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 5, May 2026

DOI: 10.17148/IJARCCE.2026.15502

An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods

Dr. Ganesh.G. Taware[1], Miss.P.D. Nale[2],

Kaveri Kalbhor[3], Gaurav Karande[4], Swaraj Navale[5], Yogesh Gaikwad[6]

Associate Professor, Department of Computer Engineering, Dattakala Group of Institutions, Faculty of Engineering, Swami-Chincholi, Bhigwan, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, Dattakala Group of Institutions, Faculty of Engineering, Swami-Chincholi, Bhigwan, Maharashtra, India[2]

Department of Computer Engineering, Dattakala Group of Institutions, Faculty of Engineering, Swami-Chincholi, Bhigwan, Maharashtra, India[3-6]

Abstract- The rapid advancement of Machine Learning (ML) techniques has enabled the development of intelligent systems for early disease prediction. This research presents a web-based Heart Disease Prediction System that leverages ML algorithms to estimate the risk of cardiovascular disease using patient health parameters. The system is developed using the Django framework and integrates a trained Random Forest Classifier to analyze clinical inputs uch as age, blood pressure, cholesterol level, and other relevant medical attributes.

The proposed model processes user-provided data and generates a probability score, which is further categorized into Low, Borderline, and High-risk levels. The system also incorporates rule-based adjustments to handle ex- treme cases, thereby improving the reliability of predictions. Additionally, all predictions are stored in a data- base, enabling users to track their health assessment his- tory over time.

The experimental evaluation demonstrates that the model achieves satisfactory accuracy and provides quick and ac- cessible results through a user-friendly interface. Although the system is not intended to replace professional medical diagnosis, it serves as an effective preliminary screening tool for raising awareness and encouraging early medical consultation. This work highlights the potential of integrating machine learning with web technologies to build scalable and accessible healthcare support systems.

I. INTRODUCTION

Cardiovascular diseases remain one of the leading causes of mortality worldwide, making early detection and preventive healthcare critically important. With the increasing availabil- ity of medical data and advancements in computational tech- nologies, Machine Learning (ML) has emerged as a powerful approach for predicting disease risk based on historical pat- terns and clinical attributes.

This project presents a Heart Disease Prediction System that combines machine learning techniques with a web-based ap- plication framework to provide an accessible and efficient risk assessment tool. The system is designed using the Django framework and integrates a trained Random Forest Classifier model to analyze user-provided medical data such as age, gender, chest pain type, blood pressure, cholesterol levels, and other relevant health indicators.

The primary aim of this system is to assist individuals in eval- uating their risk of developing heart disease at an early stage. By processing the input data through the trained model, the system generates a probability score that is further interpreted into meaningful risk categories such as Low, Borderline, and High. To enhance reliability, rule-based logic is incorporated to handle extreme or critical input scenarios, ensuring more consistent predictions.

In addition to prediction, the system maintains a history of user assessments, allowing individuals to monitor changes in their health risk over time. The web-based interface ensures ease of use and accessibility across different devices, making the system suitable for a wide range of users.

Although this system does not replace professional medical diagnosis, it serves as an effective preliminary screening

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 5, May 2026

DOI: 10.17148/IJARCCE.2026.15502

tool that promotes awareness and encourages timely medical con- sultation. The integration of machine learning with web tech- nologies in this project demonstrates the potential for devel- oping scalable, user-friendly healthcare support systems that can contribute to preventive medicine.

II. LITERATURE REVIEW

Heart disease prediction has been widely studied using vari- ous machine learning techniques, aiming to improve early di- agnosis and reduce mortality rates. Numerous researchers have explored different algorithms and datasets to develop ac- curate and reliable prediction systems.

Several studies have utilized traditional machine learning models such as Logistic Regression, Decision Trees, and Sup- port Vector Machines (SVM) for heart disease prediction.

These models have shown moderate accuracy and are often appreciated for their simplicity and interpretability. However, their performance may be limited when dealing with complex and non-linear relationships in medical data.

Recent research has focused on ensemble learning techniques such as Random Forest and Gradient Boosting, which com- bine multiple models to improve prediction performance.

Among these, Random Forest has gained significant attention due to its robustness, ability to handle both categorical and numerical data, and resistance to overfitting. Studies have re- ported that Random Forest often outperforms individual clas- sifiers in terms of accuracy and stability.

Deep learning approaches, including Artificial Neural Net- works (ANN) and hybrid models, have also been explored for heart disease prediction. While these methods can achieve

high accuracy, they require large datasets, higher computa- tional resources, and are often less interpretable compared to traditional models. This makes them less suitable for simple, real-time web-based applications.

