Influence of Data Processing on the Algorithm Fairness vs. Accuracy Trade-off

Building Pareto Fronts for Equitable Algorithmic Decisions

Andres David Salvi[1]
Supervisors: Jie Yang, Sarah Carter[1]
[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2024

Name of the student: Andres David Salvi
Final project course: CSE3000 Research Project
Thesis committee: Jie Yang, Sarah Carter, Marcus Specht, Stefan Buijsman

Abstract

Algorithmic bias due to training from biased data is a widespread issue. Bias mitigation techniques such as fairness-oriented data pre-, in-, and post-processing can help but usually come at the cost of model accuracy. For this contribution, we first conducted a literature review to get a better insight into the potential trade-offs. We followed by implementing a Python program to test how the *Disparate Impact Remover (DIR)* pre-processing and *Reject Option Classification (ROC)* post-processing techniques impacted the fairness and accuracy metric values of a *Logistic Regression* model trained on data from the Adult Income dataset. The implementation also allows for building Pareto fronts that trade off fairness and accuracy metrics of choice, thus offering a blend of perspectives on fairness. Our findings give insight into how combined fairness methods influence the trade-off, but our implementation can be extended to explore such trade-offs using other datasets, models, and fairness methods.

1 Introduction

Our society increasingly relies on algorithms; from machine learning (ML) models detecting data patterns, chatbots, navigation systems, and facial recognition. However, this dependency raises concerns about **algorithmic bias** distorting outcomes unintentionally, which can be defined as systemic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others [1]. Various solutions can help mitigate such bias, such as using more diverse and representative data, documentation transparency, and implementing statistical data fairness manipulation [1; 2]. This paper analyzes the fairness improvement from various statistical data manipulation methods, namely data pre-, in-, and post-processing, since these typically come at the cost of model accuracy, thus leading to a potential trade-off. As AI systems are commonly used for daily functions like mortgage approvals and recommender systems, maintaining their reliability despite these interventions is crucial to avoid significant potential losses [3; 4].

Despite the uncertainty on how statistical fairness methods influence model accuracy, multiple papers exist that help tackle this gap. A study on an ML model for home mortgage approvals revealed an inherent bias against black applicants compared to white ones. Attempts to correct this using an alternative model exacerbated the bias, and efforts to apply ML fairness methods resulted in negative outcomes for all stakeholders, underscoring the need for human oversight in ML implementations [4]. Another paper introduces a universal framework to identify the optimal trade-off between fairness and accuracy using diverse measures, analyzing how different techniques were applied in other studies to reduce data bias. Although it tested the effectiveness of several fairness methods in a case study, only a limited subset was examined individually, leaving plenty of room for further research [5].

1.1 Research Scope

Our study will explore how two selected methods working together can enhance this trade-off across various metrics, potentially revolutionizing bias mitigation overall. We will plot the trade-offs using Pareto fronts, which are graphs that illustrate a "set of solutions that are non-dominated to each other but are superior to the rest of solutions in the search space" [6]. Therefore, our main research question is:

How can combining pre- and post-processing techniques optimize the fairness vs. accuracy trade-off within a Pareto front framework?

The sub-questions that we aim to answer in this research paper are:

- How much could the selected fairness methods mitigate the group bias of a given (biased) dataset and system?
- How could we use a Pareto-front to set a good accuracy vs. fairness trade-off?
- How generalizable would the proposed fairness framework be with other datasets, models, and fairness methods?

1.2 Contribution

Combined with our literature review's findings, we built a Python program that uses *Disparate Impact Removal (DIR)* for pre-processing and *Reject Option Classification (ROC)* for post-processing, separately and together. These techniques aim to reduce the data imbalance on the tested binary classification Adult Income dataset which we used to train a *Logistic Regression* model with each protected attribute separately accounted for.

Our findings revealed an overall improvement by combining both fairness methods, but future research can be done to see how other data processing techniques and metrics can build a Pareto front depending on the application, along with testing with other ML models and datasets with different bias types. We justify our method and model selection in subsection 3.2 and subsection 3.3, respectively.

1.3 Structure

To answer our main research question and related sub-questions, we conducted a literature review to understand existing fairness methods and their impact on accuracy and other metrics, alongside a fairness experiment on our dataset. This experiment investigated how our selected fairness methods, both separately and combined, affect the trade-off between fairness and accuracy using various metrics, thus providing insights for implications and future research to bridge our research gap.

The structure of this paper is as follows. First, in section 2, we give a more detailed overview of our research

scope and relevant findings. Next, in section 3, we describe our experimental setup, further explaining the fairness methods and metrics we used for building our trade-off analysis. In section 4, we explore the results of our experiment, comparing the performance of our selected fairness methods and their respective Pareto fronts built using various metrics. We follow in section 5 with a discussion of our findings, their implications, and their impacts. Then, in section 6, we give our conclusions and insights for future work. Finally, in section 7, we discuss relevant ethical matters of this research.

2 Literature Review

We found various literature that explored the fairness vs. accuracy trade-off using various bias mitigation techniques. We first discuss our general findings in subsection 2.1, followed by a summary of closely related work in subsection 2.2, and finally briefly summarizing additional related work in subsection 2.3. We will further discuss the relevance of these findings in section 5.

