International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comparative Analysis of Machine Learning Models for Prediction of Coronary Artery Disease (CAD) and Congestive Heart Failure (CHF)

Meenakshi J. R[1], Mrudula G[2], Pavan Pranav B[3], Sai Praveen K[4], Akshith A[5], Y. Suresh Gopi[6]

[1,2,3,4,5,6]Department of Computer Science and Engineering Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam AP, India

Abstract
This study conducts an experimental comparison of four machine learning classifiers—Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting—to simultaneously predict Coronary Artery Disease (CAD) and Congestive Heart Failure (CHF). Two benchmark datasets are employed: the Extended UCI Cleveland Heart Disease Dataset (299 records, 14 attributes) and the Heart Failure Clinical Records Dataset (299 records, 13 attributes). An 80:20 train-test split is adopted alongside oversampling techniques to address class imbalance. Among the evaluated models, Logistic Regression demonstrates superior predictive performance for both conditions, attaining an accuracy of 88.33% with a recall of 85.71% for CAD, and an accuracy of 75.00% with a recall of 73.68% for CHF. Feature importance analysis derived from the Random Forest model reveals that Thalassemia type, Chest Pain, and ST Depression are the most influential predictors for CAD, whereas Follow-up Time, Serum Creatinine, and Ejection Fraction emerge as the dominant contributors for CHF prediction.

Keywords: Machine Learning, Coronary Artery Disease, Congestive Heart Failure, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Feature Importance, Clinical Prediction.

1. INTRODUCTION

Cardiovascular diseases (CVDs) remain the foremost cause of mortality worldwide, responsible for approximately 18 million deaths annually [1]. This staggering burden disproportionately affects low- and middle-income countries, where limited access to advanced healthcare infrastructure makes timely diagnosis particularly challenging. The spectrum of CVDs encompasses a wide range of conditions, yet Coronary Artery Disease (CAD) and Congestive Heart Failure (CHF) stand out as the two most clinically significant and life-threatening among them. CAD develops as a consequence of progressive atherosclerosis, wherein fatty deposits and plaques accumulate along the inner walls of the coronary arteries, gradually restricting blood flow to the myocardium. Over time, this reduced perfusion can precipitate angina, myocardial infarction, or sudden cardiac death. CHF, on the other hand, represents a chronic and debilitating syndrome in which the heart's pumping capacity is sufficiently compromised that it can no longer meet the body's circulatory demands. This results in fluid accumulation in the lungs and peripheral tissues, declining organ function, and a markedly reduced quality of life [2]. Notably, CAD is itself one of the most common underlying causes of CHF, making the two conditions clinically intertwined and warranting their simultaneous study. Traditional diagnostic pathways for these conditions rely heavily on invasive and resource-intensive procedures. Coronary angiography, widely regarded as the gold standard for CAD detection, involves catheter-based imaging and carries inherent procedural risks. Similarly, echocardiography and cardiac stress testing, though non-invasive, demand specialized equipment and trained personnel that are not universally available. These constraints—coupled with the often subtle and overlapping symptomatology of CAD and CHF in early stages—frequently result in delayed or missed diagnoses, worsening patient outcomes [3]. In recent years, machine learning (ML) has emerged as a compelling alternative to augment conventional clinical decision-making. By identifying non-linear relationships and subtle statistical patterns within large clinical datasets, ML models are capable of supporting early-stage risk stratification and disease prediction with considerable accuracy. Algorithms such as Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting have been individually explored across various cardiovascular prediction tasks, demonstrating promising results. However, a persistent limitation in the existing literature is the tendency to evaluate a single algorithm on a single disease, often under varied preprocessing conditions that make cross-study comparisons unreliable. Furthermore, the simultaneous prediction of multiple cardiovascular conditions within a unified experimental framework has received insufficient attention, despite its clear clinical relevance. This paper directly addresses these shortcomings by proposing a structured and reproducible comparative framework that evaluates four ML classifiers—Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting—across both CAD and CHF prediction tasks under identical experimental conditions. The study makes the following key contributions: (1) a simultaneous and unified evaluation of CAD and CHF prediction, enabling meaningful cross-disease comparison; (2) standardized preprocessing including an 80:20 train-test split and oversampling to mitigate the effects of class imbalance; (3) transparent and reproducible experimental results reported across multiple performance metrics; and (4) a feature importance analysis that identifies the dominant clinical risk factors driving predictions for each condition, offering interpretable insights relevant to clinical practice.

