**MITIGATING BIAS IN AI FOR HEALTHCARE APPLICATIONS**

*Submitted in partial fulfillment of the requirements for the degree of*

Bachelor of Technology

in

**Information Technology**

*by*

**Mridul Jain 21BIT0377**

**Under the guidance of**

**Prof. / Dr. Kishore Raja PC**

**Score VIT, Vellore.**



October, 2024

## Declaration

I hereby declare that the thesis entitled “Mitigating the selection bias in AI for healthcare applications.” I submitted for the award of the degree of *Bachelor of Technology in Information Technology* to VIT is a record of bonafide work carried out by me under the supervision of Kishore Raja PC

I further declare that the work reported in this thesis has not been submitted and will not be

submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place : Vellore

Date : 26.10.2024

**Signature of the Candidate**

## Certificate

This is to certify that the thesis entitled “Mitigating the selection bias in AI for healthcare applications.” submitted by Mridul Jain (21BIT0377) VIT, for the award of the degree of *Bachelor of Technology in Information Technology*, is a record of bonafide work carried out by him/her under my supervision

during the period, 01. 12. 2018 to 30.04.2019, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The

thesis fulfills the requirements and regulations of the University and in my opinion, meets the necessary standards for submission.

Place : Vellore

Date : 26.10.2024 **Signature of the Guide**

**Internal Examiner External Examiner**

## Acknowledgement

I would like to express my deepest gratitude to my project advisor, Prof. Kishore Raja PC for his unwavering support, guidance, and encouragement throughout this research. Their insightful feedback and expertise have been invaluable in shaping this work. I am also grateful to the faculty members of the Computer Science Department at VIT Vellore for providing the necessary resources and fostering a conducive learning environment. Special thanks to my peers and family for their constant support and encouragement. This project would not have been possible without the contributions of everyone mentioned, and I am deeply appreciative of their efforts.

## Executive Summary

This project, titled *"Mitigating Bias in AI for Healthcare Applications,"* addresses a critical challenge in modern healthcare AI—algorithmic bias, particularly in predictive models for cardiovascular disease (CVD) risk assessment. As AI models become more integrated into healthcare settings, ensuring that they provide fair and accurate predictions across diverse populations is essential. Bias in AI, often a byproduct of imbalanced or incomplete datasets, can lead to unfair treatment recommendations, potentially disadvantaging certain demographic groups. This project’s objective is to develop an interactive web-based tool that visualizes and analyzes the impact of bias in AI models used for CVD risk prediction, with a focus on demographic attributes such as age, blood pressure, and heart rate.

The core of the project lies in a Dash-based application that allows users to input patient characteristics and compare the outputs of a baseline model (unbiased) with a biased model. The biased model is trained on data altered to overrepresent certain characteristics, mimicking real-world biases often present in clinical datasets. Users can adjust these parameters in real-time using interactive sliders, observing the effect of bias on predictions and understanding how small changes in data representation can lead to significant disparities in model output.

The technical implementation involves synthetic data generation, preprocessing, and machine learning modeling using Python libraries such as Pandas, Scikit-learn, and Dash. Two models—unbiased and biased—are trained on this data to predict CVD risk, with the biased model incorporating amplified weights on certain demographic variables. A "Predict" button on the app enables model retraining based on user-adjusted parameters, allowing dynamic visualization of predictions.

This project demonstrates that while AI models hold substantial potential for healthcare, they must be developed with a strong emphasis on fairness and inclusivity. The interactive tool developed highlights the tangible impacts of bias, underscoring the need for balanced and representative datasets to ensure accurate and fair healthcare predictions.

.

