate threshold-sensitive trade-off plots (e.g., Figures 1a and 1b), which offer a fine-grained view of how fairness and utility evolve across decision thresholds.

Due to space limitations, we present the results from the pre-processing stage for three datasets and four fairness-aware pre-processing methods. For the benchmarking stage, we report the comparative results of the original and processed versions of the Adult dataset using Logistic Regression. The full experimental results can be reproduced via the link provided in the abstract.

3.2 Results for Pre-processing Stage

Table 1 summarises fairness metrics computed on the original and processed datasets across three benchmark datasets and four pre-processing methods. Several consistent trends emerge. (1) RW improves group fairness across all datasets, achieving perfect disparate

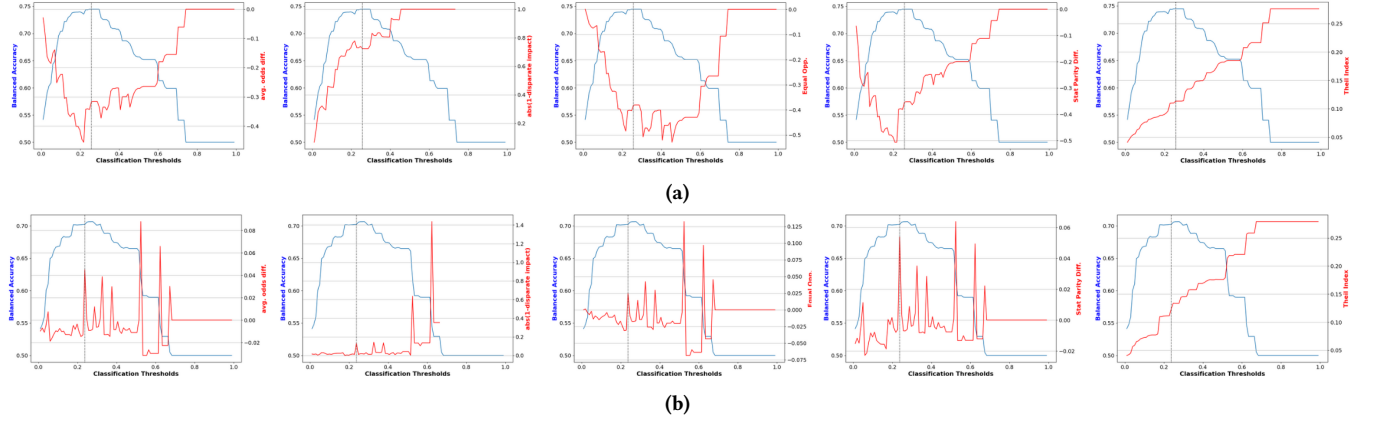


Figure 1: Balanced Accuracy (left axis, blue) versus fairness metrics (right axis, red) across varying classification thresholds for five different fairness metrics. Each subplot corresponds to a specific metric. (a) Results obtained from the original Adult dataset. (b) Results obtained after applying the Reweighting pre-processing method to the Adult dataset. Logistic Regression is used as the predictive model in both settings.

impact (1.0) and eliminating statistical parity difference (0.0), while leaving label distributions unchanged. This outcome aligns with its design, which adjusts instance weights without modifying features or labels, making it a practical choice when preserving original data semantics is important. (2) LFR introduces substantial changes to the data, often removing all positive labels (Adult, COMPAS) or setting all to positive (German). While this results in ideal fairness metrics, such extreme shifts make the data unrealistic and potentially unusable for downstream tasks. The high empirical differences further indicate that fairness is achieved at the cost of data fidelity. (3) DIR produces minimal change in group-level fairness metrics, with little to no improvement in disparate impact or statistical parity. This suggests that its rank-preserving transformations alone are insufficient for correcting entrenched disparities. (4) OPP strikes a more balanced trade-off by improving fairness metrics while maintaining a reasonable label distribution. Its empirical differences are significantly lower than those in the original data, indicating more stable and effective bias mitigation.

In summary, RW and OPP offer controlled and practical improvements, whereas LFR, despite its effectiveness on fairness metrics, may compromise data integrity. DIR shows limited standalone impact and may benefit from being used in conjunction with other techniques.

3.3 Results for Benchmarking Stage

Figure 1 illustrates how balanced accuracy and various fairness metrics evolve as classification thresholds change, using the Adult dataset and Logistic Regression as the predictive model. Figure 1a presents the results on the original dataset, which exhibits substantial variability and sensitivity to threshold selection. Fairness metrics such as equal opportunity and average odds difference fluctuate considerably, especially in mid-range thresholds. The Theil index increases sharply at more lenient thresholds, indicating rising outcome inequality. These trends suggest that fairness behaviour

on the original dataset is highly unstable and threshold-dependent, complicating model deployment in practice.

By contrast, Figure 1b shows the corresponding results after applying the RW. The processed dataset yields notably better fairness trends despite some fluctuations. Most metrics, including statistical parity difference and disparate impact, remain closer to zero across the threshold range, reflecting improved group-level parity. Although some residual fluctuations persist in average odds difference, the overall behaviour is more robust. Importantly, balanced accuracy remains comparable to the original, indicating that fairness gains are achieved without sacrificing model utility. These findings highlight RW’s effectiveness not only in improving fairness metrics at fixed thresholds but also in enhancing the stability of fairness-utility trade-offs across decision boundaries.

