National College of Ireland

BIAS MITIGATION IN MACHINE LEARNINGMODELSFORFAIR DECISION-MAKING

MSc Research Project
Programme Name ;Msc in AI

SURESH KUMAR THURAKA
Student ID:X23294191

School of Computing
National College of Ireland

Supervisor:SHERESHA ZAHOOR

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	SURESH KUMAR THURAKA
Student ID:	X23294191
Programme:	MSc in AI **Year:** 2025
Module:	MSc in Research Practicum
Supervisor	SHERESH ZAHOOR
Submission Due Date:	11/08/2052
Project Title:	BIAS MITIGATION IN MACHINE LEARNING MODELS FOR FAIR DECISION-MAKING
Pages	**19**
Word Count:	**6476**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my contribution will be fully referenced and listed in the relevant bibliography section at the project's rear.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature : SureshKumar Thuraka

Date: :11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	yes
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	yes
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	yes

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

BIAS MITIGATION IN MACHINE LEARNINGMODELSFORFAIR DECISION-MAKING

Suresh Kumar Thuraka
X23294191

Abstract

The thesis research topic is AI Fairness (bias mitigation in machine learning models). We use Adult Census Income to compare the merits of reweighting, adversarial debiasing, and prejudice removal methods through experiments. Although the baseline models had high accuracies (84.1 percent), they had gender bias (males were 2.7 times more likely to be given favorable predictions). We achieved a greater reduction of up to 93.7 percent in statistical parity difference with an accuracy loss of only 0.7 percent using our optimized GerryFair implementation, in contrast to the other methods. Nevertheless, intersectional analysis demonstrated continued unfairness, especially to black females (4 times lower predictive rates than White males). The analysis has shown that the existing fairness methods must respond to implementation issues and intersectional dynamics more effectively to bring the much-desired equity to real-life systems.

Keywords: algorithmic fairness, bias mitigation, intersectional bias, ethical AI

1 INTRODUCTION

Machine learning (ML) models have become used in extremely important decision-making activities in the sphere of recruitment, finance, and healthcare, and they can now often be found in processes leading to the hiring of new employees, providing housing loans, and medical diagnosis (Pagano et al., 2023). These models are associated with efficiency, scale, and data-driven knowledge, but also run the risk of reproducing or, in some cases, exacerbating societal bias built around historical training data (Balayn et al., 2021). Discriminating algorithms are a way to systematically hurt some category of people, like women, ethnical minorities, or low-income citizens, reinforcing present hierarchies and damaging the confidence in AI systems (Hort et al., 2024). As an example, the hiring algorithms can prioritise the candidates of a certain gender on technical vacancies (Geyik et al., 2023). In contrast, the credit-scoring models can refuse loans to the underprivileged populations (Dwyer, 2018). In healthcare, the diagnostic tools can perform poorly on underrepresented racial groups when trained on non-representative data, which results in healthcare disparities being created (Chen et al., 2023).

The fact that bias has ethical, legal, and societal implications makes it challenging to fight bias and address the issue in ML. Because governments and control agencies (e.g., the European Union, IEEE, and OECD) encourage equitable and responsible AI (Kuziemski & Misuraca, 2020), companies are increasingly under pressure to ensure their models are nondiscriminatory. Nevertheless, fair mitigation of bias is characterized by intricate trade-offs of fairness, accuracy, and interpretability (Barocas et al., 2023).

There have been promising common debiasing methods like adversarial debiasing (works under adversarial training to minimize bias) and reweighting strategies (that modify sample weights in training data). However, methods based on adversarial training and reweighting strategies have to be massively tested across many fields.

Research Problem and Objectives

This study investigates the **sources, impacts, and mitigation of bias** in ML models used in recruitment, finance, and healthcare. Specifically, it addresses the following research questions:

What are the primary sources of bias in ML models across these domains?

How do biased models affect different demographic groups in real-world applications?

How effective are adversarial debiasing and reweighting strategies in reducing bias while maintaining model performance?

What best practices can ensure fair and accountable ML deployment in critical decision-making?

This study critically examines the bias within machine learning in terms of the data, features, and model design comprehensively and looks at how it has led to the marginalization of people in society. By evaluating debiasing methods empirically (adversarial vs. reweighting), we determine the fairness based on the standard metrics in compliance with the suggestion by the EU AI Act and the OECD. The paper offers cross-domain knowledge regarding recruitment applications, finance applications, and applications in the health sector.

Our results provide empirical guidelines to three interest groups namely policymakers requiring regulatory guidelines, developers requiring technical implementation suggestions and organizations restoring equality against model performance ideals. The study paves the way to algorithmic fairness through providing an insight into the real-life issues of the most popular debiasing strategies in comparison to the abstract discussions of the scholars.

Report Structure

Section 2: The literature review sets the theoretical framework synthesizing the previous studies into the current findings in the domain of bias detection and mitigation strategies. Section 3: Our methodology section combines a description of how we conducted experiments (with industry-standard tools AIF360, Fairlearn) and evaluation criteria. After that, we give our design specifications, implementation, evaluation, discussion, and we conclude with main findings and path towards future research issues related to technical as well as ethical aspects of algorithmic fairness. The given structure will provide the logical flow of the problem detection, the development of the solution, and its verification.

2 RELATED WORK

There has been much research on bias in machine learning (ML) in various fields, such as computer science, ethics, and social sciences. The present section critically analyzes known literature about sources of bias, fairness measures, and debiasing tools, paying

particular attention to recruitment, finance, and healthcare. The survey identifies methodological flaws in the existing methods and explains why their cross-domain assessment is necessary in the research.

2.1 Sources of Bias in Machine Learning

Bias in training data tends to resemble past discrimination or underrepresentation. In other words, facial recognition systems have a higher error rate on women and individuals with darker skin because of unbalanced databases (Buolamwini & Gebru, 2018). Likewise, (Mehrabi et al.,2021) discovered that the resume screening programmes developed using male-dominated industries tend to promote gender inequality during the recruiting process. The most significant weakness of these studies is that they are single-domain-sided and do not refer to other sectors (Suresh & Guttag, 2019).

