Lucas Lee, Antonio Vieira, Lola Serdan

Heart Disease Prediction Model

Overview:

Our project uses a neural network to predict heart disease based on clinical features from real patient data from UC irvine. The analysis includes data exploration, correlation patterns, model training, evaluation using accuracy and ROC-AUC, and interpretation of feature importance. Our goal is to build a reliable model that can identify patients at higher risk and understand which factors contribute most to predictions.

Dataset Overview

We combined four UCI Heart Disease datasets from Cleveland, Hungary, Switzerland and VA Long Beach. Together there was 920 patient records.

Features (13 clinical features):

1. age – Age of the patient
2. sex – Sex (1 = male, 0 = female)
3. cp – Chest pain type
4. trestbps – Resting blood pressure (mm Hg)
5. chol – Serum cholesterol (mg/dl)
6. fbs – Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
7. restecg – Resting electrocardiographic results
8. thalach – Maximum heart rate achieved
9. exang – Exercise-induced angina (1 = yes, 0 = no)
10. oldpeak – ST depression induced by exercise relative to rest
11. slope – Slope of the peak exercise ST segment
12. ca – Number of major vessels colored by fluoroscopy (0–3)
13. thal – Thalassemia status

Dataset Overview

Target Variable:

- Binary classification
  - 0 = No disease
  - 1 = Disease

Preprocessing Notes:

- Missing values handled using median imputation
- Features standardized before training
- Dataset split using a stratified 80/20 train–test split

Model Architecture:

Model Type: Neural Network Built with PyTorch

Input Layer:

- 13 input features (clinical measurements)

Hidden Layers

- Dense layer: 32 neurons
  - Batch Normalization
  - ReLU activation
  - Dropout (0.2)
- Dense layer: 16 neurons
  - Batch Normalization
  - ReLU activation
  - Dropout (0.2)

