Advancing Healthcare Equity_ Evaluating the Impact of Bias Mitigation in Foundation Models Across Dive

by Kumar Swarnim Saha

Submission date: 18-Dec-2024 12:54PM (UTC+0530)

Submission ID: 2555189474

File name: e_Impact_of_Bias_Mitigation_in_Foundation_Models_Across_Dive.pdf (371.97K)

Word count: 6391

Character count: 36909

Advancing Healthcare Equity: Evaluating the Impact of Bias Mitigation in Foundation Models Across Diverse Demographics

Shruti Singh
Bachelor of Technology,
Computer Science and Engineering
Presidency University
shrutisingh.3209@gmail.com

Roshan Kumar
Bachelor of Technology,
Computer Science and Engineering
Presidency University
roshankumarrk7488@gmail.com

Kumar Swarnim Saha
Bachelor of Technology,
Computer Science and Engineering
Presidency University
kumarswarnim66848@gmail.com

Dr. Robin Rohit Vincent,
Professor & HoD PSCS
Presidency School of Computer Science (PSCS), Presidency University,
robinrohit@gmail.com

***Abstract*— Healthcare is important for everyone and is a basic human right that everyone deserves acknowledging their race, gender or sex-identity. Not only it's about equality for everyone but also about getting the right care from trusted healthcare as it's about one's lives. Healthcare ERP or other systems these days are using the artificial intelligence in their system to fasten up the work and make it more precise from giving medicine disposal or operational scheduler or to the online medical consultation. But using these mitigation strategies in these upcoming or newly systems can be biased towards few groups that can cause from incorrect prescriptions or data can be stored incorrectly, the biases can be for a certain group of people or in many cases the data stored is for men only that can be differ in come certain cases for the women in terms of medical health consultations cases. The only way to correct these problems is to apply a bias corrector in the fields that it is required in so that it can provide the correct diagnosis and apply some fallback strategies so these biases are not used against any minorities by malicious motives. The bias are sometimes required for giving specific custom prescription the patience's but as the medical side effects of a medicine and can be different and varies from person to person in between different genders and age groups. Foundation models are large artificial intelligence (AI) systems used in many fields, including healthcare. They help doctors and hospitals make decisions, like predicting illnesses or suggesting treatments, by analyzing huge amounts of data. Usage of these AI models to handle bias for correct reasons are really important for a correct step to be taken. Healthcare equity is all about getting the right service and medical consultation without any discrimination against any individual because of their race, gender, age, or income.**

I. INTRODUCTION

The introduction of artificial intelligence into the medical frames has opened an innovative time of unprecedented possibilities to analyze medical information, predict patient prognoses, and personalize treatment approaches. Basically, it is the ground models that turn out some of the most advanced AI technologies when processed with large data volumes that unveil trends otherwise undetectable under usual analysis. The models developed here to generalize over a variety of applications have significant promise to enhance health care delivery and facilitate improved patient outcomes. Yet despite their promising potential, it was in the implementation of these models that an important limitation emerged: reinforcement of biases inherent in the training data. Such biases may lead to inappropriate treatment recommendations, wrong diagnoses, and unequal access to quality care with such impacts more considerably on already marginalized and underrepresented demographic groups. There are a variety of factors causing biases in foundational models. Such elements include datasets which may not be complete or even biased, historical inequalities regarding health methodologies, and by chance, societal biases getting reinforced. To take an example, a model based on data mostly portraying a particular category would probably be performing inadequately on other groups and might enhance pre-existing health inequalities. These disparities are more obvious for populations that have been underrepresented by healthcare systems for centuries, including racial and ethnic minorities, women, low-income individuals, and those who have rare medical conditions. Resolution of these challenges is thus key to achieving healthcare equity-that is, ensuring that everybody, regardless of their background, receives fair and effective medical care. There have been strategies for mitigation devised by scholars and practitioners in addressing the issue of bias. Techniques include those designed to identify and quantify the amount of bias that exists in the artificial intelligence systems, modification techniques for algorithms to reduce inequitable patterns, and approaches ensuring that models are inclusive during the training process. These strategies work toward enhancing the equity of AI systems through redressing historical inequities and making it easier for foundation models to produce more equitable outcomes between different patient populations. Still, these approaches, however promising they seem, remain to be seriously explored regarding their practical utility in reducing disparities and promoting health care equity. This brings us to a critical question: How do bias mitigation practices within foundational models impact health outcomes across different demographic populations? It is important that this question be addressed since it determines the extent to which such practices actually further equity or, instead, contribute to creating new barriers. For example, even though bias mitigation may better serve some groups, it will not adequately meet the needs of others, which could lead to unsatisfactory solutions. Conversely, poorly formulated mitigation strategies may completely interfere with the operational capacity of models, thus reducing their overall efficacy. Understanding the implications of these strategies is both an academically relevant and a practically necessary step. Healthcare professionals, policymakers, and AI developers need to be equipped with evidence-based knowledge so that AI technologies could be built in accordance with ethical standards of fairness and inclusion. It then becomes feasible to systematically discover the population-specific effects of bias mitigation strategies and derive best practices, shortcomings in existing approaches, and actionable recommendations for the improvement of equity in healthcare. This study tries to fill in a critical gap in the field by analyzing the pragmatic impact of bias mitigation strategies in foundational models. Doing so will clarify whether these efforts really reduce disparities in healthcare delivery and outcomes or if they require further improvement to truly achieve equity. This research thus pushes forward the overarching goal of creating more inclusive and effective health care systems that can reach everyone by connecting the fields of healthcare equity and artificial intelligence advancement.