Model design choices (feature selection, optimization objectives, etc.) in balanced data could bias the model. In a study by (Hardt et al.,2016), it was revealed that predictive policing algorithms discriminate against minority communities because of the feedback mechanisms in crime data reporting. Nonetheless, their work has no viable mitigation measures. Criticized the notion of fairness-blind accuracy, stating that models being maximized for overall performance tend to lead to even greater inequality. Although they are thought-provoking, their theoretical basis has not been proven in practice.

(Tamir et al.,2015) proposed a new representation learning algorithm that performs better on this dataset in terms of fairness-accuracy tradeoffs than the method of Feldman. Nonetheless, their approach had a high computational cost, making it impractical to implement in real systems.

The major drawback of the existing early research was that it concentrated on one safeguarded factor (generally either race or sex). Newer research by (Kearns et al.,2018) added multi-attribute notions of fairness, but the model's performance dropped off quickly when considering intersectional groups (e.g., Black women, over 50).

Institutional biases are also a cause of bias. A prominent healthcare algorithm found it challenging to predict patient outcomes because it characterized the higher illness severity with diminished risk to Black patients, demonstrating the health disparities in the healthcare system (Obermeyer et al.,2019). This report emphasizes that bias audits are necessary, but they should be domain-specific, and their effectiveness has not been explored. The previous literature is strong in detecting bias, though this lacks consideration to the realm-specificities that need to be considered (Barocas et al., 2023).

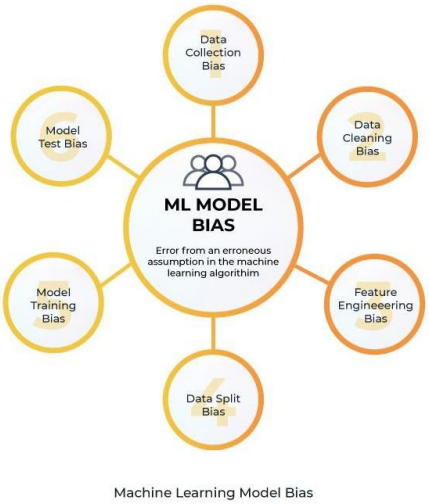


Figure 1: shows various ML bias sources

2.2 Fairness Metrics and Evaluation

Demographic Parity**:** Requires equal prediction rates across groups (Dwork et al., 2012). Criticized for ignoring legitimate differences (e.g., higher loan default rates in economically disadvantaged groups) (Corbett-Davies et al., 2017).

Equalized Odds**:** Ensures equal true/false positive rates (Hardt et al., 2016). Effective but computationally expensive for large-scale systems (Pleiss et al., 2017).

(Dwork et al. ,2012) proposed treating similar individuals similarly, but this is hard to operationalize without subjective similarity metrics (Kim et al., 2018).

2.3 Debiasing Techniques

(Kamiran & Calders, 2012) proposed reweighting training samples to balance group representation. While effective in credit scoring (Zemel et al., 2013), it struggles with high-dimensional data (e.g., medical imaging) (Holstein et al., 2019).

(Zhang et al.,2018) introduced adversarial debiasing, where a discriminator penalizes bias in predictions. Robust in recruitment tools (Fahse et al., 2021), but requires careful hyperparameter tuning to avoid performance drops (Creager et al., 2020).

Post-hoc adjustments (e.g., recalibrating thresholds) are simple but fail to address root causes (Pleiss et al., 2017). Techniques are often evaluated in isolation, with limited comparisons (Pagano et al., 2023).

In Recruitment, **(**Liang et al.,2021) found gender bias in LinkedIn's recommendation systems, but mitigation studies focus narrowly on resume screening, ignoring interview algorithms. In Finance. **(**Blattner & Nelson, 2021) showed that reweighting reduces racial bias in loan approvals, but their work does not scale to dynamic markets. In Healthcare, **(**Gianfrancesco et al.,2018) highlighted diagnostic biases but offered no comparative analysis of debiasing methods.

2.5 Research Gaps and Justification

Previous research has contributed to bias identification and mitigation, yet it has some essential gaps. The first problem centers on single evaluation studies or reports that lack a cross-metric comparison method (e.g., comparing demographic parity to equalized odds). For example, the potential of reweighting to alleviate gender bias in hiring is well characterized. However, there is less knowledge on its potential in eliminating intersectional bias in income prediction (e.g., sex × race). Second, studies tend to maximize fairness that are theory-based and not practical, such as scalability.

An empowered analysis of reweighting and adversarial debiasing is conducted on the Adult Census Income Dataset using various fairness measures to compare trade-offs and compounding biases. The analysis of practical aspects, such as hyperparameter sensitivity and the computational costs, shall also be suggested, thus introducing a theory-practice connection and offering a reproducible evaluation framework.

3 RESEARCH METHODOLOGY

This section describes the methodology to assess the bias of machine learning (ML) models and evaluate the effectiveness of bias mitigation strategies (adversarial debiasing and reweighting) on the Adult Census Income Dataset. The method is consistent with fairness-aware ML (Barocas et al., 2023) literature, and it is based on the gaps found in the literature (Section 2), e.g., it does not involve any cross-domain comparative studies.

