Ensuring Fair Play: An Exploration of Bias Mitigation Methodologies across the lifecycle of a Machine Learning Model

Srinivasaraghavan Sundar, Preethi Ramesh, Shruthi Ganesh, Sruthi Balu

Department of Statistics, Rutgers University.

Contributing authors: ss4805@scarletmail.rutgers.edu; pr597@scarletmail.rutgers.edu; sg2057@scarletmail.rutgers.edu; sb2538@scarletmail.rutgers.edu;

Abstract

The ubiquity in the utilisation of Machine Learning models to aid in making critical decisions across multiple sectors such as banking, insurance, hiring, and prison sentencing has placed a significant emphasis on ensuring fairness in AI models. In an era that has been defined by exponential acceleration of this new technology, it is important to anticipate and mitigate any potential bias that may arise in a given system. However, this has proven to be a pertinent problem, given that the biases that are found in the real world often permeate into the datasets used to train ML models. This in turn creates a situation where biases and prejudices prevalent in society are exacerbated by the algorithm.

This project aims to mitigate the biases in the data and the ML Model at various points in the development lifecycle of the model.

Bias is handled in the preprocessing stage by utilising several algorithms that modify the data prior to training the ML model. The methods that we have focused on are: Disparate Impact Removal[1]– which ensures that the selection process does not discriminate in terms of the outcomes obtained for different subgroups, Reweighing[2]– which attaches a weight for each data sample to reduce the bias in the dataset. We have also included a method to mitigate bias by Learning Fair Representations[3] of the dataset. In order to alleviate bias while training the ML model, Inprocessing fairness algorithms are used. These methods include: Prejudice Remover[4]– wherein a fairness aware regularisation term is used in tandem with any probabilistic machine learning model to reduce bias, Adversarial Debiasing[5]– where two models are trained to compete with one another to guarantee adversarially reweighed learning in the classification model. More advanced

concepts such as preventing fairness gerrymandering by developing Gerryfair classifier[6] has also been explored in this project. As for mitigating bias in an environment where the training data and the trained model cannot be modified, we have examined post processing bias mitigation strategies such as Equalized Odds Postprocessing[7], and Reject Option Classification[8]. In order to carry out our experiments, we have utilised several packages such as AIF360[9], tensorflow, keras, and scikit-learn on Python. We visualise our results by using plotly, matplotlib, and seaborn packages on Python. We examine the results obtained on the MEPS19 Dataset, a publicly available benchmark dataset. Finally, we also propose a novel Discrimination Aware Loss Function which aids in minimizing bias and maximizing fairness in Neural Network Models. We conclude with a detailed analysis of all methodologies.

Keywords: Fairness, Preprocessing, PostProcessing, Bias Mitigation, Disparity, Reweighing, Regularizer, Prejudice Remover, Adversarial Debiasing, Discrimination

1 Introduction

1.1 Relevance of Fairness

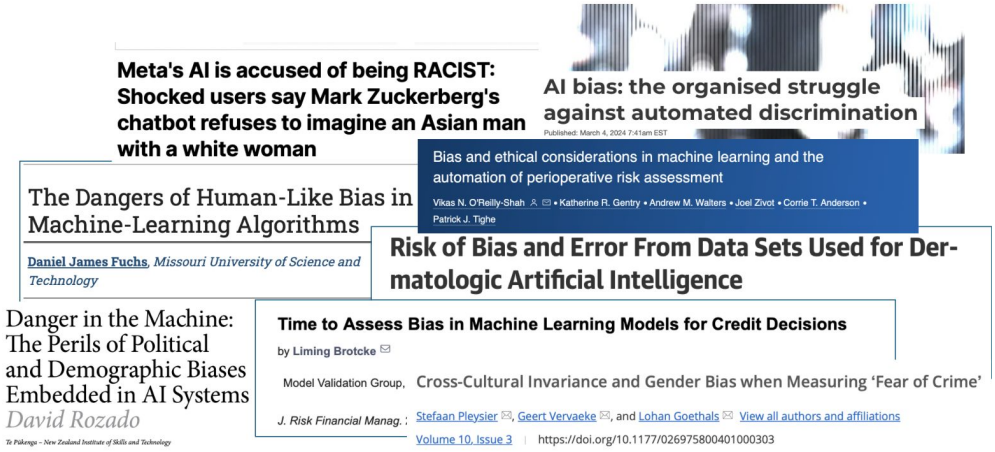

Fig. 1 Bias in Research

The rapid advancement of artificial intelligence (AI) has resulted in numerous benefits, but it also brings forth a range of prospective risks and complexities. Among the most pressing concerns is the detrimental impact of bias embedded within AI systems, which can negatively affect both individuals and society at large. AI bias, also known as machine learning or algorithm bias, describes the tendency of AI systems to

deliver results that replicate and sustain societal biases, both historical and current, often stemming from social inequalities. The irony lies in the fact that a system that was devised to eliminate bias is now exacerbating existing disparities, which in turn reinforces patterns of discrimination against marginalized groups and restricting their access to crucial services, opportunities, and resources. This can be viewed in Figure 1

1.2 Bias Mitigation in ML Pipeline

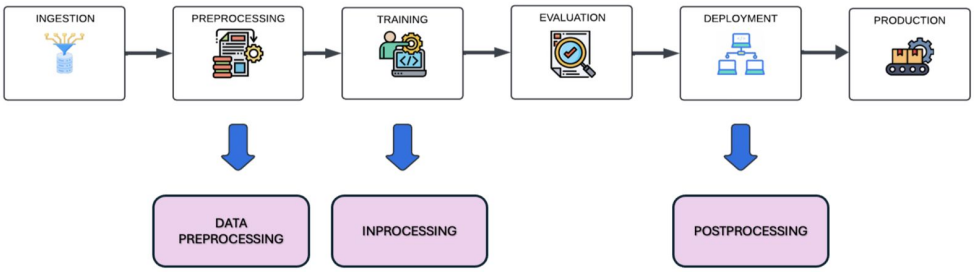

Fig. 2 Bias in Research

Figure 2 depicts where the bias may occur within the machine learning (ML) pipeline. Data collection and preparation can introduce bias through unrepresentative sampling, historical inequities, or poor data quality. Feature selection and engineering may unintentionally create biases if chosen features are closely linked to sensitive attributes. Algorithms can carry inherent biases if their design includes assumptions or optimization criteria that prioritize accuracy over fairness. During model training and evaluation, focusing solely on overall accuracy might mask biased performance across subgroups, particularly when test data lacks diversity. In deployment, user interactions often reflect their biases, which feedback loops may reinforce over time. These biases can be mitigated during Preprocessing, In-processing and Post-processing stages as depicted below.

In this research, bias has been tackled by implementing both preprocessing and in-processing techniques. Preprocessing techniques involve manipulating or augmenting the data before model training to minimize biases in the input features, thus improving data representativeness. In addition, a novel discrimination-aware loss function has been introduced during model training to specifically address and mitigate biases that arise within the algorithm itself, effectively reducing in-processing bias. By combining these strategies, the methodology ensures more equitable predictions while maintaining model performance and reliability.

2 Literature Review

Feldman et al. [1] delve into disparate impact, a legal concept aimed at addressing unintended discrimination when neutral procedures yield unequal outcomes across protected groups. They define disparate impact using the "80% rule" established by the U.S. Equal Employment Opportunity Commission (EEOC), which posits that a selection rate disparity above 80% may signal discrimination. The paper shifts focus from analyzing algorithms to assessing their outcomes, aligning with legal standards that emphasize the effects rather than the intentions of decision-making processes. It details a strategy to adjust datasets by altering non-protected attributes to minimize biases and comply with legal standards, without sacrificing decision-making accuracy. Tests on real-world datasets demonstrate the effectiveness of this method in identifying and mitigating disparate impacts while maintaining a balance between fairness and utility. The datasets, containing crucial demographic information for decisions like hiring and loan approvals, show that the method reduces the likelihood of legal challenges related to disparate impact. The paper also addresses the trade-offs between maintaining data utility and ensuring fairness, a key consideration in practical applications.

