

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

# Introduction:

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

Our research aims to develop an Artificial Intelligence (AI) and Machine Learning (ML) framework to detect early indicators of ischemic heart disease—also referred to as Coronary Artery Disease (CAD), Coronary Heart Disease (CHD), or Congestive Heart Failure (CHF). IHD refers to conditions that restrict blood flow to the heart, most often due to arterial stenosis, aneurysms, or embolic (stroke) events. While other heart diseases (e.g., arrhythmias, congenital defects, valvular disease, cardiomyopathy, and endocarditis) share overlapping symptoms, their clinical documentation is seldom present in publicly available datasets, limiting their inclusion in ML applications.

Initial exploration of approaches included building a “Heart Disease Classification” model, which would learn feature delineations between the presentations of heart disease, and assign the most likely diagnosis for a given training example. However, cardiac datasets with diagnosis details are not publicly available, and the recurrence of severe limitations in dataset availability, documentation, and quality, meant we ultimately selected the “Framingham” and “UCI” datasets as the most robust and diverse representation of heart disease patients.

Our research goal is to develop an Artificial Intelligence (AI) and Machine Learning (ML) framework to detect and diagnose early risk factors, signs, and symptoms of ischemic heart disease, also known as Coronary Artery Disease (CAD), Coronary Heart Disease (CHD), Cardiovascular disease, or Congestive Heart Failure (CHF). Ischemic heart disease refers to the subset of diseases that cause insufficient blood to reach the heart, which may be caused by stenosis (narrowing or blockages of arteries), aneurysms (tear in artery causing blood to leak), or ischemic stroke (embolic blood clot). Other types of heart disease include: arrhythmias, congenital heart defects, heart valve defects, cardiomyopathy (disease of heart muscle), rheumatic heart disease, and endocarditis. As these non-ischemic heart diseases have overlapping symptoms and can be difficult to diagnose without modern imaging equipment and clinical context, their documentation is extremely limited in publicly available datasets.

# Background

The momentous Framingham Heart Study, initiated in 1948 under President Truman’s “National Heart Act,” was the first long-term epidemiological empirical research of cardiovascular disease in the United States, which killed 1 in 3 Americans at the time. The Framingham Heart Study was designed to identify risk factors for CAD, quantify the longitudinal expression of CAD in initially healthy adult populations, and determine risk factors that predisposed the development of CAD. The town of Framingham, MA was selected due to its proximity to Harvard Medical School, and its largely European middle-class citizens were considered representative of American demographics at the time. Risk factors were measured, assessed, and quantified through clinical and lab exams conducted every two years, and observations evaluated against the target outcome by two-year and 30-year long term follow-up (McKee et al., 1971).

The first major findings published by McKee, et al, 1971, was that high blood pressure (systolic >= 160/95 mmHg) resulted in an almost four times higher chance of a CHD incident, establishing that it was systolic, not diastolic, hypertensive blood pressure that had a significant correlation to CHD. This was in contradiction to widespread beliefs at the time, which either disregarded hypertension, or focused on high diastolic Blood Pressure (BP).

In 1971, McKee, et al developed the first multivariable logistic model to compute risk scores, given an individual's current age, sex, and risk factor status. This multivariable logistic analysis facilitated the development of “risk profiles”, through the establishment of the “Framingham Risk Score for CHD”. Previously, only multiple cross-classification analysis was used, where each cell in the table corresponded to a combination of risk factors. However, this method of storing every possible combination of risk factors (similar to Dynamic Programming) is not scalable to handle larger numbers of features. Seamlessly implementing their translational work, McKee et al replaced continuous risk factor values with categorical values, allowing clinicians to quickly obtain risk estimates using lookup tables (Mahmood, et al, 2014).

Since current research has identified specific biomarkers, lab results, and imaging results to be indicative of CAD, our literature review began by exploring existing cardiac datasets, study design, and analysis results as proofs of concept.

### 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)** performed a meta-analysis of 17 studies and 285,213 patients with Cardiovascular Diseases, with the goal of comparing their Deep Learning (DL) model against other machine learning models. Their results showed that the DL model performed well, in comparison to more established models.