4 Conclusion

In this paper, we present FairPrep, a modular and extensible benchmarking framework for fairness-aware pre-processing in tabular data. By supporting a wide range of datasets, transformation techniques, and predictive models, FairPrep provides a unified and reproducible pipeline for evaluating the trade-offs between fairness and utility. Our empirical results demonstrate the distinct impacts of common pre-processing methods on both data characteristics and downstream model behaviour. While FairPrep standardises evaluation and simplifies extensibility, one limitation is its focus on binary classification tasks with group fairness metrics. Future work will extend the framework to accommodate multi-class and regression tasks, incorporate individual fairness measures, and support integration with modern deep learning pipelines.

Acknowledgments

This work was supported by the research support package from the School of Computing Technologies at RMIT University.

GenAI Usage Disclosure

During this research, generative AI tools are used to enhance the quality and efficiency of specific tasks, all within the permitted scope defined by the ACM policies on the use of GenAI. These tools primarily assist in refining the manuscript’s linguistic elements, such as grammar, syntax, and minor stylistic adjustments, to ensure clarity and conciseness. Additionally, during the implementation phase, AI suggests relevant programming libraries and provides guidance on Python syntax, which helps accelerate development while following established coding practices. All uses are strictly limited to supporting functions and do not involve the generation or interpretation of research data, methodology, or conclusions.

References

- [1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, Vol. 80. 60–69. <http://proceedings.mlr.press/v80/agarwal18a.html>
- [2] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, Vol. 97. 120–129. <http://proceedings.mlr.press/v97/agarwal19d.html>
- [3] Agency for Healthcare Research and Quality (AHRQ). 2016. Medical Expenditure Panel Survey (MEPS) Panel 21 Longitudinal Data File. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp U.S. Department of Health & Human Services.
- [4] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5XW20>.
- [5] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. <https://arxiv.org/abs/1810.01943>
- [6] Flávio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. 3992–4001. <https://proceedings.neurips.cc/paper/2017/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html>
- [7] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2015*. 259–268. doi:10.1145/2783258.2783311
- [8] Sahin Cem Geyik, Stuart Ambler, and Krishnamurthy K. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019*. ACM, 2221–2231. doi:10.1145/3292500.3330691
- [9] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. 2024. FFB: A Fair Fairness Benchmark for In-Processing Group Fairness Methods. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. <https://openreview.net/forum?id=TzAJbtCIAz>
- [10] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*. 3315–3323. <https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html>
- [11] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5NC77>.
- [12] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2011), 1–33. doi:10.1007/S10115-011-0463-8
- [13] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In *12th IEEE International Conference on Data Mining, ICDM 2012*. 924–929. doi:10.1109/ICDM.2012.45
- [14] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012*, Vol. 7524. 35–50. doi:10.1007/978-3-642-33486-3_3
- [15] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, Vol. 80. PMLR, 2569–2577. <http://proceedings.mlr.press/v80/kearns18a.html>
- [16] Renqiang Luo, Huafei Huang, Shuo Yu, Zhuoyang Han, Estrid He, Xiuzhen Zhang, and Feng Xia. 2024. FUGNN: Harmonizing Fairness and Utility in Graph Neural Networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024*. 2072–2081. doi:10.1145/3637528.3671834
- [17] Rita P. Moro, S. and P. Cortez. 2014. Bank Marketing. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5K306>.
- [18] Brodie Oldfield, Sevvandi Kandanaarachchi, Ziqi Xu, and Mario Andrés Muñoz. 2025. An Item Response Theory-based R module for Algorithm Portfolio Analysis. *SoftwareX* 31 (2025), 102239. doi:10.1016/J.SOFTX.2025.102239
- [19] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. 5680–5689. <https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526ffbebd39ab038d1cd7-Abstract.html>
- [20] ProPublica. 2016. ProPublica COMPAS Recidivism Dataset. <https://github.com/propublica/compas-analysis>
- [21] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*. 2239–2248. doi:10.1145/3219819.3220046
- [22] Hilde J. P. Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research* 24 (2023), 257:1–257:8. <https://jmlr.org/papers/v24/23-0389.html>
- [23] Ziqi Xu, Sevvandi Kandanaarachchi, Cheng Soon Ong, and Eirini Ntouts. 2025. Fairness Evaluation with Item Response Theory. In *Proceedings of the ACM on Web Conference 2025, WWW 2025*. 2276–2288. doi:10.1145/3696410.3714883
- [24] Ziqi Xu, Jixue Liu, Debo Cheng, Jiuyong Li, Lin Liu, and Ke Wang. 2023. Disentangled Representation with Causal Constraints for Counterfactual Fairness. In *Advances in Knowledge Discovery and Data Mining - 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023*, Vol. 13935. 471–482. doi:10.1007/978-3-031-33374-3_37
- [25] Ziqi Xu, Chenglong Ma, Yongli Ren, Jeffrey Chan, Wei Shao, and Feng Xia. 2025. Towards Better Evaluation of Recommendation Algorithms with Bi-directional Item Response Theory. In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025*. 1455–1459. doi:10.1145/3701716.3715540
- [26] Zhenlong Xu, Ziqi Xu, Jixue Liu, Debo Cheng, Jiuyong Li, Lin Liu, and Ke Wang. 2022. Assessing Classifier Fairness with Collider Bias. In *Advances in Knowledge Discovery and Data Mining - 26th Pacific-Asia Conference, PAKDD 2022*, Vol. 13281. 262–276. doi:10.1007/978-3-031-05936-0_21
- [27] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, Vol. 28. 325–333. <http://proceedings.mlr.press/v28/zemel13.html>
- [28] Sichen Zhao, Wei Shao, Jeffrey Chan, Ziqi Xu, and Flora Salim. 2025. FairDRL-ST: Disentangled Representation Learning for Fair Spatio-Temporal Mobility Prediction. *arXiv preprint arXiv:2508.07518* (2025).