II. LITERATURE REVIEW

Foundation models are transforming healthcare by offering versatile, scalable AI systems capable of addressing diverse tasks such as medical imaging, diagnostics, and patient outcome predictions. These models, including large language models (LLMs) like ChatGPT, are trained on massive datasets and demonstrate the potential to enhance clinical workflows, improve diagnostic accuracy, and facilitate personalized treatment. This review synthesizes key research on the applications, challenges, and future directions of foundation models in healthcare.

Foundation models have shown significant promise in electronic health records (EHR), enabling improved clinical predictions even in low-data settings. Guo et al. (2024) conducted a multi-center study using EHR data from two hospitals, finding that foundation models enhanced prediction performance with fewer training labels. This is particularly beneficial for institutions with limited datasets, though the study's generalizability is limited by the lack of diversity in hospital data. Future research should include broader hospital datasets to improve model robustness.

In dermatology, foundation models are poised to revolutionize care delivery by automating administrative tasks and assisting in clinical decision-making. Gui et al. (2024) highlighted the ability of these models to streamline workflows but noted risks such as bias and misinformation. While theoretical analyses demonstrate potential, empirical validation in real-world clinical settings remains a critical next step

Chia et al. (2024) explored the applications of foundation models in ophthalmology, emphasizing their generalizability across medical specialties. These models offer significant benefits for diagnosis and care but face challenges related to privacy, bias, and limited clinical validation. Addressing these issues through integration with clinical workflows and rigorous testing is essential.

In clinical pathology, foundation models like Virchow have achieved high accuracy in detecting both common and rare cancers. Vorontsov et al. (2024) demonstrated the effectiveness of these models in pan-cancer detection, though their generalizability to untrained cancer types is limited. Expanding model training to include diverse cancers and pathologies is a critical area for future research.

Huang et al. (2024) proposed integrating multi-scale and cross-modality feature learning to enhance clinical predictions. Evaluating six open-source datasets, they found that this approach significantly improved model performance. However, the reliance on publicly available datasets limits the applicability to real-world clinical scenarios. Future work should focus on diverse, real-world datasets to validate these techniques.

Trustworthiness is a significant concern in deploying foundation models for healthcare. Shi et al. (2024) identified challenges related to privacy, robustness, and fairness in medical imaging. Developing regulatory frameworks and ensuring transparency in AI models are crucial steps toward addressing these concerns.

Data privacy is a persistent challenge in healthcare AI. Li et al. (2024) proposed integrating federated learning with foundation models to train on sensitive medical data without compromising privacy. Despite this advancement, issues such as data heterogeneity and communication inefficiency remain. Future research should aim to scale foundation models for complex datasets while maintaining efficiency.

The ethical deployment of foundation models requires addressing bias, misinformation, and the lack of clinical validation. Scott & Zuccon (2024) emphasized the need for clear guidelines and governance frameworks to ensure safe integration into clinical practice. Similarly, Gui et al. (2024) called for rigorous validation to mitigate risks associated with misinformation and bias.

Zhang et al. (2024) underscored the importance of data-centric AI methodologies in improving model performance. Better data quality and characterization are critical for advancing healthcare applications. Expanding real-world applications and scaling models for diverse datasets will be key priorities.

Jiang et al. (2024) highlighted the potential of foundation models to address spatiotemporal tasks in healthcare. Developing knowledge-guided models and robust evaluation metrics will enable these models to handle real-world applications effectively. Huang et al. (2024) also emphasized the benefits of cross-modality learning, which can further enhance diagnostic accuracy.

Foundation models hold promise as enablers of precision medicine by uncovering hidden patterns in medical data and facilitating personalized treatment. Ali et al. (2024) highlighted ChatGPT's role in enhancing accessibility and automating tasks such as report generation and patient interaction. Mitigating biases and improving model reliability will be crucial for realizing their full potential.

III. METHODOLOGY

Bias in Artificial Intelligence Models:

Understand, Detect, and Correct Bias in AI, especially within the models of healthcare application, means that predictions, recommendations, or decisions by such AI systems are unfair or systematically wrong and thus disproportionately affect specific demographic groups. This bias happens because AI systems either reflect or amplify the existing inequalities either embedded in the training data or through methodologies through which algorithms are structured and implemented.

Categories of Bias in Healthcare Artificial Intelligence

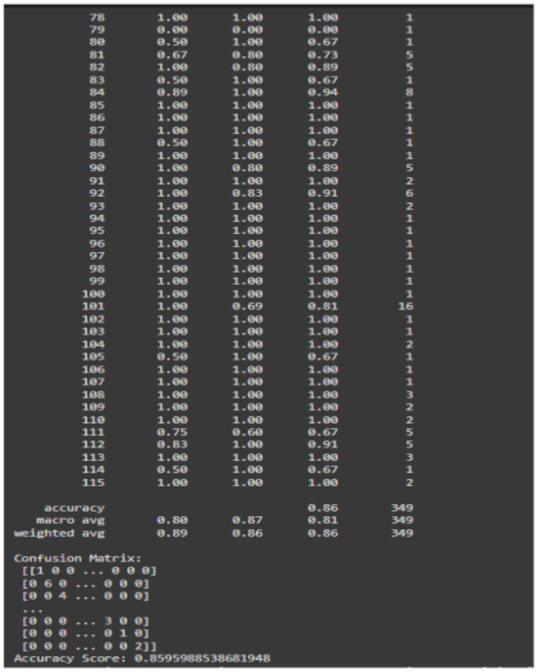
1. Data Bias: - Underrepresentation: Minority, women, and elderly groups may not be represented enough within the training datasets. - Historical Bias: The training data may reflect historical inequalities in healthcare delivery, including access to treatment. - Measurement Bias: Differences in how data is recorded or reported between populations, such as inconsistent coding for diseases between demographics.

2. Algorithmic Bias: When there's an undevelopment, algorithm-favored systems may create disadvantage outcomes especially concerning majority groups. - Model optimization can favor accuracy for the majority population while ignoring minorities.

