Investigating the impact of bias mitigation methods when applied to real-world binary classification scenarios

Kejun Dai[1] **and Natania Thomas**[2]

[1]**kdai332**
[2]**ntho111**

ABSTRACT

Machine learning models' ability to derive accurate predictions from existing data attracts many business and government organisations to use them in their decision-making process. However, it also introduces the fairness problem, where models propagate bias in the data and make unfair predictions that have the capability to alter someone's life significantly. Fairness machine learning is an emerging research field that improves the training process of models to solve the fairness problem. There is great progress in proposing effective bias mitigation methods but little progress in testing these methods in real-world scenarios. To address the knowledge gap, our research investigates the impact of bias mitigation methods in such scenarios. We use the AIF360 framework to benchmark pre-processing, in-processing and post-processing methods with datasets sourced from Kaggle.com. Our main finding is that all methods' performance is varied and depends on the dataset's characteristics in relation to sensitive attributes.

Keywords: Fairness Machine Learning, Bias Mitigation, Benchmark

1 INTRODUCTION

The rapid development of machine learning technology allows computers to extract information from data and make accurate predictions on par with experts in the same field. Nowadays, we generate 2.5 quintillion bytes of data per day [1]. Under the Big Data era, this technology is implemented to assist the decision-making by various businesses and organisations in multiple critical areas like loan approval [2], job recruits [3], school admission [4], and credit risk prediction [5].

However, multiple pieces of evidence suggest that machine learning models will rely on certain sensitive attributes like gender or race during their training and can propagate bias that exists in the data to a greater extent in their prediction. For example, an algorithm to promote jobs in the Science, Technology, Engineering and Mathematics (STEM) field in a gender-neutral way prompted fewer ads to women than to men due to gender imbalance in STEM fields [6].

Since machine learning is now involved in making decisions that would significantly alter people's lives, there is a rapidly growing interest in improving the machine learning process to mitigate bias in the training data. According to [7], there are already 341 publications proposing bias mitigation methods for machine learning classifiers. As an emerging research field, there is still much work to be done to evaluate the fairness performance of those proposed methods. Currently, several benchmark frameworks are developed, like AI Fairness 360 [8] and Fairlearn [9], and several experiments comparing a handful of methods. However, Hort et al. pointed out that there is a need for the experiment to extend to more datasets and applications to real-world scenarios can be explored.

We address the knowledge by investigating the impact of bias mitigation methods when applied to real-world binary classification scenarios. The research report will be structured in the following way. In Chapter 2, we will have brief descriptions of related work and their main findings. Chapter 3 details the selection procedure of our experiment's datasets and their characteristics. Chapter 4 explains the methodology we use in our experiment, including the structure of the benchmark and the methods we choose to evaluate. Chapter 5 demonstrates the results of our benchmark experiments and in Chapter 6 we discuss the implications of our results. We conclude our report in Chapter 7 with a summary of our research.

2 RELATED WORK

Friedler et al. [10] investigated the stability of four bias mitigation methods and provided an open-source framework to facilitate future benchmarking. The experiment's datasets are limited to real-life fairness ML datasets, like Adule incomes and German.

The main finding of the work is that fairness metrics are correlated with each other, Algorithms make significantly different tradeoffs, and Algorithms tend to be sensitive to variations in the input

Reddy [11] investigated how bias mitigation methods on neural networks respond to different correlations between sensitive values and target labels by artificially modifying real-world datasets. The main finding of the work is that the model will exploit the bias most when the unprivileged group is underrepresented or there are sensitive attributes and target labels. Furthermore, the work finds that the models still rely on information of sensitive attributes after applying the methods, meaning it may still lead to unfair predictions when used in downstream decision-making.

Zong et al. [12] construct a benchmark framework for evaluating fairnessaware models to predict medical imaging. Because of that, the paper uses imaging-related datasets and evaluates methods concerned with image recognition. The main finding is that no method can perform statistically better than the baseline model ERM in the Area Under Curve.

3 DATASETS

The datasets in the experiments are sourced from Kaggle websites to represent real-world applications. We first search for datasets in Kaggle that are tasked with binary classification, contains sensitive attribute like age, gender or race, and provides scenarios that are relevant to fairness machine learning. Ultimately, we chose the Job Application, Bank Churn, and College datasets. They are different from each other to represent diverse potential real-world applications. All these datasets share the same sensitive attribute of gender where the privileged groups are males, and the unprivileged groups are females (and non-binary).