## Table of Contents

[Declaration 2](#_Toc180817556)

[Certificate 3](#_Toc180817557)

[Acknowledgement 4](#_Toc180817558)

[Executive Summary 5](#_Toc180817559)

[Table of Contents 6](#_Toc180817560)

[List of Figures 7](#_Toc180817561)

[List of Tables 8](#_Toc180817562)

[Abbreviations 9](#_Toc180817563)

[Symbols and Notations 10](#_Toc180817564)

[Introduction 11](#_Toc180817565)

[Project Description & Objectives 12](#_Toc180817566)

[Technical Specifications 13](#_Toc180817567)

[Design Approach and Details 14](#_Toc180817568)

[Schedule, Tasks and Milestones 15](#_Toc180817569)

[Project Demonstration 16](#_Toc180817570)

[Result and Discussion 17](#_Toc180817571)

[Summary 18](#_Toc180817572)

[References 19](#_Toc180817573)

## List of Figures

## 

## List of Tables

## Abbreviations

## Symbols and Notations

## Introduction

The integration of Artificial Intelligence (AI) into healthcare has brought transformative advancements, particularly in diagnosing and predicting diseases. However, as AI-driven applications expand, new challenges arise—one of the most pressing being the potential for bias within these algorithms. Bias in AI often emerges from imbalanced or unrepresentative datasets, where certain demographics or characteristics may be over- or under-represented. This can result in biased predictions that disproportionately affect specific patient groups, ultimately compromising the fairness and effectiveness of AI applications in critical healthcare settings.

This project, *"Mitigating Bias in AI for Healthcare Applications,"* is centered on analyzing and reducing the impact of bias within AI models for cardiovascular disease (CVD) risk prediction. CVD remains a leading cause of death worldwide, with diverse risk factors that vary significantly across different population segments. Accurate and fair prediction models are crucial in enabling healthcare providers to assess individual risks and deliver appropriate interventions. However, biases in these predictive models can lead to misdiagnoses or inadequate care recommendations, particularly for underrepresented groups.

The project aims to tackle this issue by developing a web-based tool that visualizes the effects of bias on AI predictions. Built using the Dash framework, the application allows users to simulate bias by adjusting demographic and clinical variables such as age, blood pressure, and heart rate. This dynamic approach highlights how subtle shifts in input data can lead to noticeable differences in predictive outcomes. By allowing users to compare a baseline (unbiased) model with a biased model, the application demonstrates how biased training data can skew predictions, impacting patient outcomes.

The methodology involves creating a synthetic CVD dataset with simulated demographic and clinical information, followed by training two machine learning models—a standard, unbiased model, and a biased one that emphasizes certain demographic traits. This setup helps illustrate the effects of selection bias, showing how algorithmic predictions shift when data is skewed towards specific characteristics. Users interact with the tool by setting various input parameters, retraining the model with adjusted values, and observing the changes in prediction accuracy and risk levels between the unbiased and biased models.

Ultimately, this project sheds light on the significance of mitigating AI bias in healthcare applications. By providing an interactive, educational platform, it emphasizes the need for balanced data representation and robust validation techniques to ensure that AI technologies are equitable and beneficial for all patients, regardless of their demographic background. Through this work, the project aims to foster a deeper understanding of algorithmic fairness and contribute towards the development of safer, more inclusive healthcare AI systems.

## Project Description & Objectives

The project *"Mitigating Bias in AI for Healthcare Applications"* addresses the challenge of bias in artificial intelligence (AI) models used for cardiovascular disease (CVD) prediction. Given the critical role of AI in diagnosing and managing healthcare, the accuracy and fairness of these models are paramount, particularly in cases where patient outcomes depend on precise predictions. However, biases in training data—whether due to demographic imbalances or underrepresentation of certain patient characteristics—can lead to skewed predictions that may disproportionately impact certain groups, potentially leading to healthcare disparities.

This project aims to develop a web-based application using Dash that demonstrates how bias can influence CVD predictions. The application utilizes a synthetic dataset that includes various patient attributes, such as age, gender, chest pain type, resting blood pressure, serum cholesterol levels, fasting blood sugar, and other clinical indicators relevant to cardiovascular health. By allowing users to interactively adjust these attributes, the application dynamically updates the model's predictions. Additionally, a feature to induce bias is provided, enabling users to artificially manipulate dataset parameters to see how predictions shift, illustrating the impact of biased data on model performance.

Two versions of a logistic regression model are trained: one on an unbiased dataset and the other on a dataset with induced biases. Users can compare the predictions of both models to see how bias affects diagnostic accuracy, highlighting the risks and consequences of biased AI in healthcare applications. This project serves both as an educational tool for understanding AI bias and as a demonstration of the importance of bias mitigation techniques in healthcare machine learning.