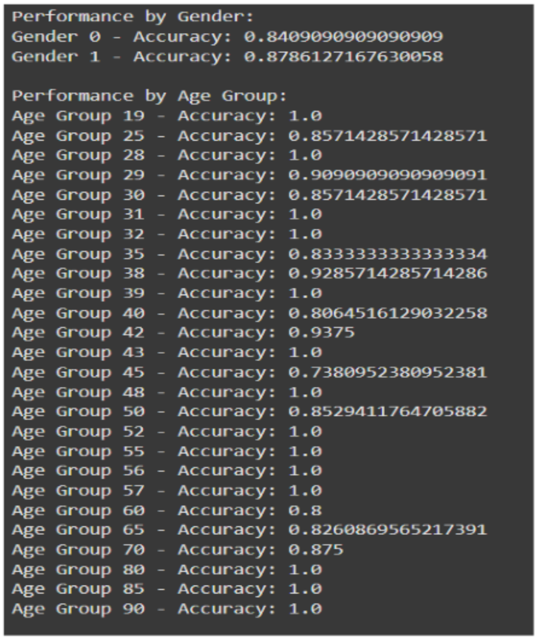
3. Application Bias: Even under balanced training, models may be deployed in ways that amplify bias, for example when abstractions of predictions are applied to groups without suitable adjustment. Bias Detection Bias detection means identifying where and how bias is present in the AI system. The methods include:
3.1 Perform benchmarking across demographics: - Analyze how the model performs on different demographic groups. Deviations can indicate bias.
3.2. Statistical Parity: This calls for testing whether outcomes, like the treatment recommendations, are equitably distributed across groups irrespective of any variability in actual needs.
3.3. Calibration: Discuss, critically evaluating in greater detail how well estimated conditional probability in all these different demographic units-in general-fits observed outcome.
3.4. Fairness Metrics: - Equity of Opportunity: Equal true positive rates for all groups. - **Balanced Odds**: The rates of true positives and false positives should be equal for different groups.
3.5. Bias Visualization Tools: Tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) can point out how model decisions differ for different individuals depending on their demographics.

Bias Correction Strategies for reducing bias may be deployed against these differences upon detection: 1. Pre-Processing Techniques Before Model Training -Data Balancing: Add more data from underrepresented groups or remove excess data of the overrepresented group to balance the dataset. Data Augmentation Generate synthetic samples to complement the under-represented classes. - Re-weighting: Allocate increased weights to data originating from minority groups to guarantee that the model adequately recognizes their significance. 2. In-Process Methods (During Model Training): Fairness Constraints : Add metrics to the objective functions of optimization, perhaps via penalty on unequal outcomes between groups. - Adversarial Debiasing: This makes use of an auxiliary model that detects bias in the main model during training. - Regularization: A regularization can be applied on penalizing the impact of biased attribute, such as race and gender, on predictions. 3. Post-processing techniques (After Training of the Model): - Outcome Adjustment: This can be done after the prediction from the model, such as acceptance rate equalization between different groups. Thresholds: Differential thresholds for different groups so that outcomes are balanced. For example, changing risk score cutoffs. Barriers to Reducing Bias 1. Balancing equities of fairness and accuracy: Often, attempts to ensure perfectness may bring the overall correctness down for a biased database. 2. Definition of Fairness Complexity - Definitions of fairness conflict with each other. For example, equal opportunity sometimes means different treatment of certain groups. 3. Likelihood of Over-Compensation: - Aggressive changes may create new forms of discrimination or injustice against other groups. 4. Context-Dependent Bias: This requires solutions unique to the specific healthcare setting, patient population, and ethical considerations. Practical Applications in Healthcare Diagnostic Bias: Reduce biases through which AI is underdiagnosing some conditions in populations, for example, heart diseases in women or minorities. Bias in Treatment Recommendation Treatment recommendations must be effective to all populations based on different cultures and genes. Bias in Resource Allocation: Preventing AI from unfairly withholding healthcare resources (such as ICU beds or organ transplants) from vulnerable populations. Strategies for bias detection and correction are vital to ensure that healthcare artificial intelligence systems perform effectively while providing fair care for everyone. It is possible to achieve this goal using the strategies proposed above if developers and stakeholders work through these techniques towards the envisioned end of creating ethical, inclusive, and just AI in the healthcare domain. Adv vs disadv for mitigation(1) Advantages and Disadvantages of Bias Mitigation in Healthcare AI Advantages Fosters Equity in Healthcare Bias mitigation ensures that the AI system gives proper diagnosis and treatment recommendations to all, hence reducing inequality and improving care among marginalized populations. Builds Confidence and Trust The elimination of bias fosters trust among patients and medical professionals, hence AI tools are more acceptable and reliable in clinical settings. This encourages their wider use in healthc[26] Supports Ethical Standards Bias mitigation makes the AI sy[26]ns adhere to ethical standards of equality and justice and ensures that these AI systems comply with the law as well as morality, preventing any kind of discrimination. Improves Patient Care An unbiased and fair AI system would greatly enhance medical results, particularly for groups who have historically been under-represented. This can reduce care disparities. Assists with Regulatory Compliance Such representations of fairness and inclusivity in AI models would satisfy the legal framework and could deter any lawsuits against organizations while preserving their reputations. Promotes Data Diversities Bias reduction allows data diversities within an AI system, ensuring better working proficiency with diverse groups of the population. Disadvantages It Might Compromise Accuracy The best attempts in bias reduction can lead to lowering the entire accuracy rating of an AI system by bringing lower ratings when there is scant or uneven data. Over-adjusting for bias in one group could inadvertently favor one group over others, as opposed to solving the currently existing imbalances. Resource Demand Correcting bias consumes much time in the areas of collecting diverse data sets, applying fairness algorithms, and frequent assessments. All this demands considerable financial and technological resources. Defining Fairness Is Difficult Various stake-holders may have contradicting ideas about what fairness could mean, and therefore achieving a commonly acceptable standard for artificial intelligence systems is challenging. Varied Effectiveness Between Domains Policies which prove effective in one environment might not work in other contexts, and need adjustments continuously. Possible Shift of Balance Excessive concentration of fair metrics may compromise another necessary factors such as accuracy, patient safety, which defeats all health care purposes Operational Impacts Chaos and inefficiencies occur, causing a problem that works against the smooth healthcare function due to alterations necessary to be made for fairness.