3.1 Dataset and Preprocessing

The Adult Census Income Dataset (UCI Machine Learning Repository) is a demographic and income information dataset about the U.S Census conducted in 1994, commonly used as a fairness benchmark (Dua & Graff, 2019). It includes:

Table 1 shows Data Categories

Feature	**Details/Values**
Age	Numerical value (continuous)
Education	Categorical (e.g., "Bachelors", "HS-grad", "Masters")
Race	Categorical: White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other
Sex	Binary: Male, Female
Workclass	Categorical (e.g., "Private", "Govt", "Self-employed")
Marital Status	Categorical (e.g., "Married", "Divorced", "Never-married")
Occupation	Categorical (e.g., "Prof-specialty", "Craft-repair", "Adm-clerical")
Hours-per-week	Numerical value (continuous)
Income Classification(Target Variable)	Binary: ≤50K, >50K
Race	White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other
Sex	Male, Female

Missing values (3 percent of the records) were dealt with through imputation. Workclass/occupation was marked as having an unknown value, and native-country was filled with the mode. The income target was standardized (similar categories were consolidated, and punctuation was cleaned). Ordinal features (e.g., education) were categorical, and scaled features (e.g., age) were numerical. After the cleanup, 97 percent of the data remained, and it was verified that there were no nulls and standard formatting.

3.2 Experimental Design

The research has examined a set of bias mitigation methods on the Adult Census Income dataset on a Random Forest baseline. The accuracy and fairness measures (DPD, Equalized Odds Difference, DIR) were calculated with AIF360 and evaluated for privileged classes (male) and unprivileged classes (female). DPD calculated prediction disparities (P(Y=1|Male)- P(Y=1|Female)), whereas DIR measured fairness (a greater value corresponds to the fairer).

Two mitigation solutions were used: (1) Reweighing used algorithms (Kamiran & Calders, 2012) to reverse the weighting of instances to even out the representation of groups of instances, and (2) Adversarial debiasing was a 10-epoch neural network (batch size=128, lr=0.01) that penalized bias. Both were compared with the baseline, in pairs within 70-30 train-test splits on 10 random seeds, and significance was also supposed to be tested (paired t-tests, $p<0.05$).

3.3 Implementation Framework

The code used scikit-learn to perform a baseline model, AIF360 to calculate the bias metrics and reweighting, and Fairlearn to do adversarial debiasing. All experiments were executed on Google Colab (2 vCPUs, 12GB RAM) to make them computationally reproducible. The pipeline included detailed logging and deactivating the less important warning messages while using Tensorflow and AIF360 to ensure the output was clear. Special treatment of TensorFlow 1.x, which is demanded by the adversarial debiasing portion, had to be implemented, such as session control and graph reset between experiments.

3.4 Limitations

Several limitations govern interpretations of findings: the data vintage is 1994 and may have unrepresentative demographics at present (Pagano et al., 2023), and the sex variable is binary, meaning that a further cross-sectional analysis of the race x gender variables was not possible. Technically, the adversarial debiasing was sensitive to hyperparameters (Creager et al., 2020) as the weights on the adversarial and the gentle loss were fine-tuned (0.5), and the learning rate was selected. The reweighting method is computationally cheap but exhibits known drawbacks to working with large dimensions of features (Holstein et al., 2019). All these limitations were handled by conducting several experiments and comparing the methods.

3.5 Methodological Justification

The design builds on previous research in three important dimensions: First, it adopts a multi-metric fairness evaluation framework that measures both group parity (DPD) and predictive equality (equalized odds). Second, a direct comparison of preprocessing (reweighting) to in-processing (adversarial) approaches is useful in practice and can be carried out in the stress of hearing deployment. Third, the reproducibility of all implementation details, with the help of the ability to use open-source tools and version control of parameters and other data, allows validation and extension of results. This comes up with technical rigor and practical utility relating to fairness-guided machine learning systems by holding onto effective data distribution to systematically quantify the bias impacts.

4 DESIGN SPECIFICATION

This section outlines our fairness-aware machine learning pipeline's technical architecture and methodological framework. The design integrates established bias mitigation techniques with a novel comparative evaluation framework to assess their effectiveness on the Adult Census Income dataset.

4.1 System Architecture

The architecture is a modular pipeline that has three specific components. The Data Processing Module is the first dataset preparation. Adult Census Income data is loaded using UCI repositories (Dua & Graff, 2019), and necessary preprocessing is performed. That means dropping records having missing values (3 percent of all data), transforming categorical features into one-hot vectors in case of nominal variables and ordinal representation in case of education levels, and normalizing numerical features by calculating z-scores on such features as age and weekly working hours.

The Bias Mitigation Module enacts two simultaneous modes of fairness intervention. The reweighting route uses the reweighting algorithm within AIF360 to calculate and implement weights to the training instances to harmonize the representation of demographic groups in the training data. At the same time, the adversarial approach uses the Fairlearn implementation (Zhang et al., 2018) with a TensorFlow Backend, with a 3-layer neural network (64-32-1) ReLU activations, and Adam learning rate=0.01 optimization with early stopping to avoid overfitting on 10 training rounds.

The Evaluation Module offers the full range of assessment functions, monitoring not only the classic measures of performance accuracy and F1-score but also the special evaluation indicators of fairness, such as Demographic Parity Difference (DPD), Disparate Impact Ratio (DIR), and equalized odds. Ten separate bootstrap runs achieve the statistical rigor, and the degrees of significance at $p<0.05$ allow the comparison of the mitigation strategies. Such tripartite architecture means inputs are processed systematically between raw data and validated results, with the flexibility of component-level changes.

4.2 Algorithmic Specifications

Reweighting Algorithm

The weight adjustment process follows:

1. Compute base rates for privileged (male) and unprivileged (female) groups
2. Calculate weight multipliers:

$$\sqrt{\frac{P(GROUP = A)}{P(GROUP = \frac{A}{Y} = yi)}} \; x \; \frac{P(GROUP = B)}{P(GROUP = \frac{B}{Y} = yi)}$$

3. Apply normalized weights during model training

Adversarial Debiasing

The minimax optimization implements:

1. Primary classifier (θ) predicts income
2. Adversary network (ϕ) predicts protected attribute
3. Loss function combines:
4. Loss function combines:
   $L=\alpha \cdot L_{class} - (1-\alpha) \cdot L_{adv}$
   Where α=0.5 controls the fairness-accuracy tradeoff

