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

Mentor: Emily Ramond eramond@deloitte.com

Mentor: Greg Thein gthein@deloitte.com

UC San Diego™

HALICIOĞLU DATA SCIENCE INSTITUTE

Introduction

- Artificial Intelligence (AI) and Machine Learning (ML) are increasingly being employed in healthcare **to classify and diagnose patients** [1]
- These technologies offer immense potential but introduce **significant challenges in avoiding bias and fairness**
- Our project focuses on **gender bias in the prediction of Coronary Heart Disease (CHD)**, where female patients experience much higher rates of misdiagnoses and sub-optimal care [2]

Objectives

- Develop **a fair and unbiased algorithm** for CHD classification
- Maintain clinically relevant accuracy while **reducing gender-based disparities**
- Contribute to the growing body of literature on ethical AI use in healthcare
- **Offer a novel approach to mitigating demographic biases** in cardiovascular disease diagnostics

Methods

We implemented a neural network (NN) model using an **adversarial configuration** to tackle this challenge, involving:

1. A **primary neural network** [3] model for CHD classification
2. A **secondary "discriminator" model** to detect and penalize gender-based biases

Ultimately, we produced an **Adversarial Neural Network** model capable of predicting CHD at a clinically relevant accuracy **while significantly reducing gender bias**.

Technical Details

- Our dataset of around 30,000 individuals was acquired from the NHANES survey conducted by the CDC [4]
- Our primary model is a Feed Forward NN implemented in TensorFlow
- Our secondary model is an SGD Classifier, supporting logistic regression, perceptron, and SVM
- Model performance was measured using accuracy and balanced accuracy
- Fairness and bias were measured using Demographic Parity Difference, Equal Opportunity Difference, and Disparate Impact

Website

Visit our website for more details!

chd-adversarial-nn.github.io

References

[1] Mihan et. al. - Mitigating the Risk of Artificial Intelligence Bias in Cardiovascular Care https://doi.org/10.1016/S2589-7500(24)00155-9

[2] Al Hamid et. al. - Gender Bias in Diagnosis, Prevention, and Treatment of Cardiovascular Diseases: A Systematic Review https://doi.org/10.7759/cureus.54264

[3] Dutta et. al. - An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction https://doi.org/10.1016/j.eswa.2020.113408

[4] https://www.cdc.gov/nchs/nhanes/about/index.html

Results

We conducted hyper-parameter tuning by experimenting with learning rate, batch size, lambda, (weight of adversarial model in loss function), adversarial model architecture. Our best model was found as follows:

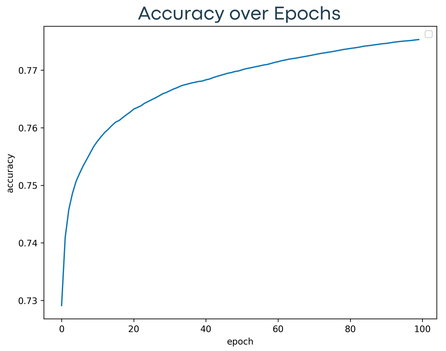
- Learning Rate: **0.001**
- Batch Size: **32**
- Lambda: **0.05**
- Adversarial Model Architecture: **Logistic Regression**

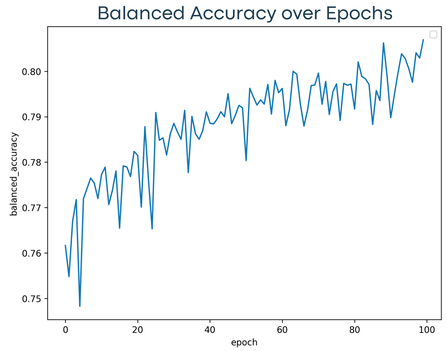
This model achieved performance metrics of:

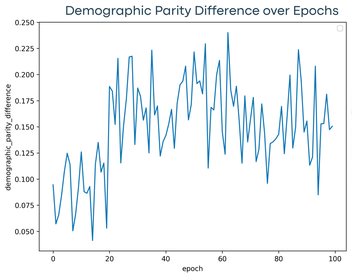
- Accuracy: **0.7715**
- Balanced Accuracy: **0.7722**

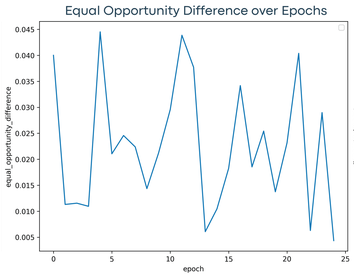
This model achieved fairness metrics of:

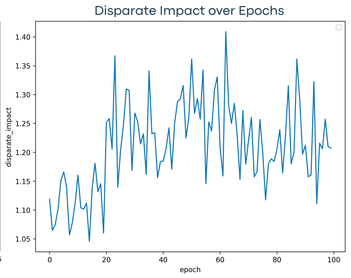
- Demographic Parity Difference: **-0.1876**
- Equal Opportunity Difference: **0.0106**
- Disparate Impact: **1.4245**











Analysis

Compared to our baseline Logistic Regression model, our model performed as follows:

Metric	Ideal	Baseline	Adversarial	% Improved
Accuracy	1.0	0.8789	0.7715	**-12.22%**
Balanced Accuracy	1.0	0.7181	0.7722	**7.53%**
Demographic Parity Difference	0.0	-0.1876	0.2436	**-29.85%**
Equal Opportunity Difference	0.0	-0.2409	-0.0106	**95.60%**
Disparate Impact	1.0	0.466	1.4245	**20.51%**

The decrease in test accuracy is likely associated with the level of imbalance in our dataset. In training our adversarial model, we implemented sampling techniques for better representation of the minority group at the cost of decreased accuracy.

Key Takeaways

- Balanced accuracy showed a moderate improvement
- Our model was trained to minimize equal opportunity difference, and **successfully decreased it by 95.6%**
- The two other fairness metrics showed unintended fluctuations of around 20%
- In our evaluation, we were able to successfully combine our NN predictor with adversarial de-biasing 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 well-being are at stake, it is of vital importance to consider fairness

Limitations

Despite promising results, our model has some limitations:

- This model is only capable of **protecting one feature**
- This framework primarily focuses on **a single fairness metric**. While in our results changes in other fairness metrics are observed, those changes may not always occur
- Given our **limited computational resources**, our model would not have been able to accommodate for a larger dataset

Next Steps

- Further develop the model to **support protection of multiple features**, for instance in scenarios where multiple features like gender, race, and socioeconomic status may all be sensitive
- Further develop the model to **support optimizing multiple fairness metrics** simultaneously
- **Improve model efficiency**, leverage additional computational resources, and test performance on **larger datasets**
- Test our adversarial approach in other contexts