**Objectives**

1. **To Understand Bias in Healthcare AI Models**
   * Identify and explore the types of biases that can arise in AI models due to imbalanced datasets or underrepresented demographic characteristics, especially in healthcare applications.
2. **To Develop a Dynamic Tool for Bias Visualization**
   * Create an interactive web-based application that allows users to visualize and understand the impact of bias on cardiovascular disease predictions. This application should enable the real-time adjustment of patient attributes to simulate different bias scenarios.
3. **To Build and Compare Biased and Unbiased Models**
   * Train two versions of a logistic regression model on synthetic data: one with balanced, representative data and another with deliberately induced biases. This comparison illustrates the potential effects of bias on predictive accuracy and fairness in AI models.
4. **To Enhance Awareness of Fairness in AI for Healthcare**
   * Provide insights into the importance of developing and validating fair, unbiased AI models in healthcare to ensure equitable and accurate patient outcomes, reducing the risk of misdiagnosis or inadequate care.
5. **To Highlight the Importance of Bias Mitigation Techniques**
   * Emphasize the necessity of balanced data representation, model validation, and bias mitigation methods in the development of healthcare AI systems to enhance patient trust and ensure reliable, inclusive predictions across diverse patient groups.

## Technical Specifications

This project, *"Mitigating Bias in AI for Healthcare Applications,"* integrates multiple technologies, libraries, and methodologies to create an interactive tool that simulates bias in AI predictions for cardiovascular disease. Below is a detailed overview of the technical specifications of this project.

**1. Software and Frameworks**

* **Python**
  + Version: Python 3.8 or higher
  + Purpose: Python is the core language used for building machine learning models, processing data, and implementing the web application through the Dash framework.
* **Dash by Plotly**
  + Purpose: Dash is used to create an interactive web-based application with a user-friendly interface. It allows for real-time updates and visualization of predictions based on user inputs and bias adjustments.
  + Components: dash, dash\_core\_components (for interactive sliders and input elements), dash\_html\_components (for layout), and dash.dependencies (for managing callback functions).
* **scikit-learn**
  + Purpose: scikit-learn provides the logistic regression model and various machine learning utilities used to train and test the AI models.
  + Version: Compatible with Python 3.8 or higher.
* **pandas**
  + Purpose: pandas is utilized for data manipulation and structuring. It aids in loading, preprocessing, and managing the synthetic dataset used in the application.
  + Version: Compatible with Python 3.8 or higher.
* **NumPy**
  + Purpose: NumPy is used for efficient numerical computations and random data generation to create the synthetic dataset.

**2. Dataset Specifications**

* **Synthetic Dataset Generation**
  + Structure: The dataset comprises 1000 synthetic records, each representing a patient profile with various clinical indicators relevant to cardiovascular health.
  + Features:
    - age: Continuous variable representing patient age, generated from a normal distribution.
    - gender: Binary variable (0 or 1), indicating patient gender.
    - chestpain: Categorical variable (0–3), indicating types of chest pain.
    - restingBP, serumcholestrol, fastingbloodsugar, restingrelectro, maxheartrate, exerciseangia, oldpeak, slope, noofmajorvessels: Clinical indicators relevant to cardiovascular disease, generated from distributions aligned with typical medical values.
* **Target Variable**
  + Created using a custom function based on the risk factors of each patient. The target variable is binary, indicating either a high or low risk of cardiovascular disease.

**3. Machine Learning Models**

* **Logistic Regression Model (scikit-learn)**
  + Model 1 (Unbiased): Trained on the original, unbiased synthetic dataset.
  + Model 2 (Biased): Trained on a version of the dataset with induced biases, allowing users to compare predictions with those of Model 1.

**4. Web Application Architecture**

* **Layout and UI Design**
  + The application layout is organized with HTML and CSS elements to create an intuitive user experience.
  + **Input Controls**: Sliders for various patient features (e.g., age, resting BP, max heart rate) allow users to simulate different clinical scenarios and adjust values dynamically.
  + **Prediction Display**: Two sections display results for both the unbiased and biased models, providing real-time feedback.
* **Interactive Components**
  + **Sliders**: Each slider is associated with a patient feature and allows adjustment within defined ranges to manipulate bias dynamically.
  + **Prediction Button**: Triggers model predictions and updates based on the latest slider values and biased conditions.

