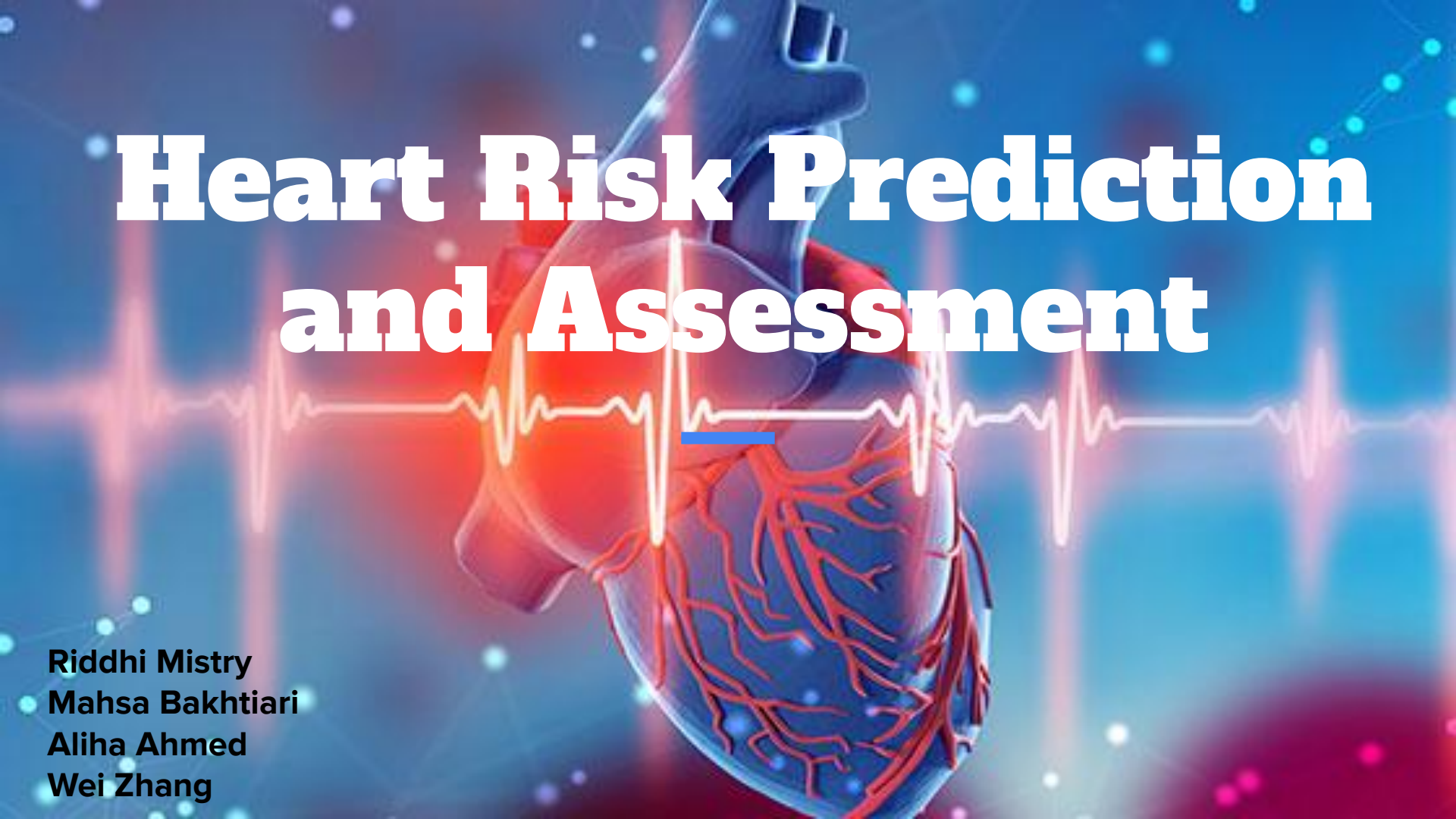
Heart Risk Prediction and Assessment



Riddhi Mistry
Mahsa Bakhtiari
Aliha Ahmed
Wei Zhang

Agenda

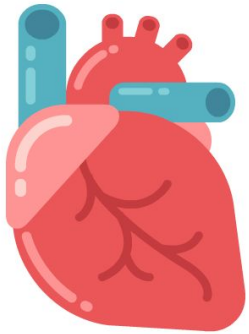
- Project Overview
- Data Limitation
- ETL Pipeline
- ECG Image Classification
- Heart Attack Risk Visualization
- Heart Attack Risk Analysis
- Conclusion

Project Focus

Heart disease or myocardial infarction remains a major global health problem that requires a deeper understanding of its precursors and possible contributing factors.

