Influence of Label and Selection Bias on Fairness Interventions

MAGALI LEGAST, Université catholique de Louvain, Belgium
TOON CALDERS, Universiteit Antwerpen, Belgium
FRANÇOIS FOUSS, Université catholique de Louvain, Belgium

Bias can be introduced in different ways in machine learning datasets, with the bias type influencing the effectiveness of fairness interventions. In this work, we model fair worlds and their biased counterparts by introducing controlled label and selection bias in real-life datasets with low discrimination. We then analyze the resulting prediction models, with or without bias mitigation. Our results provide some guidance on the use of reweighing, massaging and *Fairness Through Unawareness*, and show that other dataset characteristics also play a role on fairness intervention efficiency, calling for further research.

Keywords: Algorithmic fairness, Bias mitigation, Fair classification, Label bias, Selection bias, Artificial dataset

Reference Format:
Magali Legast, Toon Calders, and François Fouss. 2025. Influence of Label and Selection Bias on Fairness Interventions. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 6 pages.

1 Introduction

It is now well documented that machine learning models are subject to biases that can lead to discriminatory outputs [15, 16]. The field of algorithmic fairness has seen the development of many fairness metrics, and even more bias mitigation methods [8]. This prolific development of potential fairness interventions has however not been match with a thorough understanding of those methods. Interestingly, applying the same bias mitigation method in different contexts can lead to different results, both regarding model accuracy and fairness improvement [1, 6, 11]. It is thus needed to better understand how exactly the context affects the performances of fairness interventions. A few comparative studies have been made in that direction, such as [2, 6, 9, 13, 17]. Most of them share some limitations that are common in the field of fairness. The most prominent one in relation to this work is the use of datasets for which there is no precise control or understanding of their embedded biases. (See [4] for a discussion on biases in fairness datasets.) Further, the lack of unbiased ground truth leads to biased results and prevents to correctly evaluate how far the predictions are from a fair outcome [5, 14, 20], leading to an artificial fairness-accuracy trade-off. Those elements thus make it hard to draw accurate conclusions on the fairness intervention studied.

In parallel, [1, 5] show that the way bias was introduced in a dataset influences the effectiveness of different fairness interventions. Building on the theoretical findings in [5] as well as bias modelization and toy experiment in

Authors' Contact Information: Magali Legast, magali.legast@uclouvain.be, Université catholique de Louvain, Louvain-la-Neuve, Belgium; Toon Calders, toon.calders@uantwerpen.be, Universiteit Antwerpen, Antwerpen, Belgium; François Fouss, francois.fouss@uclouvain.be, Université catholique de Louvain, Louvain-la-Neuve, Belgium.

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

[1], this ongoing work further explores the relationship between bias types and mitigation methods. To counter the shortcomings explained above, we introduce artificial and controlled bias in a dataset that we consider to be our fair baseline, then study the models trained on the resulting data. We present here our findings about the mitigation of label and selection bias using the preprocessing techniques reweighting, massaging [10] and *Fairness Through Unawareness* (FTU).

2 Fairness framework and bias modelization

In our work, we consider the *Fair World Framework* [5], assuming the existence of a fair underlying world in which fairness is satisfied. The introduction of biases leads to a distorted version of this world. In this framework, a fair model not only respects certain fairness criteria, but is also accurate regarding to the fair world.

To model this theoretical fair world, we use baseline datasets that present very low discrimination level and make the assumption that they represent the fair world accurately. We then artificially introduce bias in them. This approach allows to modelize and analyze the difference between a fair world, represented by the original dataset, and the distorted unfair worlds, represented by the biased versions of that dataset. To represent the biasing process, we modelize two types of biases[1] : ***Label bias*** (or *measurement bias on label* [1]) occurs when the label in the training data doesn't correspond to that of the fair world [5]. Based on [1], we model it as a penalty on the unprivileged group : $P_L = L - \beta_m * S * scale + N$ where L is the original label values, β_m controls the bias intensity, S is the sensitive attribute (1 for unprivileged and 0 for favored), *scale* is half of the difference between the maximal and minimal values observed and N is a normally distributed random noise. We impose this biasing procedure on numerical values prior to label binarization. ***Selection bias*** (or *representation bias* [1]) occurs when the sample in the training data doesn't accurately represent the distribution of the fair world [5]. We model it as random undersampling of both the unprivileged group with positive label and the favored group with negative label. The parameter p_u controls the proportion of each group to be removed.

