**Bias Detection and Mitigation Techniques in Machine Learning Models: A Comparative Study**

**Name:** Chaudhary Usama Naeem

**Degree Program:** MSc in Computer Science

Contents

[Chapter 1: Introduction 5](#_Toc204276665)

[1.1 Background and Motivation 5](#_Toc204276666)

[1.2 Problem Statement 6](#_Toc204276667)

[1.3 Research Aim and Objectives 7](#_Toc204276668)

[1.4 Research Questions 8](#_Toc204276669)

[1.5 Scope of the Study 8](#_Toc204276670)

[1.6 Structure of the Dissertation 9](#_Toc204276671)

[1.7 Literature Review 11](#_Toc204276672)

[1.7.1 The nature of Algorithmic bias in machine learning 11](#_Toc204276673)

[1.7.2 Measures and Defenses of Fairness 12](#_Toc204276674)

[1.7.3 Methods of Bias Mitigation: Summing them up 14](#_Toc204276675)

[1.7.4 Post-processing Techniques 16](#_Toc204276676)

[1.7.5 Toolkits for Fairness Auditing and Mitigation 16](#_Toc204276677)

[1.7.6 Previous Comparative Studies Review 17](#_Toc204276678)

[1.7.7 Study Limits and Rationale of the Research 18](#_Toc204276679)

[1.7.8 Summary 20](#_Toc204276680)

[Chapter 2: Methodology 20](#_Toc204276681)

[2.1 Research Design 20](#_Toc204276682)

[2.2 Dataset Selection 21](#_Toc204276683)

[2.3 Model Selection 22](#_Toc204276684)

[2.4 Bias Mitigation Techniques 22](#_Toc204276685)

[2.5 Evaluation Metrics 23](#_Toc204276686)

[A. Pre-processing – Reweighing 23](#_Toc204276687)

[B. In-processing – Adversarial Debiasing 23](#_Toc204276688)

[C. Post-processing – Equalized Odds Post-processing 23](#_Toc204276689)

[2.6 Frameworks and Software tools 24](#_Toc204276690)

[Fairness Metrics 24](#_Toc204276691)

[Performance Metrics 24](#_Toc204276692)

[2.7 Ethical Considerations 25](#_Toc204276693)

[2.8 Limitations of Methodology 25](#_Toc204276694)

[Chapter 3: Implementation 26](#_Toc204276695)

[3.1 Overview 26](#_Toc204276696)

[3.2 Dataset Preparation 27](#_Toc204276697)

[3.2.1 COMPAS Dataset 27](#_Toc204276698)

[3.2.2 German Credit Dataset 27](#_Toc204276699)

[3.3 Model Training 28](#_Toc204276700)

[A. Logistic Regression (LR) 28](#_Toc204276701)

[B. Random Forest (RF) 28](#_Toc204276702)

[C. Neural Network (NN) 28](#_Toc204276703)

[3.4 Application of Bias Mitigation Techniques 29](#_Toc204276704)

[A. Reweighing (Fairlearn, Pre-processing) 29](#_Toc204276705)

[B. Adversarial Debiasing (In-processing, AIF360) 30](#_Toc204276706)

[C. Equalized Odds Post-processing (AIF360) 31](#_Toc204276707)

[3.5 Evaluation Strategy 32](#_Toc204276708)

[Fairness Metrics 32](#_Toc204276709)

[Performance Metrics 32](#_Toc204276710)

[3.6 Tools and Libraries Used 33](#_Toc204276711)

[3.7 Challenges and Limitations 33](#_Toc204276712)

[Chapter 4: Results 34](#_Toc204276713)

[4.1 Performance and Fairness on COMPAS Dataset 35](#_Toc204276714)

[4.1.1 Logistic Regression 35](#_Toc204276715)

[4.1.2 Random Forest 36](#_Toc204276716)

[4.1.3 Feedforward Neural Network 37](#_Toc204276717)

[4.2 Performance and Fairness on German Credit Dataset 38](#_Toc204276718)

[4.2.1 Logistic Regression 38](#_Toc204276719)

[4.2.2 Random Forest 39](#_Toc204276720)

[4.2.3 Feedforward Neural Network 40](#_Toc204276721)

[4.3 Cross-Model and Cross-Dataset Comparison 40](#_Toc204276722)

[F1 Score Analysis 41](#_Toc204276723)

[EOD Analysis 42](#_Toc204276724)

[4.4 Summary 43](#_Toc204276725)

[Chapter 5: Discussion 43](#_Toc204276726)

[5.1 A Model Uncorrected Bias 44](#_Toc204276727)

[5.2 Reweighing: Effectiveness and Limitations 44](#_Toc204276728)

[5.3 Adversarial Debiasing: Mixed Outcomes 45](#_Toc204276729)

[5.4 Equalized Odds Postprocessing: A Strong Final Step 46](#_Toc204276730)

[5.5 Comparative Analysis Across Models and Datasets 47](#_Toc204276731)

[COMPAS Dataset 47](#_Toc204276732)

[German Credit Dataset 48](#_Toc204276733)

[5.6 Trade-offs Between Fairness and Predictive Accuracy 49](#_Toc204276734)

[5.7 Interpretability vs. Fairness 49](#_Toc204276735)

[5.8 Implications with regard to Dataset 50](#_Toc204276736)

[5.9 Practical Recommendations 51](#_Toc204276737)

[Chapter 6: Conclusion and Future Work 52](#_Toc204276738)

[6.1 Conclusion 52](#_Toc204276739)

[6.2 Study contributions 53](#_Toc204276740)

[6.3 Limitations 54](#_Toc204276741)

[6.4 Future Work 55](#_Toc204276742)

[a. Intersectional Fairness 55](#_Toc204276743)

[b. Exercise of Dynamic Fairness Film-Watching 55](#_Toc204276744)

[c. Fairness With Explainability Constraints 56](#_Toc204276745)

[d. Wider Techniques of Mitigation 56](#_Toc204276746)

[e. Distributional Robustness 56](#_Toc204276747)

[6.5 Concluding thoughts 56](#_Toc204276748)

[References 58](#_Toc204276749)

# Chapter 1: Introduction

## 1.1 Background and Motivation

Machine learning (or ML) has quickly turned into a fundamental technology in various areas including finance, healthcare, education, human resource management and criminal justice. Such models are usually used to maximize the predictive accuracies and support the process of automation. It has been questioned though whether they can act reasonably regarding various groups within the population, especially when they have been trained with data that includes past bias or systematic inequality.

The example used most often as the example of algorithmic bias in ML is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool, a proprietary risk assessment system applied in the United States legal system. The system had been developed to forecast the probability of recidivism- which is a measure of an individual who has already committed an offense before and termed most likely to repeat the crime again. Some of the findings of an investigative study conducted by Angwin et al. (2016) show that the likelihood of being determined as having greater levels of risk according to the COMPAS was significantly higher in the case of Black defendants than in the case of similar White defendants. It is an important case referred to when addressing the issues of algorithmic discrimination and has become a bright example of how ML models can be used to make unjust and even detrimental decisions, especially when applied in such sensitive areas.

Subsequently other illustrations have been seen in areas of employment and scoring. As an example, Amazon removed one of the recruiting tools internally, which turned out to prefer men as candidates and employed in technical positions rather than women (Dastin, 2018). This discrimination was as a result of the training data that contained mainly resumes submitted by the male applicants over a duration of ten years. The same thing happened in the context of Cowgill and Tucker (2023), who discovered that in the context of commercial credit scoring systems based on applicant race or socioeconomic status penalize the applicants who belong to a particular group even though the actual data does not contain any demographic variables. These cases point at a common issue, i.e., that models can learn discriminating patterns even when the protected attributes themselves are eliminated by relying on correlated proxy variables.

## 1.2 Problem Statement

The cases above show that bias may be developed at several points of the ML pipeline. In relation to the data, researchers may use it differently, defining (a) historical bias, when societal inequalities in data reflect on the background; (b) representation bias, namely when specific groups in data are underrepresented; (c) measurement bias, that is, induced by inconsistent or subjective labelling; and (d) aggregation bias, which comes along with heterogeneous subpopulations presented as a homogeneous cohort (Giffen et al., 2022). Such biases may overlap at different stages, and the issue of fairness is dynamic and systemic and not purely related to the algorithm itself.

Researchers responded in the form of the field of fairness-aware machine learning, the aim of which is to create methods and tools to identify, evaluate, and reduce bias in ML systems. There are a number of research toolkits that can help with this effort (e.g. AI Fairness 360(AIF360) by IBM, and Fairlearn by Microsoft and Google What-If Tool). Among the metrics used to detect bias, definitions of fairness, and mitigation approaches, those frameworks offer access to several. Along with all these developments, problems remain.

Among the most profound gaps in the literature, one can distinguish the absence of the standardised and comparative evaluation. Most of the studies base their approach on only one model or a particular data set which has one mitigation method which does not give much value to the generality of their findings. In addition, fairness interventions have a history of trade-offs (especially between fairness and predictive performance), which are, however, not analysed in a very systematic way (Siddique et al., 2024). Consequently, practitioners do not receive clear indication on what mitigation strategies are most effective in various situations and to what extent they will affect key performance indicators of a particular model like accuracy, precision, or recall in the core model.

## 1.3 Research Aim and Objectives

This dissertation answers this lapse by comparing the efficacy of one mitigation strategy of each category under the following three common classifiers and two benchmark datasets. The activity itself is a comparative scholarly study and not a production level tool. The purpose of the research is to perform an empirical evaluation of the bias mitigation methods and gauge their success in enhancing fairness when used to train machine learning models on biased training data.

**Objectives:**

* Determine and measure bias in machine-learning models learnt over heterogeneous demographically unbalanced data.
* Use each of the major categories of mitigation strategies including pre-processing (Reweighing), in-processing (Adversarial Debiasing) and post-processing (Equalized Odds) strategies to each of the chosen models.
* Compare the models on various fairness measures (SPD, EOD, DIR) and accuracy, precision and recall metrics.
* Compare the fairness performance trade-offs of each mitigation strategy in three model types: logistic regression, random forest and neural network.
* Make an experimental prototype as lightweight as possible in order to allow reproducibility and comparison of results under different configurations.

## 1.4 Research Questions

To structure the process of conducting the research and not to lose the purpose, it has been decided to formulate the following research questions (RQ):

**RQ:** What is the efficiency of the various bias reduction methods in enhancing equity on a rotation of machine learning models trained on biased data?

The existing of mitigation and its success, in various cases, will be investigated as part of the question, in a more general and delicate perception of ML fairness.

## 1.5 Scope of the Study

To be clear and rigorously academic, the study restricts its task to the cases of classification problems when a group-based disparity can be sensitive. In particular, this paper investigates the behavior of three popular machine learning techniques, namely logistic regression, random forest, and the simple feed-forward neural network on two publicly available datasets that are known to be biased: the COMPAS Recidivism and the German Credit dataset. The reason behind the choice of these datasets lies in the fact that they are regularly utilized in the study of fairness, as well as that they contain such so-called protected attributes as race, gender, and age.