* Deep Learning (DL): AUC = 0.843; CI = [0.840–0.845]
* Gradient Boosting Machine (GBM): 91.1% accuracy
* Artificial Neural Networks (ANN): OR = 0.0905; CI = [0.0489–0.1673]
* Support Vector Machine (SVM): OR = 25.08; CI = [11.48–54.78]
* Random Forest (RF): OR = 10.85; CI = [4.74–24.83]

Our approach leaned heavily on reviewing preexisting analyses to help identify a high-quality dataset, pre-processing the raw data to convert categorical fields into numeric representations, and highlight known inconsistencies within datasets. This allowed our research scope to focus on developing a comprehensive Machine Learning (ML) approach, prioritizing interpretability, to determine which features most strongly predict the target outcome of a CAD diagnosis, without being limited by preexisting clinical expertise.

# Exploratory Data Analysis

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

At this time, numerous risk factors for CAD have been identified, and medical imaging studies have advanced to provide more revealing insights into disease progression and diagnosis. Based on preliminary research, we identified three main categories of relevant patient data that are relevant to our predictive model’s accuracy and generalizability:

1. Biomarkers/Labs:

* HDL (High-Density Lipoprotein – "good" cholesterol)
* LDL (Low-Density Lipoprotein – "bad" cholesterol)
* apoA-I (apolipoprotein A-I)
* HbA1C (Hemoglobin A1C – measures average blood sugar level as %)
* Troponins\*
* D-dimer\*

Regular monitoring of patients’ laboratory values can be strong indicators of the start of a disease progression. The two lab tests marked “\*” are not standard of care, and usually only performed if the patient is suspected of acute heart failure. However, elevated Troponins and D-dimers are strong biomarkers of heart inflammation and disease and, if available, would be valuable predictors of Coronary Artery Disease (CAD). The remaining lab tests are part of standard metabolic and comprehensive blood panels likely performed whenever a patient is due for blood work.

1. Clinical Procedure Reports:

* EchoCardioGram (ECG)
* Cardiac Catheterization
* TEE (Transesophageal Echocardiogram)
* TTE (Transthoracic Echocardiogram)
* CT/MRI imaging reports
* Stress Test\*

Interpretation of imaging reports must be done by a trained professional. To incorporate imaging data, we would need access to the physician interpretation of results, and then manually label the areas of the image that correspond to certain clinical findings or occlusions. This was deemed outside of our project’s scope, therefore, “Stress Test\*” results are the only data from clinical procedure reports included in our raw training and test datasets, at this feature is represented as a numeric integer value. The feature “thalach” from the Cleveland UCI dataset represents the maximum heart rate achieved during the Exercise Stress Test.

1. Prior History of Diagnosis (indicators that patient has already developed some kind of CAD):

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

### Dataset Evaluation

Based on existing research and analysis performed on publicly available cardiac datasets, we identified three core cardiac datasets that have been analyzed by the research and Kaggle community, and assessed each datasets’ training potential for inclusion in our final model.

1. CDC Behavioral Risk Factor Surveillance System (BRFSS): survey data collected annually for 400,000+ adults over the phone, totaling 330 features collected annually since 1984. We explored the dataset and attempted to train preliminary models, however discovered that the sheer quantity of data made model training infeasible; given our computational limitations, our most powerful desktop was not able to train a single model.
2. University of California Irving (UCI) Dataset (1,190 patients): the most extensively used and studied clinical cardiac dataset for machine learning. The UCI dataset consists of merged data aggregated from five studies over 11 common features, subsets reported below:

* Cleveland: 303 observations
* Hungarian: 294 observations
* Switzerland: 123 observations
* Long Beach VA: 200 observations
* Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Unfortunately, our research uncovered that the UCI dataset is extremely convoluted, contains hundreds of duplicates, and that the raw datasets hosted by UCI are corrupted and unavailable upon request (Simmons, 2021). Despite purportedly being the most popular cardiac dataset for Machine Learning (ML) applications, it was discovered that all ML training was only performed on the “Cleveland” subset of 303 patients. Further inspection of the other four datasets showed duplicates, missing features, and a lack of metadata documentation on the pre-processing performed by UCI. Per discussions with the professor, we were advised to focus our analyses and model training on the Cleveland and Framingham datasets.

1. Framingham Heart study: only longitudinal study, not as high-quality data (first empirical research study in USA), but foundationally important findings and implications. Started in 1948 with continual interim analyses and additional cohorts.

Moving forward with our initial Framingham dataset, we decided to test train five machine learning models: Logistic Regression, XGBoost, KNN, Random Forest, and Neural Networks.