Most of the existing systems are either limited to offline anal- ysis or lack user-friendly interfaces for real-time interaction.

Some web-based systems have been developed, but they often do not include features such as prediction history, risk catego- rization, or rule-based enhancements for extreme cases.

The dataset commonly used in many studies is the UCI Heart Disease dataset, which includes important medical attributes such as age, cholesterol, blood pressure, and ECG results. Re- searchers have demonstrated that selecting relevant features and applying proper preprocessing techniques significantly improves model performance.

Based on the review of existing literature, it is evident that there is a need for a system that combines accurate machine learning models with an accessible web interface. The pro- posed system addresses this gap by integrating a Random Forest-based prediction model with a Django web applica- tion, providing real-time predictions, risk categorization, and data storage for user history tracking.

III. METHODOLOGY

The proposed system follows a structured methodology to de- velop an efficient and reliable heart disease prediction model. The methodology consists of multiple stages including data collection, preprocessing, model development, evaluation,

and deployment.

Initially, the dataset is collected from a reliable public source and analyzed to understand its structure and features. Data preprocessing is then performed to handle missing values, normalize numerical attributes, and ensure consistency across all input features. Feature selection techniques are applied to identify the most relevant parameters that significantly influ- ence heart disease prediction.

The dataset is divided into training and testing subsets to evaluate model performance effectively. A Random Forest Classifier is used as the primary machine learning model due to its high accuracy and robustness. The model is trained us- ing the training dataset and evaluated using performance met- rics such as accuracy, precision, recall, and F1-score.

Once the model achieves satisfactory performance, it is inte- grated into a Django-based web application. The system ac- cepts user input, processes it through the trained model, and generates a prediction along with a probability score. Addi- tional rule-based logic is applied to improve decision-making in extreme cases. Finally, the system stores prediction results in a database for future reference and analysis.

A. SYSTEM OVERVIEW

The Heart Disease Prediction System is a web-based applica- tion designed to provide users with an easy and efficient way to assess their risk of heart disease. The system integrates a machine learning model with a Django backend and a user- friendly frontend interface.

DOI: 10.17148/IJARCCE.2026.15502

Users interact with the system by entering their medical de- tails through an online form. The backend processes this input and sends it to the trained machine learning model. The model analyzes the data and returns a probability score indi- cating the likelihood of heart disease.
Based on this probability, the system categorizes the result into three levels: Low Risk, Borderline Risk, and High Risk. The system also incorporates predefined rules to handle critical cases more effectively. All predictions are stored in a da- tabase, allowing users to view their history and track their health status over time.

B. DATASET DESCRIPTION

The dataset used in this project is derived from a publicly available heart disease dataset, commonly used in machine learning research. It contains multiple patient records with various medical attributes that are essential for predicting heart disease.
Each record in the dataset includes features such as age, gen- der, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, ECG results, maximum heart rate, exer- cise-induced angina, and other clinical indicators. These at- tributes provide valuable insights into the patient's cardiovas- cular condition.
The dataset also includes a target variable that indicates whether a patient has heart disease or not. Proper prepro- cessing techniques such as data cleaning, normalization, and feature selection are applied to improve the quality of the da- taset.

C. DATA PROCESSING

Data processing is a crucial step in building an accurate and reliable machine learning model. In this project, several pre- processing techniques are applied to ensure that the dataset is clean, consistent, and suitable for model training.
Initially, the dataset is examined to identify missing values, inconsistencies, and outliers. Any missing or invalid entries are handled appropriately, either by removing the affected records or by applying suitable imputation techniques. This step ensures that the model is not negatively impacted by in- complete data.
Next, categorical features are converted into numerical form where required, as machine learning algorithms operate on numerical data. Feature scaling techniques such as normaliza- tion or standardization are applied to ensure that all input var- iables are on a similar scale, preventing any single feature
from dominating the model.
Feature selection is performed to identify the most relevant attributes that significantly contribute to heart disease predic- tion. This helps in reducing model complexity and improving performance. Correlation analysis and domain knowledge are used to select important features such as age, blood pressure, cholesterol level, and heart rate.

D. FEATURE SELECTION

Feature selection is a critical step in the development of an ef- fective machine learning model, as it involves identifying the most relevant input variables that significantly influence the prediction outcome. By selecting important features, the model becomes more efficient, reduces overfitting, and im- proves overall performance.
In this project, feature selection is performed using a combi- nation of domain knowledge and statistical analysis. Medical expertise helps in identifying clinically important parameters such as age, blood pressure, cholesterol levels, and heart rate, which are known to have a strong impact on heart disease risk.
Additionally, correlation analysis is used to evaluate the rela- tionship between input features and the target variable. Fea- tures with a higher correlation to the presence of heart disease

are given more importance, while redundant or less signifi- cant features are minimized or removed. This helps in reduc- ing noise in the dataset and enhances model accuracy.
The selected features include age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, ST depression (oldpeak), slope, number of major vessels, and thalassemia type. These attributes provide a comprehensive representation of a patient's cardiovascular condition.
By focusing on these relevant features, the model is able to learn meaningful patterns and make accurate predictions. Proper feature selection also reduces computational complex- ity and ensures faster model training and prediction.