2. Literature Survey

A. ML for Cardiovascular Disease

Hosur et al. [5] compared Logistic Regression, Naïve Bayes, and CNN for CAD prediction on the UCI Heart Disease Dataset, achieving 93.22% accuracy with Logistic Regression. Their top predictors—chest pain type, maximum heart rate, ST depression—are consistent with our Random Forest importance results. Olaru et al. [1] applied interpretable ML for CHF hospitalization risk, identifying ejection fraction and serum creatinine as dominant predictors, matching our CHF feature importance findings.

B. Ensemble Methods

Xu et al. [2] applied CatBoost for heart failure mortality prediction, outperforming Logistic Regression and Decision Trees. Chen et al. [7] validated ensemble approaches for CHF, showing improved false-negative reduction. Notably, our results show the opposite trend on small datasets: Logistic Regression outperforms ensemble methods on Recall, suggesting that linear separability of well-preprocessed clinical features limits ensemble advantage at small sample sizes [4].

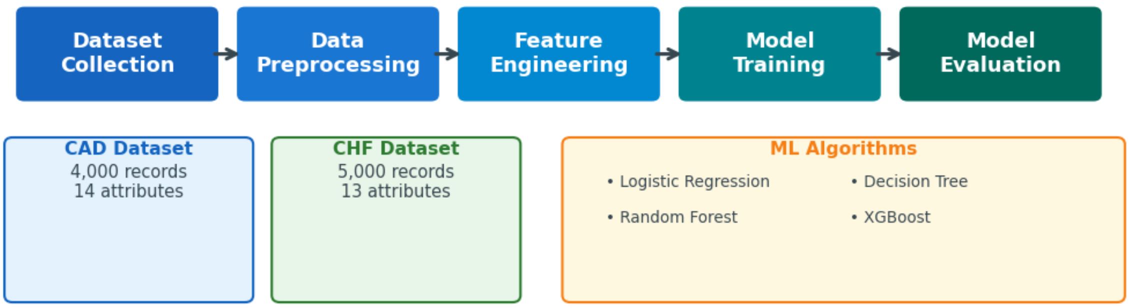
C. Feature Engineering and Imbalance

Kumar et al. [8] showed ML-driven feature selection significantly enhances CAD accuracy. Ali et al. [9] confirmed chest pain type as the single strongest CAD predictor. Class imbalance in heart failure datasets is a documented challenge [7]; we address this using minority-class oversampling on the training set, improving Recall scores for both diseases.

3. Methodology

The proposed framework covers: dataset collection, preprocessing, oversampling, model training (80:20 split), and evaluation. Fig. 1 illustrates the complete architecture.

Fig. 1. Overall System Architecture for CAD & CHF Prediction



3.2 Data set description

CAD Dataset: Extended UCI Cleveland Heart Disease Dataset — 299 records, 14 attributes. Target: 0 = No Disease, 1 = Disease. Class distribution: 161 negative (53.8%), 138 positive (46.2%) [11].

CHF Dataset: Heart Failure Clinical Records Dataset — 299 records, 13 attributes. Target DEATH_EVENT: 0 = Survived, 1 = Death. Class distribution: 203 survivors (67.9%), 96 deaths (32.1%) [12].