Kamiran, F., et al. [2] address the Discrimination-Aware Classification Problem, where training data containing unlawful discrimination, such as biases against sensitive attributes like gender or ethnicity, are used to develop a classifier. The focus of the study is on scenarios involving a single binary sensitive attribute within a binary classification framework. The research initially explores the theoretical balance between accuracy and non-discrimination in pure classifiers. Following this theoretical analysis, the paper discusses algorithmic approaches that preprocess the data to eradicate biases before the learning phase of the classifier begins. The preprocessing methods discussed include suppressing the sensitive attribute, altering the dataset by modifying class labels (massaging), and adjusting the data distribution through reweighing or resampling techniques without changing the labels. These methods were implemented using a customized version of the Weka software, and the paper presents experimental results using real-world data to demonstrate the efficacy of these preprocessing techniques in reducing discrimination. All approaches remove the discrimination from the training data and subsequently a classifier is learned on this unbiased data. Experimental evaluation shows that indeed these preprocessing approaches allow for removing discrimination from the dataset more efficiently than simple methods such as, e.g., removing the sensitive attribute from the training data. All methods have in common that to some extent accuracy must be traded-off for lowering the discrimination. This trade-off was studied and confirmed theoretically.

Zemel et al. [3] proposed an approach that frames fairness as an optimization problem, aiming to find an intermediate representation that effectively encodes data while concealing information about protected demographic groups. This dual objective is achieved by learning a set of "prototypes," creating a new space where data points lose sensitive attributes while retaining useful information. When compared to approaches like Fair Naive Bayes (FNB) and Regularized Logistic Regression (RLR), the Learned Fair Representations (LFR) model significantly reduces bias between protected and non-protected groups while maintaining strong predictive performance.

Kamishima, T., et al. [4] discuss three primary sources of bias in data analysis: prejudice, which involves the correlation between sensitive and other data quantified using mutual information; underestimation, referring to inaccuracies due to classifiers not fully converging; and negative legacy, which are biases stemming from unfair sampling or labeling in the training data. To tackle these issues, the authors propose a "prejudice remover regularizer" designed to maintain the classifier's independence from sensitive information. Results from logistic regression tests using this regularizer demonstrate its effectiveness in reducing indirect prejudice, outperforming standard methods.

Zhang, B.H., et al. [5] explore a framework to reduce biases in machine learning models by using adversarial learning techniques. The framework employs an adversarial network where the predictor model attempts to predict an outcome accurately and the adversary tries to predict the protected variable from the predictor's output. Both the models were trained using the Adam optimizer. The goal was to maximize the accuracy of the main model while minimizing the adversary's ability to predict the protected variable. The model was trained twice—once using a debiasing technique and once without it to compare the effect of the debiasing process on the model's ability to predict income brackets fairly across different gender groups.The results showed that while debiasing slightly reduced the model's accuracy, it made the model's errors more evenly distributed across gender groups, without introducing significant bias toward any group.

Kearns et al. [6] introduce a new family of fairness definitions that balance statistical and individual fairness. It tackles fairness gerrymandering by defining fairness over a comprehensive collection of subgroups, ensuring that machine learning models don't discriminate against specific subgroups while appearing fair across broad groups. The research proposes a zero-sum game framework where a learner minimizes error while an auditor identifies subgroups with fairness violations. Experiments on the Communities and Crime dataset demonstrate the framework's effectiveness in preventing discriminatory biases that are often overlooked by traditional fairness metrics.

Hardt, et al. [7] present a framework to reduce discrimination in machine learning models using "equalized odds" and "equal opportunity" criteria. Equalized odds ensures that true and false positive rates are consistent across protected and unprotected groups, while equal opportunity focuses on fairness in positive outcomes. The post-processing method adjusts any trained model to meet these criteria, minimizing bias without significantly compromising accuracy. Applying this approach to FICO credit scores reduced discrimination and maintained performance, incentivizing better data collection and feature selection for fair classification.

Kamiran,et al. [8] present two methods for enhancing fairness in machine learning: Reject Option-based Classification (ROC) and Discrimination-Aware Ensemble (DAE). ROC relabels low-confidence instances near the decision boundary to minimize discrimination, while DAE uses classifier disagreements to favorably label deprived groups. Both methods outperform previous fairness-aware models, as demonstrated through experiments on real-world datasets, reducing bias while maintaining accuracy. Moreover, they can handle multiple sensitive attributes, providing versatile solutions across various domains.

3 Dataset Description

The Medical Expenditure Panel Survey (MEPS) provides crucial national estimates on healthcare use, costs, payment methods, and insurance coverage for the noninstitutionalized U.S. population for the year 2015. The Household Component (HC) gathers data on health status, demographics, socioeconomic factors, employment, healthcare access, and satisfaction with care, offering insights into individuals, families, and specific subgroups using computer-assisted personal interviewing (CAPI). This dataset allows for long-term trend analysis since it's comparable to previous surveys from 1977 and 1987. With an annual sample of roughly 15,000 households, data is analyzed at individual or event levels and needs to be weighted to produce national estimates. Each MEPS HC panel is a subset of households from the prior year's National Health Interview Survey (NHIS), managed by the National Center for Health Statistics. The NHIS includes an oversampling of Black and Hispanic populations, and since 2006, Asian households have been included too. Connecting MEPS with NHIS provides extra data for long term analysis.

The MEPS dataset contains over 1,000 variables that capture extensive healthcare information related to patients. For the purposes of this project, a subset of columns relevant to the specific research objectives has been selected to ensure meaningful analysis and results. The data has been pruned and 44 columns have been retained. These include demographic information such as region, age, race, etc, and health indicators such as physical health and mental health scores, chronic diseases etc. Sex is a categorical variable here where "1" indicates male and "2" indicates female. The dataset contains two fields pertaining to race - one is a categorical variable that classifies the individual as either white or non-white and the other field describes the ethnicity of that individual. The dataset also contains the Physical Component Summary (PCS42) score, which is a standardized score derived from the SF-12 survey, reflecting an individual's overall physical health. Similarly, the Mental Component Summary (MCS42) score focuses on summarizing an individual's overall mental health. The K6SUM42 column, ranging from 0 to 24 represents the Kessler 6 score which measures the psychological distress a person has endured over the past 30 days. A new column called "utilization," representing how frequently patients accessed healthcare services is introduced. The total score is calculated by summing the values of office-based medical visits, outpatient visits, emergency room visits, inpatient nights and home health care used by an individual. The column categorizes the utilization as "1" if the score is greater than or equal to 10, otherwise "0". Any negative values in the dataset, which indicates null values, are dropped to ensure accurate analysis.

The bar graph in figure 3 and 4 shows the distribution of races in the MEPS dataset, grouped into two broad categories: "Non-White" and "White." The Non-White category (all racial groups except White) makes up the majority, with over 10,000 individuals and he White category represents a smaller portion, with a count of around 6,000. Exploring further based on ethnicity, it is observed that the majority of the population constitute of Caucasians, which is then followed by African-Americans and the other groups constitute less than 20% of the population.