Output Layer

- 2 neurons. Represents probability of:
  - No disease
  - Disease

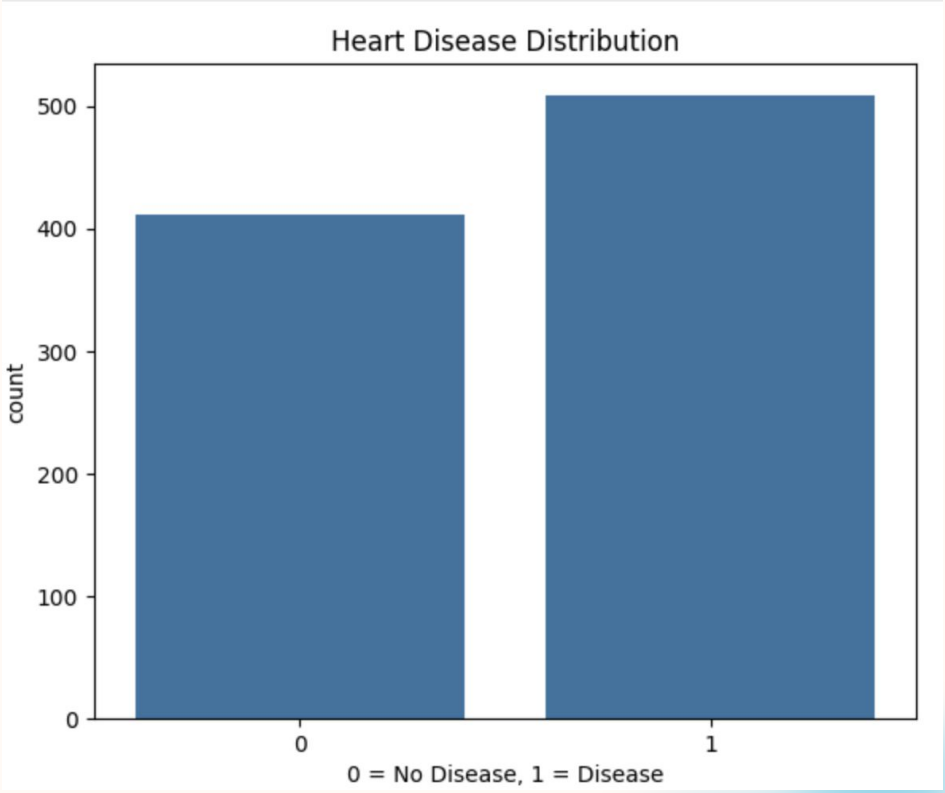
Training Details

Loss function: Cross-Entropy Loss with class weighting

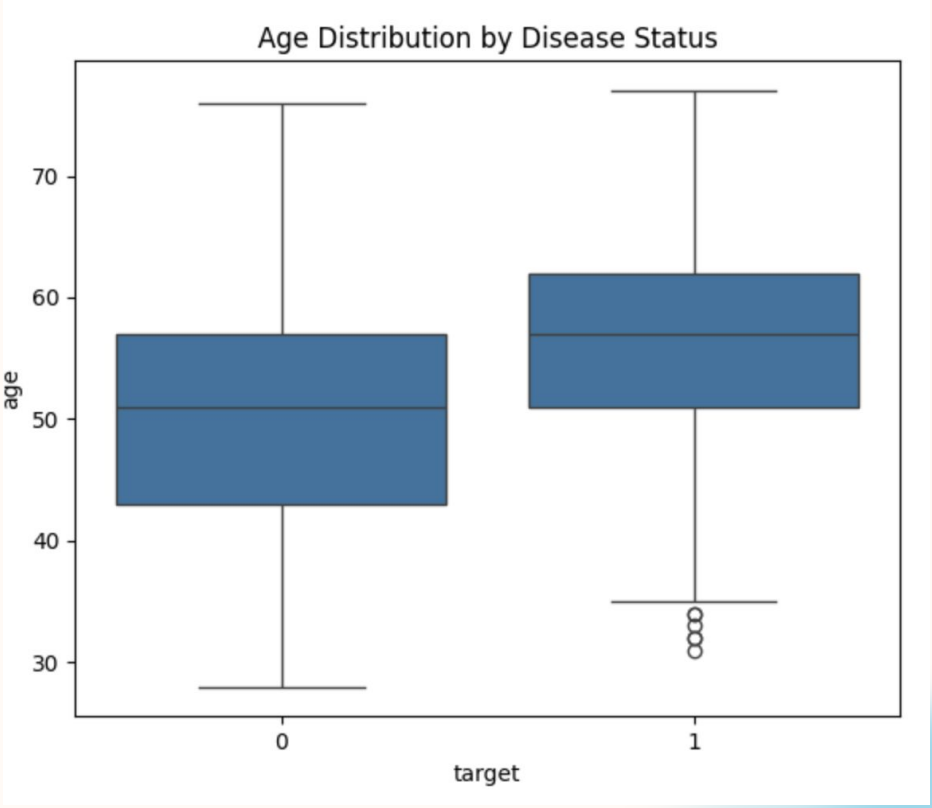
Optimizer: Adam (learning rate = 0.001)

