

FAI Healthcare Project

Initial Project Proposal



Amanda Pang, Shreyas Raman, Devansh Thakkar

Khoury College of Computer Science, Northeastern University

CS5100: Foundations of Artificial Intelligence

Dr. Rajagopal Venkatesaramani

March 27, 2025

Topic: Developing an AI framework to detect and diagnose early risk factors, signs, and symptoms of ischemic heart disease, also known as Coronary Artery Disease (CAD).

Ischemic Heart Disease has been cited by the WHO as the leading cause of death worldwide, responsible for 13% of total deaths in 2021 (WHO, 2025). The risk factors, causes, signs, and early symptoms of the disease are well documented in medical literature, driven by robust research funding.

Due to the prevalence of ischemic heart disease worldwide, and the wealth of research that has already been conducted to identify patients at risk of developing heart disease, we have plenty of prior literature from which to draw study designs and proofs of concept. Since current research has identified specific biomarkers, lab results, and imaging results to be indicative of CAD, our group can start our exploratory analysis in these areas. This will allow our group to focus on developing a comprehensive Artificial Intelligence (AI) approach to determine which features most strongly predict the outcome of CAD, without being limited by preexisting clinical expertise.

A limitation of any healthcare project is the availability of high-quality patient medical record data. Per discussion with the professor, we have proposed our strongest idea for developing an accurate, robust, and helpful AI agent, and will require assistance obtaining a high-quality, validated medical record dataset to use as input data. We may need to rely on synthetic datasets generated for experimental AI projects, but will defer until we have confirmed our project topic.

### Existing Approaches:

* Deep Learning
* Pooled network forest, funnel plots, league table (Bashar et al, 2022)
* Random forest, DL Long short-term memory (LSTM) (Yu et al, 2020)

(Bashar et al, 2022)

“Seventeen studies, with a total of 285,213 patients with CVDs, were included in the network meta-analysis. The statistical evidence indicated that the

* DL algorithms performed well in the prediction of heart failure with AUC of 0.843 and CI [0.840–0.845], while in the
* ML algorithm, the gradient boosting machine (GBM) achieved an average accuracy of 91.10% in predicting heart failure.
* An artificial neural network (ANN) performed well in the prediction of diabetes with an OR and CI of 0.0905 [0.0489; 0.1673].
* Support vector machine (SVM) performed better for the prediction of stroke with OR and CI of 25.0801 [11.4824; 54.7803].
* Random forest (RF) results performed well in the prediction of hypertension with OR and CI of 10.8527 [4.7434; 24.8305].
* The findings of this work suggest that the DL models can effectively advance the prediction of and knowledge about heart failure, but there is a lack of literature regarding DL methods in the field of CVDs.
  + As a result, more DL models should be applied in this field. To confirm our findings, more meta-analysis (e.g., Bayesian network) and thorough research with a larger number of patients are encouraged.”

### Known CAD Indicators (Advocate Healthcare, 2025):

Biomarkers/Labs:

* HDL (bad cholesterol)
* apoA-I (apolipoprotein A-I)
* LDL (good cholesterol)
* A1C
* Troponins
* D-dimer

Reports:

* ECG
* Cardiac Catheterization
* TEE (Transesophageal Echocardiogram)
* CT/MRI imaging reports
* Stress test

Diagnosis Indicators (patient has CAD):

* History of MI (myocardial infarction), stroke, ischemia, aneurysm
* History of arrhythmias, flutter, bradycardia, tachycardia

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Feature Attributes** | **Feature Description** | **Coding** |
| 1 | Age | age of the patient | [years] |
|  | Sex | sex of the patient | [M: Male, F: Female] |
|  | ChestPainType | chest pain type | [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic] |
|  | RestingBP | resting blood pressure | [mm Hg] |
|  | Cholesterol | serum cholesterol | [mm/dl] |
|  | FastingBS | fasting blood sugar | [1: if FastingBS > 120 mg/dl, 0: otherwise] |
|  | RestingECG | resting electrocardiogram results | [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria] |
|  | MaxHR | maximum heart rate achieved | [Numeric value between 60 and 202] |
|  | ExerciseAngina | exercise-induced angina | [Y: Yes, N: No] |
|  | Oldpeak | oldpeak = ST | [Numeric value measured in depression] |
|  | ST\_Slope | the slope of the peak exercise ST segment | [Up: upsloping, Flat: flat, Down: downsloping] |
|  | HeartDisease | output class | [1: heart disease, 0: Normal] |

# 

# Framingham Report

The methodology in this section utilizes all measurements determined on Exams 1 through 15 for those risk factors recorded virtually every two years and relates the risk factors to the occurrence of an event within two years after the exam. Other characteristics measured on only a few examinations are not included in -this monograph. We refer to this approach of employing the- biennial observations as the cross-sectional pooling method.