2.1 General Findings

When searching for ways to build our fairness vs. accuracy trade-off framework, we made two key observations. One, there is a plethora of data pre- and post-processing techniques with slightly different use cases and applicable situations [7; 8; 9]. And two, such technique selection can substantially impact how the trade-off would look depending on the tested dataset and metrics.

We found many different data processing techniques aimed at improving the fairness of the input data and predictions of a machine learning (ML) model. For pre-processing, these include *Disparate Impact Removal (DIR)*, *Learning Fair Representations (LFR)*, *Optimized Pre-Processing*, *Reweighing*, and several more. And for post-processing, we found ones such as *Reject Option Classification* and *Equalized Odds* [5; 8; 10]. As our focus is to examine how effective these two data processing types are, we will be selecting one of each in our model experiment as explained in section 3.

To visualize the trade-off, we can use a Pareto front. We found in various applications that it tends to work well for finding an optimal trade-off between two competing variables, which is also fairly easy to visualize [3; 5; 11; 12]. Therefore, this will require selecting one fairness metric and one accuracy metric to evaluate the model under test and generate its respective Pareto front, which is further explained in section 3.

2.2 Closely Related Work

Two related studies are particularly relevant to our research due to their focus on binary classification bias and similar data processing techniques. Leying Zou et al. looked into discrimination in mortgage application acceptance between white and black applicants [4]. Christian Haas established a framework for exploring the fairness vs. accuracy trade-off, testing various fairness methods and metrics for building Pareto fronts [5]. We will summarize each study's approach, methodology, and key findings.

Mortgage Application Discrimination

Leying Zou et al. explored racial discrimination in mortgage application acceptance towards black applicants, which has been a historic trend from the value/risk proposition perspective [4]. They analyzed an existing dataset to detect racial bias, trained an off-the-shelf ML model with the data, evaluated its bias, and attempted to debias it.

The above methodology gave various insights. Firstly, it confirmed the existence of racial bias against black people in the dataset, highlighting the importance of mitigating any biases present in datasets used for training ML models. Secondly, it tested an off-the-shelf AI model trained with the dataset to see how it would perform, finding that it unintentionally further amplified the existing bias. Finally, they tested the model's fairness and accuracy with and without the fairness methods, finding that the methods worked but at the cost of accuracy, particularly with a substantial increase in false positives.

The findings highlight two key points. First, ML models can propagate or even amplify biases from their training datasets, highlighting the need for pre-processing to enhance pre-training fairness. Second, the fairness vs. accuracy trade-off is critical, particularly when accuracy loss could heavily impact stakeholders; in this case, the increased false positives raised risks for lenders, highlighting the importance of checking even small changes on either side. Their conclusion stresses the importance of exercising caution when training and using ML models, along with considering potential long-term risks.

Framework to Explore Trade-Offs in Algorithmic Fairness

Christian Haas developed a framework intended for exploring the fairness vs. accuracy trade-off when debiasing datasets and ML models [5]. It explores various fairness methods for pre-, in-, and post-processing, to then apply these methods and metrics in two combinations on a dataset to construct a Pareto front for each combination.

The step-by-step process of the framework is illustrated in Figure 1. The documented approach then proceeds to apply the framework using various fairness methods and metrics on the Statlog (German Credit risk) dataset. Specifically, it runs in four configurations, each using one of the following fairness methods. First, it models on a *Support Vector Machine (SVM)* without applied fairness methods. Two other configurations test both *Reweighing* and *Reject Option Classification (ROC)* as fairness pre- and post-processing techniques respectively, separately on the same model. The last configuration runs solely on the *Meta-Fair* classifier. It then tests these four configurations using *Statistical (aka. Demographic) Parity* and *Theil Index* as fairness metrics for two scenarios, both compared against their respective *Area Under Curve (AUC)* values to generate their Pareto fronts. Furthermore, it gives other statistics showing each configuration's performance and fairness metric values, such as *Recall*, *Equalized Odds*, *Disparate Impact (DI)*, etc.

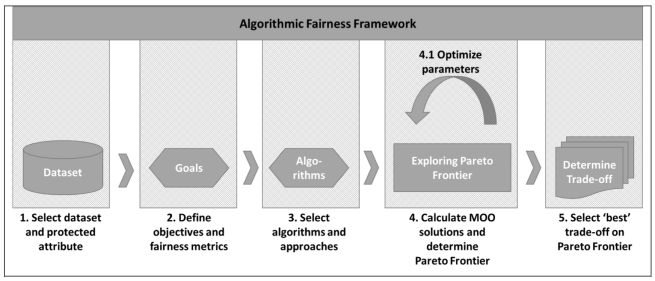

Figure 1: Framework to explore Algorithmic Fairness Trade-offs [5].

The above methodology led to two key takeaways. One, it confirmed their hypothesized non-linear trade-off between various fairness methods and model accuracy, as evidenced by the Pareto fronts, giving the ability to perform cost-based decision-making. Two, different fairness methods influenced different fairness metrics at various magnitudes, with the ROC post-processing technique showing superior performance in both scenarios. Their conclusion is that various fairness methods can mitigate data and model bias, each with its trade-offs and different Pareto-optimal levels dependent on the tested dataset and fairness metrics used.