Early stopping based on validation loss

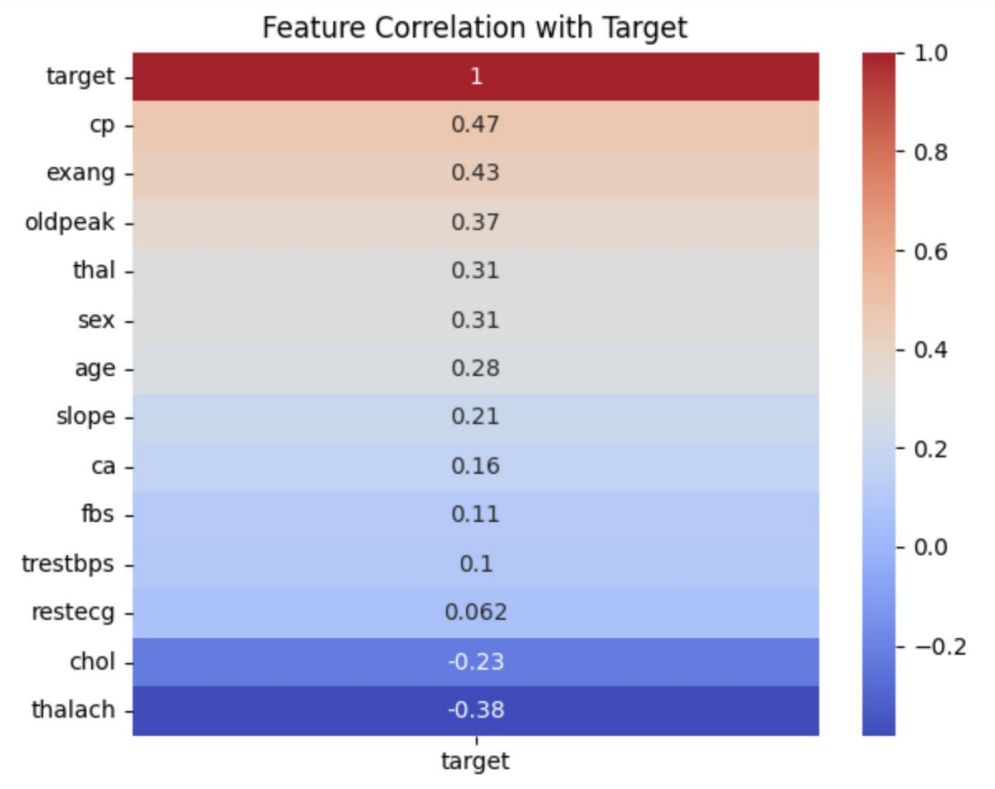
Graph Analysis



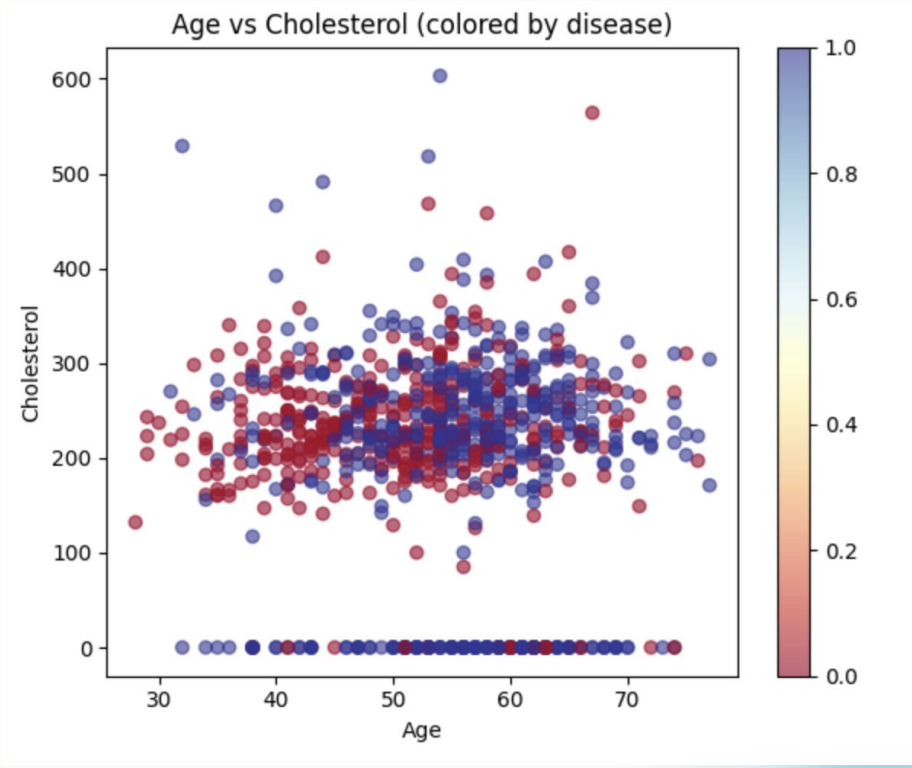
This bar chart shows the number of patients with and without heart disease in the dataset. In the data, there are 411 patients with no disease and 509 with cardiovascular disease.