This project will focus on the following:

- Use ECG images to identify heart defects
- The factors would impact heart attack risk



https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

https://data.mendeley.com/datasets/gwbz3fsgp8/2

Data Limitation

Removing excess noise
- Dropping 'Country' column
- Dropping ‘Blood Pressure’ column as we saw it was unnecessary

-Converted string values in ‘Sex', 'Continent', 'Hemisphere' columns to boolean using the get_dummies function

- Manual replacement for trinomial ‘Diet’ column to numeric values [0, 1, 2]

- Scaled the rest of the numeric columns using StandardScaler().fit_transform from the SciKit-Learn library for better consumption for the model

ETL Pipeline

- Extract

Data downloaded from the Kaggle source as a CSV

- Transform

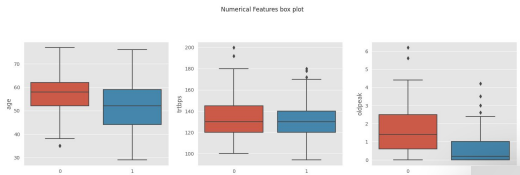
This project used Pandas to clean the data.

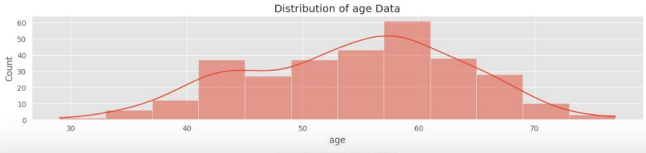
- Load

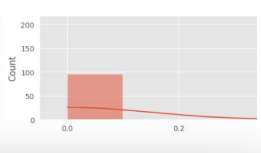
Making spark sql query based on the data and create database. Then deploy the database to Python Flask powered API.

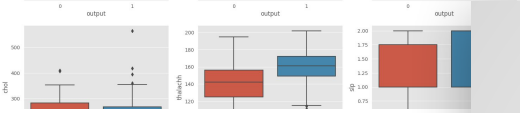
This part will analyze on two aspects from our dataset:

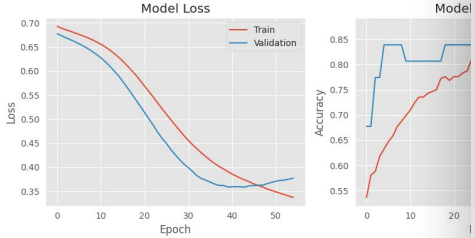
- Demographic analysis covers ages and gender
- Lifestyle choices that related to heart health

