HEART DISEASE PREDICTION

Final Report

Amanda Pang, Shreyas Raman, Devansh Thakkar

Khoury College of Computer Science, Northeastern University
CS5100: Foundations of Artificial Intelligence
Dr. Rajagopal Venkatesaramani
April 18, 2025

1. Introduction:

Coronary Artery Disease (CAD), also known as Ischemic Heart Disease (IHD) has been cited by the World Health Organization (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), Congestive Heart Disease (CHD), and Cardiovascular disease (CVD). 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). Congestive Heart Failure (CHF) is now considered end stage CAD, when your heart can no longer pump sufficient blood throughout the body. 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 (inflammatory condition), and endocarditis (infection of heart muscle), (Mayo Clinic, 2025). 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.

2. 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.

3. 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:

a.) 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*

b.) Clinical Procedure Reports:

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

c.) 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.

A. 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.

B. 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.

C. 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.

- While modern studies use the acronym "CHD" to mean Congenital Heart Disease, the Framingham researchers use CHD (Congestive Heart Disease) and CHF (Congestive Heart Failure) interchangeably.

Framingham Methods & Results: McKee 1971

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

Findings:

- 75% of heart failure cases (CHD) were preceded by hypertension
- CHD diagnosis on prior exam: 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, were 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). The term "risk score" was coined and risk profiles were established, from their published: "Framingham Risk Score for CHD". Methodology describes how the researchers documented: "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 consolidate(d) this by the cross-sectional pooling method, into an average annual incidence rate by age, sex, and level of the risk factor", (McKee et al., 1971). Objective diagnosis criteria for the target endpoint of heart disease or heart failure is outlined in Appendix A, Tables 2. The preliminary list of 18 identified cardiac risk factors produced by the McKee, 1971 study is displayed in Appendix A, Figure 1.

Framingham Methods & Results: Levy, et al., 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)
- Used as a valuable 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 (15 + 1 Target)

#	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 patient is taking any hypertensive 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 heart squeezing		BP Numerator
12	**diaBP**	Diastolic BP	Force from heart at rest (to prevent blood backflow)		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]	
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

Figure 1: Framingham Numeric Features Distribution (n = 4240).

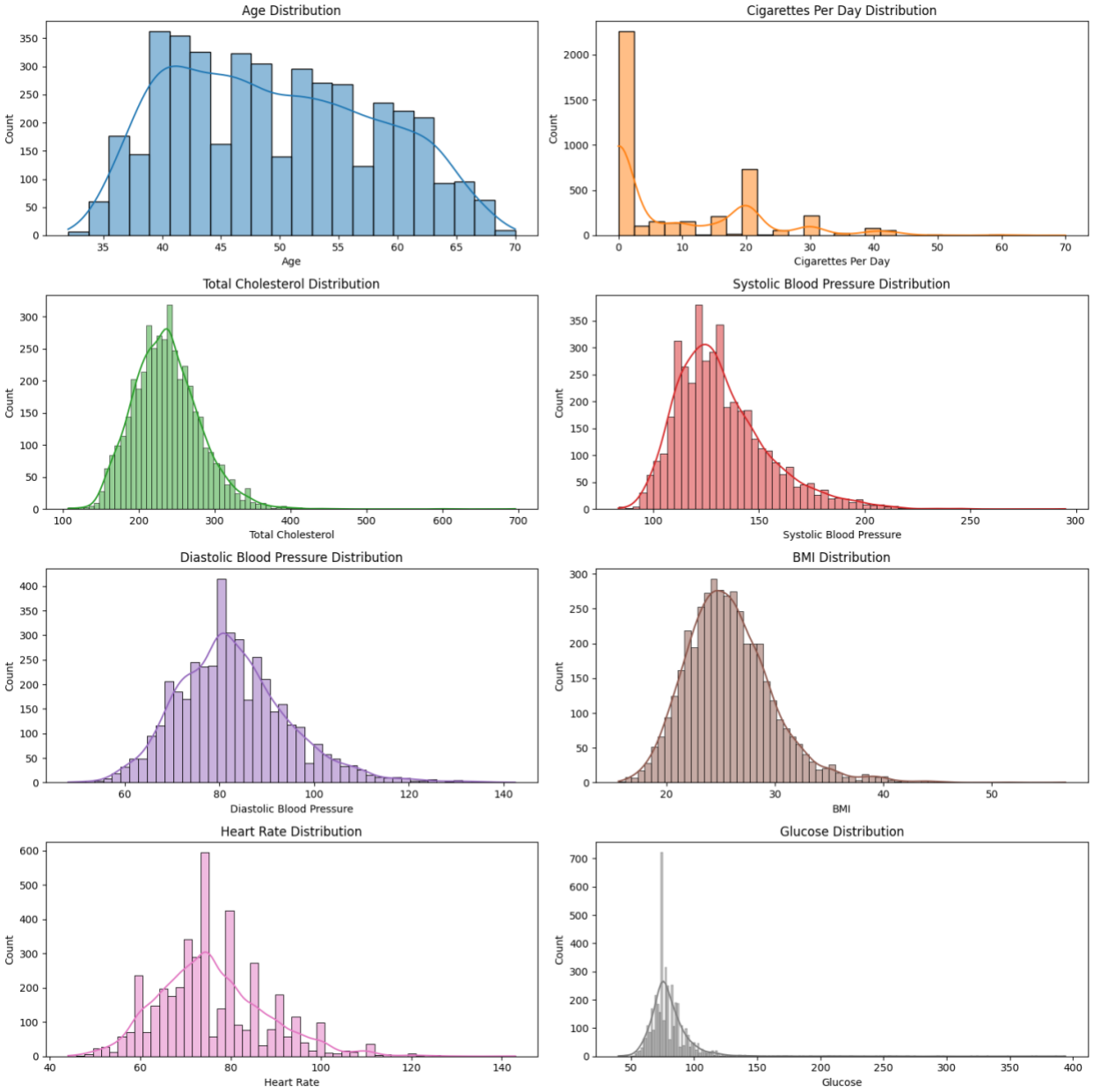


Figure 2: Framingham Categorical Features Distribution (n = 4240).

