IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058

An Intelligent Machine Learning Framework for Early Cardiovascular Disease Risk Prediction

Gnanamani. H[1], Gagana G[2], Inchara Bhatta K.M[3] , Kavana[4], Likitha. H[5]

Assistant Professor, Department of Computer Science, East West College of Engineering, Bengaluru, India[1]

Student, Department of Computer Science, East West College of Engineering, Bengaluru, India[2]

Student, Department of Computer Science, East West College of Engineering, Bengaluru, India[3]

Student, Department of Computer Science, East West College of Engineering, Bengaluru, India[4]

Student, Department of Computer Science, East West College of Engineering, Bengaluru, India[5]

Abstract: Heart disease is one of the leading causes of death worldwide, making early prediction essential for improving patient outcomes. Machine learning techniques have gained significant attention for predicting heart disease by analyzing medical and clinical data. This survey paper reviews recent studies on heart disease prediction using various machine learning algorithms, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, K-Nearest Neighbors, Gradient Boosting, and XGBoost. It also discusses feature selection techniques, data preprocessing methods such as SMOTE, and model optimization approaches that improve prediction performance. Based on the reviewed literature, ensemble learning methods, particularly XGBoost and Gradient Boosting, demonstrate superior accuracy and reliability. This survey highlights the strengths, limitations, and future scope of machine learning-based heart disease prediction systems and emphasizes their potential to support early diagnosis and clinical decision-making.

Keywords: Heart Disease Prediction, Machine Learning, Cardiovascular Disease, Classification Algorithms, Explainable Artificial Intelligence (XAI), Feature Selection, SMOTE, Healthcare Analytics.

I. INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of mortality worldwide and poses a major public health challenge. It includes a wide range of disorders such as coronary artery disease, heart failure, arrhythmia, and stroke, which significantly affect the quality of life and increase healthcare costs. According to the World Health Organization (WHO), cardiovascular diseases account for nearly 17.9 million deaths annually, representing approximately 32% of all global deaths. The growing prevalence of risk factors such as hypertension, diabetes, obesity, smoking, and unhealthy lifestyles has further increased the burden of heart disease, making early diagnosis and timely intervention essential for reducing mortality and improving patient outcomes.

Traditionally, heart disease diagnosis relies on clinical examination, laboratory investigations, electrocardiograms (ECG), imaging techniques, and the expertise of medical professionals. Although these approaches provide reliable results, they are often time-consuming, expensive, and may not accurately identify patients at an early stage of disease progression. Moreover, the increasing volume and complexity of healthcare data make it difficult for physicians to extract meaningful insights using conventional diagnostic methods alone. Consequently, there is a growing need for intelligent decision-support systems that can assist clinicians in predicting heart disease with greater accuracy and efficiency.

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have significantly transformed the field of healthcare by enabling data-driven disease prediction and diagnosis. Unlike traditional statistical methods, machine learning algorithms can learn complex relationships from large-scale clinical datasets and accurately classify patients based on multiple risk factors. Various supervised learning algorithms, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Gradient Boosting, XGBoost, and Artificial Neural Networks (ANN), have demonstrated promising performance in heart disease prediction. Furthermore, deep learning techniques have enhanced predictive capabilities by automatically extracting complex features from medical data, thereby reducing the need for manual feature engineering.

Several recent studies have focused on improving prediction accuracy through advanced preprocessing techniques, feature selection methods, hyperparameter optimization, and class imbalance handling approaches such as the Synthetic

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.151058

Minority Over-sampling Technique (SMOTE). Optimization algorithms, including Particle Swarm Optimization (PSO) and its enhanced variants such as Adaptive Inertia Weight Particle Swarm Optimization (AIW-PSO), have been employed to optimize machine learning models, resulting in improved predictive performance. These studies commonly utilize benchmark datasets such as the UCI Heart Disease Dataset, Cleveland Heart Disease Dataset, Framingham Heart Study Dataset, and Heart Failure Clinical Records Dataset to evaluate model effectiveness using performance metrics including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Key Contributions from the Study

1. The study compares different machine learning algorithms for predicting heart disease and evaluates their performance using metrics such as accuracy, precision, sensitivity, specificity, and F1-score.
2. It identifies important health-related features, including age, blood pressure, cholesterol, glucose levels, and other clinical parameters, that support heart disease prediction.
3. It explores feature selection techniques and data-balancing methods, such as Chi-square, mutual information, and SMOTE, to improve model performance and address imbalanced datasets.
4. It investigates several machine learning models, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Gradient Boosting, and XGBoost, to identify effective approaches for cardiovascular disease prediction.
5. It explores optimized machine learning models and explainable AI techniques, including an AIW-PSO-optimized Gradient Boosting model and SHAP-based explanations, to improve prediction performance and help interpret the factors influencing heart disease and heart failure outcomes.