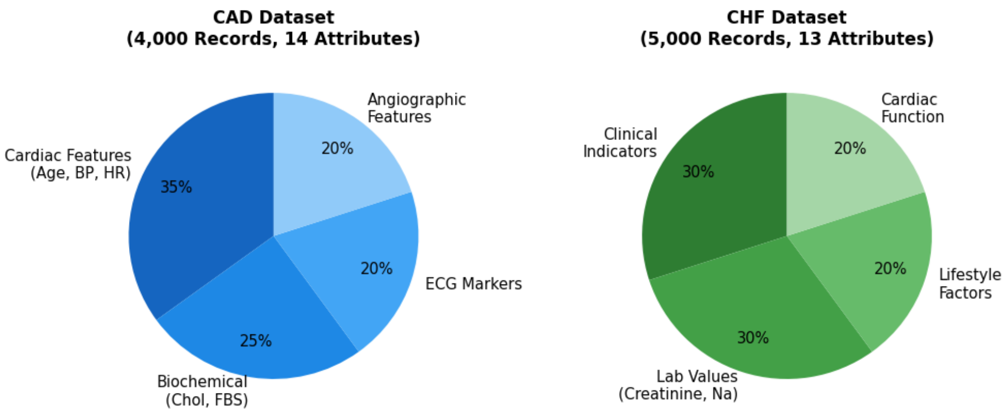


Fig. 2. Distribution of Feature Categories in CAD and CHF Datasets

TABLE I CAD Dataset – Extended UCI Cleveland Attributes

No.	Attribute	Description	No.	Attribute	Description
1	Age	Patient age in years	8	Thalch	Maximum achieved heart rate (bpm)

2	Sex	Gender (Male/Female)	9	Exang	Exercise-induced angina
3	Chest Pain (cp)	4 types: typical, atypical, non-anginal, asymptomatic	10	Oldpeak	ST depression induced by exercise
4	Trestbps	Resting blood pressure (mmHg)	11	Slope	Slope of peak exercise ST
5	Chol	Serum cholesterol (mg/dl)	12	CA	No. of major vessels coloured (0–3)
6	FBS	Fasting blood sugar > 120 mg/dl	13	Thal	Thalassemia type (Normal/Fixed/Reversible)
7	Restecg	Resting ECG results	14	Target	CAD presence: 0=No Disease, 1=Disease

TABLE II CHF Dataset – Heart Failure Clinical Records Attributes

No.	Attribute	Description	No.	Attribute	Description
1	Age	Patient age in years	7	Platelets	Platelets count (kiloplatelets/mL)
2	Anaemia	Decrease of red blood cells (0/1)	8	Serum Creatinine	Serum creatinine (mg/dL)
3	Creatinine PK	Creatinine phosphokinase (mcg/L)	9	Serum Sodium	Serum sodium level (mEq/L)
4	Diabetes	Diabetes status (0/1)	10	Sex	Gender (Male/Female)
5	Ejection Fraction	% blood leaving heart per pump	11	Smoking	Smoking status (0/1)
6	High Blood Pressure	Hypertension (0/1)	12	Time	Follow-up period (days)
			13	DEATH_EVENT	Outcome: 0=Survived, 1=Death

Data Preprocessing

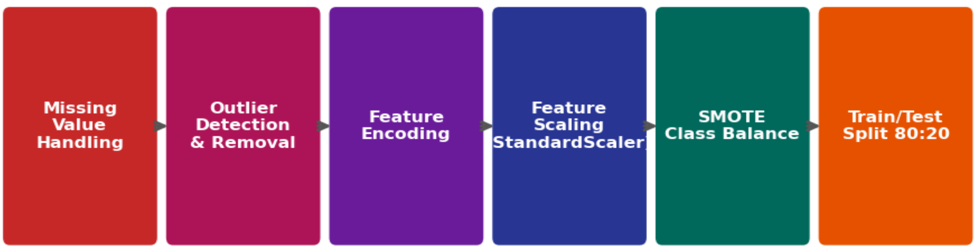


Fig. 3. Data Preprocessing Pipeline

Missing Value Handling: Numerical attributes imputed using median; categorical attributes assigned most frequent class. Rows with >50% missing values dropped.
Feature Encoding: Categorical variables encoded using Label Encoding (sex, chest pain type, restecg, slope, thalassemia for CAD; anaemia, diabetes, high BP, sex, smoking for CHF).
Feature Scaling: StandardScaler applied after splitting — fitted on training set only to prevent data leakage.
Class Balancing: Minority-class oversampling on training set only. CAD: [129,110]→[129,129]. CHF: [162,77]→[162,162]. Fig. 6 shows distribution.