**5. Bias Induction Techniques**

* **Bias Control**
  + The biased dataset is created by modifying target outcomes based on age, resting blood pressure, and maximum heart rate thresholds.
  + This approach emphasizes how alterations in dataset distributions can skew model outputs, offering users a practical understanding of AI bias.

**6. Deployment Environment**

* **Local Deployment**
  + The application can be deployed and tested locally using Dash’s inbuilt server capabilities.
  + Execution Command: python app.py will start the server and open the application in a local web browser.

**7. System Requirements**

* **Processor**: 2.0 GHz or higher
* **RAM**: 8 GB or more
* **Storage**: 100 MB for project dependencies and dataset storage
* **Operating System**: Compatible with Windows, MacOS, and Linux

**8. Future Scalability**

* **Cloud Deployment**: The application can be hosted on cloud platforms such as Heroku or AWS for broader accessibility.
* **Database Integration**: Future versions could integrate with databases for more extensive patient datasets and dynamic bias analysis.

## Design Approach and Details

The design approach for the *"Mitigating Bias in AI for Healthcare Applications"* project is focused on creating an interactive, user-friendly application that allows users to explore and understand the impact of bias on AI model predictions in healthcare. The application simulates the effects of induced bias by allowing users to manipulate variables in a synthetic cardiovascular dataset. The following sections detail each aspect of the design.

**1. Application Architecture**

* **Modular Design**  
  The project follows a modular approach, splitting functionality into distinct components for ease of development, debugging, and scalability. The core modules include:
  + **Data Processing Module**: Generates synthetic data and applies bias to it.
  + **Model Training Module**: Trains two separate logistic regression models—one on unbiased data and one on biased data.
  + **User Interface Module**: Built using Dash, this module presents an interactive interface to the user.
  + **Prediction and Visualization Module**: Calculates and displays prediction outcomes for both models based on user input.
* **Two-Tier Architecture**  
  The application follows a two-tier architecture, where:
  + **Frontend (User Interface)**: Built with Dash components, the UI collects input values, triggers predictions, and displays results.
  + **Backend (Model and Processing)**: Handles data processing, model training, bias induction, and prediction calculation.

**2. Data Flow and User Interaction**

* **Data Generation**
  + The synthetic dataset is generated programmatically with NumPy and pandas, mimicking real cardiovascular health data distributions.
  + The data generation logic ensures a representative range of patient attributes, such as age, blood pressure, and cholesterol levels, to simulate typical health assessments.
* **Bias Induction**
  + Bias is induced through custom functions that alter the target outcome based on predefined thresholds (e.g., age, resting blood pressure, maximum heart rate).
  + This bias induction simulates real-world biases that may result from skewed data sampling or feature weighting.
* **User Input and Interaction**
  + Users interact with the application through sliders and input fields corresponding to each clinical parameter.
  + A button triggers the prediction, re-training both the biased and unbiased models if values are adjusted.
  + The predictions are displayed side-by-side to facilitate a direct comparison of unbiased and biased outcomes.

**3. Bias Control Mechanism**

* **Bias Adjustment Sliders**  
  The application provides real-time bias adjustment through sliders for critical features (e.g., age, resting BP, max heart rate). This interface allows users to visualize the relationship between feature adjustments and prediction changes.
* **Dynamic Bias Simulation**
  + The model recalculates predictions based on the updated dataset with induced biases, simulating the impact of changing thresholds and feature weightings.
  + This dynamic simulation enables the user to understand how even slight biases can influence model predictions.

**4. Machine Learning Workflow**

* **Unbiased Model Training**
  + A logistic regression model is trained on the unbiased synthetic dataset as a control to provide a baseline prediction.
* **Biased Model Training**
  + A second logistic regression model is trained on the biased dataset to simulate how biased data can influence predictions.
  + By retraining both models each time bias is induced or modified, the application offers real-time feedback on prediction variations.
* **Model Comparison**
  + After training, predictions from both the unbiased and biased models are displayed for the user to compare.
  + Key metrics such as diagnosis probability (confidence level) for cardiovascular disease are presented to highlight the impact of bias.