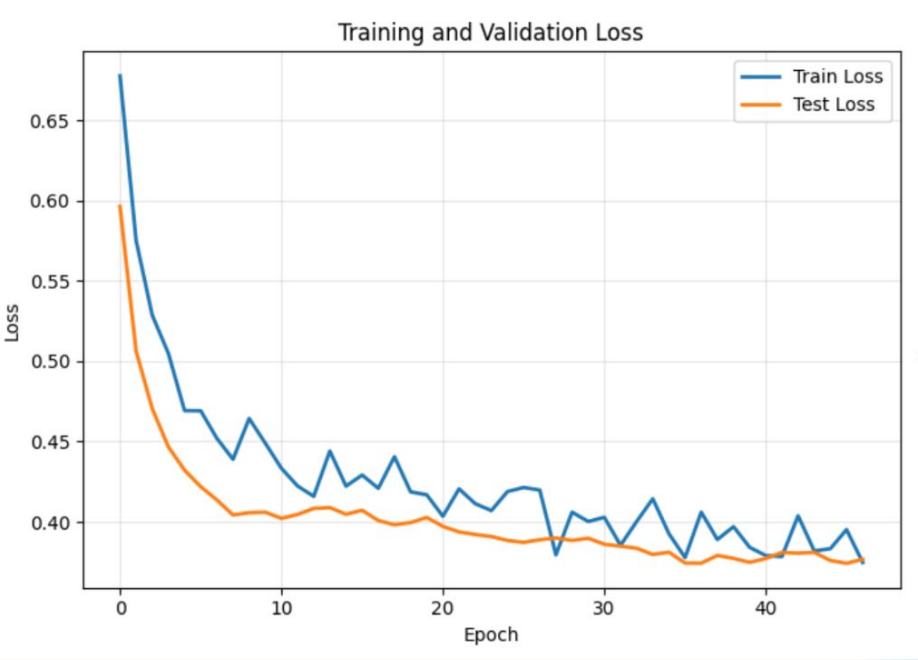
This boxplot compares the age distributions of patients with and without heart disease. Patients with heart disease tend to be slightly older on average, and their age distribution is wider.



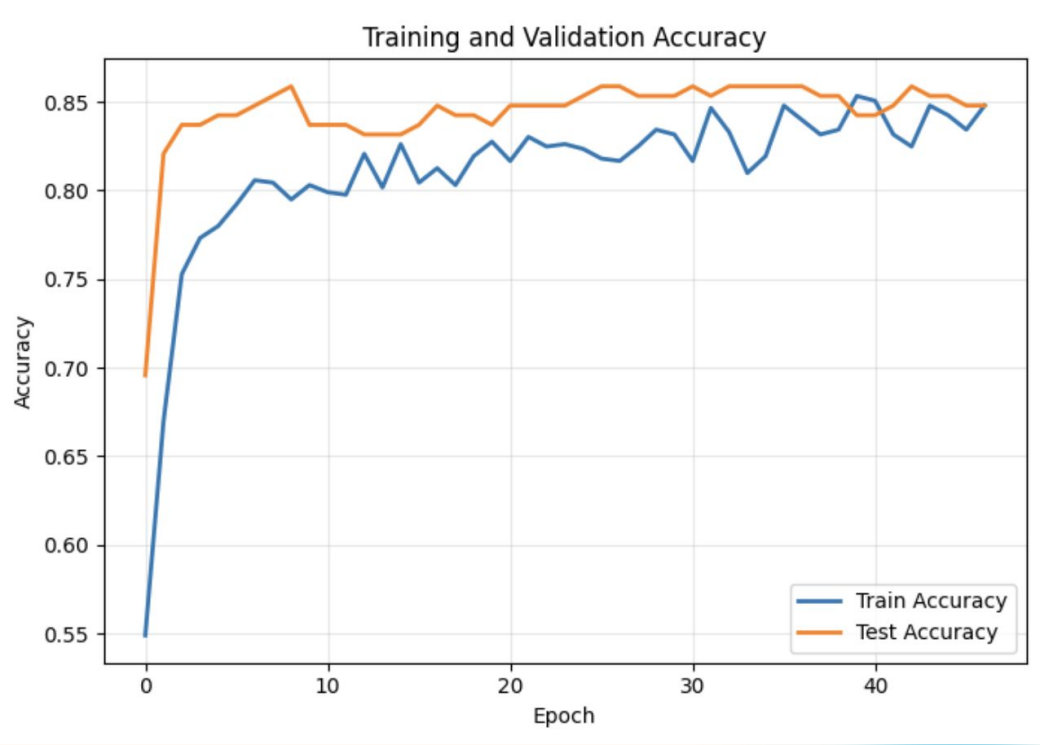
The heatmap displays the correlation between each clinical feature and the target variable (heart disease). The color scale ranges from strong positive correlations (red) to strong negative correlations (blue). Features such as cp (chest pain type), thal, oldpeak, and exang show higher positive correlations with heart disease, while features like thalach (maximum heart rate achieved) show a negative correlation.