Bank dataset

The **Bank dataset** [1] contains information on customers of ABC Multistate Bank, including their balance, credit card usage and estimated salary. The learning task of the dataset is to predict whether the customer will leave the bank after a while. Bank managers may use the dataset to investigate the market audience that tends to leave the bank.. If done inappropriately, it may result in bank managers enacting a retention campaign targeting females specifically. It has 10k data entries with 10 features, which is below the volumes of the Job dataset and Adult datasets but sufficient to train machine learning models. The privileged group to unprivileged group ratio is 55:45. Among the privileged group, the positive-to-negative label ratio is 16:84, and among the unprivileged group, the positive-to-negative label ratio is 25:75. The dataset has a balanced representation for both groups. However, for both groups, the distribution of labels is skewed towards negative labels, and unprivileged groups are more correlated with positive labels than privileged groups.

College dataset

The **College dataset** [2] is a synthetic dataset created for a college project. The data contains student's academic information and parent's information. The learning task of the dataset is to predict whether the student will go to college after graduating from high school. The dataset's scenario can still be relevant to machine learning fairness, as school counsellors may target male students more for university counselling and ignore female students after training such a model. It only has 1k data entries with 10 features, which may not provide sufficient data to train machine learning models. The privileged group to unprivileged group ratio is 52:48. Among the privileged group, the positive-to-negative label ratio is 48:52, and among the unprivileged group, the positive-to-negative label ratio is 52:48. The dataset has a balanced representation as well as a balanced label distribution, it may represent a scenario where bias presented in the dataset is minimal.

Job dataset

The **Job dataset** [3] contains information on StackOverflow users, including their education levels, coding experience and skills, and previous salary. The learning task of the dataset is to predict whether the user is being employed. The dataset may be used by recruiters to assess the employability of the new application quickly. If done inappropriately, the recruiter may favour male applicants over female ones.It has 74k data entries with 12 features, which has similar volumes of data with standard fairness machine learning datasets like the Adult datasets [13]. The privileged group to unprivileged group ratio is 93:7. Among the privileged group, the positive-to-negative label ratio is 54:45, and among the unprivileged group, the positive-to-negative label ratio is 47:53. Despite the label distribution being similarly balanced for unprivileged group and privileged group, the unprivileged group is extremely underrepresented for the dataset.

[1] https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset/data
[2] https://www.kaggle.com/datasets/saddamazyazy/go-to-college-dataset
[3] https://www.kaggle.com/datasets/ayushtankha/70k-job-applicants-data-human-resource

4 METHODOLOGY

The research utilises the open-source framework AI Fairness 360 (AIF360), as it provides sets of implementations of widely-known methods that are representative of the current knowledge sphere. We perform five-fold cross-validations in our experiment. We first split the dataset into five equally-sized folds and chose one of the folds as our test sets and the rest of the folds as training sets. Before training, these data are handled by the AIF360 framework for transforming categorical features and singling out sensitive attributes. In addition, we perform min-max normalisation on features with a large range of values.

We use Scikit-learn's implementation of Random Forest Classifier as our baseline model. The random forest is the state-of-the-art algorithm in the binary classifier and will be a contender for other fairness-unaware models in the real-life scenario. All other methods' models will be compared with it in terms of fairness-performance tradeoff.

Metrics

For the following definition of the performance or fairness metrics, we refer $\hat{Y}$ to the model's predicted labels and Y to their actual labels. Their values are 1 if the label is positive and 0 if the label is negative. We refer D to the values of the sensitive attributes. Its value is 1 if it is in the privelleged groups and 0 otherwise.

$$Accuracy = P(\hat{Y} = Y)$$

$$Precision = \frac{P(\hat{Y} = 1, Y = 1)}{P(\hat{Y} = 1)}$$

$$Recall = \frac{P(\hat{Y} = 1, Y = 1)}{P(Y = 1)}$$

The performance metrics used in the benchmark are **Accuracy**, **Precision**, **Recall**. Accuracy test the correctness of the model's overall prediction, while Precision test the reliability of model's positive prediction, and Recall test the capability of the model to recognize data with positive labels. These capitulate real-world expectation for an accurate, reliable predication models. The closer all of the performance metrics are to 1, the better the models' performance is.