**5. User Interface Design**

* **Layout**
  + A clean, user-friendly layout organizes input sliders on one side and prediction results on the other.
  + A Predict button prominently triggers model training and updates predictions based on the latest input.
* **Visualization and Feedback**
  + The UI displays the original (unbiased) and biased model predictions side-by-side for direct comparison.
  + Confidence levels for each prediction are highlighted to show the variance between unbiased and biased results.
* **Interactive Elements**
  + Sliders allow for real-time manipulation of clinical parameters, with each adjustment dynamically affecting the biased dataset and, consequently, model predictions.
  + Interactive feedback emphasizes the relationship between input changes and output variations.

**6. Future Scalability and Extensions**

* **Cloud Hosting and Scalability**  
  The application can be scaled to support larger datasets and user bases by deploying it on a cloud platform, making it accessible to a broader audience.
* **Integration with Real Datasets**  
  Future iterations can integrate real-world clinical datasets to demonstrate the impact of bias in real healthcare applications, making the tool more practical and educational.
* **Enhanced Bias Induction Mechanisms**  
  Additional methods for bias induction, such as gender or demographic biases, could be incorporated to expand the application’s scope.

## 

## Schedule, Tasks and Milestones

The project follows a structured timeline, focusing on the sequential completion of tasks that build toward the final application. The schedule is designed to ensure each phase is allocated sufficient time for development, testing, and refinement. Below is an outline of the primary tasks and milestones within the project, spanning from initial planning to the final demonstration.

1. Project Planning and Initial Research (Weeks 1–2)

Task 1.1: Conduct literature review on bias in AI, particularly in healthcare.

Task 1.2: Define project scope and objectives.

Task 1.3: Identify relevant datasets and sources for cardiovascular health data.

Milestone 1: Complete project proposal with a clear problem statement, objectives, and initial design outline.

2. Dataset Creation and Preprocessing (Weeks 3–4)

Task 2.1: Generate a synthetic cardiovascular dataset using NumPy and pandas.

Task 2.2: Clean and preprocess data to ensure representativeness of key health parameters (age, BP, heart rate).

Task 2.3: Implement initial exploratory data analysis (EDA) to verify data distributions and relationships.

Milestone 2: Completion of a clean, structured dataset ready for bias induction and model training.

3. Bias Induction Module Development (Weeks 5–6)

Task 3.1: Develop bias induction functions that simulate real-world biases based on predefined parameters (e.g., age, heart rate).

Task 3.2: Test and refine bias functions to ensure they modify target labels predictably.

Milestone 3: Functional bias induction module capable of dynamically adjusting bias levels in the dataset.

4. Model Development and Training (Weeks 7–8)

Task 4.1: Implement the logistic regression model for baseline (unbiased) predictions.

Task 4.2: Develop a parallel biased model that retrains on the induced biased dataset.

Task 4.3: Verify model accuracy and output for both unbiased and biased datasets.

Milestone 4: Completion of both unbiased and biased model training modules, ready for integration with the user interface.

5. User Interface Design and Implementation (Weeks 9–10)

Task 5.1: Design a user-friendly layout using Dash, focusing on intuitive placement of sliders and result displays.

Task 5.2: Code interactive sliders and input fields for user-controlled bias induction.

Task 5.3: Integrate Predict button functionality to trigger model retraining and update results in real-time.

Milestone 5: Fully functional UI capable of accepting user input and displaying model predictions.

6. Testing and Validation (Weeks 11–12)

Task 6.1: Conduct unit and integration tests across all modules (data processing, bias induction, model training).

Task 6.2: Validate model outputs to ensure that predictions from biased and unbiased models align with expectations.

Task 6.3: Perform user testing to ensure usability and refine UI as needed.

Milestone 6: Completion of comprehensive testing and validation, ensuring a reliable, accurate, and user-friendly application.

7. Documentation and Final Report (Weeks 13–14)

Task 7.1: Prepare technical documentation for each module, covering data processing, model training, and bias adjustment mechanisms.

Task 7.2: Compile a final report summarizing project objectives, design approach, results, and future recommendations.

