Frontiers in Health Informatics www.healthinformaticsjournal.com
ISSN-Online: 2676-7104
2024; Vol 13: Issue 4 Open Access

Mitigating Data Bias In Machine Learning: Enhancing Model Transparency Through Fairness-Aware Techniques

Harshal Dalvi[1*], Meera Narvekar[2], Shehal Shah[3], Priyal Donda[4], Jeniel Shah[5]
[1*]Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, harshal.dalvi@djsce.ac.in,
[2]Professor, Dept. of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, meera.narvekar@djsce.ac.in
[3,4,5]Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, shehalshah264@gmail.com, priyaldonda3019@gmail.com, jeniels72@gmail.com

Cite this paper as: Harshal Dalvi , (2024) Mitigating Data Bias In Machine Learning: Enhancing Model Transparency Through Fairness-Aware Techniques. *Frontiers in Health Informatics*, 13 (4), 715-729

Abstract: As the Machine Learning algorithms are increasingly influencing the decision-making process across multiple domains like healthcare and finance, it is posing an important challenge of the bias in data. Having a highly accurate AI system working on dataset containing bias will not only have skewed predictions and unintended societal consequences but can adversely affect the transparency and reliability of the system as well. This research is aimed at exploring biases in medical data within machine learning models that even exhibit ethical concerns. Bias generally refers to systematic favouring or prejudice often due to imbalances in medical records used to train algorithms. The results capture the effects of bias in the medical dataset and foretell likely outcomes of the model while addressing the problem at its root. We used the MEPS medical dataset, which showed disparities based on protected classes such as race and ethnicity or age and calculated measures of bias for the classification model's predictions. Further, we used some of the open-source tools for generating reports about biased outcomes and were calculating bias by emphasizing metrics such as Disparate Impact using IBM's fairness measures. The study compares four machine learning models: Logistic Regression, Decision Tree, Random Forest, and SVM and measures the correlation of model complexity to bias in predictions. To mitigate it, various approaches such as re-weighing, adversarial debiasing, ROC post-processing, and prejudiced remover were deployed which showed trade-offs in terms of bias reduction and model accuracy. Future research directions are concluded to advance ethical and responsible AI for medical decision-making.

Keywords: Bias in data, Bias Mitigation, Ethical AI, Fairness measurement, Responsible AI, Transparent AI.

Introduction

Machine learning is ubiquitous within various domains and introduces transformative capabilities to revolutionize the way complex problems are approached and decisions in daily life are made today. However, real-world adoption of machine learning models has raised a critical challenge: bias pervades the approach. This research paper addresses the critical question of data bias, which has a heavy bearing on the operation of machine learning models. In the context of the term, bias in machine learning refers to systematic and often unintentional favouritism or discrimination that reveals itself in the predictions and recommendations produced by these algorithms. It is a multifaceted issue with its roots in the intricate web of the machine learning pipeline. The training data is the foundation of bias in machine learning. When machine learning models are trained, they learn patterns and relationships from the data that has been provided. If the training data contains biases, whether as a result of historical imbalances, societal prejudices, or cultural biases, the model will inherit and perhaps aggravate them. The model perceives and interprets the world via the prism of the training data. If this lens is contaminated with biased data, the model's predictions and judgments may reflect and propagate such biases. Understanding and correcting bias in training data is critical for creating fair, transparent, and ethical machine learning algorithms. It entails careful data preprocessing, knowledge of potential biases, and the use of bias mitigation and management measures throughout the model construction process. For instance, a hiring model trained dataset might inadvertently favour specific demographics or educational backgrounds. This simply means that the model will predict future hires in terms of biases, perpetuating those imbalances again. Machine learning bias is also one of the major ethical and social concerns because it extends beyond the technical barriers in much the same way that significance derives from its potential to continue and exacerbate imbalances and inequalities into the dimensions with which people navigate their lives. The more machine learning models are applied to make decisions, the more two-edged sword it becomes as far as the presence of bias is concerned. While these models assure objective conclusions grounded in data, on the other hand,

the ingrained biases in the data sources might unknowingly perpetuate and expand societal disparities unless they are properly addressed. It captures the conceptual concern of something fundamentally desirable in this issue, in that models should not only be technically correct but also conceptually fair, just, and capable of leading to a well-balanced society. The wide applicability of machine learning models into all areas of applications, from simple recommendation systems to complicated autonomous vehicles, underlines how sharply impact modern society and that such models can transform how we go about interacting with technology and making decisions. Their quality and representativeness of the data they have been trained on are intricately tied to their effectiveness and fairness. Bias in datasets is a big, persisting concern that resonates throughout different industries. Biased datasets might result in not-so accurate predictive power, reinforce discriminatory societal norms, or culminate in unfair, discriminatory outputs. Addressing biases in datasets of machine learning is an important milestone while obtaining true responsibility and equity in deploying these models. It is about discovering and understanding existing biases but also about how to devise efficient strategies toward data collection, preprocessing, and model testing. As machine learning is increasingly becoming controlling in lots of aspects of human life, its ability to minimize bias is the key for the successful development of fair, transparent, and trustworthy AI. This study contributes to that effort to help get rid of bias among all the machine learning algorithms. What's most critical is that there will be objective testing and comparison of how all these different models perform in the context of biased datasets, through thorough probing on what bias entails and exploring various mitigation approaches; hope is that it might even end up informing the moulding of a more trustworthy future for AI systems-one where decisions are made with fairness and transparency.

Literature Survey

Related Studies and Findings

Advances in healthcare through machine learning models and data-driven algorithms have meant a whole shift in decision-making at a healthcare level, especially when one talks of disease prediction and the allocation of healthcare resources. These advancements, however, have opened other avenues that show very critical biases that can promote inequalities in healthcare settings. Several seminal studies investigated these biases and attempted to unmask how AI models can influence the health status of minority populations. The bias in AI models for healthcare considered by Abraham, MD [1], the criminal risk assessment system discussed by (Angwin et al. 2016) [2], salary prediction algorithms highlighted by the BBC in 2018 [3] and the loan approval decision making process reported by Swarns in 2015 (Swarns 2015) [4] are all examples of how these algorithms can inadvertently perpetuate societal biases. Among the most highly relevant and impactful studies of this kind was that by Obermeyer et al. 2019 [5]. The study described the analysis of the algorithm deployed to the U.S. healthcare system for resource allocation according to predicted health needs. The algorithm consistently assigned lower health risk scores to Black patients compared to white patients, despite both groups having similar medical conditions. The root cause was the algorithm's reliance on healthcare spending as a proxy for health status. Hence, the risk for Black patients, who, due to systemic barriers, have historically spent less in healthcare, was underestimated. This therefore went into a feedback cycle, whereby the algorithm kept on reinforcing healthcare inequalities by giving less resources to the people that needed them most. The findings from this study show that the dangers of using cost-based metrics instead of clinical health indicators led to the redesigning of AI models to focus on equitable measures such as clinical outcomes instead of monetary data. Another significant study is that of Celi et al. [6]. It is a systematic review of AI-based diagnostic tools through the unearthing of wide performance gaps when these algorithms were applied to minority populations, their research cut across a wide range of diseases, including diabetes, cardiovascular conditions, and respiratory illnesses. A common setback of these diagnostic models was that they underperformed in minority patients, since they had been largely trained on homogeneous datasets, leading to increased misdiagnosis and delayed treatments. This lack of diverse training data meant the algorithms were less accurate for underrepresented populations, ultimately exacerbating healthcare inequalities for those groups. Ghassemi et al. [7] have studied gender bias in AI treatment recommendation systems specific for cardiovascular diseases. They have found that women were treated with less aggressive treatments as compared to men who had similar profiles. Bias was observed due to not having adequate training data to precisely portray female patients. The research also showed that such kinds of gender biases were present in chronic conditions like diabetes and chronic kidney disease diagnosis. This has brought out a need for validating AI models across different demographic categories to ensure treatment recommendations are equitable rather than being biased from flawed historical data.