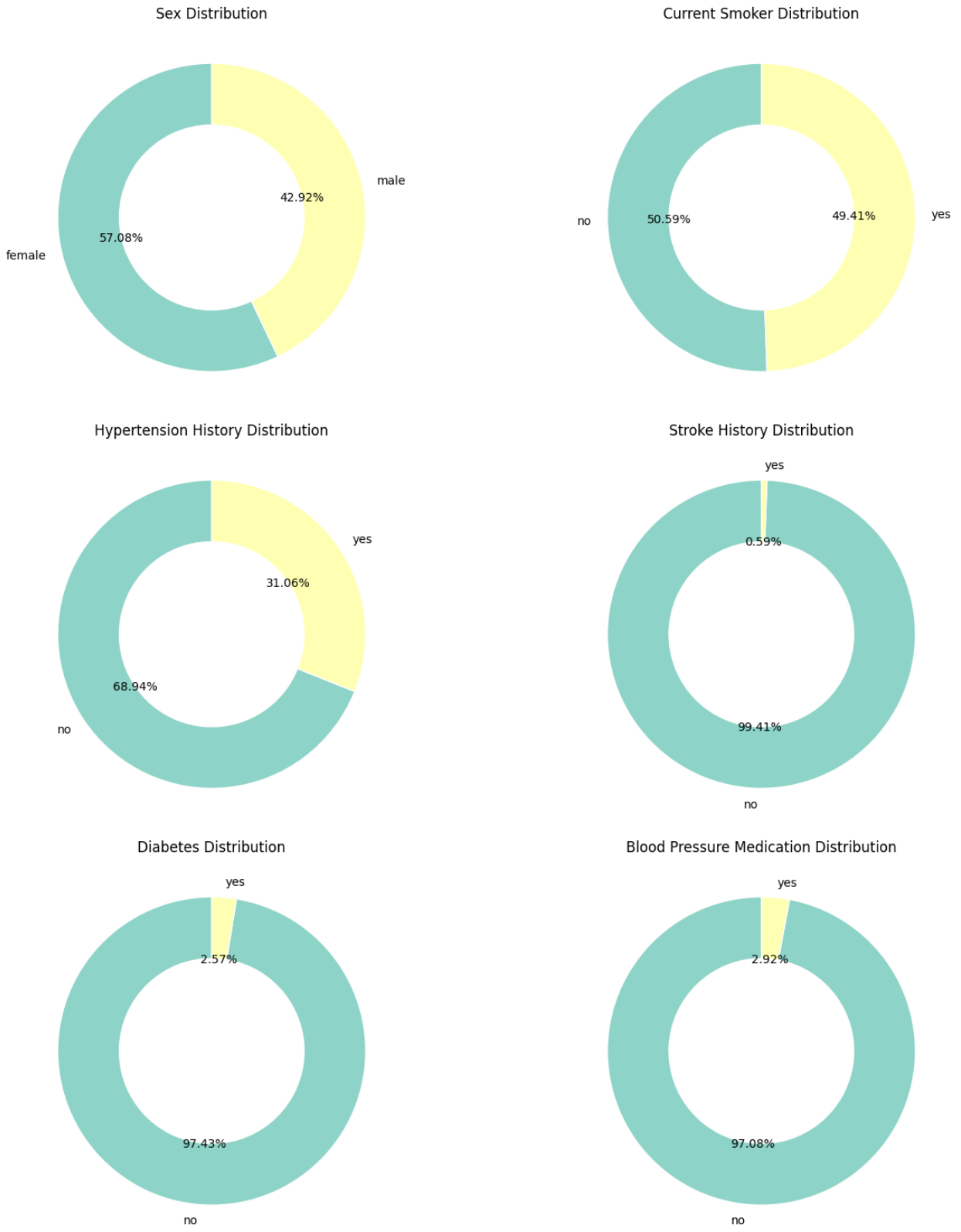
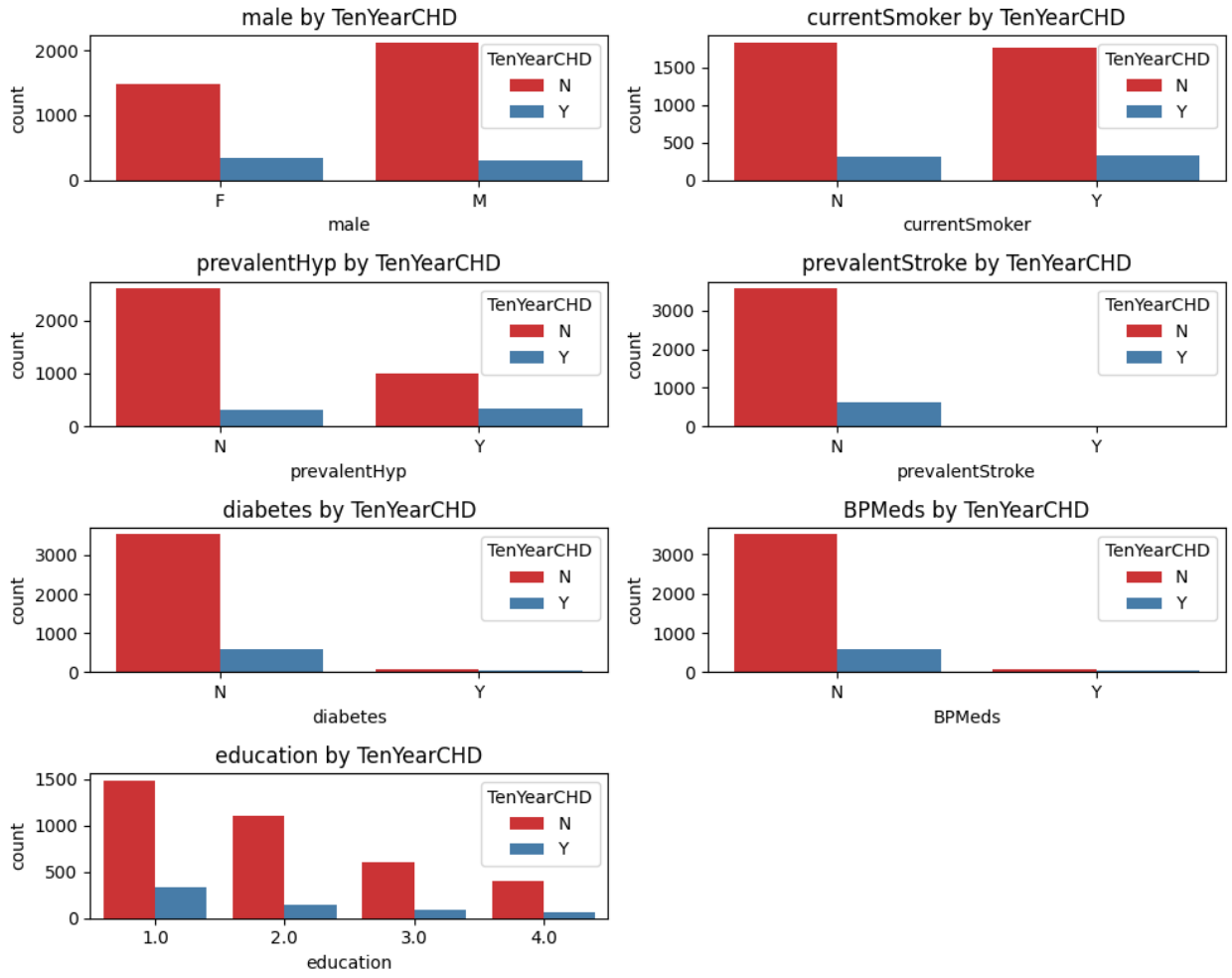