This method evaluates each two-year interval as a new short-term followup study. After being characterized at entry into the study, persons are characterized anew at each following biennial examination. Hence, a person who attended twelve of the fifteen examinations during the 30-year followup contributes the information of twelve persons who enter- the study at the beginning of a two-year cycle with the risk factors measured at the twelve examinations. More accurately, this person contributes the information of twelve person exams. To implement this approach, an observation is generated for each examination. The information obtained on the 15 two-year intervals is then pooled to obtain a file from which two-year predictions can be examined.

This method is to be distinguished from the long-term perspective described earlier in which observations only from exam 1 are employed to examine the development of disease over 30 years of followup. The cross-sectional pooling method as implemented in this section considers only the next two years of followup, given an individual's current age, sex, and risk factor status. The inherent assumption is that only the current risk factor status of an individual is needed to predict the risk of disease in the next two years.

A table takes into account risk factors on all of the first fifteen examinations and the incidence of the specified event in the fifteen biennial intervals of the 30-year followup and consolidates this by the cross-sectional pooling method into an average annual incidence rate by age, sex, and level of the risk factor. The statistics shown are:

A white paper with black text

Description automatically generated

A white background with black text

Description automatically generated

The numbering scheme on the tables indicates the number of the event first and then the number of the risk factor. For example, the table for Myocardial Infarction and Cigarettes smoked per day is numbered as **3-13**.

Each **table displays** for each sex descriptive information on the relationship between the risk factor and the risk of the specified subsequent event in a two-year period.

Logistic regression coefficients, for the risk factor, including univariate, bivariate with age at exam, and multivariate **analyses,** are displayed at the bottom of the table.

The descriptive portion of the table displays annual rates for 6 age groups: 35-44, 45-54, 55-64, 65-74, 75-84, 85-94. Age-adjusted annual rates computed by the direct method are also given for age groups 35-64 and 65-94.

Age-specific logistic regressions, using specified the above lo-year age groups, and regressions for age groups 35-64 and 65-94 are also based upon the t-do-year cross-sectional pooling method and indicate the risk of the event in the next two years of follow-up among persons free of the event at the beginning the two-year interval. However, the rates are expressed as the average annual rate per 1000.

The header of each table indicates the sex, event, risk factor, and the population at risk for each table. Each risk factor-event combination is presented on one page with the top half for males and the bottom for females.

**Risk Factor Description:**

The range of each risk factor, from the lowest to the highest value observed in the fifteen exams, is displayed in each table. The tables for hematocrit and vital capacity-height index are the only tables which have different ranges for men and women.

Each individual is characterized by his or her value at an exam. If that value is unknown, the most recent, known value at a previous exam is used. An exception to this rule is diabetes mellitus (risk factor 7). Once a person is diagnosed as being diabetic, that person retains that diagnosis on all subsequent exams.

# Data Dictionary (Framingham)

1. Systolic BP
2. Diastolic BP
3. Hypertension

At examination, a subject had two blood pressure readings taken by the examining physician(s). If both readings were “abnormal” the subject had definite hypertension; if both readings were “normal” the subject had normotension; with any other combination of readings the subject had borderline hypertension.

A blood pressure reading was “abnormal” if either the systolic or the diastolic component was “abnormal”. The reading was “normal” if both systolic and diastolic parts were “normal”. A systolic pressure was called “normal” when under 140 mm Hg, “abnormal” when 160 or greater. A diastolic pressure was called “normal” when under 90 mm Eg, “abnormal” when 95 or greater.

Thus, a person was a definite hypertensive if both of the following conditions held:

1. the first systolic pressure was 160 or greater **or** the first diastolic pressure was 95 or greater.

AND

2. the second systolic pressure was 160 or greater **or** the second diastolic pressure was 95 or greater.

For purposes of analysis in this monograph, hypertension was evaluated in two fashions: (1) one utilizing only blood pressure readings as described above and (2) one in which individuals not treated with anti-hypertensive medication were classified on the basis of blood pressure while persons treated with anti-hypertensive medication were considered as definite hypertensives.

