Heart Attack Risk Prediction Using Data Mining Techniques Final Report

Michael Jung
michael.jung@emory.edu
Emory University
, Atlanta , Georgia , USA

Caleb Jennings
cbjenn2@emory.edu
Emory University
, Atlanta , Georgia , USA

Sungho Kim
sungho.kim@emory.edu
Emory University
, Atlanta , Georgia , USA

1 ABSTRACT

1.1 Topic background

Cardiovascular disease—particularly heart attacks—continues to be one of the leading causes of death worldwide. Early detection and accurate risk prediction are essential for reducing strain on healthcare systems and improving patient outcomes. This project applies data mining techniques to predict the likelihood of heart attacks, with a particular focus on how risk factors vary across age groups and health conditions.

To ensure comprehensive and unbiased results, we employ a diverse set of models. By training three classifiers—Logistic Regression, Neural Networks, and Random Forests—we aim to minimize model-specific biases while evaluating the predictive contribution of different health indicators. The dataset is also stratified by age group to determine whether certain features carry more predictive weight within specific demographics. Finally, we identify and compare the top five most important features from each model to uncover shared patterns that offer deeper insights into cardiovascular risk.

1.2 Motivation

The primary goal of this project is to build a predictive model for heart attack risk, with a focus on how this risk is shaped by age and underlying health conditions. The dataset contains 300 clinically validated records, each with 14 relevant health-related attributes. Although relatively small in size, the dataset's quality supports robust and meaningful analysis.

Data mining has already shown substantial promise in medical diagnostics and predictive healthcare. This project extends those applications to cardiovascular risk assessment, seeking to understand how age modifies the impact of individual health indicators. Ultimately, this work aims to support more personalized and targeted approaches to heart disease prevention.

1.3 Results

The performance of the models on the general dataset, evaluated using accuracy, showed that Logistic Regression achieved the highest accuracy at 84.43%, followed by Random Forest at 81.48%, and Neural Network at 81.46%. These results demonstrate that Logistic Regression was the most accurate model for heart disease prediction on the general dataset, likely due to the linearly separable nature of the data and strong individual predictors.

When analyzing the top 5 influential features across all three models, there were several shared features. Chest pain type (cp) and number of major vessels colored (ca) were consistently ranked among the top 5 features, indicating their strong predictive value across models. These shared features highlight their importance in heart disease diagnosis and suggest they are reliable indicators of heart disease risk.

Notably, there were differences in feature importance across age groups. For example, in Logistic Regression, sex (sex) and thalassemia (thal) were highly influential in the under 55 age group but not among the top predictors in the older group. Conversely, resting electrocardiographic results (restecg) and slope of the peak exercise ST segment (slope) became more influential in the 55 and older group. This indicates that certain features contribute differently to heart disease risk prediction depending on age, emphasizing the importance of age-specific modeling.

Further results and detailed analysis of model performance and feature importance are discussed in the Performance Report section.

2 INTRODUCTION

2.1 Background

Heart attacks are responsible for millions of deaths around the world each year. While extensive research exists in the area of heart disease prediction, significant gaps remain in our understanding of how demographic and lifestyle factors contribute to risk across different age groups.

This project investigates a wide range of possible risk factors and runs several experiments to isolate variables and determine which features are the most predictive within specific age brackets. By exploring these patterns, we aim to identify more nuanced and age-sensitive predictors of heart disease.

2.2 Challenges Faced

Biomedical datasets often contain features with specialized clinical terminology that can be difficult to interpret without domain-specific medical knowledge. This posed a challenge during feature engineering and model interpretation, as it was difficult to determine the relative importance of certain attributes without consulting medical literature or experts.

Additionally, datasets with non-clinical and less specific features—such as amount of sleep, smoking status (yes/no), and income level—led to inconsistent patterns in prediction, often introducing noise rather than a meaningful signal. These features, while potentially relevant, lacked the granularity needed to contribute reliably to clinical predictions, and their presence sometimes conflicted with the more objective clinical indicators.

Another challenge was data imbalance, particularly in binary classification tasks where the number of healthy individuals often exceeded the number of patients with heart disease. This imbalance

made it harder for models to accurately detect positive cases without overfitting to the majority class.

3 RELATED WORK

3.1 Heart Disease UCI Diagnosis & Prediction

In this study, Hardik Deshmukh employed logistic regression to predict heart disease using the UCI Heart Disease dataset, which comprises approximately 300 instances and 14 key attributes. The model achieved an accuracy of 87%, effectively assessing the likelihood of heart attacks. The research emphasizes the utility of logistic regression in medical diagnostics, providing a straightforward yet effective approach to heart disease prediction.