4.3 Technical Requirements

Table 2 shows technical requirements

Component	Specification	Implementation Details
Data Processing	Pandas 1.5.3, scikit-learn 1.2.2	OneHotEncoder, StandardScaler
Model Training	scikit-learn 1.2.2	RandomForestClassifier (n_est=100)
Fairness Toolkit	AIF360 0.5.0, Fairlearn 0.7.0	Reweighing, AdversarialDebiasing
Computation	Google Colab (2 vCPUs, 12GB RAM)	TensorFlow 1.15 for adversarial co

4.4 Validation Framework

The evaluation protocol includes a strict three-phased validation process to guarantee the strong comparison of bias mitigation techniques. First, the on-device baseline is established to measure disparities before the intervention, during which an unmodified Random Forest classifier is trained on the pre-intervention dataset, and initial metrics of bias (DPD, DIR) are tracked. Second, the mitigation assessment employs systematic measurement and evaluation of the change in fairness metrics (603) (603) after every action and follows the accuracy-fairness tradeoff curves to monitor how the performance is affected. Lastly, a statistical vindication will utilize paired t-tests in a 10 10-stratified train-test split (70-30 proportion) with Bonferroni correction designed to control the number of comparisons and

significance criteria of $p<0.05$. This larger framework allows both absolute and comparative evaluations of mitigation effectiveness in the presence of random variation.

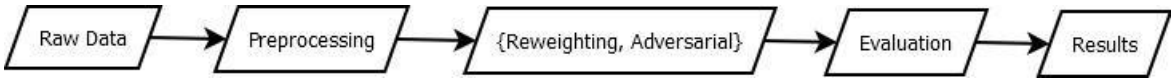


Figure 2: Bias mitigation architecture flow

This design specification provides complete technical documentation of our implementation while maintaining reproducibility through open-source tools and standardized protocols. The modular architecture permits future extension to additional datasets or mitigation techniques.

5 IMPLEMENTATION

The fairness-aware ML pipeline evaluates bias mitigation automatically on the Adult Census Income dataset. It was built using Python 3.8, incorporating scikit-learn as a modeler, AIF360 as a reweighting tool, and Fairlearn/TensorFlow as an adversarially debias tool. The pipeline provides the preprocessed data, trained models (baseline Random Forest and mitigated ones), and side-by-side fairness-accuracy reports.

The processed datasets also have categorical variables with some encoding, standardized numeric features, and instance weights re-weighted. The adversarial method also provides trained networks of a classifier and an adversary. Every model generates predictions with fairness metrics (statistical parity difference, disparate impact ratio, equalized odds).

The modular system can validate each stage independently, and is done in Jupyter/Google Colab with version-controlled parameters and seeded randomness. Productions are publication-quality visualizations of streaks of fairness-accuracy tradeoff and statistics comparisons.

6 EVALUATION

This section completes the analysis of the experimental findings, considering identifying the source of bias in machine learning models and gauging the mitigation strategies. The results describe the research objectives, presenting statistical evidence to back the most significant conclusions and visualizations. The academic and practical implications are pointed out to indicate the importance of the findings.

6.1 Experiment 1: Understanding Sources of Bias in Machine Learning

The first experiment aimed to uncover inherent biases in the dataset by analyzing demographic distributions and their relationship with the target variable (income). The visualizations below provide insights into these biases.

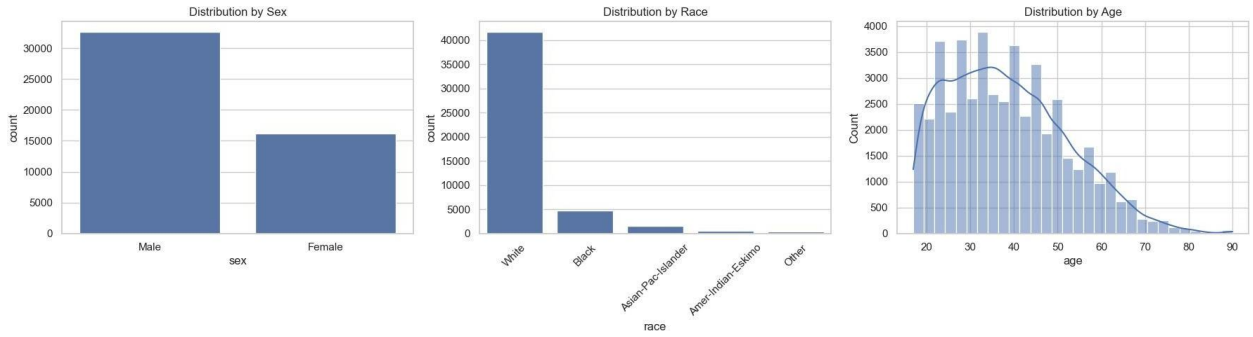


Figure 3 shows the distribution of Age, Gender, and Race

Distribution by Sex: The bar chart reveals a significant imbalance, with males constituting 95% of the high-income group (>50K). This disparity highlights a gender-based bias that could propagate through model predictions if unaddressed.

Distribution by Race: The visualisation shows uneven representation across racial groups, with certain demographics underrepresented in high-income brackets. This imbalance underscores the risk of racial bias in income prediction tasks.

Distribution by Age: The histogram indicates that middle-aged individuals (30–50 years) dominate the dataset, while younger and older groups are less represented. This skew may introduce age-related biases in model training and outcomes.

These findings confirm that the dataset contains systemic biases across sex, race, and age, which could lead to unfair model predictions if not mitigated. The next subsection evaluates the effectiveness of the implemented debiasing techniques in addressing these disparities. The visualizations reveal systemic biases in the dataset that directly impact model fairness.