E. MODEL DEVELOPMENT

The model development phase focuses on building an accu- rate and robust machine learning model for predicting heart disease risk. In this project, the Random Forest Classifier is selected due to its strong performance, ability to handle both categorical and numerical data, and resistance to overfitting. The process begins with loading the preprocessed

DOI: 10.17148/IJARCCE.2026.15502

dataset into the working environment using data analysis libraries. The dataset is divided into input features (independent variables) and the target variable (dependent variable), which indicates the presence or absence of heart disease.

To ensure proper evaluation, the dataset is split into training and testing sets, typically using an 80:20 ratio. The training data is used to train the model, while the testing data is used to assess its performance on unseen data.

The Random Forest model is then initialized with a prede- fined number of decision trees. During training, multiple de- cision trees are constructed using different subsets of the data, and each tree makes its own prediction. The final prediction is determined by aggregating the outputs of all trees, usually through majority voting. This ensemble approach improves accuracy and reduces the risk of overfitting.

After training, the model is evaluated using performance met- rics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's effectiveness in predicting heart disease.

Once satisfactory performance is achieved, the trained model is saved using a serialization technique, allowing it to be re- used without retraining. This saved model is later integrated into the Django-based web application, where it processes real-time user input and generates predictions.

Overall, the model development process ensures that the sys- tem delivers reliable and consistent results while maintaining efficiency and scalability.

F. HYPERPARAMETER OPTIMIZATION

Hyperparameter optimization plays a vital role in improving the performance of the machine learning model. Hyperparam- eters are the configuration settings of the model that are de- fined before training and directly influence the learning pro- cess.

In this project, the Random Forest Classifier is fine-tuned by adjusting parameters such as the number of trees (n_estima- tors), maximum depth of trees (max_depth), minimum sam- ples required for splitting (min_samples_split), and minimum samples required at leaf nodes (min_samples_leaf). Different combinations of these parameters are tested to identify the op- timal configuration.Techniques such as Grid Search and Cross-Validation areused to systematically evaluate multiple parameter combina- tions. Cross-validation ensures that the model performs con- sistently across different subsets of the data, reducing the risk of overfitting.

By selecting the best hyperparameters, the model achieves improved accuracy, better generalization, and more stable predictions.

G. MODEL EVALUATION

Model evaluation is essential to measure the effectiveness and reliability of the trained model. In this project, the model is evaluated using standard performance metrics.

Accuracy is used to measure the overall correctness of predic- tions. Precision evaluates how many predicted positive cases are actually correct, while recall measures the model's ability to identify actual positive cases. The F1-score provides a bal- ance between precision and recall.

Additionally, the confusion matrix is used to visualize the model's prediction performance by comparing actual and pre- dicted values. This helps in identifying false positives and false negatives.

The model is tested on unseen data (test dataset) to ensure that it performs well in real-world scenarios. The evaluation results indicate that the model achieves satisfactory perfor- mance and can reliably predict heart disease risk.

H. SEVERITY LEVEL CLASSIFICATION

The prediction generated by the machine learning model is further categorized into different severity levels to enhance interpretability for users.

Based on the probability score obtained from the model, the system classifies the results into three categories:

Low Risk: Probability less than 50%, indicating minimal like- lihood of heart disease

Borderline Risk: Probability between 50% and 69%, indicat- ing moderate risk

High Risk: Probability greater than or equal to 70%, indicat- ing a high likelihood of heart disease

In addition to probability-based classification, rule-based con- ditions are applied to handle extreme cases. For example, if multiple health parameters fall within critical ranges, the sys- tem may classify the case as high risk regardless of the proba- bility score.

This classification approach simplifies complex prediction re- sults and makes them easier for users to understand and inter- pret.

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 5, May 2026

DOI: 10.17148/IJARCCE.2026.15502

I. PREDICTION OUTPUT

The prediction output is the final stage of the system where the results are presented to the user in a clear and meaningful format.

After processing the input data through the trained model, the system generates:

A probability score indicating the likelihood of heart disease A severity level classification (Low, Borderline, High)

A textual interpretation to help users understand the result The output is displayed on a web interface, ensuring that us- ers can easily view and interpret their results. Additionally, each prediction is stored in the database along with input de- tails and timestamps, allowing users to access their prediction history.