This paper underlines the necessity of such scrutiny and mitigation while in the development and use of such models, to have their deployment yielding fair and just results, without making space for once-stigmatized practices, either for sustaining and furthering them or to let them live on. Another paper by Ueda G et al. [8] that discusses fairness in AI models for healthcare especially radiology identifies main slight sources of bias for example, data bias where the datasets are imbalanced, user bias where underrepresentation of the minorities is noted. The authors provide guidance on how more varied data should be used, how transparency should be enhanced, and how fairness measures should be incorporated during validation. They also emphasize the need for constant evaluation of AI models to maintain high quality and eliminate bias in minority groups in the usage of AI in healthcare.

Few instances highlight the point that machine learning model development and use necessitates scrutiny and bias checking. These algorithms require appropriate efforts not just to predict the outcomes accurately but also to predict them fairly and without bias, to avoid perniciousness in the sense of social inequality and discrimination. This kind of recognition calls for a broader set of considerations on ethical issues, and transparency in handling the use of such AI technologies within decision-making processes that have real and palpable impacts on a person or community. New constraints on algorithms will require innovative research approaches stemming from data science, artificial intelligence, and machine learning. Bias is as much a problem with the data used for training as it is an outcome from the practice of machine learning in real-world application spaces. Amazon's AI recruitment tool encountered substantial challenges as it exhibited bias against women in the hiring process which was highlighted by Dastin et al. [9]. The tool used for recruitment purposes was criticized for displaying discriminatory tendencies. However, the company recognized the gender bias that was built into the system and decided to discontinue it. This was all set against the ethical considerations and exposure to the potential harm that biased algorithms can do, especially in relation to hiring, an area where fairness and impartiality are extremely critical. The incident underscores how vigilance and transparency in AI technologies, especially those related to human decisions, should be at their peak. Detection and mitigation of bias are also important to ensure that this does not enable systems like these to simply perpetuate or even amplify societal inequality. The ethical dilemmas and health equity implications of artificial intelligence (AI) application in public health and medicine are highlighted by the authors Dankwa-Mullan et al. [10]. The paper explains how, without appropriate AI design and oversight, AI systems can entrench current health inequities between groups, particularly minority groups. The paper also stresses that the specific standards of ethical concerns must be created for AI in healthcare – and they must be comprehensible, fair, and inclusive. Based on the paper's findings, the implementation of AI-based innovations should transpire in partnership with policies, technology experts, and healthcare organizations to enhance equitable and preparatory accessibility of the technology for all populations based on SES and ethnicity.

TABLE 1: ML models and type of bias reported

AI model	Year	Reported bias
Recidivism assessor (COM- PAS)	2016	Racial discrimination
Google's Ads	2013	Racial discrimination
Amazon's AI recruiting tool	2018	Gender discrimination
Face Recognition	2018	Poor accuracy for minori- ties
Twitter's image preview	2020	Racial discrimination
Facebook's job ads	2021	Gender discrimination

Amazon's decision to discontinue the tool reflects a commitment to responsible AI practices and highlights the ongoing challenges in creating algorithms that are fair, unbiased, and aligned with ethical principles. Table 1 shows a few examples of models that seemed to report bias.

Bickel, Hammel, and O'Connell (1975) [11] highlighted this complexity as early as 1975. Another widely used fairness metric is Equal Opportunity, which mandates that individuals who genuinely qualify for an opportunity (the true positive group) should have an equal likelihood of being predicted as such, irrespective of their group affiliation.

Prior works on ML fairness have established statistical fair- ness metrics such as statistical parity, equalized odds, disparate impact, equal opportunity and test fairness to evaluate notions of fairness presented by Mishraky et al. [12]. Notable studies have had groundbreaking work that shed light on the gender and racial biases present in commercial AI systems. To address these biases, several techniques of mitigation have been developed. Adversarial debiasing indicates that an adversarial approach can successfully train models in such a way as not to compete with accuracy but to decrease discrimination.

It could be insightful to give a review of the bias mitigation techniques in depth to tell of strategy development for proper and just AI systems:

1) **Pre-processing:** Preprocessing will pre-process the training data so it can be treated by the model. This would now eliminate bias when training that model because this preprocessing could include re-sampling, reweighting, and even data augmentation.
2) **In-processing:** This type of method relies on varying the learning algorithm at model fitting time such that the bias may be controlled. In-processing methods are most effective when bias from the algorithm itself is likely to be present.
3) **Post-processing:** This is the modification of the output by the trained model after the training. It is most often applied to correct bias induced during the training procedure.

Some contributions were written to discuss datasets alone when it comes to bias and unfairness studies, not engaging with matters of mitigation (Fu et al. 2023) [13]. Some studies explore whether neighbourhood socioeconomic status influences recidivism returns to distressed neighbourhoods are indeed associated with higher recidivism rates even when controlling

for individual characteristics, as the following study by Charis E. Kubrin and Eric A. Stewart reports [14]. A categorization of fairness definitions has been formulated by machine learning researchers to address the prevalent bias in AI systems and promote fairness. AIF360 [15], Aquetias (aequitas, n.d.) [16], Themis-ML [17] and FairLearn [18] have developed valuable tools that play a significant role in identifying and mitigating biases and unfairness in optimizing search criteria. These tools are a collection of the most important resources to be found in research and literature towards being able to investigate these issues. Fairlearn aims at assisting practitioners to assess and improve the fairness of AI systems with the provision of tools to analyse the output of the model by different populations as well as algorithms for addressing fairness challenges in a sociotechnical setting [17]. In 2016, Angwin et al. published an investigation on ProPublica whereby COMPAS shows discrimination against the blacks behind the bars, as concluded from one analysis that applied Florida inmate data [18]. While the performance of COMPAS is equivalent for both white and black inmates regarding its accuracy, as it has been concluded from the study's conclusions, COMPAS makes a specific type of error that signifies unfair treatment of the black inmates, as concluded by Angwin. There happens to be an extremely well-known healthcare dataset that speaks volumes for the problems seen with COMPAS in criminal justice, where predictive healthcare algorithms are developed from the Optum dataset which skews their AI healthcare tools against Black patients. An example would be with the 2019 study where AI, developed with the Optum dataset, underassessed the need for Black patients to be considered for health care through using healthcare costs as a proxy for medical need. This led to stark racial inequalities in whom, at least statistically, required most of the additional care, in that Black patient, due to the typically systemic nature of disparities, also spent less money on healthcare. There has been a myriad of open-source contributions around fairness in AI and mitigation tools that have been developed as shown in Table 2.

TABLE 2: Literature Summary of Open-Source Contributions Around Fairness in AI

Mitigation Toolkit	Year	Contributors
Toolkit with comprehensive set of metrics and tools to mitigate bias	2018	IBM
Issues a bias report given the relative distinct attributes	2018	Saleiro et al.
Dashboard to visualize fairness metrics and algorithms to mitigate bias	2020	Bird et al.
Tools around fairness, accountability, and transparency	2020	Sokol et al.
Environment to simulate and provide training on fairness metrics	2020	D'Amour et al.