Figure 3: Framingham Categorical Features grouped by target outcome "10YearCHD". Education is included here to see the relationship with the target outcome, but mapping of values to labels is not available.



Cleveland Data Dictionary

Table 2: Cleveland Features (13 + 1 Target)

#	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	Target outcome	[1: heart disease, 0: Normal]	Value 0: < 50% diameter narrowing; Value 1: > 50% diameter narrowing (stenosis)

Figure 4: Cleveland Features Numeric Distribution (n = 303).

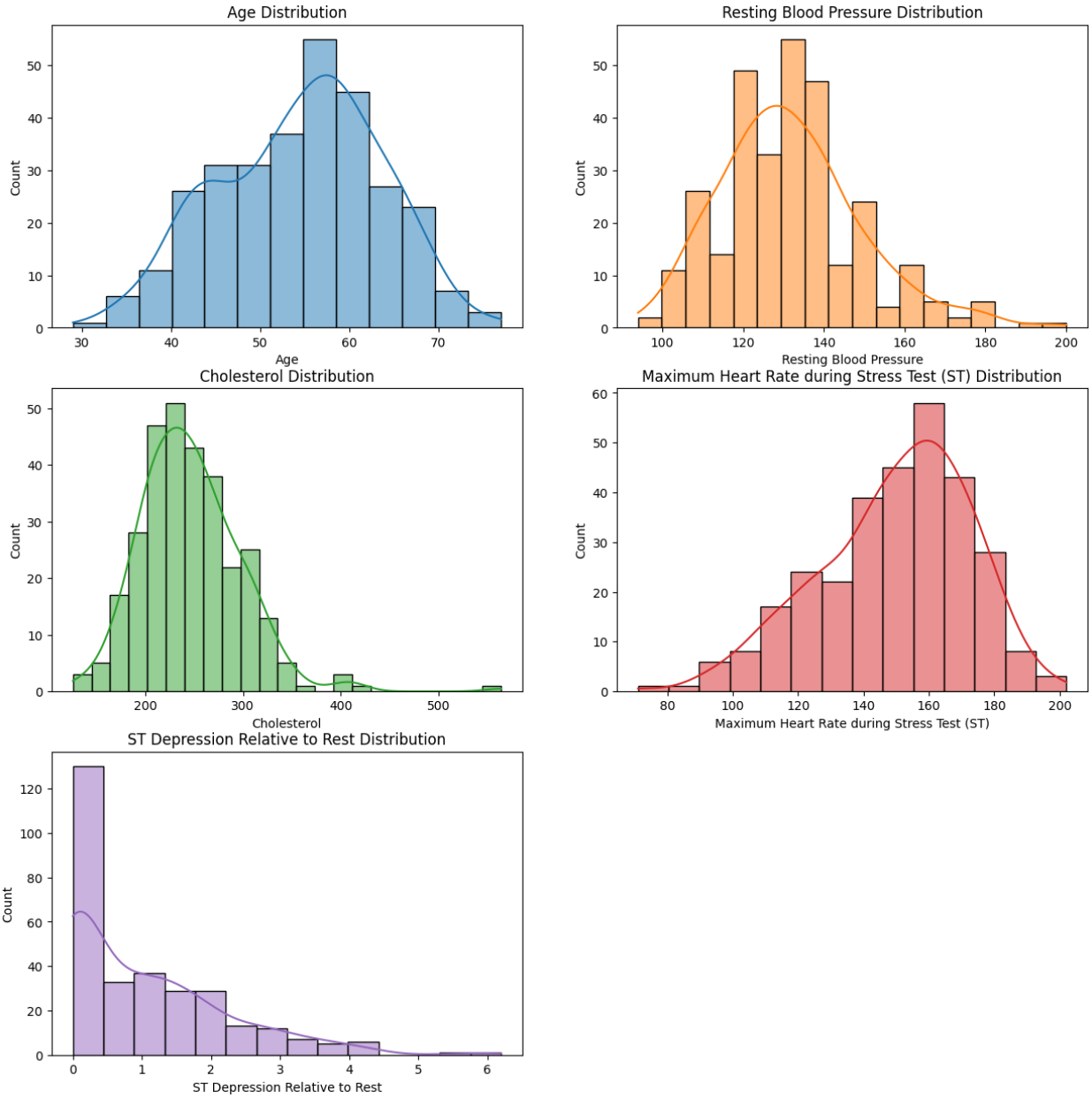


Figure 5: Cleveland Features Categorical Distribution (n = 303).