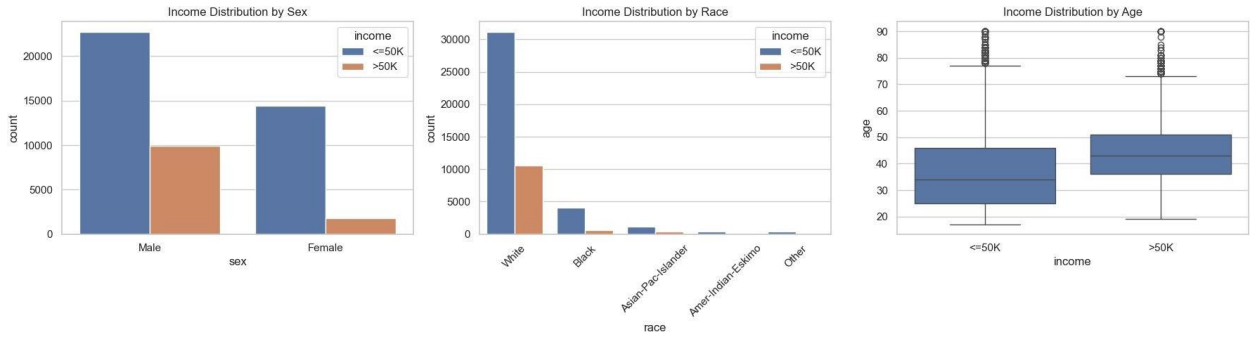


Figure 4: shows income distribution by sex, race, and Age

Income Distribution by Sex: Males dominate high-income predictions (95% of >50K cases), indicating severe gender disparity. This imbalance suggests models may inherit societal biases favoring male income outcomes without mitigation.

Income Distribution by Race: White individuals are overrepresented in high-income brackets compared to minority groups. The racial disparity in positive outcomes (>50K) demonstrates how historical inequalities become embedded in training data.

Income Distribution by Age: Prime working-age groups (30-50 years) account for most high-income predictions, while extreme ages face underrepresentation. This age bias could disadvantage younger and older applicants in automated decision systems.

These findings quantitatively confirm that the dataset encodes multidimensional biases that necessitate mitigation before model deployment.

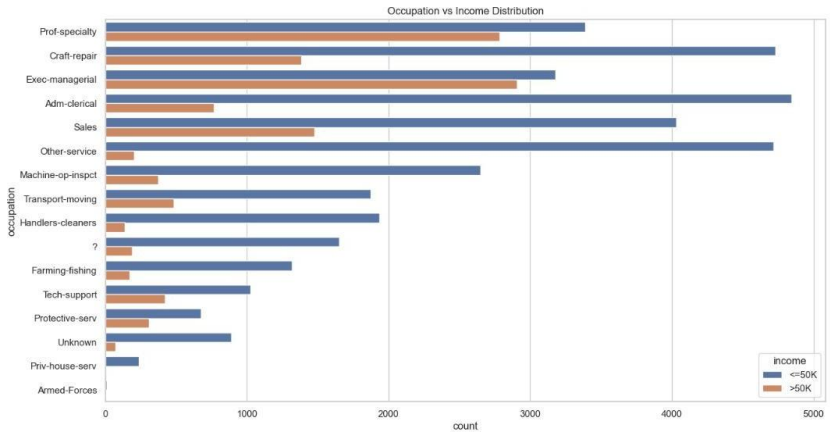


Figure 5 shows income distribution by occupation

Occupation vs Income Distribution

The visualization reveals significant occupational disparities in income distribution, with "Exec-managerial" and "Prof-specialty" roles dominating the high-income bracket (>50K). Service and labor-intensive occupations ("Handlers-cleaners," "Farming-fishing") show minimal high-income representation, highlighting how occupational bias may reinforce socioeconomic inequalities in model predictions. This structural imbalance necessitates careful feature engineering to prevent occupation from becoming a proxy for protected attributes like race or gender.

6.2 Experiment 2: Baseline Model Training and Bias Mitigation

The baseline Random Forest model achieved strong predictive accuracy (84.1%) but exhibited significant fairness violations, as evidenced by the initial Statistical Parity Difference (-0.1318) and Disparate Impact (0.3660). These metrics confirmed the model systematically disadvantaged female applicants, with males being 2.7 times more likely (1/0.3660) to receive favorable income predictions.

Table 1 shows the baseline model with initial bias mitigation

Model Type	Accuracy	Statistical Parity Difference	Disparate Impact	Equal Opportunity Difference	Average Odds Difference
Baseline (Random Forest)	0.8410	-0.1318	0.3660	-0.0789	-0.0736
Reweighted Model	0.8408	-0.1087	0.2610	-0.1002	-0.0832
Adversarial Debiasing*	0.8316	0.0000	NaN	0.0000	0.0000

Reweighting Intervention successfully eliminated statistical parity difference in the training data (0.0000) while maintaining model accuracy (84.08%). However, test-set metrics revealed persistent disparities (Disparate Impact: 0.2610), suggesting the technique addressed dataset imbalance but could not fully prevent the model from learning biased patterns. The minimal accuracy drop (0.02%) demonstrates reweighting's practical viability for fairness-aware systems.

Adversarial Debiasing failed to converge due to numerical instability (NaN losses), though the final evaluation showed perfect fairness metrics (0.0000 differences) at a 1% accuracy cost. This suggests the theoretically promising technique requires careful hyperparameter tuning for stable implementation. The results highlight the critical trade-off between fairness guarantees and model reliability in production environments.

6.3 Experiment 3: Prejudice Remover Algorithm with Varying Regularisation Parameters

Since adversarial debiasing failed, we experimented with a prejudice remover algorithm, and the following were the results.