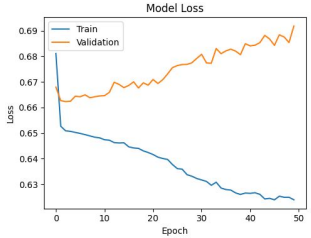


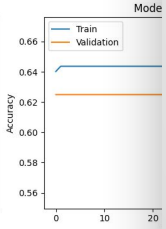


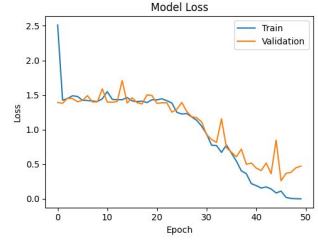


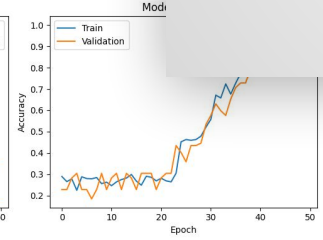


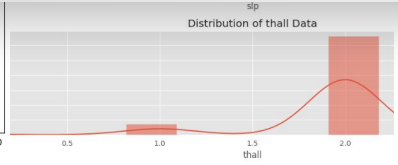


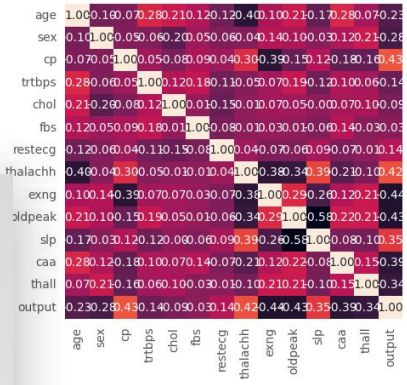










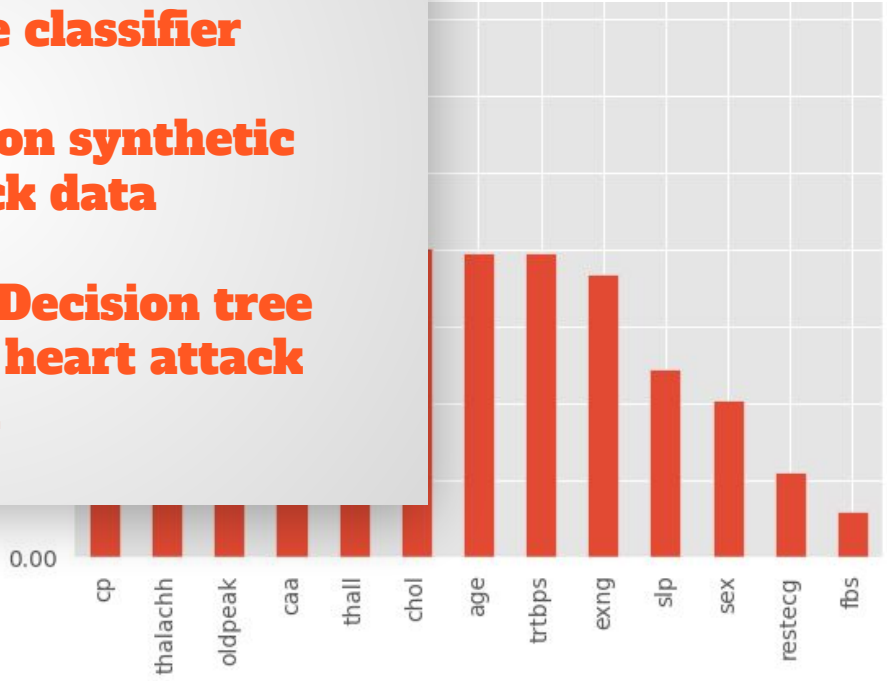


4 models in total

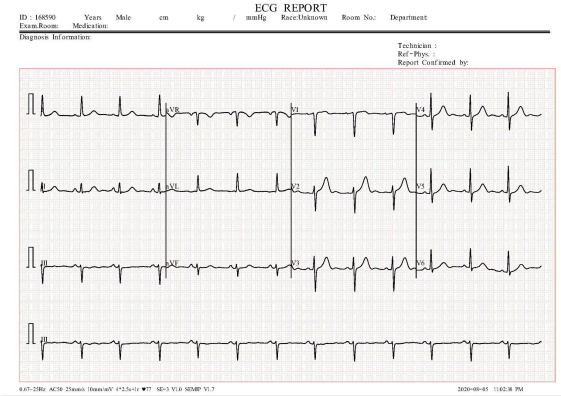
One ViT image classifier

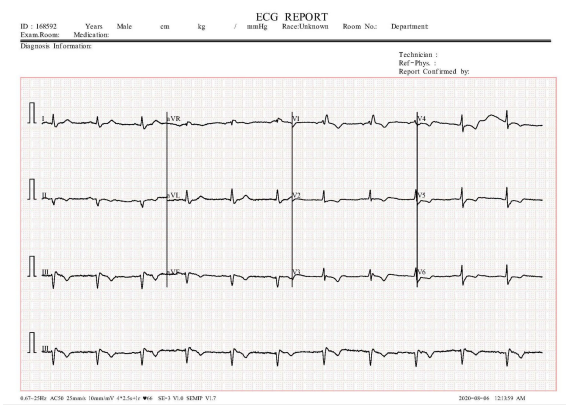
One NN models on synthetic heart attack data