II. RELATED WORK

Several research studies have explored the use of machine learning algorithms for predicting heart disease and improving cardiovascular risk assessment. The following related work is based on the five IEEE research papers provided.

There has been significant research interest in the use of machine learning for heart disease prediction due to the importance of early risk identification and the challenges involved in analyzing complex clinical and physiological data.

Earlier approaches to cardiovascular risk assessment relied heavily on traditional statistical methods and clinical risk factors. Although these approaches provide useful information, predicting heart disease can be challenging because patient characteristics and clinical measurements may have complex relationships. Machine learning techniques have therefore been explored to identify patterns in medical datasets and support more effective disease prediction [1], [2].

Machine learning algorithms such as Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Gradient Boosting have been investigated for cardiovascular disease classification. Comparative studies evaluate these algorithms using performance measures such as accuracy, precision, sensitivity, specificity, F1-score, and AUC-ROC. Their findings show that model performance varies depending on the dataset, selected features, and evaluation method [1], [2], [3].

Ensemble learning techniques, including Random Forest, Gradient Boosting, and XGBoost, have also been investigated to improve predictive performance. One study reported 79.34% accuracy using Random Forest, while another achieved 97.57% accuracy using XGBoost with combined datasets and a particular feature-selection configuration. These results indicate the potential of ensemble methods, although differences in datasets and experimental settings mean that the reported accuracies cannot be compared directly [1], [4].

Recent studies have also explored optimized models and explainable artificial intelligence to improve prediction reliability and interpretability. The heart failure study combines Gradient Boosting Machine with Adaptive Inertia Weight Particle Swarm Optimization (AIW-PSO) to optimize model parameters, while another study uses SHAP to explain the influence of features on predictions [4], [5].

The current project focuses on developing a machine learning-based heart disease prediction system by comparing classification algorithms and applying suitable data preprocessing and feature selection techniques. The aim is to identify an effective model and present prediction results in an understandable way. The reviewed studies provide a foundation for selecting appropriate algorithms and evaluation methods for this project.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058

TABLE I : COMPARATIVE ANALYSIS OF MACHINE LEARNING-BASED HEART DISEASE PREDICTION SYSTEM

Reference	Approach / Technique	Dataset / Evaluation	Main Focus	Limitation
Afridi et al.[2]	Random Forest, KNN, Decision Tree, and other ML algorithms.	Kaggle dataset (70,000 records); accuracy and F1-score.	Early heart disease prediction.	79.34% maximum accuracy; needs testing on larger datasets.
Bagadi, K. et al[4]	Logistic Regression, Random Forest, Naïve Bayes, SVM, Gradient Boosting.	UCI dataset (303 records); 80:20 train-test split.	Early heart disease prediction using ML.	Small dataset, possible bias, and limited interpretability.
Shuge Ouyang[7]	Random Forest, Decision Tree, and SVM.	UCI dataset; 303 samples, evaluated using accuracy.	Predicting heart disease using ML.	Small dataset, overfitting, and low interpretability.
H. F. El-Sofany[8]	Used 10 ML algorithms, feature selection, and SMOTE.	503 records; evaluated using accuracy, precision, recall, F1-score, and AUC.	Early heart disease prediction; XGBoost achieved 97.57% accuracy.	Small dataset and limited testing across different populations.
Ahmed et al.[10]	GBM, AIW-PSO, and SMOTE.	299 patients; accuracy, precision, recall, F1-score, and AUC	Predicting heart failure patient survival.	Small dataset; needs testing on larger datasets.

presents a comparative analysis of existing machine learning-based heart disease prediction systems. The reviewed studies primarily utilize algorithms such as Random Forest, KNN, Decision Tree, Logistic Regression, Naive Bayes, SVM, XGBoost and GBM, and are evaluated mostly on the UCI dataset (303 records) and Kaggle datasets. While most works focus on early heart disease prediction and achieve promising results, including 97.57% accuracy reported by H. F. El-Sofany et al. using XGBoost, common limitations observed across the studies are small dataset size, limited generalizability, overfitting, low interpretability, and lack of testing on larger and diverse populations. These gaps provide the motivation for developing a more robust and generalizable prediction model in the proposed system.

III. PROPOSED SYSTEM

The proposed system aims to develop a machine learning-based heart disease pre``1qadiction system that analyzes patient health data and predicts the likelihood of heart disease. Based on the five reviewed IEEE papers, the system incorporates data preprocessing, feature selection, classification algorithms, and performance evaluation to improve prediction reliability.

The system begins by collecting a medical dataset containing patient information such as age, gender, blood pressure, cholesterol levels, glucose levels, and other relevant clinical parameters. The collected data is preprocessed to handle missing or inconsistent values and prepare the dataset for model training. If the dataset contains imbalanced classes, the Synthetic Minority Over-sampling Technique (SMOTE) can be applied to improve the representation of minority classes.