As seen in one of the findings of Elhanan Mishraky, the prediction accuracy seemed to decline as the age got younger. Another finding stated that the correlation scores for gender prediction in relation to race are exceptionally high, with a weighted average of 0.97. This infers an extremely strong relationship between the racial background and the gender of the person, meaning that if it were the person's racial background on which the intent of prediction rested, the gender prediction may be quite accurate. (Mishraky et al. 2022) [12]

Debate on whether the desired disparate impact was attained was part of the work of the Serbian researchers, who were working on optimizing to find out if this deviation resulted in the desired disparate impact. This means that an explicit criterion or threshold is not applicable to the determining criteria in the optimization procedure, thus resulting in uncertainty or ambiguity in the decision-making process [19]. They attain precision at its best when the considerations for fairness are not very important and then drop with the importance of fairness. This trade-off between accuracy and fairness is a common challenge in developing machine learning models that need to balance predictive power with ethical considerations.

Another paper by Haytham Sailea [20] reflects the application of the "SHIFT" framework where it concerns how to make AI more responsible in the context of healthcare. Evaluates the ethical, social and operational concerns on AI use, and reasserts the value of patient safety, human-centred system, bias free system. It is necessary to measure the concept of fairness because increasing inequality in access to healthcare is unwarranted and undesirable, and transparency will help solve the crisis of trust. The authors call for a partnership among policymakers, health care executives, and developers to achieve a common goal of promoting safe and ethical use of AI in health care services.

As could be noticed across this study, the common themes running for this paper are principle-based ethical AI are transparency and accountability. Data as one of the key critical components will have to be developed in more representative terms and less biased so as to remove biases that could be done through what could now be called "algorithmic accountability," which is a technique toward facilitating the transcription and interpretability of AI systems.

Proposed Fairness Assessment Metric, "Fairwash," Arrives to Help Assesses the Fairness of Machine Learning Models

Drawing from all the literature detailing mitigation of data biases and handling of data drift within ethical AI, it becomes

quite evident that a room for improvement exists in how models perform while staying true to the more ethically stringent standards and adapting to situations presented by drifting data conditions.

Previous Work on MEPS Dataset

In a related study, the authors Moninder Singh et al [21], on racial bias in healthcare utilization and expenditure, examined by drawing data from the Medical Expenditure Panel Survey (MEPS). It studies disparities for health services accessed between different racial and ethnic groups with an eye to the contribution the latter makes to overall health inequities. Statistical methods assess the degree of bias in access to health care, its quality, and outcomes. The data show considerable disparities in healthcare utilization patterns among ethnic groups, indicating systemic biases that influence healthcare decisions and spending.

Another study is "Risk adjustment and observation time: comparison between cross-sectional and 2-year panel data from the Medical Expenditure Panel Survey (MEPS)", Health Information Science and Systems, 2014 [22]. Authors have investigated how risk adjustment methods compare the effectiveness of risk adjustment methods across cross-section versus 2-year panel data from MEPS. They have emphasized the impact that selection of data type has on estimating healthcare costs and the related risks. Through their analysis, they demonstrate how panel data would better expose the individual's healthcare cost and risk over time through a good judgment of healthcare utilization patterns. This paper, in the present work, shows that cross-sectional data can be advantageous in giving pictures about the perspective of healthcare cost; however, it may miss out on the major trends and changes in the temporal positions. Such authors would, thus, advocate for the use of longitudinal data in health economics studies to enhance the accuracy of risk-adjusted models in policy making and resource allocation in health services.

Another study analysed the short-run effect that the EITC had on healthcare spending for adult US citizens. The author discusses policy implications with the interpretation that financial tools such as EITC, which in this case constitutes a safety net program, have been known to be effective as they improve access to healthcare and increase health gains for the poor. Hamad R et al. [23] analysed the data to determine how the receipt of the EITC impacts healthcare spending with special emphasis on low-income people. The results showed that EITC significantly reduces healthcare expenditures, meaning that the extra amount of money that the tax credit provides allows people to spend more on healthcare needs. Additional studies also need to be conducted to investigate the more prolonged effects of such financial interventions on health outcomes and healthcare utilization.

Watanabe JH [24] further delves into factors affecting pharmacist labour supply in the United States. Its focus is on implications for these factors: an aging demographic and increased medication usage and changes in pharmacist roles within the broader healthcare system. The author tackles elements of the trend driving demand for pharmacists: increasing prevalence of chronic disease and the increasing complexities of medication management. The study further assesses the learning and working environment of pharmacy practice and emphasizes this direction towards patient entered care models. Watanabe calls for strategic workforce planning in order to ensure that there are enough pharmacists available to meet the future healthcare needs. Finally, implications for educational preparedness are where education and training in the profession of pharmacy can be revitalized to suit the changing landscape of healthcare to engage pharmacists in effective patient care and medication management.

Mark Baunthavong et al. [25] evaluates the health care costs from Crohn's disease by using a perspective of data obtained from United States Medical Expenditure Panel Survey data for the years 2003 through 2013. The researchers look at how the cost of health care associated with managing Crohn's has been changing over time, focusing on those direct medical care expenses, such as hospital and physician office visits and drugs. The results put forward estimates of high healthcare utilization and cost burdens for treating patients with Crohn's disease. The paper concludes by emphasizing the implications for healthcare policy and resource allocation of such findings by discussing the need to implement targeted interventions that enhance how the disease is controlled and reduce costs. Research authors are nudging researchers to study further and find the cause of the increase in the cost of healthcare that patients with Crohn's bring about, thereby improving care strategies and enhancing the outcome for the patient.

Although the studies dealing with healthcare expenditures and dynamics in the healthcare system elucidate fundamental issues, none of them take into consideration the possible biases arising from the MEPS dataset. The research also lacked an understanding of the inherent biases that exist within the MEPS data. This, in turn, may result in the quality of findings and generalizability of these findings towards different populations. Further strategies toward potential biases are needed in future studies so that analyses based on data from the MEPS would precisely replicate the medical care experiences and costs of the different racial and ethnic groups. Our approach provides a way of detecting and mitigating the biases that are present in the MEPS data.

Our Approach

Vision and Research Objective

This research is concerned with ethical AI, as well as algorithmic bias, and the intricacies of model complexity with an attempt to define the accuracy-bias curve. The primary goal is to investigate the impact of model complexity and identify how various forms of bias mitigation – before, during, and after model building – influence it. During the study, a thorough examination of numerous features was conducted, using careful preprocessing techniques to make the dataset better and focus on a smaller number of variables as suggested by Dressel and Farid [26].

Considering the recent social questions arising over bias in AI models, the paper focuses on evaluating the effectiveness of the bias mitigation technique like reweighing, prejudice remover, adversarial debiasing, and ROC post processing in models of varying complexities including logistic regression, random forests, XGBoost, and neural networks. Employing the biased MEPS dataset, the paper looks at how these methods maintain fair and accurate estimations across various models.

Dataset: Medical Expenditure Panel Survey (MEPS)

The types of datasets to be used also plays a significant role in adhering to the biases present in an AI model. Arora et al [27] proposes the high quality, consistent datasets to be used on how data heterogeneity, poor documentation, and inconsistent data labelling contribute to bias in AI systems. The Medical Expenditure Panel Survey (MEPS) is one of the most famous data resources, collected by the Agency for Healthcare Research and Quality (AHRQ) [28], to study the healthcare expenditures, insurance, and utilization in the United States. Since its inception in 1996, MEPS has been used to assess the healthcare utilization performance, the cost, as well as the biases and unfairness in machine learning. This is due to the fact that the dataset includes sensitive attributes such as race, ethnicity, age, gender, and income, which are crucial for analysing fairness.

