Proceedings of Machine Learning Research 279:96–118, 2024 Algorithmic Fairness Through the Lens of Metrics and Evaluation

Different Horses for Different Courses: Comparing Bias Mitigation Algorithms in ML

Prakhar Ganesh* PRAKHAR.GANESH@MILA.QUEBEC
McGill University, Mila

Usman Gohar* UGOHAR@IASTATE.EDU
Iowa State University

Lu Cheng LUCHENG@UIC.EDU
University of Illinois Chicago

Golnoosh Farnadi FARNADIG@MILA.QUEBEC
McGill University, Mila

Editors: Miriam Rateike, Awa Dieng, Jamelle Watson-Daniels, Ferdinando Fioretto, Golnoosh Farnadi

Abstract

With fairness concerns gaining significant attention in Machine Learning (ML), several bias mitigation techniques have been proposed, often compared against each other to find the best method. These benchmarking efforts tend to use a common setup for evaluation under the assumption that providing a uniform environment ensures a fair comparison. However, bias mitigation techniques are sensitive to hyperparameter choices, random seeds, feature selection, etc., meaning that comparison on just one setting can unfairly favour certain algorithms. In this work, we show significant variance in fairness achieved by several algorithms and the influence of the learning pipeline on fairness scores. We highlight that most bias mitigation techniques can achieve comparable performance, given the freedom to perform hyperparameter optimization, suggesting that the choice of the evaluation parameters—rather than the mitigation technique itself—can sometimes create the perceived superiority of one method over another. We hope our work encourages future research on how various choices in the lifecycle of developing an algorithm impact fairness, and trends that guide the selection of appropriate algorithms.

Keywords: Fairness in ML; Responsible AI; Evaluation; Multiplicity

1. Introduction

Over the past decade, concerns about fairness and discrimination in Machine Learning (ML) systems have emerged as critical issues, driving extensive research into the development of fair ML practices, including mitigation algorithms and fairness criteria (Mehrabi et al., 2021; Gohar and Cheng, 2023; Barocas et al., 2023). This has led to emerging global AI regulation focused on mitigating discrimination in AI/ML systems, mandating the reporting of fairness metrics of algorithms in compliance with various anti-discrimination laws such as the disparate impact doctrine (Justice., 2023).

However, despite the regulatory efforts, recent research has increasingly shown that the fair ML pipeline suffers from instability and high variance in fairness measures, which

* Equal contribution

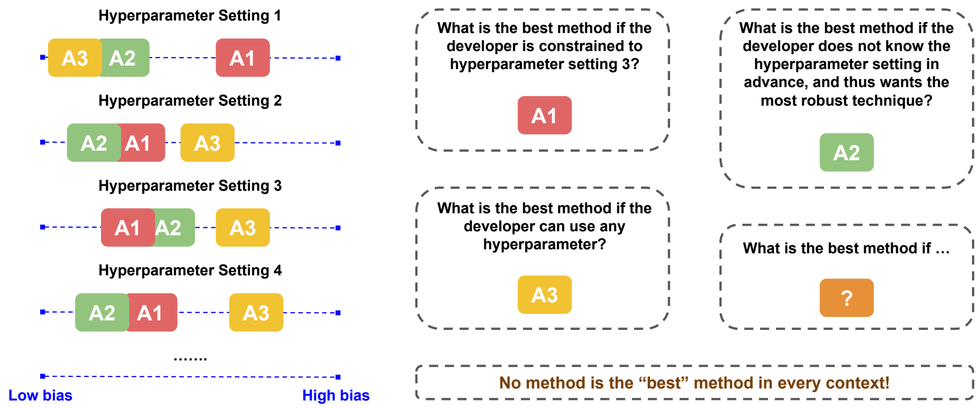

Figure 1: Motivation behind context-aware benchmarking of bias mitigation techniques, instead of using a uniform evaluation setup or finding the "best" technique.

can mask the underlying unfairness while creating an illusion of fairness (Black et al., 2023). For instance, recent work has pointed out how fairness measures vary across different training runs or between training and deployment, challenging the effectiveness, reliability, and utility of existing methods (Baldini et al., 2021; Black and Fredrikson, 2021; Friedler et al., 2019; Ganesh et al., 2023). Additionally, the multitude of mitigation techniques and fairness metrics further complicate accurate benchmarking. Therefore, from both a regulatory perspective and best practices, such variances must be taken into account to accurately represent the performance of these systems and fairness intervention methods.