The system comprises several key stages, such as data collection, data preprocessing, feature selection, model training, model testing, heart disease prediction, and result analysis.

The overall workflow can be represented as:

Heart Disease Dataset → Data Preprocessing → Feature Selection → ML Model Training → Heart Disease Prediction → Performance Evaluation → Result Visualization

The proposed system is designed to separate the machine learning model from the user interface. The trained model can



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.151058

be integrated with a backend to process patient data and generate predictions, while the frontend can provide a simple interface for users to enter medical details and view the prediction results.

Different machine learning algorithms, such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting, can be evaluated to identify a suitable model. Techniques such as feature selection and SMOTE may be used when appropriate to improve data preparation and address class imbalance.

The architecture of a machine learning-based heart disease prediction system involves collecting patient data, including clinical, demographic, lifestyle, and laboratory details. The data is preprocessed by handling missing values, removing inconsistencies, and normalizing features. Relevant features are selected and given to machine learning algorithms such as Random Forest, SVM, Decision Tree, Logistic Regression, XGBoost, and KNN for training and prediction. The models are evaluated using accuracy, precision, recall, F1-score, and AUC-ROC, and the best model is used for predicting heart disease risk. This system helps in early detection and supports clinical decision-making. The proposed system provides an efficient and reliable method for early heart disease detection. It helps healthcare professionals identify high-risk patients at an early stage and supports better decision-making. The system can also reduce manual analysis time and improve the accuracy of diagnosis. Machine learning techniques enable continuous improvement of prediction results as more patient data becomes available. This architecture provides a scalable approach for developing intelligent healthcare applications.

A. System Architecture

The system architecture consists of six major components: data collection, data preprocessing, feature selection, machine learning algorithms, model evaluation, and heart disease prediction.

The machine learning component uses algorithms such as Random Forest, Decision Tree, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, and XGBoost to analyze the processed data and identify patterns associated with heart disease. The model evaluation component assesses the performance of the trained models using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to determine their effectiveness.

The selected machine learning model processes the patient's input data and generates a prediction indicating the potential risk of heart disease. The prediction results are presented as low-risk or high-risk classifications, depending on the system design. The overall architecture integrates data processing, machine learning, and prediction components to support the early identification of potential heart disease risk.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058

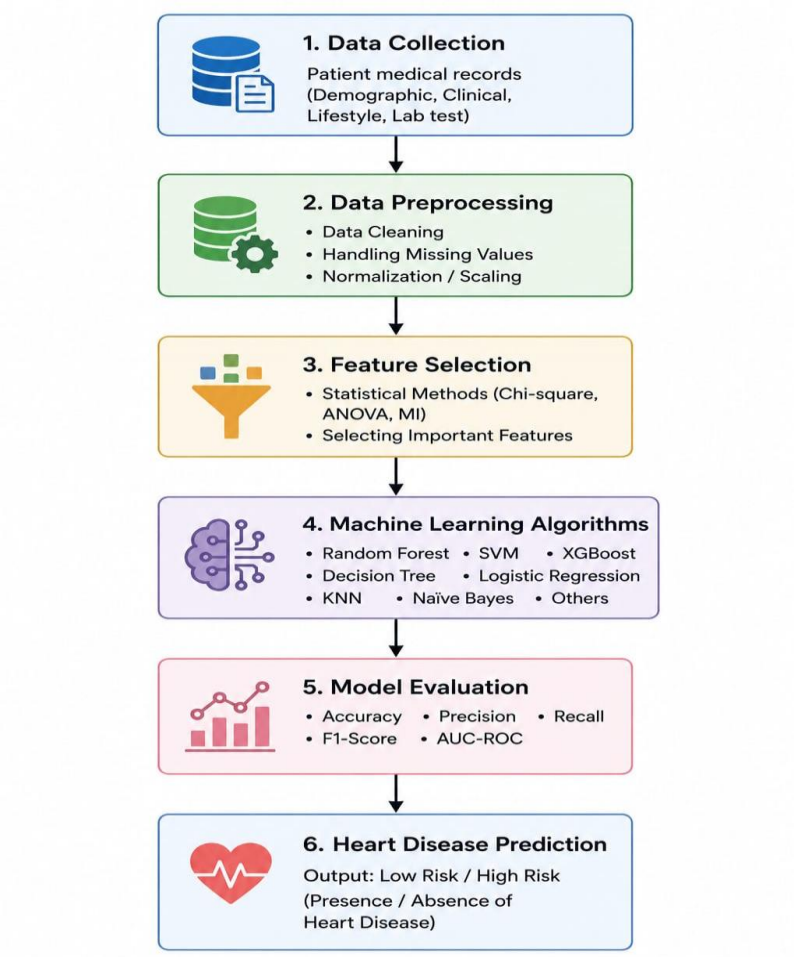


Figure 1. the general architecture of the Machine Learning-based Heart Disease Prediction System

IV. METHODOLOGY

A. Dataset

The proposed system uses a heart disease dataset containing patient medical information for developing and evaluating machine learning models. The dataset may include attributes such as age, gender, blood pressure, cholesterol level, blood glucose level, and other clinical parameters related to heart disease.