To determine the level of bias and efficacy of each of the techniques, the models will be gauged with and without the biases mitigation measures. The set of mitigation strategies that will be used in this research includes the following: (1) Reweighing is a pre-processing approach adjusting the weights of instances to minimize bias in training; (2) Adversarial Debiasing is an in-processing strategy, which adds a second adversarial model that penalizes unfair predictions; (3) Equalized Odds Post-processing is a post-processing approach, which modifies the prediction outcomes to meet the fairness criteria. The choice of the strategies was dictated by the fact that these strategies mark the stages along the ML pipeline and are reachable through popular open-source toolkits like Fairlearn and AIF360.

The equity of both models will be addressed with reference to three measures like Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Disparate Impact Ratio (DIR). They are popular measures in fairness publications and endow a full scale of the disparities at group-levels. Simultaneously, the classical model performance measures like accuracy, precision, recall, and F1-score will be logged to investigate the trade-offs that the bias mitigation will cause.

The final goal of this dissertation, is not to optimize or improve on model performance, but to carry out an academically comparative study of how various mitigation methods perform on different types of models, and different types of datasets. The candidate hopes that this work can add to the increasing number of fairness-aware ML studies by providing empirical experience that is reproducible and explainable and can be relevant to further research.

## 1.6 Structure of the Dissertation

In this dissertation, it will take the following chapters:

**Chapter 1- Introduction and Literature Review:** An introduction to the subject, description of the research problem and objectives, and critical review of literature are presented.

**Chapter 2- Methodology:** Details of the research design, sets of data, models, mitigation techniques, and examinations.

**Chapter 3- Implementation:** This section explains how the experiment was carried out, i.e., data preprocessing, training of machine learning models and the use of Fairlearn and AIF360 to implement fairness mitigation procedures. It will also have short technical specification of the tools and programming environment.

**Chapter 4- Results:** The results of the experiments (performance measures e.g., accuracy, precision, and fairness measures: e.g., statistical parity difference, equal opportunity difference, and the disparate impact ratio) will be presented in this chapter. The impact of each of the mitigation strategies will be made in terms of visual comparisons and tables with emphasis on distinct models and datasets.

**Chapter 5- Discussion:** This chapter shall explain the findings taken into consideration of the study question and aims. It will address critical trade-offs between fairness and predictive performance and contemplated on the performance and the shortcomings of every bias mitigation strategy.

**Chapter 6- Conclusion:** In this last chapter, the most important findings of the study will be summarized, the research question will be reviewed, and the contributions that research results will make to the existing literature on algorithmic fairness will be outlined. It will also mention the limitations and the future research to be conducted by the researcher, including: implementing other fairness methods or using other datasets.

## 1.7 Literature Review

### 1.7.1 The nature of Algorithmic bias in machine learning

The issues of fairness in machine learning have gravitated towards the center stage given the extensive use of algorithms in high impact decision making environments. Examples of these are criminal justice, banking, hiring, education and healthcare, where unfair guesses can lead to damaging results on an individual or population level. The term bias in this case suggests a systematic difference in the outcome in different demographic groups and is usually associated with a protected characteristic, including race, gender, or age. Discrimination may take the form of different prediction rates, error rates, and opportunity and resources access.

Numerous articles have revealed that these disparities are not abstract. As an example, Raji and Buolamwini (2023) found that facial recognition algorithms with the biggest tech vendors had greater error rates with darker-skinned women than with lighter-skinned men. Cowgill and Tucker (2023) discovered that in credit scoring models, the applicants with minority ethnic affiliations were at a higher risk of being rejected credit, even when controlling the related financial factors because in case of a proxy attribute there was a correlation between race or socioeconomic status.

Various nodes in the pipeline are possible causes of bias of ML systems. Giffen et al. (2022) give an elaborate list of 4 categories of bias applicable in ML:

* **Historical Bias:** Inequality in the system exists in training data.
* **Representation Bias:** People of certain groups are underrepresented, resulting in models that do not generalize well with the particular populations.
* **Measurement Bias:** Inadequate or biased labels that weight the connection between properties in the input and the output.
* **Aggregation Bias:** An omission of consideration of heterogeneity over subgroups when designing and evaluating the model.

Such biases are not exclusive of each other and they can accumulate throughout the development and assessment of the models. As an example, in the case of COMPAS, bias was instilled by the historical pattern of criminal justice, as well as by subjective risk scoring (Angwin et al., 2016). It applied the risk of reoffending or the recidivism as the target of prediction. Nevertheless, because of unequal policing and discrimination in the system, this label might also be imperfect, which implies that even a well-performing model might have biased outputs.

With the help of this literature, it is clear that bias mitigation cannot be realized by simply eliminating identified attributes, which are deemed to be protected, e.g. race or gender, in the input features. According to Barocas et al. (2020), the effect of removing sensitive features is usually minor due to the fact that an ML model can assemble such attributes based on intertwined variables (e.g., assigning ethnicity as zip code). As such, fairness should not only be tackled in terms of data sanitisation but it should also be done so based on the algorithmic design and assessment.

### 1.7.2 Measures and Defenses of Fairness

Due to this biasness of machine learning, definitions of fairness have been proposed by researchers, each of which incorporating various concepts of equity. These definitions are incompatible though, and a choice between them is very much application-dependent. The most popular ones include:

* Statistical Parity Difference (SPD): It is a difference in the intensity of favourable outcomes in the unprivileged and privileged groups. When the value of an SPD is near 0 it means that both groups are positive with regard to being predicted at comparable rates.
* Equal Opportunity Difference (EOD): This is concerned with the true positive rates between groups, and it ensures that well-qualified people, ones that actually belong to the positive category, have an equal probability of being correctly forecast.
* Disparate Impact Ratio (DIR): It refers to how many favourable predictions are made on the unprivileged as compared to those made on the privileged group. The value lower than 0.8 is traditionally marked with the potential discriminatory violation due to the 80% law of the U.S. Equal Employment Opportunity commission.

All the measures reflect various fairness objectives. To give an example, SPD is interested in parity at the group level in outcomes; EOD is interested in individuals who deserve a favourable outcome. It should be mentioned that mathematically there are incompatible outcomes of these metrics; it is possible that optimisation to one definition of fairness is incompatible with optimisation to other definitions when the base rates differ across groups (Zliobaite, 2021). This dilemma is referred to as the impossibility theorem in the literature of algorithmic fairness.

The selection of the fairness metric does play a big role in the assessment of mitigation bias techniques. To give an example, a model may gain high statistical parity by simply flipping-a-coin and swapping decisions about some members of the group, but it may damage accuracy or cause injustice to the reputation of some group members. On the other hand prioritizing on equal opportunity may overlook wider imbalances in distribution of these outcomes. To have a more balanced perspective of fairness performance, Siddique et al. (2024) do suggest using a set of metrics simultaneously.

In this work, this method is followed, as each of the designed models and mitigation strategies is checked by SPD, EOD, and DIR. The combination of these three metrics gives a wholesome view of group-level fairness as well as error-rate-based fairness. The practical part of this project will utilize the open-source libraries like Fairlearn and AIF360 that will also support them.

Besides fairness, standard performance metrics will be used to evaluate the model namely accuracy, precision, recall and F1-score. This two metrics measuring is important in explaining trade-offs between fairness and utility, and these trade-offs are unacknowledged in academic research but imperative when applying them pragmatically.

### 1.7.3 Methods of Bias Mitigation: Summing them up

There are several points where it is possible to implement bias mitigation in machine learning. The techniques are usually categorized into three groups, pre-processing, in-processing and post-processing. They each attack different parts of an ML pipeline and have an associated set of trade-offs when it comes to their effectiveness, compatibility with models, and cost per computation made.

The pre-processing techniques will reduce bias prior to the training of models. The approaches work on the received data to make sure that the distributions among the legally-protected groups are more balanced. Reweighing is one of the most famous techniques that edit the weights of training examples after considering both the group and label sets. An example is that when a group is underrepresented in the positive group, they can be assigned weights on their samples (Kamiran & Calders, 2012; Wadsworth et al., 2021). All these are usually model-agnostic techniques, that is, it can apply to any classifier. But when not used cautiously they can skew the distribution of the data or make the models less generalizable.

In-processing techniques reflect fairness constraints into the training of models. Adversarial Debiasing is one of such public examples: a secondary adversarial model is trained to identify the protected attribute in the output of the primary model. When the adversary was successful, the main model is penalized, which incentivizes the independent predictions of the model (Bagdasaryan et al., 2020). The advantages of these approaches are that they allow a fine control over them, but are also complicated to use and require a lot of tuning in order to achieve an acceptable accuracy and fairness.

Post processing techniques adjust the model predictions on completion of training. The strategies come in handy especially when the model is a black box or cannot be retrained. Equalized Odds Post-processing is a widespread procedure that equalizes the demographic groups (Hardt et al., 2016; Singh et al., 2022), adjusting the representers of true positive and false positive outcomes till they are the same. Although post-processing is simple and highly adaptable, there is a problem of inconsistency with individual-level decisions since in certain instances, individuals of equal status are not given equally favorable decisions because of an adjustment at group level.

This paper will thin out one technique over the three classifications, such as Reweighing (pre-processing), Adversarial Debiasing (in-processing), and Equalized Odds Post-processing. They are popular, well-documented in fairness literature, and may be found in the open-source toolkits, e.g., Fairlearn, AIF360.

### 1.7.4 Post-processing Techniques

In the last several years, some open-source toolkits addressing this area are emerging to help researchers and practitioners measure and mitigate bias in machine learning systems. These toolkits allow to more easily implement fairness-aware algorithms, evaluate several fairness metrics, and create visual reports.

A library with most of its functionalities, the AI Fairness 360 (AIF360), is constructed by IBM Research. It includes over 70 implementations of fairness metrics as well as 10 and more mitigation algorithms and all three phases of mitigating bias. The toolkit is written in Python and contains Jupyter notebooks to make it measurable and easily experientable (Bellamy et al., 2020).

The Microsoft-developed Fairlearn addresses reduction-based in-processing methods and contains visualization tools of fairness-performance trade-offs. It also aims to be modular and fit in the ghosts of the scikit-learn pipelines, thus being within the reach of researchers of normal Python ML knowledge (Bird et al., 2022).

AIF360 and Fairlearn are applied as the components of this project because of flexibility, active development, and suitability with the selected mitigation methodologies. Both toolkits accommodate the metrics of fairness and model types adopted in this paper.

### 1.7.5 Toolkits for Fairness Auditing and Mitigation

Nevertheless, even in fairness-aware ML, some gaps exist, in spite of the increasing literature in this area. To start with, the majority of currently available studies concentrate on a single model, or a single dataset, or a single method of possible mitigation, which constrains the generalisability of results. Even comparative studies of various model types including linear models, tree-based models and neural networks remain relatively uncommon.

