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

**Name:** Chaudhary Usama Naeem

**Degree Program:** MSc in Computer Science

Contents

[1.1 Introduction 3](#_Toc202308512)

[Background and Motivation 3](#_Toc202308513)

[Problem Statement 4](#_Toc202308514)

[Research Aim and Objectives 4](#_Toc202308515)

[Research Questions 5](#_Toc202308516)

[Scope of the Study 5](#_Toc202308517)

[Structure of the Dissertation 6](#_Toc202308518)

[1.2 Literature Review 6](#_Toc202308519)

[1.2.1 The nature of Algorithmic bias in machine learning 6](#_Toc202308520)

[1.2.2 Measures and Defenses of Fairness 7](#_Toc202308521)

[1.2.3 Methods of Bias Mitigation: Summing them up 8](#_Toc202308522)

[1.2.4 Post-processing Techniques 9](#_Toc202308523)

[1.2.5 Toolkits for Fairness Auditing and Mitigation 10](#_Toc202308524)

[1.2.6 Previous Comparative Studies Review 11](#_Toc202308525)

[1.2.7 Study Limits and Rationale of the Research 11](#_Toc202308526)

[1.2.8 Summary 12](#_Toc202308527)

[Chapter 2: Methodology 12](#_Toc202308528)

[2.1 Research Design 12](#_Toc202308529)

[2.2 Dataset Selection 13](#_Toc202308530)

[A. COMPAS Recidivism Dataset 13](#_Toc202308531)

[B. German Credit Dataset 13](#_Toc202308532)

[Data Preprocessing Steps 14](#_Toc202308533)

[2.3 Model Selection 14](#_Toc202308534)

[2.4 Bias Mitigation Techniques 14](#_Toc202308535)

[A. Pre-processing – Reweighing (Fairlearn) 14](#_Toc202308536)

[B. In-processing – Adversarial Debiasing (AIF360) 14](#_Toc202308537)

[C.Post-processing Equalized Odds Post-Processing (AIF360) 15](#_Toc202308538)

[2.5 Evaluation Metrics 15](#_Toc202308539)

[2.6 Frameworks and Software tools 15](#_Toc202308540)

[2.7 Ethical Considerations 16](#_Toc202308541)

[2.8 Limitations of Methodology 16](#_Toc202308542)

[References 17](#_Toc202308543)

# 1.1 Introduction

## Background and Motivation

Machine learning (ML) is currently in the new decade that is taking over power in many fields such as finance, medics, employment, criminal justice, and academics to name but a few. The systems are usually touted as efficient, predictive accurate, and data-driven decision-making systems. Though, in parallel with these advantages, an urgent problem that has become the topic of discussion has appeared namely- algorithmic bias, where ML systems generate results that in a systematic manner disfavor certain groups of people, especially on the basis of their protected characteristics while race, gender, or age are just examples.

Among the most popular terms of algorithm bias is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism prediction used in the U.S judicial system. Research found it labeled Black defendants as being at high risk of recidivism when adjusted by whip match defendants of a similar case characteristic than white defendants (Feldman et al., 2015). In its recruitment sector, Amazon has a hiring tool, which was developed as an experiment and allegedly discriminated against women candidates in all the technical jobs because of the tendencies identified in past recruitment practices that categorized men. A similar observation has been done with credit scores that show lending decisions go against minority applicants based on a historical record of lending that is skewed with underlying discrimination.

These scenarios illustrate one of the deep problems in ML: a model suggested to a biased historical set of data will most likely reproduce or even magnify the same. All these biases do not originate only in data but also in feature selection, model architecture, and loss functions and evaluation metrics. Furthermore, biases come out apparent at times when biased models have been introduced into the real world environment and it is too late to effect such results which may be colossal and irreversible.

With the growing adoption of ML models in delicate decision-making tasks, there are calls of transparency, fairness, and accountability. As noted by regulators, including the European Union (GDPR) and the proposed AI Act, regulatory frameworks are starting to Default to formulate such concerns as the introduction of legal consequences to algorithmic decision-making. At the same time, academic and technical communities have begun to investigate, identify and address bias in ML systems, which has resulted in the development of the discipline in fairness-aware machine learning.