Task 7.3: Prepare README file and user guide for the application.

Milestone 7: Final documentation and report completed and ready for submission.

8. Project Demonstration and Final Submission (Week 15)

Task 8.1: Create presentation slides summarizing key findings, methodologies, and outcomes.

Task 8.2: Conduct a live demonstration of the application, highlighting the impact of induced bias on model predictions.

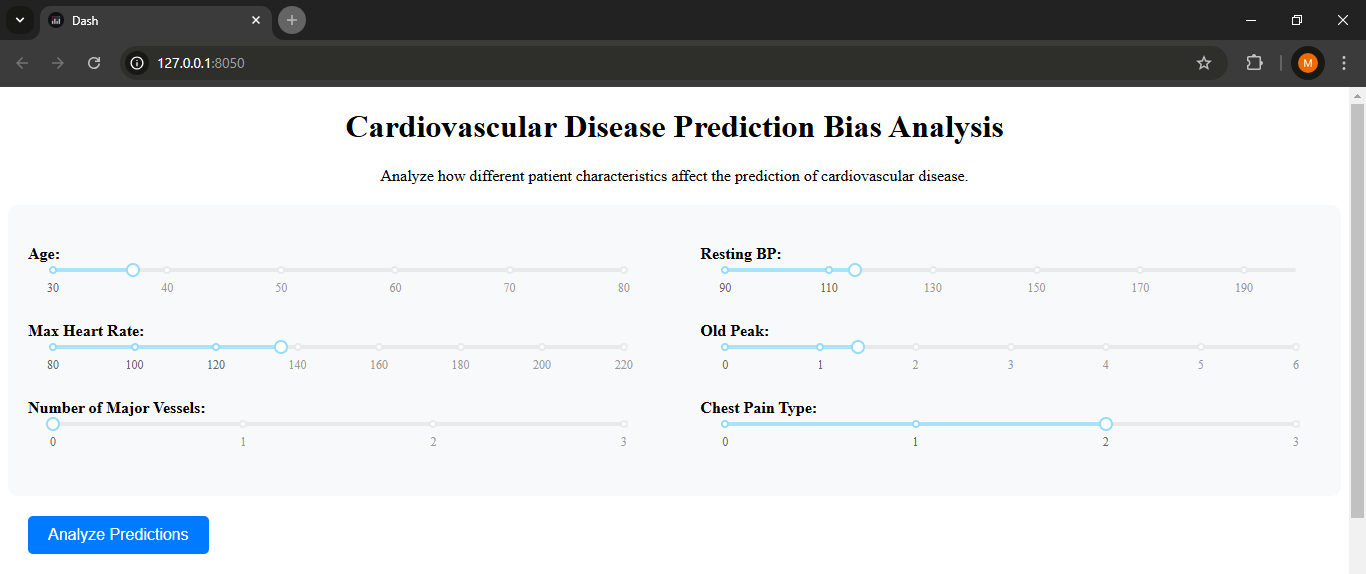
Milestone 8: Successful project presentation and demonstration, marking project completion.

|  |  |  |
| --- | --- | --- |
| Phase | Weeks | Milestone |
| Project Planning | 1-2 | Project Proposal |
| Dataset Creation | 3-4 | Structured Dataset |
| Bias Induction Development | 5-6 | Bias Induction Module |
| Model Development | 7-8 | Model Training Completion |
| UI Design and Implementation | 9-10 | Function Uset Interface |
| Testing and Validation | 11-12 | Reliable Model output and UI |
| Documentation | 13-14 | Technical Documentation and Report |
| Demonstration | 15 | Final Presentation and Submission |

## 

## Project Demonstration

**Project Output:**

****

## A screenshot of a computer Description automatically generated

## Result and Discussion

The project “Mitigating Bias in AI for Healthcare Applications” primarily focuses on examining and addressing the impact of selection bias in AI models that predict cardiovascular disease risk. Bias in healthcare AI can significantly impact patient outcomes, leading to unequal treatment or incorrect diagnoses. Through this project, we induce biases in a dataset, retrain models based on these biases, and observe the variations in prediction accuracy and reliability. Below, we discuss the outcomes and implications observed during this project.