While recent works have highlighted the issue of variance in fairness, existing fairness benchmarks predominately operate under a single identical training environment (e.g., hyperparameters, random seed, etc.) to ensure more accurate and fair comparisons (Han et al., 2023). However, this fails to consider the sensitivity of fairness to hyperparameter choices, which may mask the nuances of various bias mitigation techniques, favoring one over another (Dooley et al., 2024), as illustrated in Figure 1. We postulate that after accounting for the variance in fairness due to the hyperparameter choice, most bias mitigation methods achieve comparable performance. This further raises important questions about the one-dimensional nature of existing fairness evaluations. Is the "best" model simply the one that performs optimally under specific hyperparameter configurations, or should fairness assessments take a more holistic approach? For instance, how do we account for trade-offs between fairness, interpretability, stability, and resource constraints? Should fairness evaluations prioritize consistency over best performance, or should context-specific factors like deployment environments and real-world implications dictate the criteria for success? These questions highlight the need for more nuanced and multidimensional fairness benchmarks beyond traditional measures.

Contributions of our work. We show that bias mitigation algorithms are highly sensitive to several choices made in the learning pipeline. Consequently, without incorporating the broader context, a one-dimensional comparative analysis of these algorithms can create a false sense of fairness. We highlight that most mitigation algorithms can achieve comparable performance under appropriate hyperparameter optimization and, therefore, advocate going beyond the narrow view of the fairness-utility tradeoff to explore other factors that impact model deployment. We conduct a large-scale empirical analysis to support our claims. We hope our work inspires future research that explores the interplay between bias mitigation algorithms and the entire learning pipeline, rather than studying them in isolation, as has often been the case in the literature.

2. Related Work

Researchers have developed a range of mitigation techniques and notions to address unfairness in ML systems, targeting different stages of the ML pipeline, including pre-processing, in-processing, and post-processing methods (Mehrabi et al., 2021; Pessach and Shmueli, 2022; Gohar et al., 2024). Pre-processing balances data distribution across protected groups, reducing variance, while post-processing modifies model outputs without accessing internal algorithms, ideal for black-box models. In contrast, in-processing methods impose fairness constraints on the model or modify the objective function to mitigate bias. There is inherent randomness in the training process, which, while vital for convergence and generalization (Noh et al., 2017), can be a source of high-fairness variance. Moreover, these techniques also have control parameters (for example, regularization weight) that must be optimized for the training data, further introducing variance (Bottou, 2012). In this work, we limit our focus on in-processing techniques due to the high variance exhibited by these methods (Baldini et al., 2021; Black and Fredrikson, 2021; Friedler et al., 2019; Ganesh et al., 2023; Perrone et al., 2021).

There is increasing evidence in the literature of the instability of fairness metrics associated with non-determinism in model training and decisions (Black et al., 2022; Baldini et al., 2021; Friedler et al., 2019). This includes identical training environments with small changes such as different random seeds (Black and Fredrikson, 2021), sampling (Ganesh et al., 2023), hyperparameter choices (Ganesh, 2024; Gohar et al., 2023), or even differences in train-test-split (Friedler et al., 2019), which have been shown to lead to meaningful differences in group fairness performance across different runs. The issue arises when this instability is used to report higher fairness performance, which calls into question the effectiveness of fairness assessments (Black et al., 2024). A few recent works have explored this issue from a regulatory fairness assessment standpoint where, inadvertently or deliberately, an actor can misrepresent the fairness performance. For instance, a parallel work by Simson et al. (2024) proposes using multiverse analysis to keep track of all possible decision combinations for data processing and its impact on model fairness. We build on a similar argument, focusing instead on decisions made during the algorithm design, highlighting the instability of various fairness intervention methods.