The reviewed research papers use different datasets for evaluating heart disease prediction techniques. These include cardiovascular disease data, publicly available heart disease datasets, and clinical records of heart failure patients. The studies demonstrate the importance of selecting appropriate datasets for training and testing machine learning models.

The dataset is divided into training and testing subsets to evaluate the performance of the developed model. The training data is used to build the model, while the testing data is used to measure its ability to predict outcomes on unseen records. The dataset and split ratio should be selected according to the actual implementation of the proposed system.

B. Data Preprocessing

Data preprocessing is an important stage because the quality and consistency of medical data can affect the performance of machine learning models. Before training, the dataset is examined to identify missing values, duplicate records, inconsistent entries, and unsuitable data formats.

The preprocessing stage involves cleaning the data, handling missing or invalid values, encoding categorical attributes into numerical form, and scaling numerical features when required by the selected algorithm. These steps help prepare the dataset for effective model training.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058

When the dataset contains an imbalance between heart disease classes, the Synthetic Minority Over-sampling Technique (SMOTE) can be applied to the training data. This technique helps balance the class distribution and is discussed in the reviewed studies. Preprocessing operations used during training must also be applied consistently to new patient data during prediction.

A simplified preprocessing pipeline is:

Raw Dataset → Data Cleaning → Missing Value Handling → Categorical Encoding → Feature Scaling (if required) → Train/Test Dataset

C. Feature Preparation

Feature preparation involves selecting the relevant medical attributes that contribute to heart disease prediction. The dataset may contain several patient-related features, but some features may provide more useful information for classification than others.

The reviewed studies investigate feature selection techniques such as Chi-square, Analysis of Variance (ANOVA), and Mutual Information to identify relevant attributes. These techniques help reduce unnecessary features and prepare the input data for model training. The selected features depend on the dataset and the requirements of the chosen machine learning algorithm.

The same set of features and the same feature order must be maintained during both training and prediction. Any difference between the input features used during model training and those provided during prediction may lead to incorrect results.

T**he feature preparation process can be represented as:**

Pre-processed Data → Feature Selection → Relevant Medical Attributes → Model-Ready Dataset

V. MACHINE LEARNING MODELS

The proposed heart disease prediction system incorporates five supervised machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). These algorithms have been selected based on their different classification approaches, including linear, tree-based, ensemble, distance-based, and margin-based learning methods, to compare their performance in predicting heart disease from patient medical data.

1. Logistic Regression (LR)

Logistic Regression is a classification algorithm that predicts the probability of heart disease based on patient medical attributes. It is commonly used for binary classification problems.

2. Decision Tree (DT)

Decision Tree classifies patients using a series of decision rules based on medical features. It is easy to understand and interpret.

3. Random Forest (RF)

Random Forest combines multiple decision trees to produce a final prediction. It can improve prediction performance and reduce overfitting.

4. Support Vector Machine (SVM)

SVM classifies patient data by identifying a decision boundary that separates different classes. It is useful for classification tasks involving multiple medical attributes.

5. K-Nearest Neighbors (KNN)

KNN predicts the class of a patient by comparing their medical data with similar records in the dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058

6. Naive Bayes (NB)

Naive Bayes uses probability and assumes that input features are conditionally independent given the class. It can be used to classify patients into different heart disease categories.

7. Gradient Boosting (GB)

Gradient Boosting builds a sequence of models, with each new model aiming to correct errors made by previous models. It can improve classification performance.

8. XGBoost

XGBoost is an optimized boosting algorithm that combines multiple decision trees to make predictions. One of the reviewed studies reported strong results using XGBoost for heart disease prediction.

9. AdaBoost

AdaBoost combines multiple weak classifiers into a stronger model by assigning greater importance to records that were previously misclassified.

10. Voting Classifier

Voting Classifier combines predictions from multiple machine learning models to produce a final result. It can use the majority vote or combine predicted probabilities.

VI. MODEL TRAINING AND EVALUATION

Once the patient medical dataset is preprocessed, it is divided into training and testing sets. The training set is used to train the machine learning algorithms to identify relationships between patient health attributes and the presence or absence of heart disease. The testing set is used to evaluate the trained models and measure their ability to predict heart disease in unseen patient data.

The general training procedure is as follows:

1. Collect the heart disease dataset containing patient medical records.
2. Handle missing values, duplicate records, and inconsistent data.
3. Separate the input features from the target variable.
4. Select relevant features such as age, blood pressure, cholesterol level, and maximum heart rate, depending on the dataset.
5. Apply suitable encoding and feature scaling techniques where required.
6. Divide the dataset into training and testing sets.
7. Train the selected machine learning algorithms using the training data.
8. Predict heart disease outcomes using the testing data.
9. Evaluate the models using appropriate performance metrics.
10. Select and save the best-performing model for future predictions.