3.2 Early Prediction of Heart Disease Using PCA and HGA with k-Means

This research, conducted by Md. Touhidul Islam, Sanjida Reza Rafa, and Md. Golam Kibria, focuses on early prediction of heart disease by integrating Principal Component Analysis (PCA) for dimensionality reduction and a Hybrid Genetic Algorithm (HGA) with k-means clustering. The PCA technique reduced the dataset's attributes to two principal components, simplifying the data structure. Subsequently, the HGA enhanced the clustering process by avoiding local optima, a common issue with standard k-means clustering. The combined approach achieved a prediction accuracy of 94.06%, demonstrating its effectiveness in early heart disease detection.

3.3 Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm

This study, "Prediction of Heart Disease Based on Machine Learning Using the Jellyfish Optimization Algorithm" by Polat and Ahmad, applied four machine learning classifiers—ANN, Decision Tree, AdaBoost, and SVM—on the Cleveland dataset. It used the Jellyfish optimization algorithm for feature selection, achieving the highest accuracy (98.47%) with SVM, demonstrating the effectiveness of optimization algorithms in improving heart disease prediction.

4 METHODOLOGIES

4.1 Dataset

Initially, we used a larger heart disease risk dataset from Kaggle. However, after observing unreliable model performance and inconsistencies in the data, we decided to switch to a more reputable and polished dataset. As a result, we transitioned to a cleaned and preprocessed version of the original **UCI Heart Disease Dataset**.

This dataset contains **302 instances** and **14 attributes**, offering a compact yet rich source of clinical information relevant to heart disease prediction. It includes demographic and clinical features such as age, sex, chest pain type, cholesterol level, resting blood pressure, maximum heart rate achieved, fasting blood sugar, and exercise-induced angina.

The dataset required minimal cleaning and came with no missing values. We removed duplicate entries using `df.drop_duplicates()` to ensure data integrity. All features were either numerical or converted into numerical values through appropriate mappings for machine learning model compatibility.

Binary Feature Mapping:

- `Sex` (Male / Female) → 1 / 0
- `Fasting Blood Sugar` (Lower than 120 mg/ml / Greater than 120 mg/ml) → 0 / 1
- `Exercise-induced Angina` (Yes / No) → 1 / 0

Categorical Feature Mapping:

- `Chest Pain Type` (Typical angina / Atypical angina / Non-anginal pain) → 0 / 1 / 2
- `Resting ECG` (Normal / ST-T wave abnormality / Left ventricular hypertrophy) → 0 / 1 / 2
- `Slope of ST Segment` (Upsloping / Flat / Downsloping) → 0 / 1 / 2
- `Vessels Colored by Fluoroscopy` (Number of vessels: 0–4) → Integer values 0–4
- `Thalassemia` (No / Normal / Fixed Defect / Reversible Defect) → 0 / 1 / 2 / 3

Overall, this polished version of the UCI dataset provided a clean and compact foundation for heart disease risk classification, with all features structured for efficient preprocessing and model training.

4.2 Further preprocessing per model

The data preprocessing phase is crucial for improving the performance and efficiency of the machine learning models. The following steps will be performed:

- **Logistic Regression:**
  - **One-Hot Encoding of Categorical Features:** Categorical variables (`cp`, `restecg`, `slope`, and `thal`) were transformed using one-hot encoding to convert them into a binary format suitable for logistic regression. To avoid multicollinearity, the first category from each feature was dropped (`drop_first=True`).
  - **Exclusion of Binary Features from Normalization:** Binary columns, including the one-hot encoded ones (e.g., `sex`, `fbs`, `exang`), as well as the target variable `target`, were excluded from normalization to preserve their original 0/1 values.
  - **Z-Score Normalization of Numerical Features:** All remaining numerical features (e.g., `chol`, `thalach`, `oldpeak`) were standardized using Z-score normalization to ensure zero mean and unit variance. This helps logistic regression converge more efficiently and treat all features on a comparable scale.
  - **Special Handling of Age Feature:** The original `age` column was retained in unnormalized form to split the dataset into age-based groups accurately. A separate Z-score normalized `age` feature was used for training and evaluating the model on the full dataset.

- **Random Forest:**
  - **No further preprocessing steps taken:** Default random forest model finds the optimal binary splitting point on its own.

- **Neural Network:**
  - **No further preprocessing steps taken:** Batch normalization is only helpful for large datasets, so this particular model didn't need additional preprocessing.