Several attempts have been made to benchmark bias mitigation algorithms (Bellamy et al., 2019; Bird et al., 2020; Han et al., 2023), including those conducted by papers proposing new mitigation techniques (Kamishima et al., 2012; Li et al., 2022; Madras et al., 2018; Adel et al., 2019; Zhang et al., 2018). As highlighted by Han et al. (2023), every paper that proposes a new bias mitigation algorithm often introduces its own experimental setup. This lack of standardization can make it difficult to compare different algorithms effectively. To overcome this issue, Han et al. (2023) created the FFB benchmark, offering a comprehensive evaluation over a wide range of datasets and algorithms to allow fair comparison. While FFB provides an excellent foundation to compare and benchmark bias mitigation algorithms, its attempts to find an algorithm that provides the best fairness-utility tradeoff (which they claim is HSIC) comes with two important caveats, (a) FFB only considers a single hyperparameter setting, which, while useful for standardization, overlooks the sensitivity of fairness scores to hyperparameter choices (Perrone et al., 2021), and (b) FFB aggregates results across multiple datasets, which can obscure the nuances of performance on individual datasets. In our work, we demonstrate that the fairness variability due to hyperparameter choices can often mask unfairness and raise concerns about existing evaluation techniques.

3. Variance in Bias Mitigation

In this section, we argue that benchmarking under different settings can reveal trends that are lost when sticking to only a single standardized hyperparameter setting or aggregating results across multiple datasets, as done by Han et al. (2023). Thus, a more nuanced approach to benchmarking bias mitigation techniques is needed to capture the strengths and limitations of various algorithms.

3.1. Experiment setup

For our experiments, we borrow from Han et al. (2023), using their open-source code [1]. We focus on the seven tabular datasets and the seven bias mitigation algorithms used in their benchmark (not including the standard empirical risk minimization without fairness constraints). In addition to varying the random seed for training and the control parameter for bias mitigation as done in the benchmark, we also vary the batch size, learning rate, and model architecture to explore several different hyperparameter settings. More details on the experiment setup are delegated to Appendix A.

3.2. Case Study: Adult Dataset

We start with a case study on the Adult dataset (Becker and Kohavi, 1996) and show the changing trends across different hyperparameters. We plot fairness as demographic parity and utility as accuracy, studying the fairness-utility tradeoff across various settings in Figure 2. Additional discussions on other datasets and fairness metrics are present in Appendix B.

We first observe the absence of a clear winner among bias mitigation algorithms; no single method consistently and significantly outperforms the others. While HSIC achieves better tradeoffs under the hyperparameter setting used by Han et al. (2023), other techniques such as PRemover and DiffDP perform equally well – or even better – under other hyperparameter settings. Thus, a comparative analysis limited to just one combination of hyperparameters fails to capture the competitive performance of other algorithms.

Another interesting set of results comes from the hyperparameter setting with a large batch size and a high learning rate. These settings are particularly relevant in scenarios where the emphasis is on rapid convergence and minimizing the number of training steps, even at the cost of performance, for instance, during rapid prototyping, edge computing with