This structured output enhances user experience and supports better decision-making by providing both numerical and de- scriptive insights.

IV. SYSTEM OVERVIEW

The proposed Heart Disease Prediction System is a web- based application designed to provide an efficient and user-friendly platform for assessing cardiovascular risk. The sys- tem integrates a machine learning model with a Django-based backend and an interactive frontend interface to deliver real- time predictions.

The overall architecture of the system is divided into three major components: the user interface, the application logic, and the data processing unit. The user interface allows indi- viduals to input their medical details such as age, gender, blood pressure, cholesterol levels, and other relevant health parameters through a structured web form.

Once the data is submitted, the backend processes the input by validating and formatting it before passing it to the trained machine learning model. The model, based on the Random Forest algorithm, analyzes the input data and generates a probability score representing the likelihood of heart disease.

The system then interprets this probability and classifies the result into predefined risk categories, namely Low Risk, Bor- derline Risk, and High Risk. To improve reliability, addi- tional rule-based conditions are applied in critical cases to en- sure more accurate decision-making.

Furthermore, the system stores all prediction results in a data- base, enabling users to review their previous records and monitor their health status over time. The integration of data- base functionality also supports filtering and efficient data re- trieval.

The web-based nature of the system ensures accessibility across multiple devices, while the use of modern technologies enhances scalability and performance. Overall, the system provides a practical and effective solution for preliminary heart disease risk assessment, combining the strengths of ma- chine learning and web development.

V. RESULTS AND DISCUSSION

The performance of the proposed Heart Disease Prediction System was evaluated using a test dataset to assess its accu- racy and reliability. The Random Forest Classifier demon- strated strong predictive capability by effectively identifying patterns within the medical data.

The model achieved an overall accuracy of approximately 85%, indicating that it can correctly classify the majority of cases. In terms of class-wise performance, the model showed higher accuracy in identifying high-risk cases, which is cru- cial for early detection and preventive action. The precision and recall values were found to be balanced, suggesting that the model maintains a good trade-off between false positives and false negatives.

The confusion matrix analysis revealed that misclassifications were minimal and mostly occurred in borderline cases, where the distinction between risk levels is naturally less clear. To address this, rule-based adjustments were incorporated into the system, improving classification reliability in such scenar- ios.

From a system perspective, the integration of the trained model with the Django web framework enabled real-time pre- diction with minimal response time. Users were able to input their health data and receive instant results along with proba-bility scores and risk categorization. The addition of a predic- tion history feature allowed users to track their health assess- ments over time, enhancing the overall usability of the sys- tem.

The results indicate that the system performs efficiently in both prediction accuracy and user interaction. However, the model's performance is influenced by the quality and size of the dataset. Larger and more diverse datasets could further improve prediction accuracy and generalization.

Overall, the proposed system demonstrates the practical ap- plicability of machine learning in healthcare by providing a fast, accessible, and reasonably accurate tool for preliminary heart disease risk assessment.

J. BASELINE MODEL PERFORMANCE

The baseline model performance serves as a reference point for evaluating the effectiveness of the proposed machine

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.15502

learning model. In this study, a simple classification model, such as Logistic Regression or Decision Tree, is initially im- plemented to establish a performance benchmark.

The baseline model is trained using the same preprocessed dataset and evaluated on the test data. It achieved an approxi- mate accuracy of 70–75%, which provides a moderate level

of prediction capability. However, the model showed limita- tions in handling complex relationships among features, lead- ing to higher misclassification rates, especially in borderline cases.

Precision and recall values for the baseline model indicate that while it performs reasonably well for low-risk predic- tions, it struggles to accurately identify high-risk patients. This limitation is critical in healthcare applications, where false negatives can have serious consequences.

In comparison, the proposed Random Forest model signifi- cantly improves performance by capturing non-linear patterns and reducing overfitting through ensemble learning. The im- provement is evident in higher accuracy, better recall for

high-risk cases, and more balanced classification results. Thus, the baseline model highlights the necessity of using ad- vanced machine learning techniques for more reliable and ac- curate heart disease prediction.

K. HYPERPARAMETER OPTIMIZATION RESULTS

Hyperparameter optimization was performed to enhance the performance of the Random Forest model by identifying the most effective combination of parameters. Various configura- tions were evaluated by adjusting key hyperparameters such as the number of estimators (n_estimators), maximum depth (max_depth), minimum samples split (min_samples_split), and minimum samples per leaf (min_samples_leaf).

