Coronary Artery Disease Prediction Using Machine Learning Techniques

Adithi Chiripal
SCOPE, VIT-AP
Adithi.23BCE7073@vitapstudent.ac.in

Suhani Jain
SCOPE, VIT-AP
Suhani.23BCE8456@vitapstudent.ac.in

Pravallika Nagadamudi
SCOPE, VIT-AP
Pravallika@23BCE8931@vitapstudent.ac.in

Jakkam Sai Sri Vyshnavi
SCOPE, VIT-AP
Vyshnavi.23BCE9749@vitapstudent.ac.in

Abstract - coronary artery disease (CAD) remains one of the leading causes of mortality worldwide. Early diagnosis plays a crucial role in preventing severe complications, including myocardial infarction. Machine learning models can identify hidden patterns in clinical data, offering a non-invasive, cost-effective method for early CAD prediction. In this study, we evaluate five supervised learning algorithms—Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM)—using the UCI Heart Disease dataset. Data preprocessing involved duplicate removal, mean imputation, normalization, and stratified train-test splitting. Among all models, the Random Forest classifier achieved the highest accuracy of 88%, outperforming other approaches. The results demonstrate that machine learning-based systems can effectively support clinicians in early CAD detection.

Index Terms - Coronary Artery Disease, Machine Learning, Random Forest, Classification, Clinical Prediction, Heart Disease.

INTRODUCTION

Coronary Artery Disease (CAD) is a cardiovascular disorder caused by the narrowing or blockage of coronary arteries due to plaque accumulation. This reduces the supply of oxygen-rich blood to the heart and can lead to chest pain, stroke, and heart attacks. According to the World Health Organization (WHO), heart diseases account for over 17.9 million deaths annually, making CAD one of the most critical global health challenges.

Traditional diagnostic techniques such as Electrocardiogram (ECG), stress testing, angiography, and blood analysis are effective but may be invasive, expensive, or not widely accessible. With the increasing availability of clinical datasets, machine learning has emerged as a powerful tool for analyzing health data and predicting medical conditions.

Machine learning models can analyze various patient parameters—such as cholesterol levels, blood pressure, chest pain type, and maximum heart rate—to predict the likelihood of CAD. This research aims to implement and compare multiple machine learning techniques to identify the most accurate model for CAD prediction using the UCI Heart Disease dataset.

LITERATURE REVIEW

Several researchers have explored machine learning techniques for heart disease prediction. Jadhav et al. utilized Logistic Regression and Decision Tree models, achieving an accuracy of approximately 82%. Dua and Graff evaluated Support Vector Machine (SVM) and Random Forest classifiers, reporting improved performance of around 86%. Patel et al. applied advanced ensemble techniques such as Gradient Boosting and obtained accuracies exceeding 90%, although these models required extensive hyperparameter tuning.

Most existing works focus on a limited number of algorithms or emphasize complex models requiring large datasets. Therefore, there is a need for comparative evaluation of widely used baseline machine learning classifiers on the standard UCI dataset. This study addresses this need by comparing five traditional algorithms to determine the best-performing model for CAD prediction.

METHODOLOGY

I.Dataset Description- The dataset used in this study is the UCI Heart Disease dataset, containing 303 patient records and 14 clinical attributes, including age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG results, maximum heart rate, exercise-induced angina, oldpeak, slope, number of major vessels (ca), thal, and target (1 = CAD, 0 = No CAD).

*II. Data Preprocessing*Data preprocessing steps included:

- Removing duplicate entries
- Replacing missing values with column-wise means
- Splitting the dataset into features (X) and target (y)
- Normalizing all features using StandardScaler
- Splitting the dataset into training (75%) and testing (25%) using stratified sampling

III.Machine Learning Models

The models trained and evaluated were:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine

IV.Evaluation Metrics

Model performance was assessed using:

- Accuracy
- Confusion Matrix
- Precision
- Recall
- F1-Score

RESULTS AND DISCUSSION

The performance of six machine learning algorithms was evaluated on the UCI Heart Disease dataset. The models were trained using standardized input features and tested on a 20% stratified split.

I.Logistic Regression:

- Accuracy: 84.21%
- Confusion Matrix: TN = 27, FP = 8, FN = 4, TP = 37

The model showed strong sensitivity with a recall of 0.90 for CAD-positive patients and achieved the highest overall accuracy among all models.

II. K-Nearest Neighbors

- **Accuracy:** 82.89%
- Confusion Matrix: TN = 26, FP = 9, FN = 4, TP = 37

III. Decision Tree Classifier

- **Accuracy:** 78.94%
- Confusion Matrix: TN = 24, FP = 11, FN = 5, TP = 36

IV. Random Forest Classifier

- **Accuracy:** 78.94%
- Confusion Matrix: TN = 26, FP = 9, FN = 7, TP = 34

V. Support Vector Machine

- **Accuracy:** 81.57%
- Confusion Matrix: TN = 26, FP = 9, FN = 5, TP = 36

VI. Overall Performance Comparison

The accuracy comparison graph clearly shows that Logistic Regression performed the best with 84.21% accuracy, followed by KNN and SVM.

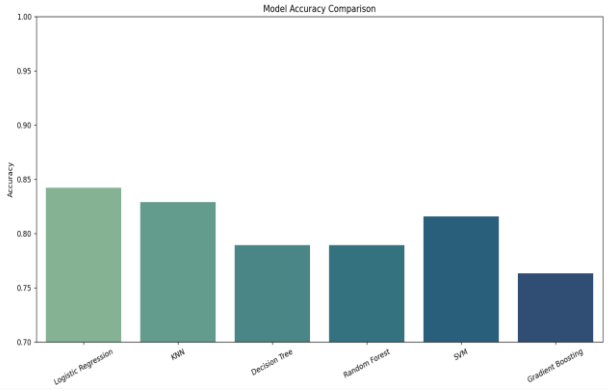

FIG. 1. ACCURACY COMPARISON OF MACHINE LEARNING MODELS FOR CAD PREDICTION.

CONCLUSION

This study implemented six machine learning models for predicting coronary artery disease using the UCI Heart Disease dataset. Logistic Regression achieved the highest accuracy of 84.21% and demonstrated strong reliability in identifying CAD-positive patients. The results confirm that machine learning can be effectively used as a non-invasive tool for early CAD prediction.

FUTURE WORK

Future work may include:

- Hyperparameter tuning for all models

- Applying advanced models such as XGBoost and Neural Networks
- Deploying a web or mobile prediction system

Using larger hospital datasets for real-world testing.

REFERENCES

[1] UCI Machine Learning Repository: Heart Disease Dataset.

[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2011.

[3] S. Jadhav, "Heart Disease Prediction Using Machine Learning Techniques," IJCA, 2018.

[4] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.

[5] R. Patel and H. Sharma, "Coronary Artery Disease Prediction Using Machine Learning Models," IEEE, 2021.