An Adversarial Framework for Mitigating Gender Bias in Coronary Heart Disease Prediction

Diego Silva
d1silva@ucsd.edu

Patrick Salsbury
psalsbury@ucsd.edu

Kai Ni
c5ni@ucsd.edu

Emily Ramond
eramond@deloitte.com

Abstract

This study investigates the mitigation of gender bias in machine learning models for Coronary Heart Disease (CHD) diagnosis. Utilizing data from the National Health and Nutrition Survey (NHANES), we develop a neural network (NN) model employing an adversarial configuration to address disparities in CHD prediction. The primary NN model for CHD classification is coupled with a secondary discriminator model designed to penalize gender-based biases. Our results demonstrate a significant reduction in bias compared to conventional approaches while maintaining clinically relevant accuracy. Although a marginal decrease in overall predictive performance was observed, the model efficiency remained comparable to existing methodologies. This research contributes to the growing body of literature on the ethical use of artificial intelligence and machine learning in healthcare, offering a novel approach to mitigating demographic biases in cardiovascular disease diagnostics. The findings suggest potential for improving equitable healthcare outcomes, particularly for underrepresented patient populations in CHD diagnosis and treatment.

Website: https://chd-adversarial-nn.github.io
Code: https://github.com/patsals/CHD-adversarial-nn

1 Introduction

1.1 Intro

Coronary Heart Disease (CHD) is a type of Cardiovascular Disease historically associated with significant gender disparities in the diagnosis and treatment of patients. In our review of existing literature, we found studies showing that CHD can be efficiently and effectively classified using Convolutional Neural Networks (CNNs), but the fairness and bias of these models have not been adequately explored or addressed. In general, while the intersection of machine learning and healthcare has shown promising advancements in patient classification and diagnosis, the emergence of biases in these models, especially those discriminating against underrepresented groups based on sex, gender, race, and socioeconomic status, raises ethical concerns.

In this project, we worked with health statistics data from the National Health and Nutrition Survey (NHANES) to train and develop a Neural Network (NN) model that can accurately predict CHD among patients. From there, we worked to detect and mitigate potential biases in the diagnosis and classification process. Ultimately, our goal was to develop and implement a fair and unbiased algorithm to classify patients with CHD at a comparable accuracy to existing models. Our approach leverages neural networks (NNs) as a primary model to predict CHD. In order to mitigate bias, we employed an adversarial configuration and utilize a secondary "discriminator" model to detect and penalize any potential bias, focusing particularly on gender-based biases. This methodology encourages the primary model to focus on relevant diagnostic features while discouraging reliance on sensitive attributes thereby promoting more equitable healthcare outcomes.

1.2 Literature Review

In preliminary research, we found several insightful papers discussing the gender disparities between male and female patients in the treatment and diagnosis of not only Coronary Heart Disease (CHD) but cardiovascular diseases (CVD) at large.

To start, a systematic review conducted by Al Hamid et al. (2024) provides evidence supporting the need for fairness in our development of models to predict CHD. Through a comprehensive analysis of 19 studies, Al Hamid et al. reveal significant gender disparities in the diagnosis, prevention, and treatment of cardiovascular diseases CVDs. Their findings indicate that women were less likely to be diagnosed with CVDs, received fewer diagnostic tests such as coronary angiography and ECGs, and were prescribed fewer cardiovascular medicines compared to men. This gender bias in clinical practice could potentially be reflected in the NHANES data we are using, highlighting the importance of carefully examining gender distribution in our dataset. The study's observations on differences in symptom presentation and risk factor consideration between genders aligns with our model, which incorporates gender-specific features to improve accuracy and fairness in predicting CHD across demographics.

Schenck-Gustafsson (2009)'s study on risk factors for CVD in women showcases the contrast of CVD being the most common cause of death in women, while it is often neglected in women's health management. Schenck-Gustafsson identifies key risk factors for CVD in women, including dyslipidemia, hypertension, smoking, stress, diabetes, obesity (especially abdominal fat distribution), physical inactivity, and poor eating habits. Importantly, the study points out unique risk factors for women, such as older age at presentation and higher likelihood of co-morbidities like diabetes and hypertension.

Maserejian et al. (2009)'s study on disparities in physicians' interpretations of heart disease symptoms by patient gender offers crucial insights into the challenges of diagnosing CHD in women. Through a factorial experiment using videotaped CHD symptoms, the researchers systematically altered patient characteristics and examined physicians' diagnostic decisions. The study reveals significant gender-based disparities in clinical decision-making:

1. Physicians were less certain about the underlying cause of symptoms in female patients, regardless of age.
2. For middle-aged women specifically, physicians showed significantly less certainty in diagnosing CHD.
3. Mental health conditions were more frequently considered as the most certain diagnosis for middle-aged women (31.3%) compared to their male counterparts (15.6%).
4. An interaction effect indicated that high-SES females were most likely to receive a mental health diagnosis as the most certain.

These findings highlight a concerning trend where middle-aged female patients, particularly those of high socioeconomic status, are at risk of misdiagnosis. The tendency to attribute symptoms to mental health conditions in these cases could lead to delayed or missed CHD diagnoses – our model could potentially mitigate these disparities and improve diagnostic accuracy across all demographic groups.

Beery (1995)'s study on gender bias in coronary artery disease diagnosis and treatment highlights a significant disparity in cardiovascular care, noting that women often receive fewer referrals for diagnostic and therapeutic procedures despite being at high risk for cardiovascular disorders. This bias is particularly concerning as many current procedures and therapies were developed primarily for men, potentially limiting their efficacy for women. The study reveals that women typically undergo angioplasty or bypass grafting at more advanced ages and in poorer health conditions, receive fewer advanced treatments like implantable cardioverter defibrillators and heart transplants, and experience poorer outcomes overall. These findings underscore the importance of our project's focus on incorporating gender-specific features and mitigating bias in our model.

Narrowing in on the referral system as an avenue for bias to manifest, we found a particularly interesting paper highlighting the referral system utilized by primary care physicians (PCPs).