A comparison of Physical Component Summary (PCS42) scores across multiple racial/ethnic groups has been visualized through a box plot in figure 5. Asian Indian

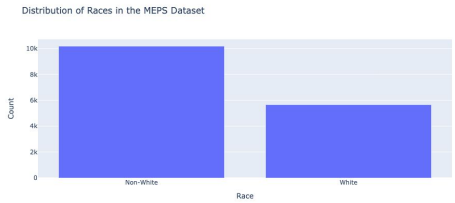

Fig. 3 Distribution of Races in MEPS (a)

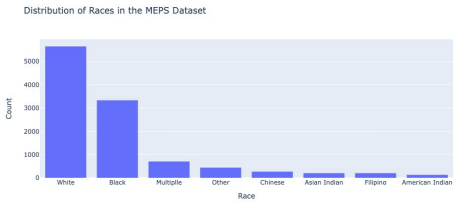

Fig. 4 Distribution of Races in MEPS (b)

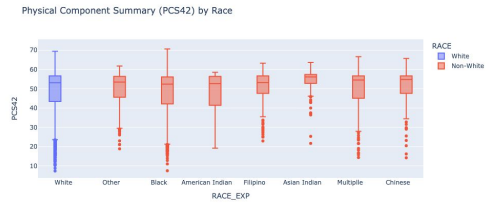

Fig. 5 PCS 42 By Race

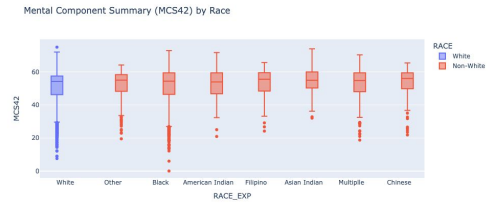

Fig. 6 MCS 42 By Race

and Chinese groups tend to show higher medians and narrower IQRs, indicating generally better and more consistent health scores. White, Black, and other groups show more variability, with a broader range and more outliers. Potential differences in mental health status across racial groups have also been visualized in figure 6. The White group displays a wider range of scores, indicating greater variability in mental health status. This range reflects more diverse experiences or disparities in mental health within the White population. Asian Indian and Chinese Groups have higher median scores, suggesting relatively better average mental health compared to others. The presence of low outliers (scores below 20) across different groups highlights individuals struggling with significant mental health challenges. The graphs given in Figure 7 and 8 depict healthcare utilization by race and the color coding distinguishes between those who have more than 10 visits (red) versus those who don't (blue). One can observe that 25.5% of White people make the most effective use of the available healthcare services compared to the rest of the White population. But when we compare it with the healthcare utilization of the Non-White group, only 12.5% avail these services, which indicates a huge disparity across these groups.

Fig. 7 Percentage of Healthcare Utilization by Race

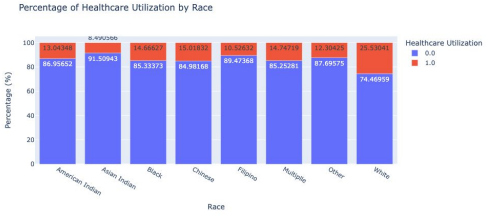

Fig. 8 Percentage of Healthcare Utilization by Race Expanded Values

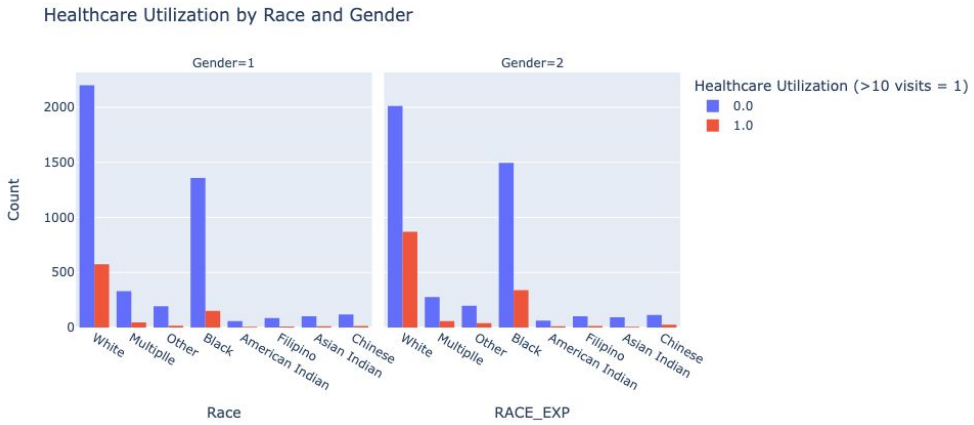

Fig. 9 Healthcare Utilization By Race and Gender

The utilization of these healthcare services by each gender across different racial groups has also been illustrated in figure 9. As inferred from the previous graphs, White individuals tend to have higher counts of healthcare visits compared to other racial groups and this pattern is consistent across both genders.

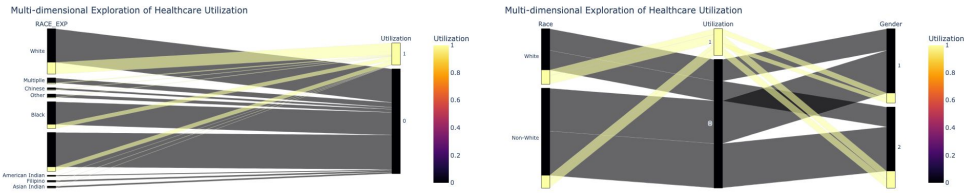

Fig. 10 Multidimensional Utilization of Resources

The plot in Figure 10 portrays healthcare utilization based on race. There is a noticeable disparity in utilization between the White and Non-White racial groups, with the Non-White group having overall lower utilization. Among the non-White groups, the Multiple/Chinese/Other category seems to have the next highest utilization levels. The African-American group has relatively moderate utilization, whereas the American Indian, Filipino, and Asian Indian groups appear to have the lowest utilization rates overall. The plot in Figure 10 portrays the multi-dimensional exploration of healthcare utilization based on race and gender. This 3D plot explores healthcare utilization based on race (x-axis) and gender (y-axis). It is evident that White males utilize healthcare services the most when compared to White females. When comparing the usage between the White and Non-White groups, utilization of these services is very low across Non-White groups, especially by the Non-White females. The differences observed in healthcare utilization as in Figure 11 suggest potential inequalities which arise due to systemic issues like health care accessibility or differences in how various groups perceive and prioritize healthcare services.

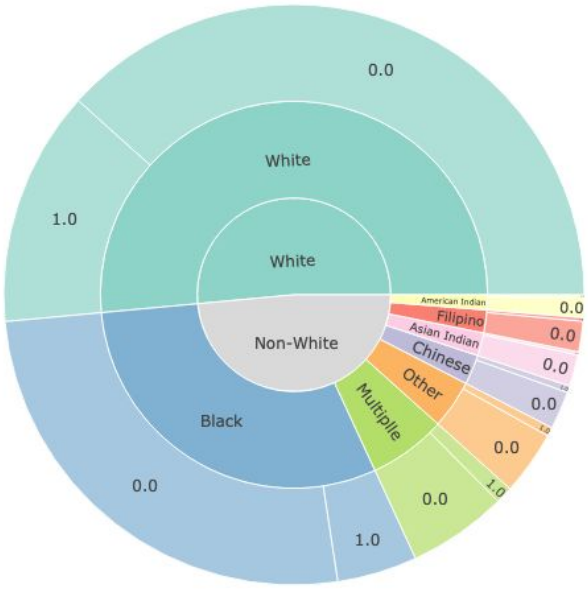

Fig. 11 Healthcare Utilization By Race

4 Methodology

4.1 Bias Detection in Base LR Model

Here we manually identify and specify the sensitive attribute, which is crucial for analyzing the bias in our model. Using the sensitive attributes, we define the privileged and unprivileged groups in the dataset. We split the data into an 80% training set and a 20% test set to train the model. By maneuvering over the Scikit-Learn libraries, we have built a Logistic Regression model using the training dataset with an L2 penalty and the 'lbfgs' solver to optimize the classification. After predicting on our test data using the model that we have built, we have implemented 5 Statistic Test to Detect bias in our Dataset:

Statistical Parity Difference: This metric measures the difference in the probability of positive outcomes between the privileged and unprivileged groups. A value of zero indicates perfect fairness. It is used to assess whether both groups have equal opportunities for receiving positive predictions, irrespective of their privileged or unprivileged status.