$$DisparateImpact = P\frac{P(\hat{Y} = 1|D \neq 1)}{P(\hat{Y} = 1|D = 1)}$$

$$DisparateParity = P(\hat{Y} = 1|D \neq 1) - P(\hat{Y} = 1|D = 1)$$

$$EqualOpportunity = P(\hat{Y} = 1, Y = 1|D \neq 1) - P(\hat{Y} = 1, Y = 1|D = 1)$$

$$EqualizedOdds = P(\hat{Y} = 1, Y \neq 1|D \neq 1) - P(\hat{Y} = 1, Y \neq 1|D = 1)$$

The fairness metrics used in the benchmark are **Disparate Impact**, **Disparate Parity**, **Equal Opportunity (TPR rate difference)** and **Equalized Odds (FPR rate difference)**. Disparate Impact and Disparate Parity test the independent criteria of fairness where the model's prediction is independent of the data's sensitive attribute [14]. Equal opportunity and Equalized odds test the separation criteria of fairness where the model's prediction is independent of the data's sensitive attribute on condition to its actual outcome [14]. They are some of the standard fairness metrics acorss the research field [7]. Except Disparate Impact which is closer to 1, when all other fairness metrics are closer to 0, the models' predictions are much fairer.

Methods candidates

Bias mitigation methods can be classified into three categories based on their mechanism: **Pre-processing**, **In-processing** and **Post-processing**. For each category, we choose 2 methods for benchmarking.

Pre-processing methods mitigate bias by analysing the data and transforming their labels in relation to their sensitive attribute value so that the models trained from the transformed models will learn less information from the sensitive attribute. The pre-processing methods we evaluate are Reweighting [15] and DisparateImpactRemover [16]. After transforming the data with pre-processing methods, we train a Random Forest model from Scikit-learn and use it for evaluation. We train one model with Reweighting and three with DisparateImpactRemover with different repair levels (0.5, 0.7 and 1.0).

In-processing methods mitigate bias by interfering with the training process and regulating the trained parameter so that the model relies less on sensitive attributes during training phases. As a result, we directly use the model produced by the method for evaluation. The in-processing methods we evaluate are AdverservialDebiasing [17] and GridSearchReduction [18] [19]. We train one model with AdverservialDebiasing and two with GridSearchReduction with different metrics criteria (EqualizedOdds and DisparateImpact)

Post-processing methods mitigate bias by learning prediction outcomes from a model and then transforming the label in relation to its sensitive attribute values to improve the fairness metrics The post-processing methods we evaluate are Calibrated-EqOddsPostprocessing [20] and RejectOptionClassification [21]. We train one model with CalibratedEqOddsPostprocessing and two with RejectOptionClassification with different metrics criteria (EqualizedOdds and DisparateImpact)

5 RESULTS

Graphical Representations

Bank Dataset

Figure 1 is indicating the relationship between Recall and the four different fairness metrics. There are three methods that are distanced further from the baseline, Adversarial Debias, GridSearchReduction(DP), GridSearchReduction(EO). These methods did the best when looking at the fairness metric, at least for Equal Opportunity and Equalized Odds, specifically Adversarial Debias, which scored the closest to zero, which indicates high fairness. All methods scored close to the baseline in recall, the range being 46% to 47%, whereas Adversarial Debias scored around 41% in recall compared to baseline, which could indicate that it won't pick up on the positive instances as well as the other methods. In regards to recall, the pre and post processing methods performed the best, with two of the in processing methods. However the in-processing methods had the best fairness.