However, as with most datasets, MEPS captures societal biases inherent to the US healthcare system. These biases are socio-economic and racial in nature and can be passed to the machine learning models trained on the data. For example, even though black patients often face worse health outcomes and have higher rates of chronic illness, their healthcare expenditures are typically lower than those of white patients. Such a difference can lead to prejudiced forecasts, meaning that black patients will not be included in the forecast of the resources to be allocated to their healthcare. Thus, it becomes critical to implement bias reduction measures to prevent machine learning models from consolidating such discriminations.

Such models can be dependent on expenditure-related features when trained on biased expenditure data. However, expenditure is a poor way of representing the need for healthcare services, especially for the marginalized sections of the population experiencing healthcare inequality. All these biases need to be addressed to ensure that the AI systems being developed and deployed are more equitable.

Black patients have heavier utilization of ER visit and inpatient stay compared to white patients regardless of expenditure level. Black patients also generally show worse health indicators than other racial groups in both the overall and the top decile of healthcare spenders, experiencing longer inpatient stays, as well as more chronic conditions. This unequal treatment fosters disparate impact for the black patients when the machine learning algorithm is trained using MEPS datasets and with low probability of high costly health care predictions, thus limiting black patient's opportunity to get interventions that can improve their health status.

Measuring the extent to which people are in the first decile of the health care spenders forms the subject of the outcome variable for this study. During the assessment of bias, special focus is accorded to such features as race and gender. The MEPS dataset provides a unique chance to study healthcare inequities and justice and to make the way AI works less prejudiced and more correct, both in terms of race and ethnicity in healthcare.

Data Preprocessing

Standard set of procedures for data preprocessing were used for the MEPS dataset to clean the data and make it align with the objectives of this study. This entailed data pre-processing to clean the data and prepare it for bias correction and model training including addressing bias arising from race, gender or healthcare utilization.

In order to examine racial bias, participants were split into two groups based on their race, White and Non-White (Non-Hispanic). Such a type of division was critical for understanding differences in the healthcare expenditure between the groups. The study was conducted on Panel 19 (2014-2015), which contains more or less successive interviews of the same respondents.

Column renaming was required to achieve uniformity across the data set and normalize the parameters such as healthcare conditions, mental health, and employment status. Any record that contained missing or invalid values in the key attributes (region, age, marital status, health conditions) were excluded from the analysis to maintain data quality.

Health care utilization, which was one of the main areas of interest, was calculated by summing up measures like office visits, emergency visits, and inpatient days. This was dichotomized into high utilizers (1) with visit frequency of 10 or above and low utilizers (0), consistent with studies on high-risk patient profiling.
Racial bias was measured using the protected attribute race distinguishing privileged group (White) and unprivileged group (Non-White). This was important for measuring and assessing fairness using measures such as statistical parity difference, equal opportunity difference, disparate impact and then directing the bias reduction process.

Methodology
This research uses machine learning models of varying complexity—Logistic Regression, Random Forest, Neural Networks, and XGBoost—to examine the impact of racial bias in the MEPS dataset. Each model is evaluated using bias-related and traditional performance metrics like accuracy. The goal is to assess how these models handle biased data and the trade-offs between fairness and accuracy. The AIF360 framework is used throughout the study to measure and mitigate bias.

Model Training
The models were trained on Panel 19 of the MEPS data that was split into training, validation and test sets. A preprocessing pipeline with Standard Scaler and machine learning classifiers was applied, ensuring consistency in scaling and to take into consideration sample weights to represent the population structure. This helped to protect minority representation by providing proportional representation. The racially biased MEPS data was used to train Logistic Regression, Random Forest, Neural Networks, and XGBoost using this approach.

1. Logistic Regression: Logistic Regression is a binary classification model with a linear dependence of features, and the logarithm of odds of the outcome. The model was trained using the liblinear solver, which was specifically designed for small scale data sets.
2. Random Forest: Random Forest which is an ensemble method was developed with 500 decision trees, using sample weights to handle the issue of dataset imbalance. To minimize overfitting, at least 25 samples per leaf were used.

Random forests also model non-linearity and are less sensitive to bias than logistic regression and other models.

3. XGBoost: XGBoost, a gradient-boosting algorithm, was trained with default hyperparameters and sample weights to reduce bias. It is good at managing complicated relations, so it is excellent for working with structured data, but it also intensifies biases derived from the data set.
4. Neural Networks: A Multi-Layer Perceptron (MLP) with one layer of 100 neurons was trained for 300 epochs. The complexity of the model allows for more detailed representation of the data, but this feature can lead to the aggravation of the problem of bias if it is not adequately controlled.

After training, threshold tuning was performed to balance accuracy and fairness. Each model's probability estimates were converted into binary labels based on various thresholds. A threshold array from 0.01 to 0.5 was used, and metrics were evaluated at each threshold to determine the optimal balance between performance and fairness.

Evaluation Metrics
Each model's fairness and performance were evaluated using metrics from the AIF360 framework [29], offering a comprehensive view of their accuracy and fairness. Metrics such as Manhattan and Euclidean distances, as well as differential effect and equal opportunity distance, are also useful for evaluating both individual and group fairness by Broder1 and Berton1 [30]. Another metric suggested that instances from the disadvantaged group of the protected attribute with a positive label are assigned higher weights compared to those with a negative label, while instances from the favoured group with a positive label are given lower weights than those with a negative label by Kamiran and Calders [31]. These measures provide several viewpoints for assessing the fairness of machine learning models, which adds to a more thorough assessment of the ethical ramifications of these models.

1. Balanced Accuracy:
Balanced accuracy is calculated as the average of the true positive rate (TPR) and the true negative rate (TNR):

$$Balanced\ accuracy = \frac{TPR+TNR}{2} \quad \text{(i)}$$

2. Average Odds Difference (AOD):
AOD measures the difference in false positive and true positive rates between privileged and unprivileged groups. An ideal AOD value is 0, indicating equal treatment in prediction errors.

3. Disparate Impact (DI):
Disparate impact measures whether a model favours one group over another by comparing the ratio of favourable outcomes for unprivileged versus privileged groups. A value of 1 indicates perfect fairness, while deviations reflect bias.

$$Disparate\ Impact = \frac{Favorable\ Outcome\ Rate\ for\ Underprivileged\ Group}{Favorable\ Outcome\ Rate\ for\ Privileged\ Group} \quad \text{(ii)}$$

4. Equal Opportunity Difference (EOD):
EOD focuses on differences in true positive rates between groups, specifically measuring fairness in correctly identifying positive instances:

$$EOD = TPR\ unprivileged - TPR\ privileged \quad \text{(iii)}$$

5. Theil Index:
The Theil index is an entropy-based measure of inequality, commonly used to assess fairness across groups. Lower values indicate less inequality in the model's predictions.

All the four models were evaluated based on the thresholds using the computed metrics and the results show the best trade-off between fairness and performance. The best performing threshold was identified based on balanced accuracy, and other fairness metrics were then considered. Model performances were measured and compared from a set of metrics on the test set so that each model's handling of a data set bias with consideration to accuracy and fairness can be assessed. Four bias mitigation techniques were employed to address bias in the dataset: They include: Reweighing, Prejudice Remover, Adversarial Debiasing and ROC post processing. These techniques aim at bias at different levels of the ML pipeline; data preparation, during the model training, and during inference. These strategies were implemented through the AIF360 framework that provided a way through which the effects on fairness and accuracy could be judged.