1. https://github.com/ahxt/fair_fairness_benchmark

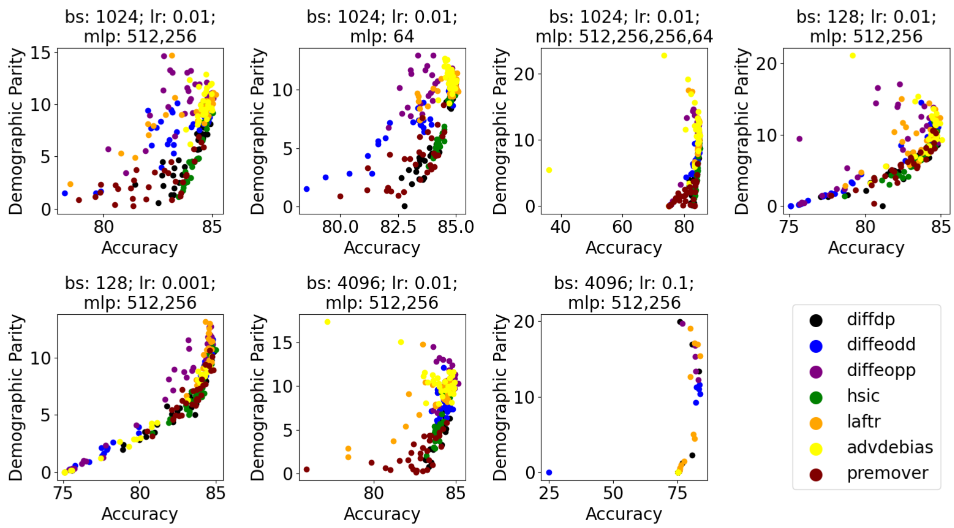

Figure 2: Fairness-utility (demographic parity-accuracy) tradeoff across various settings for the Adult dataset. Each graph represents a different combination of hyperparameters, and each dot in the graph represents a separate training run. Multiple dots for the same mitigation algorithm in the same graph represent runs with changing random seeds and control parameters.

limited compute cycles or federated learning. Most bias mitigation methods that performed well in other settings fail to even converge under these conditions. Instead, methods like AdvDebias and LAFTR, which did not stand out in other settings, provide good fairness scores when trained under these constraints.

Our findings highlight the importance of considering a diverse range of choices in the learning pipeline when evaluating bias mitigation techniques, as different algorithms excel under different settings. Further interesting trends can be extracted from the comparative analysis of various algorithms across different hyperparameter settings, left for future work. We now turn to a similarly nuanced comparative analysis of bias mitigation techniques, focusing on changing trends across datasets.

3.3. Changing Trends Across Datasets

In the previous section, we observed different techniques perform better than others under changing settings within a single dataset. We now extend this observation to multiple datasets to show that even the trends across different datasets can vary significantly, and the choice of combining results from all datasets, as done by Han et al. (2023), can obscure these trends. We record the fairness (demographic parity) and utility (accuracy) across different datasets and bias mitigation algorithms in Figure 3. Results for other fairness metrics are present in Appendix C.

We begin again with HSIC, which Han et al. (2023) identifies as the algorithm offering the best tradeoff overall. While HSIC does well on most datasets, it is noticeably not the

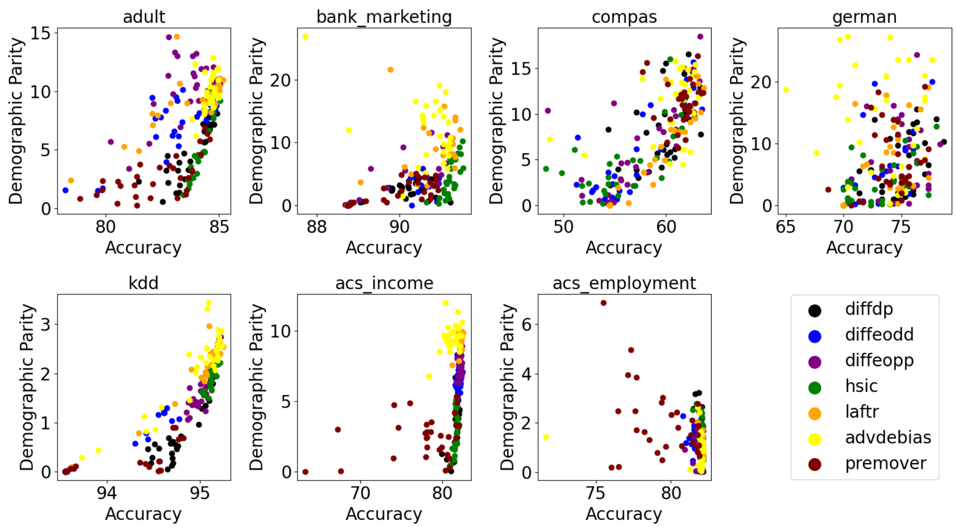

Figure 3: Fairness-utility (demographic parity-accuracy) tradeoff across various datasets, under their default hyperparameters. Each dot in the graph represents a separate training run with changing random seeds and control parameters.

unanimous top choice for both the COMPAS and German datasets. Interestingly, both of these are small datasets with smaller batch sizes used for training. Given that HSIC relies on pairwise similarity in a batch to estimate dependence, its superiority with larger batch sizes but its lackluster performance with smaller batch sizes is not surprising. Most of the datasets analyzed by Han et al. (2023) were big datasets with large batch sizes. Thus, combining results from multiple datasets overshadowed the trends present only in smaller datasets, effectively hiding HSIC's shortcomings.