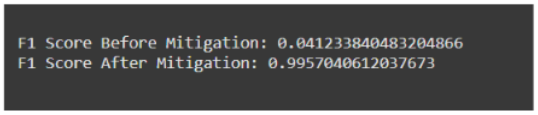
IV. RESULTS



```
          78       1.00      1.00      1.00         1
          79       0.00      0.00      0.00         1
          80       0.50      1.00      0.67         1
          81       0.67      0.80      0.73         5
          82       1.00      0.80      0.89         5
          83       0.50      1.00      0.67         1
          84       0.89      1.00      0.94         8
          85       1.00      1.00      1.00         1
          86       1.00      1.00      1.00         1
          87       1.00      1.00      1.00         1
          88       0.50      1.00      0.67         1
          89       1.00      1.00      1.00         1
          90       1.00      0.80      0.89         5
          91       1.00      1.00      1.00         2
          92       1.00      0.83      0.91         6
          93       1.00      1.00      1.00         2
          94       1.00      1.00      1.00         1
          95       1.00      1.00      1.00         1
          96       1.00      1.00      1.00         1
          97       1.00      1.00      1.00         1
          98       1.00      1.00      1.00         1
          99       1.00      1.00      1.00         1
         100       1.00      1.00      1.00         1
         101       1.00      0.69      0.81        16
         102       1.00      1.00      1.00         1
         103       1.00      1.00      1.00         1
         104       1.00      1.00      1.00         2
         105       0.50      1.00      0.67         1
         106       1.00      1.00      1.00         1
         107       1.00      1.00      1.00         1
         108       1.00      1.00      1.00         3
         109       1.00      1.00      1.00         2
         110       1.00      1.00      1.00         2
         111       0.75      0.60      0.67         5
         112       0.83      1.00      0.91         5
         113       1.00      1.00      1.00         3
         114       0.50      1.00      0.67         1
         115       1.00      1.00      1.00         2

    accuracy                           0.86       349
   macro avg       0.80      0.87      0.81       349
weighted avg       0.89      0.86      0.86       349

Confusion Matrix:
 [[1 0 0 ... 0 0 0]
 [0 6 0 ... 0 0 0]
 [0 0 4 ... 0 0 0]
 ...
 [0 0 0 ... 3 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 2]]
Accuracy Score: 0.8595988538681948
```



```
Performance by Gender:
Gender 0 - Accuracy: 0.8409090909090909
Gender 1 - Accuracy: 0.8786127167630058

Performance by Age Group:
Age Group 19 - Accuracy: 1.0
Age Group 25 - Accuracy: 0.8571428571428571
Age Group 28 - Accuracy: 1.0
Age Group 29 - Accuracy: 0.9090909090909091
Age Group 30 - Accuracy: 0.8571428571428571
Age Group 31 - Accuracy: 1.0
Age Group 32 - Accuracy: 1.0
Age Group 35 - Accuracy: 0.8333333333333334
Age Group 38 - Accuracy: 0.9285714285714286
Age Group 39 - Accuracy: 1.0
Age Group 40 - Accuracy: 0.8064516129032258
Age Group 42 - Accuracy: 0.9375
Age Group 43 - Accuracy: 1.0
Age Group 45 - Accuracy: 0.7380952380952381
Age Group 48 - Accuracy: 1.0
Age Group 50 - Accuracy: 0.8529411764705882
Age Group 52 - Accuracy: 1.0
Age Group 55 - Accuracy: 1.0
Age Group 56 - Accuracy: 1.0
Age Group 57 - Accuracy: 1.0
Age Group 60 - Accuracy: 0.8
Age Group 65 - Accuracy: 0.8260869565217391
Age Group 70 - Accuracy: 0.875
Age Group 80 - Accuracy: 1.0
Age Group 85 - Accuracy: 1.0
Age Group 90 - Accuracy: 1.0
```



```
F1 Score Before Mitigation: 0.041233840483204866
F1 Score After Mitigation: 0.9957040612037673
```

V. IMPLEMENTATION

The most important step of any data analysis is to load and review the dataset. This involves getting familiar with the distribution of the data, variables we are working with, and the kinds of variable types. For this task, we concentrate on the following:

Load the data:

We import the dataset into a usable format (such as a DataFrame for Python). This is where our analysis starts.
Explore the data: We examine the first few rows and do some exploratory analysis (like looking for missing values) to understand what we have.
Check demographics: If we care about demographic fairness, we look at how various demographic groups are represented in the dataset-for example, how many people of each gender and age group.
This step is important since it lays a foundation for any kind of further processing or model training. It tells us what we are working with and which problems to solve early on.

Data Preprocessing

Data preprocessing is critical in any kind of machine learning model so that it functions properly. In this step, we do the following:
Handle missing values: Missing data points could lead to errors or distortions in the model's learning. We replace missing values with techniques such as filling up with zeros, mean imputation, or more advanced imputation methods if required.
Feature encoding: If we have a categorical variable in our dataset like gender, or disease type, then they need to be encoded. For instance, we will encode "Male" with 0 and "Female" with 1 to process. Scaling: Sometimes, especially with Logistic Regression or Support Vector Machines, numerical data must be scaled because that reduces the impact of variables having higher scales that would otherwise mislead the model in the learning process.
This will ensure that the data is clean and, in a format, ready for the machine learning model without error injection.