3 Experimental setup

We use the three following baseline datasets. **Student** [3] (or Student Performance) with 649 students, positive label *pass* (grade ≥ 10) and sensitive attribute *sex* with boys as unprivileged (41% of data). The original Statistical Parity Difference (SPD) is -0.057 for a base rate of 0.85. **OULADstem** is a subset of the Open University Learning Analytics Dataset [12] with 7040 students in a specific STEM course, and using features described in [18]. The sensitive attribute is *gender* with men as unprivileged (81.7% of data). We remove instances with missing values and first occurrences of duplicated students (who had repeated the course). The positive binary labels are *Pass* and *Distinction* and negative ones are *Fail* and *Withdrawn*. The SPD is -0.025 and base rate 0.48. **OULADsocial** is another OULAD subset with 7632 students (11.6% of unprivileged men) in a specific social science course. It is preprocessed like OULADstem. The SPD is -0.026 and base rate 0.49.

We create distorted versions of those datasets, introducing either label or selection bias with varying intensity (controlled by β_m and p_u). For each of those datasets, we train fairness agnostic models and models mitigated using respectively reweighting [10], massaging [10], and FTU. We report the mean values of cross-validation with 10 folds for the label bias scenarios and 5 folds for selection bias. The test sets used are always unbiased.

[1]The different bias types can be modelized in different ways, so our results should not be generalized to other manifestations of one bias type.

4 Results

Our results[2] are shown in Figure 1. We analyze the models with accuracy (proportion of True Positive), SPD (group metric based on predicted outcome), Equalized Odds Difference (EOD) [7] (based on group-conditioned accuracy [6]) and Generalized Entropy Index (GEI) [19] (individual metric based on accuracy). Overall, we observe that the effect of both bias introduction and mitigation is not homogeneous across the datasets. The effect of bias is thus also sensitive to the baseline datasets characteristics, such as the proportion of the unprivileged group.

Reweighing has a very positive effect on selection bias reduction, also being the best method on both OULAD datasets. Expectedly, the highest gain is on SPD, the metric it aims to optimize. The method also improves accuracy in most cases, except for a few exceptions with very strong bias. The results for label bias mitigation are more heterogeneous. The positive effect on group metrics, obtained in most cases, is sometimes at the cost of individual fairness (GEI increase or accuracy decrease). Out of the three methods, it has the worst fairness metrics results.

Massaging applied on label bias gives heterogeneous results on accuracy improvement. The effect on the SPD and EOD is however very positive, largely outperforming both other methods in the case of Student. The results for GEI are more nuanced, with positive impact for Student and OULADstem, but the group fairness gain is at the expense of individual fairness for OULADsocial. The mitigation of selection bias doesn't work as well, often causing a decrease in accuracy, thus deviating from the fair world. This negative impact can also be observed on group fairness metrics, except for Student, and for individual metric when the discriminated group is a minority (OULADsocial). Massaging is thus by far the worst for selection bias mitigation.

FTU overall closely follows the good performances of reweighing when addressing selection bias. The impact is also mostly positive against label bias, except for OULADstem, where the majority group is disadvantaged and the biasing is thus very strong. This echos [1, 5] where FTU is considered positively to address label bias when the label is independent from the sensitive attribute in the original data.

5 Discussion and conclusion

Our work studied the impact of bias types on the efficiency of fairness intervention, through experiments involving label and selection bias and the preprocessing methods reweighing, massaging and FTU. By using datasets that represent the fair world and in which controlled bias was introduced, our approach avoided the use of criticized and biased test sets that lead to biased metric values and a fairness-accuracy trade-off. Our results encourage the use of reweighing to mitigate selection bias in various contexts, while massaging should be avoided. Label bias, on the other hand, is better addressed by massaging, even though other factors have a significant impact.