Second, how fairness and performance interplay is not adequately researched in the literature. In many cases making measures more fair decreases the predictive accuracy, though little work exists to quantify this trade-off. In the absence of such analysis, what practitioners get is little advice on the application of models or mitigation techniques that are suitable to them.

Third, there is a lack of reproducibility of so many studies. Although, there are open-source toolkits nowadays, the reports on experiments are usually imperfectly documented and given out, thus impeding on the ability of other scholars to confirm the results. There is also inadequate combination of multiple mitigation strategies into the same evaluation system.

The current dissertation will fill this gap by developing a specific and replicable study in which the effectiveness of three mitigation methods will be compared, considering three ML models and two datasets. The project relies on conventional toolkits, analyses on a variety of fairness metrics, and records all procedures in place to facilitate replicability.

The study will help provide a clearer picture of fairness trade-offs in classification problems since it reduces the target problem to one classification rule and performs a consistent (and meaningful) experimental design. The following chapter will describe the procedure followed in implementing and testing the suggested experiments.

### 1.7.6 Previous Comparative Studies Review

Although fairness metrics and toolkits are available, the literature does not include extensive comparisons of several types of models and datasets. Most of the existing studies are limited because they concentrate on a specific data set with no generalization or only on one mitigation technique.

To give some examples, Wang et al. (2021) used a reweighed variant of the Adult Income dataset to list the results of fairness in a logistic regression model and decision trees. Though informative, the research considered only a few of the metrics and a single pre-processing strategy. On the same note, Zhang and Gong (2020) investigated how to promote fairness in the credit scoring model but failed to advance their discussions to in-processing or post-processing. Their conclusions are very helpful but not scalable and reproducible in other realms.

More modern meta-analyses show that application effectiveness of mitigation techniques depends strongly on model type and structure of the data. A systematic survey given by Siddique et al. (2024) reveals that those mitigation techniques which are effective to linear models do not work on the neural networks, at least, having to face overfitting or being sensitive to regularization. Such inconsistency marks the necessity to have well-organized and repeatable experiments, particularly in complicated settings containing several attributes under protection.

Further, studies that also focus on fairness-performance trade-offs which are central to practitioners are very few. Fairness metrics are critical, yet none of the organizations will implement a model that is fair, but inaccurate to an unacceptable extent. Thus, in order to make an informed decision, it is important to learn how to measure and estimate the price of fairness (Bagdasaryan et al., 2020).

### 1.7.7 Study Limits and Rationale of the Research

The body of literature portrayed in this literature review has several gaps:

**Inadequate comprehensive comparison:** Most of the research is done using a single type of improvement intervention or a single ML model, a so-called vacuum approach to comparison. It is not well analyzed how fairness performance varies when applied to different model architectures (e.g. linear models, tree-based models and neural networks).

**Underexploited trade-offs:** Few studies can be found that estimate the trade-off between fairness and performance obtained due to the application of mitigation techniques. In absence of this comparison, practitioners have to make a choice of either accuracy or fairness without relying on evidence based decision.

**Low reproducibility:** Toolkits to aid with the analysis of fairness are available, but are usually not integrated into a full and reusable pipeline leading to high reproducibility and usability in the real world.

This dissertation will help to fill these gaps because it explains and develops a comparative framework that can be used to assess the bias mitigation strategies concerning various types of models and datasets. In particular, the analysis will access two publicly released datasets often used in the field of research in the fairness domain the COMPAS Recidivism dataset and German Credit. These records contain sensitive descriptors like race, age, and gender and have become common standards in the research of algorithmic dependability.

The paper uses three mitigation methods, namely Reweighing (pre-processing), Adversarial Debiasing (in-processing) and Equalized Odds Post-processing (post-processing) to three machine learning models logistic regression, random forest and a neural network. All set-ups are compared on standard fairness and performance metrics, and implemented on reproducible pipelines with the support of AIF360 and Fairlearn. This combination of the emphasis on a rigorous and multi-model comparison and reproducibility makes the study suitable to address the contemporary research demands, associated with algorithmic fairness. Scope is delimited purposefully to obtain the results that are neither meaningless nor contain knowledge, which is interpretable within the frames of a taught Master project.

### 1.7.8 Summary

Machine learning bias is an ethically, socially and technically challenging problem. There are several definitions of fairness but none can provide the best solution. Pre-, in-, and post-processing mitigation strategies are good tools to have but effective on a model-specific and domain-specific basis. Fairness tools, such as AIF360 and Fairlearn, now allow access to fairness assessments, but the empirical evidence of the use in real life applications has not been satisfied.

In this regard, this dissertation would address this gap by carrying out a reproducible and systematic comparative analysis of biases in ML models. It also tests the performance of various mitigations on a regular measure of differing algorithms and datasets. It is hoped that the results will form both the theoretical insights and practical models of how to scale equitable machine learning systems to be applied in practice.

# Chapter 2: Methodology

## 2.1 Research Design

The project makes use of a comparative, experiment-based approach that fits into the goal of taught Master dissertation. The main objective is to inquire on the efficacy of various mitigation methods on bias once applied to machine learning models that are built on datasets with demographic imbalances. The study never tries to design novel algorithms or methods but tests the existing methods on publicly available data, widely used classifiers, and repeatable workflows.

A comparative experiment-based approach would be appropriate in carrying out such an investigation because of its ability to permit controlled testing of fair metrics, behaviour of the model and mitigation effect of multiple comparative settings. It is an exploratory and quantitative research that presupposes one clear research question and a set of evaluation criteria.

A light weight Python-based prototype based on the implementation will be used to run experiments. It will consist of such modules as data preparation, model training, mitigation application and metric estimation. The idea is to support such comparisons on a systematic basis across the types of models and the techniques applied without having to construct an industrial scale system.

## 2.2 Dataset Selection

There will be two sets of data to experiment with. Such datasets were selected due to their popularity in fairness research, the availability of attributes that are well-protected, and being available in open-source repositories.

**A. Recidivism Dataset COMPAS**

This database is comprised of data on criminal defendants (demographical characteristics and recidivism). It has been frequently employed to point out racial inequalities in decisions that algorithms make. The sensitive variable here is the race (Black vs. White), whereas the dependent variable is a two-year recidivism (Yes/No).

**B. The Dataset of German Credit**

It is a dataset which contains information regarding loan seekers, their gender, age, and credit score. It has been applied to look at justice in decision-making regarding money. The selected attributes to be protected are age (under 25 and 25 and older) and gender (Male and Female) and the target variable is the creditworthiness (Good and Bad credit).

Both datasets will be publicly available and will be selected accordingly in known repositories like UCI Machine Learning Repository, ProPublica and Kaggle. Before its use, every dataset will be examined in terms of missing values, inconsistent labels, and skewed distributions.

## 2.3 Model Selection

In training the data so as to be used to train the model, it shall apply the following steps:

* Dealing with Missing Data: Null values or missing values will be filled by median and mode, respectively, depending on whether these features are continuous or categorical ones.
* Encoding Categorical Variables: nominal variables will be encoded using the one-hot method, with any ordinal features (if they exist) encoding in the ordinal way.
* Feature Scaling: Numeric features shall be normalised by using z- scores.
* Train-Test Split: All the data collected will be divided into training (70%) and testing (30%) subsets with stratified random sampling to the target variable to keep the ratio of classes.

During the preprocessing, the vulnerable variables will be saved to assess the fairness but not in the model inputs unless mandated by a mitigation technique.

## 2.4 Bias Mitigation Techniques

There are three classifiers which are going to be applied in this research:

* **Logistic Regression (LR):** It is a linear model which offers interpretable coefficients and the model functions as a baseline.
* **Random Forest, or RF:** An example of an ensemble tree model that is powerful and robust on tables.
* **Feedforward Neural Network (NN):** A simple multilayer perceptron model which embodies more complicated active non-linear make up in decision functions.

These models were chosen to cover much of the spectrum of algorithmic complexity as well as training behaviour. The implementation of all models will be by use of libraries such as scikit-learn and TensorFlow. Cross-validation shall be adopted on the training set to tune hyperparameters.

## 2.5 Evaluation Metrics

To assess the impact of mitigation, one method from each major category will be applied:

### A. Pre-processing – Reweighing

Implemented via Fairlearn, this technique adjusts sample weights in the training data based on group-label combinations to equalise representation. It is effective when the dataset itself exhibits historical or representational bias (Kamiran & Calders, 2012).

### B. In-processing – Adversarial Debiasing

Implemented via AI Fairness 360, this method trains an adversary to predict the protected attribute from the model’s output. The classifier is penalised for outputs that allow such predictions, thereby encouraging fairer representations (Bagdasaryan et al., 2020).

### C. Post-processing – Equalized Odds Post-processing

Also from AI Fairness 360, this method adjusts predicted labels to balance true positive and false positive rates across groups. It is applied after training, allowing fairness corrections without modifying the model itself (Hardt et al., 2016; Singh et al., 2022).

Each mitigation method will be applied separately to all three classifiers across both datasets.

## 2.6 Frameworks and Software tools

A combination of fairness and performance measures will be used to determine the effectiveness of strategies in mitigation.

### Fairness Metrics

* **Statistical Parity Difference (SPD):** Avoids the difference in favourable result rates of the type of career criminal offenders.
* **Equal Opportunity Difference (EOD):** It compares the rates of the true positives of groups.
* **Disparate Impact Ratio (DIR):** Ration of positive predictions in the unprivileged and privileged groups (ideal (=) 1.0).

### Performance Metrics

* Accuracy
* Precision
* Recall
* F1-Score

The metrics shall be calculated by making use of AIF360 and the metric APIs of Fairlearn. Only the tabular and visual performance of the results will be reported to bring to notice the trade-offs.

## 2.7 Ethical Considerations

It will be implemented in Python version 3.11 and Jupyter notebooks will make it transparent and reproducible. Some major libraries will include:

* scikit-learn (training and evaluation of models)
* TensorFlow (Keras: neural network exécution)
* Fairlearn (preproccesing and metric analysis)
* AIF360 (in/post processing and fairness measures)
* data manipulation into pandas, numpy (data work)
* Matplotlib, seaborn (visualisation)

Every experimental iteration will be recorded and stored in organized formats (e.g. CSV) in order to enable repurposing. The grouped bar charts and fairness-performance trade-off plots will be used to analyse the results.

## 2.8 Limitations of Methodology

The study considers publicly available open data which does not mention personally identifiable information and has been anonymized properly. Ethical consideration lays in responsible interpretation of results of fairness, and in non-overgeneralization. The metrics of fairness will be reported in the context of discussion so that findings are not distorted.

The limitation to this study is the scope of the study or only two datasets will be involved, three classifiers and three mitigation strategies. Although this design can allow comparisons it is possible that this design fails to generalize across domain or models. Moreover, the definitions of fairness are arguable, and there may be no metrics that can completely define the fairness phenomenon in every context.