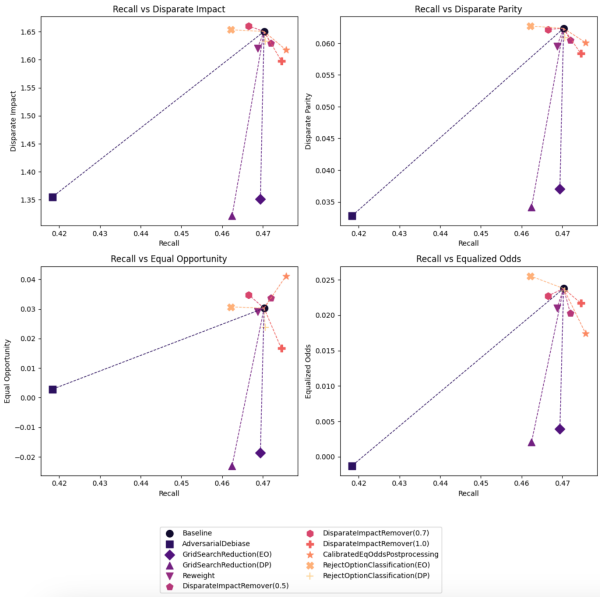

Figure 1. Recall vs Fairness Metric

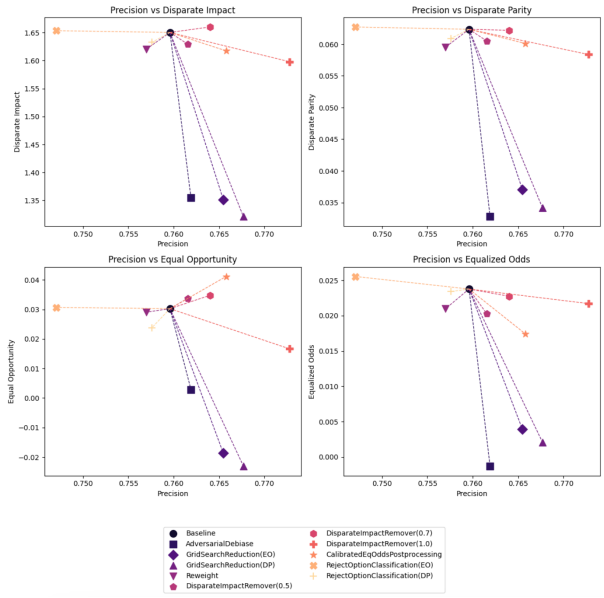

Figure 2. Precision vs Fairness Metric

Figure 2 is indicating the relationship between Precision and the four different fairness metrics. Once again the methods that stood out are the Adversarial Debias as well as the two GridSearchReduction methods. Adversarial Debias once again scores high in fairness for Equal Opportunity and Equalized Odds. One thing to consider is there often can be a tradeoff between precision and recall. In this case, there seems to be some support for this from the In-Processing methods, as Adversarial Debias scored lower than the other methods in Recall, however seems to score similar to the Baseline with 76% of precision. All methods seemed to perform close to the baseline in regard to precision with values ranging from 74% to 76% precision. RejectOptionClassification(EO) had the lowest precision. In regards to precision all the methods seemed to have good precision, while the in-processing methods had the best fairness, specifically Adversarial Debias.

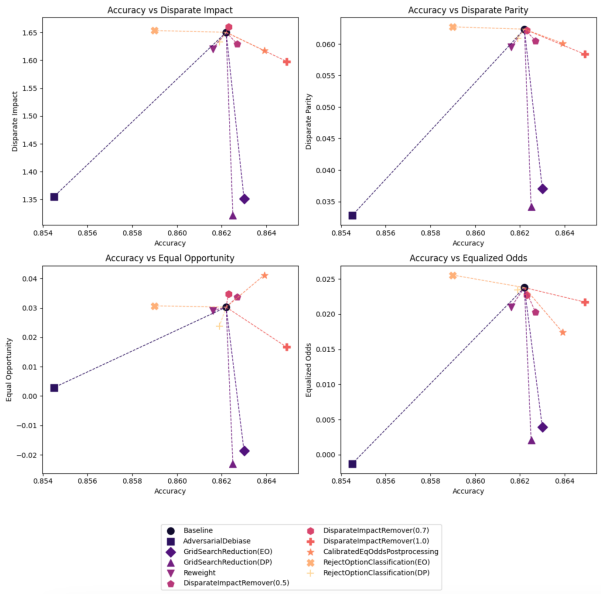

Figure 3. Accuracy vs Fairness Metric

Figure 3 is indicating the relationship between Accuracy and the four different fairness metrics. Adversarial Debias has high fairness in both Equal Opportunity and Equalized Odds fairness metrics. It however has the lowest accuracy. The highest accuracy model was the DisparateImpactRemover(1.0). When comparing the accuracy to the baseline, all methods are relatively close to the baseline with values of accuracy ranging from 85% to 86%.

Overall in the Bank Dataset, the in-processing methods seemed to perform the best with high fairness, low recall, higher precision and relatively high accuracy. There did seem to be a tradeoff between the recall and precision of the methods, specifically the in-processing methods. With a downward trend, we do expect there to be high fairness and high accuracy/precision/recall, and the data seems to support this to an extent. The in-processing methods however performed best in the fairness in the Equal Opportunity and Equalized odds fairness metrics.

College Dataset

Figure 4 is indicating the relationship between Recall and the four different fairness metrics. Here the data seems to have over-corrected a bit, where the ebay fairness was in the Disparate Parity with the Adversarial Debias method. All methods but the Adversarial Debias Method, had quite high recall, some higher than the baseline itself. It's difficult to determine which methods did the best in fairness and recall as most of the methods had a better recall performance than fairness.

