Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Enhancing Cardiovascular Disease Diagnosis through Data-Driven Feature Analysis and Cross-Validated Machine Learning Models

Abhilash Butola
M.Tech Computer Science and Engineering
Department of Computer Science and Engineering

Abstract **- Cardiovascular diseases are a major global health problem, accounting for 17.9 million deaths per year and constituting 32 percent globally. According to the World Health Organization, the disease in people is due to an unhealthy diet,such as the intake of more junk food along with alcohol and smoking consumption; high cholesterol levels; and a reduction in people 's lives. And the diagnostic techniques that are used ,like ECG, echocardiography, angiography, and various laboratroy test and are very costly and time-consuming. It will be in limited in rural areas, so it will be as the limited no of infrastructure, and their will be a delay of detection of disease and the increases of chances of heart attack and strokes. In this study, we used machine learning to produce digital records,predict heart disease, reduce human errors, and improve decision-making. The methodology used the data set from Excel and Panda, using missing values and conducting the analyses'visualization of gender-based frequency and tabulating the variables chest pain and diagnosis and evaluating the scatter plots and heart risk, which will be associated with a risk pattern. Now, we can analyze with a correlation matrix, manage redundancy, and assist in feature selection. The development of the model is now divided into training and testing sets, 60 percent and 70 percent, and includes and now compares the logistic regression, K-Nearest Neighbors, and Random Forest models using F1-score and ROC-AUC, and we have a cross-validation matrix. And to achieve cross-validation and hyperparameters, such as GridSearchCV or RandomizedSearchCV, which will be included for optimization and will be a combination of ML variables, such as blood pressure, cholesterol levels, chest pain,and blood sugar level, and also to enhance the accuracy and improve the interpretability through features and for CVD and better outcomes and also the limitations on it.**
Keywords: Cardiovascular diseases, Machine learning, Heart diseases, Prediction accuracy, Ensemble learning, Artificial Neural Networks, Decision Tree.

I. INTRODUCTION

Cardiovascular diseases are known as heart diseases. It was a very high, wide world where many people suffered from a lot of diseases. The World Health Organization (WHO) states that cardiovascular diseases are responsible for 17.9 million deaths in every year, which accounts for 32percent of global fatalities. Heart diseases include heart failure and congenital heart diseases. The increase in heart diseases is spreading worldwide, in which people are now eating wrong food, such as unhealthy junk food and oily food, and also drinking large amounts of alcohol. The side effects in the body will be diabetes and cholesterol levels now increasing, so it will be critical for the health of people, and the life of the patient is not good [1], [2]. Heart disease diagnosis is based on a combination of various clinical methods, such as electrocardiography (ECG), echocardiography, angiography, and laboratory tests. All methods are used to detect heart diseases... In rural areas, healthcare is in demand, and hospitals and advanced techniques are not available to detect heart disease. However, access to advanced diagnostic methods is limited in rural settings. In which the people are in now early stages; they do not know inside the body that symptoms are will increase, but they were not noticeable in the early stage. Due to the noticeable severity, several symptoms can experience a heart attack or stroke. the development of advanced machines in medical health care to find and predict heart diseases before symptoms of heart disease or stroke [3], [4]. In the new era, medical digital records are increasing rapidly, and there will be a large amount of labeled data. The medical field is now combining artificial intelligence and machine learning in heart diseases to predict the risk. In addition, the hidden patterns and relationships among the variables were determined. These algorithms are now used to predict the historical data and to develop the prediction models with high precision. MI-based decision models to reduce human error, which will be helpful for people to provide evidence for the risk of heart diseases [5, 6]. Machine algorithms are now predicting heart diseases, including: various algorithms, such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), K Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Gradient Boosting, Extreme Gradient Boosting (XGBoost), and various techniques. The algorithms, such as ensemble learning, basically predicts the various classifiers and reduce overfitting. , They will use a machine learning algorithm that is affected by the quality of the dataset; thus, when we choose the model, it will be challenging in research [7], [8]. In medicine, the advanced intelligent system is machine learning or AI. The current research has several various limitations. They have a small dataset, and we have a limited number of algorithms. We are currently focusing on improving the prediction accuracy of the existing models. The combination of AI and medical diagnosis is highly accurate in real- world medical applications [10], [11]. The challenges in various machine learning algorithms can be used, and it could be used as a framework for heart disease and also to analyze the data analysis feature analysis and combination of various machine learning algorithms. And we have a classifier to check the performance of the metrics precision, Recall,F1-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Score, ROC-AUC, and confusion matrices, and it should be used to improve the outcomes of patients, whether they have heart disease or not [7], [20].