Although this attention is increasing, one vital gap still exists: most projects that have been proposed to mitigate bias are either at theoretical stages or are tested on small and isolated operations. Model comparison has been highly informal, and does not standardize an dataset, or a fairness measure or test. Moreover, there is a common neglect in the trade-off favoring increased fairness and predictive performance. Such a trade-off is critical in applied ML, where accuracy is a major factor in whether or not a model is deployed and at the expense of fairness.

The necessity to fill those gaps, analyzing the bias of machine learning models comprehensively, comparatively, reproducing the results and reaching conclusions is the driving force that inspired this dissertation. It aims at measuring the behavior of various models when subjected to biased data, the effectiveness of various bias-mitigation strategies, in correcting unfairness, and the impact of such interventions to the overall performance of the models.

## Problem Statement

As machine learning models start entering the lives of people more deeply in the form of decision-making, the power they possess on societies increases as well. Nonetheless, majority of available research deals either with diagnosing bias in particular models, or application of a single mitigation method and sometimes in an unsystematic and uncomparative manner. With such a siloed mentality, we do not have enough information about the wider implications of bias in all forms of ML models and use cases.

In addition, the evaluation of ML fairness is usually limited to isolated measures or narrow datasets, which limits the ability to generalize. The few pieces of the literature that exist fail to examine how fairness metrics engage with predictive performance tasks (e.g., accuracy, precision, recall) and the trade-offs that are involved in the effort of improving fairness. This complicates the job of practitioners in formulating informed decisions in choosing models, bias mitigation approaches, and sensitive applications.

This is why more holistic approach is needed, one that e.g. compares several models with several datasets with standard fairness and performance measurements, and uses variety of mitigation measures to generate useful comparisons.

## Research Aim and Objectives

This dissertation seeks to explore, compare, and reduce bias during machine learning model development, in a framework, which considers all the types of models, measures of fairness, and mitigation activities.

With the purpose of achieving this goal the following research objectives have been established:

* Evaluate how much bias we have in typical ML models when we train them on biased data that can be classified in terms of known demographic imbalances within this data.
* Apply bias mitigation strategies of three types, namely pre-processing, in-processing and post-processing, and compare them.
* Examine the trade-offs in fairness to predictive accuracy brought about by application of these mitigation strategies.
* Create modular program artefact that will serve to automate the bias testing and bias mitigation methodology deployment and display them in graphical form.
* Provide decision-ready and practice-recommendations to practitioners and researchers on how to enhance the fairness of ML without disproportionately degrading its utility.

## Research Questions

In order to organize the research process and to ensure that the purpose does not get lost, the research questions (RQs) outlined below have been formulated:

RQ1: How biased are machine learning models during the training phase when being fed with datasets with demographic biases?

RQ2: What is the effectiveness of the various types of bias mitigation strategies in enhancing fairness of the diverse ML models?

RQ3: What is the model fairness-model performance tradeoffs of using such bias mitigation strategies?

The questions will incorporate an investigation of the existence and the success of mitigation, in a variety of cases, as a more subtle and all-encompassing view of ML fairness.

## Scope of the Study

This study is based on classification research that employs socially sensitive data like race, gender and age. Well-known datasets, with published bias, like the COMPAS Recidivism dataset or the German Credit dataset are used. The research is restricted to three popular families of models, including logistic regression, random forest, and feed-forward neural networks that were selected due to their popularity and various levels of interpretability and complexity.

The study takes two mitigation strategies in each of the three main categories: •

**Pre processing:** Converting data prior to training the model.

**In-processing:** training-time alterations to an algorithm.

**Post-processing:** Tuning of prediction following train-on-the-model.

Fairness Fairness is scored with various dimensions such as statistical parity difference, equal opportunity difference, and disparate impact ratio as well as with conventional model performance scores.

## 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:** Describes the architecture and development of the software artefact and progress with sprints.

**Chapter 4- Results:** it introduces empirical data of the performed experiments and visualizations of fairness and accuracy results.