Overall, we could observe that the efficiency of the bias mitigation methods varies according to different factors. First, the type of bias, with performance changing for label and selection bias, confirming the results in [1, 5]. Second, the dataset characteristics, including the proportion of unprivileged and favored groups. Third, the bias intensity, with the same method sometimes increasing bias for lower bias levels while reducing it for higher ones.

These findings call for further research to better understand the effect of these elements on the efficiency of fairness intervention. Future work should also address the effect of combined bias types, which are closer to real-life scenario, and the possibility to cumulate fairness interventions to better address them.

[2]Further results and the experiment code can be found at https://github.com/Magalii/ControlledBias/tree/EWAF2025

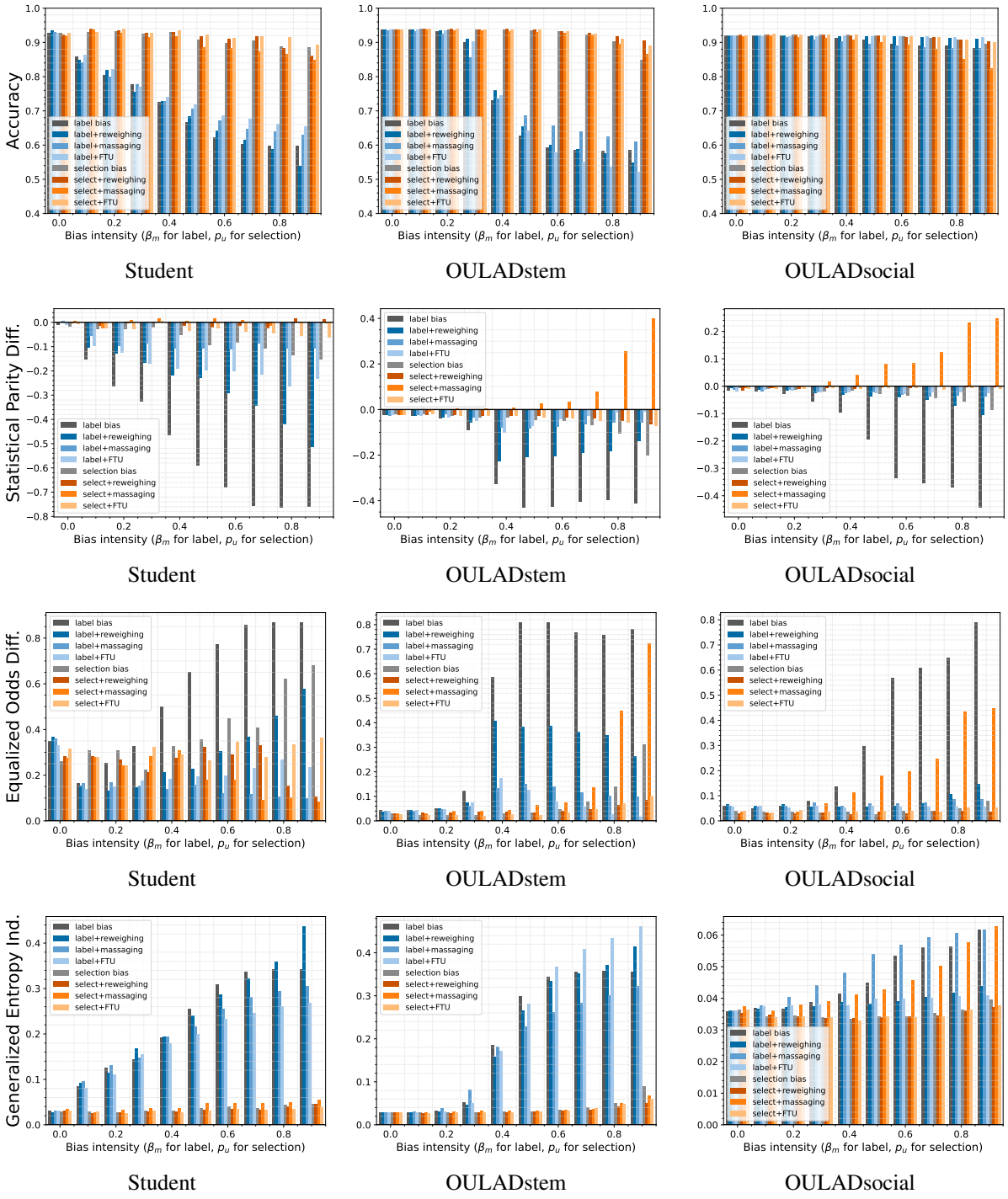

Fig. 1. Evolution of models accuracy and fairness metrics with the increase of bias intensity in the training sets

References

[1] Joachim Baumann, Alessandro Castelnovo, Andrea Cosentini, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. 2023. Bias On Demand: Investigating Bias with a Synthetic Data Generator. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (Macau, SAR China). International Joint Conferences on Artificial Intelligence Organization, 7110–7114. https://www.ijcai.org/proceedings/2023/828

[2] William Blanzeisky, Padraig Cunningham, and K. Kennedy. 2021. Introducing a Family of Synthetic Datasets for Research on Bias in Machine Learning. (2021). https://doi.org/10.48550/arXiv.2107.08928

[3] Paulo Cortez. 2008. Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T

[4] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2074–2152. https://link.springer.com/10.1007/s10618-022-00854-z

[5] Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. 2023. How to be fair? A study of label and selection bias. 112, 12 (2023), 5081–5104. https://doi.org/10.1007/s10994-023-06401-1

[6] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta GA USA). ACM, 329–338. https://dl.acm.org/doi/10.1145/3287560.3287589

[7] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[8] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. 1, 2 (2024), 11:1–11:52. https://doi.org/10.1145/3631326