II. PROBLEM STATEMENT

number off people in world it will have a financial impact on the healthcare system. Modern advanced techniques such as ECG, angiography (MRA), and other tests are costly and time-consuming in the medical field. In rural areas, hospitals have limited advanced tools, which delay heart disease diagnosis and patient outcomes [1], [2]. The traditional medical diagnosis has various parameters, such as age, blood pressure, cholesterol level, and pain in the chest. The manual diagnosis has the hidden relationships among patients. It will be erroneous and inconsistent [3], [4]. The amount of electronic health data is now increasing, increasing the large amount of data on it, and these amounts of data are not analyzed using the traditional statistical methods. All the various algorithms are now used to predict the heart diseases. They face challenges such as inappropriate processing, invalid improper feature selection, and imbalance. It is also used to integrate practice and advanced and improve its accuracy of it [5], [6]. The intelligent machine learning framework is now used to predict heart disease and collect clinical methods, and it can also be used to identify patterns, and it could be now decrease cardiovascular heart disease [7].

III. ROLE OF MACHINE LEARNING IN HEART DISEASE PREDICATION

Machine learning tools can enhance the diagnosis of diseases and assist in decision-making in the medical healthcare field. Machine learning is used to analyze patient data and high-dimensional data sets to identify the hidden patterns in heart diseases [5], [6]. Various machine learning algorithms, such as linear regression and random forest, have been used to predict heart diseases. Each algorithm predicts the accuracy of the model, which is used for clinical methods [3], [5]. Machine learning has several advantages, such as the availability of large amounts of data, which reduces time and allows easy analysis and enhances the decrease in human interpretation, will, and advanced features, which improve the accuracy and reduce overfitting [10], [12].

IV. RESEARCH GAP

Various studies on the application of machine learning techniques in cardiovascular diseases have shown that it will be an issue in small data sets across diverse patients, as there will be a limited number of machine learning algorithms to identify the classifier in medical diagnoses [5], [10]. It will be as now class imbalances, missing values after deployment, and reduced classification performance in many studies to predict accuracy in other critical evaluations, such as Ecall, F1-score, and ROC-A in model performance [6], [11]. Deep or ensemble learning has high prediction accuracy and operates in a black- box system. It can be incorporated now to incorporated in the pre-process in many combinations of machine learning algorithms to identify heart disease for real-world clinical practice [7], [12].

V. METHODOLOGY

VI. MACHINE LEARNING FRAMEWORK FOR HEART DISEASE PREDICTION

The prediction of heart diseases by combining AI and machine learning enhances the models, evaluates the patient data record, and determines the information of heart disease conditions. This approach addresses various problems in heart diseases, such as blood pressure, cholesterol levels, type of chest pain, and increased the blood sugar levels. And the ML algorithm is identifies the patient data by classifying whether they have a heart disease. Due to the early people, now take decisions to check whether the people will reduce the errors and improve the patient's life. This provides accurate efficiency and improves medical healthcare.

VII. DATASET ACQUISITION AND PREPARATION GENDER-BASED DISTRIBUTION OF HEART DISEASE

We used the publicly available "Heart Disease Prediction" dataset (Kaggle);and therefore, the heart disease dataset can now be imported through Python libraries such as pandas. It is a tool for handling the unlabeled data, after which we can import data and analyze that we have missing value data types, and now they will be used to analyze the data and predict the data and analysis and predict the model. Heart disease frequency according to gender In the data, we can now analyze whether the disease differs between males and females. Therefore, we can analyze heart disease in each sex/gender group (male or female) through distribution by charts or statistical data. Many people are using visualization techniques to analyze the data of heart disease between males and females, and it could be used in medical diagnoses that easily analyze the data and should be as it used in health recommendations in patient data.