4.3 Experiments

- **General Dataset Risk Evaluation and 5-Fold Cross Validation:**
  - To maximize the utility of the limited dataset, each model was evaluated using 5-fold cross-validation. Several datasets were considered throughout the process, including some that attempted to increase instance count through data augmentation. However, these approaches ultimately produced inconsistent and unreliable model behavior. The dataset that delivered the most consistent and meaningful results was one with a smaller but clinically validated set of instances.
  - Models were assessed using five standard evaluation metrics: accuracy, AUROC, precision, recall, and F1-score. The results indicated strong model performance with minimal signs of either overfitting or underfitting, suggesting a well-balanced approach across all classifiers.

- **Age-Based Evaluation:** Under 55 vs. 55 and older
  - The original dataset showed a noticeable imbalance in patient distribution, particularly in the 30–40 age range. This skew made it difficult to analyze age-specific trends and model behavior. By dividing the dataset at age 55, two balanced subgroups were created, enabling clearer, more meaningful comparisons.
  - This stratification allowed for a deeper investigation into how age influences the predictive power of various risk factors. It also highlighted how certain features contribute differently to heart disease prediction depending on the age group being studied.

5 PERFORMANCE REPORT

5.1 Evaluation Metrics

The model's performance will be assessed using the following metrics:

- **Accuracy**: To evaluate the overall performance of the model in making correct predictions.
- **Precision and Recall**: To assess the model's ability to correctly identify true positives and true negatives, especially in the context of imbalanced classes.
- **F1-Score**: To provide a balanced measure of precision and recall, particularly useful when dealing with imbalanced datasets.
- **AUC (Area Under the ROC Curve)**: To assess the model's ability to distinguish between positive and negative classes, particularly useful in binary classification tasks.

5.2 Performance and Feature Importance for the Entire Dataset

(1) **Logistic Regression:**
5-Fold CV Results for Full Dataset
- Accuracy: 0.8443 ± 0.0404
- AUROC: 0.9110 ± 0.0215
- F1-Score: 0.8431 ± 0.0401
- Recall: 0.8443 ± 0.0404
- Precision: 0.8549 ± 0.0428

Top 5 Features (Coefficient Magnitude)
- cp_2: 1.513228
- cp_3: 1.278922
- sex: -1.274889
- exang: -0.915756
- ca: -0.781951

(2) **Random Forest:**
5-Fold CV Results for Full Dataset
- Accuracy: 0.8148 ± 0.0400
- AUROC: 0.8123 ± 0.0401
- F1-Score: 0.8302 ± 0.0416
- Recall: 0.8415 ± 0.0844
- Precision: 0.8268 ± 0.0489

Top 5 Features (Feature Importance)
- cp: 0.1290
- thalach: 0.1274
- thal: 0.1161
- ca: 0.1144
- oldpeak: 0.1061

(3) **Neural Network:**
5-Fold CV Results for Full Dataset
- Accuracy: 0.8146 ± 0.0285
- AUROC: 0.8731 ± 0.0433
- F1-Score: 0.8294 ± 0.0335
- Recall: 0.8404 ± 0.0425
- Precision: 0.8201 ± 0.0408

Top 5 Features (Feature Importance)
- exang: 70.2552
- cp: 68.8643
- oldpeak: 67.7219
- thalach: 64.2378
- ca: 60.2627

5.3 Performance and Feature Importance for Age Group Under 55

(1) **Logistic Regression:**
5-Fold CV Results for Age Group Under 55
- Accuracy: 0.8394 ± 0.0405
- AUROC: 0.9309 ± 0.0399
- F1-Score: 0.8406 ± 0.0372
- Recall: 0.8394 ± 0.0405
- Precision: 0.8635 ± 0.0073

Top 5 Features (Coefficient Magnitude)
- cp_2: 1.457772
- sex: -1.331324
- thal_2: 1.203793
- thal_3: -1.020499
- oldpeak: -0.977969

(2) **Random Forest:**
5-Fold CV Results for Age Group Under 55
- Accuracy: 0.8882 ± 0.0335
- AUROC: 0.8439 ± 0.0353
- F1-Score: 0.9222 ± 0.0242
- Recall: 0.9600 ± 0.0374
- Precision: 0.8879 ± 0.0226

Top 5 Features (Feature Importance)
- thal: 0.1746
- thalach: 0.1435
- cp: 0.1424
- oldpeak: 0.0987
- trestbps: 0.0905