**Chapter 5- Discussion:** Formats the findings in the terms of research questions and literature by paying attention to the trade-offs, challenges, and implications.

**Chapter 6- Conclusion:** It presents the summary of the main findings, addresses limitations and provide suggestions on further work.

## 1.2 Literature Review

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

Machine learning (ML) bias Machine learning bias results when machine-learned models make systematic or systematic outputs biased against, or in favor of, accepting or rejecting specific groups, usually defined by such attributes as race, gender, age, or socioeconomic status. This is alarming even more considering that the most prevalent use of ML systems comes in life-saving applications such as healthcare triage, credit scores, hiring, and judicial sentencing (Pagano et al., 2023). Even when such systems are improper, they may further the talents gap instead of addressing them.

Not all bias is induced by an explicit discrimination. In many cases it manifests as a seemingly harmless pattern in the design, or a historical trend in the data. Giffen, Herhausen, and Fahse (2022) specify four common culprits so far in the introduction of algorithmic bias: (1) Historical bias, sequential to the inequalities (societal) that were prerecorded in the training set; (3) Representation bias, produced by the low representation of the selected populations; (3) Measurement bias, ensuing inappropriate or unequally applied labels and adjustments; and (4) Aggregation bias, when dissimilar subpopulations are totalized, disregarding their diver.

Cases of discrimination by implemented ML are not new. A recent review noted that facial recognition technology had a much greater misclassification rate among darker-skinned people as compared to the lighter-skinned counterparts thanks to biased training data (Raji et al., 2020). Even in the financial aspect, research results indicate that neurocybernetic algorithms working in the field of credit scoring tend to discriminate against minority borrowers, an example of historic marginalization (Cowgill & Tucker, 2020). It is on such instances that the necessity of making fairness operational and integrating bias detection mechanisms into ML pipelines resounds.

The bias may come in during any phase of the model development such as during data collection, feature selection, during the training of the model, and during its evaluation. Therefore, the pursuit of fair ML has become a cross-disciplinary endeavor comprising of statistics, ethics, computer science, and law. However, the current state of mitigation is quite dispersed, as according to Raji and Buolamwini (2022), there is no unified definition of fairness or much application of fairness-aware tooling in reality. This underscores the fact that there is a gap in consolidated, comparative framework, which this research attempts to fill.

### 1.2.2 Measures and Defenses of Fairness

It is hard to define fairness in machine learning on its own and there is no general definition of it. The selection of metric to use on fairness hinges on the domain of application, social aim of the application, and the distribution on which the data is being applied (Siddique et al., 2024). Three popular fairness measures are listed below and they will be used in the present research as a guide toward measurement:

* **Statistical Parity Difference (SPD):** The difference between the percentages of positive outcomes of unprivileged and privileged groups. The lower SPD means more fairness.
* **Equal Opportunity Difference (EOD):** Measures the difference between rates in true positives between groups. It guarantees that the chances of qualified people belonging to different groups being properly classified become equal.
* **Disparate Impact Ratio (DIR):** The difference between favorable results of the non-privileged and privileged can be calculated with the help of Disparate Impact Ratio (DIR). The lower the DIR, the more the fairness.

Such measurements are mathematically differentiated and are likely to contradict with each other. As an example, maximizing statistical parity can hurt overall model accuracy, or decreased individual fairness that is defined by different outcomes given to an individual of equal qualification. Siddique et al. (2024) stress that is beyond doubt that fairness cannot be a one-size-fits-all goal; on the contrary, it should be modified to the stakeholder expectations, legal contexts, and requirements of particular domains.

It has also been noted in the recent literature that there is an impossibility theorem of fairness: when base rates are different among groups, it is impossible to achieve several different fairness criteria, e.g., calibration, equal opportunity, or demographic parity concurrently (Zliobaite, 2021). This tension makes equity something that is bargained, rather than a formula, and the argument that models should be gauged on a number of measures.

The study will be a multi-metric study in order to achieve an overall perception of interrelation of fairness and performance varying multiple modeling and mitigating techniques.