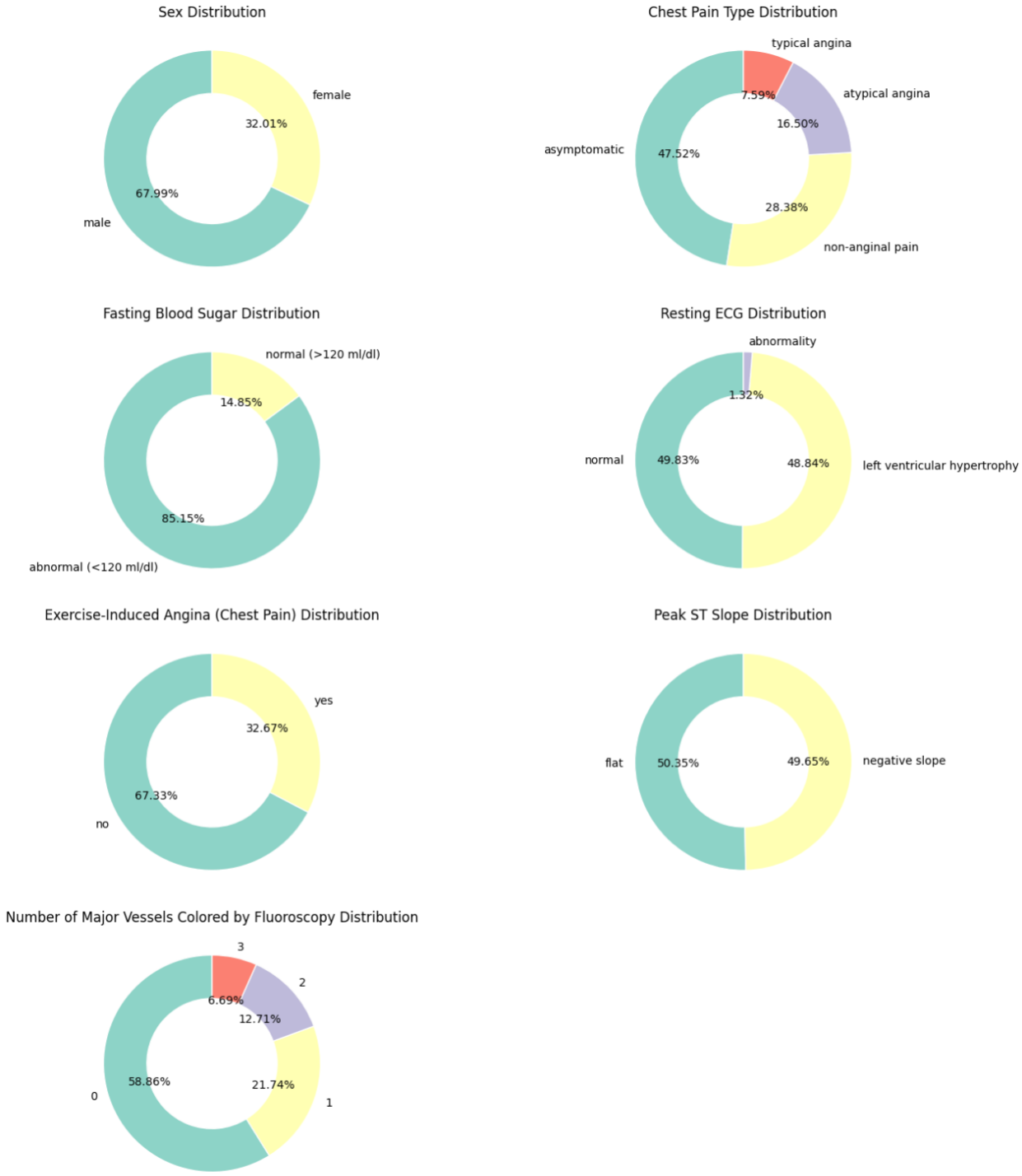
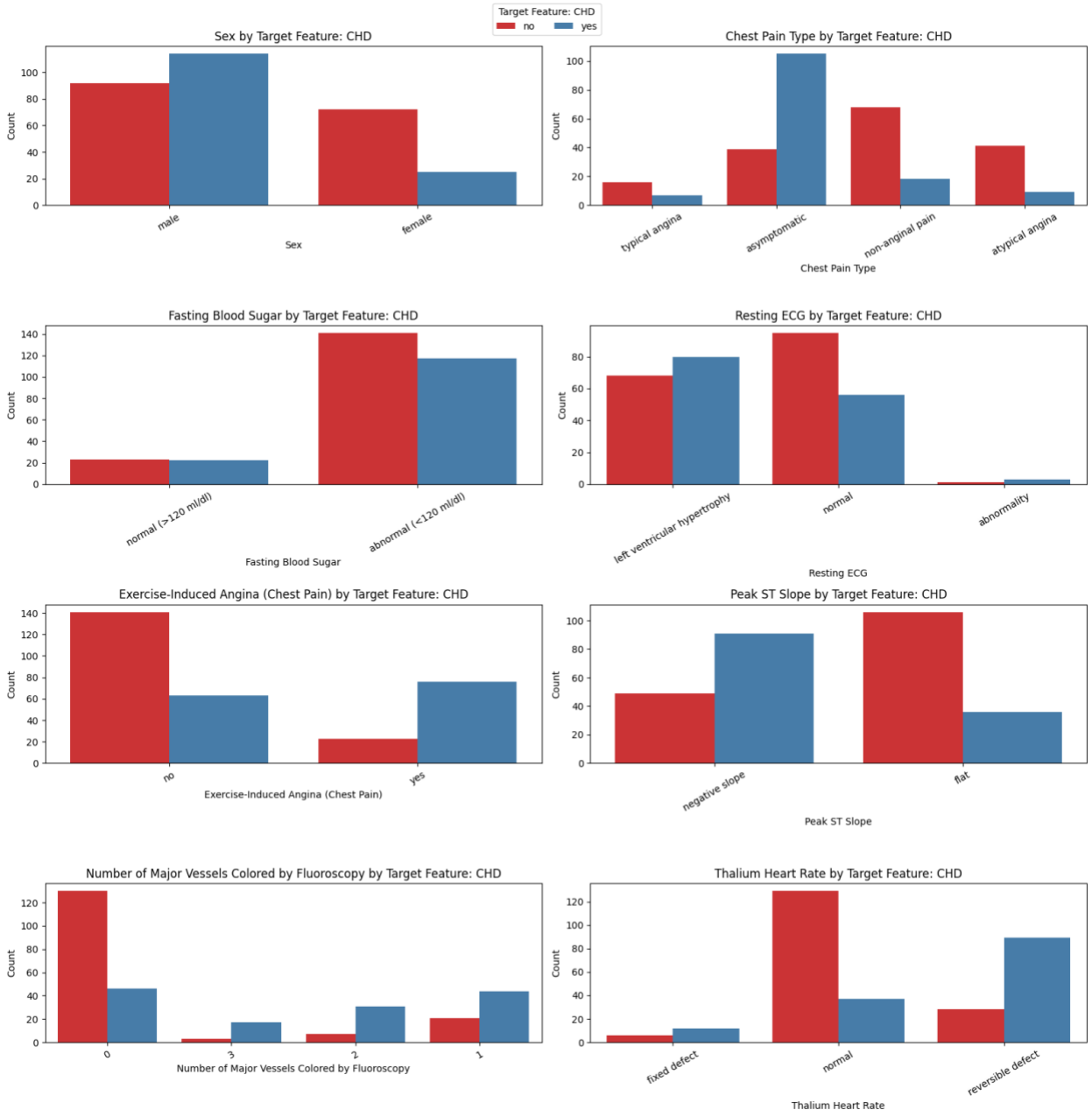