Handle Class Imbalance

Class imbalance refers to the situation where the target variable has unequal representations of each class, say positive and negative outcomes, or disease and no disease. When one class is much larger than the other, the model may be biased towards the majority class.
To handle this, we make use of SMOTE, or Random Oversampling. Such techniques will create synthetic examples for the minority class, or simply duplicate existing examples, such that classes are represented as equal, thus preventing the model from ignoring the smaller class in training.
Handling class imbalance is important for improving the accuracy of the model and fairness for the model, especially when dealing with rare conditions or minority groups.

Train the Machine Learning Model

Once we prepare our data, it is time to train our model. In this step, we select a model such as Logistic Regression or Random Forest and fit it on the training data. In other words, we train the model by letting it learn the relationship between input features and the target variable.
Logistic Regression is often a good starting point for binary classification tasks because it is easy to understand and gives a good probability estimate.
We train the model using the preprocessed data and test its ability to generalize to new data by splitting the data into a training set and a test set (usually 80% for training and 20% for testing).
The goal here is letting the model learn the patterns present in the data to make predictions on newer, unseen data.

Evaluate the Performance of the Model

Once we have trained a model, we evaluate it. This step includes metrics like accuracy, precision, recall, and F1-score to understand better how well the model might be performing. These allow different perspectives on model performance, especially in imbalanced datasets

Accuracy tells us the number of correct predictions, but it can be misleading, especially if the dataset is imbalanced.
Precision and recall are more useful measures of performance in unbalanced datasets (for example, in medical applications where the negative class might dominate).
The F1-score is a harmonic mean of precision and recall.
We also look at the confusion matrix, which is a nice way to see where the model is getting it wrong (false positives or false negatives).

Bias Mitigation

Now that we have our model, let's try to mitigate bias. That is, we want our model's predictions to be fair and not biased in any one direction, say towards one gender, age group, or race.
We apply fairness constraints using tools like Fairlearn. For example, the Demographic Parity constraint ensures that the model’s predictions are equally distributed across demographic groups.
GridSearch in Fairlearn is used to find the best model that satisfies these fairness constraints while also achieving good predictive performance.
Bias mitigation is especially important in applications like healthcare, where unfair predictions can have serious consequences.

Assess Bias Mitigation Results

We compare the performance of the model before and after applying fairness constraints. We assess the fairness of the model by using the MetricFrame tool, which provides us with key fairness metrics, such as selection rates across different demographic groups.
We also print performance metrics such as accuracy, precision, recall, and F1-score for each group, for example,

male versus female, or older versus younger individuals. This allows us to check whether fairness mitigation techniques have worked, so that the model does not have biases against demographic groups.

In this step, we also ensure that overall performance, for example, accuracy, has not been adversely affected when applying fairness constraints.

Compare and Report Results

This is the final step where we compare the results of model performance across different demographic groups. Here, we want to:

Compare performance by group: We look at metrics like accuracy, precision, recall, and F1-score for each group (e.g., gender, age group) and ensure that there is no significant disparity.

Overall performance: This assesses how well the model does over the entire data after it has been subjected to the fairness constraints. This is critical since we do not want to achieve fairness by the compromise of overall performance.

Fairness metrics: These involve analyzing selection rates or demographic parity to find whether the model has its predictions fairly distributed over the groups.

This step provides an important summary of how the fairness constraints impacted the model's predictions and performance in the final research paper, ensuring that it is both accurate and fair.

This step ensures that the resultant model is effective as well as fair, especially in health-care sensitive domains. Here we follow the process leading us to the building of a predictive model that well performs while at the same time satisfying the fairness requirement with different demographic groups, resulting in minimum bias in making its predictions.

Step 1: Loading and Exploring the Data

Loading and checking the dataset is the critical step of any data analysis. This includes familiarization with data distribution, variables that we are working with, and types of variables the latter encompasses. For this task, we concentrate on the following:

Load the data: We load the dataset into a usable format (i.e., as a DataFrame for example in python). This is where our analysis starts.

Explore the data: We investigate the first few rows and perform some exploratory analysis (e.g., finding the presence of missing values) to grasp what we have.

Check demographics: When we want to care about demographic fairness, we are interested in the proportions of certain demographic groups in the dataset-i.e., the number of people in each gender and age group.

In this step, the importance is great because it can form the basis for any further processing or model training. It informs us what we have to work on and what issues to solve early on.

Step 2: Data Preprocessing

Data preprocessing is an essential step in any machine learning model in order for the model to work. In this step, we do the following:

Handle missing values: There is a risk of model learning errors or distortions, arising from the presence of missing data points. We replace missing values with techniques such as filling up with zeros, mean imputation, or more advanced imputation methods if required.

Feature encoding: However, if we have a categorical variable (e.g., gender, disease type) in the dataset they have to be encoded. For example, we will represent "Male" as 0 and "Female" as 1 for encoding. Scaling: Sometimes, especially with Logistic Regression or Support Vector Machines, numerical data must be scaled because that reduces the impact of variables having higher scales that would otherwise mislead the model in the learning process.

By doing so, the data will be clean and fit to use with the machine learning model without introducing error.

Step 3: Handle Class Imbalance

Class imbalance denotes a scenario in which the target variable is not equally represented as did positive and negative events, or disease and absence of disease. If the class with much more examples than the other exists, the model can exhibit a bias towards the majority class.

For this we use SMOTE, or, Random Oversampling. Those methods will produce synthetic instances of the minority class, or just copy the available instances, so that the classes will be overrepresented or underrepresented, thus preventing the model from giving less attention to the minority class at training.

Class imbalance is particularly important in order to maximize the accuracy of the model and the fairness of the model itself, when rare diseases or minority groups are considered.