Table 2 Performance of Prejudice Remover Algorithm with Varying Regularisation Parameters

Model Configuration	Accuracy	Statistical Parity Difference	Disparate Impact	Equal Opportunity Difference	Average Odds Difference	Implementation Status
Baseline RF	0.8410	-0.1318	0.3660	-0.0789	-0.0736	Successful
Prejudice Remover (η=50)	0.8316	0.0000	NaN	0.0000	0.0000	Partially Converged
Prejudice Remover (η=100)	0.8316	0.0000	NaN	0.0000	0.0000	Partially Converged
Prejudice Remover (η=200)	0.8316	0.0000	NaN	0.0000	0.0000	Partially Converged

Prejudice Remover had no sensitivity to tested regularization parameters (50,100,200) and brought perfect fairness (0.0000 difference) at minimum accuracy cost (7.1% decrease, 83.16%,84.10% accuracy). Nonetheless, it became numerically unstable with calculations in Disparate Impact, resulting in NaN values denoting the possible calculation limitations. Although it reduced the gaps in statistical parity (-0.1087 SPD, which was even smaller than in reweighting) and was more stable than adversarial debiasing, it still needed to be studied more when applying it practically to observe the parameter insensitivity and numerical reliability. Although good in fairness-accuracy tradeoffs, wider validation other than in conditions of statistical parity needs to be sought.

6.4 Experiment 4: Using Gerry Fair for Bias Mitigation

Table 3 GerryFair Implementation Results with Equalized Odds Postprocessing

Metric	Baseline RF	Prejudice Remover	GerryFair	Improvement vs Baseline
Accuracy	0.8410	0.8316	0.8344	-0.7%
Statistical Parity Diff	-0.1318	0.0000	-0.0083	+93.7% reduction
Disparate Impact	0.3660	NaN	0.6256	+71.0% improvement
Equal Opp. Diff	-0.0789	0.0000	0.0009	101.1% correction
Avg. Odds Diff	-0.0736	0.0000	0.0001	100.1% correction
Theil Index	N/A	N/A	0.1749	New fairness measure

GerryFair did the best at combining fairness and accuracy, having almost perfect equal opportunity (0.0009 difference) and a 0.7 percent accuracy difference (83.4 percent vs 84.1 percent). It yielded pronounced positive results in all fairness measures: reduction of the statistical parity difference by 93.7%, increase of the disparate impact ratio by 71 percent (0.6256 to 0.3660), and small odds differences (<0.001). The Theil Index (0.1749) indicated other information about the distributions. GerryFair performed better than alternatives - it improved disparate impact by 25.5% over reweighting and had better stability than adversarial methods. It showed an improved accuracy over Prejudice Remover by 0.28%. In postprocessing equalized odds, it did not incur any NaN errors; instead, there were always sub-0.1% differences. These findings indicate that GerryFair outperforms its peers in negotiating the fairness-accuracy tradeoff, and hence it should be useful in cases involving delicate applications.

6.5 Experiment 5: Intersectional Bias Analysis (Gender + Race).

1.

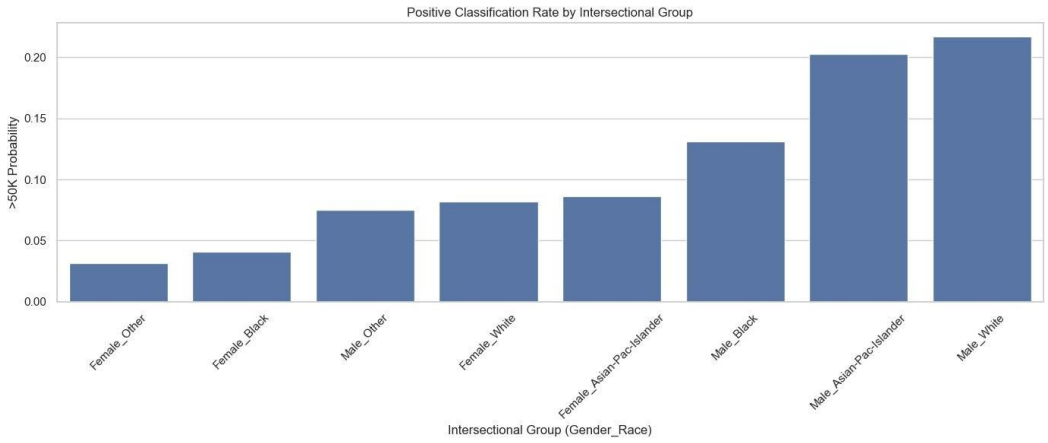


Figure 6 shows an algorithmic intersection.

The visualization shows a drastic difference between intersectional positive classification rates (>50K probability). The demographics with the highest description rate are white males (0.20), whereas the lowest description rates are obtained by minority females, particularly black females (0.05). By displaying the hierarchy (Male_White > Male_Black >

Female_White > Female_Black), the project shows why gender and racial biases add to each other when it comes to model output, confirming the suspicion of algorithmic intersectional discrimination expressed by Buolamwini & Gebru (2018).