Figure 6: Cleveland Features Categorical Distribution – grouped by Target outcome (n = 303).



After analyzing both the Framingham and Cleveland datasets, and grouping them by the target outcome of CHD, we can see that both show similar skewed distribution patterns towards healthy patients and bias towards male patients, although the Cleveland dataset contains far more positive cases (Figures 3 and 6). For overlapping features between both datasets (Sex, Age,

Smoking, etc), we can see similar distribution patterns and features captures, suggesting the results of one model may be generalizable to the other.

4. Methods:

After identifying the Framingham and Cleveland datasets as the most 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 towards always predicting the majority class. We used a **two-step resampling strategy** to compensate for imbalance:

a. SMOTE (Synthetic Minority Over-sampling Technique)

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

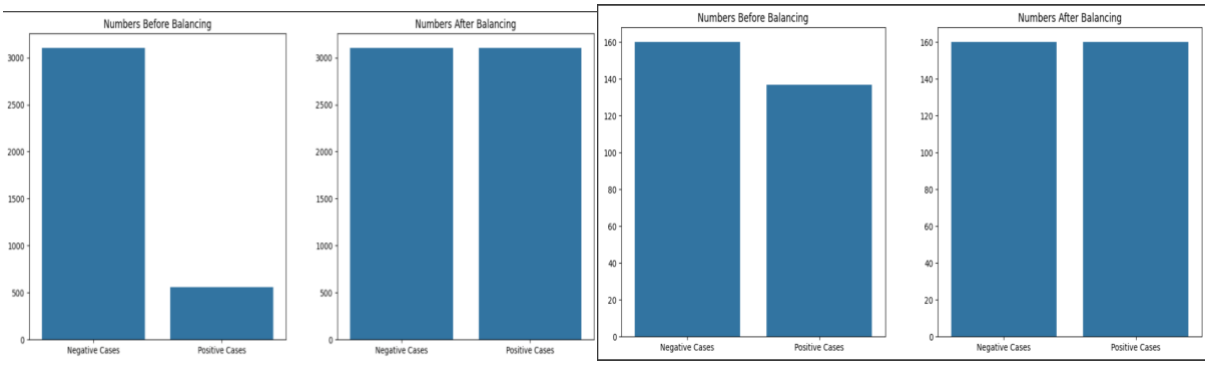
- Selecting a minority class example.
- Identifying its k-nearest neighbors.
- Generating a new, synthetic point by interpolating between the original and one of its neighbors.

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.

Figure 7. *Visual Results after data balancing in Framingham Dataset and UCI Datasets*



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**. All features were scaled independently to ensure uniformity and to prevent models from overweighting features with larger numerical ranges. This was especially important for algorithms that rely on distance calculations or gradient-based optimization, such as K-Nearest Neighbors and Neural Networks.

4. Model Selection and Training

We evaluated five different classification models, each chosen for their unique advantages and to compare performance across complexity levels. Each model was trained on the balanced and scaled dataset, split into 80% training and 20% testing subsets.

a. Logistic Regression: 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: 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): 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): 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.

5. Hyperparameter Tuning

To ensure optimal performance, each model underwent **hyperparameter tuning** using GridSearchCV with 5-fold cross-validation. We ultimately selected the configuration that achieved the best validation performance for each model:

- 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

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). The incorporation of SHAP provided both Global and Local Interpretability respectively by: a.) Identifying which features most influenced predictions across the entire dataset; and b.) Identifying why a particular individual was predicted to have (or not have) heart disease. SHAP also provided intuitive plots showing how each feature — such as systolic blood pressure, age, smoking status — pushed the model's prediction higher or lower.

5. 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 Results (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