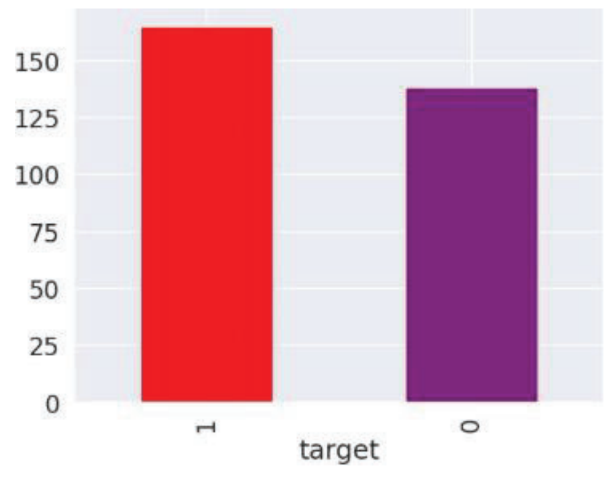


Figure 1.As shown in the dataset contains a higher proportion of male patients than female patients

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

VIII. CROSS-TABULATION ANALYSI

A cross-tabulation visualization technique will be used to show the relationship between categorical variables, which could be in frequency distribution or graphical format. The cross-tabulation was used to compare heart disease or pain in the chest with the diagnosis. Therefore, graphical tools were used to enhance the visualization between the male and female heart disease patients, and cross- tab analysis was used to explore the disease prediction.

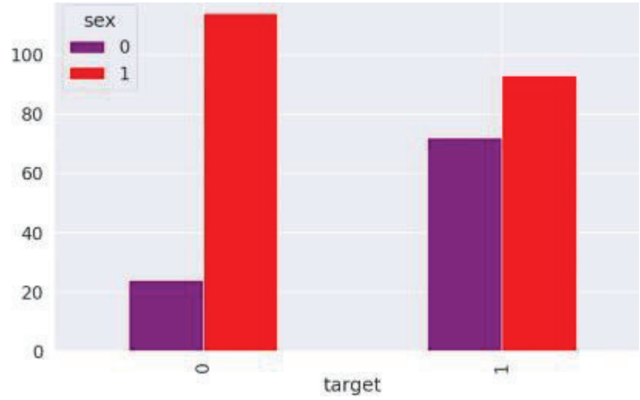


Figure 2 Presents the frequency of heart disease among male and female patients

IX. AGE AND MAXIMUM HEART RATE RELATIONSHIP ANALYSIS

The term "Age vs Max Heart Rate" has a relationship between age and peak heart rate in patients with heart disease. We have a scatter plot in which whether have heart disease is present or absent without heart disease as age increases and the heart rate is now decreased; this typical pattern shows that they will be at a risk of cardiovascular disease. In the future, we can compare the age or maximum heart rate for prediction, which will also enhance the performance, and now they will be as evaluated on the models. Heart disease frequency per chest pain type Chest pain is another clinical condition in which we examine the patient for heart disease symptoms; they will be chest pain-like, angina or non-angina, and it would be, as a condition, a bar chart now to compare the heart disease from the categories, and we also analyze the chest pain associated with the heart disease, and we enhanced machine learning models for diagnosis in cardio vascular disease.