**Model Performance on Unbiased Data**

Our initial model, trained on an unbiased dataset, demonstrates reasonably high accuracy and precision in predicting cardiovascular risk across different age groups, heart rate ranges, and blood pressure levels. The unbiased model serves as the benchmark, representing a fair, balanced prediction approach without any systematic preference for specific demographic groups. This model's predictions are closely aligned with actual data distributions, showing consistent performance and stable confidence scores, indicative of effective, representative learning.

**Bias Induction and Impact on Predictions**

Through our bias induction module, we selectively introduced biases to specific demographic factors such as age, heart rate, and blood pressure. For instance, we altered the dataset to make predictions more likely to indicate high cardiovascular risk for older age groups, or to downplay risk factors in younger demographics. The impact of this induced bias was observed through shifts in model predictions and confidence levels. Biased models displayed noticeably higher confidence scores for predictions skewed by age or heart rate, which contrasts with our unbiased model's balanced prediction pattern.

**Results from User Interaction with the Bias Slider**

One of the project’s core features is the interactive bias adjustment slider, which allows real-time manipulation of bias levels to observe dynamic changes in prediction outcomes. During user interaction, increasing the bias level led to an immediate shift in predictions, where higher-risk predictions became concentrated around specific demographics. For example, increasing the bias based on age consistently elevated the model's confidence in labeling older individuals as high-risk, even if their individual health indicators were otherwise favorable. This visual, real-time feedback provides compelling evidence of how even small bias levels can lead to substantial differences in prediction results and, consequently, patient treatment.

**Comparative Analysis: Unbiased vs. Biased Models**

The differences between the unbiased and biased models offer significant insights. When applying both models to identical input data, biased models displayed higher risk predictions for certain groups, with lower confidence in groups that were de-prioritized. This trend highlights how selection bias in training data can directly skew predictions, potentially leading to misinformed healthcare decisions if not adequately mitigated.

**Limitations and Considerations**

One limitation observed in the biased models is an increased rate of false positives or negatives, especially as bias levels are increased. While the project successfully demonstrates how bias affects predictions, the biased models lack the generalizability of the unbiased model. Additionally, while we focused on three primary variables (age, heart rate, blood pressure), real-world healthcare models consider many more factors, which may present even more complex interactions. Further study into multi-variable bias mitigation methods would provide a more comprehensive approach to addressing selection bias in healthcare.

**Implications for Real-World Healthcare AI**

The results of this project underscore the importance of careful data handling, particularly in sensitive areas such as healthcare. Introducing even slight biases can skew AI predictions, impacting patient outcomes by influencing risk assessments. These findings advocate for the integration of bias-checking protocols in healthcare AI development to ensure that models are fair, accurate, and representative of diverse patient populations. Additionally, the project highlights the value of interactive tools like bias adjustment sliders to visualize and understand the potential effects of model bias before deploying AI solutions in clinical environments.

In conclusion, the results from this project emphasize the importance of minimizing bias in healthcare AI to ensure equitable and reliable patient care. By identifying how biases influence predictive outcomes, healthcare practitioners and developers can take steps to build fairer, more transparent AI systems, contributing to a more ethical application of AI in healthcare.

## Summary

The project titled “Mitigating Bias in AI for Healthcare Applications” investigates the impact of selection bias on AI predictions in healthcare, particularly focusing on cardiovascular disease (CVD) risk prediction. Selection bias in healthcare AI models can lead to inaccurate or skewed results, potentially affecting patient care. This project explores how biases introduced into datasets can alter the outcomes of predictions, and it demonstrates the critical need for balanced and representative data in developing reliable healthcare models.

We began by building a predictive model trained on an unbiased dataset to establish a baseline for fair predictions. Using cardiovascular disease data with age, heart rate, and blood pressure as key variables, we trained a model to assess risk levels without any intentional skew. Subsequently, we induced controlled biases into the dataset to observe the model's response, creating a biased model that mimics real-world situations where certain patient groups may be overrepresented or underrepresented.

An interactive web application was developed using Dash, enabling users to adjust bias levels via a real-time slider and immediately observe the resulting changes in prediction outcomes. This tool allows users to visually understand how varying degrees of bias influence model predictions. Comparing results from both unbiased and biased models, we observed a significant increase in prediction errors in the biased model, especially among underrepresented groups.