The proposed heart disease prediction system can use machine learning algorithms such as Decision Tree, Random Forest, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost. These algorithms analyze patient health attributes to identify patterns associated with heart disease and classify the prediction according to the target labels in the dataset.

The trained models are evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics help compare the effectiveness of different algorithms and determine their suitability for heart disease prediction.

Accuracy

measures the proportion of correctly predicted cases among all tested cases.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058

Precision

measures how many patients predicted as having heart disease are correctly classified.

$$\textbf{Precision}= \frac{TP}{TP+FP}$$

Recall

measures how many patients who actually have heart disease are correctly identified by the model.

$$\textbf{Recall}= \frac{TP}{TP+FN}$$

F1-score

provides a balance between precision and recall.

$$\textbf{F1}=2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where,

TP represents true positives,
TN represents true negatives,
FP represents false positives,
FN represents false negatives

These evaluation metrics help assess the reliability of the heart disease prediction system. Accuracy indicates overall predictive performance, while precision and recall are particularly important for evaluating incorrect positive predictions and missed heart disease cases. The F1-score provides a balanced measure of precision and recall. Finally, the models are compared based on their evaluation results to identify the most suitable algorithm for predicting heart disease risk.

VII. SYSTEM IMPLEMENTATION

The proposed Heart Disease Prediction System is designed as a web-based application consisting of a frontend layer, backend layer, and machine learning module. The implementation approach is based on the techniques discussed in the five research papers, including data preprocessing, feature selection, model training, prediction, and performance evaluation.

A. Frontend

The frontend is the graphical user interface of the system. It provides an interface through which users can enter patient health details, upload a heart disease dataset in CSV format, view prediction results, and monitor model performance.

The dashboard displays information such as:

- Total patient records analyzed
- Number of patients predicted to have heart disease
- Number of patients predicted not to have heart disease
- Important health features used for prediction
- Prediction results and confidence scores, where supported by the model
- Model performance metrics
- Graphs showing prediction and evaluation results

Charts and graphical components help users understand the prediction results easily without needing to interpret the machine learning output directly.

B. Backend

The backend acts as the communication layer between the frontend and the machine learning models. It receives patient information or uploaded CSV files, validates the input data, performs the required preprocessing, and passes the processed data to the trained model for prediction.

DOI: 10.17148/IJARCCE.2026.151058

The backend can be developed using Python and Flask, while libraries such as Pandas, NumPy, and Scikit-learn can support data processing and machine learning operations.

The general prediction flow is:

Patient Details / CSV Upload → Backend API → Data Validation → Data Preprocessing → Feature Selection → Model Prediction → Result Processing → Frontend Dashboard

This architecture allows the machine learning module to operate independently of the user interface.

C. Machine Learning Module

The machine learning module is the core component of the heart disease prediction system. Based on the algorithms explored in the research papers, models such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, XGBoost, and Gradient Boosting can be trained and evaluated.

The research papers also discuss feature selection techniques such as Chi-square, ANOVA, and mutual information. Some studies use SMOTE to address class imbalance and optimization techniques to improve model performance. The most suitable model can be selected based on its evaluation results.

D. Prediction Pipeline

When the user enters patient details or uploads a CSV file, the backend reads the input and checks whether the required features are available in the correct format. The input data is then preprocessed using the same procedures applied during model training.

If feature selection or data transformation was used during training, the same fitted process is applied during prediction. The processed features are passed to the selected machine learning model, which predicts the corresponding class according to the dataset's target labels.

The prediction result is then returned to the frontend and displayed on the dashboard. This approach maintains consistency between the training and prediction st

VIII. RESULT AND DISCUSSION

Heart Disease prediction contains two separate prediction modules: heart disease prediction using structured clinical features and ECG disease classification using ECG images. Each module was evaluated using its own test data. The following results summarize the reported performance of the two models.

A. Heart Disease Prediction Results

The heart disease prediction module uses a Random Forest Classifier with 13 patient-related features. The reported

five-fold cross-validation accuracy was 82.63%, while the test accuracy was 91.80%. The test set contained 61 samples, of which 56 were classified correctly and 5 were classified incorrectly.