### Framingham Methods & Results: McKee 1971

Population: 5,192 individuals (ages 30–62), followed for 16 years

Findings:

* 75% of heart failure cases were preceded by hypertension
* CHD incidence: 39%
  + accompanied by hypertension in 29% of all cases
* Rheumatic heart disease: 21% of cases
  + accompanied by hypertension in 11% of all cases
* 5-year mortality from CHD: 62% in men, 42% in women

### Framingham Methods & Results: Levy 1993

Population: 9,405 participants (47% male)

Findings:

* CHF developed in 652 individuals
  + 331 men; 321 women
* Mean age at diagnosis: 70.0 ± 10.8 years
* Median post-diagnosis follow-up time: 1.8 years (mean 3.9 ± 5.4 years; range 0–35.8 years)

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.

# Exploratory Data Analysis (Framingham)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Feature Name (label)** | **DescriptiveName** | **Description** | **Raw Coding** | **Definitions** |
| 1 | **sex** | Sex | sex of the patient | [1: Male, 0: Female] |  |
| 2 | **age** | Age | age of the patient | [years] |  |
| 3 | **education** | Education |  |  |  |
| 4 | **currentSmoker** | Current Smoker |  | 0: no smoker;  1: yes smoker |  |
| 5 | **cigsPerDay** | Cigarettes per Day | Number of Cigarettes Smoked per Day |  |  |
| 6 | **BPMeds** | Blood Pressure Medication |  |  |  |
| 7 | **prevalentStroke** | Stroke (recorded or diagnosed) |  |  |  |
| 8 | **prevalentHyp** | Hypertension (high blood pressure) | Diagnosed by  1.) abnormal BP on exam;  2.) Taking anti-hypertensive medications |  |  |
| 9 | **diabetes** | Diabetes mellitus | Diagnosed by  1.) blood glucose > 150 mg/100mL; 2.) receiving treatment for Diabetes; 3.) Record of diagnosis | 0: no;  1: yes |  |
| 10 | **totChol** | Cholesterol | serum cholesterol | [mm/dl] | (millimeters per deciliter) |
| 11 | **sysBP** | Systolic BP |  |  |  |
| 12 | **diaBP** | Diastolic BP |  |  |  |
| 13 | **BMI** | Body Mass Index |  |  |  |
| 14 | **heartRate** | Heart Rate | heart rate per minute, recorded by ECG |  |  |
| 15 | **glucose** | Blood Glucose levels |  | [mg/100mL] | miligrams per mililiter |
| 16 | **TenYearCHD** | Presence of Coronary Heart Disease | Target Outcome | 0: no CHD;  1: yes CHD | Qualifying Events:myocardial infarction, coronary insufficiency, angina pectoris, sudden death from CHD, non-sudden death from CHD |

# Exploratory Data Analysis (Cleveland)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Feature Name (label)** | **DescriptiveName** | **Description** | **Raw Coding** | **Definitions** |
| 1 | **age** | Age | age of the patient | [years] |  |
| 2 | **sex** | Sex | sex of the patient | [1: Male, 0: Female] |  |
| 3 | **cp** | ChestPainType | chest pain type | [0: Typical Angina;  1: Atypical Angina;  2: Non-Anginal Pain;  3: Asymptomatic] |  |
| 4 | **trestbps** | RestingBP | resting blood pressure (on admission to hospital) | [mm Hg] | (millimeters of mercury) |
| 5 | **chol** | Cholesterol | serum cholesterol | [mm/dl] | (millimeters per deciliter) |
| 6 | **fbs** | Fasting Blood Sugar | fasting blood sugar | [1: if FastingBS > 120 mg/dl, 0: otherwise] |  |
| 7 | **restecg** | RestingECG | resting electrocardiogram results | [0: Normal; 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2: showing probable or definite left ventricular hypertrophy by Estes' criteria] | The Romhilt-Estes (RE) score assigns points for the presence of certain ECG findings, and a score of 4 is considered probable LVH, while a score of 5 or greater indicates definite LVH. Left ventricular hypertrophy (LVH) means the muscle of the heart's main pump (left ventricle) has become thick and enlarged. |
| 8 | **thalach** | MaxHR | maximum heart rate achieved during Stress Test | [Numeric value between 60 and 202] |  |
| 9 | **exang** | ExerciseAngina | exercise-induced chest pain | [1: Yes, 0: No] |  |
| 10 | **oldpeak** | Oldpeak | oldpeak = ST depression induced by exercise relative to rest | [Numeric value measured in depression] | ST depression induced by exercise relative to rest |
| 11 | **slope** | ST\_Slope | the slope of the peak exercise ST segment | [Up: upsloping; Flat: flat; Down: downsloping] |  |
| 12 | **ca** | NumVessels | Number of major vessels colored by Fluoroscopy | [0-3] | Use continuous X-rays and contrast dyes to visualize how blood flows (or does not flow) through vessels |
| 13 | **thal** | Congenital | Normal or abnormal heart | 3 = normal; 5 = fixed defect; 7 = reversable defect |  |
| 14 | **num** | HeartDisease | output class | [1: heart disease, 0: Normal] | Value 0: < 50% diameter narrowing; Value 1: > 50% diameter narrowing (stenosis) |