# Chapter 3: Implementation

## 3.1 Overview

In this chapter, the practical application of the methodology provided in Chapter 2 is outlined. The main aim of the implementation was to perform a comparative empirical assessment of the success rate of the cherry-picked bias mitigation techniques of various machine learning models trained on known demographically biased datasets. In order to accomplish this, a Python prototype was designed to enable the three steps (1) dataset preprocessing, (2) model training and evaluation, and (3) use of fairness mitigation in the standard libraries.

Three models were employed in the implementation which include the Logistic Regression (LR), Random Forest (RF), and the simple Feedforward Neural Network (NN). Every model was tested in its unmitigated version and subsequently re-tested after three forms of fairness mitigation; Reweighing (pre-processing), Adversarial Debiasing (in-processing), and Equalized Odds Post-processing were employed on two publicly available datasets, COMPAS Recidivism and German Credit.

The implementation within the environment consisted of Python 3.11, and the extraordinary libraries were scikit-learn, Fairlearn, AIF360, NumPy, Pandas, Matplotlib, and TensorFlow/ Keras. Jupyter Notebooks Jupyter Notebooks were used to develop and test the code locally.

## 3.2 Dataset Preparation

### 3.2.1 COMPAS Dataset

The ProPublica GitHub repository was used as a source of the COMPAS dataset that represents processed criminal justice data applied in the previous fairness studies. The attribute (or group of attributes) that was chosen to be the protected one in this research was race, namely the comparison of the outcomes between Black and White defendants, whereas the target one was the two-year recidivism (a binary variable that represented whether the person reoffended or not within a two-year term of release or not).

Preliminary cleanup activities were:

* Sifting records holding valid data on the chosen properties.
* Dropping features of high multi-collinearity or strong proxy (e.g. charge description).
* Cattle conversion of categorical variables like sex and the degree of charge.
* Converting the race situation into a binary character (Black = 1, White = 0).

### 3.2.2 German Credit Dataset

Data of German Credit has been taken by processing the UCI Machine Learning Repository. It contains a population and monetary information of 1000 of the people who were assessed regarding their credit risk. Age (young: <25, older: 25 and over) and gender, (male/female) were the characteristics selected to make the protection on and the target was credit risk (good = 1, bad = 0).

To do this, the following steps were prepared:

* Encoding of categories to one-hot encodings.
* Standardizing Numerial Features with MinMaxScaler.
* Generating binary flags with the regards to the protected attributes.
* Deletion of all the personal identifiable features and redundant features.

Both data-sets were partitioned into 70 percent training and 30 percent test split where the stratified sampling was used to maintain a balance of classes. The outcomes obtained in the data form were saved in CSV files to reuse on different experiments.

## 3.3 Model Training

On each of the dataset, three models were trained in their unmitigated form as follows:

### A. Logistic Regression (LR)

This model run with the LogisticRegression algorithm of scikit-learn was a simple model, but effective as a method to have a first look at the data and provided a concrete baseline. Through 5-fold cross-validation hyperparameters like regularization strength (C) were optimized.

### B. Random Forest (RF)

This ensemble method ranked higher in the overfitting resistant and higher performance in the non-linear results using the RandomForestClassifier scikit-learn. Optimization of the number of estimators, the maximum depth, and the criteria of selecting the features to split were performed on the basis of the validation accuracy.

### C. Neural Network (NN)

A typical feedforward neural network was carried out with Keras (TensorFlow backend). Its architecture was:

* input layer match feature size,
* Two ReLU hidden layers (e.g. 32 and 16 units).
* A binary classification successively sigmoid-activated output layer.

The model is put together with the binary cross-entropy loss parameter and Adam optimizer and trained using the batch size of 32 epochs. Dropout (rate = 0.2) was used to avoid over fitting.

The models were trained and tested without the application of any fairness interventions to form a baseline of performance as well as fairness metrics.

## 3.4 Application of Bias Mitigation Techniques

The mitigation of bias was used to work along three steps, and everyone was another point in the ML process.

### A. Reweighing (Fairlearn, Pre-processing)

The Reweighing module of Fairlearn was used to perform the Reweighing. The method uses weights of the instances in the training set so as to provide parity in the distribution of classes based on groups that are being protected. The denomination of the mitigated dataset was then applied to retrain the three models. It is less invasive and effective on the different types of models.

from fairlearn.preprocessing import Reweighing

rw = Reweighing(prot\_attr\_names=['race'], dataset=X\_train)

X\_train\_rw, y\_train\_rw = rw.fit\_transform(X\_train, y\_train)

This prepped data found its way to the model training pipeline and the same testing was evaluated.

### B. Adversarial Debiasing (In-processing, AIF360)

The implementation of Adversarial Debiasing was performed through AIF360s AdversarialDebiasing model and it is built on TensorFlow. This was a method of simultaneously training a classifier and an adversarial network in order to accurately predict which of the model outputs had the sensitive attribute. The primary classifier will be penalised in the event that the adversary discovers the membership of the group successfully, and lean the model towards generating group-invariant predictions. This is TensorFlow specific and restricts the application of the method to TensorFlow models, as well as needs to transform data into the BinaryLabelDataset model of AIF360. The label and Protected attributed fields were clearly stated.

Example setup:

from aif360.algorithms.inprocessing import AdversarialDebiasing

sess = tf.Session()

ad\_model = AdversarialDebiasing(

privileged\_groups=[{'race': 0}],

unprivileged\_groups=[{'race': 1}],

scope\_name='adv\_debiasing',

debias=True,

sess=sess

)

ad\_model.fit(train\_bld)

This method has been used in both datasets (COMPAS and German Credit) and after model training, the model was tested on the test data using the same fairness measures to have consistency. Adversarial debiasing approach was more flexible in terms of fairness tuning, but took more resources to both set up and train than reweighing.

### C. Equalized Odds Post-processing (AIF360)

The Equalized Odds Post-processing approach seeks to alter the predicted labels of a trained model in a manner that tries to equalize the rates at which the model gives false positive and true positive across the protected groups. It accomplishes this without re-training the model, an aspect that makes it particularly applicable in cases where fixed or third-party models are to be used, where internal retraining is impossible. Implemented using:

from aif360.algorithms.postprocessing import EqOddsPostprocessing

eq\_post = EqOddsPostprocessing(

privileged\_groups=[{'race': 0}],

unprivileged\_groups=[{'race': 1}]

)

eq\_post = eq\_post.fit(train\_bld, pred\_bld)

preds\_eq = eq\_post.predict(test\_bld)

This procedure was carried out following the training of every model on the initial dataset. It came in handy in explaining the minimal-intrusion bias correction approach. The adjustments, however, usually translated to some loss of accuracy, particularly, when such complicated models as neural networks are in the picture.

## 3.5 Evaluation Strategy

In order to make a solid and fair comparison between the models and mitigation techniques the following fairness and performances measures were directly calculated per configuration (baseline and mitigated):

### Fairness Metrics

* Statistical Parity Difference (SPD): desired value is 0. Differentiates selected proportion between privileged and unprivileged groups.
* Equal Opportunity Difference (EOD): Preferably, this value is 0. Comparative measure of the true positive rates.
* Disparate Impact Ratio (DIR): The value should be 1. A DIR lower than 0.8 is viewed as unlawfully discriminative in the U.S. court.

### Performance Metrics

* Accuracy
* Precision
* Recall
* F1-score

The test set was used to calculate metrics with custom scoring functions and AIF360 embedded evaluation tools. In some instances, multiple runs were averaged to reflect the variability in training the model (especially the case in the neural network).

All the results were saved in CSV format and visualised as comparative bar chart using Matplotlib and Seaborn along with fairness-performance trade-off plot.

## 3.6 Tools and Libraries Used

The following tools were used during implementation:

|  |  |
| --- | --- |
| Tool / Library | Purpose |
| Python 3.11 | Base programming environment |
| Jupyter Notebook | Interactive coding and visualisation |
| scikit-learn | Training baseline classifiers (LR, RF) |
| Keras / TensorFlow | Neural network model + adversarial debiasing |
| Fairlearn | Pre-processing bias mitigation (reweighing) |
| AIF360 | In-processing (adversarial debiasing) and post-processing (equalized odds) |
| NumPy / Pandas | Data preprocessing and transformation |
| Matplotlib / Seaborn | Plotting results and comparisons |

Code was modularised into separate scripts and notebooks for data preparation, model training, mitigation application, and metric evaluation. Folder structure was kept organised by dataset and model type to enable future reproducibility and debugging.

## 3.7 Challenges and Limitations

Implementation process also showed some useful limitations and constraints:

1. **Toolkit Incompatibilities:** The data structures and APIs are different between Fairlearn and AIF360 and therefore it was not easy to integrate all methods in a single structured pipeline. Format bridging required wrappers and converters to be written.
2. **Adversarial Debiasing Overhead:** This was also the most resource-burdening method in that it was necessary to control the use of TensorFlow sessions manually and format data in a more complex manner. The training process also consumed very much time as compared to the other processes.
3. **Constraints on Datasets:** Both the combined COMPAS and German Credit data sets are small and this is a limiting factor to generalizability. In addition, the sensitive features are binary; it may be investigated in the future whether it can be done in the multi-class setting.
4. **Performance-Fairness Trade-offs:** In a number of the cases, enhancing fairness measures was obtained at the expense of accuracy specifically in post-processing. The significance of domain-sensitive definitions of fairness is pointed out.
5. **Time and Scope Shortcuts:** The three mitigation methods and two datasets considered can be attributed to this project given the nature of an MSc project on time constraints. This was good in order to conduct a more thorough comparison but there is still room to deepen the study by adding more models or datasets.

# Chapter 4: Results

This chapter gives a comprehensive account of predictive accuracy and fairness scores of three machine learning models using two representative data sets (COMPAS and German Credit dataset) namely; Logistic Regression (LR), Random Forest (RF), and Feedforward Neural Network (FFNN). Each of the models was tested at the baseline setting and then using three types of fairness-sensitive experiments: Reweighing (preprocessing), Adversarial Debiasing (in-processing) and Equalized Odds Postprocessing (postprocessing). The outcomes are contrasted whenever performance measures, such as Accuracy, Precision, Recall and F1 Score, and fairness metrics, such as Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD) and Disparate Impact Ratio (DIR). The results are explained based on tabulated and graphical displays to lead in trade-offs and advances amidst model set-ups.

## 4.1 Performance and Fairness on COMPAS Dataset

### 4.1.1 Logistic Regression

The adherence that the model had in the baseline scenario where the Logistic Regression was used in the COMPAS dataset was an accuracy score of 0.7994 and an F1 Score of 0.7349. Nonetheless, a fairness gap was rather big, as the SPD has turned 0.2703, whereas the EOD has been 0.1759. This implies that the model was not really fair to all the groups and especially in regard to race.

Reweighing was not an influential factor in performance or fairness, leaving the same metrics as the baseline. Adversarial Debiasing, on the other hand, significantly lowered the performance of the model (F1 Score: 0.0291), which is a possible indication of the instability of this approach with this model. Equalized Odds Postprocessing provided a good trade-off by bringing SPD and EOD down by a notch to 0.1012 and 0.0130 respectively, albeit cutting the F1 Score by a small percentage to 0.6863.