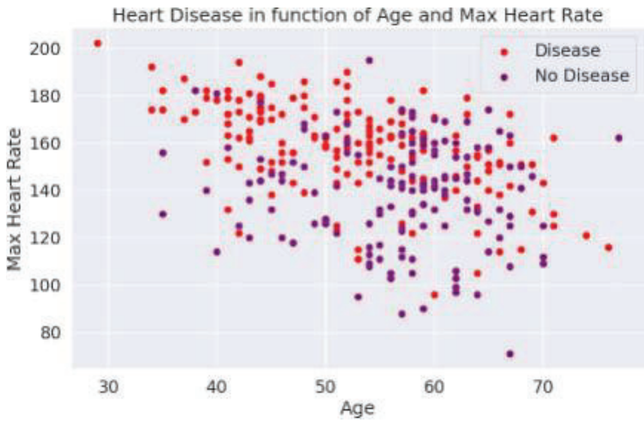


Figure 3 shows the relationship between age and maximum heart rate

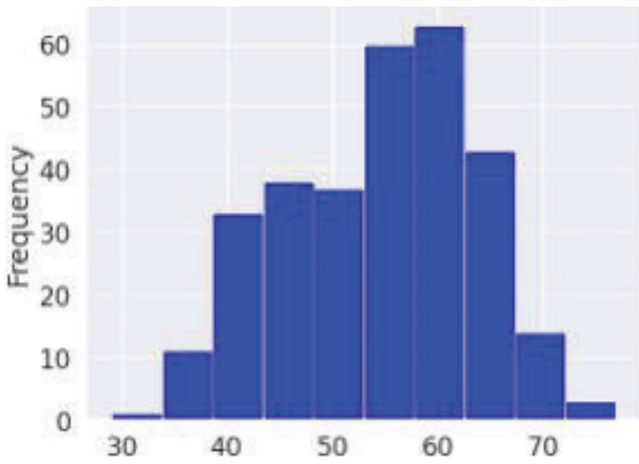


Figure 4 illustrates the normal distribution of the selected features

X. FEATURE CORRELATION ANALYSIS

Correlation analysis examines the relationships between independent variables in a dataset. A correlation matrix is typically employed to determine the relationship between -1 and +1; thus, features have a high correlation redundancy, in which the model performance is used to select the relevant features to improve the model performance to provide valuable insights into the data set underlying the structure and development of machine learning models.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

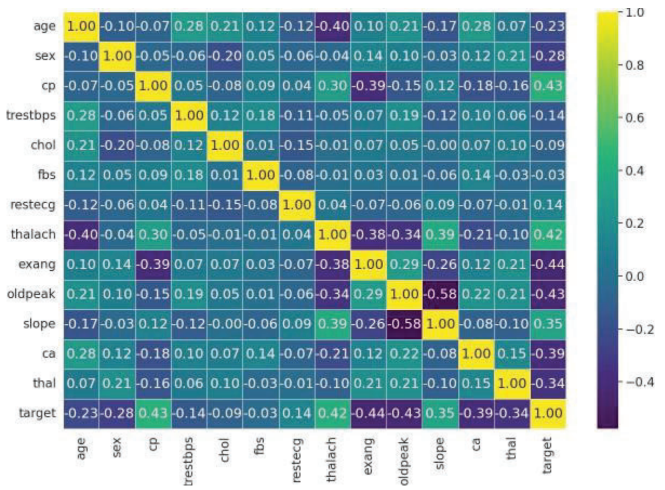


Figure 5.As shown in Fig. 5, the correlation matrix identifies relationships among independent variables

XI. PREDICTIVE MODEL DEVELOPMENT

Modeling is now used to train the machine learning models to identify the patterns of heart disease. In this phase, we have a classification algorithm connection between the disease status and the patient to develop a capable model of the patient's new record accurately. There are various algorithms to determine pre-processing the data to accurately relay it for health care applications.

XII. DATASET PARTITIONING STRATEGY

The model performance in which the dataset is divided into training and testing 60 Percent and 70 percent of the data is now collected from training the model to learn from the historical data pattern, generalize to new substances, and overfitting to predictive and estimate of predictive performance, a well-organized way to predictive the performance A well-balanced data split is essential for developing an ML model.

XIII. SELECTION OF CLASSIFICATION ALGORITHMS

We are now choosing a model for predicting heart disease, including logistic regression and K-nearest neighbors (KNN). Each algorithm is now used to improve the accuracy and efficiency, and it will be advantageous to the accuracy and robustness to determine the best algorithm for capturing the dataset. The model is selecting the captured dataset, underlining and evaluating the metrics of F1-Score, ROCAUC, precision, and recall

XIV. COMPARATIVE PERFORMANCE EVALUATION

In the model comparison, we used logistic regression, K-nearest neighbors (KNN), and random forest comparison with accuracy, F1 - score, ROC-AUC and compared the algorithms that achieved the accuracy and generalization. And Charts, performance tables, and comparisons of the selected model were used to predict the outcome and support its application in the decision system.