In a study on determinants of referral for suspected coronary artery disease, Winkler et al. (2023) explore PCPs' referral decisions and the factors influencing these decisions, using a sample of 26 cases from nine practices in Hesse, Germany. Their findings reveal that referral decisions are influenced by various factors beyond patient characteristics, including

practice environment, PCP-related factors, and non-diagnostic patient characteristics. The study highlights the complexity of the referral process, with PCPs considering factors such as proximity to specialist practices, relationships with colleagues, and concerns about overtreatment. Notably, the authors found that most PCPs were unaware of formal guidelines and relied on informal local consensus, which was largely influenced by specialists. This reliance on informal consensus rather than standardized guidelines indicates potential value of our model in providing more consistent, evidence-based support for referral decisions.

Finally, we reviewed two papers more directly related to our project scope discussing the potential benefits and pitfalls of using AI/ML techniques in predicting patient diagnoses.

Mihan, Pandey and Van Spall (2024)'s paper on mitigating AI bias in cardiovascular care demonstrates that AI algorithms, while transformative in cardiovascular healthcare delivery, can introduce and perpetuate biases when trained on homogeneous data or inequitable healthcare processes. This bias can manifest at various stages: algorithm development, testing, implementation, and post-implementation. The consequences of such algorithmic bias are significant, potentially leading to missed diagnoses, disease misclassification, incorrect risk prediction, and inappropriate treatment recommendations. Importantly, these adverse effects disproportionately impact marginalized demographic groups, exacerbating existing health disparities. Mihan et al. propose strategies to mitigate bias during AI algorithm training, testing, and implementation, emphasizing the need for an AI health equity framework.

Dutta et al. (2020)'s efficient neural network (NN) for CHD serves as a key reference for our project, despite some differences in approach. The authors propose a convolutional neural network (CNN) model to classify highly imbalanced clinical data for CHD prediction. While our project focuses on traditional NNs rather than CNNs, several aspects of their methodology are noteworthy:

1. The authors address the challenge of class imbalance, a common issue in medical datasets, which our project must also consider.
2. They employ a two-step approach, first using LASSO for feature selection, followed by homogenization of important features through a fully connected layer. We consider this in our data cleaning and processing steps.
3. Their model achieves a balanced accuracy of 79.5%, outperforming traditional machine learning methods like SVM and random forest. We will consider this in reference, although we will expect notably lower figures in exchange for improved fairness and reduced bias.

Our project's adversarial configuration, with a primary NN model and a rival "discriminator" model, presents a novel approach to ensuring fairness and reducing reliance on sensitive features – an aspect not addressed in Dutta et al.'s study. We implement this adversarial setup to enhance our model's ability to provide unbiased predictions while maintaining reasonable accuracy across all demographic groups.

Our review of existing literature accumulated ample evidence showcasing the persistent gender disparities in CHD diagnosis and treatment, and points toward the potential for AI/ML models to both perpetuate and mitigate these biases. These studies emphasize the

need for our project's focus on developing an unbiased model by incorporating gender-specific features, addressing data imbalances, and implementing an adversarial configuration. Ultimately, our approach aims to contribute to more equitable and accurate CHD prediction across all demographic groups.

1.3 Data Description

We acquired our data from the National Health and Nutrition Survey (NHANES) conducted by the Nation Center For Health Statistics (NCHS), a unit of the Centers for Disease Control and Prevention (CDC).

The NHANES collects data to understand the health of adults and children in the United States. It is a comprehensive survey that includes data on participant dietary habits, supplements, and blood work. As part of their survey, participants undergo health exams, laboratory tests, and nutritional interviews. Since 1999, the NCHS has conducted a continuous survey collecting data from 5,000 participants, including adults and children, in different communities throughout the United States. The NCHS states that they follow a "random, scientific process to select the people [they] invite to participate. This process ensures that this group of people can accurately represent the health and nutritional status of everyone in our diverse nation."

To begin the data collection process, the NCHS notifies local governments of each location about an upcoming survey. The households in the community receive a notice about the survey from the NCHS director. The NCHS then sends teams of nutrition and health interviewers, nurses, and health technicians to the community. Health and diet interviews are conducted in participants' homes, while health exams are conducted in mobile exam centers. The NCHS uses an advanced computer system to track and store their data throughout the process.

Each year, six main categories of data are collected: demographics, dietary, examination, laboratory, questionnaire, and limited access data. For this project, we focus only on demographics, laboratory, and questionnaire data. From these collections, we selected a total of 35 variables to create our dataset for this project. Demographic variables include age and gender. Laboratory variables include iron, glucose, protein, uric acid, creatinine, etc. Lastly, the questionnaire variables include questions about frequency of moderate work and vigorous work, diabetes, coronary heart disease, blood-related stroke, etc. Reference Table 1 on page 6 for the full table of variables and relevant descriptions and units where applicable.

We used this dataset to train our neural network model for Coronary Heart Disease classification with an adversarial model based on SGD Classifier to mitigate gender bias in the neural network.

Table 1: Full list of variables

Attribute	**Description**
Gender	Male/Female
Age	Years
Systolic	Systolic: Blood pressure (first reading) mm Hg
Diastolic	Diastolic: Blood pressure (first reading) mm Hg
Weight	Pounds (lbs)
Body mass index	
White blood cells	White blood cell count: SI
Basophils	Basophils number
Red blood cells	Red cell count SI
Hemoglobin	Hemoglobin (g/dL)
Platelet count	Platelet count (%) SI
Mean volume of platelets	Mean platelet volume (fL)
Red blood cell width	Red cell distribution width (%)
Aspartate aminotransferase (AST)	AST (U/L) AST: SI (U/L)
Alanine aminotransferase (ALT)	ALT (U/L) ALT: SI (U/L)
Creatinine	Creatinine (umol/L)
Glucose	Glucose (mg/dL)
Gamma-glutamyl transferase (GGT)	GGT (U/L) GGT: SI (U/L)
Iron	Iron (umol/L)
Lactate dehydrogenase (LDH)	LDH (U/L)
Phosphorus	Phosphorus (mmol/L)
Bilirubin	Bilirubin, total (umol/L)
Protein	Protein, total (g/L) Total protein (g/L)
Uric acid	Uric acid (umol/L)
Triglycerides	Triglycerides (mmol/L)
Albumin	Albumin (g/L)
Alkaline phosphatase (ALP)	Alkaline phosphotase (U/L)
High-density lipoprotein (HDL)	HDL-cholesterol (mmol/L)
Cholesterol	Total cholesterol (mmol/L)
Glycohemoglobin	Glycohemoglobin (%)
Vigorous-work	How often did you do tasks requiring vigorous effort in the last 30 days?
Moderate-work	How often requiring moderate effort?
Diabetes	Yes/No
Blood related diabetes	Were any of your close biological ever told by a health professional that they had diabetes?
Blood related stroke	Yes/No
Coronary heart disease	Yes/No