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

Algorithms can mitigate algorithmic bias in three broad categories which include: pre-processing, in-processing, and post-processing. All categories fall in to various stages of ML lifecycle and each presents unique advantages and restrictions.

#### Pre-processing Methods

Pre-processing is when the training dataset is changed prior to model building to reduce bias. Such methods are:

* **Reweighting:** Set sample weights to match the representation of groups of people who are protected.
* **Disparate Impact Remover:** Altering features in such way that they will no longer correlate with the features that are protected or correlation level will be minimized.
* **Fair representation learning:** injuring features in a transformed dimension that is independent of group affiliation.

Importantly, pre-processing techniques tend to be model-free and simple to incorporate with the current pipelines.

Nevertheless, they are potentially introducing distortion in the data and their performance may differ according to the distribution hidden behind the features (Wadsworth et al., 2021). Moreover, overcorrection can decrease model utility since the utility of any model trained on a protected feature is diminished when it correlates with the target label in anything other than a neutral manner.

#### B. In-processing Techniques

Fairness constraints are hard-coded into learning algorithms in in-processing strategies.

Examples include:

* **Adversarial Debiasing:** Training a second network, which will identify and punish classification mistakes of bias obtained by the main model.
* **Prejudice Remover Regularization:** Providing fairness-penalties to the goal.

Such techniques provide a finer-grained control in fairness-accuracy trade-offs at the expense that they tend to be algorithmically specific and can demand substantial computing resources. Furthermore, they cannot be always interpreted as they are often not understandable in hyper-scale structures such as deep neural network (Bagdasaryan et al., 2020).

### 1.2.4 Post-processing Techniques

Post-processing methods work on models that have been trained on prior data and override its predictions to ensure fairness. Such techniques can be quite helpful when a model has to be retrained but it is not possible to or even undesirable. In contrast to the in-processing methods, post-processing does not alter either the data or the parameters describing the model (but acts upon the results of the prediction).

A popular post-processing method is Equalized Odds Post-processing, in which the threshold of decisions made to distinguish between various demographic groups is modified in order to balance the rate of true positives and the rate of false positives. The last approach is the Reject Option Classification, which alters the consequences of those on the boundary of the decisions, operating in favor of disadvantaged groups under uncertain predictions. Both approaches seek to compromise on predictive performance and fairness, usually by introducing trade-offs on the classification outcomes (Singh et al., 2022).

The post-processing algorithms are usually model-agnostic, i.e. they can be applied to arbitrary black-box classifier. This qualifies them as appealing in industry environments in which the internal model structure is proprietary or frozen. But sometimes post-processing results perhaps in inconsistency and diminished confidence in personal judgment especially when it seems that results are various in identical persons because of group affiliation (Lohaus et al., 2021).

Furthermore, post-processing can be very useful in cases where data sets are balanced and decision boundaries are well-established. Such methods can have difficulties in producing significant gains in fairness in high-dimensional or sparse datasets, without sacrificing accuracy significantly. Thus, post-processing, though helpful, should be understood as one of the instruments in an existing huge fairness plan.

### 1.2.5 Toolkits for Fairness Auditing and Mitigation

As a consequence of the increased demand of standardized, transparent, and reproducible fairness assessments, multiple toolkits have been created in an open-source format. Such libraries give access to implementations of fairness metrics, bias mitigation algorithms, and visualization tools that have become standard to the work of fairness-aware machine learning.

* **AI Fairness 360 (AIF360):** AIF360 was created by IBM Research staff and contains more than 70 fairness metrics and additional than 10 bias reduction algorithms in each of the three classes (pre-, in-, and post-processing). It enables the work on Python and Jupyter environments and has full tutorials on practical use (Bellamy et al., 2020).
* **Fairlearn:** An open-source library created by the Microsoft, Fairlearn is targeted at algorithmic fairness to the reduction-based in-processing. It allows visualizing the fairness-accuracy trade-off through an interactive dashboard and allows practitioners to interface with decision boundaries in a controlled way (Bird et al., 2022).
* **What-If Tool:** A What-If Tool is a TensorBoard plugin developed by Google that lets developers visually inspect model predictions across a wide range of groups without writing code or performing any instruction. It facilitates counterfactual inferences and subgroups investigation, which makes it particularly valuable to non-technical stakeholders (Wexler et al., 2021).