Reweighing (Preprocessing)
The preprocessing technique used in this study is reweighing which aims at changing the weights of samples depending on the groups (privileged or unprivileged) in order to achieve a better balance in the demographic data during the training session. In this study, reweighing was used on the raw MEPS data set to address inherent racial prejudice. Initially, the dataset's Disparate Impact (DI) was 0.47554, indicating significant bias. After reweighing, the DI increased to 0.998, reflecting a near-perfect balance.
After reweighing, all four machine learning models, namely, Logistic Regression, Random Forest, Neural Network, and XGBoost, were retrained on the new weights dataset. The same measures of fairness and accuracy were used in the re-evaluation of their performance so that a like-for-like comparison could be made between the models before and after reweighing. It was observed that reweighing enhanced the fairness parameters of all the models under consideration while incurring a small degradation in the balanced accuracy.

Mathematical Explanation for Reweighing:
Reweighing assigns different weights to samples based on the following formula:

$$Weight = \frac{P(Class|Group)}{P(Group)} \quad \text{(iv)}$$

This ensures that underrepresented groups are given higher importance during training, thus mitigating bias.

Prejudice Remover (In-Processing)
The Prejudice Remover algorithm is an in-processing bias mitigation technique that adds a fairness constraint to the model's objective function through a regularization term. This is controlled by a hyperparameter, η (eta).
In this study, it was applied to the Logistic Regression model in order to determine its effect on the fairness. The sensitive attribute was encoded and the model trained on a scaled dataset with η=25 to focus on the reduction of bias.
The Prejudice Remover caused a positive shift in the values of the fairness measures: Statistical Parity Difference and Equal Opportunity Difference, while causing a marginal drop in cross-entropy loss, which confirms the utility of the proposed approach in terms of making the model fair during the training process.
The objective function of the model is modified to include a fairness term:

$$Objective\ Function = Loss\ Function + \eta \times Fairness\ Penalty \quad \text{(v)}$$

Adversarial Debiasing (In-Processing)
Adversarial Debiasing is an in-processing method that try to remove bias during the training of the model by using an adversarial approach. It has a classifier which predicts and an adversary which tries to predict the sensitive attribute (race) from those predictions. The classifier is trained to minimize prediction loss while making it difficult for the adversary to detect bias.

In this study, adversarial debiasing was applied to the Logistic Regression model. First, the model was trained with debiasing turned off (debias=False) to provide a benchmark, and the adversarial process was handled using a TensorFlow session.

Adversarial debiasing effectively reduced Average Odds Difference and Disparate Impact, promoting fairness during training. Still, as with any in-processing technique, it slightly affected the accuracy of the results overall.

ROC Postprocessing (Post-Processing)

Reject Option Classification Postprocessing is a method used after the training of a model to control the decision boundaries in order to produce more fair classifications based on group membership. This method gives favorable predictions to the client in the unprivileged group when the model is ambiguous.

In this study, ROC was applied to the validation set predictions of the Logistic Regression model to determine thresholds for fairness, with low and high classification thresholds set at 0.01 and 0.99.

ROC postprocessing is beneficial in scenarios where fairness cannot be attained in training or preprocessing, adjustments in decision-making can be done without retraining the model. However, it only modifies predictions and does not address inherent biases from the training process.

The Logistic Regression model served as the baseline, while only reweighing was applied to the other models (Random Forest, Neural Networks, and XGBoost).

Results showed that reweighing effectively reduced bias across all models, while Prejudice Remover and Adversarial Debiasing offered targeted fairness adjustments. ROC Postprocessing provided flexibility in predictions but was limited in addressing training biases.

Each technique has unique advantages and trade-offs, with their selection based on the specific context of the application, offering valuable insights for achieving fairness in biased datasets like MEPS.

Results and Discussion

This study investigated the effectiveness of bias mitigation methods on machine learning models of varying complexity using the biased MEPS dataset. It focused on how models (Logistic Regression, Random Forest, Neural Networks, and XGBoost) performed before and after applying reweighing. Key fairness and accuracy metrics evaluated included Disparate Impact (DI), Average Odds Difference (AOD), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and the Theil Index. Below are the results.

Disparate Impact: Before and After Reweighing

The original dataset had a Disparate Impact (DI) of 0.47554 signifying that the privileged group had been favoured. When the data was reweighed, the DI was 0.998, indicating that the groups are somewhat balanced. The above result indicates that reweighing indeed significantly reduces inherent bias and can be used as a preprocessing method for fairness enhancement.

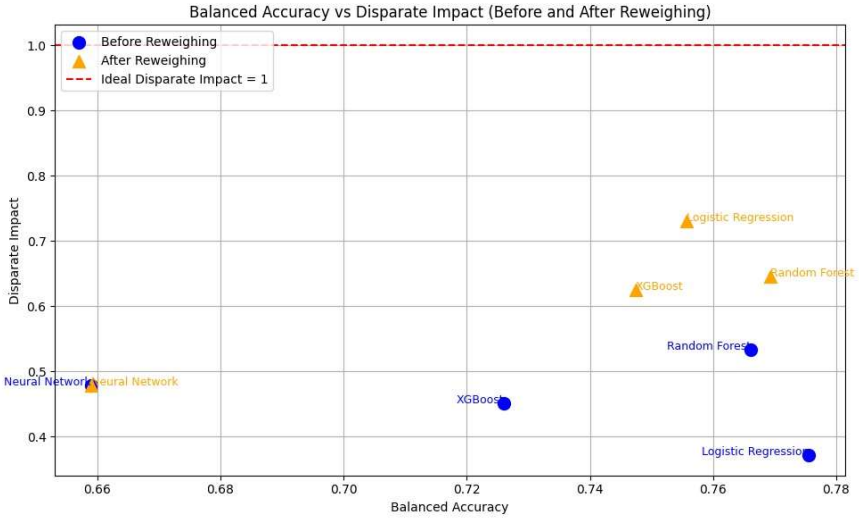

Fig. 1: Balanced Accuracy vs Disparate Impact (Before and After Reweighing) Model Performance: Accuracy vs. Fairness Trade-off

To demonstrate the trade-off between accuracy and fairness, balanced accuracy was calculated before and after reweighing. Simpler models like Logistic Regression and Random Forest consistently achieved higher accuracy than more complex models such as Neural Networks and XGBoost, both before and after reweighing.

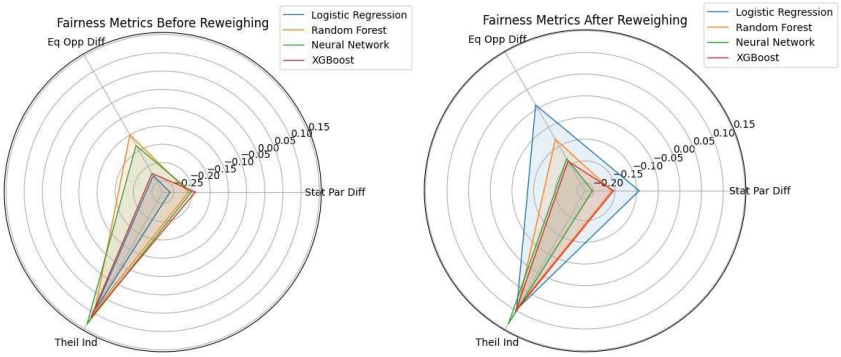

Fig. 2: Fairness Metrics (Before and After Reweighing)

- After reweighing, Logistic Regression had a lower balanced accuracy of 0.7556 as compared to 0.7756 before reweighing, but its fairness metrics like Average Odds Difference (AOD) and Disparate Impact (DI) were considerably enhanced, indicating that it benefited from reweighing at the cost of some accuracy.
- Random Forest had a slight improvement in balanced accuracy, 0.7661 to 0.7694, as well as significant improvements in the general fairness metrics. DI increased from 0.5335 to 0.6451 and AOD reduced from -0.1220 to -0.0812 which also confirms the bias reduction.
- XGBoost had a slight increase in balanced accuracy (from 0.7261 to 0.7474) and fairness metrics after reweighing. The DI improved from 0.4510 to 0.6250, and AOD decreased from -0.1651 to -0.1055.
- Neural Networks maintained a balanced accuracy of 0.6590 before and after reweighing, suggesting limited benefits from reweighing due to its complexity. The DI remained at 0.4789, with other fairness metrics, including AOD and Equal Opportunity Difference (EOD), remaining stable. These results suggest that reweighing is effective in increasing fairness across all models without a substantial loss of accuracy. However, the degree of improvement varies by model; simpler models like Logistic Regression exhibit a more pronounced reduction in bias, while more complex models like Neural Networks show less improvement in accuracy but some gains in fairness.