2 Methods

2.1 Dataset

We compiled our dataset using the publicly available NHANES API. To account for inconsistencies in formatting across a multiple-year range of records, we referenced the official documentation to track renaming of features between years.

We observed significant missingness of around 40% of observations in various features, and notably almost 85% in history of strokes in blood-related family members. We took a minimal approach in handling missing values using a combination of dropping certain features with effectively unusably high amounts of missing values and imputing mean values for features where fewer observations record nulls.

Our final dataset retains around 37,000 observations.

Additional pre-processing was conducted on the final dataset prior to model development and training. First, numerical variables were normalized to ensure gradient convergence, stabilize gradients to avoid exploding or vanishing, and ensure that all features are considered proportionally on the same scale. Categorical variables such as whether a patient has Diabetes or CHD were converted to numerical variables using one-hot encoding.

It is important to note that our dataset is highly class-imbalanced such that there are significantly more CHD negative patients than there are CHD positive patients. This can cause models to become biased towards the majority class during classification. Additionally, it can lead to poor generalization because models may not be able to learn meaningful patterns from the minority class due to the low number of observations. To address this issue, minority oversampling was used to randomly generate new minority samples, better balance the dataset, and prevent bias in favor of the majority class.

2.2 Baseline Model

We developed basic implementations of three models to potentially serve as baseline: random forest, logistic regression, and a simple neural network. Based on superior results in test accuracy and fairness metrics as well as ease of use, we proceed with our logistic regression model.

2.3 Final Model

2.3.1 Main Model

Our primary predictive model is a Feed-Forward Neural Network implemented via TensorFlow, initialized as indicated in the table below.

Our primary model uses Binary Cross-Entropy Loss and Adam Optimizer.

Input Layer	35 Units
Dense Layer	32 Units, ReLU
Dropout Layer	30% Dropout
Dense Layer	16 Units, ReLU
Output Layer	1 Unit, Sigmoid Activation (Binary Classification)

We tracked Binary Accuracy and Balanced Accuracy as performance metrics.

2.3.2 Adversarial Model

The secondary adversarial or "discriminator" model in our adversarial configuration attempts to predict the sensitive attribute, in our case gender, from the prediction output of the main model. If the main model is biased, the adversarial should be able to correctly predict the sensitive attribute, and the main model will be penalized.

During early epochs of the training process, the main model should have large loss values, indicating inaccuracy, while the discriminator model should generally have small loss values, indicating bias in the main model. As the model is trained, the main model's loss should decrease as it improves in accuracy and learns to rely less on sensitive features, while the discriminator model's loss should increase as a result.

For this project, the main model outputs the probability that a patient has CHD, which is then used as input for the adversarial model. Since only the predicted label is used as input this adversarial model specifically aims to reduce Demographic Parity Difference. Other fairness metrics, such as odds equality, could be targeted using true and predicted labels as input for the adversarial model as demonstrated by Yang et al. (2023).

Our discriminator model is a Stochastic Gradient Descent (SGD) Classifier, which supports logistic regression, perceptron, and SVM (determined by parameter adv_model_type). The model uses SGDClassifier with the appropriate loss function and learns to predict sensitive attributes (z) from the main model's predictions.

2.3.3 Training Process

During each Epoch, our model is trained as follows:

1. Neural Network makes a Forward Pass on a batch
2. Neural Network Computes Loss (Binary Entropy Loss)
3. SGD Classifier uses NN prediction probabilities to predict sensitive attribute
4. Compute Binary Entropy Loss of Adversarial Model
5. Compute Combined Loss Function
6. Compute Gradients with Combined Loss
7. Update Weights with new gradients

2.3.4 Combined Loss Function

A combined loss function was used to compute the new gradients of the main model and update its weights. Developed by Yang et al. (2023), this combined loss function features L_P and L_A, representing the loss of the main model and the loss of the adversarial model respectively.

α represents a tunable hyper-parameter that determines the relative significance of the adversarial model in protecting the sensitive feature, z. Higher α values result in the adversarial model having a larger impact. Lastly, $\frac{L_P}{L_A}$ represents a correction term to ensure that the combined loss at the beginning of training is large, to incentivize the main model to minimize L_P while the adversarial model would maximize L_A.

At the beginning of training, ideally, L_P is large while L_A is small. As the training process progresses, the adversarial model will begin to penalize the main model when it can predict the sensitive features from the predicted labels from the main model's output. This will encourage the main model to rely less on sensitive features and focus on other features.

Over epochs of training, L_P would begin to decrease and L_A would increase. Eventually, the $\frac{L_P}{L_A}$ correction term would converge to 0 resulting in the combined loss function becoming $L_P - \alpha L_A$.

Per Yang et al. (2023), we compute combined loss as follows:

$$L = L_P + \frac{L_P}{L_A} - \alpha L_A$$

2.3.5 Metrics

We measure our model's predictive performance using:

- Accuracy
- Balanced Accuracy

Accuracy provides a general indication of the performance of our model. Due to our use of a highly imbalanced dataset, it should not be relied on alone. Balanced Accuracy serves as a more robust metric that accounts for true positives, true negatives, false positives, and false negatives. Given the context of healthcare, we took special care to factor in false negatives as it is preferable to diagnose patients who do not have CHD as positive than to diagnose patients who do as negative.