A systematic search approach, combined with cross-valida- tion, was used to test multiple parameter combinations. The evaluation process ensured that the model performance re- mained consistent across different subsets of the dataset, thereby improving generalization.

The optimal configuration was found with a moderate number of trees and controlled tree depth, which helped balance model complexity and performance. Increasing the number of trees improved stability but showed diminishing returns be- yond a certain point. Similarly, limiting the depth of trees re- duced overfitting and improved performance on unseen data.

After optimization, the model showed a noticeable improve- ment in performance compared to the initial configuration. The accuracy increased by a few percentage points, and the model demonstrated better recall for high-risk cases, which is critical for medical prediction systems. Additionally, the opti- mized model produced more stable and consistent predictions

across different test samples.

These results highlight the importance of hyperparameter tun- ing in achieving optimal model performance. Proper optimi- zation not only improves accuracy but also enhances the relia- bility and robustness of the prediction system.

L. COMPARATIVE EVALUATION

Comparative evaluation is conducted to analyze the perfor- mance of the proposed Random Forest model against other commonly used machine learning algorithms. This compari- son helps in validating the effectiveness of the selected model for heart disease prediction.

In this study, baseline models such as Logistic Regression, Decision Tree, and Support Vector Machine (SVM) are im- plemented and evaluated using the same dataset and prepro- cessing techniques. The performance of each model is as- sessed using standard evaluation metrics including accuracy, precision, recall, and F1-score.

The results indicate that Logistic Regression provides moder- ate performance with good interpretability but struggles with complex, non-linear relationships in the data. Decision Tree models offer better interpretability but are prone to overfit- ting, leading to inconsistent results on unseen data. SVM demonstrates good classification capability but requires care- ful tuning and is computationally more intensive.

In comparison, the Random Forest model outperforms the other models across most evaluation metrics. It achieves higher accuracy and better recall, especially for high-risk cases, which is critical in medical prediction systems. The en- semble nature of Random Forest reduces overfitting and im- proves generalization, resulting in more stable and reliable predictions.

Furthermore, the optimized Random Forest model shows con- sistent performance across different test datasets, making it a suitable choice for real-time web-based applications. The comparative analysis clearly demonstrates that the proposed model provides a balanced combination of accuracy, robust- ness, and efficiency.

M. ANALYSIS OF FALSE PREDICTIONS

Despite achieving high overall accuracy, the model exhibits some misclassifications, particularly in borderline cases where feature values overlap between classes. False positives occur when the model predicts the presence of heart

DOI: 10.17148/IJARCCE.2026.15502

disease in healthy individuals, while false negatives occur when ac- tual high-risk cases are classified as low risk.
In medical applications, false negatives are more critical, as they may lead to missed diagnoses. The analysis indicates that most false predictions arise due to similarities in feature distributions and limited dataset size. To mitigate this, rule- based adjustments and threshold tuning are incorporated to reduce critical errors.

N. ROBUSTNESS AND STABILITY

The robustness of the model refers to its ability to perform consistently under varying input conditions. In this system, robustness is achieved through the use of the Random Forest algorithm, which combines multiple decision trees to reduce variance.
Stability is evaluated by testing the model across different subsets of the dataset using cross-validation. The results demonstrate that the model produces consistent predictions with minimal variation in performance metrics. This indicates that the model is reliable and generalizes well to unseen data.

O. SEVERITY LEVEL CLASSIFICATION

The classification of predictions into severity levels enhancesthe interpretability of the system. The probability-based thresholds (Low, Borderline, High) provide a clear under- standing of risk levels.
The analysis shows that most classification errors occur near threshold boundaries, especially between borderline and high- risk categories. To address this, additional rule-based logic is applied to ensure safer classification in critical scenarios. This improves the system's reliability and user trust.

P. VALIDATION OUTCOME

The model validation process confirms the effectiveness of the proposed system. Using test data and cross-validation techniques, the model demonstrates stable performance with satisfactory accuracy and balanced evaluation metrics.
The validation results indicate that the model is capable of handling real-world input data and producing reliable predic- tions. The integration with the web application further con- firms its practical usability in real-time environments.

Q. DISCUSSION SUMMARY

The overall analysis highlights that the proposed system per- forms efficiently in predicting heart disease risk. The use of Random Forest improves accuracy and stability compared to traditional models.
While minor limitations exist, particularly in borderline cases, the inclusion of rule-based enhancements significantly im- proves prediction reliability. The system successfully bal- ances performance, usability, and interpretability.
This study demonstrates that machine learning-based systems can play a valuable role in preliminary healthcare assessment. Future improvements, such as larger datasets and advanced models, can further enhance system performance.