1. Serum cholesterol (mg/100ml)
2. Hematocrit
3. Blood Glucose
4. Diabetes mellitus
5. Glucose in urine
6. Glucose Intolerance
7. Metropolitan Relative Weight
8. Vital Capacity
9. Heart Rate
10. Cigarettes smoked per day
11. Albumin in urine
12. Heart enlargement by X-Ray
13. Left Ventricular Hypertrophy
14. Intraventricular conduction defect
15. Nonspecific T-wave or ST-segment abnormality by ECG

# A screenshot of a computer Description automatically generatedEDA UCI Dataset

A graph with numbers and letters

Description automatically generated with medium confidence

## A graph of different colored squares Description automatically generated with medium confidenceImpute

## A screenshot of a computer Description automatically generated

# A group of graphs showing different sizes and colors Description automatically generated with medium confidence

# Exploratory Data Analysis (EDA)

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 4240 entries, 0 to 4239

Data columns (total 16 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 male 4240 non-null int64

1 age 4240 non-null int64

2 education 4135 non-null float64

3 currentSmoker 4240 non-null int64

4 cigsPerDay 4211 non-null float64

5 BPMeds 4187 non-null float64

6 prevalentStroke 4240 non-null int64

7 prevalentHyp 4240 non-null int64

8 diabetes 4240 non-null int64

9 totChol 4190 non-null float64

10 sysBP 4240 non-null float64

11 diaBP 4240 non-null float64

12 BMI 4221 non-null float64

13 heartRate 4239 non-null float64

14 glucose 3852 non-null float64

15 TenYearCHD 4240 non-null int64

dtypes: float64(9), int64(7)

memory usage: 530.1 KBA graph of age

Description automatically generated

from matplotlib import pyplot as plt

df['age'].plot(kind='hist', bins=20, title='age')

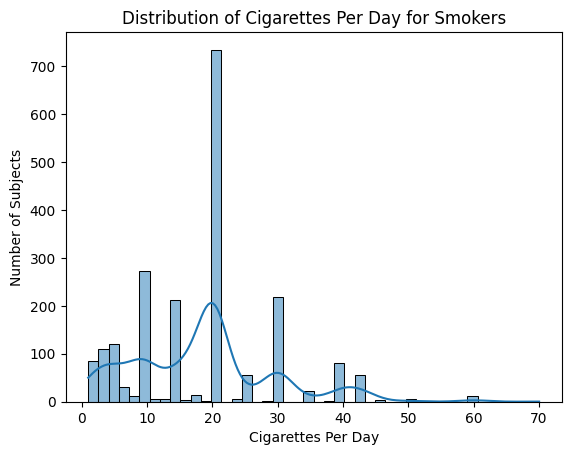
plt.gca().spines[['top', 'right',]].set\_visible(False)

A group of graphs showing different sizes of data

Description automatically generated with medium confidence

A group of graphs and diagrams

Description automatically generated



# A blue and orange pie chart Description automatically generatedCategorical

A blue and orange pie chart

Description automatically generated

A blue and orange pie chart

Description automatically generated

## Key Insights from the Correlation Matrix

1. **Age vs. HeartDisease (0.2254)**
   * Highest positive correlation with heart disease.
   * Suggests higher age is somewhat linked to a higher risk of heart disease.
2. **Education vs. HeartDisease (-0.0542)**
   * Strongest negative correlation with heart disease.
   * Indicates more education is slightly associated with a decreased chance of heart disease.
3. **Systolic BP vs. HeartDisease (0.2163)**
   * Slight positive correlation.
   * Suggests higher Systolic BP is somewhat linked to a higher risk of heart disease.
4. **Prevalent Hypertension vs. HeartDisease (0.1775)**
   * Weak positive correlation.
   * Diagnosis of High Blood Pressure may be associated with increased heart disease risk.
5. **Diastolic BP vs. HeartDisease (0.1451)**
   * Weak positive correlation.
   * Suggests higher Diastolic BP is somewhat linked to a higher risk of heart disease.
6. **Glucose vs. HeartDisease (0.1256)**
   * Weak positive correlation.
7. **Diabetes vs. HeartDisease (0.0973)**
   * Weak positive correlation.
8. **Sex vs. HeartDisease (0.0884)**
   * Weak positive correlation.
   * Suggests men are at slightly higher risk of heart disease.
9. **Blood Pressure Medication vs. HeartDisease (0.0875)**
   * Weak positive correlation.
   * Suggests men are at slightly higher risk of heart disease.
10. ?**Notable Feature Interactions**
    * **Age and MaxHR** are negatively correlated (-0.382), reflecting the natural decline in max heart rate with age.
    * **FastingBS and Cholesterol** are negatively correlated (-0.261), indicating complex metabolic or treatment-related factors.

