

FAI Healthcare Project

Final Report



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

# New Notes

* Framingham was selected as ideal due to proximity to Harvard Medical School. Town of 28,000 middle class predominantly European, considered to be representative of the US at the time.
* Active recruitment for random sampling
* First major findings was that high blood pressure (systolic >= 160/95 mmHg resulted in an almost 4x higher chance of a CHD incident. Showed that it was systolic, not diastolic blood pressure that had a significant correlation to CHD.
  + This was in contradiction to widespread beliefs at the time, that disregarded hypertension, or focused on high diastolic BP.

Framingham: Started in 1948 by President Truman, who signed the “National Heart Act”, for President Roosevelt, who died of cardiovascular disease (along with 1 in 3 Americans).

* Huge grant awarded for a 20 year epidemiological heart study, to study the expression of CAD in “normal” unselected population, and determine risk factors that predisposed the development of CAD, through clinical and lab exams, and long term follow-up.
* First major findings was that high blood pressure (systolic >= 160/95 mmHg resulted in an almost 4x higher chance of a CHD incident. Showed that it was systolic, not diastolic blood pressure that had a significant correlation to CHD.
  + This was in contradiction to widespread beliefs at the time, that disregarded hypertension, or focused on high diastolic BP.
  + Framingham analysis established that hypertension was the leading risk factor for heart failure.
  + Led to many sub-studies that recruited family members and descendents of original cohort, laying the foundation for genetic risk factor analysis.
  + Led to many sub-studies that recruited family members and descendents of original cohort, laying the foundation for genetic risk factor analysis. In 2006, NIH funded project to sequence the genomes of the entire Framingham study, supporting global effort to identify common genetic risk variants.

* Epidemiological cohort studies like Framingham, contributed towards the shift in medical attitudes/perceptions. Moving away from treating patients only after they develop cardiovascular disease.
  + More focus on preventing disease development in identifiable high risk populations, and early interventions.
* The term “Risk Factor” was coined during this time, as researchers identified CHD risk factors including: hypertension, hyperlipidemia (high cholesterol), and diabetes.
  + Developed the first multivariable logistic model to compute risk scores (previously only multiple cross-classification was used, where each cell in the table corresponds to a combination of risk factors). Led to development of “risk profiles”: Framingham Risk Score for CHD. They replaced continuous risk factor values with categorical values, allowing clinicians to quickly obtain risk estimates using lookup tables.
* As the study progressed, and more patients developed heart failure, researchers were able to identify and quantify a set of clinical criteria for heart failure, still in use today.

**The Natural History of Congestive Heart Failure: The Framingham Study**

<https://www.nejm.org/doi/10.1056/NEJM197112232852601?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed>

In 1971, investigators McKee, et al published their findings, using data collected over 16 years, from 5192 “healthy” people, ages 30-62, without congestive heart failure. They observed that obvious evidence of congestive heart failure developed in 142 people, with an overall higher incidence rate in men.

Findings: (across all cases)

* Hypertension preceded heart failure in 75% of all cases.
* Incidence of CHD was 39%. In 29% of all cases, CHD was accompanied by hypertension.
* Rheumatic Heart disease was present in 21% of all cases, which was accompanied by hypertension in 11% of all cases.
* Probability of dying of CHD within 5 years from the time of onset was 62% for men, and 42% for women.

Subsequent studies using the ongoing Framingham Heart Study data later identified additional cardiac risk factors including: increased left ventricle diameter, asymptomatic LV systolic dysfunction, diabetes, hyperlipidemia (high cholesterol), all still highly focused on today. The quantification of various presentations and progressions of heart failure led to standardized assessments and diagnosis criteria, strengthening future data collection, analysis and treatments.

The Framingham data was valuable as a control comparison in the 1990s, to demonstrate the efficacy of new medications: beta blockers and ACE-inhibitors (Levy et al) in reducing 5-year mortality prognosis. One of the most valuable contributions was the demonstration that non-rheumatic atrial fibrillation was a strong risk factor for stroke / ischemic heart disease, leading to controlled trials on newer classes of medications: anticoagulants and anti-arrhythmia, for treatment of AFib.  
  
Later cohorts included the family members and descendents of original participants, laying the groundwork for the future identification of genetic risk factors.

Abstract

The natural history of congestive heart failure was studied over a 16-year period in 5192 persons initially free of the disease. Over this period, overt evidence of congestive heart failure developed in 142 persons. In almost every five-year age group, from 30 to 62 years, the incidence rate was greater for men than for women. Although the usual etiologic precursors were found, the dominant one was clearly hypertension, which preceded failure in 75 per cent of the cases. Coronary heart disease was noted at an earlier examination in 39 per cent, but in 29 per cent of the cases it was accompanied by hypertension. Precursive rheumatic heart disease, noted in 21 per cent of cases of congestive heart failure, was accompanied by hypertension in 11 per cent. Despite modern management, congestive heart failure proved to be extremely lethal. The probability of dying within five years from onset of congestive heart failure was 62 per cent for men and 42 per cent for women.

Levy Paper

Among 9,405 Framingham Heart Study participants (47% male) followed up from September 1948 to June 1988, congestive heart failure developed in 652.

The mean age at the diagnosis of heart failure was 70.0 + 10.8 years. These 331 men and 321 women were followed up for a median of 1.8 years after the diagnosis of congestive heart fails (mean 3.9 + 5.4 years; range 0 days to 35.8 years). There were 17 additional cases of congestive heart failure diagnosed at the first Framingham Heart Study examination.

### A screenshot of a report Description automatically generatedFramingham Population Cohorts

A screenshot of a medical report

Description automatically generated

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

# EDA UCI Dataset

## 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')

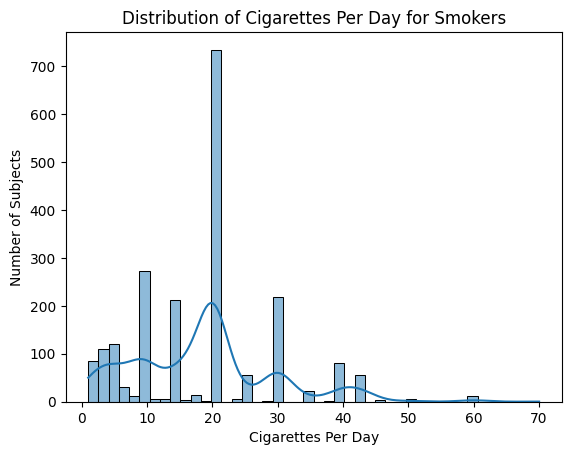
plt.gca().spines[['top', 'right',]].set\_visible(False)

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>

## https://pmc.ncbi.nlm.nih.gov/articles/PMC4159698/