2.3.6 Fairness Metrics

We measure our model's fairness using:

- Demographic Parity Difference
- Equal Opportunity Difference
- Disparate Impact

Demographic Parity Difference is a metric that measures the difference in prediction rates between groups, per Fairlearn Team (2024). This metric encourages models not to make predictions dependent on whether observations are in a sensitive group. In the context of our project, our model's prediction of whether someone has CHD should not depend on whether someone is male or female. An ideal value for demographic parity difference is 0; larger values can indicate of bias.

Equal Opportunity Difference is a metric that compares the true positive rates between groups, per Fairlearn Team (2024). This metric encourages all groups in the dataset to be equally likely to receive a positive prediction. In the context of our project, male patients should have an equal true positive rate as female observations. While this metric does not consider differences in false positives, in the context of CHD, false positives would cause less harm overall compared to missed true positives. It is better to conduct further testing on patients who not have CHD than to incorrectly determine that patients who do have CHD are safe. An ideal value for equal opportunity difference is 0, large absolute values indicate bias.

Disparate Impact is a metric that measures the ratio between positive predictions between groups, per IBM Cloud Pak for Data (2024). This metric encourages equal positive prediction rates between the non-sensitive and sensitive groups. In the context of our project, the rate of positive CHD predictions between male and female patients should be close to equal. An ideal value for disparate impact is 1. Any number lower or higher than can indicate bias.

2.4 Hyper-parameter Optimization

Several hyper-parameters that can be adjusted in our model:

lambda_tradeoff	Weight of penalty from the adversarial model
epochs	Maximum number of epochs to train
learning_rate	Rate at which the model adjusts for each epoch
patience	Terminates training if loss stagnates or increases
adv_model_type	Architecture of adversarial/discriminator model

Due to our model's simultaneous usage of several different machine learning libraries, standard procedures such as grid search cross-validation were not be utilized. Instead, each model was trained and evaluated iteratively with different hyper-parameter values to find a "most" accurate and fair model or models.

3 Results

3.1 Baseline Model

Our baseline Logistic Regression model performed as follows:

Accuracy	0.8789
Balanced Accuracy	0.7181

Given the large gap between the two metrics, the model likely struggled to make correct predictions for the minority class. As a result, the model may overlook CHD positive patients and be biased towards predicting patients as CHD negative. This was predictable and expected considering the severity of imbalance in the dataset. The baseline model demonstrates this disparity without any additional data processing or de-biasing techniques.

Demographic Parity Difference	-0.1876
Equal Opportunity Difference	-0.2409
Disparate Impact	0.466

Keeping in mind that the ideal score is 0 for Demographic Parity Difference and Equal Opportunity Difference, and 1 for Disparate Impact, the baseline model fails to provide equal representation between groups within the sensitive classes. In the context of our problem, the negative demographic parity and equality of opportunity differences indicate that male or female patients are receiving positive predictions at a difference in rate of approximately 20%. The Disparate Impact reiterates the same result, emphasizing that the minority group is receiving fewer positive predictions compared to the majority population.

3.2 Adversarial Neural Network

3.2.1 Best Balanced Model

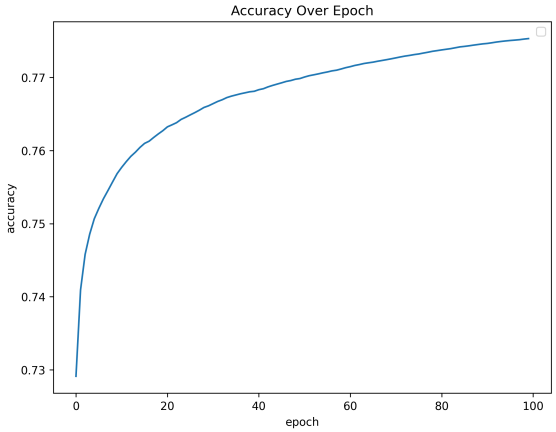
In training our adversarial neural network, we found the configuration with the best balance in performance and fairness with hyper-parameters as follows:

Adversarial Model Architecture	Logistic Regression
Learning Rate	0.001
Lambda	0.05
Batch Size	32

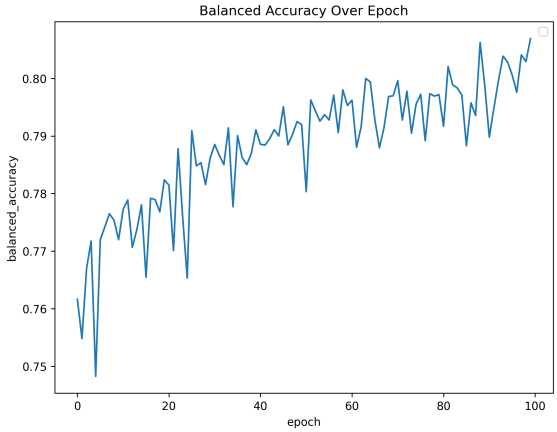
With these hyper-parameters, it was able to achieve performance as follows:

Accuracy	0.7715
Balanced Accuracy	0.7722

Comparing these numbers with our baseline model, it is clear that the test accuracy score decreased significantly from 0.8789 to 0.7715 but the test balanced accuracy increased from 0.7181 to 0.7715. Both changes are likely associated with the high level of imbalance in the original dataset and the implementation of balanced sampling in our final adversarial model.