2.3 Additional Related Work

We found other papers that also gave great insights, three of which we will summarize. As done in the previous subsection, we will briefly discuss the methodology and key takeaways of each.

The first paper, from Lele Sha et al., proposes measures that can be taken to detect and reduce bias in education environments and how to mitigate them with minimal accuracy loss [13]. They selected five datasets, each with their respective classification tasks to test, and they then measured their respective distribution (Dist-bias) and hardness biases. They pre-processed their datasets using multiple *Class Balancing Techniques (CBTs)* and *Balanced Data Samples by Task and Demographic Labels* with gender and first-language background as protected attributes. Afterward, they tested the dataset with and without different combinations of their pre-processing techniques applied, showing a noticeable improvement in model fairness. Overall, the paper highlighted the importance of checking for and mitigating any dataset bias before using it to train ML models, along with showing how their tested data processing techniques led to significant bias reduction with minimal accuracy loss (less than 1%).

The second paper, from Patrick Janssen et al., examines the effects of different fairness methods on model accuracy and social welfare from the debiased model [8]. Unlike the other papers we have studied, it generated synthetically biased datasets with one protected attribute to train and test a *Logistic Regression* ML model. It subsequently used the *AIF360* library to train the models with the raw and debiased datasets, using two techniques from the pre-, in-, and post-processing technique categories respectively (a total of six techniques). With the trained models, they then evaluated their overall accuracy, false-positive rates, and social welfare scores (calculated using various other metrics). They found that the different fairness techniques collectively reduced overall bias as intended with a minor reduction in accuracy with some outliers having a stronger reduction or even an improvement. A notable decrease in social welfare from the different techniques was also found, highlighting the importance of also assessing this metric when building fairness frameworks.

The final paper, from Zhenpeng Chen et al., highlights the importance of checking the fairness and accuracy regarding unconsidered protected attributes when applying fairness metrics to other protected attributes [10]. They individually tested eleven fairness pre-, in-, and post-processing techniques (including the ones we tested) on five different datasets to train four different ML models. For measuring overall accuracy, combined with the typical metrics used in other papers, they also used the *F1-Score* metric for a clearer view of the indirect recall and precision impact of the tested techniques. Overall, they highlighted the effects that different fairness methods could have on unrelated protected attributes along with the accuracy and values of other metrics, such as the *F1-Score*, underscoring the importance of using multiple metrics aside from accuracy to thoroughly measure a model's performance.

3 Methodology

In our investigation on dataset bias, we first explain the dataset we worked with in subsection 3.1. We then tested one fairness data pre-processing and one post-processing technique, separately and together, to see how they would reduce our model's prediction bias, being judged by a core fairness metric and a few others to give further insights. The experiment's process is explained in subsection 3.2, and the evaluation process of our trained model is explained in subsection 3.3.

Our experiment was coded in the *Python* language with various libraries and frameworks used to run our data processing algorithms, modeling, and data collection. These include *AIF360* for data processing and metrics, *TensorFlow* for neural network modeling, *scikit-learn* for the remaining machine learning models we tested, and *Matplotlib* for generating our plots. Our pre-processing and modeling implementations are loosely based on the AIF360's example of using the *Disparate Impact Remover (DIR)* whereas our post-processing was written from the ground up [14].

3.1 Dataset

We tested our implementation on the open-source *Adult Income* dataset provided by the UCI Machine Learning Repository [15]. It comprises data extracted from the U.S. 1994 Census Bureau database, which includes demographic information from thousands of individuals. The prediction task on this dataset is to determine whether or not a person's income exceeds $50k/year based on their listed stats.

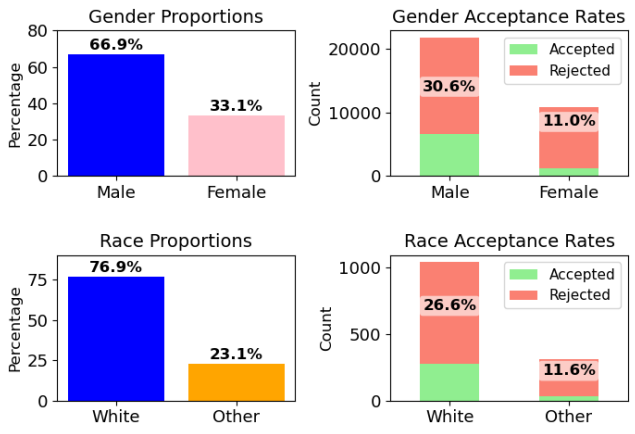

Figure 2: Adult dataset's statistics on gender and race (simplified). The percentages for the graphs on the right indicate the acceptance rate for the listed group.

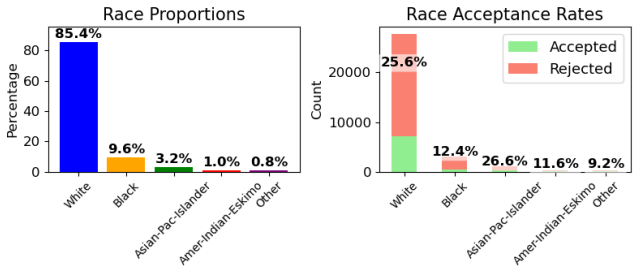

Figure 3: Same structure as Figure 2 but with statistics for all races.