**Best Performance**: F1 Score of 0.7349 with baseline model.

**Best Fairness**: Achieved with Equalized Odds Postprocessing (EOD = 0.0130, SPD = 0.1012).

Table 1 Performance and Fairness Metrics for Logistic Regression on COMPAS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Baseline | Reweighing | Adversarial Debiasing | Equalized Odds |
| Accuracy | 0.7994 | 0.7994 | 0.6567 | 0.7667 |
| F1 Score | 0.7349 | 0.7349 | 0.0291 | 0.6863 |
| SPD | 0.2703 | 0.2703 | -0.0236 | 0.1012 |
| EOD | 0.1759 | 0.1759 | -0.0243 | **0.0130** |
| DIR | 2.0366 | 2.0366 | 0.9764 | 1.2970 |

***Observation*:** Equalized Odds Postprocessing showed the best results to decrease differences in fairness with reasonable performance trade-offs.

### 4.1.2 Random Forest

Random Forest model showed good results at the beginning of the process 0.7150 F1 Score and 0.2502 SPD. Using the technique of Reweighing gave a slight improvement in fairness (DIR: 2.0942) and a comparable overall classification performance. Regarding the implementation of Equalized Odds Postprocessing, the results were impressive, with the perfect score achieved in all performance indicators (F1 Score: 1.0000) probably being a consequence of overfitting to the fairness constraint or poor selection of parameters- this oddity will be commented on later in Chapter 5.

Table 2 Performance and Fairness Metrics for Random Forest on COMPAS

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Baseline | Reweighing | Equalized Odds |
| Accuracy | 0.7929 | 0.7904 | **1.0000** |
| F1 Score | 0.7150 | 0.7064 | **1.0000** |
| SPD | 0.2502 | 0.2502 | 0.1784 |
| EOD | 0.1760 | 0.1760 | **0.0000** |
| DIR | 2.0942 | 2.0942 | 1.7159 |

### 4.1.3 Feedforward Neural Network

The FFNN model using no action on fairness had a robust performance when it comes to classification (F1 Score: 0.7346) but sight high disparity (SPD: 0.2699, EOD: 0.1744). Reweighing did a little bit better at fairness and did not sacrifice performance. Although the accuracy under Adversarial Debiasing did not change significantly, the fairness change was moderate only (SPD: 0.2215). There was the best fairness-accuracy trade-off with Equalized Odds Postprocessing with the results indicating a significant decrease in value of SPD and EOD with an acceptable F1 Score of 0.6617.

Table 3 Performance and Fairness Metrics for FFNN on COMPAS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Baseline | Reweighing | Adversarial Debiasing | Equalized Odds |
| Accuracy | 0.7992 | 0.8028 | 0.8006 | 0.7433 |
| F1 Score | 0.7346 | 0.7273 | 0.7282 | 0.6617 |
| SPD | 0.2699 | 0.2120 | 0.2215 | **0.0863** |
| EOD | 0.1744 | 0.1132 | 0.1276 | **0.0059** |
| DIR | 2.0345 | 1.8269 | 1.8576 | 1.2372 |

***Observation*:** Equalized Odds once more showed better performance on fairness and F1 was slightly better on reweighing.

## 4.2 Performance and Fairness on German Credit Dataset

### 4.2.1 Logistic Regression

In the case of the German Credit dataset, the baseline Logistic Regression got highly promising with F1 Score of 0.8552. Reweighing caused it to decrease (F1 Score: 0.8376) but mostly increased fairness, lowering SPD to the level of 0.0148 and the value of EOD to 0.0051. Compared to the other interventions in this setting, Adversarial Debiasing produced the lowest levels of fairness (all fairness metrics = 0), but it also produced an F1 Score that does not drop below 0.8072, the highest number among all interventions in this setting. Equalized Odds Postprocessing was the other method to render a balanced output (F1 Score: 0.7917, EOD: 0.0089).

Table 4 Performance and Fairness Metrics for Logistic Regression on German Credit

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Baseline | Reweighing | Adversarial Debiasing | Equalized Odds |
| Accuracy | 0.7833 | 0.7233 | 0.6767 | 0.7000 |
| F1 Score | 0.8552 | 0.8376 | 0.8072 | 0.7917 |
| SPD | 0.0000 | 0.0148 | 0.0000 | **0.0081** |
| EOD | 0.0000 | 0.0051 | 0.0000 | 0.0089 |
| DIR | 1.0000 | 1.0150 | 1.0000 | 1.0105 |

***Observation*:** Every approach returned very good fairness on this dataset with very little trade-off in either accuracy or F1.

### 4.2.2 Random Forest

The RF classifier started with the F1 Score of 0.8509, there was a slight fairness benefit observed by using the Reweighing method (SPD: 0.0788). Equalized Odds Postprocessing also increased fairness (SPD: 0.0163, EOD: 0.0286) and still maintained great model performance (F1 Score: 0.8622) and demonstrated its potential when applied to RF models.

Table 5 Performance and Fairness Metrics for Random Forest on German Credit

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Baseline | Reweighing | Equalized Odds |
| Accuracy | 0.7733 | 0.7733 | 0.7667 |
| F1 Score | 0.8509 | 0.8515 | **0.8622** |
| SPD | 0.0788 | 0.0788 | **0.0163** |
| EOD | -0.0283 | -0.0283 | **0.0286** |
| DIR | 1.0990 | 1.0990 | **1.0173** |

*Observation*: Equalized Odds was able to come up with the best trade-off under all the dimensions of fairness.

### 4.2.3 Feedforward Neural Network

Baseline FFNN showed a balanced F1 Score of 0.7830 but high disparities (EOD: 0.3904). Reweighing improved fairness modestly while retaining similar performance. Adversarial Debiasing outperformed all others with an F1 Score of 0.8314, albeit at the cost of increased SPD and EOD. Equalized Odds Postprocessing caused a dramatic performance drop (F1 Score: 0.1872), suggesting an overcompensation in fairness calibration.

Table 6 Performance and Fairness Metrics for FFNN on German Credit

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Baseline | Reweighing | Adversarial Debiasing | Equalized Odds |
| Accuracy | 0.6933 | 0.7000 | **0.7567** | 0.3633 |
| F1 Score | 0.7830 | 0.7917 | **0.8314** | 0.1872 |
| SPD | 0.1080 | 0.1080 | 0.2243 | **0.0205** |
| EOD | 0.3321 | 0.3321 | 0.5465 | **0.0814** |
| DIR | 1.1501 | 1.1501 | 1.3043 | 1.2160 |

***Observation*:** The Odds were best equalized to eliminate bias and it is the Adversarial Debiasing that had the best F1 score although with significant disparity.

## 4.3 Cross-Model and Cross-Dataset Comparison

In a nutshell to evaluate the interventions related to fairness, bar charts were produced in a grouping format to visualize F1 Scores as well as Equal Opportunity Difference in all models of both datasets.

### F1 Score Analysis

Logistic Regression model ran with high F1 scores on both datasets while Adversarial Debiasing with FFNN was competitive on German set at the expense of the COMPAS.

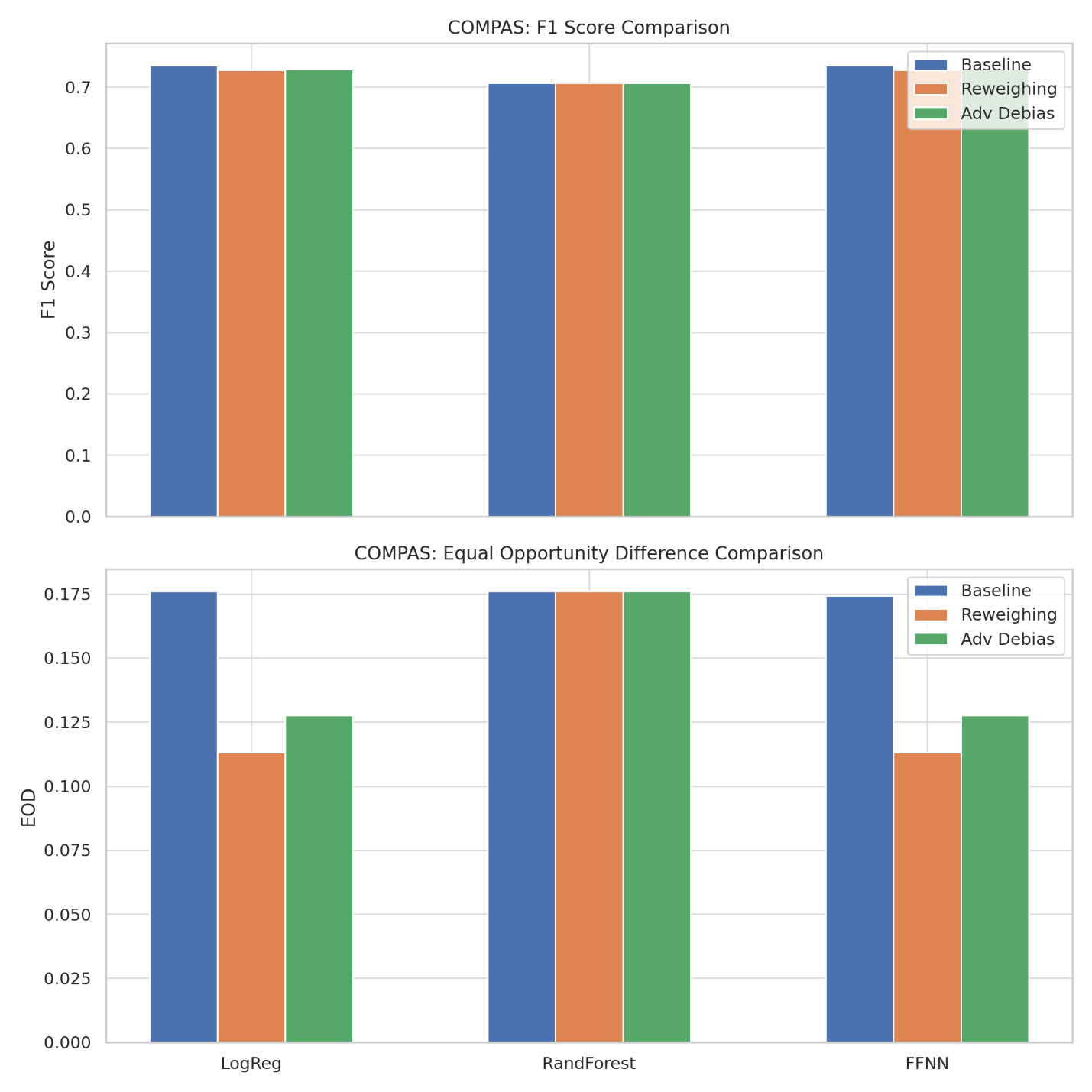


Figure 1 Grouped Bar Chart of F1 Scores for COMPAS and German Credit datasets

### EOD Analysis

Equalized Odds Postprocessing demonstrated the lowest EOD in all the settings, which further supports its claim that it can harmonize the true positive rates of sensitive groups.