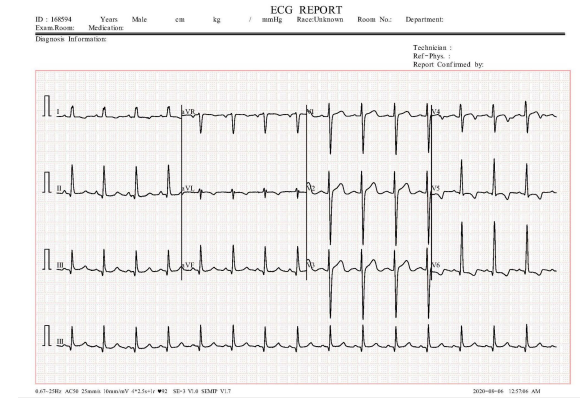
One NN and one Decision tree classifier on real heart attack data

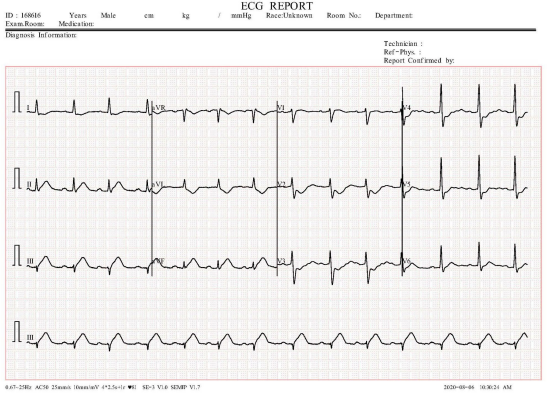


Predicting Heart Defects from ECG Images









Labels:

- Normal
- Abnormal Heart Rate
- myocardial infarction (MI)
- History of MI

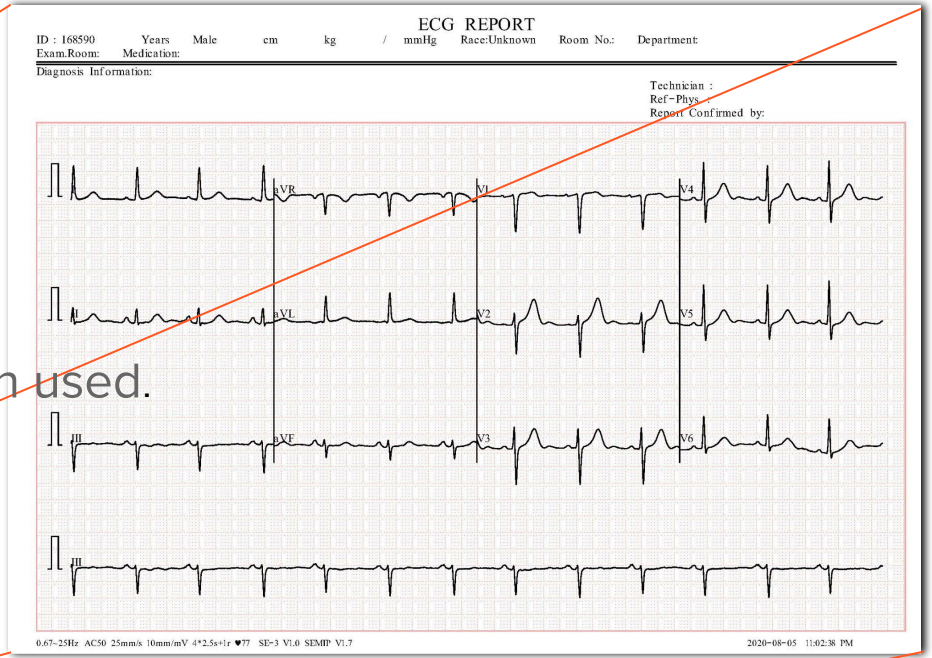
Dataset:

- ~1000 images
- 2213 x 1572 resolution
- Labeled by human experts

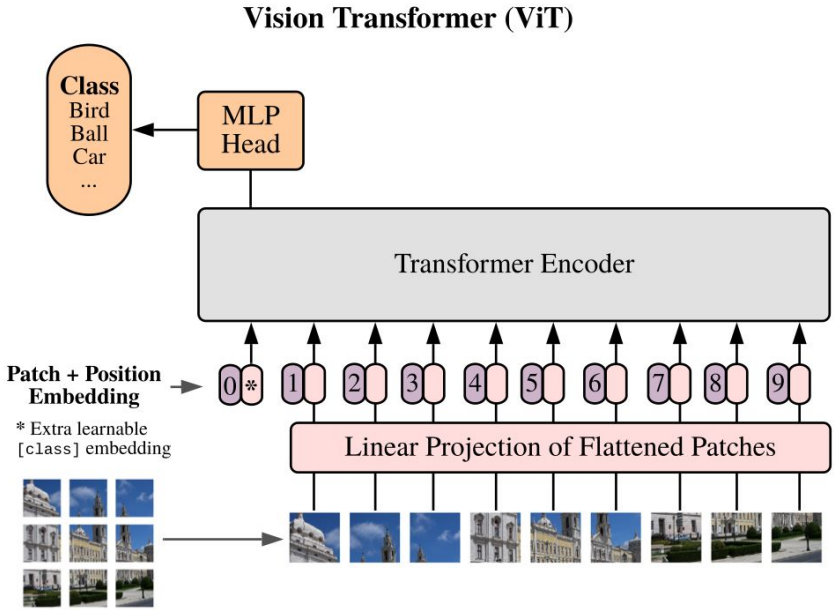
Data Preprocessing

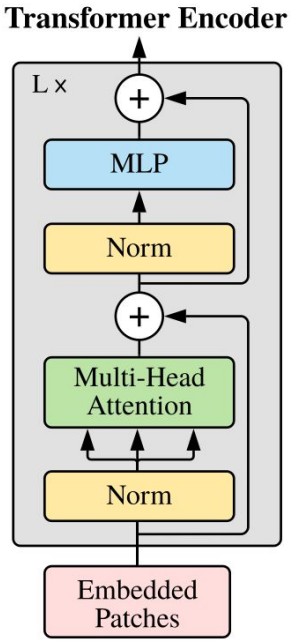
Image resizing:

- 182 x 256 resolution
- Image resizing hyper parameters:
  - Antialiasing is enabled.
  - “Area” algorithm for interpolation used.



Vision Transformer Architecture



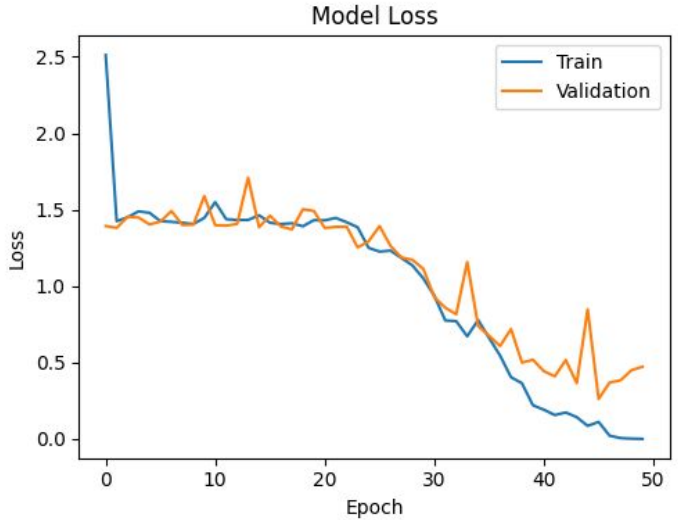


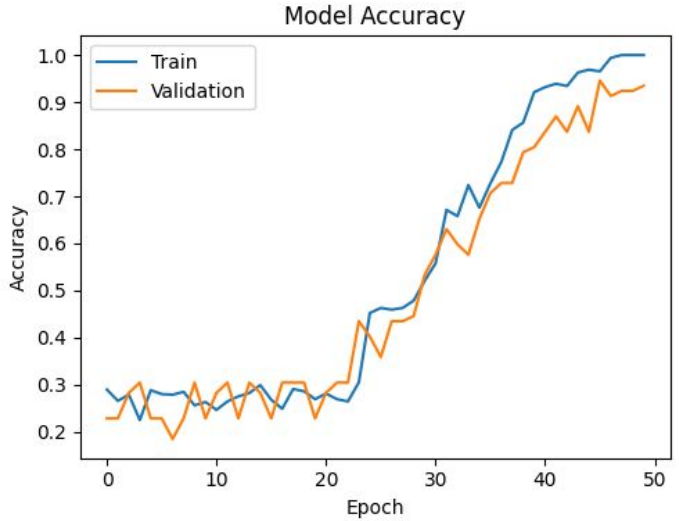
Steps:

1. Patchify
   a. 7x8 Conv kernel
   b. 7x8 Stride
2. Flattening
3. ViT Blocks
   a. Layer Norm
   b. Self Attention
   c. MLP Block
   d. Residuals
4. Final Layer Norm
5. Dense
   a. Softmax
6. Cross Entropy Loss

Modeling Hyper-parameters & Results

- 25 million trainable parameters
- Adam optimizer
  - Weight Decay = 1.0 to address overfitting
- 35 epochs
- 32 batch size
- 91% accuracy on the training set
- 86% accuracy on the validation set





Heart Attack Risk Data Features

Age

Cholesterol

Heart Rate

Diabetes

Family History

Smoking

Obesity

Alcohol Consumption

Exercise Hours Per Week

Previous Heart Problems

Medication Use

Stress Level

Sedentary Hours Per Day

Income

BMI

Triglycerides

Physical Activity Days Per Week

Sleep Hours Per Day

Heart Attack Risk

systolic

diastolic

Sex

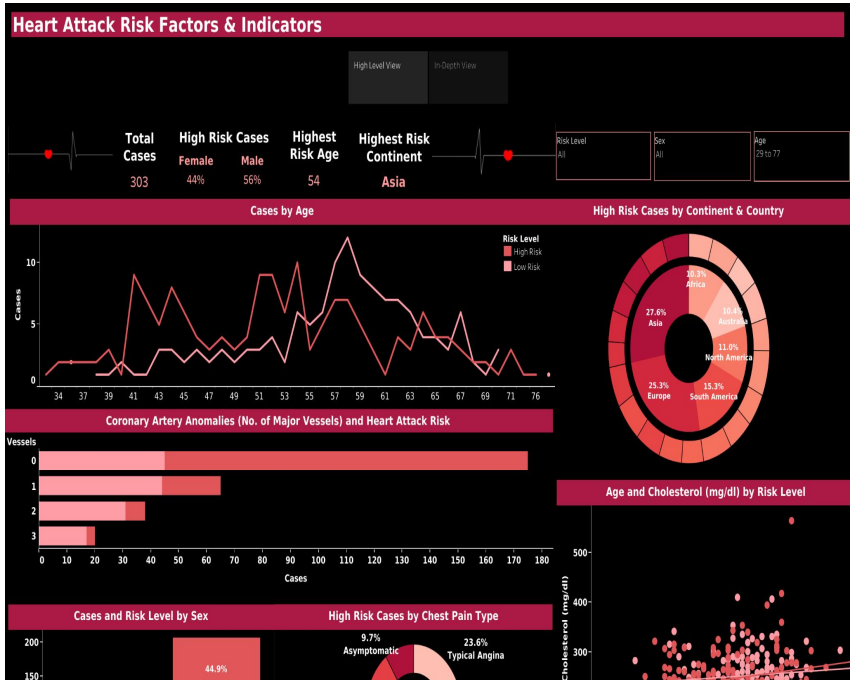
Diet

Country

Continent

Hemisphere

Visualization: Tableau Story



Visual Insights

- Visual indicators agree with broader literature and research highlighting “male” as the higher risk gender for heart attacks, as well as higher mortality rates associated with Central Asian and Eastern European populations
- Peak ages for high risk cases across male and female patients are 54 and 52, respectively
- Lacking all major vessels, the most severe form of a coronary artery anomaly indicated in the dataset, presented the highest heart attack risk out of other categories (0-3)
- Linear regression coefficient for age and cholesterol relationship in high risk group found to be roughly double that of low risk group
    - High risk group: 1.39611 (R-Squared: 0.0617129)
    - Low risk group: 0.69766 (R-Squared: 0.0126011)

Data Analysis

8,763 patients data distribute between different genders:

Sex	Number_of_Patients
Female	2652
Male	6111

Presence of heart attack risk between different genders:

Sex	Number_of_Patients
Female	944
Male	2195

Data Analysis

Age > 50, potential heart attack with (1) and without (0) pervious heart issue

Sex	Older_Than_50
Female	237
Male	597

(1)

Sex	Older_Than_50
Female	252
Male	592

(0)

Age < 30, potential heart attack with (1) and without (0) pervious heart issue

Sex	Yonger_Then_30
Female	94
Male	171

(1)

Sex	Yonger_Then_30
Female	75
Male	201

(0)

