ABCFair: an Adaptable Benchmark approach for Comparing Fairness methods

MARYBETH DEFRANCE, MAARTEN BUYL, and TIJL DE BIE, Ghent University, Belgium

Numerous methods have been implemented that pursue fairness with respect to sensitive features by mitigating biases in machine learning. Yet, the problem settings that each method tackles vary significantly, including the stage of intervention, the composition of sensitive features, the fairness notion, and the distribution of the output. Even in binary classification, these subtle differences make it highly complicated to benchmark fairness methods, as their performance can strongly depend on exactly how the bias mitigation problem was originally framed.

Hence, we introduce ABCFair, a benchmark approach which allows adapting to the desiderata of the real-world problem setting, enabling proper comparability between methods for any use case. In this extended abstract provide a summary from the results of applying ABCFair to a range of pre-, in-, and postprocessing methods on both large-scale, traditional datasets and on a dual label (biased and unbiased) dataset to sidestep the fairness-accuracy trade-off.

Keywords: Fairness, Benchmark, Comparability, Bias Mitigation

Reference Format:
MaryBeth Defrance, Maarten Buyl, and Tijl De Bie. 2025. ABCFair: an Adaptable Benchmark approach for Comparing Fairness methods. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 6 pages.

1 Introduction

Fairness has become a firmly established field in AI research as the study and mitigation of algorithmic bias. Thus, the range of methods that pursue AI fairness is now broad and varied [8, 26]. Many have been implemented in large toolkits such as *AIF360* [2], *Fairlearn* [30], *Aequitas* [28], or in libraries with a narrower focus like *Fair Fairness Benchmark* (FFB) [15], *error-parity* [9], and *fairret* [5].

So, which of these methods performs 'best'? Benchmarks in the past [3, 9, 14, 15, 17, 18, 23] tend to search for the best method *per dataset*. We argue for a more thorough benchmarking approach that can account for desiderata similar to those seen in a real-world context, allowing for proper comparison.

Moreover, prior work commonly observes a trade-off: the more fair a model, the less accurate its predictions. Unsurprising, as fairness is typically pursued in settings where the training data *and the evaluation data* are both assumed to be biased [4]. Removing bias from predictions degrades of the biased accuracy measure [31]. However, theoretical work has shown that this is not necessarily the case when evaluating on less biased data [13, 31].

Authors' Contact Information: MaryBeth Defrance, marybeth.defrance@ugent.be; Maarten Buyl, maarten.buyl@ugent.be; Tijl De Bie, Ghent University, Ghent, Belgium, tijl.debie@ugent.be.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

Contributions. We formalize four types of desiderata that could arise in real-world classification problems. These include **1)** the stage where the method intervenes, **2)** the composition of sensitive groups, **3)** the exact definition of fairness, and **4)** the distribution that is expected from the model output.

We distinguish our evaluation approach for two types of datasets. For large-scale, traditional datasets (where the labels are often biased) such as the folktables datasets [12], we compute large sets of tables that allow for an easy lookup of the best performing model for a certain fairness level. Yet, we also consider a *dual* label dataset, which contains both biased and unbiased labels for evaluation [24]. Such datasets allow us to train a method with biased labels (simulating real-world settings), while still evaluating accuracy and fairness over less biased labels.

We briefly discuss our results in this extended abstract. The full validation of the approach on 10 methods, 7 fairness notions, 3 formats of sensitive features, and 2 output formats and detailed information on the ready-to-use pipeline can be found in the full paper [10].

Related Work. The novelty of our benchmarking approach is found in the comparability challenges we identify in Sec. 2. A quantitative comparison in Tab. 1 validates that our benchmark's coverage of methods and fairness notions is on par with prior work.

2 Comparability Challenges in Fairness Benchmarking

Out of all machine learning tasks, AI fairness research has been mainly focused on binary classification [8, 26]. Our discussion and benchmark therefore also focus on binary classification, though the underlying concepts apply more broadly. We discuss four desiderata below: *Stage of Intervention*, *Composition of Sensitive Features*, *Incompatibility of Fairness Notions*, and *Distributions of the Output*.