We selected this dataset for its relevance to our research, which focuses on binary classification involving two protected attributes: race and gender. Our copy had 32,561 out of over 48,000 total entries, therefore still providing a diverse range of data. However, it's imbalanced, predominantly featuring males and white people, which incurs a risk for bias propagation during model training, which is a key issue we aim to address and test. Figure 2 and Figure 3 give an overview of the race and gender acceptance rates. Notice how males and white people make up the majority of the dataset entries and have higher acceptance rates, highlighting the dataset's selection and historical biases [16; 17].

3.2 Experiment Setup

To ensure data integrity for our model training and testing, we cleaned our dataset by removing the redundant "education" feature, converting categorical features to numeric values, and eliminating duplicate or incomplete rows. This process only reduced the dataset from 32,561 to 32,537 rows, ensuring a minimal data loss of 24 rows.

We then tested our machine learning model, separately with "race" and "gender" as protected attributes, in four different configurations:

1. No fairness methods applied.
2. Only fairness pre-processing applied.
3. Only fairness post-processing applied.
4. Both fairness pre- and post-processing applied.

For fairness pre-processing, we used the *Disparate-Impact Remover (DIR)* technique at ten different repair strength levels between 0 and 1. For post-processing, we used the *Reject Option Classification (ROC)* guided by the *Statistical Parity Difference (SPD)* metric. We selected the DIR due to its simplicity and because the other techniques had too many parameters, were too accommodating for protected attributes, or unintentionally intervened with the model [18]. And we chose the ROC for its proven effectiveness [5]. We used the *Logistic Regression* classifier as our ML model since out of the five models we initially tested, only this and the *Support Vector Machine* classifier performed well, and the latter was already tested by Haas [5].

In summary, our experiment was conducted as follows:

1. Cleaned up our dataset to prepare it for further processing.
2. If applicable, apply the DIR pre-processing to the training data.
3. Train the model with the prepared data.
4. Run the data's test set through the model. If applicable, tweak the model's predictions using the ROC post-processing technique to optimize its final predictions.
5. Proceed with using our fairness and accuracy metrics to evaluate the model's performance, as explained in subsection 3.3.

3.3 Model Evaluation

We evaluated our model in each configuration using various fairness and accuracy metrics, allowing us to construct their respective Pareto fronts. Our implementation generates two Pareto fronts per run, one with and one without the ROC applied. Each front has 10 data points corresponding to their respective DIR repair level. Therefore, each generated front pair displays the results across all four configurations for the selected protected attribute and metrics.

Each generated front's points are ranked using a linear scoring function that calculates a point's accuracy minus any deviation from the optimal fairness metric value (0 for all except *Disparate Impact (DI)*, which is 1). Each point is then annotated with its score ranking and DIR repair level, where then the best scores from the pre- and post-ROC fronts determine if the ROC yields an optimal result. Therefore, our implementation allows for selecting the optimal DIR repair level and assessing the ROC's effectiveness for each metric pair. In addition, one can also adjust the score function's fairness penalty weight to emphasize fairness as needed.

Our implementation allows for selecting any of the following **fairness metrics**:

- Statistical parity difference (SPD)
- Disparate impact (DI)
- Average odds difference
- Equal opportunity difference
- Equalized odds difference
- Theil index

And the following **accuracy metrics**:

- Recall (i.e., True positive rate; TP)
- False positive rate (FP)
- True negative rate (TN)
- False negative rate (FN)
- Accuracy: $(TP + TN)/(P + N)$[1] [18]
- Balanced accuracy: $(TP + TN)/2$
- Precision: $TP/(TP + FP)$ [18]

In our analysis, we will focus on using the DI and SPD metrics for fairness, and *Accuracy* and *Balanced Accuracy* for accuracy, aiming to ensure equal acceptance rates between majority and minority groups. We hypothesized that guiding the ROC using the SPD metric over the other two possible metrics would yield the best outcome, which we will prove in subsection 4.1.

4 Results

Running our experiment on both protected attributes separately yielded robust results. We will focus primarily on the "gender" model as it provided clearer visualizations of our metrics, though the results for "race" were similar. We first present our core fairness metric results from all four configurations in subsection 4.1, where we also show our findings when we test our post-processing with different guiding metrics. We then detail our generated Pareto fronts using a subset of our metrics in subsection 4.2.

4.1 Data Modeling and Processing

Across all four configurations for the "gender" model, the results largely met our expectations. Without fairness methods applied, the model generally had the worst fairness metric scores and highest accuracy. Increasing our DIR preprocessing repair level slightly improved our fairness metrics at a slight cost of accuracy. However, applying ROC postprocessing significantly improved our fairness metrics across all repair levels. We will summarize the findings of our core fairness metrics and then for the accuracy metrics to then conclude with results from testing the ROC alone.

Fairness Results

For fairness, both of our data processing techniques were generally effective at mitigating the dataset's bias. Figure 4 shows the trends for our metrics for each DIR repair level before and after using the ROC post-processing.