Equal Opportunity Difference: This metric evaluates the difference in true positive rates between the unprivileged and privileged groups. A lower value closer to zero suggests less bias. Equal Opportunity Difference specifically focuses on the fairness of outcomes for those individuals who should receive a positive outcome, ensuring that one group is not favored over another.

Average Absolute Odds Difference: This metric averages the absolute differences in false positive rates and true positive rates between the privileged and unprivileged groups. It provides a more comprehensive view of both type I and type II errors, offering a broader assessment of how fair the model is across different decision thresholds.

Disparate Impact: Disparate Impact measures the ratio of positive outcomes received by the unprivileged group to that of the privileged group. A value of 1 indicates no disparate impact. It is useful for determining whether a model's predictions are disproportionately favoring one group over another, which is crucial for complying with legal standards in many regions.

Theil Index: The Theil Index is a statistical measure of inequality, adapted to measure inequality in model predictions. A value of zero represents perfect equality. It quantifies the disparity in model outcomes across individuals, highlighting scenarios where the model may be overly favorable or unfavorable to specific groups.

4.2 Bias Mitigation Strategies

4.2.1 Reweighing

Reweighing is a preprocessing method designed to reduce bias. This approach corrects historical injustices by assigning more significance to appropriate instances and less weight to erroneous (i.e., discriminatory) ones. Reweighing assumes that a fair dataset would exhibit no conditional dependence of the outcome on a protected attribute. Utilizing the sensitive attribute 'Race', we identified privileged and unprivileged groups and applied new weights using the Reweighing method from the AIF360 Toolkit.

The transformed data was loaded into a dataframe and incorporated into a Logistic Regression model, where instance weights were used as sample weights to mitigate bias. After predicting on our test data using the model built with reweighed data, we performed five statistical tests to determine whether bias had been mitigated in our dataset.

4.2.2 Disparate Impact Remover

Disparate Impact Remover (DIR) is another preprocessing method designed to reduce bias. When implementing the technique and some contrivance, we understand that this technique adjusts feature values to balance distributions across groups, in contrast to re weighing, which changes the weights of data points to balance their influence in model training without altering the data itself. Utilizing the sensitive attribute 'Race', we identified privileged and unprivileged groups and applied the Disparate Impact Remover method from the AIF360 Toolkit. The transformed data is loaded into a dataframe and fit into a Logistic Regression model, where instance weights are used as sample weights to mitigate bias. After predicting on our test data using the model built with the modified data, we performed five statistical tests to determine whether bias was mitigated in our dataset.

4.2.3 Prejudice Remover

Prejudice Remover is an in-processing method designed to mitigate bias directly during the model training phase.The technique incorporates a fairness oriented regularization term based on the sensitive attribute 'Race'. This regularization term penalizes the learning algorithm when it detects dependency of the predictions on this sensitive attribute. The strength of this regularization is controlled by a parameter , which can be tuned based on the desired level of bias mitigation. We are implementing this using the AIF360 Toolkit. After training, the model is evaluated on test data using statistical tests to confirm whether bias has been effectively mitigated. This technique, like Reweighing and the Disparate Impact Remover, aims to ensure more equitable outcomes in predictive modeling.

4.2.4 Adversarial Debiasing

Adversarial Debiasing is an in-processing technique that seeks to reduce bias by incorporating adversarial learning. This method leverages a neural network that predicts the outcome while simultaneously minimizing the ability for an adversary to determine the sensitive attribute (Here, 'Race') from the predictions. Utilizing TensorFlow, we first identified privileged and unprivileged groups, then trained an Adversarial Debiasing model within a dedicated session. The model was trained for 100 epochs with the aim of debiasing, using the AIF360 Toolkit. After training, the model was evaluated on test data and subjected to five statistical tests to verify the effectiveness of the bias mitigation.

4.2.5 The Discrimination-Aware Loss Function

The Discrimination-Aware Loss Function is an in-processing method designed by us to directly address bias during the training phase of a machine learning model. This technique involves modifying the traditional loss function to include a term that accounts for discrimination based on sensitive attributes, such as 'Race'. The aim is to penalize the model more heavily when discrepancies in outcomes between privileged and unprivileged groups are detected, thereby encouraging the model to learn fairer representations.

Model Setup: A binary classification neural network model is defined with two linear layers. The input passes through a ReLU activation function in the first layer and a sigmoid activation in the output layer to predict binary outcomes.

Custom Loss Function: The loss function combines Binary Cross-Entropy (BCE) with a discrimination term. The BCE computes the model's accuracy in predicting binary outcomes. The discrimination term is calculated based on the difference in probabilities of positive outcomes between the privileged and unprivileged groups. This term is raised to the power of k (a hyperparameter) and multiplied by lambda (another hyperparameter), which controls the strength of regularization against discrimination.

$$\lambda \left[\bar{P}_{\text{Priv}} - \bar{P}_{\text{unPriv}} \right]^{k}$$

Fig. 12 Formula for Computing Discrimination

$$\text{Loss} = (1 + D) \cdot \text{BCE}$$

Fig. 13 Formula for Computing Discrimination Aware Loss

Training Process: During training, the model weights are updated by minimizing the custom loss function. The function penalizes the model more as the discrimination between groups increases. Sensitive features are identified (e.g., using a threshold on a particular attribute), and these features are used to calculate the discrimination term in the loss function.

Evaluation and Adjustment: Adjustments to the hyperparameters lambda and k were made on the discrimination and observed. After training, the model was evaluated on test data and subjected to five statistical tests to verify the effectiveness of the bias mitigation.

5 Results

5.1 Comparison of Existing Methodologies

The proposed methodology in this project to mitigate bias has been implemented and its results have been evaluated against the preexisting methods. To compare the efficacy of the suggested methodology and the preexisting methods, a logistic regression model without any bias mitigating algorithms has been formulated to predict the utilization of healthcare services by individuals.

5.1.1 ROC Curves And Confusion Matrices

A Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system by plotting the True Positive Rate (sensitivity) against the False Positive Rate at various threshold settings. It is used to assess a model's ability to distinguish between classes, with the Area Under the Curve (AUC) providing a single metric to summarize overall performance. The x-axis represents the False Positive Rate (FPR) and the y-axis shows the True Positive Rate (TPR), or Recall. The dashed diagonal line represents the performance of a random classifier (a model with no discriminative power).

Baseline LR Model The baseline logistic regression model exhibits an AUC of 0.86, as depicted in the ROC curve analysis. This indicates a strong performance with a high True Positive Rate (TPR) achieving 0.8 at a False Positive Rate (FPR) of approximately 0.1. The model's robust predictive ability sets a high benchmark for the subsequent debiasing techniques. The confusion matrix for the baseline model predicting healthcare utilization shows 3777 true negatives (TN), 174 false positives (FP), 340 true positives (TP), and 458 false negatives (FN). With an accuracy of 88.67%, the model has high specificity in identifying non-utilizers. However, the high false negative count suggests underprediction of healthcare utilization, which could result in biases impacting specific groups and leading to disparities in care access.

