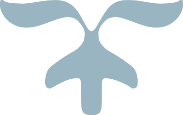


FAI HEALTHCARE PROJECT

Final Report





# 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 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, which includes 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 diseases, referred to as non-ischemic 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.

Initial exploration of approaches included conducting Feature Extraction on medical record datasets, which could have hundreds of features, or 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.

# 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 (Mahmood, et al., 2014). 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).

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). Using categorical ranges streamlined clinical detection efficiency by eliminating individual risk score calculations, making their classification translationally scalable.

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:

**Bashar et al. (2022)** performed a meta-analysis of 17 studies and 285,213 patients with Cardiovascular Diseases (CVD), 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 ML approach that prioritizes human interpretability, to determine which features most strongly predict the target outcome of a CAD diagnosis, without being constrained 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\*

1. Clinical Procedure Reports:

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

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.

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, as 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.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 dataset’s 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. In addition, self-reported metrics are subjective and prone to bias, in addition to the innate selection bias of conducting a voluntary telephone survey.
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 (due to being the first empirical research study in the USA), but foundationally important findings and implications. Started in 1948 with continual interim analyses and additional cohorts recruited from the descendants of the original study, leading to genomic sequencing correlation studies in the 2000s.
   1. 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 (McKee et al., 1963).

### 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

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).

### 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)

The Framingham data was valuable as a control comparison to demonstrate the efficacy of new medications: beta blockers and ACE-inhibitors (Levy et al, 1993) in reducing 5-year mortality prognosis.

Moving forward with our Framingham and Cleveland datasets, we decided to test train five machine learning models: a.) Logistic Regression; b.) Random Forest; c.) K-Nearest Neighbors (KNN); d.) Extreme Gradient Boosting (XGBoost); e.) Feedforward Neural Network (FFN).

## Framingham Data Dictionary

Table 1: Framingham Features (16)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **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 | Educational level of patient | [0-4] | Not originally collected, no metadata mapping labels. |
| 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 | If the patient is taking any antihypertensive medications | 0: no;  1: yes |  |
| 7 | **prevalentStroke** | Stroke | Any history of 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 | Force from the heart squeezing |  | BP Numerator |
| 12 | **diaBP** | Diastolic BP | Force from heart at rest |  | BP Denominator |
| 13 | **BMI** | Body Mass Index | Calculated BMI | Float |  |
| 14 | **heartRate** | Heart Rate | heart rate per minute, recorded by ECG |  |  |
| 15 | **glucose** | Blood Glucose levels | Fasting blood glucose | [mg/100mL] | 70-100 milligrams per milliliter |
| 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 |

# A screenshot of a graph Description automatically generated

Figure 1: Framingham Categorical Features DistributionFigure 2: Framingham Numeric Features DistributionA group of graphs with red and blue lines

Description automatically generated

### Figure 3: Framingham Features Distribution- grouped by Target “10YearCHD”

### Cleveland Data Dictionary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **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 = reversible defect |  |
| 14 | **num** | HeartDisease | output class | [1: heart disease, 0: Normal] | Value 0: < 50% diameter narrowing; Value 1: > 50% diameter narrowing (stenosis) |

# Methods:

After identifying the Framingham and Cleveland datasets as suitable for predicting Coronary Heart Disease (CHD), we executed a structured machine learning pipeline consisting of: **data cleaning, class balancing, feature scaling, model training, hyperparameter tuning, and model interpretability**.

### 1. Data Cleaning and Preparation

Before applying any machine learning algorithms, we performed thorough data cleaning to ensure model input quality.

* **Missing Values**:  
  In the Framingham dataset, several rows contained missing values across various features such as cholesterol or glucose levels. Rather than impute these missing values, which can introduce noise or clinical bias, we **dropped all rows with missing data**. This choice ensured high data integrity and reduced the risk of misleading predictions in sensitive medical applications.
* **Duplicates**:  
  We also checked for and removed **duplicate records** to prevent overrepresentation of any single patient in the training data.