Hypothesis: Bias vs. Model Complexity

One of the main assumptions to be tested was that bias will grow with the model complexity. This was partially confirmed by evaluating models on the biased dataset before reweighing. Logistic Regression, the simplest model, had the highest Disparate Impact of 0.3717, followed by Random Forest (0.5335), XGBoost (0.4510), and Neural Networks (0.4789). These results imply that less complex models are more vulnerable to data bias because they are based on linear associations which only exacerbate differences.

However, complex models like Neural Networks and XGBoost, although demonstrating less bias initially, were harder to correct for fairness. This indicates that although complex models may start with less bias, they are less responsive to bias correction techniques like reweighing.

Average Odds Difference (AOD):

AOD, which measures the difference in error rates between privileged and unprivileged groups, improved across all models after reweighing. Logistic Regression's AOD decreased from -0.2024 to -0.0085, with similar reductions in Random Forest and XGBoost, demonstrating balancing of error rates.

Equal Opportunity Difference (EOD):

EOD, expressed as the true positive rate difference, demonstrated the highest change in the Logistic Regression (from -0.2284 to 0.0094), while Random Forest and XGBoost showed moderate gains. Neural Networks experienced minor variation.

Theil Index, which measures inequality of prediction distribution,, saw minimal changes across models. Logistic Regression's Theil Index was slightly higher after reweighing from 0.0958 to 0.0974 meaning that reweighing did not impact distributional fairness though it enhanced other measures.

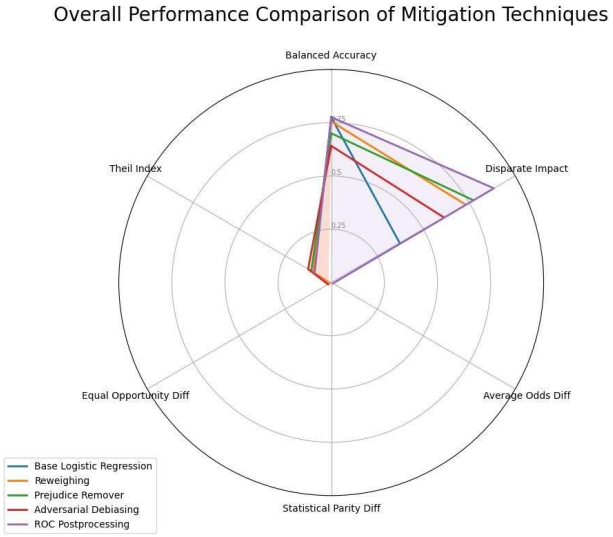

Fig. 3: Overall Performance Comparison of Mitigation Techniques

The results highlight the trade-offs between accuracy and fairness across machine learning models. Although, reweighing enhanced fairness measures for all models, the effect on balanced accuracy varied. For models such as Logistic Regression we can see that they incurred very little in terms of accuracy loss while improving fairness considerably. On the other hand, other models such as the Neural Networks responded poorly to the reweighing especially on the aspect of accuracy. Next, the research contrasts four bias mitigation techniques—Reweighing, Prejudice Remover, Adversarial Debiasing, and ROC Postprocessing—against the Logistic Regression model, which served as the baseline without bias mitigation. The objective is to assess the degree fairness and accuracy are met by each of the methods.

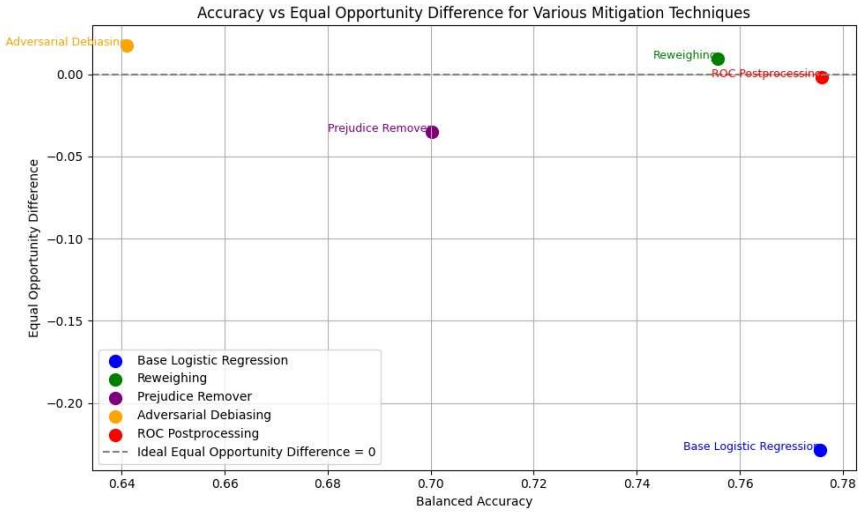

Fig. 4: Accuracy vs Equal Opportunity Difference for Various Mitigation Techniquess

Logistic Regression (No Mitigation)

Without bias mitigation, the Logistic Regression model achieved a balanced accuracy of 0.7756, with significant disparities in fairness. The Disparate Impact (DI) was 0.3717, indicating bias against unprivileged groups. The Average Odds Difference (AOD) was -0.2024, and the Statistical Parity Difference (SPD) was -0.2608, both of which showed the differences between the groups. The Equal Opportunity Difference (EOD) was -0.2284, due to disparate true positive rates. These results clearly imply that the baseline model has a highly negative bias towards the unprivileged population.