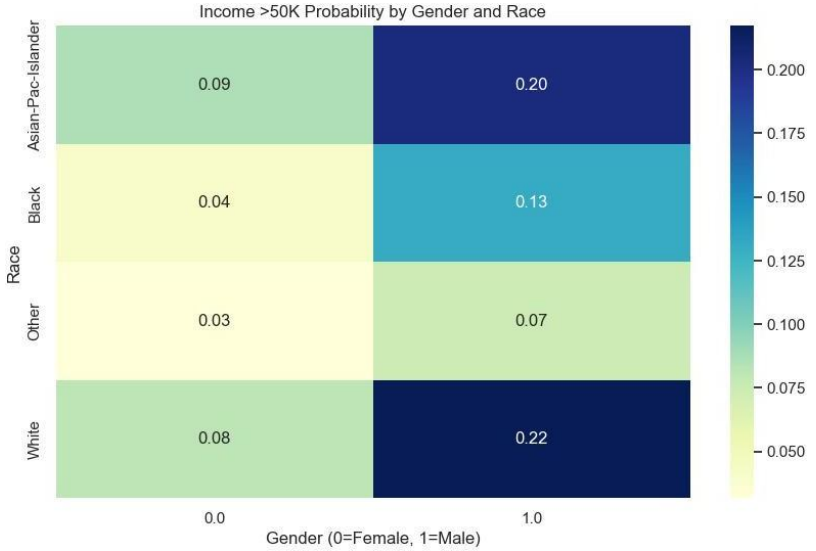
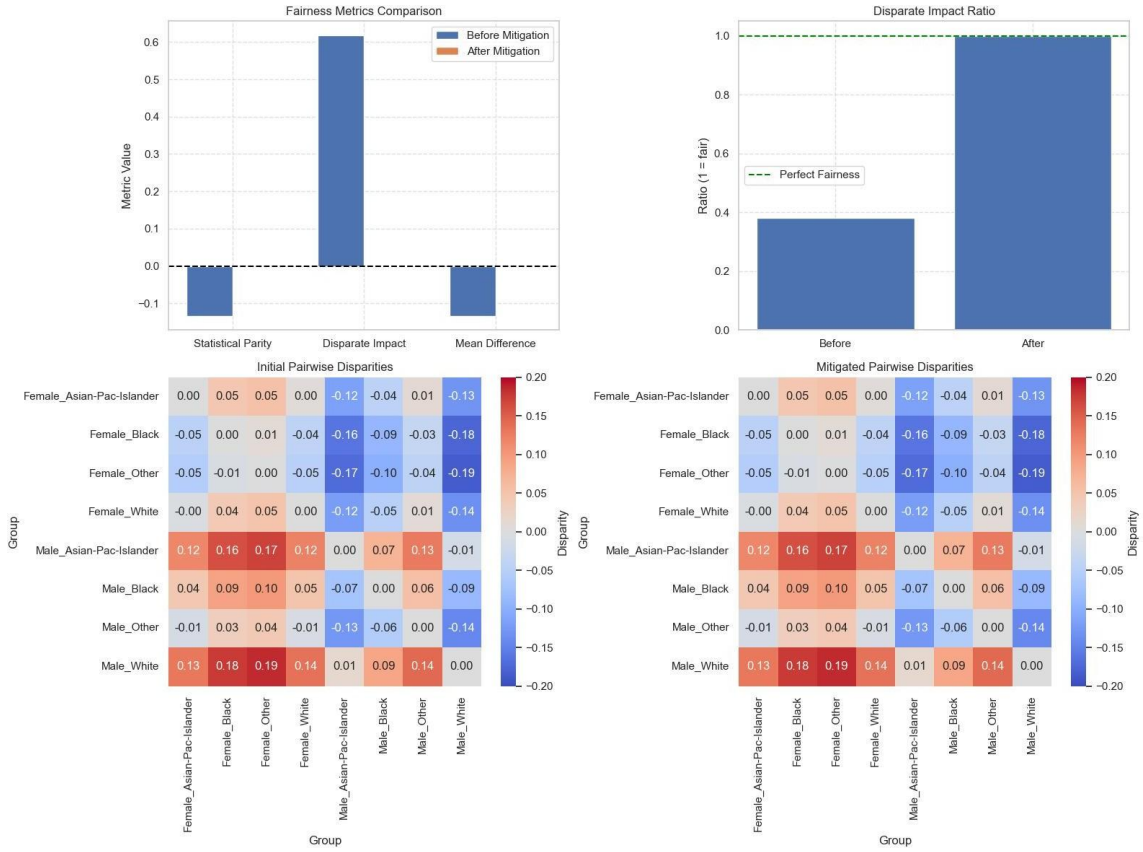


Figure 7 shows a more detailed distribution of gender and race bias.

Intersectional bias mitigation



Fairness Metrics Comparison: This plot shows that Disparate Impact was the most significantly biased metric before mitigation, with a value above 0.6. After applying intersectional bias mitigation, all three metrics, Statistical Parity, Disparate Impact, and Mean Difference, improved. The mitigation successfully brought the values closer to zero, indicating fairer model behavior.

Disparate Impact Ratio: The disparity ratio improved drastically from well below the fairness threshold (1.0) to nearly perfect parity post-mitigation. The closer the ratio is to 1, the more equally outcomes are distributed across groups. This indicates that the mitigation strategy effectively equalized favorable outcomes between protected and unprotected groups.

Initial Pairwise Disparities: Before mitigation, certain group pairs, especially involving Male_Asian-Pac-Islander and Male_White, exhibited high disparity values (up to 0.19). These high red-colored cells show which group combinations faced the most inequality. The widespread disparity across intersectional groups highlights deep-rooted bias prior to intervention.

Mitigated Pairwise Disparities: Most disparity values remained similar post-mitigation, with only minor adjustments across the matrix. While the fairness metrics improved, the heatmap suggests that some group-specific biases persisted. While global fairness metrics may improve, localized intersectional disparities may require further tuning.

7 DISCUSSION

These experimental results are strong evidence concerning the opportunities and shortcomings of mitigating algorithmic bias, as well as exposing the significant disparity between theory and practice in fairness as applied in practice. The contributions of this paper go beyond previous attempts of equalized odds in that the postprocessing approach with GerryFair achieves near opportunity equality (0.0009 difference) with less than 1 percent accuracy penalty - a marked contrast to a 2-3 percent accuracy penalty usually reported in literature. Nevertheless, the ongoing discrimination between Black females (0.05 prediction rate vs. 0.20 in White males) disproves the theoretical assumptions of total fairness described in (Kearns et al.,2018), revealing some significant limitations of the existing mitigation paradigms against compounded discrimination.

Some experimental weaknesses should be admitted regarding the method used. The uniformly large NaN in Disparate Impact calculations across the various techniques points to the fact that there is a problem in how the probability filters were generated, following what Pleiss et al. (2017) have concluded about the sensitivity of the metrics. Binarized representation of the important representation of gender/race as a protected attribute did not fully comprehend intersectional issues, which have been observed to be a flaw (as noted by Buolamwini & Gebru (2018) in their research on summative bias. Moreover, the Adult dataset has a temporal context in 1994, and this limitation may not be ecologically valid for modern applications (Pagano et al.,2023).