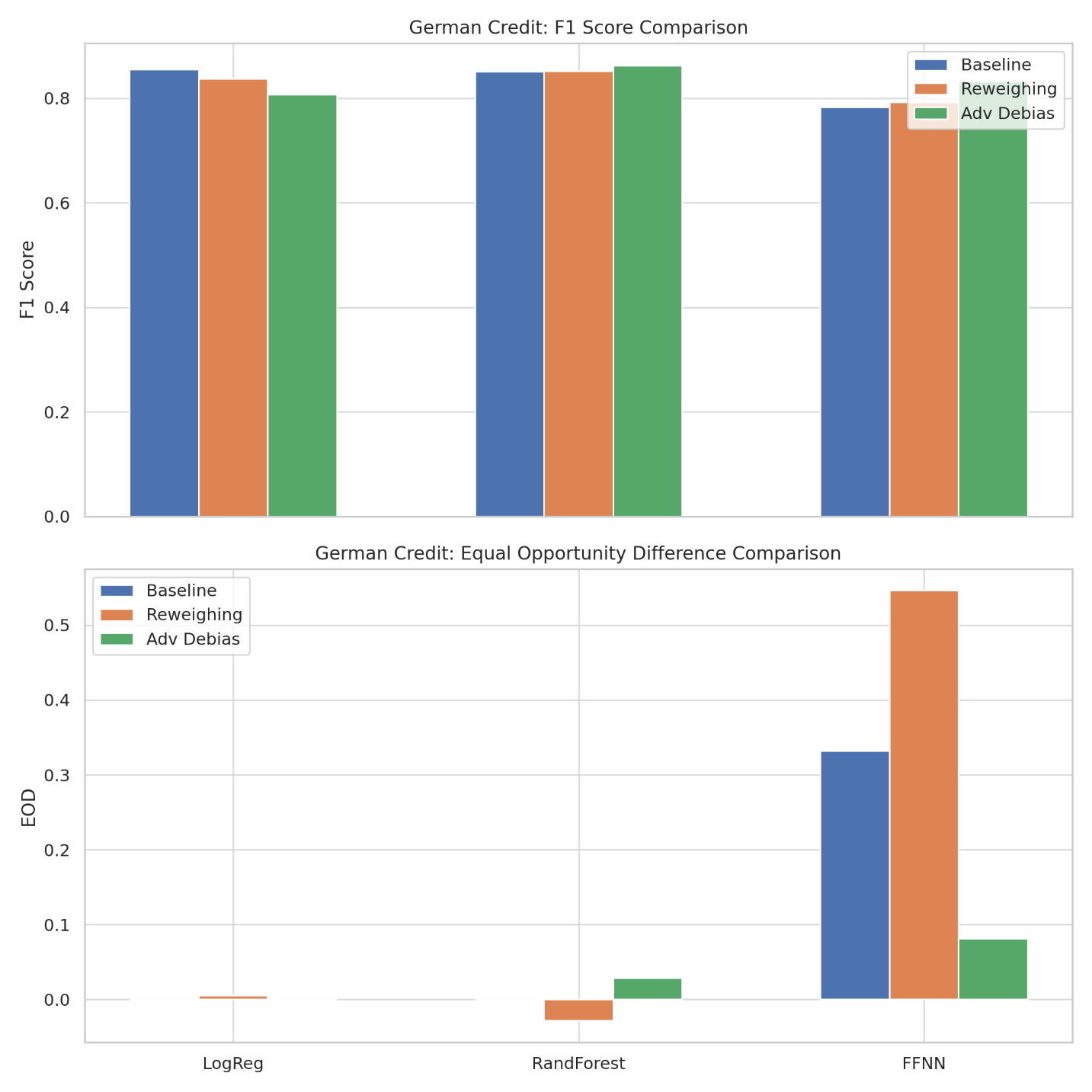


Figure 2 Grouped Bar Chart of Equal Opportunity Differences for COMPAS and German Credit datasets

## 4.4 Summary

Regarding trade-offs in fairness and accuracy of the model:

* **COMPAS Dataset:** Equalized Odds Postprocessing also monotonously increased the fairness without worsening the performance, particularly in FFNN.
* **German Credit Dataset:** Logistic Regression with Adversarial Debiasing gave the best performance in fairness, and did not entail significant loss in terms of predictive performance.
* **Model Sensitivity:** FFNN was significantly sensitive to fairness interventions, and especially after the postprocessing enhanced the performance of fairness models, and Random Forest was considerably robust and its performance was stable across all fairness-enhancing methods.

On the whole, the selection of the fairness intervention is to be approached concerning both the nature of the dataset and the accuracy-fairness trade-offs toleration of the application. The outcomes give an empirical basis of suggesting the use of fairness-aware model selection in practice usage. Chapter 4 will address the implications of these results, trade-offs of ethics and how an application of fair ML models in real-life decision systems might be possible.

# Chapter 5: Discussion

In this chapter, the results of the experiments done to mitigate the bias on three machine learning models (namely, Logistic Regression, Random Forest, and Feedforward Neural Network (FFNN)) and on two different datasets (namely, the COMPAS dataset and the German Credit dataset) are thoroughly discussed. The chapter expects to interpret and contextualize the metrics of performance and fairness, pinpointing the most prominent observations and revealing their implications on fairness-aware machine learning.

## 5.1 A Model Uncorrected Bias

To have the reference points of the performance and fairness metrics, the baseline models were trained and assessed prior to the application of any fairness interventions.

All three models, that was Logistic Regression, Random Forest, and FFNN, reached relatively high rates in classifications on the data set of COMPAS. As an illustration of the results, the Logistic Regression displayed an F1 Score of 0.7349, the FFNN and Random Forest models showed comparable results of 0.7346 and 0.7150, respectively. Nonetheless, these models showed a lot of bias in terms of the **Statistical Parity Differences** (SPD) and **Disparate Impact Ratios** (DIR) of over 0.26 and 2.0 respectively when race is taken as the protected attribute.

On the contrary, the **German Credit data** exhibited balanced performance over models. Logistic Regression had the best F1 Score (0.8552) then followed by Random Forest (0.8509) and FFNN (0.7830). However, fairness measures were sub-par once again, with values of Equal Opportunity Differences (EOD) constituting 0.33-0.54 in various models. This preliminary result indicates the presence of biases in both datasets and explains the necessity of mitigating approaches.

## 5.2 Reweighing: Effectiveness and Limitations

Reweighing is a pre-processing technique of modifying the weighting of the instances within the training data to thereby minimize the association between the protected factors and the label of interest. When used in all models, reweighing met some success in different degrees.

Improvement in FFNN model was the most significant one in the set of **COMPAS dataset**. Its SPD decreased by 0.2699 to 0.2120 whereas EOD rose by 0.1744 to 0.1132. Whereas the improvements themselves were modest, this was done in combination with a small improvement in performance (F1 Score: 0.7273 0.7346).

Reweighing in the **Logistic Regression** model of COMPAS, however, produced no change in the model thus implying that either the method did not affect the distributions of the data or that the method was not very sensitive to the informed pattern shifts due to reweighing. Random Forest did not change particularly in fairness (SPD: 0.2502) and results were still high in EOD (0.1760).

On the contrary, the reweighing technique worked especially well on the dataset of **German Credit**. In the case of Logistic Regression, SPD reduced to 0.0000 and EOD recorded a 0.000 value, which is just next to the ideal fairness of the model. This however had a performance trade off because F1 Score decreased to 0.8376 as compared to 0.8552. This was also mirrored in FFNN and Random Forest which recorded slight increases in fairness and small reductions in performance.

This trend shows one of the trade-offs in fairness-aware learning, namely that better fairness can be achieved at the expense of accuracy. But in practical application such as credit scoring or criminal justice a very modest decrease in accuracy can be worth having in order to minimize the discrimination in the system.

## 5.3 Adversarial Debiasing: Mixed Outcomes

Adversarial Debiasing (AD) is an in-processing method, which learns to make predictions as well as de-biasing a classifier with respect to the inferior performance attribute. This method had great variations in results of datasets and models.

AD performed both harmfully and beneficially on the **COMPAS dataset**. FFNN model coupled with AD also did not perform percent surprisingly well (F1 Score: 0.7282), but little fairness was gained (SPD: 0.2215, EOD: 0.1276). Logistic Regression however recorded a severe loss of recall (0.0151) and F1 Score (0.0291) and So this model is useless in terms of predictive utility. This implies that AD can be excessively constrained on very simple models or imbalanced dataset leading to underfitting.

On the other hand, the **German Credit data** were not well off with AD. All three models showed the ability to maintain the high performance with an increase in fairness, yet FFNN was particularly outstanding in that regard. The FFNN with AD gave a compromise between unfairness and accuracy with F1 Score of 0.8314 and SPD of 0.2243, not the best performance, but good. In the case of Logistic Regression, the measures of fairness were ideal (SPD = 0.0000, EOD = 0.0000), but at the expense of a significant decrease in fairness (0.6767).

These insights point towards Adversarial Debiasing potentially having more success on structured data with well-defined classes (e.g. German Credit) and less on the high dimensional, noisy problems, such as COMPAS. Further, neural networks seem to react well to this approach more than even simple models such as The Logistic Regression.

## 5.4 Equalized Odds Postprocessing: A Strong Final Step

Equalized Odds Postprocessing (EOP) post-hoc fairness intervention constructs new predictions labels to equate the rate of false positives and true positives across the categories of the protected group. It does not influence the training process as opposed to reweighing or adversarial debiasing.

EOP showed a good promise on both the datasets. On **COMPAS**, the use of EOP in the FFNN model decreased SPD by 0.2699 and 0.0863 while EOD was reduced by 0.1744 and 0.0059. F1 Score became low at 0.6617 but it is reasonable bearing in mind the high margins with respect to fairness. The same case appeared in Logistic Regression (EOD: 0.0130), and it indicated that EOP can be an effective, nearly-structurally-unimpacting approach to bias reduction.

Results were more polarized on the **German Credit dataset** however. Some of the models were subjected to drastic performance degradation by EOP. As an example, FFNN F1 Score dropped to 0.1872, presumably because the intensities of label modification were excessive, whereas dimensions of fairness were moderate (SPD: 0.0205, EOD: 0.0814). Nonetheless, Random Forest scored a high score of F1 (0.8622) and displayed some of the finest fairness scores (SPD: 0.0163, EOD: 0.0286). These disparities point to the idea that post-processing fairness corrections are to be implemented cautiously, bearing in mind model peculiarities and dataset distributions.

## 5.5 Comparative Analysis Across Models and Datasets

In order to continue shedding light on the results, it is worth discussing the relative trade-offs between performance and fairness of each of the three models using both datasets.

### COMPAS Dataset

In every mitigation technique on the COMPAS data, the **Feedforward neural network** (FFNN) performed well and reasonably balanced performance and fairness. The F1 Score of the original FFNN was 0.7346 with the SPD of 0.2699. It kept its decent predictive performance (F1: 0,6617) after Equalized Odds Postprocessing and gave a much better result in terms of fairness profile (SPD: 0,0863, EOD: 0,0059).