Another intriguing trend is present in the performance of LAFTR, the adversarial representation learning-based bias mitigation technique. Generally, LAFTR underperforms across various datasets, offering poor tradeoffs. However, it performs surprisingly well on the COMPAS dataset, providing competitive tradeoffs against other techniques. This anomaly may be linked to the data preprocessing step adopted by Han et al. (2023), where all categorical features are converted into one-hot encodings. Since LAFTR relies on representation learning at its core, this explosion in the number of features can make the representation learning task more difficult, thus hurting the eventual mitigation attempts. Notably, the COMPAS dataset, where LAFTR excels, contains the least number of categorical features and, consequently, the smallest input size compared to other datasets. Thus, LAFTR's underperformance might not be simply due to the algorithm itself but rather the choice of input feature representation used.

Finally, we also observe fairness metric-specific trends that affect the comparative analysis between different algorithms. For many datasets like Adult, COMPAS, and KDD, there is a clear tradeoff between demographic parity and accuracy, which isn't surprising since the two are not aligned. In these datasets, we find different mitigation techniques occupy distinct positions in the tradeoff curves. Thus, the choice of the mitigation technique, therefore, depends on the stakeholders' level of willingness to trade utility for fairness. Conversely, many other datasets do not have these apparent tradeoffs. Here, we find that the results are mixed, with highly overlapping trends across different mitigation algorithms.

4. Comparisons Beyond the Fairness-Utility Tradeoff

In the previous section, we discussed the limitations of a generic comparative analysis of bias mitigation algorithms that don't take into account the nuances of the entire learning pipeline. We now turn to a practical concern: *choosing the appropriate algorithm*.

We begin this section by first showing that given the opportunity to perform hyperparameter optimization, various mitigation algorithms can provide competitive models. We then discuss how, given the lack of appropriate differentiation between these algorithms in their fairness-utility tradeoff, the selection of the appropriate algorithm can prioritize other factors, like algorithm runtime, complexity, potential robustness, theoretical guarantees, etc. Consequently, selecting the appropriate algorithm and the resulting model would involve balancing these additional considerations, rather than solely focusing on just the best fairness-utility tradeoff.

4.1. Most Mitigation Algorithms are Competitive

As we saw in Figure 2 (and Appendix B), different algorithms do well under varying settings. In several real-world applications, many of these choices are flexible, and hyperparameter optimization plays an important role in model selection. Thus, when comparing different algorithms, it is important to focus on evaluating the best-performing models from each algorithm, as these are the models that would be deployed if those algorithms were used.

To perform this comparison, we only filter the models at the Pareto front for various algorithms after searching through different hyperparameters and random seeds collected in Figure 4. Trends for other fairness metrics are present in Appendix D. We find that several algorithms can provide competitive tradeoffs for almost every dataset. For instance, DiffDP, PRemover, and HSIC demonstrate excellent fairness-utility tradeoffs for the Adult dataset, while all seven bias mitigation algorithms exhibit competitive tradeoffs on the German dataset. With multiple algorithms showing similar tradeoffs, it becomes evident that simply evaluating fairness-utility tradeoffs is insufficient when choosing the most suitable bias mitigation technique. We explore these considerations further in the next section.

4.2. Choosing the Right Mitigation Technique

When several bias mitigation algorithms provide similar tradeoffs, selecting one can be challenging. In such cases, additional factors must be considered, such as the specific requirements of the task, the deployment environment, the stakeholders' expectations, etc. Here, we provide some examples of comparisons beyond the fairness-utility tradeoff that can help choose an appropriate algorithm.

Runtime: An algorithm's runtime can be a crucial factor when comparing bias mitigation techniques. Even minor differences in runtime might become relevant when multiple runs of the same algorithm are needed, for instance, to perform hyperparameter optimization.