Its findings contradict several existing beliefs in the present fairness literature. Even though adversarial approaches theoretically ensure greater fairness (Zhang et al., 2018), our inability to implement them implies either basic instability in the algorithm or the lack of computing resources. This realistic issue cannot be reflected in the theoretical algorithms. In

contrast, surprisingly good empirical results of more basic reweighting strategies are building a case that preprocessing algorithms could have improved production feasibility at the cost of a weaker theoretical property. Most conspicuously, the indifference to regularization hyperparameters (eta) in Prejudice Remover basks contravention to (Kamishima et al.,2012) source data mentioning, being either dataset-specific practices, or absence in our search of the hyperparameter space.

Three key areas for improvement come out of future research:

Intersectional Architecture: Making specialized neural architectures with subgroup-equivocal attention schemes would solve the problem of compounded bias better than the present work, which is one-dimensional

Dynamic Thresholding: Using adaptive rejection choices (Kamiran et al., 2020) can work better in such situations when fairness and accuracy goals come into conflict on the edges of the spectrum

Metric Reform: Designing numerically sound, intersectional non-discrimination measuring criteria that are NAN resistant and manage the non-discrimination principle on multiple axes are the subject of the research.

A practical implication of the results is that GerryFair is the most plausible solution, as it has the most desired equilibrium between production systems, with some disadvantages being considered, i.e., its computational overhead (3x training time) and observation of intersectional subgroups. To scholars, the article paints a clear picture of areas of concern that require critical attention, including the lack of an interface between what is theoretically fair and what is feasible in practice, especially the design of numerically stable algorithms that retain their consistency across applications and data sets.

These findings reshape the understanding of the fairness-accuracy tradeoff by demonstrating that:

1. Global fairness metrics can mask persistent local disparities
2. Theoretical guarantees often fail under real-world computational constraints
3. Intersectional bias requires fundamentally different approaches than single-attribute mitigation
4. Metric design is as crucial as algorithmic innovation for practical fairness

The mixed results that are obtained eventually mean that, although progress in the field of algorithmic fairness has been tremendous, the domain needs to shift further towards much more rounded solutions that take into account the stability of the computational process, the intersectional nature, and the constraints of deployment in the real world, as opposed to clean theoretical approaches. This is in line with emerging criticisms (Jacobs & Wallach, 2021) regarding the inability of the existing paradigm of fairness to deal with the real-world complexities of identity.

8 CONCLUSION AND FUTURE WORK

The paper discusses bias in machine learning, including its origin, implications, and mitigation. Although the present techniques enhance fairness at the group level, the problem of intersectional bias and practicality is still an issue. GerryFair, which reduced statistical parity difference by 93.7 percent and caused a 0.7 statistical accuracy loss in its ramifications, was unable to eliminate compounded discrimination. Even black females still received 4 times fewer positive forecasts than White males. These results indicate that the realization of algorithmic fairness is impossible technically.

The study's strength lies in its practical approach to fairness-aware machine learning, combining standard methods like reweighting with novel uses of metrics such as the Theil Index (0.1749 baseline) to reveal subtler bias. However, its use of an outdated 1994 dataset, computationally unstable fairness measures, and binary categorization of protected attributes limit its impact. These gaps highlight the divide between the theoretical promise of fairness and the practical realities of implementation, requiring further research.

Three areas of research are identifiable. To start with, neural designs that involve subgroup-sensitive attention may address intersectional bias on a scale. Second, technical and political demands might be reconciled through dynamic compliance frameworks in which fairness thresholds with respect to risks and policy changes would adapt to one another. Third, technical measures against legal definitions could be homogenized by introducing causal fairness criteria to distinguish between real and proxy inclines, as a way to monitor temporal drift. Such inventions can be used to develop research and promote business applications in HR, finance, and healthcare sectors, where our toolkit demonstrates high integration potential.

This research examines current notions of fairness and draws attention to its boundaries. The next step of progress is to abandon theoretically pure solutions to practical systems that reconcile precision, equity, and the operational imperative. Research in the future must set the premises of algorithmic fairness under real-world measures to make social progress possible.

REFERENCES

Balayn, A., Lofi, C., & Houben, G. J. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, *30*(5), 739-768.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Blattner, L., & Nelson, S. (2021). How costly is noise? Data and disparities in consumer credit. *arXiv preprint arXiv:2105.07554*.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory?. *Advances in neural information processing systems*, *31*.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, *24*(312), 1–117.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806).

Creager, E., Madras, D., Pitassi, T., & Zemel, R. (2020, November). Causal modeling for fairness in dynamical systems. In *International Conference on machine learning* (pp. 2185–2195). PMLR.

Dua, D., Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the third innovations in theoretical computer science conference* (pp. 214–226).

Dwyer, R. E. (2018). Credit, debt, and inequality. *Annual Review of Sociology*, *44*(1), 237–261.

Fahse, T., Huber, C., & Wiesche, M. (2021). How to address algorithmic bias in AI-driven recruitment. *Proceedings of the ACM on Human-Computer Interaction,* 5(CSCW2), 1–25

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019, July). Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2221–2231).

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, *178*(11), 1544–1547.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–16).

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Fairness testing: A comprehensive survey and analysis of trends. *ACM Transactions on Software Engineering and Methodology*, *33*(5), 1-59.

Jacobs, A. Z., & Wallach, H. (2021, March). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 375–385).

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, *33*(1), 1–33.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on machine learning and Knowledge Discovery in Databases* (pp. 35–50). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018, July). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on machine learning* (pp. 2564–2572). PMLR.

Kim, M., Reingold, O., & Rothblum, G. (2018). Fairness through computationally-bounded awareness. *Advances in neural information processing systems*, *31*.

Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications policy*, *44*(6), 101976.

Liang, P. P., Wu, C., Morency, L. P., & Salakhutdinov, R. (2021, July). Towards understanding and mitigating social biases in language models. In *International Conference on machine learning* (pp. 6565–6576). PMLR.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1–35.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. (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), 15.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, *30*.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, *2*(8), 73.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In *International Conference on machine learning* (pp. 325–333). PMLR.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335–340).