(3) **Neural Network:**
5-Fold CV Results for Age Group Under 55
- Accuracy: 0.8672 ± 0.0408
- AUROC: 0.9136 ± 0.0655
- F1-Score: 0.9068 ± 0.0321
- Recall: 0.9492 ± 0.0041
- Precision: 0.8696 ± 0.0560

Top 5 Features (Feature Importance)
- cp: 37.7684
- thalach: 37.7152
- exang: 37.1115
- thal: 33.3095
- oldpeak: 29.5509

5.4 Performance and Feature Importance for Age Group 55 and Above

(1) **Logistic Regression:**
5-Fold CV Results for Age Group 55 and Above
- Accuracy: 0.7796 ± 0.0495
- AUROC: 0.8742 ± 0.0314
- F1-Score: 0.7772 ± 0.0486
- Recall: 0.7796 ± 0.0495
- Precision: 0.8219 ± 0.0511

Top 5 Features (Coefficient Magnitude)
- restecg_1: 1.306774
- cp_3: 1.220679
- exang: -1.205701
- cp_2: 0.815003
- ca: -0.734936

(2) **Random Forest:**
5-Fold CV Results for Age Group 55 and Above
- Accuracy: 0.7984 ± 0.0397
- AUROC: 0.7897 ± 0.0402
- F1-Score: 0.7462 ± 0.0568
- Recall: 0.7385 ± 0.1427
- Precision: 0.7931 ± 0.1198

Top 5 Features (Feature Importance)
- ca: 0.1539
- oldpeak: 0.1176
- thalach: 0.1088
- cp: 0.0946
- chol: 0.0925

(3) **Neural Network:**
5-Fold CV Results for Age Group 55 and Above
- Accuracy: 0.7419 ± 0.0255
- AUROC: 0.8128 ± 0.0364
- F1-Score: 0.6798 ± 0.0394
- Recall: 0.6825 ± 0.1009
- Precision: 0.6900 ± 0.0327

Top 5 Features (Feature Importance)
- ca: 35.4371
- exang: 28.6734
- cp: 28.1558
- oldpeak: 24.3586
- slope: 16.5294

6 VISUALIZATION

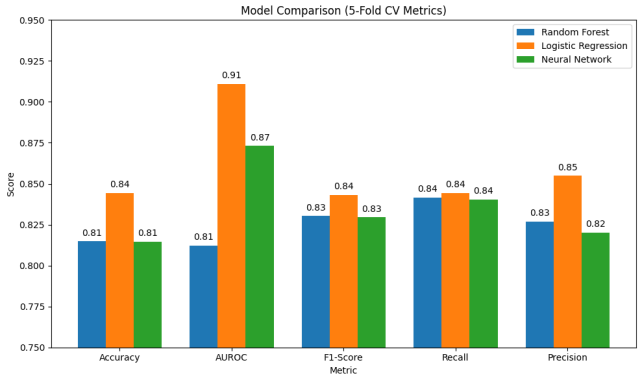

Figure 1: Bar graph showing metrics of each model side by side. Logistic Regression can be seen to outperform the Random Forest and Neural Network.

Top 5 Features by Model (Age < 55)

Random Forest	Logistic Regression	Neural Network
thal	cp_2	cp
thalach	sex	thalach
cp	thal_2	exang
oldpeak	thal_3	thal
trestbps	oldpeak	oldpeak

Figure 2: Table showing the top 5 features of each model trained on age < 55. Logistic Regression and the Neural Network have a lot of similarities.

Top 5 Features by Model (Age ≥ 55)

Random Forest	Logistic Regression	Neural Network
ca	restecg_1	ca
oldpeak	cp_3	exang
thalach	exang	cp
cp	cp_2	oldpeak
chol	ca	slope

Figure 3: Table showing the top 5 features of each model trained on age ≥ 55. There doesn't appear to be any special correlation.

7 DISCUSSION

- **Logistic Regression as Primary Insight Model:** We selected the Logistic Regression model as our primary tool for extracting insights due to its superior classification performance across all models and its ability to output interpretable feature importance rankings for the overall dataset.

- **Challenge in Determining Feature Directionality:** A key limitation of the Random Forest model is its inability to indicate whether a feature contributes positively or negatively toward the target outcome. To overcome this, we used coefficients from a trained Logistic Regression model to determine the direction (positive or negative correlation) of each feature's influence.