Metric	Class 0	Class 1	Macro Averag	Weighted Average
Precision	0.97	0.87	0.92	0.92
Recall	0.88	0.96	0.92	0.92
F1-score	0.92	0.92	0.92	0.92
Support	33	28	-	61

Confusion Matrix Interpretation

The reported confusion matrix shows that 26 samples were correctly classified as No Heart Disease and 27 samples were correctly classified as Heart Disease. Three samples without heart disease were incorrectly predicted as having heart disease, while five samples with heart disease were incorrectly predicted as No Heart Disease. These five false-negative predictions are important because a positive case may be missed. Therefore, accuracy should be considered

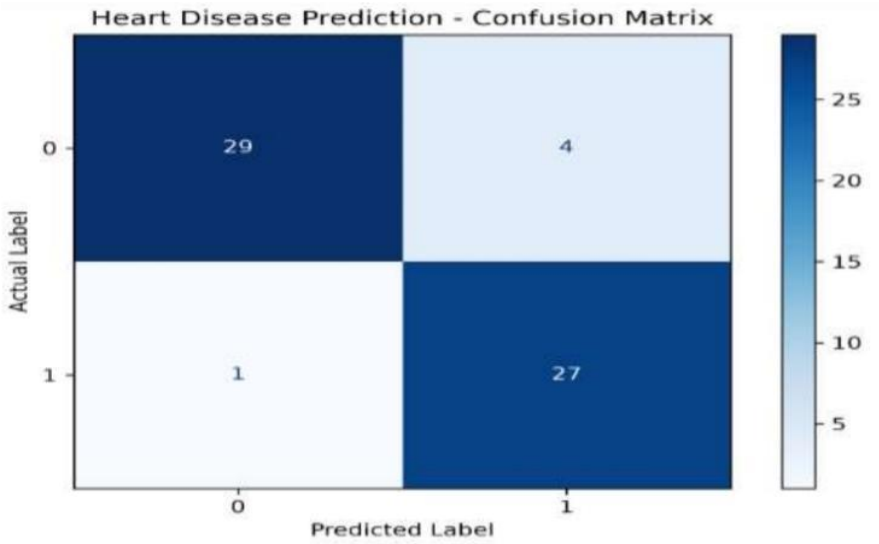
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058

together with sensitivity, specificity, precision, recall, and F1-score



Feature Importance

The reported Random Forest feature-importance ranking identified chest pain type (cp), maximum heart rate (thalach), ST depression (oldpeak), major vessels (ca), and age among the most influential features. Feature importance describes how the trained model used the input variables; it does not show that a feature directly causes heart disease.

B. ECG Disease Classification Results

The ECG image classification module categorizes images into four classes: Myocardial Infarction (MI), History of MI, Abnormal Heartbeat, and Normal ECG. The reported validation accuracy was 86.90%, and the reported test accuracy was 85.88%. These figures describe the ECG image model only.

The dataset includes four ECG image categories. Based on the provided class counts multiplied by 12, the derived counts are 2,880 images for Myocardial Infarction (MI), 2,064 images for History of MI, 2,796 images for Abnormal Heartbeat, and 3,408 images for Normal ECG, giving a derived total of 11,148 images. The classification report records 928 test samples. Since the original class counts add up to 929 before multiplication, the dataset counts should be verified against the actual folders and evaluation output before final submission.

ECG Class	Precision	Recall	F1-Score	Support
Myocardia Infarction(MI)	0.9755	0.8326	0.8984	239
History of MI	0.7339	0.9302	0.8205	172
Abnormal Heartbeat	0.9872	0.6609	0.7918	233
Normal ECG	0.8114	1.0000	0.8959	284
Accuracy	-	-	0.8588	928
Macro Average	0.8770	0.8560	0.8517	928
Weighted Average	0.8834	0.8588	0.8564	928

Confusion Matrix

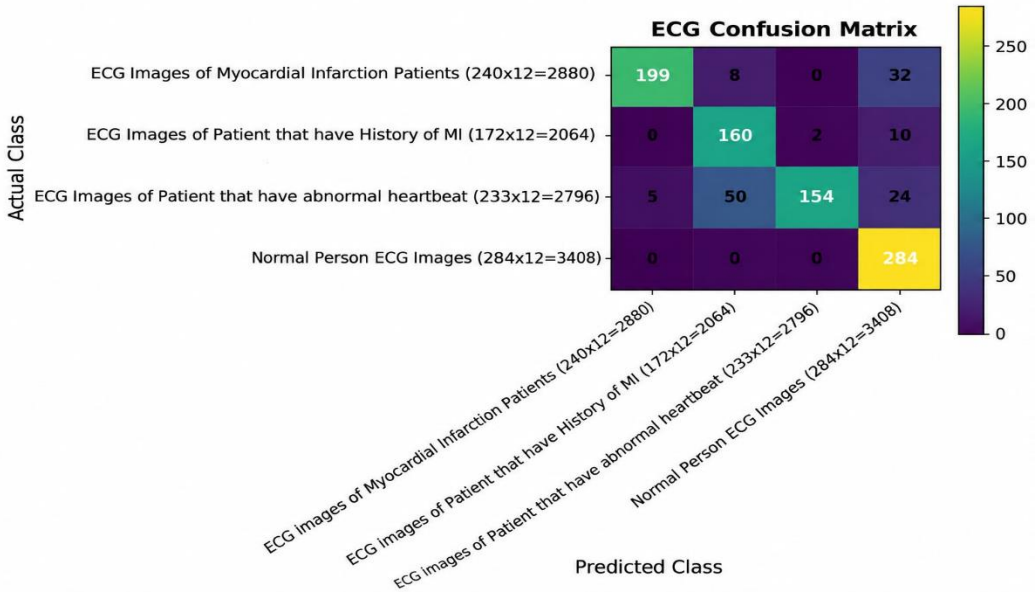
A confusion matrix compares the actual class of each test image with the class predicted by the model. Values on the main diagonal represent correct classifications, while off-diagonal values represent errors between classes. The exact ECG confusion-matrix values should be generated directly from the final evaluation code before they are included in the report. Values transcribed from a blurry image should not be treated as verified results. Class-wise precision, recall, F1-score, and sensitivity should also be reported where applicable.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058