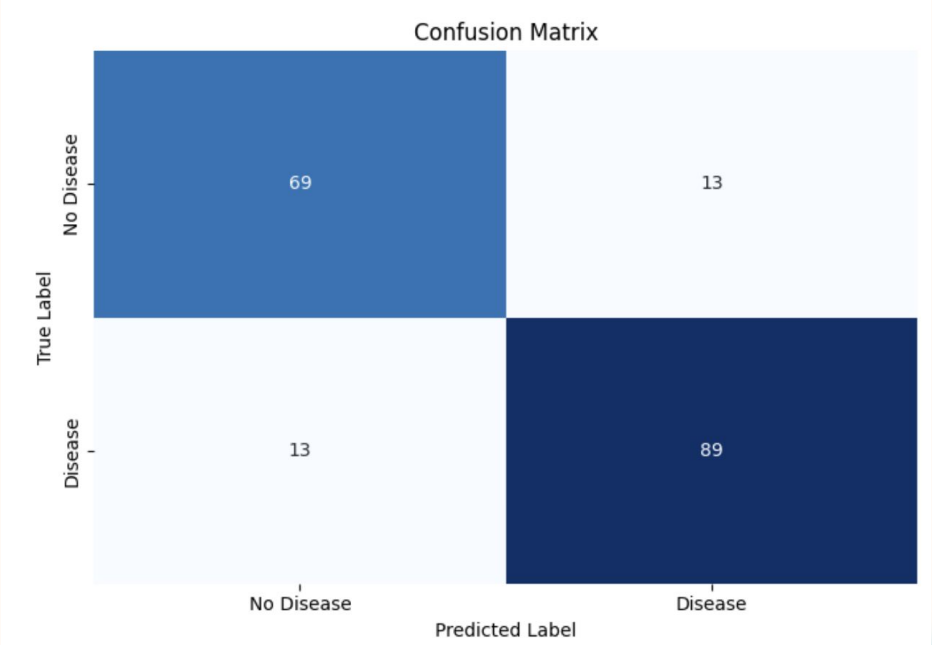
This scatterplot shows cholesterol levels vs. age, with colors indicating if they have disease or not. The points for disease and non-disease overlap a lot. The overlap suggest that cholesterol alone does not seem to not be a strong standalone predictor.



This line plot shows the decrease in training and test loss over the different epochs. Both of the losses decrease steadily and stay close to in proximity, indicating no major/considerable overfitting. The model learns effectively and generalizes competently to new data.