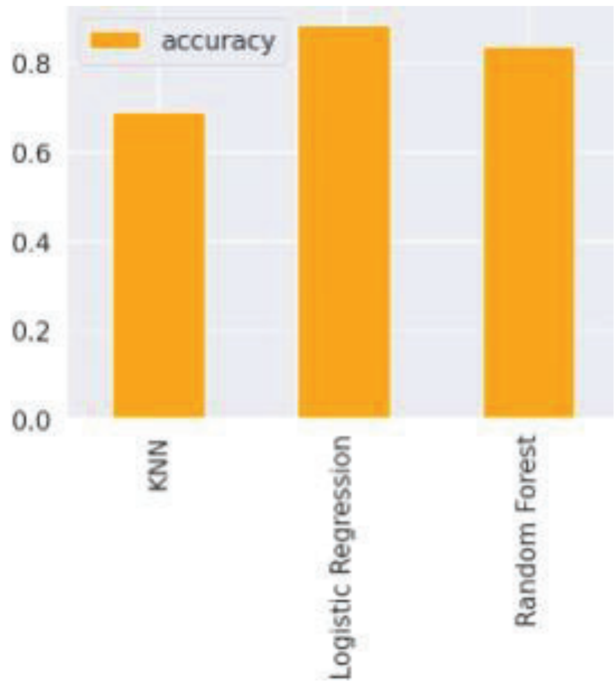


Figure 6. compares the performance of the selected machine learning models

XV. MODEL OPTIMIZATION AND CROSS-VALIDATION

Hyperparameter tuning enhances the algorithm and maximizes the predictive performance. Cross-validation divides the data into multiple parts to estimate the model performance, and the techniques used are GridSearchCV or RandomizedSearchCV, which improve the classification accuracy and make the in-heart-disease prediction model more reliable.

XVI. KNN HYPERPARAMETER OPTIMIZATION

KNN is a distance-based algorithm based on distance with the choice of numbers, and hyperparameter tuning will also identify the k and values of testing in different configurations by cross-validation. Another distance parameter, distance, can be optimized easily, and effective tuning reduces the error and predicts the accuracy. kNN models classify patients and are suitable for heart disease prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 7 presents the KNN hyperparameter tuning results

XVII. RANDOMIZED HYPERPARAMETER SEARCH

RandomizedSearchCV is a technique in which we identify the parameter and combination that must be selected from the predefined search space. An exhaustive method was used. We limited the configuration and have reduced the computational requirements. The selected configuration in cross- validation assesses the model performance and now sets the parameters for machine learning for the model final model training. This method was used to achieve results in a shorter amount of time.

XVIII. GRID-BASED HYPERPARAMETER OPTIMIZATION

GridSearchCV in hyperparameter optimization was used to determine and find every possible combination of parameter values. A configuration using cross vladiation to predict and accurately assess the performance. GridSearchCV also identifies the parameters for the machine learning model. GridSearchCV was is used to enhance the model accuracy and robustness. In this method, a grid search for heart disease prediction is performed to enhance the classification problems of overfitting.

XIX. RECEIVER OPERATING CHARACTERISTIC (ROC) AND AUC EVALUATION

The ROC is a classification model that plots the true positive rate against the false positive rate at the thresholds. The AUC is now the model for all the abilities into a single value. A higher AUC score with high performance and a score of 1 represents prediction. ROC was used for AUC analysis for the model and comparing the different classifiers

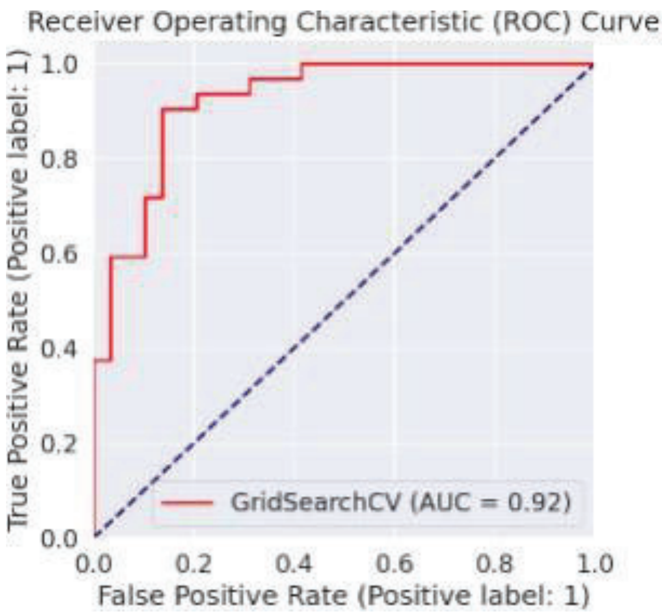


Figure 8. shows the ROC curve used to evaluate classifier performance

XX. CONFUSION MATRIX ANALYSIS

The confusion matrix is all about the performance of algorithms, predicate labels, and all actual outcomes. They comprise four components: true positive, true negative, false positive, and false negative. These are the matrices that offer the classification error models, which are made such as accuracy, precision, recall, and specificity, and also to analyze the matrix and model strength and improvement in predictive model performance on it