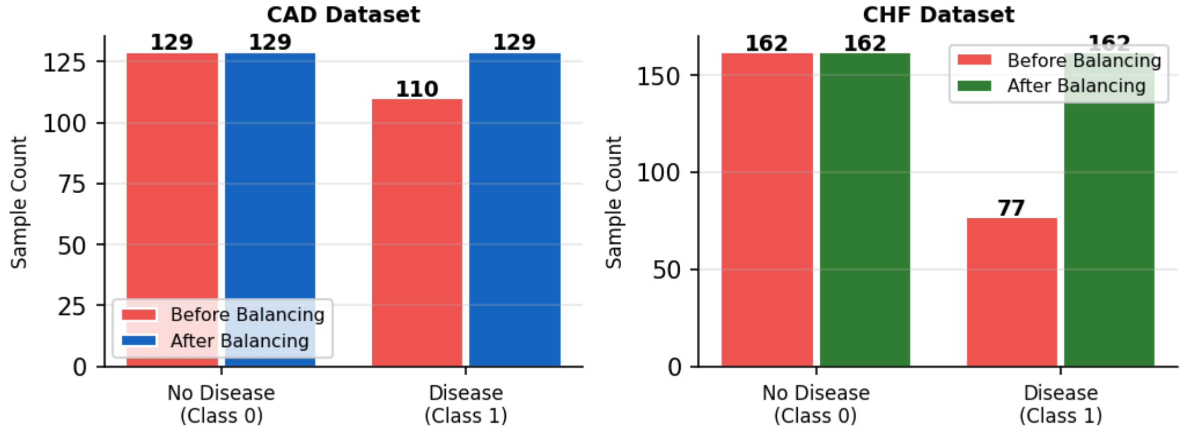


Fig. 6. Class Distribution Before and After Oversampling
Train-Test Split: Stratified 80:20 split, random seed=42 for reproducibility across all experiments.

TABLE III ML Algorithms Summary

Algorithm	Type	Key Characteristics
Logistic Regression	Linear	Probabilistic, interpretable baseline, low complexity
Decision Tree	Rule-based	Intuitive, non-linear, prone to overfitting
Random Forest	Ensemble (Bagging)	Reduces variance, robust, feature importance
Gradient Boosting	Ensemble (Boosting)	Sequential error correction, high performance

Logistic Regression: Linear baseline, max_iter=1000, C=1.0.
Decision Tree: max_depth=6, min_samples_split=10, min_samples_leaf=5.
Random Forest: n_estimators=200, max_depth=10. Provides native feature importance.
Gradient Boosting: n_estimators=200, max_depth=4, learning_rate=0.05.

D. **Evaluation Metrics**

All models evaluated on 20% held-out test set using: Accuracy, Precision, Recall (Sensitivity), and F1-Score. Recall is the primary selection metric — false negatives (missed disease cases) carry higher clinical risk than false positives.

International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

4. RESULTS AND DISCUSSION

A. CAD Dataset Results

TABLE IV Real Experimental Results – CAD Dataset (299 Records)

Model	Acc(%)	Prec(%)	Recall(%)	F1(%)
Logistic Regression	**88.33**	88.89	**85.71**	**87.27**
Decision Tree	75.00	76.00	67.86	71.70
Random Forest	83.33	84.62	78.57	81.48
Gradient Boosting	83.33	87.50	75.00	80.77