Every toolkit is strong and weak. AIF360 has the richest collection of tools however, it has the highest learning curve. Fairlearn is easy to use in-processing wise but cannot offer a lot of pre-processing procedures. What-If Tool is very user friendly yet it is not easily scalable or extendable. Collectively they appear to point to maturity and diversity of fairness-aware development tools presently available.

In this dissertation, AIF360 and Fairlearn are complementary and used together in order to have extensive coverage of bias mitigation strategies and fairness metrics. Their combination enables a modular and extensible architecture that is flexible to be used in the future experimentation.

### 1.2.6 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).

### 1.2.7 Study Limits and Rationale of the Research

This literature review presents a number of gaps in the body of literature:

* **Insufficient holistic comparison:** The majority of studies compare fairness “in a vacuum” (they test a single type of mitigation methods or a single ML model). Few studies have compared fairness performance associated with a variety of model architectures (e.g., linear models, tree-based models, and neural models).
* **Trade-offs that are underexplored:** There are also few studies estimating the loss in performance that occurs when bias mitigation methods are used. In the absence of such comparisons, the practitioners face the choice between performance and fairness as a matter of choice without empirical information.
* **Low reproducibility:** There exist toolkits to do all of this procedure, but rarely are they used in a one-stop pipeline which aids reproducibility, extensibility and feasible deployment.

The current dissertation will deal with these gaps by designing and implementing a comparative framework of bias and mitigation strategies evaluation across models and datasets. It uses common measures of fairness, publicly accessible data, and toolkits with open source code to be able to reproduce that work. In this way, the study will provide theoretical knowledge on the one hand and practical recommendations on how to construct fair and accurate ML systems on the other hand.

### 1.2.8 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

This paper follows the methodology of Design Science Research (DSR), which is appropriate to the development of realistic solutions to complicated real-life situations by creating artefacts systematically and assessing them. DSR is based on the design, development and validation of technological solutions to identified problems, i.e. of designing, developing and testing of a modular software artefact that can be used to detect and mitigate bias in ML models.

DSR process includes six stages: (1) identification of the problem, (2) definition of the objective, (3) design and development, (4) demonstration, (5) evaluation, and the (6) communication (Brocke et al., 2020). This study will follow such steps as identifying the areas of fairness gaps in the ML system, developing a framework to compare the approach to mitigation between model types and analyzing the outcomes to conclude on meaningful action. The artefact itself is an experimentation and visualisation ground of metrics and trade-offs of fairness.

Such a method favours a combination of constructive and empirical dimensions: it can build an artefact and measure its performance empirically against test data and quantitatively. The prototypic structure of DSR can also be adjusted depending on the result, which makes it close to the agile development process that was involved in the current research.

## 2.2 Dataset Selection

In a bid to guarantee a comparative and strong analysis of the biases, two publicly available datasets (based on extensive usage in fairness research and the existence of known demographical skews) were chosen:

### A. COMPAS Recidivism Dataset

* **Context:** Applied in the courts of America to find the propensity of criminals to commit crimes again.
* **Bias Concern:** Being charged with racial bias, African American defendants have been frequently rated as higher risk than white ones based on similar profiles (Raji et al., 2020).
* **Protected attribute:** Race (Black or White).
* **Target Variable:** 2 year recidivism (class: reoffended or no).

### B. German Credit Dataset

* **Context:** Information that includes people who have submitted personal loans to a bank.
* **Bias Concern:** It shows bias according to age and gender.
* **• Protected Attributes:** Age (<25 & Gender (Male and Female).
* **• Dependent:** Credit worthiness (0 or 1: either good or bad credit).

The latter datasets are properly documented within the literature, and hence, they offer perfect benchmarking of fairness interventions. Also, the datasets can be used in binomial classification, which is easy to compare performance on many models and methods.