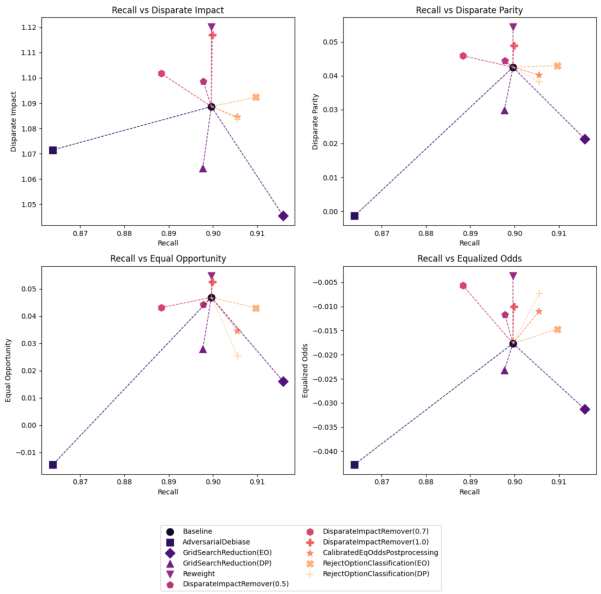

Figure 4. Recall vs Fairness Metric

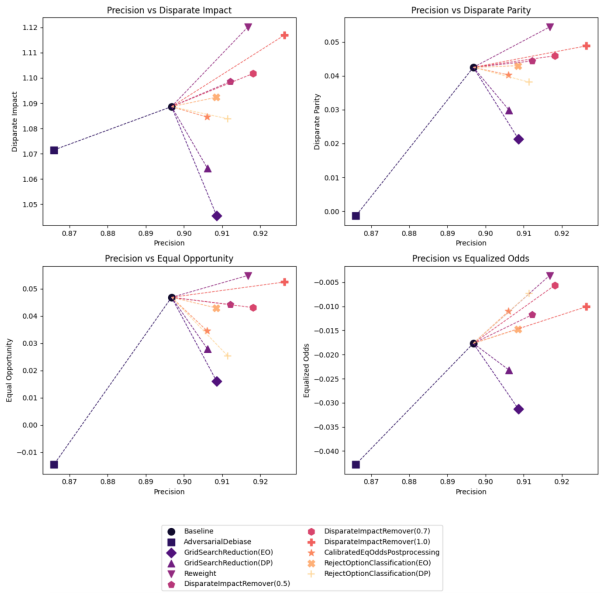

Figure 5. Precision vs Fairness Metric

Figure 5 is indicating the relationship between Precision and the four different fairness metrics. In this set of data with precision, there is a downward trend for the in-processing methods whereas an upward trend for the remainder methods. The downward trend is what we would prefer as this indicates higher precision and higher fairness. Comparing the in-processing methods, the GridSearchReduction methods seem to have better precision, but the Adversarial Debias has the best fairness.

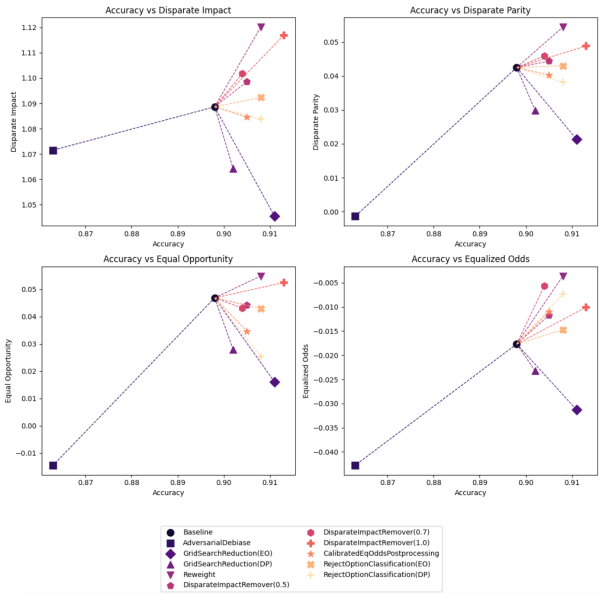

Figure 6. Accuracy vs Fairness Metric

Figure 6 is indicating the relationship between Accuracy and the four different fairness metrics. Once again we have a comparison between the in-processing methods with downward trends and the reaminded methods having upward trends. Adversarial Debias seemed to have high fairness, but had much lower accuracy than the other methods.