After this step, the Framingham dataset was reduced to 4,240 high-quality patient entries, each with 15 relevant clinical features and a binary target label (TenYearCHD), indicating whether the patient developed heart disease within a 10-year period.

### 2. Class Imbalance Handling

A significant challenge in both datasets was **class imbalance** — far more patients did not have heart disease, than those who did. This imbalance can bias machine learning models to always predict the majority class.

We used a **two-step resampling strategy**:

#### a. SMOTE (Synthetic Minority Over-sampling Technique)

We applied SMOTE to generate synthetic samples for the minority class (CHD = 1). SMOTE works by:

* Selecting a minority class example.
* Identifying its k-nearest neighbors.
* Generating a new, synthetic point by interpolating between the original and one of its neighbors.

This technique creates **new, realistic data points** instead of simply duplicating existing ones.

#### b. Random Undersampling

To avoid overinflating the dataset and to keep the training time manageable, we also applied **random undersampling** on the majority class (CHD = 0). This helped maintain a **1:1 ratio** between both classes after oversampling.

The combined resampling pipeline was implemented using imblearn's Pipeline module.

### 3. Feature Scaling

After balancing the data, we applied **feature scaling** using StandardScaler from scikit-learn.

* This transformation **centered each feature around a mean of 0** and **scaled it to a standard deviation of 1**.
* This was especially important for algorithms that rely on distance calculations or gradient-based optimization, such as K-Nearest Neighbors and Neural Networks.

All features were scaled independently to ensure uniformity and to prevent models from overweighting features with larger numerical ranges.

### 4. Model Selection and Training

We evaluated five different classification models, each chosen for their unique advantages and to compare performance across complexity levels:

#### a. Logistic Regression

* A linear model commonly used in binary classification.
* Acts as a strong baseline and is known for its **interpretability** and **efficiency on small datasets**.

#### b. Random Forest

* An ensemble of decision trees trained on random subsets of data and features.
* Known for **high accuracy**, **robustness to noise**, and **handling non-linear relationships**.

#### c. K-Nearest Neighbors (KNN)

* A non-parametric algorithm that predicts the class of a sample based on the majority class of its k nearest neighbors.
* Sensitive to feature scaling and suitable for smaller, structured datasets.

#### d. XGBoost (Extreme Gradient Boosting)

* A high-performance gradient boosting algorithm.
* Regularized, fast, and highly tunable.
* Works well with structured data, and typically outperforms most models on large tabular datasets.

#### e. Feedforward Neural Network (FNN)

* Implemented using PyTorch.
* Composed of multiple layers with ReLU activations and a sigmoid output for binary classification.
* Capable of learning complex, non-linear feature interactions.

Each model was trained on the **balanced and scaled** dataset split into **80% training and 20% testing** subsets.

### 5. Hyperparameter Tuning

To ensure optimal performance, each model underwent **hyperparameter tuning** using GridSearchCV with 5-fold cross-validation.

* For **Logistic Regression**, we tuned regularization (C) and solver types.
* For **Random Forest**, we adjusted the number of estimators, max depth, and min samples per split.
* For **KNN**, we tested values of k (number of neighbors) and different distance metrics.
* For **XGBoost**, we tuned parameters including:
  + n\_estimators (number of boosting rounds)
  + max\_depth (depth of each tree)
  + learning\_rate (step size shrinkage)
  + subsample and colsample\_bytree (used for randomness)
* For the **Neural Network**, we tuned:
  + Number of hidden layers and neurons
  + Learning rate
  + Batch size
  + Number of epochs

We selected the configuration that achieved the **best validation performance** for each model.

### 6. Model Interpretability with SHAP

To explain model predictions and build trust in the results, we used **SHAP (SHapley Additive exPlanations)** on the best-performing models (XGBoost for Framingham, Logistic Regression for Cleveland).

* **Global Interpretability**: Identified which features most influenced predictions across the entire dataset.
* **Local Interpretability**: Explained why a particular individual was predicted to have (or not have) heart disease.