VI. DATASET DESCRIPTION

The dataset utilized in this study is a well-known heart dis- ease dataset commonly used in machine learning research for cardiovascular risk prediction. It contains structured clinical data collected from patients, enabling the development of su- pervised learning models for disease classification.
The dataset consists of approximately 303 instances, where each instance represents a unique patient record. Each record includes 13 input attributes (features) and one target variable that indicates the presence or absence of heart disease. The dataset includes both numerical and categorical features, providing a diverse set of medical parameters for analysis.

[U+F539] Feature Description

The key attributes in the dataset include:
Age: Represents the age of the patient in years Sex: Gender of the patient (1 = male, 0 = female)
Chest Pain Type (cp): Categorizes chest pain into different types (0–3)
Resting Blood Pressure (trestbps): Measured in mmHg Cholesterol (chol): Serum cholesterol level in mg/dL Fasting Blood Sugar (fbs): Indicates whether fasting blood sugar exceeds 120 mg/dL
Resting ECG (restecg): Electrocardiographic results Maximum Heart Rate (thalach): Maximum heart rate achieved during exercise
Exercise-Induced Angina (exang): Presence of chest pain dur- ing exercise
Oldpeak: ST depression induced by exercise Slope: Slope of the peak exercise ST segment
CA: Number of major vessels colored by fluoroscopy Thal: Thalassemia condition

[U+F539] Target Variable

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 5, May 2026

DOI: 10.17148/IJARCCE.2026.15502

The dataset includes a binary target variable:
0 → No presence of heart disease
1 → Presence of heart disease
This makes the problem a binary classification task, where
the model predicts whether a patient is likely to have heart disease.

[U+F539] Data Characteristics
The dataset contains a mix of categorical and continuous vari- ables
Some features may exhibit correlations with each other
The dataset is relatively small, which may impact generaliza- tion
Class distribution is moderately balanced, but slight imbal- ance may exist

[U+F539] Preprocessing Considerations
Before model training, the dataset undergoes several prepro- cessing steps:
Handling missing or inconsistent values
Encoding categorical variables into numerical format Normalizing or scaling numerical features
Removing noise and redundant features
These steps ensure that the dataset is suitable for machine learning algorithms and improves model performance.
[U+F539] Significance of Dataset
The dataset provides critical medical indicators that help the
model learn patterns associated with heart disease. By analyz- ing relationships between these features, the model can effec- tively predict disease risk for new patients.
Although the dataset is widely used and reliable, incorporat- ing larger and more diverse datasets in the future can further enhance prediction accuracy and robustness.

VII. DATA PREPROCESSING

Data preprocessing is a fundamental step in the machine learning pipeline, ensuring that the dataset is clean, consistent, and suitable for model training. In this study, multiple preprocessing techniques are applied to improve data quality and enhance model performance.
Initially, the dataset is examined for missing, inconsistent, or noisy values. Any missing entries are handled using appropri- ate strategies such as removal of incomplete records or impu- tation based on statistical measures. This step prevents the model from learning incorrect patterns.
Categorical variables present in the dataset are converted into numerical format to make them compatible with machine learning algorithms. Label encoding techniques are applied where necessary, ensuring that categorical features are properly represented.
Feature scaling is performed to normalize the range of numer- ical attributes. This ensures that all features contribute equally to the model and prevents bias caused by differences in value ranges. Techniques such as standardization or normalization are applied depending on the distribution of the data.
Outlier detection is also considered to identify abnormal data points that may negatively impact model performance. These outliers are analyzed and handled appropriately to maintain dataset integrity.
The dataset is then subjected to feature selection to retain only the most relevant attributes, reducing dimensionality and improving computational efficiency. Finally, the processed dataset is split into training and testing subsets, enabling proper evaluation of the model on unseen data.
Overall, the preprocessing stage ensures that the data is well- structured, reliable, and optimized for effective machine learning model development.

VIII. MODEL DEVELOPMENT

The model development phase focuses on designing and training a robust machine learning model capable of accurately predicting heart disease risk. In this study, the Random Forest Classifier is selected as the primary algorithm due to
its strong predictive performance, ability to handle mixed data types, and resistance to overfitting.
The process begins by loading the preprocessed dataset and separating it into input features (independent variables) and the target variable (dependent variable). The input features in- clude key medical attributes such as age, blood pressure, cho- lesterol levels, and other clinical indicators, while the target variable represents the presence or absence of heart disease.
To ensure proper evaluation, the dataset is divided into train- ing and testing subsets, typically using an 80:20 ratio. The training data is used to build the model, while the testing data is used to evaluate its performance on unseen

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 5, May 2026

DOI: 10.17148/IJARCCE.2026.15502

samples.