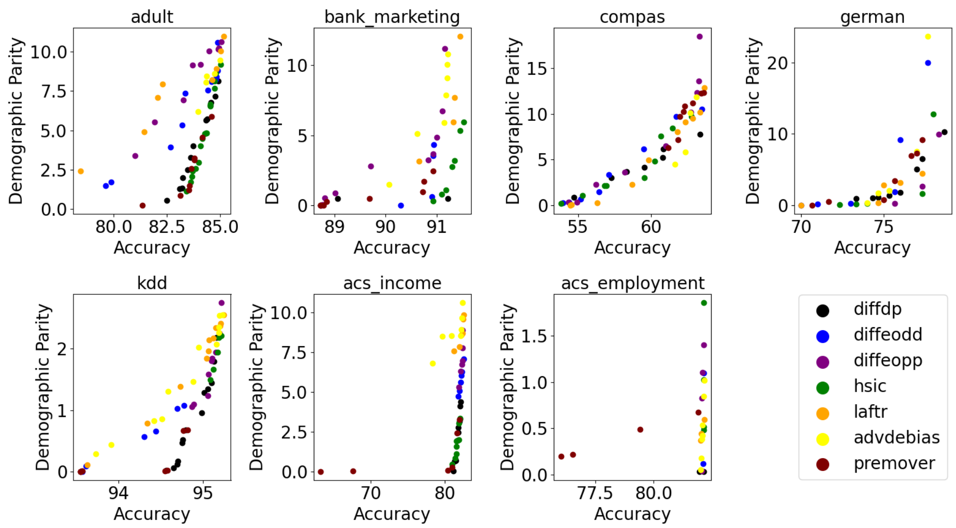

Figure 4: Pareto front of the fairness-utility (demographic parity-accuracy) tradeoff across various datasets. Each dot represents a separate training run on the pareto front with changing hyperparameters, random seeds and control parameters.

Our results, detailed in Table 1, reveal interesting trends in training runtime across various algorithms. We find the algorithms HSIC, LAFTR, PRemover, DiffEOdd, and DiffEOpp to be quite expensive, while in contrast, algorithms DiffDP and AdvDebias offer runtime comparable to the standard empirical risk minimization. Considering the competitive tradeoffs achieved by DiffDP, in addition to the lower runtime, it emerges as an appropriate choice for settings where computational efficiency is critical, surpassing other well-performing but slower methods like HSIC and PRemover.

Dataset	**Runtime (rounded to 5s intervals)**							
	ERM	**DiffDP**	**DiffEOdd**	**DiffEOpp**	**HSIC**	**LAFTR**	**PRemover**	**AdvDebias**
Bank	1m 15s	1m 15s	1m 45s	1m 45s	1m 50s	1m 46s	1m 50s	1m 25s
German	30s	30s	35s	35s	40s	35s	40s	30s
Adult	1m 40s	1m 40s	1m 45s	1m 45s	2m 0s	1m 50s	2m 0s	1m 40s
COMPAS	30s	30s	30s	30s	30s	30s	30s	30s
KDDCensus	6m 45s	6m 50s	10m 40s	10m 40s	10m 5s	10m 0s	9m 50s	6m 50s
ACS-I	7m 10s	7m 10s	9m 50s	9m 50s	9m 50s	9m 30s	10m 0s	7m 50s
ACS-E	13m 40s	13m 45s	15m 40s	15m 40s	16m 20s	15m 50s	16m 10s	13m 40s
ACS-P	7m 20s	7m 35s	9m 50s	9m 50s	9m 40s	10m 5s	10m 5s	7m 40s
ACS-M	4m 40s	4m 45s	6m 5s	6m 5s	6m 0s	6m 10s	6m 0s	4m 50s
ACS-T	7m 30s	7m 30s	10m 5s	10m 5s	10m 15s	10m 0s	10m 20s	8m 0s

Table 1: Training runtime of mitigation algorithms under default hyperparameters.

Theoretical Guarantees and Procedural Requirements: Another important consideration when selecting the appropriate algorithm is the theoretical guarantees that some techniques can offer. For instance, while adding regularizers to the training objective can be useful, it does not provide any form of guarantee for the model's final fairness scores. In contrast, methods like HSIC and LAFTR can provide theoretical bounds on the fairness of the final model, albeit limited to only simpler models (Li et al., 2022; Madras et al., 2018).