# Methods:

### Logistic Regression

### XGBoost

### KNN

### Random Forest

### Neural Networks

# Results:

### Logistic Regression

### XGBoost

### KNN

### Random Forest

### Neural Networks

# Discussion:

# Conclusion:

Epidemiological cohort studies, like Framingham, contributed towards the shift in medical attitudes and perceptions of the time. Moving away from treating patients only after they develop cardiovascular disease, more focus was placed on preventing disease development in identifiably higher risk populations, and implementing early interventions to cut off disease progression. 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 invaluable as a control comparison cohort to demonstrate the efficacy of new medications, beta blockers and ACE-inhibitors (Levy et al, 1993). One of the most valuable contributions was the demonstration that non-rheumatic atrial fibrillation was a strong risk factor for stroke and ischemic heart disease, leading to a flurry of controlled trials on newer classes of medications: anticoagulants and anti-arrhythmics, which are indispensable modern tools for managing heart disease. Later cohorts recruited the family members and descendants of original participants, laying the groundwork for the future identification of genetic risk factors.

Subsequent studies using the ongoing Framingham Heart Study data later identified additional cardiac risk factors including: increased left ventricle (LV) diameter, asymptomatic LV systolic dysfunction, diabetes, and hyperlipidemia, all still highly focused on today (Mahmood et al, 2014).

A limitation of any healthcare project is the availability of high-quality patient medical record data. This project was only able to access the most well-known and well-studied cardiac datasets, which unfortunately limits our predictive scope to reproducing and validating prior feature correlation findings, and testing various machine learning models to try and optimize our model’s predictive ability to simply detect the target outcome of present or not-present CAD. We considered generating synthetic datasets and comparing our results against authentic data, but ultimately decided that the medical diagnostic field was not the ideal setting for exploring synthetic data generation, as we are looking to identify outliers, patterns, and correlations.

# References

Advocate Health. (n.d.). *Ischemic heart disease*. Retrieved March 1, 2025 from <https://www.advocatehealth.com/health-services/advocate-heart-institute/conditions/ischemic-heart-disease>

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>

Levy, Daniel. Kannel, William. Ho, Kalon. Pinsky, Joan. (1993). *The epidemiology of heart failure: The Framingham Study*. *Journal of the American College of Cardiology. 22. 6A-13A.* 10.1016/0735-1097(93)90455-A.

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>

Mahmood, Syed. Levy, Daniel. Vasa, Ramachandran. Wang, Thomas. (2014). *The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective.* Lancet. 2014 Mar 15;383(9921):999-1008. doi: 10.1016/S0140-6736(13)61752-3. Epub 2013 Sep 29. PMID: 24084292; PMCID: PMC4159698.

McKee, Patrick. Castelli, William. McNamara Patricia. Kannel, William. (1971). *The Natural History of Congestive Heart Failure: The Framingham Study.**New England Journal of Medicine, 285*(26). <https://doi.org/10.1056/NEJM197112232852601>

Mayo Clinic. (n.d.). *Heart disease*. Mayo Clinic. Retrieved March 1, 2025 from <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>

World Health Organization (n.d.). (2024, August 7). *The top 10 causes of death*. Retrieved March 1, 2025 from

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

# Appendix A: Figures

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

A screenshot of a medical report

Description automatically generated

A white paper with black text

Description automatically generated

A white background with black text

Description automatically generated