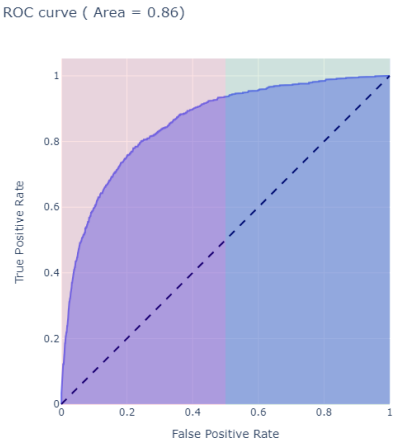

Fig. 14 ROC for Base Model

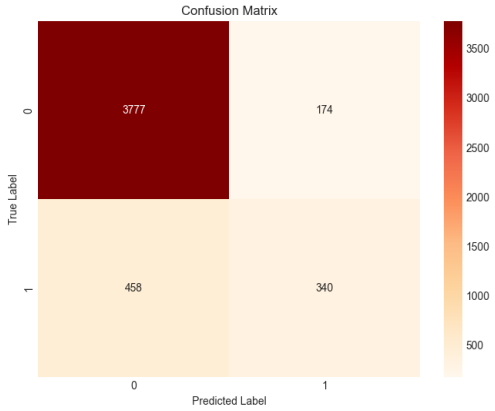

Fig. 15 Confusion Matrix for Base Model

Reweighing The ROC curve for the reweighing method presents an AUC of 0.84. This performance, slightly below the baseline, suggests a moderate trade-off between fairness and accuracy. The curve's progression indicates that while the method effectively reduces bias by adjusting class weights to correct underrepresented groups in the training data, it does so with a slight reduction in the ability to classify instances correctly. The TPR and FPR dynamics observed suggest that the model, despite being slightly less accurate, maintains a commendable balance between sensitivity and specificity. The confusion matrix for the reweighing method used in predicting bias in healthcare utilization shows True Negatives (TN) at 3735, False Positives (FP) at 216, False Negatives (FN) at 436, and True Positives (TP) at 362. The model demonstrates high specificity (94.52%) but lower recall (45.36%), indicating a bias towards predicting non-utilization. This underestimation of healthcare utilization among the minority class could negatively impact the equitable distribution of healthcare resources. The significant disparity between recall and specificity underscores a potential imbalance, highlighting the need for further refinement to enhance the model's fairness and accuracy.

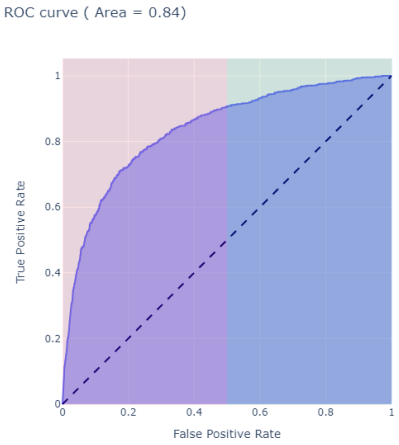

Fig. 16 ROC for Reweighing

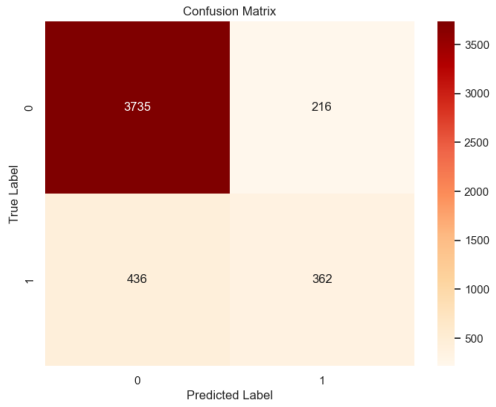

Fig. 17 Confusion Matrix for Reweighing Model

Disparate Impact Remover The ROC curve for the Disparate Impact Remover shows an AUC of 0.86. This indicates that the method effectively maintains the predictive performance of the baseline model while potentially reducing bias. The curve is closely aligned with that of the baseline model, suggesting that this debiasing approach manages to address disparities without compromising on the model's ability to differentiate between users with high and low healthcare utilization effectively. The confusion matrix for the disparate impact remover in predicting bias in healthcare utilization shows True Negatives (TN) at 3787, False Positives (FP) at 164, False Negatives (FN) at 459, and True Positives (TP) at 339. The model's high specificity (95.85%) indicates effective identification of individuals who do not utilize healthcare, but the low recall (42.48%) highlights a significant miss in detecting true positives. This underestimation could suggest a bias against predicting healthcare utilization for certain groups, leading to potentially inadequate resource allocation.

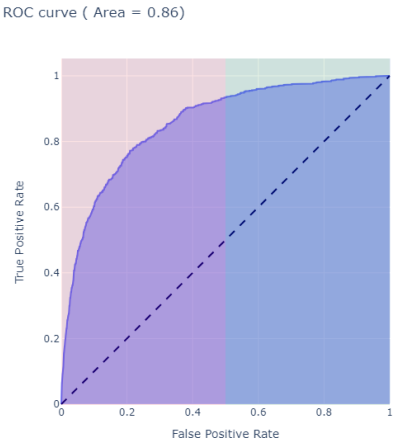

Fig. 18 ROC for DIR

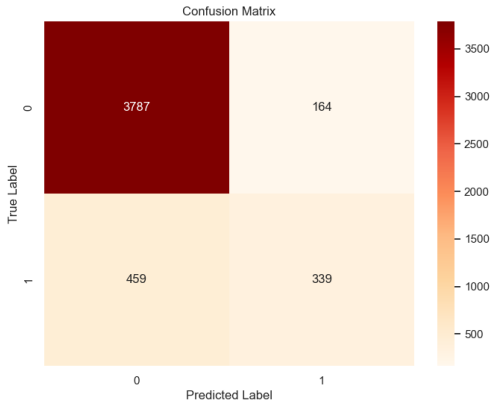

Fig. 19 Confusion Matrix for DIR

Prejudice Remover This method's performance, while slightly lagging behind with an AUC of 0.84, may offer advantages in scenarios where reducing the prejudice in predictions is crucial, even at the cost of a minor drop in accuracy. The method's approach to adjusting the decision boundary to minimize prejudice directly within the algorithm could be particularly beneficial in sensitive applications like healthcare. The confusion matrix for the Prejudice Remover technique in predicting healthcare utilization shows a high specificity of 98.25%, with 3882 true negatives and only 69 false positives, effectively minimizing unnecessary resource allocation. However, the model suffers from a significant false negative rate, as it correctly identifies only 163 true positives while missing 635 cases, resulting in a low recall of 20.43%. This suggests a potential bias against predicting healthcare utilization, which could lead to inequitable access to essential services.