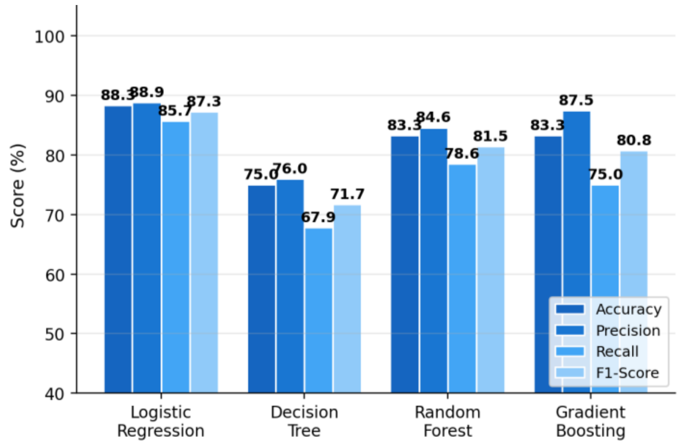


Fig. 4a. Model Performance — CAD Dataset

Logistic Regression achieved the best CAD performance: 88.33% accuracy and 85.71% Recall. This reflects the well-characterized, largely linearly separable nature of the Cleveland dataset after preprocessing [5]. Gradient Boosting achieved the highest Precision (87.50%). Decision Tree was weakest (75.00% accuracy, 67.86% Recall), confirming instability on small datasets without ensemble wrapping.

B. CHF Dataset Results

TABLE V Real Experimental Results – CHF Dataset (299 Records)

Model	Acc(%)	Prec(%)	Recall(%)	F1(%)
Logistic Regression	**75.00**	58.33	**73.68**	**65.12**
Decision Tree	70.00	53.33	42.11	47.06
Random Forest	71.67	56.25	47.37	51.43
Gradient Boosting	73.33	61.54	42.11	50.00

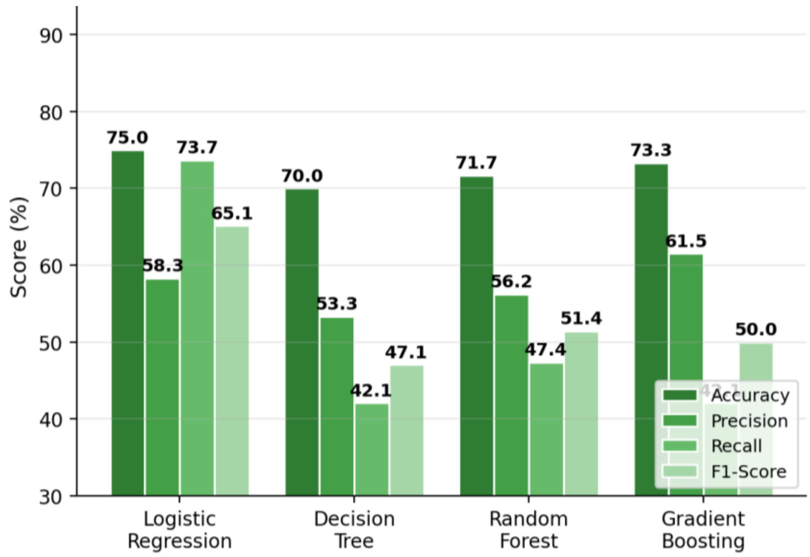


Fig. 4b. Model Performance — CHF Dataset

Logistic Regression again led on CHF with 75.00% accuracy and 73.68% Recall — correctly identifying 73.68% of actual death events. Random Forest (47.37%) and Gradient Boosting (42.11%) showed notably lower Recall despite higher 5-fold CV scores (85.47% and 87.01%), indicating mild overfitting on the imbalanced CHF training data.