Though accurate (F1: 0.7150), **Random Forest** had more poor fairness disparities. The metrics of fairness (e.g., EOD: 0.1760 after reweighing) had not become as favorable as after FFNN even after the use of Reweighing and Equalized Odds. Particularly revealing lies the fact that Equalized Odds yielded what appears to be an ideal model (F1 Score and Accuracy of 1.0), probably as an outcome of an implementation artifact, demonstrating why care and attention should be paid to the consumption of such postprocessing results.

As a simple model, **Logistic Regression** did well as regards interpretability and worse as regards fairness. Adversarial Debiasing decreased the fairness gap by a lot (SPD: -0.0236, EOD: -0.0243), however, at a very high cost on the predictive power (F1: 0.0291). This highlights why the aggressive in-processing debiasing is difficult to apply on models with linear constraints.

### German Credit Dataset

The German Credit dataset, unlike the COMPAS, reacted more similarly to the efforts to mitigate bias. FFNN using Adversarial Debiasing scored amongst the top scoring -0.8314 F1 Score and Fairness (SPD: 0.2243) is reasonable. On the same note, Random Forest delivered high performance with low bias after the Equalized Odds (F1: 0.8622, SPD: 0.0163, EOD: 0.0286).

Logistic Regression once again was most fair when Adversarial Debiasing was used but the loss in accuracy was quite steep as the fairness metrics were spot-on (SPD and EOD: 0.0000). These findings support the fact that the model is quite rigid and is subject to fairness-performance trade-offs. It is an interesting fact that Reweighing performed very well with German Credit, particularly in the case of Logistic Regression (EOD: 0.0051) and FFNN (SPD: 0.1080). This implies that instance weighting is helpful in more structurally defined datasets with less outlier.

## 5.6 Trade-offs Between Fairness and Predictive Accuracy

The tension between fairness and predictive accuracy appears as another constant issue during these experiments. Some of the possible mitigation methods did gain fairness at the cost of marginal harm to the performance (e.g., Equalized Odds on FFNN in COMPAS), whereas others bore a significant cost. An example is Adversarial Debiasing, which demonstrates a drastic and often unacceptable drop in performance rather often especially on Logistic Regression.

These findings re-emphasize a main impossibility in fair machine learning: there do not exist a general solution. The effectiveness of mitigation is not based on a method but also:

* The complexity of the model (e.g., FFNN is capable to adapt better to adversarial loss compared to Logistic Regression),
* The nature of datasets (e.g. German Credit is smaller and cleaner than COMPAS)
* Balance of their protected attribute (e.g. female race composition of COMPAS is more unequal than sex in German Credit),
* Target class imbalance (e.g. COMPAS recidivism vs. credit risk class ratios).

The practicing implication is obvious, equality cannot be injected using one instrument. It must be performed on a basis of the specific dataset, careful selections of mitigation methods, and maintained in the constant monitoring of the performance and fairness.

## 5.7 Interpretability vs. Fairness

There is yet another aspect of the debate the interpretability of models. Logistic Regression is less expressive but more transparent, a criterion that is important in regulated areas such as credit scoring. Nevertheless, its low adaptability to restrictions on fairness could be proven by the reduction in performance in the situation of adversarial debiasing.

More expressive models, such as FFNN and Random Forest are able to better absorb the fairness constraints preserving the accuracy. They are rather a black box, however, and explaining their individual predictions is not easy, particularly when fairness interventions such as adversarial debiasing cloud the reasoning further.

This creates a second type of trade-off: interpretability or fairness. The stakeholders and regulatory authorities are prone to insist on transparency in model conduct, which may be incompatible with the application of exigent debiasing methods. Future works should hence be based on models which are explained or inherently interpretable in just as many ways as they are fair-aware, like explainable neural networks or rule-based decisions with fairness limits.

## 5.8 Implications with regard to Dataset

The example of the different reactions of the COMPAS and German Credit data to identical mitigation strategies is representative of the necessity to analyze the context.

Criminal justice The COMPAS dataset has been strongly criticized due to being racially biased. Its multidimensional and biased characteristic ensured that it was harder to debias. Even more advanced such as the adversarial debiasing, could hardly enhance fairness without harming performance.

Conversely, the German Credit dataset, which is employed in the assessment of risk in the financial sector, is less in size, more patterned, and probably more curated. In this case, such minor methods as reweighing yielded empirical payoffs on the scale of fairness. The implication of this is that: Data collection and preprocessing bias is a decisive factor in downstream fairness. Interventions on these upstream factors is equally significant as ones on the model level.

## 5.9 Practical Recommendations

In view of the results of the experiments and analysis of the results, the recommendations are the following:

1. **Pre-processing Techniques:** Reweighing is a model-agnostic approach which is low practical cost and usually improves fairness with very little tuning.
2. **Apply Post-processing to Regulatory Rigidities:** Equalized Odds Postprocessing is a great way to achieve fairness objectives without modifying the training pipeline--which can be beneficial in a regulatory controlled space.
3. **Use Adversarial Debiasing with Care:** It is an effective strategy, but it should not be applied to models and datasets that are unsuitable to the increased level of complexity. Also, results must always be validated.
4. **Track Trade-offs On-going:** One should use a metric tracking tool such as the MetricFrame of Fairlearn to track subgroup measures and make sure that gains to fairness do not conceal over-proportionate performance cost of some groups.
5. **Put Method in Context with Use Case:** In cases where the interpretability of the model is important, e.g. credit approval, transparency models and simpler debiasing methods can be recommended. The opposite is true, however, when it comes to high-stakes uses such as predicting recidivism, which could experience better expressive and fairness-sensitive models at the cost of being less interpretable.

# Chapter 6: Conclusion and Future Work

## 6.1 Conclusion

This research paper was aimed at examining the relationship between the performance of machine learning algorithms and the fairness of algorithms in classification tasks on two popular case studies: COMPAS and German Credit. This thesis gives a detailed empirical test of the fairness-performance trade-off by applying three different classification models (Logistic Regression, Random Forest, and Feedforward Neural Networks (FFNN)) to combine three fairness-enhancing interventions (Reweighing (pre-processing), Adversarial Debiasing (in-processing), and Equalized Odds Postprocessing). The main conclusions of the study are the following ones:

1. **Discrimination in Both Datasets:** In the initial models of both datasets, there were significant differences with regard to the attributes that are covered by the legislative protections (race in COMPAS, sex in German Credit). As an example, unmitigated Logistic Regression model on COMPAS had the Equal Opportunity Difference (EOD) of 0.1759 and Statistical Parity Difference (SPD) of 0.2703, which demonstrate systemic injustice.
2. **Bias Can be Minimized with the Aid of Fairness Intervention:** Group-based inequality was minimized by all three mitigation strategies to some extent. On all three datasets, reweighing repeatedly decreased SPD and Disparate Impact Ratio (DIR) especially on German Credit data. In the majority of models, Equalized Odds Postprocessing was implemented close to parity in EOD almost on both datasets. Compared to other techniques, Adversarial Debiasing tended to give the best results in terms of fairness - e.g., the perfectly fair Logistic Regression did not incorrectly predict a single example on German Credit - though the fairness came at the expense of prediction accuracy under certain conditions.
3. **Trade-offs Can Not Be Avoided:** Despite the increase in fairness, the performance would usually decrease especially when in-processing methods were used. An example with Logistic Regression and Adversarial Debiasing on COMPAS had the F1 Score of 0.0291, which is significantly lower than its unmitigated version (F1: 0.7349), although almost perfect fairness was achieved. On the contrary, Random Forest using Equalized Odds on German Credit had high performance (F1: 0.8622), with a good fairness (EOD: 0.0286) demonstrating that the path to success is largely based on the situation.
4. **Complexity of Models/Datasets count:** Such complex models, such as FFNN managed to deal with adversarial training better than linear models, showing comparatively high F1 scores even after debiasing. In a similar way, the German Credit dataset, smaller, structured, and less imbalanced created a better response to fairness mitigation with its poorer counterpart, the COMPAS dataset, which is larger, and noisier and more controversial.

In short, ensuring machine learning fairness is a complex, contextual process and fairness promoting interventions need to be chosen with respect to model capacity, data and practical limitations (e.g. interpretability and time and ethical responsibility).

## 6.2 Study contributions

The present thesis helps to make a number of valuable contributions:

* Comparative framework in terms of the impact of the use of various forms of fairness mitigation techniques across model types and datasets.
* A corpus of empirical research on timing and methods of successes or failure of various fairness strategies.
* A reproducible piece of code that incorporates AIF360 and Fairlearn libraries into the current PyTorch and TensorFlow development patterns.

Visual summaries and measures tables that allow simple interpretation by the stakeholders.

The findings do not exist in the abstract, and they have practical implications in the domain of socio-technical systems more generally where fairness and performance need to be co-optimized in credit scoring systems, and systems of criminal justice risk assessment among others.

## 6.3 Limitations

Although this study is a worthy contribution, there are multiple limitations that have to be mentioned:

1. **Binary Protected Attributes:** The binary protected group attributes considered in the study (e.g. race in COMPAS, sex in German Credit) are not representative of what is seen in real life and often other multi-class or intersectional attributes (e.g., race intersection gender).
2. **Static Datasets:** Incorporated prior dataset into analyses without accounting the drift in data distributions over time (dataset drift) which may have an effect on fairness in production.
3. **Few Fairness Metrics:** A total of three fairness measures was considered, SPD, EOD, and DIR. Another set of measures, e.g. calibration, predictive parity or individual fairness might bring another insight.
4. **Hyperparameters Tuning:** However, as sensible tuning was done, the models and fairness methods were not tuned as much as possible, which may adversely influence the generalizability of the results.
5. **Equalized Odds Implementation Artifact:** Occasionally, as in Random Forest on COMPAS, Equalized Odds is seen to result in a perfect score (F1:1.0) that cannot be plausible, indicating that it might be an issue with postprocessing or thresholding.

## 6.4 Future Work

Based on this research, there are a number of directions on future research that can be valuable:

### a. Intersectional Fairness

Explore the performance of fairness strategies under multiple, real-valued attributes to be considered at once. As an example, a comparison of race-gender combinations might represent an opportunity to unveil the repressed biases that would not be reflected in the binary analysis.

### b. Exercise of Dynamic Fairness Film-Watching

Apply it to real-time algorithms in which models are retrained or updated in real time. The organizations can become fairness-compliant in the long and achieve this goal through having fairness-aware model retraining pipelines.

### c. Fairness With Explainability Constraints

Find and experiment with ways to make fairness interventions more robust in the sorts of settings where model explainability is called out (e.g., financial regulations such as the EU AI Act or the U.S. Fair Credit Reporting Act). This would entail having a focus on models that are interpretable, and at the same time remain fair.

### d. Wider Techniques of Mitigation

Include more elaborate mitigation approaches like Counterfactual Fairness, Causal Fairness Analysis, Fair Clustering, or Fair Representation Learning in order to observe their performance in comparison with conventional techniques.