[9] Gareth Jones, James M. Hickey, Pietro G. Di Stefano, C. Dhanjal, Laura C. Stoddart, and V. Vasileiou. 2020. Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. https://arxiv.org/abs/2010.03986

[10] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. 33, 1 (2012), 1–33. http://doi.org/10.1007/s10115-011-0463-8

[11] Lisa Koutsoviti Koumeri, Magali Legast, Yasaman Yousefi, Koen Vanhoof, Axel Legay, and Christoph Schommer. 2023. Compatibility of Fairness Metrics with EU Non-Discrimination Laws: Demographic Parity & Conditional Demographic Disparity. arXiv:2306.08394 [cs] http://arxiv.org/abs/2306.08394

[12] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. 2017. Open University Learning Analytics dataset. 4, 1 (2017), 170171. https://doi.org/10.1038/sdata.2017.171

[13] Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida, and Mohammed J. Zaki. 2019. A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu HI USA). ACM, 437–444. https://doi.org/10.1145/3306618.3314262

[14] Daphne Lenders and Toon Calders. 2023. 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* (Tallinn Estonia). ACM, 350–357. https://dl.acm.org/doi/10.1145/3555776.3577634

[15] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. 54, 6 (2021), 115:1–115:35.

[16] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. 10, 3 (2020), e1356. https://www.doi.org/10.1002/widm.1356

[17] Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero Soriano, Samira Shabanian, and Sina Honari. 2021. Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/2723d092b63885e0d7c260cc007e8b9d-Paper-round1.pdf

[18] Shirin Riazy, Katharina Simbeck, and Vanessa Schreck. 2020. Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments. In *Proceedings of the 12th International Conference on Computer Supported Education* (Prague, Czech Republic). SCITEPRESS - Science and Technology Publications, 15–25. https://doi.org/10.5220/0009324100150025

[19] 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* (London United Kingdom). ACM, 2239–2248. https://dl.acm.org/doi/10.1145/3219819.3220046

[20] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems* (Vancouver, Canada), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html