|  |  |
| --- | --- |
| Feature | Correlation Score |
| **age** | 0.22540774 |
| **sysBP** | 0.21637383 |
| **prevalentHyp** | 0.17745756 |
| diaBP | 0.14511159 |
| glucose | 0.12559036 |
| diabetes | 0.09734424 |
| male | 0.08837357 |
| BPMeds | 0.08751945 |
| totChol | 0.08236854 |
| BMI | 0.07530032 |
| prevalentStroke | 0.06182263 |
| cigsPerDay | 0.05775521 |
| heartRate | 0.02290661 |
| currentSmoker | 0.0194485 |
| education | -0.0542485 |
|  |  |

## Feature Importance XGBoost

# Uses best parameters: Best parameters: {'colsample\_bytree': 0.8, 'learning\_rate': 0.01, 'max\_depth': 3, 'n\_estimators': 300, 'subsample': 0.8}

weight

A graph with numbers and a bar chart

Description automatically generated with medium confidence

Accuracy: 0.8573113207547169

AUC Score: 0.7036613400616765

Confusion Matrix:

[[725 0]

[121 2]]

importance\_type = gain

A graph with numbers and a bar chart

Description automatically generated

importance\_type = cover

A graph with numbers and a bar chart

Description automatically generated

## SHAP

A graph of different colored lines

Description automatically generated

## Correlation Heatmap

A screenshot of a graph

Description automatically generated

## Pair Plot

A group of blue and orange dots

Description automatically generated

# 10 year CDH vs other Features

# A graph with red and blue bars Description automatically generated

# A graph with red and blue bars Description automatically generated

# A graph with red and blue bars Description automatically generated

# A graph with numbers and a bar chart Description automatically generated

# A graph with a bar and a number of different colored bars Description automatically generated with medium confidence

# A graph with a bar and a number of bars Description automatically generated with medium confidence

# A graph with red and blue bars Description automatically generated

# A graph with a bar and a number of people Description automatically generated with medium confidence

# A graph with red lines Description automatically generated

# A graph of a number of data Description automatically generated with medium confidence

# A graph of a number of data Description automatically generated

# A graph with red and blue lines Description automatically generated

# A graph of different colored lines Description automatically generated

# A graph of blood glucose Description automatically generated

# A screenshot of a computer screen Description automatically generatedAfter filling in NULL

# A screenshot of a computer screen Description automatically generated

A graph of a number of cholesterol levels

Description automatically generated

A graph of age distribution

Description automatically generated

A graph of heatmap

Description automatically generated

# Balancing

{0: 3596, 1: 644} {0: 3595, 1: 2876}

# A comparison of numbers and numbers Description automatically generatedRandom Forest Results

A screenshot of a computer

Description automatically generated

A graph of a graph with numbers and a number of squares

Description automatically generated with medium confidence

## References

Baashar, Yahia. Gamal Alkawsi, Hitham Alhussian, Luiz Fernando Capretz, Ayed Alwadain, Ammar Ahmed Alkahtani, Malek Almomani. (2022, February 24). *Effectiveness of Artificial Intelligence Models for Cardiovascular Disease Prediction: Network Meta-Analysis.* Wiley Online Library.<https://onlinelibrary.wiley.com/doi/full/10.1155/2022/5849995>

Luc, G. Jean-Marie Bard, Jean Ferrières, Alun Evans, Philippe Amouyel, Dominique Arveiler, Jean-Charles Fruchart, Pierre Ducimetière. (2022, May 23). *Value of HDL Cholesterol, Apolipoprotein A-I, Lipoprotein A-I, and Lipoprotein A-I/A-II in Prediction of Coronary Heart Disease: The PRIME Study.* AHA|ASA Journals. <https://www.ahajournals.org/doi/full/10.1161/01.ATV.0000022850.59845.E0>

Yu, Jaehak. Sejin Park, Soon-Hyun Kwon,Chee Meng Benjamin Ho, Cheol-Sig Pyo, and Hansung Lee. (2020). *AI-Based Stroke Disease Prediction System Using Real-Time Electromyography Signals.* MDPI Journals.

<https://www.mdpi.com/2076-3417/10/19/6791>

Advocate Health Care. (2025). *Ischemic heart disease.*

<https://www.advocatehealth.com/health-services/advocate-heart-institute/conditions/ischemic-heart-disease>

World Health Organization (WHO). (2024, August 7). *The top 10 causes of death*.

<https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>