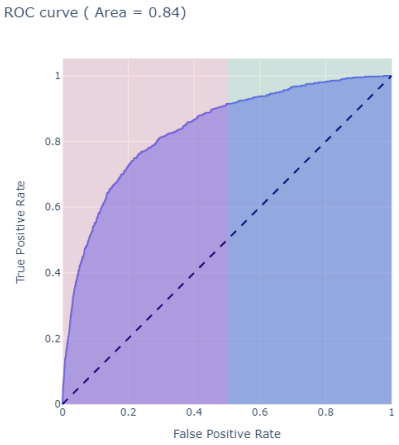

Fig. 20 ROC for Prejudice Remover

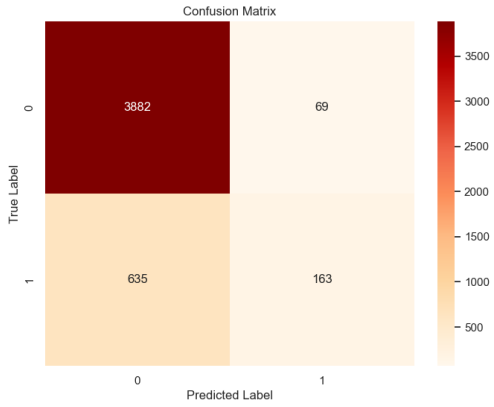

Fig. 21 Confusion Matrix for Prejudice Remover

Adversarial Debiasing The ROC curve for the Adversarial Debiasing method exhibits an AUC of 0.85. This performance is nearly on par with the baseline model, suggesting that the method effectively addresses biases without substantial loss to accuracy. The confusion matrix for the Adversarial Debiasing technique shows 3805 true negatives (TN), 146 false positives (FP), 286 true positives (TP), and 512 false negatives (FN). With this configuration, the model achieves high specificity but relatively low recall due to a significant number of false negatives. This results in an accuracy of about 85.22% . The low recall indicates potential bias against identifying healthcare utilization needs, which could result in disparities affecting resource allocation.

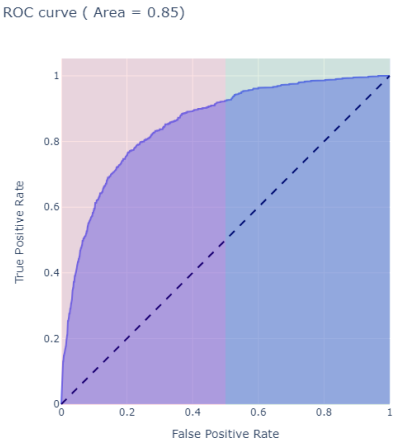

Fig. 22 ROC for Adversarial Debiasing

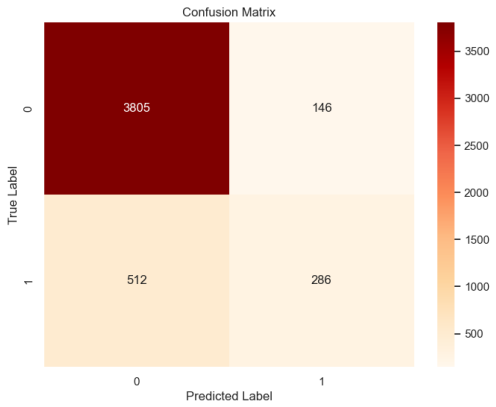

Fig. 23 Confusion Matrix for Adversarial Debiasing

5.2 Performance of our Novel Discrimination Aware Loss Function

In this part of the study, we compared two deep learning models to evaluate the impact of incorporating a discrimination-aware loss function on the model's performance metrics: accuracy, loss, discrimination, and disparity in fairness. Model A was trained with a standard binary cross-entropy loss function, while Model B included a discrimination-aware loss function aimed at reducing discriminatory outcomes across sensitive attributes.

5.2.1 Accuracy and Loss

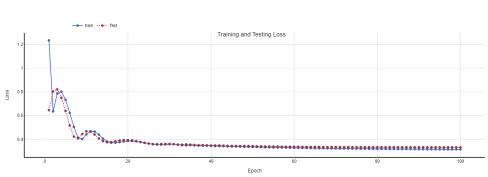

Fig. 24 Loss For Discrimination Unaware Model

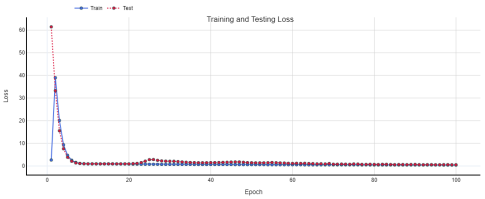

Fig. 25 Loss For Discrimination Aware Model

Both models demonstrated significant improvements in accuracy over 100 epochs. Model A achieved a stable test accuracy close to 90% after an initial fluctuation, as illustrated in the "Training and Testing Accuracy" graph. Model B showed a similar

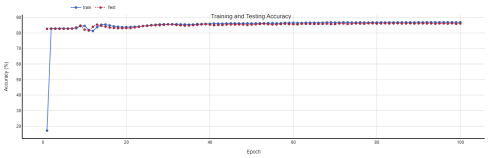

Fig. 26 Accuracy For Discrimination Unaware Model

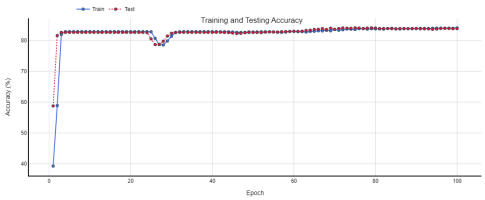

Fig. 27 Accuracy For Discrimination Aware Model

trajectory, reaching stability slightly earlier than Model A and maintaining an approximate 88-90% accuracy. This suggests that the inclusion of a discrimination-aware component did not compromise the overall accuracy of the model.

In terms of loss, both models showed a sharp decrease in the initial epochs, with Model B exhibiting a slightly faster reduction in loss, as shown in the "Training and Testing Loss" graph. This rapid decrease underscores the effectiveness of the discrimination-aware adjustments in Model B, leading to quicker optimization convergence compared to the standard loss function used in Model A.

5.2.2 Discrimination and Disparity in Fairness

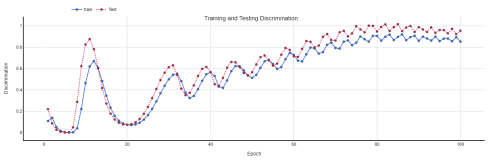

Fig. 28 Discrimination For Discrimination Unaware Model

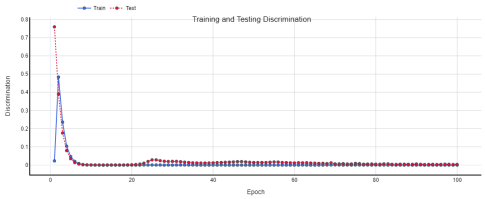

Fig. 29 Discrimination For Discrimination Aware Model

The discrimination metrics, a critical aspect of this study, were significantly different between the two models. Model A's discrimination values decreased gradually, stabilizing at around 0.1 after the initial epochs. In contrast, Model B, with the discrimination-aware loss function, rapidly approached zero discrimination, maintaining this level throughout the testing phase, as depicted in the "Training and Testing Discrimination" graph. This underscores the effectiveness of the discrimination-aware loss function in minimizing discriminatory outcomes.

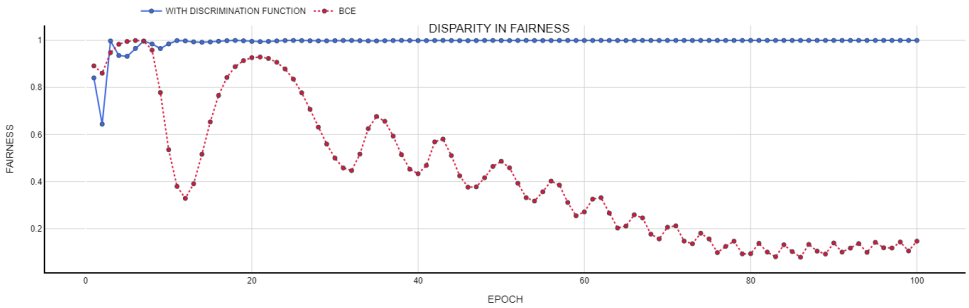

Fig. 30 Disparity in Fairness Between the Two Models

Disparity in fairness, measured as the difference in performance between subgroups within the dataset, was notably lower in Model B throughout the training process. As shown in the "Disparity in Fairness" graph, Model B maintained a near-zero disparity level after initial training epochs, highlighting its robustness in treating different groups equitably. The comparative analysis between the two models clearly illustrates the benefits of integrating a discrimination-aware loss function in deep learning models, particularly in applications where fairness is a crucial concern. While both models achieved high levels of accuracy, Model B demonstrated superior performance in reducing discrimination and disparity in fairness without compromising on other key performance metrics. This suggests that adopting discrimination-aware models can be a significant step towards ethical AI practices, particularly in sensitive applications such as healthcare, finance, and public services.