(a) Model Accuracy Over Epochs



(b) Model Balanced Accuracy Over Epochs

Figure 1: Best Model - Accuracy Metrics over Epoch

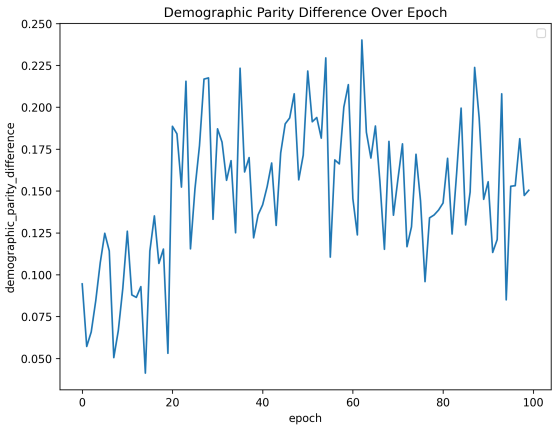
Our ”best” model achieved fairness metrics as follows:

Demographic Parity Difference	0.2436
Equal Opportunity Difference	0.0106
Disparate Impact	1.4245

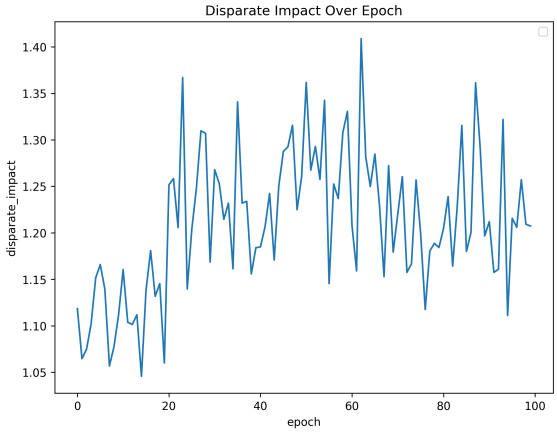
Considering the ideal values for all three fairness metrics, the model obtained a slightly worse demographic parity, moderately worse disparate impact, and significantly better equal opportunity.

Our loss function is catered towards decreasing (improving) equal opportunity difference, and it was expected that out of all three fairness metrics it would be positively affected the most. From -0.2409 to 0.0106, our adversarial configuration was able to reduce equal opportunity difference by 95.6%.

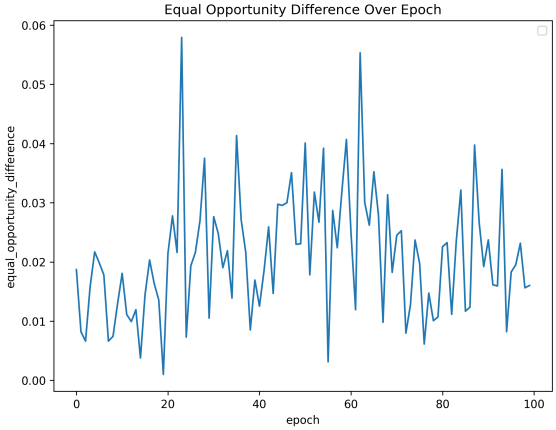
Demographic parity shifting from negative to positive may be due to oversampling the underrepresented CHD data points, diversifying the sensitive attributes of gender.



(a) Demographic Parity Difference



(b) Disparate Impact



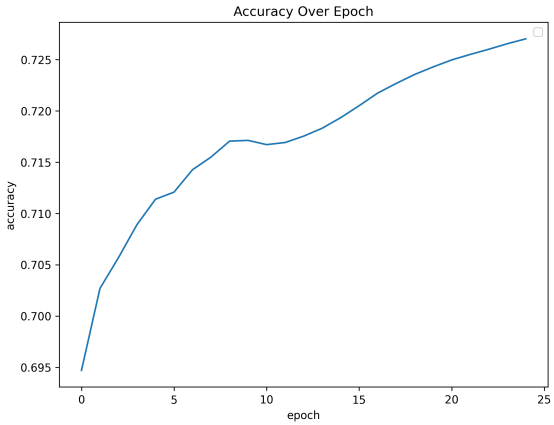
(c) Equal Opportunity Difference

Figure 2: Best Model - Fairness Metrics over Epoch

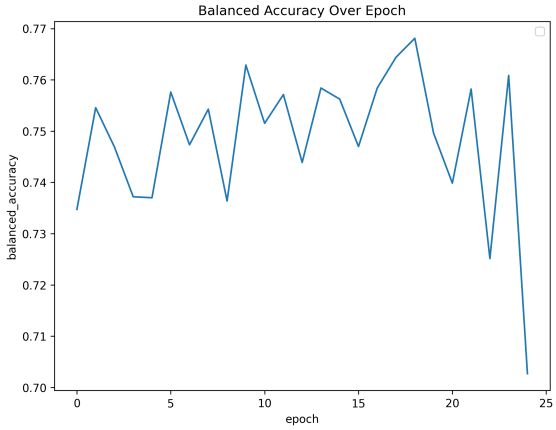
3.2.2 Most Fair Model

The model that achieved the best results in the three fairness metrics utilized a perceptron as its adversarial component, had a learning rate of 0.1, lambda trade off of 0.05, and a batch size of 64. It received a test accuracy of 0.5399 and test balanced accuracy of 0.54, showcasing a significant drop of 0.2 when compared to our baseline model. Both the accuracy and balanced accuracy over the training process can be seen in Figure 3.

Although accuracy dropped significantly, we saw the opposite trend in our fairness metrics. Achieving a demographic parity difference of 0.0563, equal opportunity difference of 0.0035 and a disparate impact of 1.0609, our model achieved near perfect values, visible in 4. This model demonstrates minimal bias when making decisions in considering patient gender. While our model showcased its ability to treat all groups of the sensitive attribute equally, its fairness came at the cost of its ability to make accurate predictions.