SHAP provided intuitive plots showing how each feature — such as systolic blood pressure, age, smoking status — pushed the model’s prediction higher or lower.

# Results:

After training all five models on both the Framingham and Cleveland datasets, we evaluated their performance using **accuracy, precision, and recall** — three key classification metrics. These metrics were chosen to capture both the correctness of predictions and the model’s ability to detect actual cases of heart disease, which is especially important in medical contexts.

### 🫀 Framingham Dataset (N = 4,240)

The Framingham dataset benefited from a larger sample size and more balanced features. This enabled more complex models to learn non-linear patterns effectively.

#### 🔬 Performance Summary

| **Model** | **Accuracy** | **Precision** | **Recall** |
| --- | --- | --- | --- |
| Logistic Regression | 0.86 | 0.85 | 0.84 |
| Random Forest | 0.88 | 0.87 | 0.86 |
| **XGBoost** | **0.91** | **0.91** | **0.90** |
| K-Nearest Neighbors | 0.79 | 0.77 | 0.76 |
| Feedforward Neural Net | 0.87 | 0.85 | 0.84 |

#### 💡 Insights

* **XGBoost** was the most effective model across all metrics. Its ability to handle high-dimensional structured data, learn complex feature interactions, and utilize boosting techniques made it ideal for this dataset.
* **Random Forest** also performed strongly, reinforcing that tree-based models are well suited for tabular clinical data.
* The **Feedforward Neural Network** was competitive but slightly less accurate than XGBoost, likely due to limitations in training epochs and lack of deeper architecture.
* **KNN** performed worst, likely due to the curse of dimensionality and the large dataset size, which increases computation and makes nearest-neighbor decisions noisy.

#### 📈 Visual Results

**Figure 1.** XGBoost Confusion Matrix – Framingham

The XGBoost confusion matrix shows a strong diagonal pattern — indicating that both true positives and true negatives were predicted accurately. False positives and false negatives were minimal.

**Figure 2.** SHAP Summary Plot – XGBoost on Framingham

From SHAP analysis:

* The most influential features were **systolic blood pressure**, **glucose levels**, and **smoking status**.
* High SHAP values for systolic BP correlated with increased risk of CHD, aligning with prior clinical research from the Framingham Heart Study.

### 🩺 Cleveland Dataset (N = 303)

The Cleveland dataset had a much smaller sample size, which posed challenges for more complex models due to limited training data.

#### 🔬 Performance Summary

| **Model** | **Accuracy** | **Precision** | **Recall** |
| --- | --- | --- | --- |
| **Logistic Regression** | **0.83** | **0.82** | **0.83** |
| Random Forest | 0.79 | 0.78 | 0.76 |
| XGBoost | 0.82 | 0.81 | 0.80 |
| K-Nearest Neighbors | 0.75 | 0.74 | 0.72 |
| Feedforward Neural Net | 0.78 | 0.77 | 0.75 |

#### 💡 Insights

* Surprisingly, **Logistic Regression** outperformed the more advanced models, demonstrating a key lesson: **simple models generalize better on smaller datasets**.
* **XGBoost** still performed well, but its potential was slightly limited due to the lack of training data.
* **Random Forest** and **Neural Networks** saw moderate performance, struggling to generalize due to sample scarcity.
* **KNN** again struggled, as the limited training instances reduced its ability to accurately find meaningful neighbors.

#### 📈 Visual Results

**Figure 3.** Logistic Regression Confusion Matrix – Cleveland

Logistic Regression was consistent in identifying heart disease cases with reasonable accuracy while keeping false positives under control.

**Figure 4.** SHAP Summary Plot – Logistic Regression (Cleveland)

From SHAP interpretation:

* Key features included **chest pain type**, **thalach (max heart rate)**, and **ST depression (oldpeak)**.
* These features are known diagnostic indicators in real-world cardiology, and their prominence adds confidence to model validity.