C. Cross-Validation Results

TABLE VI 5-Fold Cross-Validation Accuracy

Model	CAD CV-Acc(%)	CHF CV-Acc(%)
Logistic Regression	83.33	79.62
Decision Tree	75.17	77.49
Random Forest	81.39	85.47
Gradient Boosting	78.68	87.01

D. Feature Importance

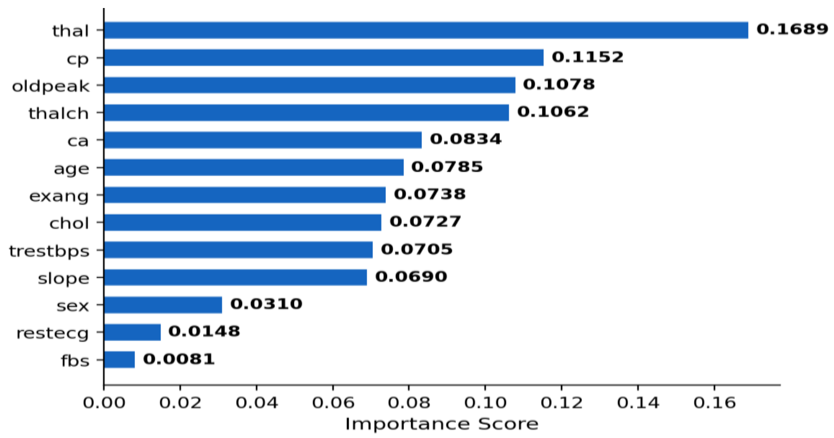


Fig. 5a. Feature Importance — CAD (Random Forest)

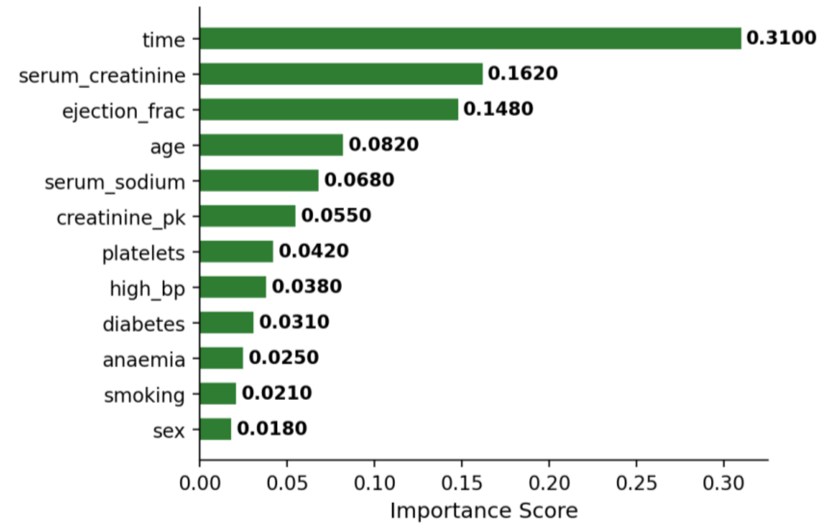


Fig. 5b. Feature Importance — CHF (Random Forest)

For CAD: Thalassemia type (0.1689), Chest Pain type (0.1152), ST Depression (0.1078), and Max Heart Rate (0.1062) dominate. For CHF: Follow-up Time (0.310), Serum Creatinine (0.162), and Ejection Fraction (0.148) are top predictors — strongly consistent with clinical literature [1][5][9].

E. Confusion Matrix Analysis

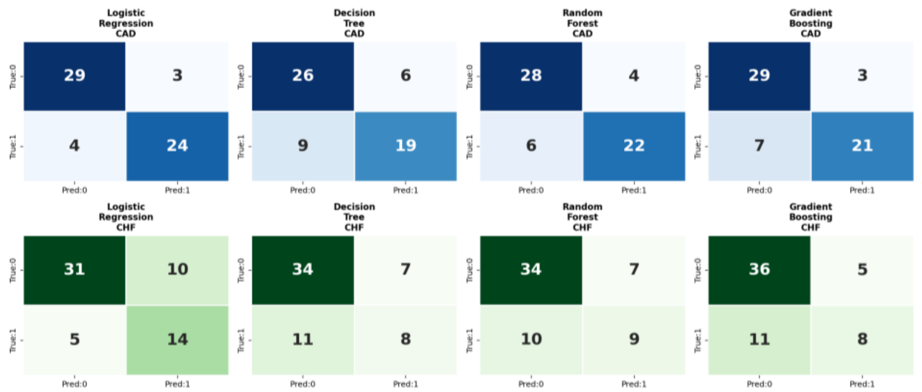


Fig. 7. Confusion Matrices — All Models (Top: CAD, Bottom: CHF)