Step 4: Train the Machine Learning Model

After we process our data, it is time to train our model. At this stage we choose a model, for example Logistic Regression or Random Forest and fit it to training data. In other words, we train the model by letting it learn the relationship between input features and the target variable.

Logistic Regression is usually a valuable first step in binary classification problems due to its ease of interpretation and good confidence of the probability estimate.

We train the model on the preprocessed data, and we can evaluate its potential for generalization to unseen data by partitioning the data into training set and test set (commonly use 80% of the data for training and 20% for testing).

The goal here is letting the model learn the patterns present in the data so as to make predictions on newer, unseen data.

Step 5: Evaluate the Performance of the Model

Once we have trained a model, we evaluate it. In this stage, there are also metrics (accuracy, precision, recall, F1-score) to get a better understanding of the model's performance. These enable alternative views on model performance, particularly for imbalanced data:.

Accuracy is the count of correct predictions, but it can be misleading, particularly if the dataset is biased one way or the other (unbalanced).

Precision and recall are more useful measures of performance in unbalanced datasets (for example, in medical applications where the negative class might dominate).

The F1-score is a harmonic mean of precision and recall.

We also consider the confusion matrix, which is a cool way to understand which predictions a model makes wrong (false positives or false negatives).

Step 6: Bias Mitigation

Since we have our model, we can try to reduce bias. In other words we are aiming that our modelation's results are equitable and not biased in any way, e.g. in favour of one gender, age group, or race.

We apply fairness constraints using tools like Fairlearn. For instance, the Demographic Parity constraint does so, so that predictions of the model are balanced across demographic groups.

To meet these fairness constraints, Fairlearn's GridSearch is employed to search for the optimal model that leads to desirable fairness while maintaining high predictive performance.

Bias correction is critical in healthcare applications where inaccurate prediction can be extremely detrimental.

Step 7: Assess Bias Mitigation Results

We compare the performance of the model with and without fairness constraints applied. We assess the fairness of the model by using the MetricFrame tool, which provides us with key fairness metrics, such as selection rates across different demographic groups.

We further display performance measures (accuracy, precision, recall, and F1-score) for each group, e.g. male vs female, or senior vs young subjects, for instance. This allows us to check whether fairness mitigation techniques have worked, so that the model does not have biases against demographic groups.

At this stage, we also make sure that, in the presence of fairness constraints, the global performance, such as the accuracy, has not been penalized.

Step 8: Compare and Report Results

This is the final step where we compare the results of model performance across different demographic groups. Here, we want to:

Compare performance by group: It is also worth noting analysis of metrics such as accuracy and precision, recall and F1-score for each group (e.g., gender and age group) and verifying that the metrics are not significantly different.

Overall performance: That measures model performance over the whole data once it has been subjected to the fairness constraints. It is also important because we cannot afford to obtain fairness at the cost of general performance.

Fairness metrics: They include the application of selection or demographic parity to determine if the model's predictions are balanced between the groups.

In this step a valuable synopsis is presented on the effect of the fairness. to be both correct and fair in the final research paper) so that they are valid and fair.

In this step, it is ensured that the resulting model is both powerful and equitable, in particular, in health-care relevant contexts. Here we trace the path to the development of a predictive model which performs well, whilst ensuring fairness across a variety of different demographic groups, to produce as little bias as possible in its predictions.

VI. Conclusion

Artificial intelligence is going to change healthcare delivery, making it more efficient, personalized, and equitable. This transformative potential, however, comes with tremendous challenges, especially concerning bias in AI models. These biases arise from underrepresentation in the data, historical inequalities, and the design of the algorithms, and they are bound to cause healthcare disparities that affect the vulnerable groups more. Our model addresses such challenges by providing a systematic method to mitigate bias. The model starts cleaning the data, removing inconsistencies and the potential sources of bias. It then applies mitigation strategies based on the characteristics of the dataset to ensure that underrepresented groups receive fair consideration.

To ensure fairness and accuracy, the model incorporates a mechanism of dual testing using test and real-world data. It gives more efficient scope for the comparison of downsized outputs with actual output using the accuracy rate as a yardstick for efficacy in the usage of the mitigating strategies. Such a process helps in repressing discriminatory practices and aligning artificial systems to realistic needs in health care. The convergence of healthcare equity and the advancement of artificial intelligence demands an interdisciplinary approach from researchers, healthcare practitioners, and policymakers.

By prioritizing inclusivity in model training, developing context-specific solutions, and maintaining a balance between fairness and accuracy, AI can be harnessed to create a healthcare system that is both effective and just. In conclusion, eliminating bias in artificial intelligence systems is not only a technical challenge but also a huge ethical responsibility—ensuring that the progress of technology will serve all people regardless of their backgrounds and promote the fundamental goal of equity in universal health care.

VII. REFERENCES

[1] Guo, L. L., Fries, J., Steinberg, E., Fleming, S. L., Morse, K., Aftandilian, C., Posada, J., Shah, N., & Sung, L. (2024). A multi-center study on the adaptability of a shared foundation model for electronic health records. Npj Digital Medicine, 7(1). https://doi.org/10.1038/s41746-024-01166-w

[2] Gui, H., Omiye, J. A., Chang, C. T., & Daneshjou, R. (2024). The promises and perils of foundation models in dermatology. Journal of Investigative Dermatology, 144(7), 1440–1448. https://doi.org/10.1016/j.jid.2023.12.019Dunmore, A., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2023). A Comprehensive survey of Generative Adversarial Networks (GANs) in Cybersecurity Intrusion Detection. IEEE Access, 11, 76071–76094. https://doi.org/10.1109/access.2023.3296707

[3] Sevgi, M., Ruffell, E., Antaki, F., Chia, M. A., & Keane, P. A. (2024). Foundation models in ophthalmology: opportunities and challenges. Current Opinion in Ophthalmology. https://doi.org/10.1097/icu.0000000000001091