Overall, the in-processing had the best graphical representation where they had downward trends compared to the other methods which had more upward trends. Adversarial Debias had the best fairness in all the metrics but the Equalized odds, as the other methods ranked much closer to the zero than Adversarial

Job Dataset

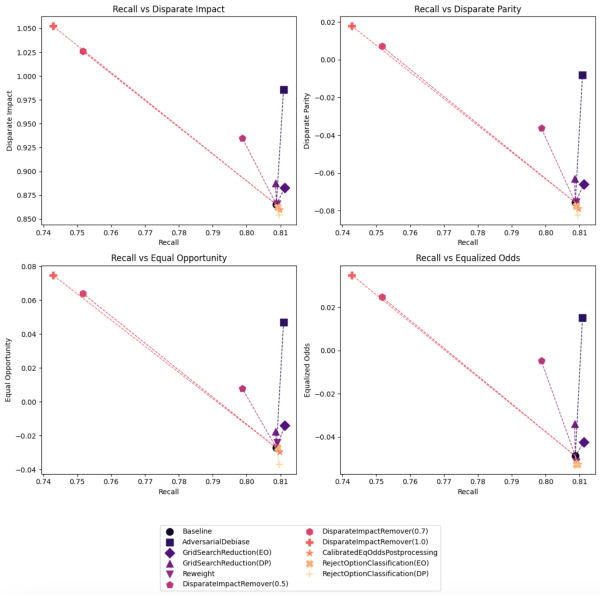

Figure 7. Recall vs Fairness Metric

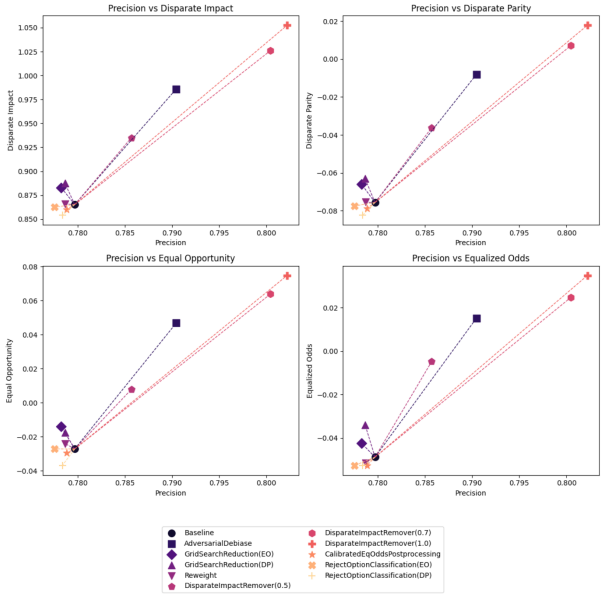

Figure 8. Precision vs Fairness Metric

Figure 7 is indicating the relationship between Recall and the four different fairness metrics. The only two methods that had quite low recall compared to the baseline, were the pre-processing methods, DisparateImpactRemover(0.7) and DisparateImpapctRemover(1.0), but they seemed to overcorrect for some metrics, but also have relatively good fairness for others, like Disparate Parity and Equalized Odds. Recall overall was very consistent with the majority of the methods scoring high.

Figure 8 is indicating the relationship between Precision and the four different fairness metrics. We can see an upward trend for this, with quite low precision, the DisparateImpactRemovers (1.0, 0.7) having the highest precision. While they had very high precision, they had also relatively good fairness compared to the other methods for Disparate Parity, however for the remainder they didn't do as well. Equal Opportunity had methods which were closer to the expected fairness value of zero, however very low precision.