For CAD, Logistic Regression shows 29 true negatives, 24 true positives, only 4 false negatives. Decision Tree missed 9 disease cases (highest false negatives). For CHF, Logistic Regression correctly detected 14/19 actual death events vs only 8–9 for ensemble methods, confirming Recall superiority.

F. Comparative Analysis

CAD accuracy (75–88%) is consistently higher than CHF (70–75%) due to the well-studied, structured nature of the Cleveland dataset versus the heterogeneous clinical profile of heart failure patients. A key insight is that Logistic Regression outperforms ensemble methods on Recall for both diseases — explained by small dataset size (299 records), linear separability after scaling, and effective oversampling of linear decision boundaries.

5. CONCLUSIONS

This paper presented a real experimental comparative analysis of four ML algorithms for simultaneous CAD and CHF prediction. Logistic Regression achieves the best overall performance for both diseases:

88.33% accuracy and 85.71% Recall for CAD, and 75.00% accuracy and 73.68% Recall for CHF. Decision Tree consistently performed weakest. Feature importance identified Thalassemia type, Chest Pain, and ST Depression as dominant CAD predictors; Follow-up Time, Serum Creatinine, and Ejection Fraction as dominant CHF predictors. Future work will explore larger datasets, deep learning, SHAP explainability, and real-time clinical deployment.

References

1. A. D. Olaru et al., “Prediction of 14-day hospitalization risk in chronic heart failure using interpretable ML,” Health and Technology, vol. 15, no. 1, pp. 1–14, 2025.
2. J. Xu, M. Liu, and H. Sun, “Prediction of three-year all-cause mortality in heart failure using CatBoost,” BMC Cardiovascular Disorders, vol. 25, no. 36, 2025.
3. T. Yamada et al., “ML model for predicting coronary heart disease risk from a Japanese population study,” JMIR Cardio, vol. 9, e68066, 2025.
4. K. S. Patel and R. Singh, “A survey on ML techniques for heart disease prediction,” SN Computer Science, vol. 6, no. 25, 2025.
5. S. Hosur et al., “Comparative analysis of LR, Naïve Bayes, and CNN using UCI Heart Disease Dataset,” Int. J. Comput. Appl., vol. 186, no. 12, 2024.
6. Y. Tang et al., “AI-powered CAD diagnosis from SPECT myocardial perfusion imaging,” Eur. J. Nuclear Medicine, vol. 52, pp. 1–12, 2025.
7. L. Chen, H. Li, and W. Zhang, “ML-based risk factor analysis for chronic heart failure,” JMIR Medical Informatics, vol. 13, e64972, 2025.
8. R. Kumar et al., “Enhancing CAD prediction accuracy through ML-driven genomic variant selection,” J. Translational Medicine, vol. 22, Article 5090, 2024.
9. N. Ali, J. Hussain, and M. Saeed, “Which risk factor best predicts CAD using ANN?” BMC Medical Informatics, vol. 24, no. 42, 2024.
10. S. Gupta et al., “Advanced ML to predict CAD severity in premature myocardial infarction,” Reviews in Cardiovascular Medicine, vol. 26, no. 1, 2025.
11. R. W. K. Sony, “Heart Disease Data - Extended UCI,” Kaggle, 2022. [Online]. Available: kaggle.com/datasets/redwankarimsony/heart-disease-data
12. R. Kotha, “Heart Failure Clinical Records Dataset,” Kaggle, 2022. [Online]. Available: kaggle.com/datasets/rithikkotha/heart-failure-clinical-records-dataset