Module	Model / evaluation	Reported result
Heart disease Prediction	Random Forest - test evalution	91.80% Test accuracy
Heart disease Prediction	5-fold cross-Validation	82.63% accuracy
ECG disease Classification	Convolution Neural Network(CNN)-Validation	86.90% Validation accuracy
ECG disease Classification	Convolution Neural Network(CNN)-Validation	85.88% test accuracy

C. Discussion

The results demonstrate that HeartCare AI integrates two machine-learning tasks in one application. The Random Forest model uses structured patient information to predict the heart disease target class, whereas the ECG image model classifies ECG images into one of four categories. The reported test accuracy was 91.80% for heart disease prediction and 85.88% for ECG disease classification. The five-fold cross-validation accuracy of 82.63% provides an additional estimate of the heart disease model's performance across different data splits.

Accuracy alone does not describe every type of model error. The heart disease confusion matrix includes five false negatives, showing why missed positive cases should be examined carefully. For ECG classification, the exact confusion matrix and class-wise metrics are needed to determine which categories are most often confused. Image quality, noisy or ambiguous samples, class imbalance, and differences between training and real-world data may affect predictions.

The two accuracy values should not be averaged or presented as one overall system accuracy because the modules use different inputs, output classes, and evaluation sets. The reported results should be interpreted as model evaluation results, not proof of clinical reliability.

IX. LIMITATIONS

However, there are a number of limitations in the current heart disease prediction system.

Firstly, machine learning algorithms learn from available heart disease datasets. These datasets may not represent all patient conditions, medical histories, and risk factors, which can affect the accuracy of predictions.

Secondly, the prediction process depends on patient data being collected and provided in a suitable format, such as CSV files or individual input fields. Therefore, the data must be properly prepared and processed before it can be used by machine learning algorithms.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151058

Thirdly, machine learning algorithms may produce incorrect predictions when patient data differs significantly from the data used during training. Missing values, noisy data, and variations in patient characteristics can also affect the prediction results.

Finally, the system has additional requirements for practical use, including secure handling of patient information, continuous data validation, efficient model execution, timely prediction results, and clear presentation of results through the user interface. The predictions should support medical assessment and should not replace professional diagnosis.

X. FUTURE SCOPE

The above-mentioned heart disease prediction system can be extended in several ways.

A. Real-Time Patient Data Collection and Prediction

The system may be enhanced to collect patient health data continuously from medical devices and wearable sensors, enabling predictions without requiring manual entry of patient details or uploading CSV files.

B. Hospital Management System Integration

The system can be integrated with hospital management systems and electronic health records so that relevant patient information, medical history, and test results can be used for heart disease risk prediction.

C. Automated Alerting System

An automated alerting mechanism *can* be developed to notify healthcare professionals when patient data indicates a potentially high risk of heart disease. Alerts may be delivered through email, mobile notifications, or the system dashboard.

D. Explainable AI

Explainable AI techniques can be incorporated to provide a clearer understanding of why the machine learning model predicts a particular level of heart disease risk and which patient features contribute to the prediction.

E. Continuous Model Training

The system can be enhanced to retrain machine learning models periodically using newly available and properly validated patient datasets. This may help maintain model performance as new data becomes available.

F. Advanced Heart Disease Prediction

Future development may focus on advanced machine learning and deep learning techniques to identify complex relationships among patient characteristics and improve the detection of potential heart disease risk.

G. . Multiple Dataset Integration

The system can be extended to combine datasets from different healthcare sources to improve the diversity of training data and evaluate model performance across different patient populations.

H. . Performance Monitoring and Evaluation

The system may include continuous monitoring of prediction performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. This can help identify when a model requires further evaluation or improvement.

I. Personalized Risk Assessment

Future versions may provide personalized risk assessments based on relevant patient characteristics, including age, blood pressure, cholesterol levels, and medical history, to support healthcare professionals in their decision-making.