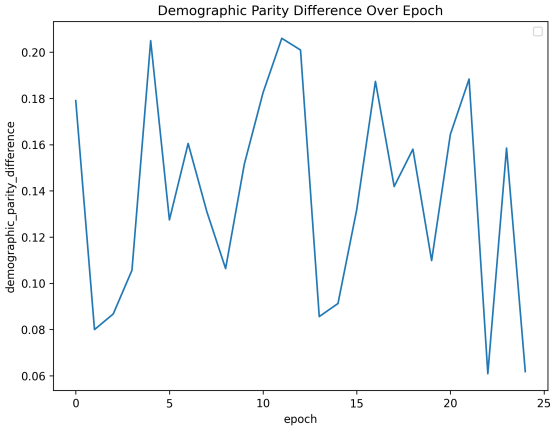


(a) Model Accuracy Over Epochs

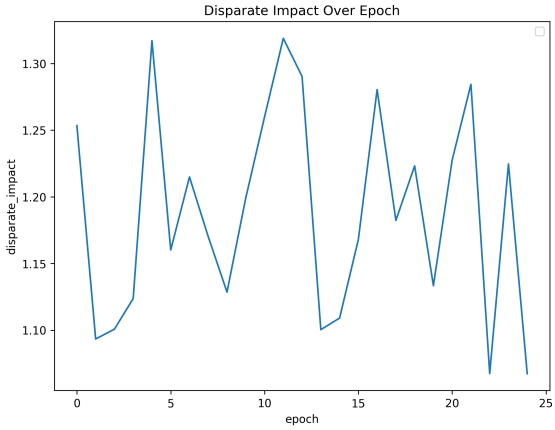


(b) Model Balanced Accuracy Over Epochs

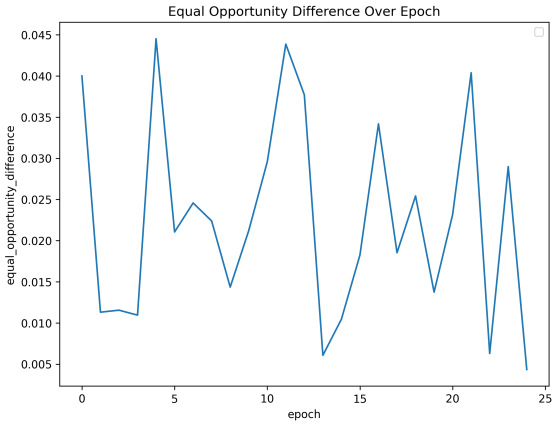
Figure 3: Most Fair Model - Accuracy Metrics over Epoch



(a) Demographic Parity Difference



(b) Disparate Impact



(c) Equal Opportunity Difference

Figure 4: Most Fair Model - Fairness Metrics over Epoch

4 Discussion

4.0.1 Overall Analysis

Our implementation of an adversarial model was designed to reduce gender bias while maintaining reasonable performance. In this pursuit, we achieved a reduction of 95.6% from -0.2409 to 0.0106 in equal opportunity difference. While our overall accuracy decreased marginally, our balanced accuracy increased a notable amount, indicating better performance with a more balanced dataset.

In our evaluation, we were able to successfully combine our neural network predictor with adversarial de-biaising to protect a sensitive feature in our dataset. While overall accuracy decreased, we believe the tradeoff is worthwhile in order to improve fairness and minimize bias. Especially in the context of healthcare, where patients health and wellbeing are at stake, it is of vital importance to consider fairness.

Importantly, we note that our model should under no context be used to conclusively diagnose patients for CHD. It is absolutely not a sufficient substitute for professional diagnosis and treatment from a licensed physician. We believe the value of our model lies primarily in our advent of the adversarial neural network configuration, and secondly as an iterative improvement on existing models used to predict CHD.

4.1 Limitations

Our results showed that reducing gender bias through an adversarial configuration is feasible even when using two different models for the main model and the adversarial component. It is important to note that there are limitations that could be addressed in future research.

Firstly, our model does not support protection of multi-class sensitive features. Only one sensitive feature can selected for the adversarial model to predict from the output of the main model. This is not necessarily a problem our dataset where the goal is only to protect one sensitive feature in gender. However, other datasets may have several sensitive features such as race, sexuality, or socioeconomic status. A more complex model may be necessary to address this.

Another limitation is that this framework focuses on improving the score of a single fairness metric. In our results, changes were observed in both other fairness metrics but those changes were unexpected and may not always occur. Future research could work to develop a model capable of addressing multiple fairness metrics simultaneously.

Lastly, model is not suitable for large data sets that contain observations in the hundreds of thousands or more. Although this model was successful in addressing gender bias, it is more suited to small or medium-sized datasets due to long training times.

4.1.1 Further Research

Further research could improve time and space-efficiency in training of our model, such as by training the adversarial on every K epoch, leveraging adversarial models that can utilize GPU hardware acceleration, or optimizing the training process as a whole.

Moreover, our adversarial framework can be applied to much more complex feed-forward neural network models with more layers. Future researchers can examine the efficacy of this framework with not just feed-forward neural networks but also other types of neural networks such as convolutional neural networks. Although this model can only reduce one specific type of bias, it can still be applied in many other healthcare applications where classification tasks can provide relevant information.

Additionally, future researchers could consider applying this framework to reduce bias in non-healthcare areas such as job applicant screening, the criminal justice system, or loan approval.

5 Conclusion

In this project, we demonstrated that adversarial de-biasing as an in-processing technique can significantly reduce gender bias for classification tasks such as CHD prediction in US patients. Our adversarial model was trained on a medium-sized data set, around 30,000 observations, using a neural network for the CHD classification task and an SGD classifier for the adversarial component. For this project, we quantify bias using fairness metrics of demographic parity, equality opportunity difference, and disparate impact.

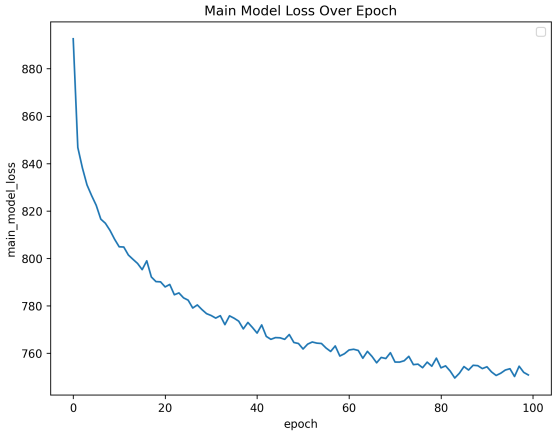
In our results, we were able to reduce the equal opportunity difference metric by 95.6%. By achieving a score closer to 0, our adversarial neural network model correctly classifies CHD more fairly and equally across male and female patients than before.

As AI/ML continues to transform healthcare delivery, our research contributes to the development of systems that are not only powerful and accurate but also fair and equitable for all patients. Although our project does not completely resolve gender bias found in the CHD healthcare space as a whole it does contribute insight and provides confidence that gender bias can be addressed in AI and machine learning models without massively sacrificing performance.