Before applying the ROC, the results aligned with our expectations. Increasing DIR repair levels modestly improved our fairness metrics but showed diminishing returns at higher levels, suggesting a potential optimal midpoint. Even at the highest repair level, both metrics were still not within an acceptable fair range, typically from 0.8 to 1.2 for DI, 0 to 0.1 for *Theil Index*, and -0.1 to 0.1 for all other metrics, but can vary depending on sociotechnical situations and preferences [19]. We also tested the *Average Odds* and *Equal Opportunity* difference metrics, with both remaining consistent at all repair levels but with suboptimal values at around -0.22 and -0.33 respectively. The "race" model gave similar results, except that the last two discussed metrics had an overall incline due to a dramatic upward shift at the center repair level, showing the threshold where the DIR became effective at bias mitigation.

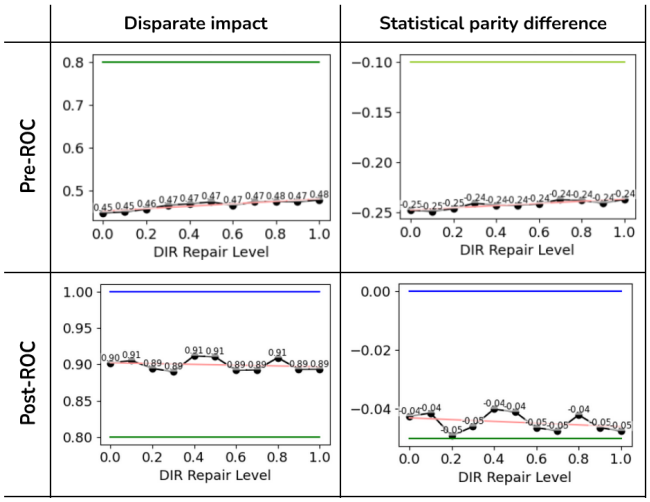

Figure 4: Each row shows the *Disparate Impact (DI)* and *Statistical Parity Difference (SPD)* of the dataset at each DIR repair level, from 0.0 to 1.0. The top row shows the measurements before applying the ROC post-processing, and the bottom one after. Each graph has a red line of best fit to show its overall trend.

After applying the ROC, all four fairness metrics we discussed earlier changed slightly at all repair levels but declined slightly by around -0.01. However, their values were far closer to their optimal levels, hovering at around 0.90 and -0.04, respectively, and with *Average Odds* and *Equal Opportunity* difference around -0.01 and -0.11, respectively. The "race" model displayed the same trend but with more variation in metric values, showing pronounced peaks and drops, likely due to the ROC's decision boundary adjustments not being tailored for these metrics. Depending on the target optimal value, such extremes would likely yield an optimal or fully suboptimal result.

Accuracy Results

As expected, both fairness methods caused a minor overall drop in accuracy for each model trained with the protected attributes. For each DIR repair level before and after using the ROC, the results of two of our four discussed accuracy metrics are plotted in Figure 5.

[1] P = Positive rate; N = Negative rate

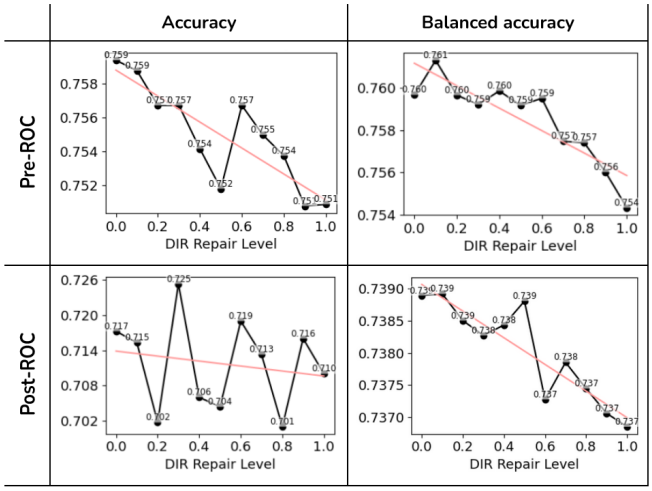

Figure 5: The same structure as in Figure 4 but showing the model's accuracy using our *Accuracy* and *Balanced Accuracy* metrics for each ROC stage.

For the pre-ROC model, both metrics displayed a downward trend, reducing by around 0.007 and 0.005, respectively, and with *Precision* reducing slightly more but with *Recall* remaining the same. However, as shown on the graph, the metrics had decent variations, which was also the case for *Precision* and *Recall*. For the "race" model, both *Accuracy* and *Precision* unexpectedly rose by around 0.005, possibly due to the DIR breaking the model's reliance on the selected protected attributes or reducing dataset noise. The other two metrics, however, experienced a drop as expected. As with the "gender" model, the values had strong variations between the repair levels.

For the post-ROC model, the first three metrics experienced a smaller decline of around 0.003 while *Recall* remained stagnant. Interestingly, all four metrics had significantly more variation between the repair levels. Peaks in these variations would likely yield either a Pareto-optimal solution or the opposite. For the "race" model, however, the results were quite different. Except for *Balanced Accuracy*, all four metrics changed direction, but with smaller changes ranging from 0.001 to 0.005 across the repair levels, likely due to the ROC reducing the model's tendency to overfit with the data at higher DIR repair levels.