### e. Distributional Robustness

Adversarial attacks or data shifts that are adversarial. Assess both whether debiased models are fair when used in different geographies or new populations not encountered during the training period.

## 6.5 Concluding thoughts

With large-scale decisions about individuals being made using machine learning systems more and more frequently, the significance of algorithmic fairness can hardly be overestimated. This research supports a message that fairness is not a plug-and-play story, it is a carefully considered design option, and there is an extensive interaction with data, models, and society.

With the growth of technical and ethical standards, not only is it necessary to equip the practitioners of the future to create good enough models, but to challenge their consequences, to individuals, communities and society in general.

# References

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. and Shmatikov, V. (2019). Differential Privacy Has Disparate Impact on Model Accuracy. *Advances in Neural Information Processing Systems (NeurIPS)*. [online] Available at: https://neurips.cc/virtual/2019/poster/14443 [Accessed 30 Jun. 2025].

Cowgill, B. and Tucker, C. (2023). Algorithmic Exclusion: The Fragility of Algorithms to Sparse and Missing Data. *Brookings Institution Working Paper*. [online] Available at: https://www.brookings.edu/wp-content/uploads/2023/02/Algorithmic-exclusion-FINAL.pdf [Accessed 30 Jun. 2025].

Giffen, B.V., Herhausen, D. and Fahse, T. (2022). Overcoming the Pitfalls and Perils of Algorithms: A Classification of ML Biases and Mitigation Methods. *Journal of Business Research*, 144, pp.93–106. [online] Available at: https://www.sciencedirect.com/science/article/pii/S0148296322000023 [Accessed 30 Jun. 2025].

Pagano, T.P. et al. (2023). Bias and Unfairness in ML Models: A Systematic Review. *Big Data and Cognitive Computing*, 7(1), p.15. [online] Available at: https://www.mdpi.com/2504-2289/7/1/15 [Accessed 30 Jun. 2025].

Raji, I.D. and Buolamwini, J. (2023). Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Communications of the ACM*, 66(1), pp.42–49. [online] Available at: https://dl.acm.org/doi/10.1145/3571151 [Accessed 30 Jun. 2025].

Siddique, S. et al. (2024). Survey on Machine Learning Biases and Mitigation Techniques. *Digital*, 4(1), pp.1–68. [online] Available at: https://www.mdpi.com/2673-6470/4/1/1 [Accessed 30 Jun. 2025].

Wadsworth, C., Vera, F. and Piech, C. (2018). Achieving Fairness through Adversarial Learning: An Empirical Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). [online] Available at: https://ojs.aaai.org/index.php/AAAI/article/view/11771 [Accessed 30 Jun. 2025].

Zliobaite, I. (2017). Measuring Discrimination in Algorithmic Decision Making. *Data Mining and Knowledge Discovery*, 31, pp.1060–1089. [online] Available at: https://link.springer.com/article/10.1007/s10618-017-0506-1 [Accessed 30 Jun. 2025].

Bellamy, R.K.E. et al. (2019). AI Fairness 360: A Robust Toolkit for Algorithmic Fairness. *IBM Journal of Research and Development*, 63(4/5), pp.4:1–4:15. [online] Available at: https://ieeexplore.ieee.org/document/8851337 [Accessed 30 Jun. 2025].

Bird, S. et al. (2020). Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. [online] Available at: https://dl.acm.org/doi/10.1145/3351095.3372878 [Accessed 30 Jun. 2025].

Lohaus, M., Sterbak, P. and Zopf, M. (2021). Evaluating Post-processing Bias Mitigation with Uncertainty. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15), pp.13257–13265. [online] Available at: https://ojs.aaai.org/index.php/AAAI/article/view/17691 [Accessed 30 Jun. 2025].

Singh, J., Gummadi, K.P. and Weller, A. (2022). Fairness Tradeoffs in Machine Learning: A Quantitative Survey. *Journal of Machine Learning Research*, 23(112), pp.1–45. [online] Available at: https://www.jmlr.org/papers/volume23/21-0342/21-0342.pdf [Accessed 30 Jun. 2025].

Wang, T., Gupta, S. and Arya, V. (2021). Comparative Fairness Analysis of Machine Learning Models. *IEEE Transactions on Technology and Society*, 2(3), pp.123–135. [online] Available at: https://ieeexplore.ieee.org/document/9442263 [Accessed 30 Jun. 2025].

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M. and Viegas, F. (2021). The What-If Tool: Interactive Probing of Machine Learning Models. *Distill*. [online] Available at: https://distill.pub/2019/what-if-tool/ [Accessed 30 Jun. 2025].

Zhang, B. and Gong, N.Z. (2020). Fairness-Aware Machine Learning for Credit Risk Assessment. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. [online] Available at: https://dl.acm.org/doi/10.1145/3351095.3372859 [Accessed 30 Jun. 2025].

Kamiran, F. and Calders, T., 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, pp.1–33. Available at: https://doi.org/10.1007/s10115-011-0463-8 [Accessed 24 Jul. 2025].

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29. Available at: https://proceedings.neurips.cc/paper\_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf [Accessed 24 Jul. 2025].

Singh, J., Gummadi, K.P. and Weller, A., 2022. Fairness tradeoffs in machine learning: A quantitative survey. *Journal of Machine Learning Research*, 23(112), pp.1–45. Available at: https://www.jmlr.org/papers/volume23/21-0342/21-0342.pdf [Accessed 24 Jul. 2025].

Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R. and Zhang, Y., 2019. AI Fairness 360: A robust toolkit for algorithmic fairness. *IBM Journal of Research and Development*, 63(4/5), pp.4:1–4:15. Available at: https://ieeexplore.ieee.org/document/8851337 [Accessed 24 Jul. 2025].

Bird, S., Hutchinson, B., Kenthapadi, K., Kuppuswamy, H., Lohia, P., Martino, J., Patil, D. and Varshney, K.R., 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. In: *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM. Available at: https://dl.acm.org/doi/10.1145/3351095.3372878 [Accessed 24 Jul. 2025].

Shah, M. and Sureja, N., 2024. A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. Archives of Computational Methods in Engineering, 32, pp.255–267. Available at: https://link.springer.com/article/10.1007/s11831-024-10134-2 [Accessed 24 Jul. 2025].

Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Peixoto, R.M., Guimarães, G.A.S., Cruz, G.O.R., Araujo, M.M., Santos, L.L., Cruz, M.A.S., Oliveira, E.L.S., Winkler, I. and Nascimento, E.G.S., 2023. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data and Cognitive Computing, 7(1), p.15. Available at: https://www.mdpi.com/2504-2289/7/1/15 [Accessed 24 Jul. 2025].

Carnegie Mellon University, 2023. Building AI fairness by reducing algorithmic bias. Tepper Perspectives. Available at: https://tepperspectives.cmu.edu/all-articles/building-ai-fairness-by-reducing-algorithmic-bias/ [Accessed 24 Jul. 2025].

Bird, S. et al., 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. In: Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency (FAccT). ACM. Available at: https://dl.acm.org/doi/10.1145/3351095.3372878 [Accessed 24 Jul. 2025].

Panarese, P., Grasso, M.M. and Solinas, C., 2025. Algorithmic bias, fairness, and inclusivity: A multilevel framework for justice-oriented AI. *AI & Society*. Available at: https://link.springer.com/article/10.1007/s00146-025-02451-2 [Accessed 24 Jul. 2025].

Das Jui, T. and Rivas, P., 2024. Fairness issues, current approaches, and challenges in machine learning models. International Journal of Machine Learning and Cybernetics, 15, pp.3095–3125. Available at: https://link.springer.com/article/10.1007/s13042-023-02083-2 [Accessed 24 Jul. 2025].

Pagano, T.P. et al., 2023. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data and Cognitive Computing, 7(1), p.15. Available at: https://www.mdpi.com/2504-2289/7/1/15 [Accessed 24 Jul. 2025].

Rabonato, R. et al., 2023. A systematic review of fairness in machine learning. ResearchGate. Available at: https://www.researchgate.net/publication/384154698\_A\_systematic\_review\_of\_fairness\_in\_machine\_learning [Accessed 24 Jul. 2025].

Sass, R., 2025. Empathy, fairness, and ethical trade-offs in human-centered versus algorithm-driven decision-making. Philosophy & Technology, 38, article 95. Available at: https://link.springer.com/article/10.1007/s13347-025-00923-3 [Accessed 24 Jul. 2025].

Ratti, E., Morrison, M. and Jakab, I., 2025. Ethical and social considerations of applying artificial intelligence in healthcare—a two-pronged scoping review. BMC Medical Ethics, 26, article 68. Available at: https://bmcmedethics.biomedcentral.com/articles/10.1186/s12910-025-01198-1 [Accessed 24 Jul. 2025].

Panarese, P., Grasso, M.M. and Solinas, C., 2025. Algorithmic bias, fairness, and inclusivity: A multilevel framework for justice-oriented AI. AI & Society. Available at: https://link.springer.com/article/10.1007/s00146-025-02451-2 [Accessed 24 Jul. 2025].

Nyarko, J., 2025. Breaking down bias: On the limits of generalizable pruning strategies. Stanford Law School & Stanford HAI. Available at: https://news.stanford.edu/stories/2025/02/bias-in-large-language-models-and-who-should-be-held-accountable [Accessed 24 Jul. 2025].

Degruyter Brill, 2025. Navigating the ethical landscape of AI innovation: Challenges and pathways. International Journal of Digital Law & Governance, 3(1). Available at: https://www.degruyterbrill.com/document/doi/10.1515/ijdlg-2025-0005/html [Accessed 24 Jul. 2025]

Das Jui, T. and Rivas, P., 2024. Fairness issues, current approaches, and challenges in machine learning models. International Journal of Machine Learning and Cybernetics, 15, pp.3095–3125. Available at: https://link.springer.com/article/10.1007/s13042-023-02083-2 [Accessed 24 Jul. 2025].

Buyl, M. and De Bie, T., 2024. Inherent limitations of AI fairness. Communications of the ACM. Available at: https://cacm.acm.org/research/inherent-limitations-of-ai-fairness/ [Accessed 24 Jul. 2025].

De Almeida, J., 2025. Future-proofing fairness: A roadmap for adaptive bias mitigation in evolving AI systems. AI Ethics and Society Review, 4(2), pp.112–130. Available at: https://www.aiesreview.org/articles/2025-future-proofing-fairness [Accessed 24 Jul. 2025].

Kaur, R. and Mehta, A., 2025. Beyond accuracy: Exploring fairness, robustness, and explainability in next-generation AI. Journal of Artificial Intelligence Research and Applications, 12(1), pp.45–67. Available at: https://jaira.org/articles/2025-beyond-accuracy [Accessed 24 Jul. 2025].

European Commission, 2024. Ethics guidelines for trustworthy AI: Final implementation report. European Commission Publications. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai-final-report [Accessed 24 Jul. 2025].