Furthermore, the deployed models may need to comply with specific procedural requirements, which can influence the choice of the mitigation algorithm. For instance, one might need to choose between algorithms focusing on outcome fairness (such as DiffDP, DiffEOpp, DiffEOdd) versus those focusing on process fairness (such as HSIC, LAFTR, PRemover, AdvDebias). The specific requirements of the application can dictate the choice of the algorithm, looking beyond the tradeoffs it can provide.

Multiplicity and Arbitrariness: Model multiplicity refers to the existence of a set of good models, which have similar performance but differ in their predictions for individuals (Marx et al., 2020; Black et al., 2022). Existing works have shown that bias mitigation can exacerbate multiplicity concerns, leading to arbitrariness in individual-level predictions (Long et al., 2024). However, the degree of multiplicity introduced can vary depending on the mitigation algorithm used. Following Long et al. (2024), we define the set of competing models as models with similar accuracy under ERM and record multiplicity using ambiguity (Marx et al., 2020), which is the fraction of data points whose predictions change across different models within the set of good models, in Table 2.

Dataset	**Ambiguity**							
	ERM	**DiffDP**	**DiffEOdd**	**DiffEOpp**	**HSIC**	**LAFTR**	**PRemover**	**AdvDebias**
Bank	0.15	0.16	0.16	0.18	0.17	0.19	0.15	0.26
German	0.55	0.57	0.57	0.63	0.55	0.59	0.60	0.87
Adult	0.17	0.30	0.37	0.42	0.28	0.32	0.34	0.47
COMPAS	0.93	0.99	1.0	1.0	1.0	1.0	0.92	0.99
KDDCensus	0.04	0.06	0.04	0.06	0.05	0.06	0.04	0.09
ACS-I	0.26	0.35	0.38	0.35	0.38	0.32	0.75	0.49
ACS-E	0.14	0.20	0.30	0.21	0.26	0.20	0.49	0.37
ACS-P	0.27	0.33	0.38	0.39	0.34	0.45	0.32	0.69
ACS-M	0.26	0.29	0.27	0.29	0.31	0.38	0.20	0.61
ACS-T	0.70	0.81	0.88	0.80	0.79	0.91	0.92	0.90

Table 2: Ambiguity scores of mitigation algorithms under default hyperparameters.

Unsurprisingly, most bias mitigation techniques exhibit higher ambiguity than ERM, which aligns with the observations made by Long et al. (2024). However, an interesting exception is the PRemover algorithm, which achieves remarkably low ambiguity scores across many datasets, distinguishing it from other algorithms. Strikingly, at the same time, PRemover also shows significantly high ambiguity in several other datasets, highlighting its behavior on both extremes. Thus, for certain datasets, PRemover could be considered a superior choice compared to other methods like HSIC and DiffDP, which, while offering similar trade-offs, tend to introduce more arbitrariness into the model. In contrast to

PRemover, the AdvDebias algorithm consistently results in very high ambiguity scores, making it a poor choice in contexts where minimizing arbitrariness is crucial.

In this section, we showed several examples of additional factors to consider when selecting an algorithm for a specific use case. Naturally, this list is not exhaustive, as additional considerations may arise depending on the specific application context. The objective of our study was to emphasize the lack of distinction between mitigation algorithms that focus solely on the fairness-utility tradeoff and the importance of choosing algorithms that offer additional advantages beyond this tradeoff. With these results, we hope to move away from the narrative of a single optimal bias mitigation technique and emphasize the need for context-dependent comparative analysis.

5. Discussion

In this paper, we underscore the limitations of current fairness benchmarking practices that rely on uniform evaluation setups. We demonstrate that hyperparameter optimization can yield similar performance across different techniques, raising questions about the effectiveness of existing benchmarks and the criteria for selecting appropriate fairness algorithms.

Context-dependent evaluation. We argue that the current one-dimensional approach to fairness evaluation may be insufficient. Given the high variability in fairness scores, relying on a single run or, conversely, simply aggregating multiple training runs, both common practices across different dimensions, may not always provide an appropriate comparison of bias mitigation techniques.