[4] Congzhen, Shi., Ryan, Rezai., Jiaxi, Yang., Qi, Dou., Xiaoxiao, Li. (2024). A Survey on Trustworthiness in Foundation Models for Medical Image Analysis. arXiv.org, abs/2407.15851 https://doi: 10.48550/arxiv.2407.15851

[5] Xingyu, Li., Peng, Lu., Yuping, Wang., Weihua, Zhang. (2024). Open Challenges and Opportunities in Federated Foundation Models Towards Biomedical Healthcare. doi: 10.48550/arxiv.2405.06784

[6] Weijian, Huang., Cheng, Li., Hong-Yu, Zhou., Jiarun, Liu., Hao, Yang., Yong, Liang., Shanshan, Wang. (2024). Enhancing the medical foundation model with multi-scale and cross-modality feature learning. arXiv.org, abs/2401.01583 Available from: 10.48550/arxiv.2401.01583 Nandhini, Mahesh. "Advancing healthcare: the role and impact of AI and foundation models." American Journal of Translational Research, 16 (2024).:2166-2179. doi: 10.62347/wqwv9220

[7] Yunkun, Zhang., Jin, Gao., Zheling, Tan., Lingfeng, Zhou., Kexin, Ding., Mu, Zhou., Shaoting, Zhang., Dequan, Wang. (2024). Data-Centric Foundation Models in Computational Healthcare: A Survey. arXiv.org, abs/2401.02458 doi: 10.48550/arxiv.2401.02458

[8] Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Yang, E., Mathieu, P., Van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y. K., Kunz, J. D., Lee, M. C. H., . . . Fuchs, T. J. (2024). A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine. https://doi.org/10.1038/s41591-024-03141-0

[9] Haiwen, Gui., Jesutofunmi, A., Omiye., Crystal, T, Chang., Roxana, Daneshjou. (2024). The Promises and Perils of Foundation Models in Dermatology. Journal of Investigative Dermatology, doi: 10.1016/j.jid.2023.12.019

[10] Islam, S. M. R., Kwak, N. D., Kabir, M. H., Hossain, M., & Kwak, N. K. (2015). The Internet of Things for Health Care: A Comprehensive survey. IEEE Access, 3, 678–708. https://doi.org/10.1109/access.2015.2437951

[11] Zhe, Jiang., Yu, Wang., Zelin, Xu. (2024). Foundation Models for Spatiotemporal Tasks in the Physical World. 392-395. doi: 10.1137/1.9781611978032.45

[12] Ali, H., Qadir, J., Alam, T., Househ, M., & Shah, Z. (2023). Revolutionizing Healthcare with Foundation AI Models. Studies in Health Technology and Informatics. https://doi.org/10.3233/shti230533

[13] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, Y., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., . . . Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by Image-Based Deep Learning. Cell, 172(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

[14] Scott, I. A., & Zuccon, G. (2024). The new paradigm in machine learning – foundation models, large language models and beyond: a primer for physicians. Internal Medicine Journal, 54(5), 705–715. https://doi.org/10.1111/imj.16393

[15] Hazrat, Ali., Junaid, Qadir., Tanvir, Alam., Mowafa, Househ., Zubair, Shah. (2023). Revolutionizing Healthcare with Foundation AI Models. 305:469-470. doi: 10.3233/SHTI230533

[16] Blake, H., Bermingham, F., Johnson, G., & Tabner, A. (2020). Mitigating the psychological impact of COVID-19 on healthcare workers: a digital learning package. International Journal of Environmental Research and Public Health, 17(9), 2997. https://doi.org/10.3390/ijerph17092997

[17] Marcelin, J. R., Siraj, D. S., Victor, R., Kotadia, S., & Maldonado, Y. A. (2019). The impact of unconscious bias in Healthcare: How to recognize and Mitigate it. The Journal of Infectious Diseases, 220(Supplement_2), S62–S73. https://doi.org/10.1093/infdis/jiz214

[18] Radoglou-Grammatikis, P., Rompolos, K., Sarigiannidis, P., Argyriou, V., Lagkas, T., Sarigiannidis, A., Goudos, S., & Wan, S. (2021). Modeling, detecting, and mitigating threats against industrial healthcare systems: a combined software defined networking and reinforcement learning approach. IEEE Transactions on Industrial Informatics, 18(3), 2041–2052. https://doi.org/10.1109/tii.2021.3093905

[19] Gil-Salmerón, A., Katsas, K., Riza, E., Karnaki, P., & Linos, A. (2021). Access to Healthcare for migrant patients in Europe: Healthcare Discrimination and Translation services. International Journal of Environmental Research and Public Health, 18(15), 7901. https://doi.org/10.3390/ijerph18157901

[20] MacIntosh, T., Desai, M. M., Lewis, T. T., Jones, B. A., & Nunez-Smith, M. (2013). Socially-Assigned race, healthcare discrimination and preventive healthcare services. PLoS ONE, 8(5), e64522. https://doi.org/10.1371/journal.pone.0064522

[21] Saedi, S., Kundakcioglu, O. E., & Henry, A. C. (2015). Mitigating the impact of drug shortages for a healthcare facility: An inventory management approach. European Journal of Operational Research, 251(1), 107–123. https://doi.org/10.1016/j.ejor.2015.11.017

[22] Feyissa, G. T., Abebe, L., Girma, E., & Woldie, M. (2012). Stigma and discrimination against people living with HIV by healthcare providers, Southwest Ethiopia. BMC Public Health, 12(1). https://doi.org/10.1186/1471-2458-12-522