5.3 Fairness Plots For All Methods

5.3.1 Baseline Model

The analysis of fairness metrics before bias preprocessing reveals that the model exhibits certain biases across racial lines. A Statistical Parity Difference of -0.12 and an Equal Opportunity Difference of -0.20 indicate underrepresentation and unequal true positive rates for certain racial groups, suggesting mild to moderate discriminatory outcomes. The Average Absolute Odds Difference at 0.12 points to slight inconsistencies in error rates between groups, while a Disparate Impact value of 0.35 highlights a significant bias against one group in receiving favorable outcomes. The Theil Index at 0.11 shows some level of inequality, though not excessively high, suggesting that while there is some bias, the overall inequality is moderate. These metrics collectively underscore the need for specific bias mitigation strategies to enhance fairness in the model's predictions across different racial demographics.

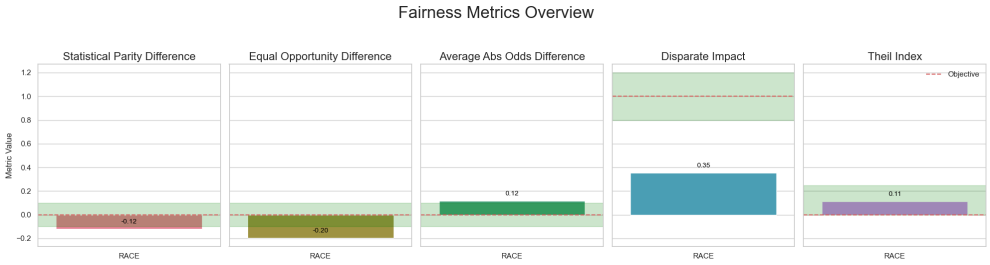

Fig. 31 Fairness Metrics in Baseline Model

5.3.2 Reweighing Model

The fairness metrics after applying reweighing techniques show significant improvements towards achieving equity across racial lines. The Statistical Parity Difference is exactly 0.00, indicating no disparity in the rate of favorable outcomes across different racial groups. The Equal Opportunity Difference is near zero at 0.01, demonstrating almost equal true positive rates between groups, which is a strong indicator of non-discriminatory behavior. The Average Absolute Odds Difference also stands at a minimal 0.01, showing that both false positive and false negative rates are nearly identical across races. Disparate Impact has achieved a value of 1.01, which is ideal, indicating no disproportionate impact on any group. However, the Theil Index remains at 0.11, suggesting there is still a slight inequality in how predictions are distributed across different groups, although this is relatively minor. Reweighing has showed us the best metrics so far.

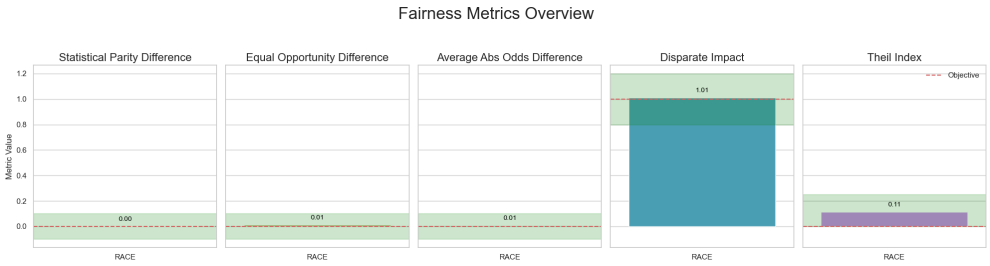

Fig. 32 Fairness Metrics in Reweighing Model

5.3.3 Disparate Impact Remover Model

The fairness metrics for the model using the disparate impact remover technique reveal significant biases. The Statistical Parity Difference of -0.12 indicates a slight bias against certain racial groups in receiving favorable outcomes. A more pronounced Equal Opportunity Difference of -0.21 highlights notable disparities in correctly predicting true positives for different races, suggesting unequal treatment. The Average

Absolute Odds Difference at 0.12 reflects small disparities in error rates across groups. A Disparate Impact value of 0.33 suggests considerable disproportionate impact, significantly below the ideal of 1.0, indicating that one group is far less likely to receive positive predictions. Despite these issues, the Theil Index at 0.12 suggests that overall inequality in prediction distribution is modest, but the model still requires adjustments for greater equity.

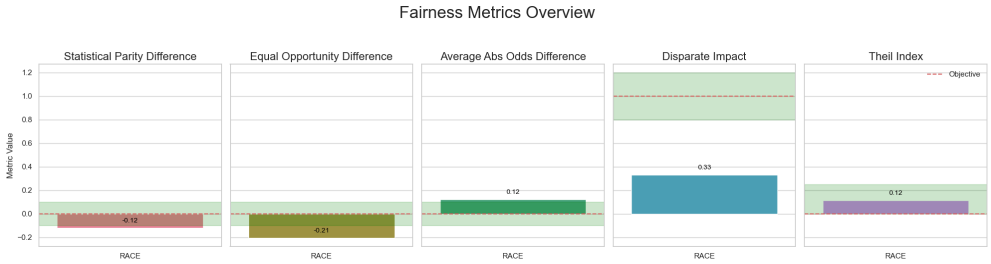

Fig. 33 Fairness Metrics in DIR Model

5.3.4 Advarserial Debiasing Model

The fairness metrics from the model using adversarial debiasing indicate modest improvements but still reflect underlying biases. The Statistical Parity Difference is -0.05, showing a slight underrepresentation of certain racial groups in favorable outcomes. The Equal Opportunity Difference at -0.04 further confirms a slight bias in accurately predicting true positives for different racial groups. The Average Absolute Odds Difference at 0.02 suggests minimal disparity in error rates across groups, reflecting a relatively fair treatment. However, the Disparate Impact value of 0.55 points to a considerable imbalance, indicating that one group is significantly less likely to receive positive outcomes than others. The Theil Index at 0.13, although low, still indicates a presence of inequality in the model's predictions across different groups, highlighting areas where further bias mitigation is necessary.

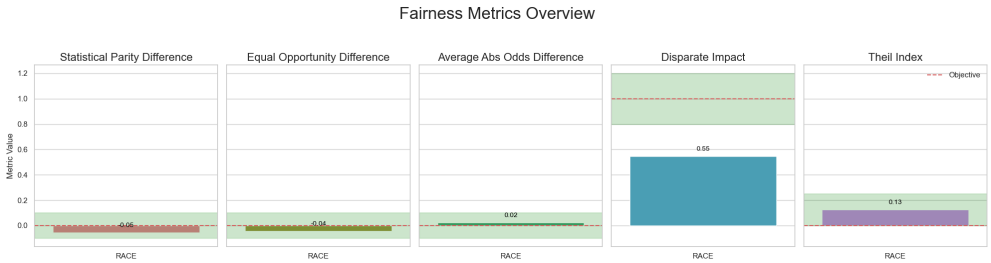

Fig. 34 Fairness Metrics in Advarserial Debiasing Model

5.3.5 Prejudice Remover Model

The fairness metrics from the model using the prejudice remover technique indicate a more balanced but still slightly biased approach. The Statistical Parity Difference of -0.02 and the Equal Opportunity Difference of 0.01 both show minimal deviation from zero, suggesting almost no bias in the rate of favorable outcomes or in the true positive rates across different racial groups. The Average Absolute Odds Difference at 0.01 further supports this near-uniform fairness in terms of both false positives and false negatives. However, a Disparate Impact value of 0.60 indicates that the model still disproportionately favors one group over another, with ideal parity being 1.0. The Theil Index at 0.15, while low, points to some residual inequality in prediction distribution among groups, highlighting an area for potential improvement.