Data Analysis

High heart attack risk with following factors:

- Diabetes
- Family History
- Smoking
- Obesity
- Alcohol Consumption

Heart_Attack_Risk	Number_of_Patients
1	284
0	494

Heart Attack Risk with Sleeping

Heart_Attack_Risk	Avg_Sleeping_Time_Per_Day
1	6.974195603695445
0	7.051031294452347

Heart Attack Risk with Exercise and Diet

Diet	Avg_Exercise_Hours_Per_Week
Unhealthy	10.179744647675276
Average	9.987016034593525
Healthy	10.132370137728234

Heart_Attack_Risk = 1

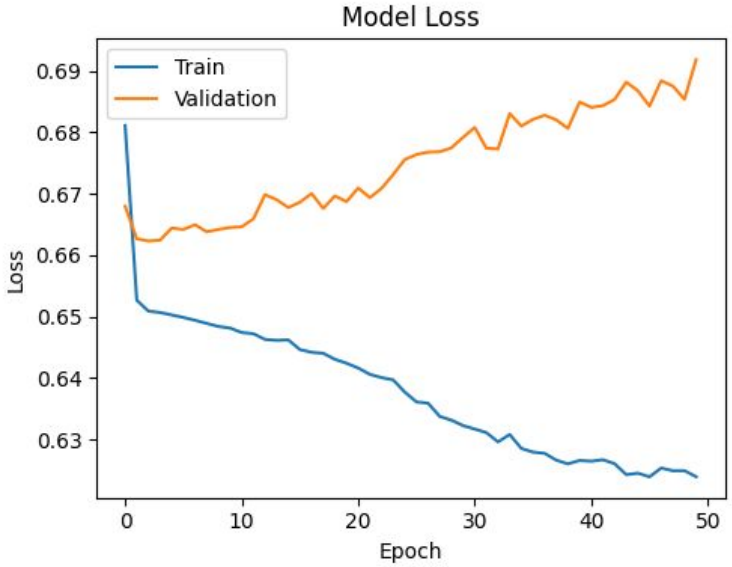
Diet	Avg_Exercise_Hours_Per_Week
Unhealthy	9.884804984010403
Average	9.934329292998868
Healthy	10.07845718923255

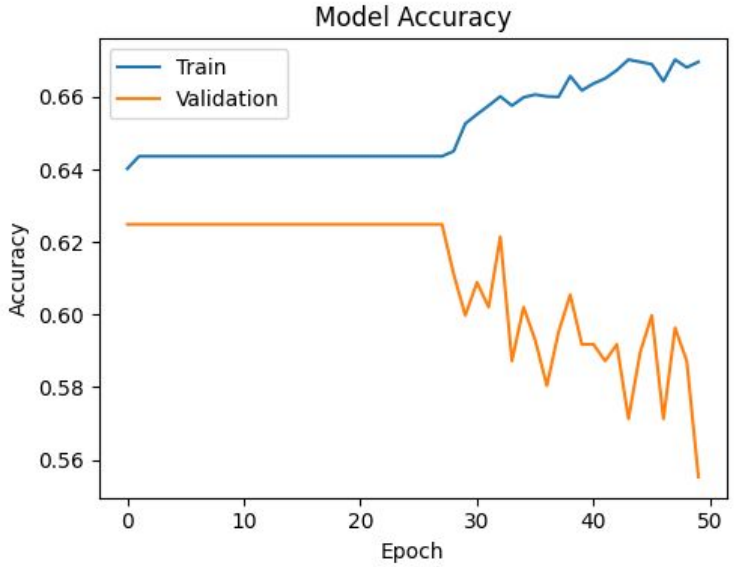
Heart_Attack_Risk = 0

Heart Attack Risk Deep Learning Modeling

A deep neural network model is created by using the TensorFlow's Keras library:

- several dense layers with ReLU activations
- a final dense layer with the sigmoid activation





Conclusion

With ECG Classification Model:

- train a ViT (Vision Transformer) model to identify different heart conditions based on the input ECG images

Based on the heart attack dataset, it seems the relationship between heart attack risk and lifestyle choices is correlative, but not a huge impact.

With Deep Learning Model on syndicated date, it used the dataset to generate random values; and this model is capable of "overfitting".

We also develop a model with dataset contains information of real patients, which is available in our GitHub.