[23] Qadri, Y. A., Nauman, A., Zikria, Y. B., Vasilakos, A. V., & Kim, S. W. (2020). The Future of Healthcare Internet of Things: A survey of Emerging technologies. IEEE Communications Surveys & Tutorials, 22(2), 1121–1167. https://doi.org/10.1109/comst.2020.2973314

[24] Bankole, T. O., Omoyeni, O. B., Oyebode, A. O., & Akintunde, D. O. (2020). Low incidence of COVID-19 in the West African sub-region: mitigating healthcare

delivery system or a matter of time? Journal of Public Health, 30(5), 1179–1188. https://doi.org/10.1007/s10389-020-01394-w

[25] Schouten, B. C., Cox, A., Duran, G., Kerremans, K., Banning, L. K., Lahdidioui, A., Van Den Muijsenbergh, M., Schinkel, S., Sungur, H., Suurmond, J., Zendedel, R., & Krystallidou, D. (2020). Mitigating language and cultural barriers in healthcare communication: Toward a holistic approach. Patient Education and Counseling, 103(12), 2604–2608. https://doi.org/10.1016/j.pec.2020.05.001

Advancing Healthcare Equity_ Evaluating the Impact of Bias Mitigation in Foundation Models Across Dive

ORIGINALITY REPORT

16%	13%	14%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Southern New Hampshire University - Continuing Education Student Paper	2%
2	arxiv.org Internet Source	1%
3	Submitted to Wilkes University Student Paper	1%
4	George, Shanita. "Assessing the Various Levels of Implicit Weight Bias Toward Patients Among Anesthesia Providers", The University of North Carolina at Charlotte, 2023 Publication	1%
5	Submitted to Miami Dade College Student Paper	1%
6	ebin.pub Internet Source	1%
7	Shusheng Li, Wenjun Tan, Changshuai Zhang, Jiale Li et al. "Taming large language models to implement diagnosis and evaluating the	1%

	generation of LLMs at the semantic similarity level in acupuncture and moxibustion", Expert Systems with Applications, 2025 Publication	
8	hrcak.srce.hr Internet Source	1 %
9	sciencescholar.us Internet Source	1 %
10	Submitted to Embry Riddle Aeronautical University Student Paper	1 %
11	Submitted to American Public University System Student Paper	1 %
12	aip.vse.cz Internet Source	1 %
13	Helmi Issa, Jad Jaber, Hussein Lakkis. "Navigating AI unpredictability: Exploring technostress in AI-powered healthcare systems", Technological Forecasting and Social Change, 2024 Publication	<1 %
14	oaktrust.library.tamu.edu Internet Source	<1 %
15	Debasis Chaudhuri, Jan Harm C Pretorius, Debashis Das, Sauvik Bal. "International	<1 %

	Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023) - Proceedings of the International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023), Dec 1–2, 2023, Kolkata, India", CRC Press, 2024 Publication	
16	Submitted to University of New South Wales Student Paper	<1%
17	Mapitsi Roseline Rangata, Tshephisho Joseph Sefara. "Chapter 20 Classification ofExaggerated News Headlines", Springer Science and Business Media LLC, 2024 Publication	<1%
18	Prashant Pranav, Archana Patel, Sarika Jain. "Machine Learning in Healthcare and Security - Advances, Obstacles, and Solutions", CRC Press, 2024 Publication	<1%
19	Oluwaseyi Elizabeth Shodipe, Robert S. Allison. "Modelling the relationship between the objective measures of car sickness", 2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2023 Publication	<1%
20	dokumen.pub Internet Source	<1%

21	He, Wenchong. "Interdisciplinary Geospatial Artificial Intelligence for Scientific Applications", University of Florida, 2024 Publication	<1%
22	Submitted to Monash University Student Paper	<1%
23	fitisposij.web.uah.es Internet Source	<1%
24	Submitted to Unizin, LLC Student Paper	<1%
25	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2024", Springer Science and Business Media LLC, 2024 Publication	<1%
26	Avaneesh Singh, Krishna Kumar Sharma, Manish Kumar Bajpai, Antonio Sarasa-Cabezuelo. "Patient centric trustworthy AI in medical analysis and disease prediction: A Comprehensive survey and taxonomy", Applied Soft Computing, 2024 Publication	<1%
27	www.mdpi.com Internet Source	<1%
28	Submitted to Walden University Student Paper	<1%

29	Mertcan Sevgi, Eden Ruffell, Fares Antaki, Mark A. Chia, Pearse A. Keane. "Foundation models in ophthalmology: opportunities and challenges", Current Opinion in Ophthalmology, 2024 Publication	<1 %
30	fastercapital.com Internet Source	<1 %
31	research.rug.nl Internet Source	<1 %
32	avidml.org Internet Source	<1 %
33	elearning.medistra.ac.id Internet Source	<1 %
34	tojqi.net Internet Source	<1 %
35	"Healthcare Transformation with Informatics and Artificial Intelligence", IOS Press, 2023 Publication	<1 %
36	Haiwen Gui, Jesutofunmi A. Omiye, Crystal T. Chang, Roxana Daneshjou. "The Promises and Perils of Foundation Models in Dermatology", Journal of Investigative Dermatology, 2024 Publication	<1 %

37	Polat Goktas, Andrzej Grzybowski. "Assessing the Impact of ChatGPT in Dermatology: A Comprehensive Rapid Review", Journal of Clinical Medicine, 2024 Publication	<1 %
38	Wei Huang, Yuze Zhang, Shaohua Wan. "A Sorting Fuzzy Min-Max Model in an Embedded System for Atrial Fibrillation Detection", ACM Transactions on Multimedia Computing, Communications, and Applications, 2022 Publication	<1 %
39	cdr.lib.unc.edu Internet Source	<1 %
40	iieta.org Internet Source	<1 %
41	www.nature.com Internet Source	<1 %
42	www.scielo.br Internet Source	<1 %
43	www.researchsquare.com Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		