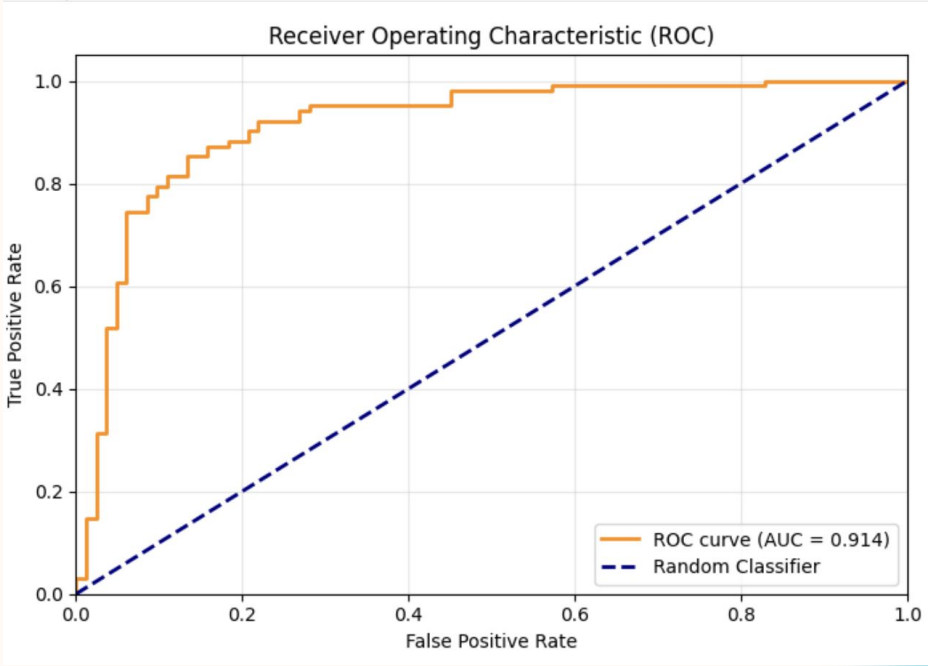
This plot shows how training and test accuracy advance over time. The close alignment between training and validation accuracy verifies the models stability. The plateau shows the model has reached its optimal learning level, without overfitting.



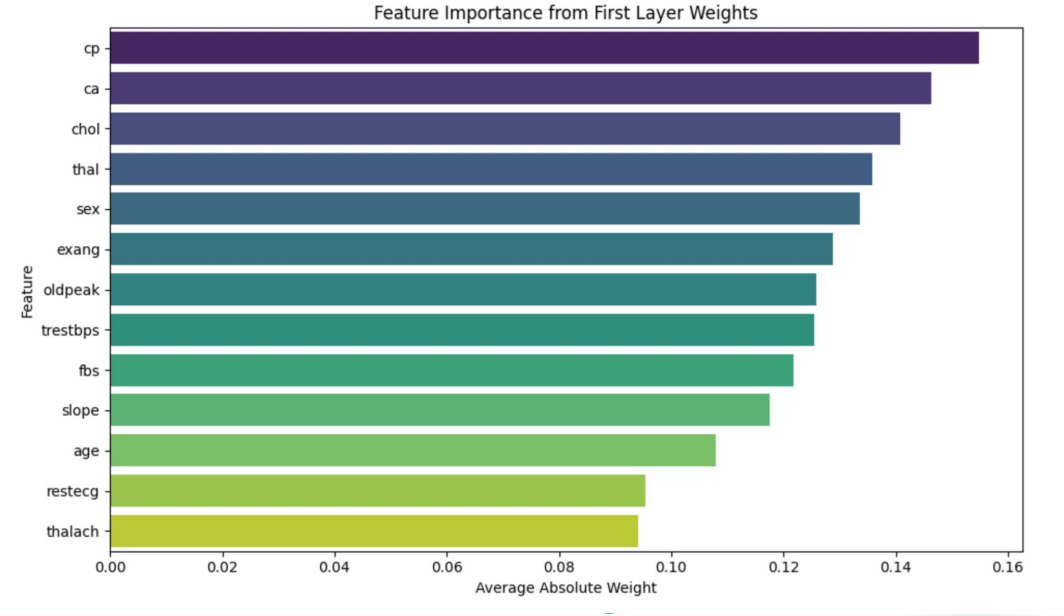
The matrix shows the model's predicted vs. actual classes:

- 69 correct “No Disease”
- 89 correct “Disease”
- 13 false positives
- 13 false negatives

This represents an overall accuracy of 86% (158/184)