Interpretation

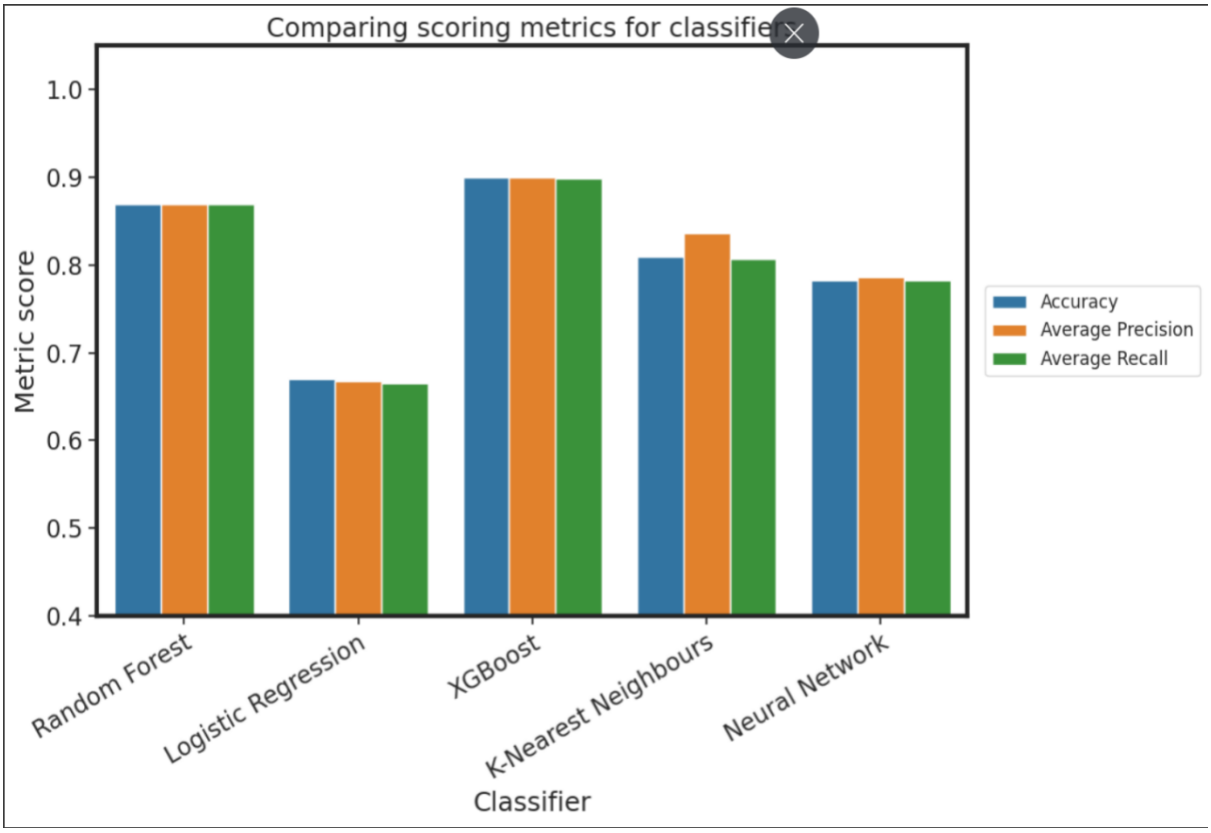
- **XGBoost** was the most effective model across all metrics. Its ability to handle high-dimensional structured data, learn complex feature interactions, and utilize boosting techniques made it ideal for this dataset.
- **Random Forest** also performed strongly, reinforcing that tree-based models are 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

The bar chart compares five classifiers — Random Forest, Logistic Regression, XGBoost, K-Nearest Neighbours, and Feedforward Neural Network — using three evaluation metrics: Accuracy, Average Precision, and Average Recall. XGBoost demonstrated the strongest overall performance across all three metrics. In contrast, the Logistic Regression shows the lowest

scores among the models evaluated. This visual comparison helps identify which classifiers generalize better on the dataset based on multiple performance aspects.

Figure 8. Classifier Performance Comparison – Framingham Dataset



Framingham SHAP interpretation (Figure 9):

- The most influential features were **education**, **cigarettes per day**, **age**, **current smoker** and **systolic blood pressure**.
- High SHAP values for systolic BP correlated with increased risk of CHD, aligning with prior clinical research from the Framingham Heart Study.