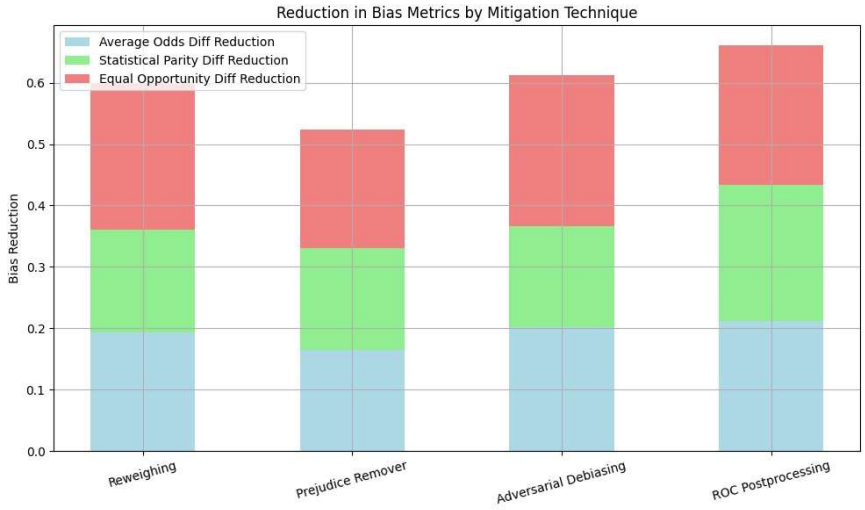

Fig. 5:Reduction in Bias Metrics by Mitigation Technique

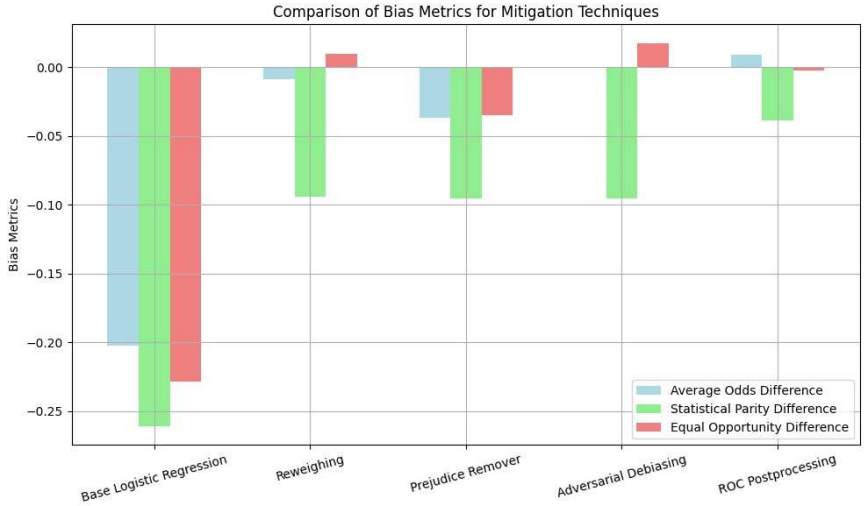

Fig. 6: Comparison of Bias Metrics for Mitigation Techniques

Reweighing (Preprocessing)

After applying Reweighing the Disparate Impact was enhanced from 0.3717 to 0.7306 which shows lesser prejudice. The Average Odds Difference was found to be -0.0085 while Equal Opportunity Difference came out to be 0.0094, which means that the group are almost at par. Although balanced accuracy reduced slightly from 0.7756 to 0.7556, the extent of fairness improvement outweigh this reduction. The Statistical Parity Difference also improved from -0.2608 to -0.0942. In total, three reweighing demonstrated the high ability to reduce bias while keeping the accuracy at a relatively high level.

Prejudice Remover (In-Processing)

The Prejudice Remover algorithm integrates fairness constraints into the model's objective function, leading to moderate improvements in fairness while experiencing a significant loss of accuracy. The balanced accuracy came down to 0.7002, reflecting the cost of prioritizing fairness. However, Disparate Impact increased to 0.7721 and Average Odds Difference was decreased to -0.0371. The Statistical Parity Difference was increased from -0.2608 to -0.0957, which also means that disparities in the favourable outcome were less profound. The major strength of the Prejudice Remover is that it operates in-processing, that is, fairness is incorporated during the training process. However, this comes with a trade-off, resulting in a compromise in accuracy.

Adversarial Debiasing (In-Processing)

Adversarial Debiasing introduced fairness through adversarial learning. While classification accuracy was high at 0.8217, balanced accuracy was lower at 0.6410, indicating imbalanced performance across classes. Disparate Impact increased to 0.6115 and the Average Odds Difference reduced to -0.0006. The Equal Opportunity
Difference increased to 0.0173, which means that both groups were treated equally.
Although Adversarial Debiasing has notable fairness gains, lower balanced accuracy implies a loss in performance regarding true positive and true negative rates.

ROC Postprocessing (Post-Processing)

ROC Postprocessing tweaked decision thresholds to improve fairness, which was accompanied by substantial improvements while maintaining a high level of accuracy. Balanced accuracy stayed at 0.7759, with Disparate Impact rose to 0.8838. The Average
Odds Difference declined to 0.0091 and Equal Opportunity Difference increased to 0.0021 showing that the models are now quite balanced in terms of true positives. Statistical Parity Difference enhanced to -0.0384, reflecting notable fairness improvements. A Theil Index slightly declined hence implying that the prediction distribution was more equally distributed. This method is effective when fairness cannot be fully achieved during training, allowing for flexible adjustments without retraining.

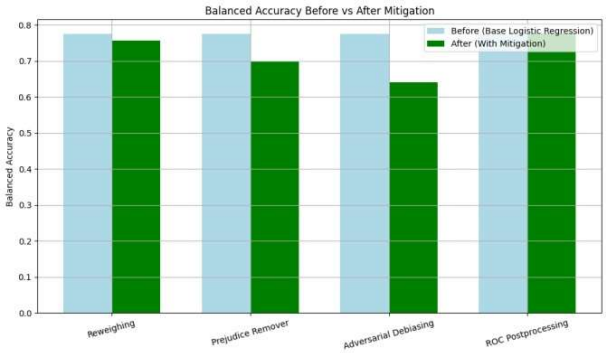

Fig. 7: Balanced Accuracy Before vs After Mitigation

Summary of Mitigation Techniques

Each bias mitigation method showed varying degrees of success in balancing the trade-off between fairness and accuracy: Reweighing offered substantial fairness improvements with minimal reduction in accuracy, making it an effective preprocessing method.
Prejudice Remover achieved fairness gains but with a more noticeable drop in accuracy, reflecting the trade-off inherent in in-processing methods.
Adversarial Debiasing significantly improved fairness, particularly Average Odds Difference and Equal Opportunity Difference, but at the cost of lower balanced accuracy.
ROC Postprocessing maintained high accuracy while offering meaningful improvements in fairness metrics, making it a robust post-processing technique.
Ultimately, the choice of bias mitigation technique depends on the specific application and the acceptable trade-offs between fairness and accuracy. Reweighing and ROC Postprocessing offer more balanced solutions, while Prejudice Remover and Adversarial Debiasing may be preferable when fairness is the primary concern.

Conclusion

In conclusion, this research examined the complex dynamics of machine learning biases, particularly how real-world data impacts model predictions. An analysis of the Medical Expenditure Panel Survey (MEPS) dataset revealed inherent racial biases, quantified using metrics such as Disparate Impact and Statistical Parity Difference. By employing four models—Logistic Regression, Random Forest, Neural Networks, and XGBoost—the study explored the relationship between model complexity, bias, and accuracy. Contrary to the initial hypothesis, findings showed that biases are not solely linked to model complexity, emphasizing the need to address biases at their source within the training data.
The evaluation of bias mitigation techniques, including Reweighing, Prejudice Remover, Adversarial Debiasing, and ROC Postprocessing, demonstrated varying impacts on fairness metrics and highlighted the trade-offs between bias reduction and accuracy. Reweighing was particularly effective, offering significant fairness improvements with minimal accuracy loss.

Future work should focus on exploring mitigation algorithms for neural networks of varying complexities, particularly for image-based data. Additionally, addressing data drift and developing novel bias mitigation strategies are essential. These efforts are vital for advancing ethical AI and ensuring fair decision-making in machine learning applications. Overall, this research aimed to illuminate the causes and impacts of biases in machine learning, paving the way for more ethical, transparent, and responsible practices in the field.

Conflict of Interest

We the authors hereby declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. We have not received any financial support or funding from any organization or entity that could influence the research findings or interpretations presented in this paper. There are no personal relationships or affiliations that could be perceived as having influenced the research reported in this paper. We confirm that no competing interests exist concerning this study, and all research was conducted independently and without bias to the best of my knowledge. The integrity and objectivity of this research were maintained throughout the study, and the conclusions drawn are solely based on the data and analysis presented.