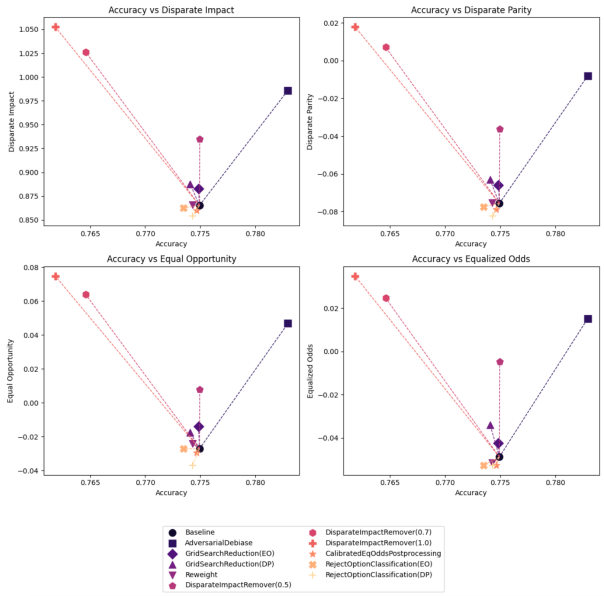

Figure 9. Accuracy vs Fairness Metric

Figure 9 is indicating the relationship between Accuracy and the four different fairness metrics. The model with the highest accuracy was the Adversarial Debias model. It also had relatively good fairness excluding the Equal Opportunity metric, which is interesting as it didn't have the same accuracy fairness balance in the other datasets. It actually had the lowest accuracy out of all methods in other datasets. Lowest accuracy were the pre-processing methods, but still some fairness involved.

Overall, all methods in this dataset had upward trends which we would assume is not supportive, as our initial idea was that a downward trend indicated better results, however in this dataset, there may have been overcorrecting or it was just the data, the data did seem to have values closer to fairness expected values. The methods that performed best were the pre and in-processing methods, but as not all methods within those two specific areas did not perform to the same level, we cannot say which method is the best.

STATISTICAL ANALYSIS

It is important to note, that the values gathered from the training were mostly zero values, which would make it difficult to do any statistical analysis on, to gain any relevant statistical conclusions. The tests that were conducted, seem to indicate that there was no significance to the relationships between the baseline value and the other average values.

The Cohen's d test is done to compare two groups and it takes the difference between the two groups to express this in the standard deviation units. The Cohen's d test is to measure the effect size which attempts to assess the stability of the algorithms. The comparison is done between the baseline and the fairness metrics. Specifically it is looking at how different the fairness metric performs compared to the baseline. The scale is based on three specific values, 0.2 (small effect), 0.5 (medium effect), 0.8 (large effect).

Performance/Metric	Cohen's d Test Value
Accuracy	0.2739
Recall	0.5331
Precision	-0.4963
Disparate Impact	1.0729
Disparate Parity	1.0595
Equal Opportunity	0.8195
Equalised odds	1.1499

Table 1. Bank Dataset - Cohen's d Test

Performance/Metric	Cohen's d Test Value
Accuracy	-0.4638
Recall	0.1221
Precision	-0.9892
Disparate Impact	0.0401
Disparate Parity	0.5247
Equal Opportunity	0.9718
Equalised odds	-0.1756

Table 2. College Dataset - Cohen's d Test

Performance/Metric	Cohen's d Test Value
Accuracy	0.4547
Recall	0.6855
Precision	-0.7705
Disparate Impact	-1.0634
Disparate Parity	-1.0786
Equal Opportunity	-1.0682
Equalised odds	-1.0831

Table 3. Job Dataset - Cohen's d Test

Some overall comments from this values. The ranges of effect size do range from small to large. There are quite a few negative values especially in the Job Dataset which could indicate that there were models that were performing poorer than others. The negative values for the performance metrics indicate that there was poorer accuracy/recall/precision in some models than others. We were measuring the recall metric depending on the model's ability to correctly identify positive cases and the precision metric, by the model's ability to performs positive predictions on the datasets.

6 DISCUSSION

In regard to the graphical representations, it is difficult to determine which method performed the best with high fairness but maintaining accuracy, precision or recall. Most methods seemed to score high on precision, however the bank dataset had the lowest recall rate compared to the other datasets, which may attribute to its imbalanced label distributions. The in-processing methods seemed to perform well in the bank and college datasets however not as well the job dataset, the adversarial debias being the exception. The pre-processing method seemed to do better in the job dataset.

Given a dataset, the fairness-performance tradeoff of all methods is relatively similar regardless of the performance metrics or the fairness metrics. Equalized Odds display the most difference with other fairness metrics. Given the same performance metrics, the relative difference between fairness-performance tradeoff of methods are much more prevelant in smaller datasets. It facilitates the claim that fairness metrics are correlated by [10]. It illustrates that, in real-world application, those metrics can be used interchangeably when comparing different bias mitigation methods.