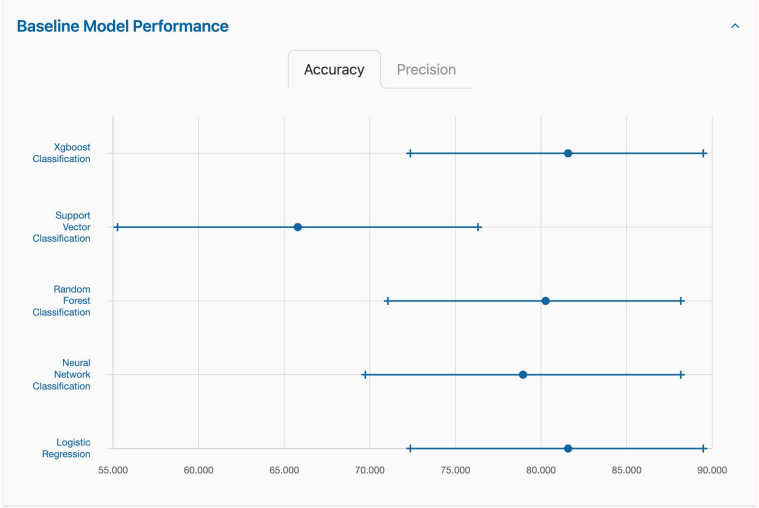
This curve compares true positive rate vs. false positive rate across thresholds. The AUC is 0.914. AUC over 0.90 is considered great by many standards. It displays the models ability to distinguish well between diseased and healthy patients, even as the classification threshold changes.

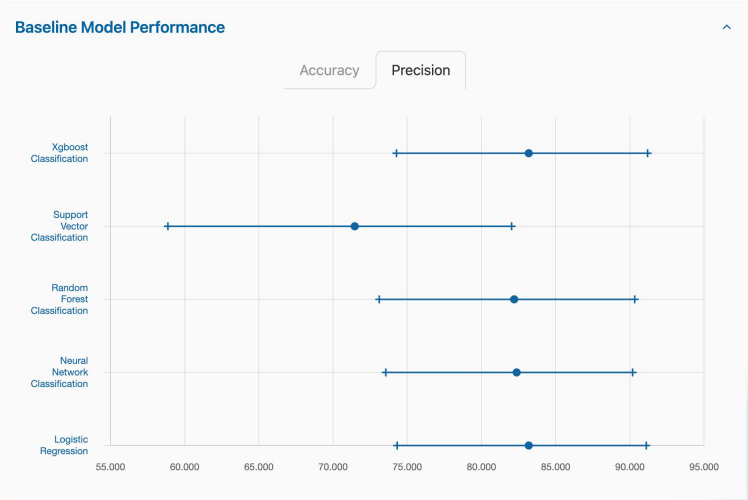


This bar chart displays the magnitude of first-layer weights, showing which features influence the model most.

The top influencing features:

1. cp
2. thal
3. chol
4. restecg
5. exang





These charts are from the website where we got out dataset; a UCI repository. They show that their Neural Network Classification achieved a accuracy of 78.947 and a precision of 82.372. Our Neural Network Classification received higher accuracy and precision scores than their models. Our test accuracy score was 83.7 and the precision was 84.89.

Global Relevance:

Cardiovascular disease is the world's leading cause of death attributing to 17.9 million deaths annually, this accounts for almost 1 of 3 deaths. In the US, this figure translates to a death every 40 seconds. On top of this tragic statistic, sits nearly $200 billion in healthcare costs. Despite these numbers, many cardio disease cases go undetected until late-stage condition, or a major cardiac event occurs (cardiac arrest, etc) hence why early detection and recognition is critical.

Conclusion:

The neural network achieved strong performance, reaching an accuracy of 85.87% and an AUC of 0.914, showing excellent ability to distinguish between patients with and without heart disease. The model maintained balanced sensitivity and specificity, and the confusion matrix demonstrated consistent accuracy across both classes. Feature importance and correlation analysis confirmed that medically relevant variables—such as chest pain type, thalassemia results, and cholesterol—played major roles in the predictions. Overall, the project shows that machine learning can effectively capture meaningful clinical patterns and support early risk detection.

Next Steps:

1: Focus on making the model more foolproof and useful for real clinical settings.

2: Use cross-validation and more structured hyperparameter tuning to allow us to get to more dependable performance estimates and incremental gains in accuracy.

3: Extend the task from binary classification to predict disease severity and test the model on fully independent datasets to further increase its clinical relevance.