References

1. Abràmoff, M.D., Tarver, M.E., Loyo-Berrios, N. *et al.* Considerations for addressing bias in artificial intelligence for health equity. *npj Digit. Med.* 6, 170 (2023). https://doi.org/10.1038/s41746-023-00913-9
2. Angwin, Julia, et al. “Machine Bias — ProPublica.” ProPublica, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminalsentencingAccessed 4 February 2024.
3. Using machine learning techniques for subjectivity analysis based on lexical and nonlexical features - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/BBC-dataset-statistics_tbl2_318225737 [accessed 5 Oct 2024]
4. Chatterjee, Swarn. (2013). Borrowing decisions of credit constrained consumers and the role of financial literacy. Economics Bulletin. 33. 179-191.
5. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science (New York, N.Y.)*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342. Accessed 7 February 2024.
6. Celi LA, Cellini J, Charpignon M-L, Dee EC, Dernoncourt F, Eber R, et al. (2022) Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. PLOS Digit Health 1(3): e0000022. https://doi.org/10.1371/journal.pdig.0000022. Accessed 7 February 2024.
7. Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. "The False Hope of Current Approaches to Explainable Artificial Intelligence in HealthCare." https://doi.org/10.1016/S2589-7500(21)00208-9. Accessed 7 February 2024.
8. Ueda, D., Kakinuma, T., Fujita, S. *et al.* Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol* 42, 3–15 (2024). https://doi.org/10.1007/s11604-023-01474-3
9. Dastin, Jeffrey. “Insight - Amazon scraps secret AI recruiting tool that showed bias against women.” Reuters, 10 October 2018, https://www.reuters.com/article/idUSKCN1MK0AG/ Accessed 4 February 2024.
10. Dankwa-Mullan I. Health Equity and Ethical Considerations in Using Artificial Intelligence in Public Health and Medicine. Prev Chronic Dis 2024;21:240245. DOI: http://dx.doi.org/10.5888/pcd21.240245
11. Bickel PJ, Hammel EA, O'connell JW. Sex bias in graduate admissions: data from berkeley. Science. 1975 Feb 7;187(4175):398-404. doi: 10.1126/science.187.4175.398. PMID: 17835295.
12. Elhanan Mishraky, Aviv Ben Arie, Yair Horesh, Shir Meir Lador, Bias detection by using name disparity tables across protected groups, Journal of Responsible Technology, Volume 9, 2022, 100020, ISSN 2666-6596,https://doi.org/10.1016/j.jrt.2021.100020. (https://www.sciencedirect.com/science/article/pii/S2666659621000135)
13. Fu, Enxian, et al. “A Comparative Study Between Traditional Algorithms and Machine Learning Algorithms in Predicting Recidivism.” A Comparative Study Between Traditional Algorithms and Machine Learning Algorithms in Predicting Recidivism, 2023, https://doi.org/10.1007/978- 981-99-6441-3_148. Accessed 7 February 2024.
14. Kubrin, C.E., Stewart, E.A.: Predicting who reoffends: The neglected role of neighbourhood context in recidivism studies*. Criminology 44(1), 165–197 (2006) DOI:10.1111/j.1745-9125.2006.00046.x
15. Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hofman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development 63, 4/5 (2019), 4–1. https://doi.org/10.48550/arXiv.1810.01943

16. Center for Data Science. “COMPAS Analysis using Aequitas.” COMPAS Analysis using Aequitas.,2018, https://dssg.github.io/aequitas/examples/compas_demo.html#Pre-Aequitas:Exploring-the-COMPAS-Dataset. Accessed 7 February 2024.
17. Niels Bantilan. 2018. Themis-ml: A fairness-aware machine learning interface for endto-end discrimination discovery and mitigation. Journal of Technology in Human Services 36, 1 (2018), 15–30. https://doi.org/10.48550/arXiv.1710.06921
18. Weerts, Hilde, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. "Fairlearn: Assessing and improving fairness of ai systems." Journal of Machine Learning Research 24, no. 257 (2023): 1-8. https://doi.org/10.48550/arXiv.2303.16626
19. Radovanovic´, Sandro, et al. “Do We Reach Desired Disparate Impact with In-Processing Fairness Techniques?, Procedia Computer Science.” Do we Reach Desired Disparate Impact with In- Processing Fairness Techniques? Procedia Computer Science, 2022,
20. Haytham Siala, Yichuan Wang, Shifting artificial intelligence to be responsible in healthcare: A systematic review, Social Science & Medicine, Volume 296, 2022, 114782, ISSN 0277-9536, https://doi.org/10.1016/j.socscimed.2022.114782.
21. Singh, Moninder, and Karthikeyan Natesan Ramamurthy. "Understanding racial bias in health using the Medical Expenditure Panel Survey data." *arXiv preprint arXiv:1911.01509* (2019).
22. Chao, YS., Wu, CJ. & Chen, TS. Risk adjustment and observation time: comparison between cross-sectional and 2-year panel data from the Medical Expenditure Panel Survey (MEPS). *Health Inf Sci Syst* 2, 5 (2014). https://doi.org/10.1186/2047-2501-25
23. Hamad R, Niedzwiecki MJ. The short-term effects of the earned income tax credit on health care expenditures among US adults. Health Serv Res. 2019 Dec;54(6):12951304. doi: 10.1111/1475-6773.13204. Epub 2019 Sep 30. PMID: 31566732; PMCID: PMC6863225.
24. Watanabe JH. Examining the Pharmacist Labor Supply in the United States: Increasing Medication Use, Aging Society, and Evolution of Pharmacy Practice. Pharmacy (Basel). 2019 Sep 19;7(3):137. doi: 10.3390/pharmacy7030137. PMID: 31546891; PMCID: PMC6789639.
25. Bounthavong M, Li M, Watanabe JH. An evaluation of health care expenditures in Crohn's disease using the United States Medical Expenditure Panel Survey from 2003 to 2013. Res Social Adm Pharm. 2017 May-Jun;13(3):530-538. doi: 10.1016/j.sapharm.2016.05.042. Epub 2016 May 20. PMID: 27263802.
26. Dressel, Julia, and Hany Farid. “The accuracy, fairness, and limits of predicting recidivism.” The accuracy, fairness, and limits of predicting recidivism, 2018, https://www.science.org/doi/full/10.1126/sciadv.aao5580. Accessed 7 February 2024. DOI:10.47611/jsrhs.v12i4.5779
27. Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, OakdenRayner L, Pfohl SR, Ghassemi M, McKay F, Treanor D, Rostamzadeh N, Mateen B, Gath J, Adebajo AO, Kuku S, Matin R, Heller K, Sapey E, Sebire NJ, Cole-Lewis H, Calvert M, Denniston A, Liu X. The value of standards for health datasets in artificial intelligence-based applications. Nat Med. 2023 Nov;29(11):2929-2938. doi: 10.1038/s41591-023-02608-w. Epub 2023 Oct 26. PMID: 37884627; PMCID: PMC10667100.
28. Medical Expenditure Panel Survey (MEPS). Content last reviewed July 2024. Agency for Healthcare Research and Quality, Rockville, MD. https://www.ahrq.gov/data/meps.html
29. IBM. “AI Fairness 360.” AI Fairness 360, 2018, https://aif360.res.ibm.com/. Accessed 7 February 2024.
30. Broder1, Renata Sendreti, and Lilian Berton1. “Performance analysis of machine learning algorithms trained on biased data.” Performance analysis of machine learning algorithms trained on biased data, 2021, https://sol.sbc.org.br/index.php/eniac/article/view/18283/18117. Accessed 7 February 2024.
31. Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1):1–33. https://doi.org/10.1007/s10115-011-0463-8