Stage of Intervention. Surveys partition fairness methods in three types [26]: *preprocessing*, *inprocessing*, and *postprocessing*. Preprocessing methods modify the training distribution to remove clear biases, such as undesired correlations between the sensitive features and the label [6, 19]. Inprocessing methods modify model training, e.g. imposing a constraint during training [32]. Postprocessing methods only modify the output, e.g. separate classification thresholds for each demographic group [9].

Table 1. Quantitative comparison of fairness benchmarks.

Benchmark	Dataset labels		Methods			Fairness notions	Multiple sens. feat.
	Dual	Biased	Pre-	In-	Post-		
L. Cardoso et al. [23]	✓	✗	5	0	0	3	✗
Friedler et al. [14]	✗	✓	1	3	0	7	✓
Biswas and Rajan [3]	✗	✓	2	2	3	6	✓
Fairea [17]	✗	✓	3	2	3	2	✗
Islam et al. [18]	✗	✓	0	2	0	4	✓
Fair Fairness Benchmark [15]	✗	✓	0	6	0	6	✗
Cruz and Hardt [9]	✗	✓	2	3	1	3	✓
ABCFair (ours)	✓	✓	4	5	1	7	✓

Composition of Sensitive Features. We distinguish three different combinations of sensitive attributes: binary, categorical, and parallel. Many methods use a *binary* format [16, 19, 22, 25, 27], thus they only consider an advantaged and a disadvantaged group. It is a rather coarse categorization in practice: the 'disadvantaged' group may contain some subgroups towards whom the bias is much less significant than others. Bias against the latter is then underestimated. A more practical composition is *categorical* which allows an arbitrary amount of disjoint demographic groups [1, 32]. It also allows for *intersections* of demographic categories, like black women [20]. Yet, categorical grouping can quickly lead to a deterioration of statistical power. A middle ground is a *parallel* format, i.e. considering multiple axes of demographic attributes independently [5]. It is therefore possible to detect if a model is i) biased against black people *or* ii) biased against women, but not for the subgroup of black women.

Incompatibility of Fairness Notions. Mathematically, bias is a broad concept that could refer to a pattern in the data, the procedural steps of an algorithm, or output of a model [26]. Defining fairness over the output is by far most popular, and a wealth of empirical and theoretical results have made it clear that such a definition can take on a wide range of mathematical forms [29], which are often mutually incompatible [11, 21].

Distribution of the Output. A final, common challenge to comparability is the distribution of the output, i.e. whether fairness is evaluated over *hard*, binary decisions or *soft* scores. Fairness definitions are almost always defined in terms of hard decisions [29], so standard libraries like Fairlearn [30] are designed with hard decisions in mind. However, soft scores may be more desireable in some real-world cases, for example if the final decision is deferred to a human decision-maker that can use the scoring of algorithm [7].

3 Experiments

We conduct two sets of experiments: one on the dual-label dataset and another on more conventional fairness datasets [12]. The dual-label dataset allows us to question the accuracy-fairness trade-off as is shown in Figure 1. We also challenge the reporting of bias mitigation methods on just the fairness-accuracy trade-off and introduce a novel method of reporting the performance of a bias mitigation method. This method is shown in Table 2, where a set of fairness violations k are chosen for which the best performance of a bias mitigation method is reported for which it does not exceed the allowed fairness violation.

From our experiments we gain several interesting results. The results of the experiments on the dual label dataset challenge the fairness-accuracy trade-off, as shown in Figure 1. The accuracy measured over the unbiased labels actually increases by applying bias mitigation methods on biased labels, thus contradicting the fairness-accuracy trade-off. We summarize the rest of the result from the main paper in the following 5 key findings:

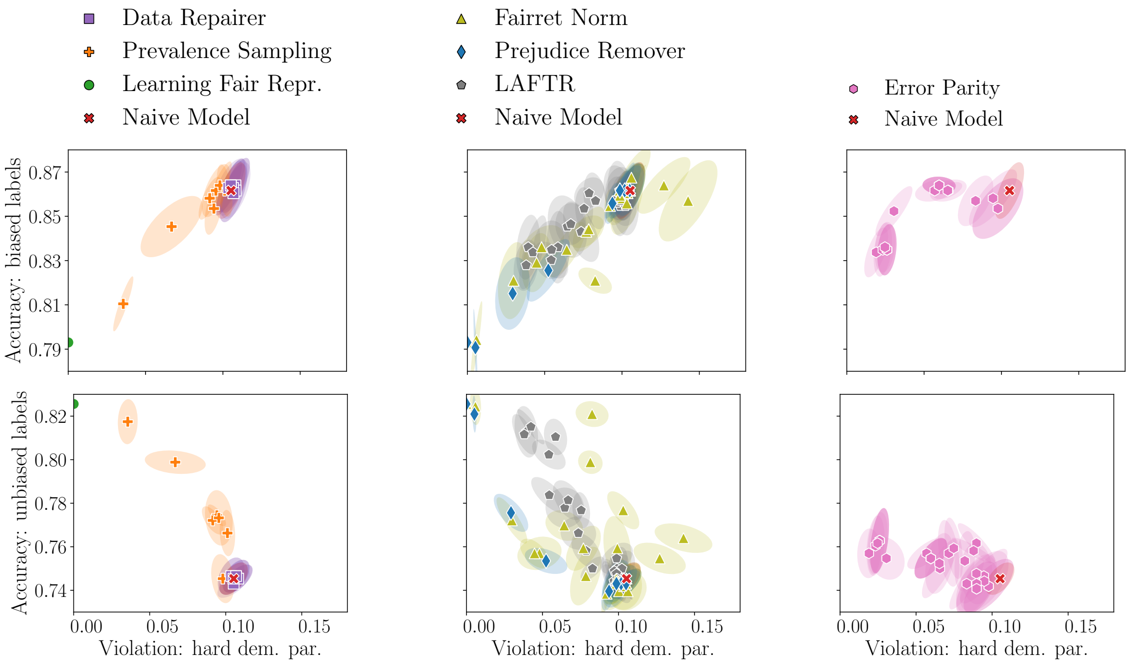
Key Finding 1: Methods that perform better on the traditional accuracy-fairness trade-off perform worse on unbiased labels.

Key Finding 2: Preprocessing methods struggle to ever obtain very low fairness violations.

Key Finding 3: More granular sens. feat. compositions lead to consistently larger fairness violations.

Key Finding 4: Preprocessing methods optimize for specific a fairness notion more efficiently than inprocessing, but those improve all fairness notions at once by improving on one.

Key Finding 5: Whether the output consists of hard or soft scores has a significant impact on trade-offs.



(a) Pre-processing methods

(b) In-processing methods

(c) Post-processing methods

Fig. 1. Fairness-accuracy trade-offs on the SchoolPerformance dataset trained on biased labels. The top row is evaluated on biased labels and the bottom row on unbiased. A marker is the mean test score over 5 random seeds, with a confidence ellipse.

Table 2. The maximal **AUROC** and standard error in %, for the fairness strength where the **(soft) dem. parity** violation is smaller than k. A naive model achieves AUROC = 81.2% and violation of Binary = 0.046, Intersectional = 0.48 and Parallel = 0.17.