Testing our Post-Processing on Different Guiding Metrics

As we observed earlier, the ROC post-processing had a significantly positive impact on our fairness metrics. As mentioned in subsection 3.2, we ran our whole experiment with only the SPD metric to guide the ROC. However, we also guided the ROC in another instance using *Average Odds* and *Equal Opportunity* difference to see how well they would perform.

Figure 6 compares the performance of the three metrics when testing the "gender" model where all three guiding metrics yielded similar results, with SPD being in the middle. However, when post-processing the model's predictions with

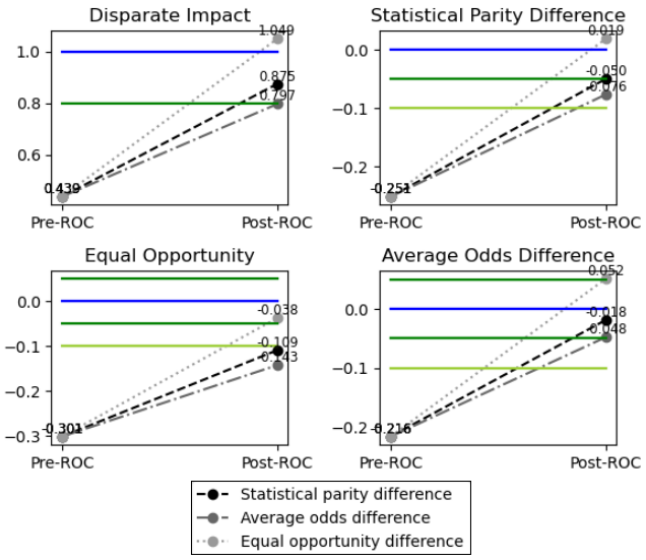

Figure 6: The performance of the ROC post-processing based on the four shown fairness metrics, tested on all three different guiding metrics.

"race" as the protected attribute, using the SPD gave by far the best fairness ratings. Overall, these findings prove our hypothesis made in subsection 3.3.

4.2 Pareto Fronts

After running our models for both protected attributes, we generated several Pareto front pairs; the left front corresponds to before the ROC post-processing and the right after. Each pair represents a combination of one fairness and one accuracy metric of choice. With six supported fairness metrics and seven accuracy metrics, our implementation allows for creating 42 combinations for extensive testing to identify optimal trade-offs. The fairness weight of the scoring function is also adjustable (default = 1.0), allowing for further customization. However, due to space constraints, we only show our four most significant Pareto front pairs of three different metric combinations, all from the "gender" model.

Figure 7 shows our first Pareto front combination goes between SPD and *Accuracy*. The optimal points of each front exhibit a trade-off, where the left point scores lower in fairness but higher in accuracy, while the right has the opposite. Overall, the right point has the best trade-off given the selected metrics and fairness weight, which is the case for most of the other generated front pairs.

Our next combination is similar to the previous one but with *Balanced Accuracy* as the accuracy metric, as shown in Figure 8. We get a similar situation as the previous pair but with a smaller front for the right graph due to it having fewer Pareto-optimal points. This occurs when the selected metrics have greater variation with jagged peaks and dips, as these would lead to more Pareto-dominated points. Adjusting the fairness weight to 0.3 yielded a different optimal trade-off point for

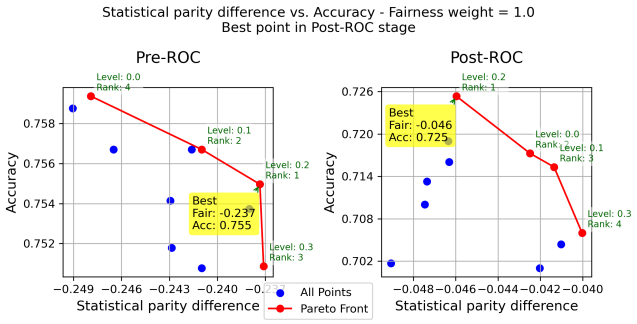

Figure 7: Our first Pareto front pair, comparing model performance before and after applying the ROC. Each point is labeled with its DIR repair level and trade-off score, and the best points additionally show their fairness and accuracy. The title indicates the fairness weight value used for the scoring function and indicates which graph presents the optimal trade-off.

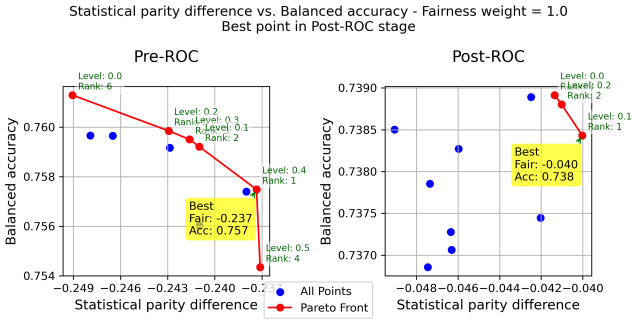

Figure 8: Another Pareto front pair; with *Balanced Accuracy* as its accuracy metric.

both fronts, as shown in Figure 9. However, once again the post-ROC point yielded the best result of both fronts.