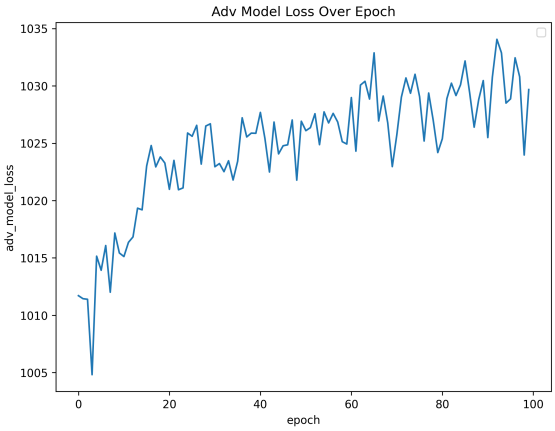
Appendices

A.1 Training Details

A.2 Additional Figures

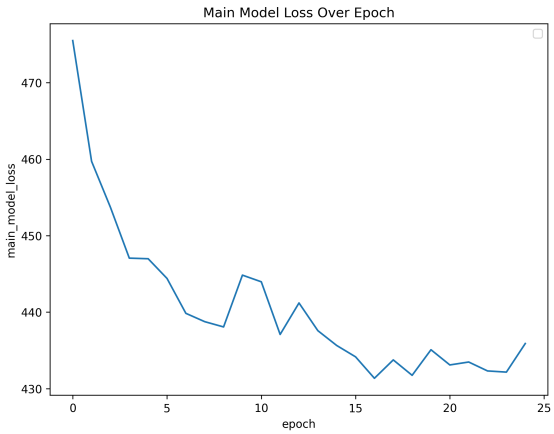


(a) Main Model Loss Over Epochs

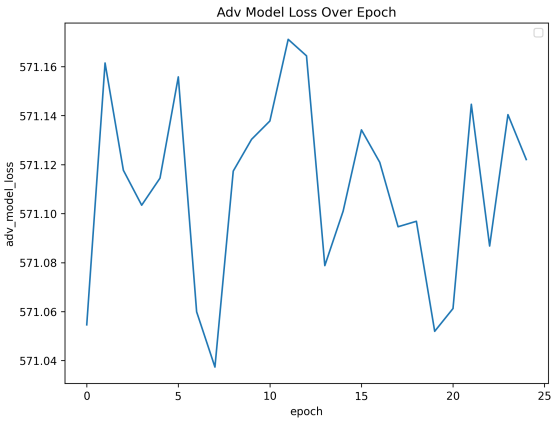


(b) Adversarial Model Loss Over Epochs

Figure A 1: Loss for Most Accurate Model

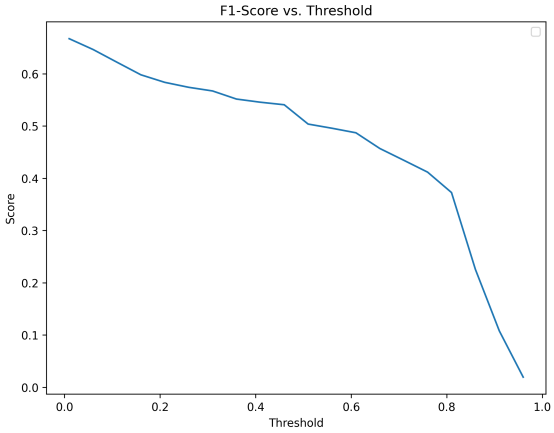


(a) Main Model Loss Over Epochs

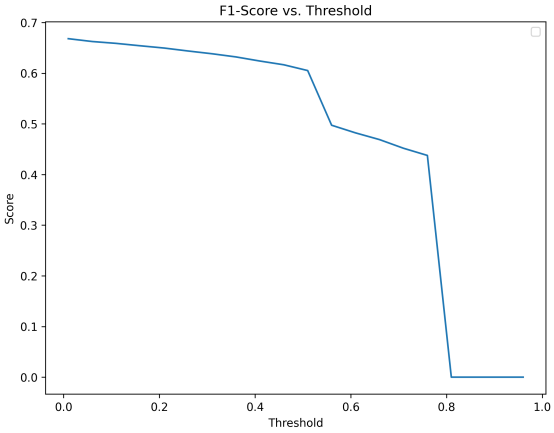


(b) Adversarial Model Loss Over Epochs

Figure A 2: Loss for Most Fair Model



(a) Most Accurate Model



(b) Most Fair Model

Figure A 3: F1-Score Over Epoch Training

A.3 Project Proposal

Initial draft of project proposal attached as follows:

DSC 180AB Group A10-2 Proposal: An Adversarial Framework for Mitigating Gender Bias in Coronary Heart Disease Prediction

Members: Diego Silva, Kai Ni, Patrick Salsbury

Problem Overview & Background

Machine learning is becoming an increasingly popular solution within the healthcare industry to classify and diagnose patients [1]. Although this has demonstrated to be an effective approach, machine learning models have been found to contain biases that discriminate underrepresented groups based on variables such as sex, gender, race, and socioeconomic status [6]. With certain diagnoses like Cardiovascular Disease (CVD), which has often been perceived as a "man's disease" [4], female patients are less likely to receive the proper treatment as they are often diagnosed with less confidence compared to male counterparts [5]. Coronary Heart Disease (CHD) is a specific type of CVD in which significant sex and gender disparities continue to be especially pervasive. Women experience more hypertension, diabetes, longer stays in intensive care units, and poorer outcomes with CVD, which has been theorized to be linked with the biases within diagnosis and referral systems [2, 3].