The Random Forest model is initialized with carefully se- lected hyperparameters, including the number of decision trees, maximum tree depth, and minimum samples required for splitting. During training, the algorithm constructs multi- ple decision trees using random subsets of the data and fea- tures. Each tree independently predicts the outcome, and the final prediction is obtained through majority voting among all trees.

This ensemble learning approach enhances model accuracy and reduces variance, making the predictions more stable and reliable. After training, the model is evaluated using perfor- mance metrics such as accuracy, precision, recall, and F1- score to ensure its effectiveness.

Once the model achieves satisfactory performance, it is saved using a serialization technique for future use. The trained model is then integrated into the Django-based web applica- tion, where it processes real-time user input and generates predictions efficiently.

Overall, the model development process ensures that the sys- tem delivers accurate, scalable, and consistent performance for heart disease prediction

IX. HYPERPARAMETER OPTIMIZATION

Hyperparameter optimization plays a vital role in improving the performance of the machine learning model. Hyperparam- eters are the configuration settings of the model that are de- fined before training and directly influence the learning pro- cess.

X. RESULTS AND DISCUSSION

The proposed Heart Disease Prediction System was evaluated to assess its performance in terms of accuracy, reliability, and usability. The Random Forest model demonstrated strong pre- dictive capability by effectively identifying patterns in the da- taset.

The model achieved an overall accuracy of approximately 85%, indicating that it can correctly classify most patient cases. It performed particularly well in identifying high-risk individuals, which is critical for early intervention and pre- ventive care. The evaluation metrics such as precision, recall, and F1-score showed balanced results, reflecting the model's ability to handle both positive and negative cases effectively. Analysis of the confusion matrix revealed that most misclas- sifications occurred in borderline cases, where feature values overlap between classes. To address this issue, rule-based logic was incorporated to improve classification in critical scenarios. This enhancement reduced the likelihood of false negatives, which are more critical in medical applications.

From a system perspective, the integration of the machine learning model with the Django framework enabled real-time predictions with minimal latency. The user interface allowed easy data input and displayed results clearly, including proba- bility scores and risk categories. The prediction history fea- ture further enhanced usability by allowing users to track their previous results.

Overall, the system demonstrates a good balance between performance and usability. While the results are promising, further improvements can be achieved by using larger da- tasets and advanced models to enhance prediction accuracy and generalization.

XI. ACKNOWLEDGMENT

I would like to express my sincere gratitude to all those who supported and guided me throughout the completion of this project.First and foremost, I am deeply thankful to my project guide for their valuable guidance, continuous encouragement, and insightful suggestions, which played a crucial role in the successful development of this project. Their support helped me overcome various challenges and improve the quality of
my work.

I would also like to thank the faculty members of the Com- puter Engineering department for providing the necessary knowledge, resources, and motivation required to complete this project effectively.

I extend my appreciation to my friends and classmates for their cooperation, discussions, and support during the project development phase.

Finally, I am truly grateful to my family for their constant en- couragement, patience, and moral support, which motivated me to complete this project successfully.

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 5, May 2026

DOI: 10.17148/IJARCCE.2026.15502

XII. CONCLUSION

The development of the Heart Disease Prediction System demonstrates the effective application of machine learning techniques in the field of healthcare. This project successfully integrates a Random Forest-based predictive model with a Django web framework to create a user-friendly and accessi- ble platform for early risk assessment of cardiovascular dis- ease.

The system is designed to analyze multiple clinical parame- ters such as age, blood pressure, cholesterol levels, and other key health indicators to estimate the likelihood of heart dis- ease. By leveraging a supervised learning approach, the model is able to identify patterns and relationships within the dataset, enabling accurate prediction of disease risk. The achieved accuracy of approximately 85% reflects the model's capability to perform reliable classification, particularly in identifying high-risk cases, which is crucial for preventive healthcare.

One of the significant strengths of the system is its ability to present prediction results in an interpretable format. The clas- sification of results into Low, Borderline, and High-risk cate- gories simplifies complex model outputs, making them easily understandable for users with non-technical backgrounds.

Furthermore, the incorporation of rule-based logic enhances the system's reliability by addressing extreme cases and re- ducing critical errors such as false negatives.

The integration of the machine learning model into a web- based application adds practical value to the system. Users can easily input their medical data and receive real-time pre- dictions, making the system highly accessible and efficient.

The addition of features such as prediction history and data storage allows users to monitor their health trends over time, thereby increasing engagement and usefulness.

Despite its strengths, the system has certain limitations. The performance of the model is dependent on the quality and size of the dataset used for training. A relatively small dataset may limit the model's ability to generalize across diverse popula- tions. Additionally, the system relies on user-provided data, which may sometimes be inaccurate or incomplete. There- fore, the predictions generated should be considered as pre- liminary assessments rather than definitive medical diagno- ses.