The final combination we will illustrate is between the *Theil Index* and *Balanced Accuracy* metrics, which is shown in Figure 10. We selected this combination because unlike the others, the **pre-ROC** front yielded the optimal trade-off of both ROC configurations, showing how combining fairness pre- and post-processing does not always give an optimal result, thus highlighting the importance of making decisions based on which metrics matter most.

The four front pairs we just discussed gave great insights, but we explored many more combinations for both models. A key observation is that for each plotted pair, both optimal points corresponded to DIR repair levels at 0.5 or below. All of our generated fronts, including for the "race" model, had their highest-ranking points within this repair range, highlighting the diminishing returns from excessive pre-processing. However, almost all generated fronts showed the post-ROC point being optimal, illustrating how in most cases, combining the DIR and ROC techniques yielded the best trade-off.

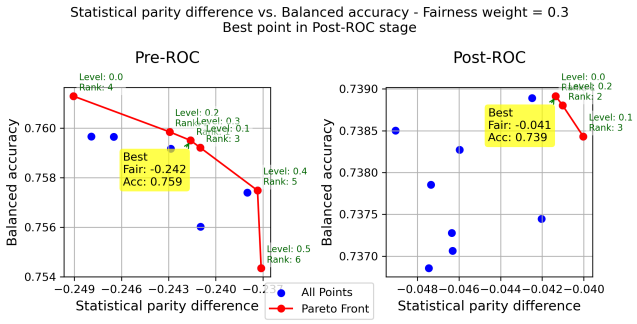

Figure 9: Same Pareto front pair as in Figure 8, but with the fairness weight set to 0.3 instead of 1.0.

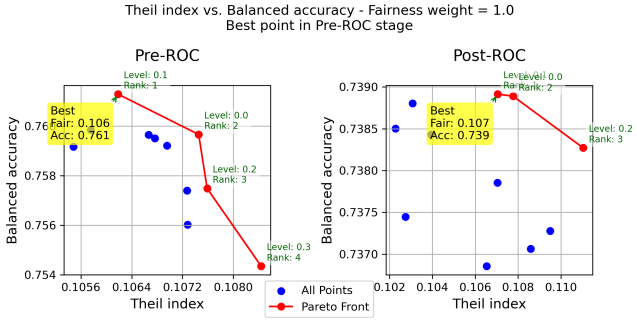

Figure 10: Pareto front pair with the *Theil Index* as its fairness metric.

5 Discussion

Our literature review and data modeling experiment produced numerous findings that generally supported each other but also revealed some differences and surprises. We first explain the key connections between our literature review and experiment findings and their implications, followed by an overview of our work's limitations and their impacts.

Our experiment's findings generally matched what we found in our studied literature. We confirmed the existence of the non-linear trade-off between fairness and accuracy, as seen by Haas [5]. However, as seen by Patrick Janssen et al., despite using different methods, we also found cases where the model accuracy unexpectedly improved after applying fairness methods [8]. Lastly, we noticed that some fairness metrics showed no improvement or even degraded after applying our fairness methods, as noted by Janssen et al., Zou et al., and others [4; 8]. As another example, in our generated *Theil Index* vs. *Balanced Accuracy* front pair, the pre-ROC stage gave the best trade-off, which shows that combining the DIR and ROC techniques is not always optimal.

Our findings led to two main implications. First, since the Pareto fronts vary greatly between fairness-accuracy metric combinations, careful metric selection is critical to prioritize specific fairness and accuracy aspects, ensuring optimal decisions regarding selecting an optimal DIR repair

level and whether also using the ROC gives desired results. For example, Zou et al. mentioned the importance of minimizing false positives to avoid hefty costs for the lenders despite the fairness implications of increased false negatives [4]. Second, while some metric pairs may show minimal trade-offs, others may not perform as well, highlighting the need to consider other combinations to achieve optimal outcomes.

Our contribution has notable limitations. While our literature review uncovered many key fairness aspects and their accuracy trade-offs, our experiment contained three limitations. First, we only tested the DIR pre-processing and ROC post-processing techniques, and different data processing techniques could yield vastly different Pareto fronts and values for the tested metrics. Second, we only tested our implementation on the Adult dataset, and different datasets could yield varying results. Third, we only thoroughly tested using a *Logistic Regression* model, and other model types could behave vastly different depending on their data input and manipulation. Fourth, due to time constraints, we were unable to generate an error bar showing the metric trends between multiple model runs, therefore increasing the margin of error of our findings. These limitations suggest that our findings, including metric findings and generated Pareto fronts, are likely not generalizable for other datasets with different biases and protected attributes.

6 Conclusions

We conclude that our research consistently demonstrates a trade-off between fairness and accuracy, with occasional exceptions where fairness methods negatively impact fairness metrics or enhance accuracy. Our implementation allows for building Pareto fronts on several metric combinations, allowing customizable evaluations of fairness through the implemented pre- and post-processing methods. However, different datasets will likely have different fairness and accuracy metric values from different pre- and post-processing methods.