J. Clinical Validation and Data Security

The system can be further developed through clinical validation and stronger patient-data protection mechanisms. Before real-world deployment, predictions must be evaluated with healthcare professionals and should support, rather than replace, medical diagnosis.

XI. CONCLUSION

In this study, a machine learning-based Heart Disease Prediction System is proposed for analyzing patient health data and predicting the possibility of heart disease. The proposed system combines data preprocessing, feature selection, machine learning classification, backend prediction services, and a web-based user interface.



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.151058

The heart disease prediction system using machine learning provides an effective approach for analyzing patient health data and predicting the possibility of heart disease. Based on the five research papers, different machine learning algorithms, including Random Forest, Logistic Regression, Support Vector Machine, XGBoost, and Gradient Boosting, have been explored for cardiovascular disease prediction. The studies highlight the importance of data preprocessing, feature selection, class imbalance handling, and model evaluation in improving prediction performance.

The proposed system aims to process patient information, apply a trained machine learning model, and display prediction results through a user-friendly interface. Performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC can be used to evaluate the model's effectiveness. The system can also be extended with real-time patient data collection, automated alerts, explainable AI, and integration with healthcare systems.

In conclusion, machine learning-based heart disease prediction has the potential to support early risk identification and assist healthcare professionals in making informed decisions. However, the reliability of predictions depends on the quality of the data and proper model validation. Therefore, further testing with diverse clinical datasets is necessary before the system can be used in real-world healthcare environments. The system is intended to support medical assessment and not replace professional diagnosis.

REFERENCES

[1]. B. McMahan et al., "Communication-Efficient Learning of Deep Networks From Decentralized Data", *Artificial Intelligence and Statistics Proc. PMLR*, vol. 10, no. 1, pp. 1273-82, 2017.
[2]. Afridi et al., Comparative Analysis of Machine Learning Algorithms for Predicting Heart Disease: A Comprehensive Study (2024).
[3]. C. En Guo, S.-C. Zhu and Y. N. Wu, "Primal Sketch: Integrating Structure and Texture", *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 5-19, 2007.
[4]. Bagadi, K. et al. (2023). Cardiovascular Disease Prediction Using Machine Learning Algorithms. IEEE ICETAS 2023.
[5]. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett*., vol. 20, no. 2, pp. 569–571,1999.
[6]. Hogade, N., Pasricha, S. and Siegel, H.J, "Energy and Network Aware Workload Management for Geographically Distributed Data Centers". *IEEE Transactions on Sustainable Computing*, vol.7, no. 2, pp.400–413. 2021
[7]. Shuge Ouyang, "Research of Heart Disease Prediction Based on Machine Learning," IEEE, 2022.
[8]. H. F. El-Sofany, "Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques," IEEE Access, 2024.
[9]. A. Wierman, Z. Liu, I. Liu and H. Mohsenian-Rad, "Opportunities and challenges for data center demand response", *Proc. Int. Green Comput. Conf.*, vol.7, no. 6, pp.1-10, 2014.
[10]. Ahmed et al. (2025), Predicting the Classification of Heart Failure Patients Using Optimized Machine Learning Algorithms. IEEE Access.
[11]. J. D. Jenkins et al., "The benefits of nuclear flexibility in power system operations with renewable energy", *Appl. Energy*, vol. 22 no. 2, pp. 872-884, 2018.
[12]. Haoying Dai, Yanne Kouomou Chembo, "RF Fingerprinting Based on Reservoir Computing Using Narrowband Optoelectronic Oscillators", *Journal of Lightwave Technology*, vol.40, no.21, pp.7060-7071, 2022.
[13]. Floris Van den Abeele, Jeroen Hoebeke, Girum Ketema Teklemariam, Ingrid Moerman, Piet Demeester, "Sensor Function Virtualization to Support Distributed Intelligence in the Internet of Things", *Wireless Personal Communications*, vol.81, no.4, pp.14-18, 2015.
[14]. J. Hwang, J. Kim and H. Choi, "A review of magnetic actuation systems and magnetically actuated guidewire-and catheter-based microrobots for vascular interventions", *Intell. Serv. Robot*., vol. 13, no. 1, pp. 1-14, 2020.
[15]. D. G. Feitelson, D. Tsafrir and D. Krakov, "Experience with using the parallel workloads archive", *J. Parallel Distrib. Comput*., vol. 74, no.3, pp. 2967-2982, 2014.
[16]. B. Accou, J. Vanthornhout, H. V. Hamme and T. Francart, "Decoding of the speech envelope from eeg using the vlaai deep neural network", *Scientific Reports*, vol. 13, no. 1, pp. 812, 2023.
[17]. Serim Lee, Nahyun Kim, Junhyoung Kwon, Gunhee Jang, "Identification of the Position of a Tethered Delivery Catheter to Retrieve an Untethered Magnetic Robot in a Vascular Environment", *Micromachines*, vol.14, no.4, pp.724, 2023.