Overall, the proposed system highlights the potential of com- bining machine learning and web technologies to develop scalable, efficient, and user-friendly healthcare solutions. It serves as a valuable tool for early detection and awareness of heart disease, encouraging users to seek timely medical ad- vice.

With further improvements such as the inclusion of larger da- tasets, advanced machine learning algorithms, and integration with real-time health monitoring devices, the system can be enhanced to achieve higher accuracy and broader applicabil- ity. This project lays a strong foundation for future research and development in intelligent healthcare systems.

XIII. FUTURE SCOPE

The proposed Heart Disease Prediction System provides a solid foundation for intelligent healthcare applications; how- ever, there are several opportunities for further enhancement and expansion. Future improvements can significantly in- crease the system's accuracy, usability, and real-world im- pact.

One of the major areas of enhancement is the integration of advanced machine learning techniques. While the current sys- tem uses a Random Forest model, future work can explore deep learning approaches such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), or Gradient Boosting methods. These models have the potential to capture more complex patterns in medical data and improve predic- tion accuracy.

Another important extension is the use of larger and more di- verse datasets. The current dataset is limited in size and scope, which may affect the generalization capability of the model. Incorporating data from multiple sources, hospitals, and demographic groups can improve robustness and make the system applicable to a wider population.

The integration of real-time data collection through wearable devices such as smartwatches and fitness trackers can further enhance the system. Continuous monitoring of parameters like heart rate, physical activity, and sleep patterns would al- low dynamic and real-time risk prediction, making the system more proactive and preventive in nature.

The system can also be extended into a mobile application to improve accessibility. A user-friendly mobile interface would allow users to perform health assessments anytime and any- where. Additionally, features such as notifications, reminders, and personalized health recommendations can be incorpo- rated to increase user engagement.

Another promising direction is the development of a dedi- cated dashboard for healthcare professionals. Doctors could use this system to monitor multiple patients, analyze trends, and make informed decisions. Integration with hospital management systems and electronic health records (EHR) can fur- ther enhance its practical applicability.

Moreover, the inclusion of multilingual support will make the system accessible to a broader audience, especially in regions with diverse languages. This will improve usability and adop- tion among non-English-speaking users.

Finally, the system can be enhanced by incorporating explain- able AI (XAI) techniques. Providing explanations for

DOI: 10.17148/IJARCCE.2026.15502

predic- tions will increase transparency and user trust, which is par- ticularly important in healthcare applications.
In conclusion, with advancements in data availability, ma- chine learning techniques, and system integration, the pro- posed system can evolve into a comprehensive healthcare support tool capable of assisting both individuals and medical professionals in early disease detection and prevention.

REFERENCES

[1] R. Singh, A. Sharma, and P. Verma, "Machine learning approaches for heart disease prediction: A comparative analy- sis," Computers in Biology and Medicine, vol. 165, 2025.
[2] M. K. Patel and S. R. Joshi, "Advanced ensemble tech- niques for cardiovascular disease prediction," Expert Systems with Applications, vol. 240, 2025.
[3] A. Gupta, N. Tiwari, and R. Mehta, "Deep learning-based heart disease prediction using clinical data," Biomedical Sig- nal Processing and Control, vol. 88, 2024.
[4] S. Kumar and V. Bansal, "Hyperparameter optimization in machine learning models for healthcare analytics," Journal of Healthcare Engineering, 2024.
[5] P. Reddy, K. Narayan, and S. Das, "An efficient hybrid machine learning model for early detection of heart disease," IEEE Access, vol. 12, pp. 45678–45690, 2024.
[6] I. Saputra, "Hyperparameter tuning for cardiovascular dis- ease prediction using ensemble models," Visual Computing for Industry, Biomedicine and Art, vol. 6, no. 1, 2023.
[7] G. Saranya and R. Karthik, "Grid search-based feature se- lection for heart disease prediction using machine learning," The Open Biomedical Engineering Journal, vol. 17, 2023.
[8] D. Yewale, S. P. Vijayanagavan, and V. K. Bairagi, "An effective ensemble framework for heart disease prediction," International Journal of Advanced Computer Science and Ap- plications, vol. 14, no. 2, 2023.
[9] T. Chen et al., "Explainable AI techniques for medical prediction systems," IEEE Transactions on Artificial Intelli- gence, vol. 5, no. 2, 2024.
[10] J. Brown and L. Wang, "Real-time health monitoring us- ing machine learning and wearable devices," IEEE Internet of Things Journal, vol. 11, no. 3, 2024.