- **Insights from the Overall Dataset:** On the full dataset, considering accuracy alone, logistic regression leads with (84.43% ± 4.00%), while random forest (81.48% ± 4.00%) and neural network (81.46% ± 2.85%) perform similarly, each about three percentage points behind. This suggests that the simpler linear boundary captured by logistic regression may generalize slightly better across the entire cohort, whereas the more flexible models achieve comparable but marginally lower accuracy.
  When we look at the top-5 features shared by all three models on the full dataset, only chest pain type (CP) and number of major vessels (CA) appear in every ranking. Neither maximum heart rate (Thalach) nor ST-segment depression (Oldpeak) nor exercise-induced angina (Exang) are universally present—only CP and CA make the cut. Their consistent prominence across logistic regression, random forest, and neural network highlights how patients describe their pain and the anatomical extent of coronary involvement are the most discriminative signals for predicting heart disease.

- **Age Group-Based Insights:** In the under-55 subgroup, three features—CP, Thal (thalassemia defect type), and Oldpeak (ST-segment depression)—emerge in the top-5 across all three models. Younger patients' risk appears driven by qualitative chest pain characteristics, specific blood disorder markers, and ischemic changes on ECG. This alignment underscores the importance of combining clinical symptomatology with physiological measurements for early risk detection in a younger demographic.
  For patients aged 55 and above, CP and CA again surface in every model's top-5, reaffirming their overarching predictive value. In addition, exercise-induced angina (Exang) and ST-segment depression (Oldpeak) feature prominently in at least two of the three models, indicating that in older adults, classical anginal triggers and ischemic burden—alongside the anatomical measure of vessel count—carry significant weight in modeling outcomes.
  Across age groups, random forest secures the highest accuracy in both cohorts—(88.82% ± 3.35%) for age under 55 and (79.84% ± 3.97%) for age 55 and above—while neural

network (86.72% ± 4.08% and 74.19% ± 2.55%) and logistic regression (83.94% ± 4.05% and 77.96% ± 4.95%) trail in that order. This pattern suggests that tree-based ensembles may more effectively capture nonlinear interactions in patient subpopulations, especially among younger individuals, whereas linear models remain competitive but slightly less adaptable. Regardless of model choice or age strata, the universal importance of chest pain descriptors and vessel involvement underscores their foundational role in clinical risk stratification for heart disease.

- **Experiments with Alternate Datasets:** We explored different datasets to test model robustness. We hypothesized that datasets with clinically consistent and medically accurate features tend to result in higher model performance.

- **Influence of Conflicting Data Instances:** In datasets containing conflicting records (e.g., individuals who consume alcohol daily and have poor sleep habits but do not have heart disease), model performance may degrade due to difficulty in learning meaningful patterns.

8 FUTURE WORKS

- **Improved modeling techniques:** While the three models employed in this project—Logistic Regression, Random Forest, and Neural Network—are widely used and effective, they represent standard approaches. Future research could benefit from developing models more specifically tailored to heart disease prediction. This may include the integration of alternative machine learning techniques, such as ensemble methods, gradient boosting, or even specialized algorithms designed for healthcare applications. Exploring these options could lead to models with improved precision and interpretability in clinical contexts.
- **Potential inclusion of clinical data:** The dataset used in this study contained 14 clinically validated attributes, providing a solid foundation for predictive modeling. However, future iterations of this research could benefit from incorporating richer clinical datasets that include additional features—such as genetic markers, detailed patient history, or lifestyle metrics—which may enhance model performance and lead to more nuanced predictions.
- **Testing with more diverse datasets (with more attributes):** To improve the generalizability and robustness of predictive models, future work should focus on acquiring and testing against larger, more diverse datasets. A broader sample population with a wider range of attributes would help increase model confidence, reduce bias, and support the development of risk stratification tools that are applicable to a broader range of patients.

REFERENCES

[1] Ahmad Ayid Ahmad and Huseyin Polat. Prediction of heart disease based on machine learning using jellyfish optimization algorithm. *Diagnostics (Basel, Switzerland)*, 13(14):2392, 2023.
[2] Hardik Deshmukh. Heart disease uci - diagnosis & prediction, 2020. https://medium.com/data-science/heart-disease-uci-diagnosis-prediction-b1943ee835a7.
[3] Ketan Gangal. Heart disease dataset (uci), 2022. https://www.kaggle.com/datasets/ketangangal/heart-disease-dataset-uci/data.
[4] Md. Touhidul Islam, Sanjida Reza Rafa, and Md. Golam Kibria. Early prediction of heart disease using pca and hybrid genetic algorithm with k-means, 2021.
[5] David Lapp. Heart disease dataset, 2019. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.
[6] Redwan Karim Sony. Uci heart disease data, 2020. https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/data.
[2] [4] [1] [3] [5] [6]