### Data Preprocessing Steps

* Processing Missing Values As medians/mode based imputation performed depending on feature type.
* **Normalization:** Insets on numerical features to be used to make models consistent with one another.
* **Encoding:** Categorical features have been one-hot encoded: protected attributes were explicitly used to perform the group-based fairness evaluation.

The entire data was divided into training (70 per cent) and testing (30 per cent) sets to make evaluation reliable and ditto on the target variable such that there is a balance of classes.

## 2.3 Model Selection

Three ML categorizers were chosen according to the popularity, interpretability, and different complexity:

* **Logistic Regression (LR):** It is a simple and clear model that is used as a baseline.
* **Random Forest (RF):** A decision tree model which is robust and performs well and is an ensemble.
* Feed-forward Neural Network (NN) A deep learning model that shows more complicated and high capacity algorithms.

Training of these models was done with default parameters after which the hyperparameter tuning was done using cross-validation which optimized the baseline performance before fairness interventions were implemented. The variety in the models used facilitates the cross sectional study of the mitigation strategies behavior in various model families.

## 2.4 Bias Mitigation Techniques

In order to compare the fairness strategies in the ML pipeline, one of the representatives of each of the following categories of mitigation methods was applied:

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

Reweighing alters the weights of instances in the training set to provide equitative representation of groups. It is very efficient in rectifying past imbalances and it is model-free (Siddique et al., 2024).

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

The approach emphasizes the use of a second adversarial network that is designed to identify membership of the protected group. A penalty of the main classifier is done in case the adversary is successful, hence encouraging fairness in hidden representations (Bagdasaryan et al., 2020).

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

This method adjusts the labels that are expected to balance the rates of truly positive and false positive among groups. It is applying when retraining is impossible or when there is no possibility to access only model predictions (Singh et al., 2022). Either technique was used on each model individually to determine its unconfounded influence on both fairness and predictive performance.

## 2.5 Evaluation Metrics

As a comprehensive measure to conclude on the behavior of models, the following metrics were employed:

1. Performance Indicators

* Accuracy: Percentage of predictions which are correct.
* Precision: Ratio of correct prediction to positive prediction.
* Recall (TPR): The percentage of true positive correctly identified.
* F1-score: It is the harmonic mean between precision and recall.

1. Metrics of the Fairness

* Statistical Parity Difference (SPD): It must be near 0.
* Equal Opportunity Difference (EOD): It should be near to 0.
* Disparate Impact Ratio (DIR): It should be near to 1.

Besides enhancing fairness scores, the objective was also to determine the trade-off between fairness and utility, especially whereby the mitigation techniques were implemented.

## 2.6 Frameworks and Software tools

The artefact was created with python 3.11, and the development and experimentation were made in Jupyter notebooks and VSCode. Important libraries are:

* scikit-learn: Modeling training and testing.
* Fairlearn: Pre-processing techniques and reporting of metrics.
* AIF360: Applying assessment of fairness and pre/post-processing agents.
* matplotlib / seaborn: Display of results.

This system is modular, which allows an expansion to new datasets, models and mitigation methods. Outputs of every experiment are recorded in CSV and presented using comparative bar plots and heat maps.

## 2.7 Ethical Considerations

The research involved publicly available (anonymized) data, and thus it did not require an ethical approval. Nevertheless, the procedures involved in all the experiments followed the ethical concept of AI, such as transparency, equality, and non-discrimination. Caution was considered in making sure that there was no unintentional damage, especially in the interpretation of fairness measurements, which can be used in the wrong context..

## 2.8 Limitations of Methodology

In its totality, the methodology contains limitations:

* Limitation of datasets: Benchmark datasets can be far removed to the real world complexities.
* Metric competition: Fairness metrics often compete with each other- optimizing one will worsen another.
* The selection of classifiers and mitigation techniques covers a narrow scope of three different strategies and three different classifiers; it does not reflect the vast landscape of the ML models as well as fairness algorithms.

In spite of this limitation, the adopted method will give meaningful insight into the comparative behavior of mitigation strategies and facilitates both reproducibility and transparency.

# 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].