Answering our research questions: for the first one, the DIR pre-processing performed modestly on most of our fairness metrics, and the ROC post-processing was substantially more effective with optimal results often combining both at lower DIR repair levels. Answering our second question, we generate Pareto fronts with metrics of choice, using a scoring function that highlights the best trade-off point with its respective DIR repair level and whether the ROC qualifies. Answering our final question, our implementation can be adapted for other datasets, models, and fairness methods by changing its dataset parser, selected model (it has 6 to choose from), and methods as desired, therefore being generalizable.

Our contribution gives plenty of room for future work. We could explore the impacts of different combinations of pre- and post-processing techniques, including in-processing adjustments directly to the model. Adding more fairness and accuracy metrics would allow for more thorough testing, along with applying our implementation to other datasets with varying biases and considering impacts on other protected attributes; modifying our dataset reading function would allow for easy adaptation to new datasets. Our framework can also be tested on any of the five other supported models or with an added one, such as experimenting with the *K-Nearest Neighbors* or *Support Vector Machine (SVM)* classifier, which we found sometimes showed promising results compared to *Logistic Regression* and could reveal additional research opportunities. Finally, having multiple model runs to create error bars showing the metric trends would reduce the margin of error and therefore increase the generalizability of the findings.

7 Responsible Research

Our research has various ethical implications that will likely influence how data and algorithms are evaluated and assessed for fairness and accuracy. To start, we used ChatGPT to help with paraphrasing, code augmentation, and debugging. All generated content was thoroughly checked for errors to ensure integrity. Additionally, to ensure that informed and thoughtful decisions are made based on our findings, we will discuss the ethical integrity and reproducibility of our methodology.

Various ethical aspects relate to both when running our implementation and interpreting its results. During the data processing, one should ensure that it does not compromise the dataset integrity or introduce new biases, which may invalidate the findings if not dealt with properly, and keep in mind that the metrics selected to generate the Pareto fronts can have significant ethical implications on influenced decisions. Therefore, including a broad metric range to cover different ethical perspectives is advised. Interpreting the results should also be taken with care, where decisions regarding the trade-offs should not rely solely on the generated Pareto fronts but also consider metrics or features outside the experiment's scope. Engaging with relevant domain experts and stakeholders to assess the real-world impacts of the findings is also advised, as such decisions will likely impact large stakeholder groups (such as those for employment, banking, and more).

Our methodology's reproducibility is well-supported. Our literature review is fully replicable by examining the relevant papers of the different tested techniques and metrics. Regarding our experiment setup, we concisely outline it in subsection 3.2. Our GitHub repository[2] contains both a README file with comprehensive implementation details and thoroughly commented code files, allowing users with some knowledge in the field to follow along. The code is easily customizable, allowing for the selection of different fairness and accuracy metrics for the Pareto fronts, selecting the ROC guiding metric, choosing different models, and modifying graph formatting to thus enhance its adaptability for various research needs.

[2]Link to repository: https://github.com/1Sulture/Fairness-vs.-Accuracy-Pareto-Front-Builder

References

[1] Simon Friis and James Riley. Eliminating algorithmic bias is just the beginning of equitable ai, Sep 2023.

[2] Stefan Buijsman. Navigating fairness measures and trade-offs. *AI and Ethics*, July 2023.

[3] Shizhou Xu and Thomas Strohmer. Fair data representation for machine learning at the pareto frontier, Nov 2023.

[4] Leying Zou and Warut Khern-am nuai. Ai and housing discrimination: the case of mortgage applications. *AI and Ethics*, 3(4):1271–1281, November 2022.

[5] Christian Haas. The price of fairness - a framework to explore trade-offs in algorithmic fairness, 2019.

[6] M. Akbari, P. Asadi, M.K. Besharati Givi, and G. Khodabandehlouie. 13 - artificial neural network and optimization. In Mohammad Kazem Besharati Givi and Parviz Asadi, editors, *Advances in Friction-Stir Welding and Processing*, Woodhead Publishing Series in Welding and Other Joining Technologies, pages 543–599. Woodhead Publishing, 2014.

[7] Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized data pre-processing for discrimination prevention, Apr 2017.

[8] Patrick Janssen and Bert M. Sadowski. Bias in algorithms: On the trade-off between accuracy and fairness, Jan 2021.

[9] Preetam Nandy, Cyrus Diciccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. Achieving fairness via post-processing in web-scale recommender systems, Aug 2022.

[10] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairness improvement with multiple protected attributes: How far are we?, Apr 2024.

[11] Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: Fairness versus accuracy: Proceedings of the 23rd acm conference on economics and computation, Jul 2022.

[12] Susan Wei and Marc Niethammer. The fairness-accuracy pareto front, Nov 2021.

[13] Lele Sha, Dragan Gašević, and Guanliang Chen. Lessons from debiasing data for fair and accurate predictive modeling in education, May 2023.

[14] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, Jul 2015.

[15] Barry Becker. Adult, 1996.

[16] Yuhong Luo, Austin Hoag, and Philip S. Thomas. Learning fair representations with high-confidence guarantees, Oct 2023.

[17] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[18] Samuel Hoffman. Ai fairness 360 documentation, 2018.

[19] Paul Meier, Jerome Sacks, and Sandy L. Zabell. What happened in hazelwood: Statistics, employment discrimination, and the 80