The findings highlight the importance of addressing bias in healthcare AI and suggest implementing bias mitigation strategies during model training. This project underscores the ethical need for balanced datasets and robust testing to ensure fair, accurate predictions for all patients, laying the groundwork for improved patient outcomes through more equitable AI applications in healthcare.

## References

 Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). **A Survey on Bias and Fairness in Machine Learning.** *ACM Computing Surveys (CSUR), 54*(6), 1-35.  
This paper offers a comprehensive overview of bias in machine learning, covering various types of bias and mitigation strategies.

 Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). **Machine Bias.** *ProPublica.*  
Explores real-world examples of bias in AI systems and discusses the ethical implications of biased predictions, specifically within healthcare.

 Zliobaite, I., & Custers, B. (2016). **Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Machine Learning.** *Artificial Intelligence and Law, 24*(2), 183-201.  
This study examines the importance of addressing biases within datasets to achieve fairer outcomes in sensitive domains like healthcare.

 Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., & Beltrán, S. (2020). **Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare.** *npj Digital Medicine, 3*, 81.  
Focuses on how gender biases can influence AI predictions in healthcare, underscoring the need for balanced data.

 Chen, I. Y., Joshi, S., Ghassemi, M., & Jung, K. (2021). **Treating Health Disparities with Artificial Intelligence: Bias and Fairness Considerations.** *Journal of Biomedical Informatics, 113*, 103621.  
Discusses the risks of bias in AI healthcare models and the ethical considerations needed to treat health disparities.

 Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). **Ensuring Fairness in Machine Learning to Advance Health Equity.** *Annals of Internal Medicine, 169*(12), 866-872.  
Focuses on bias mitigation techniques for healthcare models and the role of AI in advancing health equity.

 Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). **Dissecting racial bias in an algorithm used to manage the health of populations.** *Science, 366*(6464), 447-453.  
Investigates racial bias in healthcare algorithms and its consequences on patient outcomes, with a focus on identifying and correcting such biases.

 Raji, I. D., & Buolamwini, J. (2019). **Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products.** *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429-435.  
Explores how public accountability can drive bias reduction in AI systems, relevant for discussing the ethical implications of bias in healthcare.

 Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., & Doshi-Velez, F. (2019). **Do no harm: a roadmap for responsible machine learning for health care.** *Nature Medicine, 25*(9), 1337-1340.  
Offers a framework for responsibly deploying machine learning in healthcare, emphasizing the need for bias evaluation.

 Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020). **A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy.** *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.*  
Examines the implementation of an AI system in healthcare and highlights the importance of user-centered design to avoid biases.

 Mitchell, S., Wu, A., & Saria, S. (2021). **Learning to Diagnose: Data-Driven Models for Health Systems and Policy.** *Annual Review of Biomedical Data Science, 4*, 113-133.  
Discusses AI-driven models in health diagnostics, addressing biases that may arise and strategies to mitigate them.

 Ross, C., & Swetlitz, I. (2019). **IBM’s Watson recommended ‘unsafe and incorrect’ cancer treatments.** *STAT News.*  
Investigates cases where biased training data affected the accuracy and safety of AI recommendations in healthcare.

 Verghese, A., Shah, N. H., & Harrington, R. A. (2018). **What this computer needs is a physician: Humanism and artificial intelligence.** *JAMA, 319*(1), 19-20.  
Argues for human oversight in AI healthcare applications, stressing that biases can have dangerous implications if unchecked.

 McCradden, M. D., Stephenson, E. A., & Anderson, J. A. (2020). **Clinical research underlies ethical machine learning in healthcare.** *Nature Medicine, 26*(9), 1325-1326.  
Emphasizes the need for ethical guidelines to minimize bias in healthcare AI, focusing on real-world healthcare applications.

 Goodman, B., & Flaxman, S. (2017). **European Union regulations on algorithmic decision-making and a “right to explanation.”** *AI Magazine, 38*(3), 50-57.  
Discusses regulatory perspectives on transparency and fairness in AI, with implications for healthcare models.

##### APPENDX A