	k	Data Repairer	Label Flipping	Learning Fair Repr.	Prevalence Sampling	Fairret Norm	Fairret KL_{proj}	LAFTR	Prejudice Remover	Expon. Gradient	Error Parity
Binary	0.03	80.2$_{\pm 0.1}$	**81.2$_{\pm 0.1}$**	-	**81.2$_{\pm 0.1}$**	**81.2$_{\pm 0.0}$**	**81.2$_{\pm 0.1}$**	80.8$_{\pm 0.0}$	**81.2$_{\pm 0.1}$**	70.5$_{\pm 0.1}$	69.5$_{\pm 0.1}$
	0.05	**81.2$_{\pm 0.1}$**	**81.2$_{\pm 0.1}$**	63.0$_{\pm 1.0}$	**81.3$_{\pm 0.1}$**	**81.2$_{\pm 0.1}$**	**81.2$_{\pm 0.1}$**	**81.3$_{\pm 0.1}$**	**81.2$_{\pm 0.1}$**	71.1$_{\pm 0.1}$	69.6$_{\pm 0.1}$
Inters.	0.35	78.8$_{\pm 0.1}$	-	63.0$_{\pm 1.0}$	80.4$_{\pm 0.1}$	80.8$_{\pm 0.0}$	**81.2$_{\pm 0.1}$**	80.1$_{\pm 0.2}$	80.2$_{\pm 0.1}$	70.5$_{\pm 0.1}$	68.5$_{\pm 0.1}$
	0.50	**81.2$_{\pm 0.1}$**	**81.2$_{\pm 0.1}$**	63.0$_{\pm 1.0}$	**81.3$_{\pm 0.1}$**	**81.2$_{\pm 0.1}$**	**81.2$_{\pm 0.1}$**	**81.3$_{\pm 0.1}$**	**81.2$_{\pm 0.1}$**	71.1$_{\pm 0.1}$	68.9$_{\pm 0.0}$
Parallel	0.12	-	-	63.0$_{\pm 1.0}$	-	79.9$_{\pm 0.0}$	**81.2$_{\pm 0.1}$**	74.6$_{\pm 1.6}$	79.2$_{\pm 0.1}$	66.3$_{\pm 0.9}$	68.0$_{\pm 0.1}$
	0.16	79.7$_{\pm 0.1}$	-	63.0$_{\pm 1.0}$	**81.2$_{\pm 0.1}$**	**81.2$_{\pm 0.0}$**	**81.2$_{\pm 0.1}$**	80.8$_{\pm 0.1}$	**81.2$_{\pm 0.1}$**	71.1$_{\pm 0.1}$	68.5$_{\pm 0.1}$

4 Conclusions

We introduced *ABCFair*, a novel benchmarking approach focused on the challenges of comparing fairness methods with different desiderata. After an extensive discussion of these challenges, we provide a configurable pipeline designed to address each of these challenges. We finally provide guidance on benchmarking fairness methods accordingly, covering a wide range of configurations.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

[2] 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. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

[3] Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, page 642–653, Virtual Event USA, November 2020. ACM.

[4] Maarten Buyl and Tijl De Bie. Inherent limitations of ai fairness. *Communications of the ACM*, 67(2):48–55, 2024.

[5] Maarten Buyl, MaryBeth Defrance, and Tijl De Bie. fairret: a Framework for Differentiable Fairness Regularization Terms. In *International Conference on Learning Representations*, 2024.

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[7] Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the conference on fairness, accountability, and transparency*, pages 309–318, 2019.

[8] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, July 2024.

[9] André Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. In *The Twelfth International Conference on Learning Representations*, 2024.

[10] MaryBeth Defrance, Maarten Buyl, and Tijl De Bie. ABCFair: an adaptable benchmark approach for comparing fairness methods. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[11] MaryBeth Defrance and Tijl De Bie. Maximal fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 851–880, 2023.

[12] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. How to be fair? a study of label and selection bias. *Machine Learning*, 112(12):5081–5104, December 2023.

[14] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA, January 2019. Association for Computing Machinery.

[15] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods. October 2023.

[16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[17] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairea: a model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, page 994–1006, Athens Greece, August 2021. ACM.

[18] Rashidul Islam, Shimei Pan, and James R. Foulds. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 586–596, Virtual Event USA, July 2021. ACM.

[19] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[20] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 100–109, 2019.

[21] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[22] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862, 2018.

[23] Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida, and Mohammed J. Zaki. A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, page 437–444, Honolulu HI USA, January 2019. ACM.

[24] Daphne Lenders and Toon Calders. Real-life Performance of Fairness Interventions - Introducing A New Benchmarking Dataset for Fair ML. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 350–357, Tallinn Estonia, March 2023. ACM.

[25] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[26] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, July 2022.

[27] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.

[28] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[29] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

[30] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research*, 24, 2023.

[31] Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

[32] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.