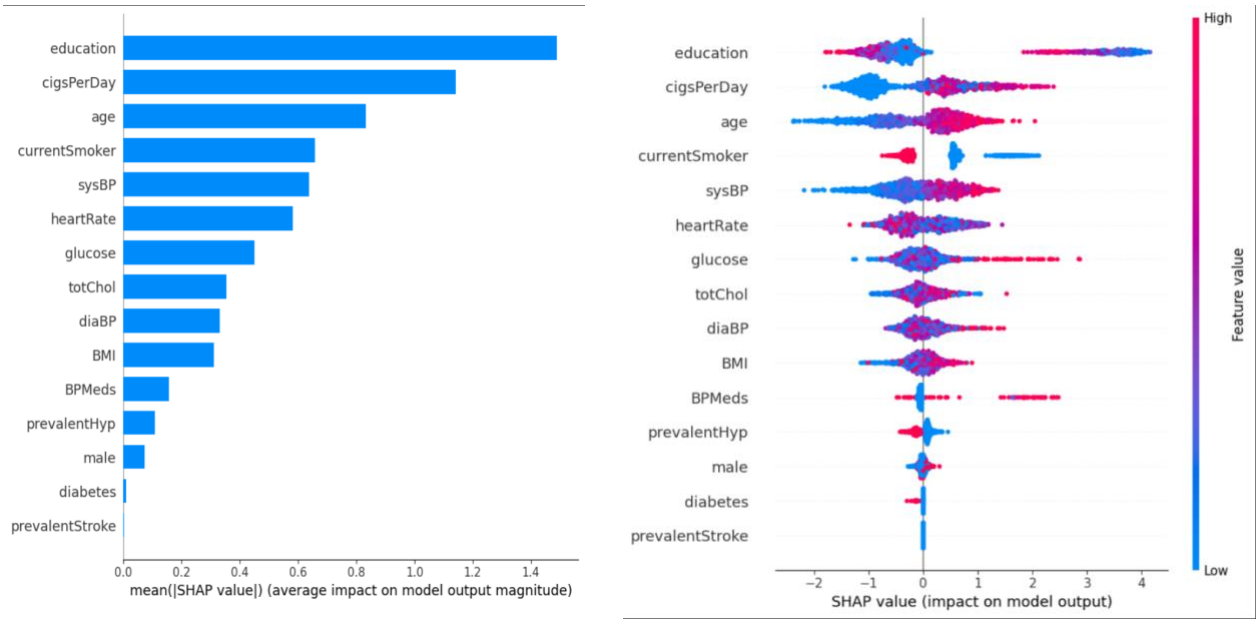


Figure 9. *SHAP Summary Plot – Framingham (Using XGBoost)*

Cleveland Dataset Results (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

Interpretation

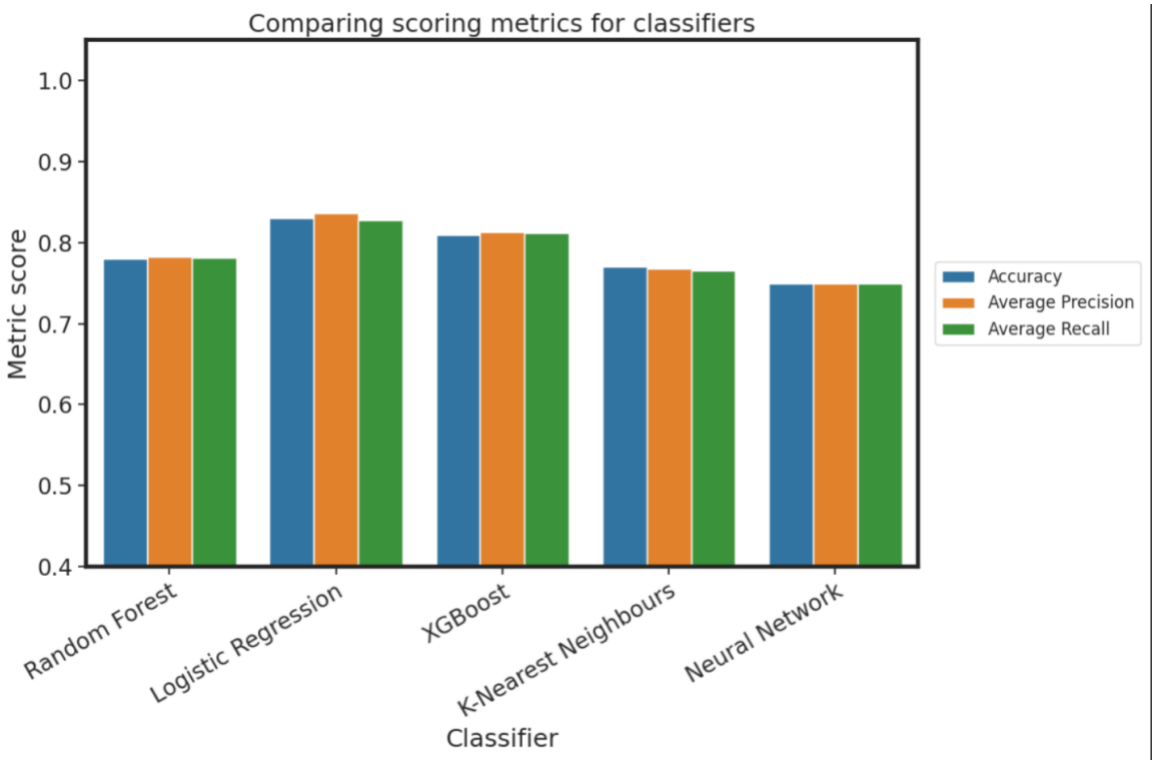
Surprisingly, **Logistic Regression** outperformed the more advanced models, demonstrating that simple models generalize better on smaller datasets. **XGBoost** still performed well, but its potential was slightly limited due to the lack of training data. **Random Forest** and

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

The bar chart in Figure 10 compares five classifiers — Random Forest, Logistic Regression, XGBoost, K-Nearest Neighbours, and Feedforward Neural Network — using three evaluation metrics: Accuracy, Average Precision, and Average Recall. Logistic Regression demonstrated the strongest overall performance across all three metrics. In contrast, the Feedforward Neural Network shows the lowest scores among the models evaluated. This visual comparison helps identify which classifiers generalize better on the dataset based on multiple performance aspects.

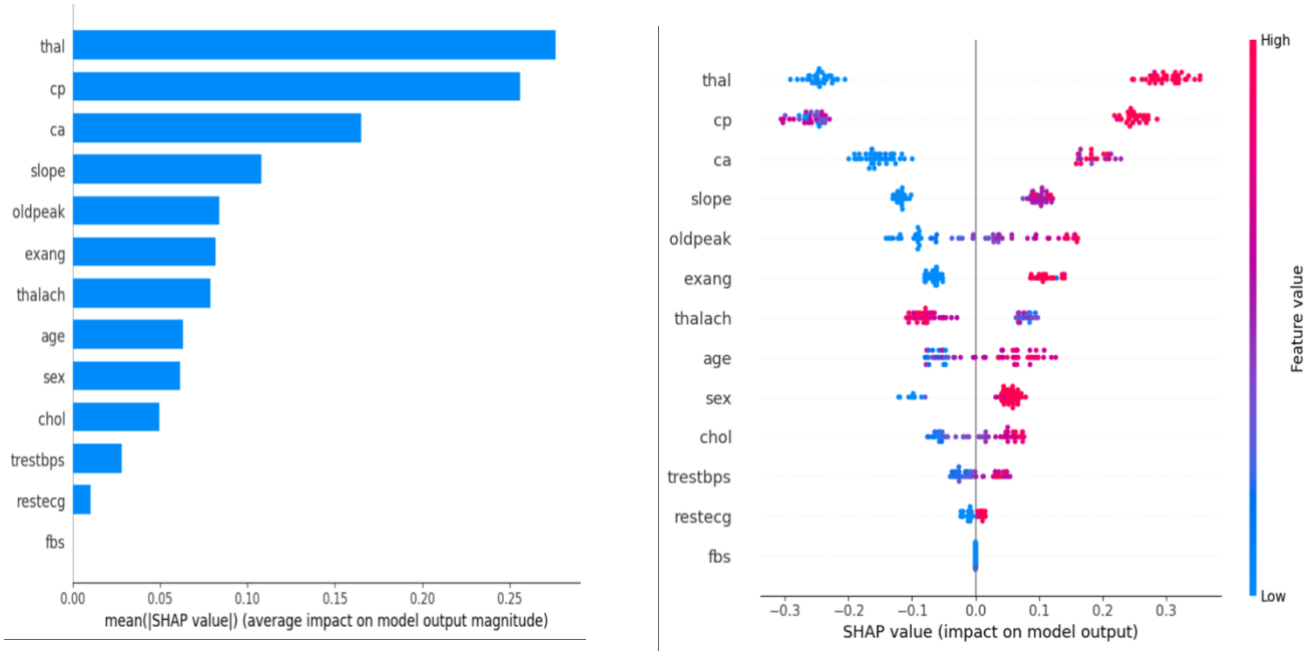
Figure 10. *Classifier Performance Comparison – Cleveland Dataset*



Cleveland SHAP interpretation:

- Key features included: **number of major vessels colored by fluoroscopy (ca), chest pain type (cp), max heart rate (thalach), ST depression (oldpeak) and slope of the the peak exercise ST segment (slope)**.
- These features are known diagnostic indicators in real-world cardiology, and their prominence adds confidence to model validity.