While there may be many determinants which can result in disproportionate rates of referral between male and female patients for CHD treatment, we would like to eliminate biases contributed by AI algorithms. Specifically, we would like to detect and mitigate biases in the development and training processes to build fair and unbiased machine learning algorithms for classifying and diagnosing patients for CHD. Previous research has explored the efficacy of using convolutional neural networks (CNN) to predict CHD, finding that CNNs can be used effectively and efficiently to produce accurate predictions [7]. CNNs are particularly flexible deep learning algorithms able to automatically learn features from datasets and analyze multiple media types. However, without appropriate supervision and necessary adjustments, they may develop significant biases and over-reliance on certain features. We believe that by using two models in an adversarial configuration, we can mitigate gender bias even in machine learning prediction of CHD. We will use a primary model to predict whether or not a patient should be referred for CHD, and we will implement a secondary discriminator model that penalizes the primary model if its predictions correlate with sensitive features within the dataset. With this approach, we should be able to encourage the primary model to focus on relevant features related to the diagnosis and discourage it to rely on sensitive features like gender that can introduce bias.

Problem Statement & Methodology

The healthcare industry faces a critical challenge in addressing sex and gender disparities in CHD prediction through machine learning models. Despite the increasing adoption of AI-driven diagnostic tools, current models may perpetuate and amplify existing biases in medical decision-making, or introduce new biases during model training and development. In this project, we will utilize adversarial debiasing, which involves leveraging two rival models – one

primary and one secondary – to reduce dependence on sensitive features or over-reliance on specific features.

This project will extend on certain aspects of our replication project, in which we examined and reduced any perceived biases in modeling healthcare utilization rates. Here, we will dive deeper and focus on CHD specifically, as opposed to the more general scope of our previous project. Our technical approach will involve the following:

$$L_{Total} = L_{Prediction} - \lambda L_{Discriminator}$$

where $L_{Prediction}$ represents the primary model's loss function for CHD prediction, $L_{Discriminator}$ represents the adversarial component's ability to predict gender from the model's internal representations, and a constant λ is used to weight the discriminator model. Our ultimate goal will be to develop a machine learning model that maintains clinical relevance while reducing gender correlation through the adversarial configuration and incorporating fairness metrics into the training objective.

Justification of Success

Based on the successful development of CNN-based models for predicting CHD [7], we believe that we will be able to maintain a competent level of performance on the same dataset with our debiased model, even after our debiasing techniques. We will explore combinations of CNN with other conventional machine learning models such as logistic regression, gradient boosting, and random forest for the primary and discriminator models, and evaluate performance in efficiency, accuracy, and bias. We will also experiment with using the same type of model for both the primary and discriminator models and compare results. We will determine which configuration yields the best outcomes, and proceed by further optimizing any relevant hyperparameters. By testing multiple combinations, we should finally be able to produce an efficient, accurate, and fair model.

Dataset: “NHANES data from 1999–2000 to 2015–2016. The dataset is compiled by combining the demographic, examination, laboratory and questionnaire data of 37,079 (CHD – 1300, Non-CHD – 35,779) individuals” [7].

Primary Output Statement

We will choose a website for our primary output format on top of our standard report. Our report will have a detailed description of our model architecture, training process, and key results and findings. Our website will include documentation that will allow anyone to reproduce our results with instructions on how to set up an environment and install necessary imports, along with our results and analysis. Based on the results of our project, we will also work to design feasible and relevant interactive modules to include in our website.

A.4 Contributions

Individual contributions attributed to members based on who has taken lead or contributed most significantly on relevant component(s); components developed or composed through collaborative effort listed under Section A.4.1.

A.4.1 Team Members Collectively

- Project proposal
- Baseline and final models
- Formal project report
- Project website

A.4.2 Diego Silva

- Develop and test baseline models
  - Random Forest
  - Logistic Regression
  - Simple Neural Network
- Develop and test final model with adversarial module

A.4.3 Patrick Salsbury

- Compile dataset over multiple year range using NHANES API
- Write automated scripts
  - Dataset download
  - Dataset cleaning
  - Final model testing

A.4.4 Kai Ni

- Write and edit proposal and report
  - Abstract
  - Introduction
  - Methodology
- Develop project website

References

Al Hamid, Abdullah, Rachel Beckett, Megan Wilson, Zahra Jalal, Ejaz Cheema, Dhiya Al-Jumeily OBE, Thomas Coombs, Komang Ralebitso-Senior, and Sulaf Assi. 2024. "Gender bias in diagnosis, prevention, and treatment of cardiovascular diseases: A systematic review." *Cureus*. [Link]

Beery, Theresa A. 1995. "Gender bias in the diagnosis and treatment of coronary artery disease." *Heart amp; Lung* 24 (6), p. 427–435. [Link]

Dutta, Aniruddha, Tamal Batabyal, Meheli Basu, and Scott T. Acton. 2020. "An efficient convolutional neural network for coronary heart disease prediction." *Expert Systems with Applications* 159. [Link]

Fairlearn Team. 2024. "Common Fairness Metrics." [Link]

IBM Cloud Pak for Data. 2024. "Disparate Impact." [Link]

Maserejian, Nancy N., Carol L. Link, Karen L. Lutfey, Lisa D. Marceau, and John B. McKinlay. 2009. "Disparities in Physicians' Interpretations of Heart Disease Symptoms by Patient Gender: Results of a Video Vignette Factorial Experiment." *Journal of Women's Health* 18 (10), p. 1661–1667. [Link]

Mihan, Ariana, Ambarish Pandey, and Harriette GC Van Spall. 2024. "Mitigating the risk of artificial intelligence bias in Cardiovascular Care." *The Lancet Digital Health* 6 (10). [Link]

Nation Center For Health Statistics (NCHS)., "About the National Health and Nutrition Examination Survey (NHANES)." https://www.cdc.gov/nchs/nhanes/about/

Schenck-Gustafsson, Karin. 2009. "Risk factors for cardiovascular disease in women." *Maturitas* 63 (3), p. 186–190. [Link]

Winkler, Katja, Navina Gerlach, Norbert Donner-Banzhoff, Anika Berberich, Jutta Jung-Henrich, and Kathrin Schlößler. 2023. "Determinants of referral for suspected coronary artery disease: A qualitative study based on decision thresholds." *BMC Primary Care* 24 (1). [Link]

Yang, Jenny, Andrew A. S. Soltan, David W. Eyre, Yang Yang, and David A. Clifton. 2023. "An adversarial training framework for mitigating algorithmic biases in clinical machine learning." *NPJ Digital Medicine* 6 (1), p. 55. [Link]