For example, when models are too large and retraining is impractical, choosing fairness interventions that prioritize stability and consistent scores may be more appropriate. On the other hand, if sufficient computational resources exist to explore hyperparameter options, selecting the best-performing model might be more valid. Additionally, explainability, runtime, and scalability constraints can significantly impact the choice of fairness assessments. Ultimately, the method of comparing algorithms depends on the context. However, in all cases, it is crucial to consider the variability introduced by hyperparameter tuning.

Future work. Our experiments were limited to in-processing techniques in bias mitigation. In the future, we plan to explore a broader range of methods, including pre and post-processing. Moreover, we have not explored the potential presence of consistent fairness trends for different hyperparameter choices covered in the experiments. It would be interesting to investigate whether we can identify patterns that guide our decisions to choose better hyperparameter settings for various bias mitigation algorithms. Finally, while evidence in the literature would suggest similar trends exist even with hyperparameters in other parts of the pipeline, for instance, data processing (Simson et al., 2024), our empirical results are limited to hyperparameter choices during training. Further work on a large-scale study of the impact of various choices in the lifetime of an algorithm design is needed.

Acknowledgments

Funding support for project activities has been partially provided by the Canada CIFAR AI Chair, FRQNT scholarship, and NSERC discovery award. We also express our gratitude to Compute Canada and Mila clusters for their support in providing facilities for our evaluations. Lu Cheng is supported by the National Science Foundation (NSF) Grant #2312862, NIH #R01AG091762, and a Cisco gift grant.

References

Census-Income (KDD). UCI Machine Learning Repository, 2000. DOI: https://doi.org/10.24432/C5N30T.

Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2412–2420, 2019.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. Your fairness may vary: Pretrained language model fairness in toxic text classification. *arXiv preprint arXiv:2108.01250*, 2021.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

Emily Black and Matt Fredrikson. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 285–295, 2021.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

Emily Black, Rakshit Naidu, Rayid Ghani, Kit Rodolfa, Daniel Ho, and Hoda Heidari. Toward operationalizing pipeline-aware ml fairness: A research agenda for developing practical guidelines and tools. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–11, 2023.

Emily Black, Talia Gillis, and Zara Yasmine Hall. D-hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 602–615, 2024.

Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436. Springer, 2012.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Samuel Dooley, Rhea Sukthanker, John Dickerson, Colin White, Frank Hutter, and Micah Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. *Advances in Neural Information Processing Systems*, 36, 2024.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations*, pages 1–14, 2016.

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*, pages 329–338, 2019.

Prakhar Ganesh. An empirical investigation into benchmarking model multiplicity for trustworthy machine learning: A case study on image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4488–4497, 2024.

Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1789–1800, 2023.

Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. *arXiv preprint arXiv:2305.06969*, 2023.

Usman Gohar, Sumon Biswas, and Hridesh Rajan. Towards understanding fairness and its composition in ensemble machine learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1533–1545. IEEE, 2023.

Usman Gohar, Zeyu Tang, Jialu Wang, Kun Zhang, Peter L Spirtes, Yang Liu, and Lu Cheng. Long-term fairness inquiries and pursuits in machine learning: A survey of notions, methods, and challenges. *arXiv preprint arXiv:2406.06736*, 2024.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

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. In *International Conference on Learning Representations*. ICLR, 2023.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

United States Dept. Of Justice. Title vi legal manual, section vii: Proving discrimination – disparate impact., Oct 2023. URL https://www.justice.gov/crt/fcs/T6Manual7.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. Propublica compas analysis—data and analysis for 'machine bias.'. *https://github.com/propublica/compas-analysis*, 2016.

Zhu Li, Adrián Pérez-Suay, Gustau Camps-Valls, and Dino Sejdinovic. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *Pattern Recognition*, 132:108922, 2022.

Carol Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. Individual arbitrariness and group fairness. *Advances in Neural Information Processing Systems*, 36, 2024.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *Advances in neural information processing systems*, 30, 2017.

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.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5K306.

Hyeonwoo Noh, Tackgeun You, Jonghwan Mun, and Bohyung Han. Regularizing deep neural networks by noise: Its interpretation and optimization. *Advances in neural information processing systems*, 30, 2017.

Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 854–863, 2021.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Jan Simson, Florian Pfisterer, and Christoph Kern. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1305–1320, 2024.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.