The fairness-performance tradeoff of all accessed methods are sensitive to the settings and characteristics of the training dataset. None of the accessed methods share a consistent fairness-accuracy tradeoff across different datasets. Most noticeably is the Adversarial Debiasing. Depending on the dataset, it may either over correct, trades much performance for fairness or both improves fairness and accuracy. It means that for different datasets, the best fairness-accuracy methods may be different. It also means that in real-world scenarios, there are no rule of thumb for picking the best fairness-performance tradeoff bias mitigation methods. And for each new and unknown dataset, a similar benchmark experiment needs to be taken to figure the most suitable methods for the dataset.

An interesting observation is that for Bank dataset where the unprivileged group is positively correlated with positive label, methods other than in-processing failed to derive meaningful fairness improvements for the model, and some of them lose accuracy at the same time. Similar to Reddy's [11] observation that bias mitigation methods do not necessarily prevent models from learning from sensitive attributes, it means that pre-processing and post processing methods may not necessarily recognize unfair predictions. It also means that in practice, there is also a risk of unfair predictions when those methods are applied without care.

7 CONCLUSION

This research took an experimental approach to test how bias mitigation methods behaved when applied to real-life scenarios. We utilised the Random Forrest method for the baseline, three performance metrics (Accuracy, Recall, Precision) and four fairness metrics (Disparate Impact, Disparate Parity, Equalized Odds, Equal Opportunity). These fairness metrics and bias mitigation methods were tested on three different datasets (Bank, College, Job).

From the results of our research, there does not seem to be exactly one method that reigns over all the methods. We do need to take into account that the real life scenarios are very different and our findings won't be necessarily generalized across them. If we were to pick a specific bias mitigation method for application we could say the Adversarial Debias Method. We could provide a far more solid conclusion with a couple more distinguished datasets to see whether some method can provide consistent high fairness-performance tradeoff across multiple datasets.

Future work includes using the same methods to test across more recent and more datasets in general. With more datasets it would help in supporting/creating more solid conclusions. It would also be interesting to compare old datasets to more recent datasets to see whether the methods perform better or worse on them. Moreover, we could in-depth analyse the bias present in old datasets and recent datasets, and see how methods react to different levels of bias present in the datasets.

REFERENCES

[1] IBM, *What is big data?* `http://www-01.ibm.com/software/data/bigdata/what-is-big-data`.

[2] A. Mukerjee, R. Biswas, K. Deb, and A. P. Mathur, "Multi-objective evolutionary algorithms for the risk–return trade–off in bank loan management," *International Transactions in operational research*, vol. 9, no. 5, pp. 583–597, 2002.

[3] E. Faliagka, K. Ramantas, A. Tsakalidis, and G. Tzimas, "Application of machine learning algorithms to an online recruitment system," in *Proc. International Conference on Internet and Web Applications and Services*, 2012, pp. 215–220.

[4] J. S. Moore, "An expert system approach to graduate school admission decisions and academic performance prediction," *Omega*, vol. 26, no. 5, pp. 659–670, 1998.

[5] I.-C. Yeh and C.-h. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert systems with applications*, vol. 36, no. 2, pp. 2473–2480, 2009.

[6] A. Lambrecht and C. Tucker, "Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads," *Management science*, vol. 65, no. 7, pp. 2966–2981, 2019.

[7] M. Hort, Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Bia mitigation for machine learning classifiers: A comprehensive survey," *arXiv preprint arXiv:2207.07068*, 2022.

[8] R. K. Bellamy, K. Dey, M. Hind, *et al.*, "Ai fairness 360: An extensible toolkit for detecting," *Understanding, and Mitigating Unwanted Algorithmic Bias*, 2018.

[9] S. Bird, M. Dudık, R. Edgar, *et al.*, "Fairlearn: A toolkit for assessing and improving fairness in ai," *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[10] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 329–338.

[11] C. Reddy, "Benchmarking bias mitigation algorithms in representation learning through fairness metrics," 2022.

[12] Y. Zong, Y. Yang, and T. Hospedales, "Medfair: Benchmarking fairness for medical imaging," *arXiv preprint arXiv:2210.01725*, 2022.

[13] R. Kohavi *et al.*, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.," in *Kdd*, vol. 96, 1996, pp. 202–207.

[14] H. Barocas and Narayanan, *Fairness and machine learning*.

[15] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[16] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[17] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[18] A. Agarwal, A. Beygelzimer, M. Dudık, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International conference on machine learning*, PMLR, 2018, pp. 60–69.

[19] A. Agarwal, M. Dudık, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*, PMLR, 2019, pp. 120–129.

[20] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in neural information processing systems*, vol. 30, 2017.

[21] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *2012 IEEE 12th international conference on data mining*, IEEE, 2012, pp. 924–929.