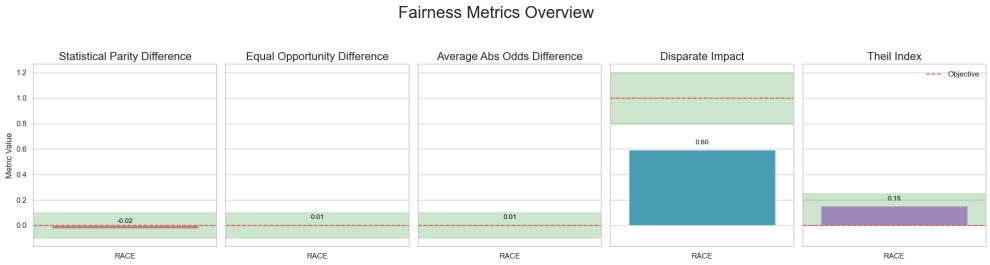

Fig. 35 Fairness Metrics in Prejudice Remover Model

5.3.6 Model Trained with novel Discrimination-Aware Loss Function

The fairness metrics from the model utilizing a Discrimination-Aware Loss Function show exemplary results, indicating an optimal level of fairness across all evaluated dimensions. The Statistical Parity Difference, Equal Opportunity Difference, and Average Absolute Odds Difference all register at 0.00, demonstrating no bias in rates of favorable outcomes, true positive rates, or error rates across different racial groups. The Disparate Impact metric achieves a perfect score of 1.00, signifying no disproportionate impact on any racial group. Additionally, the Theil Index also reports a score of 0.00, indicating absolute equality in the distribution of predictions among groups. These metrics collectively suggest that the model achieves ideal fairness, with no observable bias or inequality in treatment or outcome distribution across racial demographics.

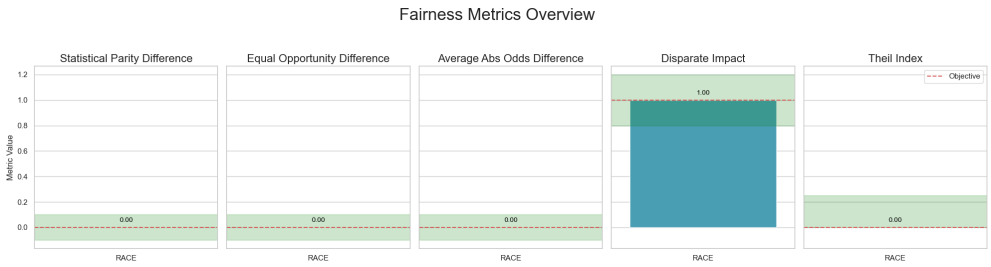

Fig. 36 Fairness Metrics in Model Trained with Discrimination Aware Loss Function

5.4 Final Comparison of Results

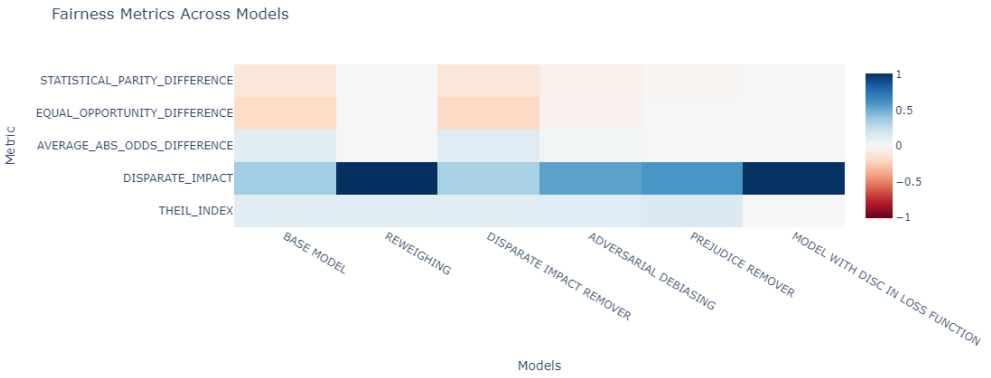

Fig. 37 Fairness Metrics Across Models

The above figure demonstrates the effectiveness of different models in addressing various fairness metrics. Notably, our novel discrimination aware loss function model shows significant improvement across multiple metrics, suggesting its effectiveness in reducing bias comprehensively.

The second figure ranks each model based on its overall fairness performance. The reweighing approach emerges as the second most effective, bested only by our novel model trained with a discrimination-aware loss function. This highlights the potential of reweighing and discrimination-aware techniques in creating more equitable algorithms.

Conclusion

We have explored various strategies to mitigate bias in machine learning models using MEPS dataset with a sensitive attribute of 'Race'. Among the five different models

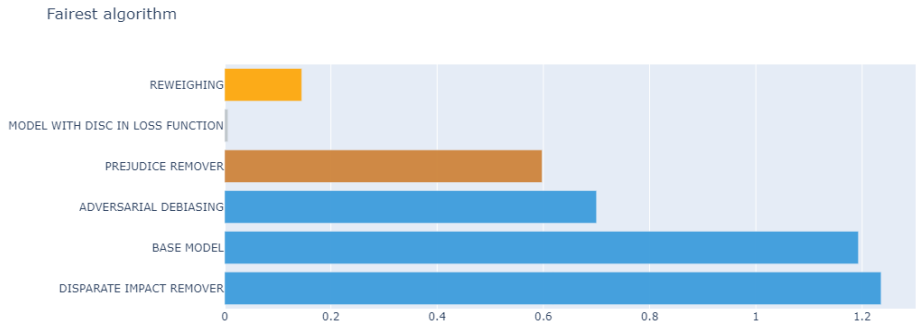

Fig. 38 Overall Fairness Assessment of Algorithms

evaluated—Base Model, Reweighing, Disparate Impact Remover (DIR), Adversarial Debiasing, Prejudice Remover—a standout contribution was the development and implementation of a novel approach, the Model with a Discrimination-Aware Loss Function. Through this project, we have highlighted the complexities of mitigating bias in machine learning, showing that various methods are needed based on specific fairness goals and data characteristics. While Reweighing was the fairest across most metrics, the novel Discrimination-Aware Loss Function proved to be an innovative and effective strategy for embedding fairness directly into the model training. This approach not only improved fairness but also advanced both the theory and practice of fair machine learning. This has emphasized the need for ongoing innovation in methods that skillfully balance accuracy and fairness. Future research could further refine this method by optimizing the discrimination term or integrating it with other bias mitigation techniques to enhance fairness outcomes.

References

[1] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015)

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

[3] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333 (2013). PMLR

[4] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Machine Learning and Knowledge Discovery in

Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23, pp. 35–50 (2012). Springer

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

[6] Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International Conference on Machine Learning, pp. 2564–2572 (2018). PMLR

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

[8] Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th International Conference on Data Mining, pp. 924–929 (2012). IEEE

[9] Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., *et al.*: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1 (2019)