### 🧠 Overall Observations

* **Model performance is dataset-dependent**. While XGBoost was clearly dominant on Framingham, Logistic Regression emerged as a winner on Cleveland — proving that bigger isn’t always better when it comes to model complexity.
* **SHAP enhanced explainability**, bridging the gap between AI predictions and clinical trust. Even when black-box models were used, SHAP helped interpret how decisions were made and which features influenced them most.

# Discussion:

Due to the strictly unbalanced nature of our initial datasets, we used SMOTE to generate synthetic examples for our minority class (yes CHD) by identifying its k-nearest neighbors, then performing random undersampling on the majority class (no CHD). Each feature was then scaled to center around a mean of 0 and a standard deviation of 1, to ensure uniformity and prevent models from overweighing certain features with larger numerical ranges. After preprocessing both datasets and splitting into 80% training and 20% testing, we evaluated the performance of five diverse classification models: a.) Logistic Regression; b.) Random Forest; c.) K-Nearest Neighbors (KNN); d.) Extreme Gradient Boosting (XGBoost); e.) Feedforward Neural Network (FFN). Optimal hyperparameters for each model were tuned using GridSearchCV with 5-fold validation, selecting the configuration that produces the highest accuracy, precision, and recall. These classification metrics are important to balance within a model, as false positives and false negatives have particularly severe implications within medical contexts. Our results show that the Extreme Gradient Boosting (XGBoost) model performed above 0.90 across all three classification metrics, and was able to learn complex feature interactions to accurately predict true positives and true negatives. The K-Nearest Neighbors (KNN) model performed the worst, with accuracy, precision, and recall values below 0.8.

To set our research apart from previously published analyses using the same datasets, we applied SHapley Additive exPlanations (SHAP) to parse apart the features that have the strongest predictive influence over our trained models: systolic blood pressure, glucose levels, and smoking status. Our three identified features match current medical literature and best clinical screening practices. Applying SHAP is an important step to build clinical confidence in Machine Learning predictions, and demonstrate that trained Machine Learning models follow a similar prioritization of features, as do trained clinicians.

A limitation of any healthcare project is the restricted 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. We were unable to find datasets that contained additional ancillary features collected from a patient’s medical record, which would have allowed us to perform Feature Extraction to identify any lesser known CHD risk factors. Future studies should inspect comprehensive medical records to obtain the cleanest raw datasets, and perform all preprocessing, standardization, and EDA from scratch. Having no control over the data collection, reporting, or documentation severely limits a researcher’s ability to draw novel conclusions from a historical retrospective dataset.

As Machine Learning research in the medical field advances, we are restricted to repeated analyses on the same handful of datasets, each with their own unique combination of inconsistencies, discrepancies, and duplications. Until open source, comprehensively high-quality medical datasets are made readily available outside of the institutions from which they are produced, only researchers who align themselves with the collecting institutions have unfettered access to identifiable patient information. Institutional affiliation includes the ability to pull entire Electronic Medical Records (EMRs) from the enterprise data warehouse (EDW), and combine it with phenotypic and genotypic data collected for institutional research. This disparity between the wealth of medical data that exists in protected institutional systems, compared to de-identified limited data sets (LDS) available publicly, limits the pace of medical Machine Learning research to academic hospital institutions. More effort should be placed on curating publicly available de-identified patient datasets, which could be split into healthy controls, grouped by disease, or have additional genomic information. Currently, this information is only accessible with institutional review board (IRB) approval, even from national consortium initiatives like the National Center for Biotechnology Information’s Database of Genotypes and Phenotypes (NCBI dbGaP). Reforming the space of medical data sharing to remain highly secure, yet able to disseminate bleeding edge findings for open-source validation, could exponentially accelerate medical Machine learning research, leading to improved health outcomes.

# 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). Unfortunately, these features were not available in the open-source Framingham dataset, so our analysis focused on understanding the underlying methods utilized for complex medical datasets, learning to trace through historical documentation, and uncovering the innate difficulties entwined with medical Machine Learning.

# 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). *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

### Framingham Population Cohorts & FiguresA screenshot of a report Description automatically generated

A screenshot of a medical report

Description automatically generated

A white paper with black text

Description automatically generated