Figure 11. *SHAP Summary Plot – Cleveland (Using XGBoost)*



6. Discussion:

Reviewing analysis on the Cleveland dataset, our results suggest two main observational findings:

1. 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.
2. 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.

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.

7. 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.

We have concluded that the stark disparity between the wealth of medical data that exists in protected institutional systems, in comparison to de-identified limited data sets (LDS) available publicly, holds responsibility in limiting 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.

8. Contributions

Based on our preliminary findings and research, we unanimously decided as a group to initially go with the Framingham Dataset. As our research progressed and we met with the professor and TA, we decided to incorporate the UCI Cleveland dataset, as well as applying SHAP to our best performing model for both datasets.

Dataset Code: Framingham & UCI Cleveland: EDA, Pre-Processing, Analysis, Model Training

Github Link: https://github.com/shrsai123/Heart_Disease_Prediction

- Amanda performed Exploratory Data Analysis for both datasets (Framingham and UCI), conducted literature review, and investigated dataset quality. Documented methodologies for preprocessing, cleaning, and model training (tested Logistic regression ROC-AUC and SHAP on Framingham).
- Devansh worked with both the Framingham and UCI Datasets, performing preliminary data cleaning and preprocessing, then training the models: KNN, Neural Networks, and SVM.
- Shreyas worked with both the Framingham and UCI Datasets, performing data cleaning and preprocessing, then training the models: Logistic Regression, Random Forest, XGBoost. Applied SHAP to both Framingham and Cleveland datasets to interpret key features.

Presentation:
Every member worked on creating slides for their respective work individually. Then, we combined all slides, discussed the scope of our presentation, and worked on the final pacing of the slides. All members contributed equally, and are very happy with the final presentation, report, and demonstration of Machine Learning applications on medical datasets.

Report:

Each member wrote up their respective part of the report, then we worked together as a group to combine all sections, tuning the report to adhere to all guidelines, length, and detail.
Throughout the entire semester, all three members were constantly connected via Teams, with regularly scheduled meetings (in person and virtual) to ensure we were all on the same page, and on pace to complete the project together. We all learned to communicate well, and work with each others' different styles, to combine our shared experiences, knowledge, and technical skills and complete a project with a scope beyond anything we could attempt individually.

9. 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.

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: Additional Figures

Appendix Table 1: Framingham substudy cohorts (Mahmood et al., 2014).

Table 1. Population cohorts for the heart study.

Cohort	First Year	Size	% Female	Salient Features
Original	1948	5209	55%	
Offspring	1971	5124	52%	Children of the Original Cohort, and their spouses
Third Generation	2002	4095	53%	Children of the Offspring Cohort
New Offspring Spouse	2003	103	54%	Spouses of Offspring Cohort who were not initially enrolled in FHS, and whose 2 children are in Third Generation Cohort. Added to improve statistical power
Omni 1	1994	506	58%	To reflect the increasing ethnical diversity of the community. African-American, Hispanic, Asian, Indian, Pacific Islander and Native American
Omni 2	2003	410	57%	Recruited in order to achieve 10% of Third Generation Cohort size

Appendix Table 2: Framingham Major and Minor definitions of Target outcome "10YearCHD" (Mahmood et al., 2014).

Table 2. Criteria for heart failure.

MAJOR	MINOR
1. Paroxysmal nocturnal dyspnea	1. Ankle edema
2. Neck-vein distension (not counting supine position)	2. Night cough
3. Rales in presence of unexplained dyspnea	3. Dyspnea on ordinary exertion
4. Cardiomegaly and pulmonary hilar congestion (by X-ray in absence of left to right shunt), or increasing heart size.	4. Hepatomegaly
5. Acute pulmonary edema described in hospital records	5. Pleural effusion
6. Ventricular gallop	6. Decreased vital capacity by one-third from maximum records
7. Increased venous pressure (>16 cm H2O form right atrium)	7. Tachycardia (≥120 beats per minute)
8. Circulation time (>24 seconds, arm to tongue)	
9. Hepato-jugular reflux	
10. Autopsy shows pulmonary edema, visceral congestion, cardiomegaly	

Open in a new tab

Minor or Major: Weight loss (≥4.5 Kg) in 5 days, in response to HF therapy.

Appendix Table 3: Framingham outcome criteria (23) and identified risk factors (18) (McKee et al. 1963).

DESCRIPTION OF TABLES

The tables in this section of the monograph series, Section 34, present data for 23 events and 19 potential risk factors. The tables are numerically sequenced according to the combination of an event and a risk factor with the numbers indexed as follows:

Numbering Scheme for Event-Risk Factor Combinations
Framingham Study Monograph 30-year Followup

EVENTS

1. Coronary Heart Disease
2. Coronary Heart Disease other than Angina Pectoris
3. Myocardial Infarction
4. Myocardial Infarction unrecognized
5. Myocardial Infarction recognized
6. Coronary Insufficiency
7. Angina Pectoris
8. Angina Pectoris uncomplicated
9. Sudden Coronary Death among persons free of CHD
10. Sudden Coronary Death among all persons
11. Coronary Heart Disease Death among persons free of CHD
12 Coronary Heart Disease Death among all persons
13. Stroke and Transient Ischemic Attack
14. Atherothrombotic Brain Infarction
15. Transient Ischemic Attack
16. Stroke Death among persons free of Stroke and TIA
17. Stroke Death among all persons
18. Intermittent Claudication
19. Congestive Heart Failure
20. Cardiovascular Disease
21. Cardiovascular Disease Death among persons free of CVD
22. Cardiovascular Disease Death among all persons
23. Death among all persons

RISK FACTORS

1. Systolic Blood Pressure, First examiner
2. Diastolic Blood Pressure, First examiner
3A. Hypertension with antihypertensive treatment
3B. Hypertension ignoring treatment
4. Serum Cholesterol
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