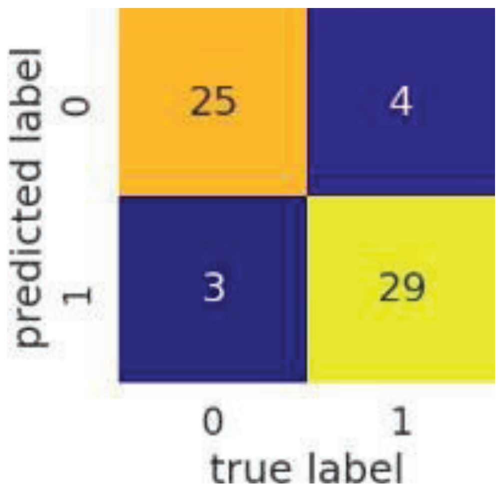


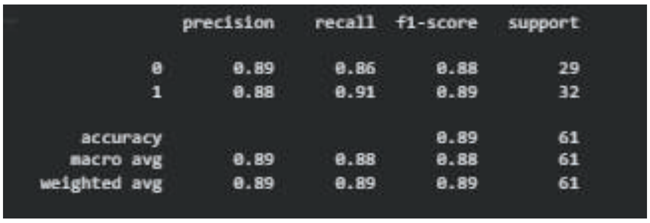
Figure 9. presents the confusion matrix of the best-performing model

XXI. CLASSIFICATION PERFORMANCE METRICS

The analysis of model performance was analyzed in terms of precision, recall, F1-score, and support for each category. Precision assessment and predication to identify the true positive instances of the F1 score could be between these

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

two metrics and denote the size of a category. Now the overall accuracy is particularly advantageous for data setswith imbalanced classes by comparing models and requiring further refinement.



	precision	recall	f1-score	support
0	0.89	0.86	0.88	29
1	0.88	0.91	0.89	32
accuracy			0.89	61
macro avg	0.89	0.88	0.88	61
weighted avg	0.89	0.89	0.89	61

Figure 10 illustrates the classification report

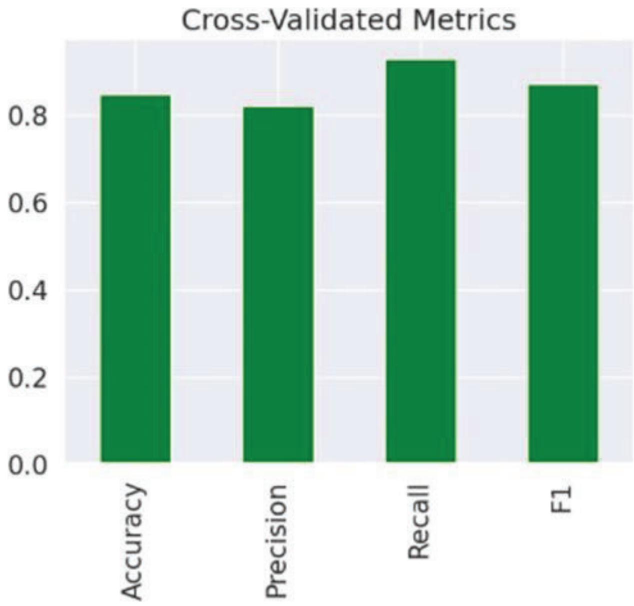


Figure 11. shows the cross-validation performance metrics

XXII. FEATURE IMPORTANCE AND MODEL INTERPRETABILITY

In the feature, the input variable was predicted by a machine learning model, and in the data, the factors such as clinical factors, age, cholesterol levels, blood pressure, chest pain, heart rate, and impact of heart disease prediction were defined. This feature also enhances the model's performance in clinical decision-making. It also enables research in less significant areas and improves the model efficiency and predictive accuracy. The analysis provides the risk factors for cardiovascular disease.

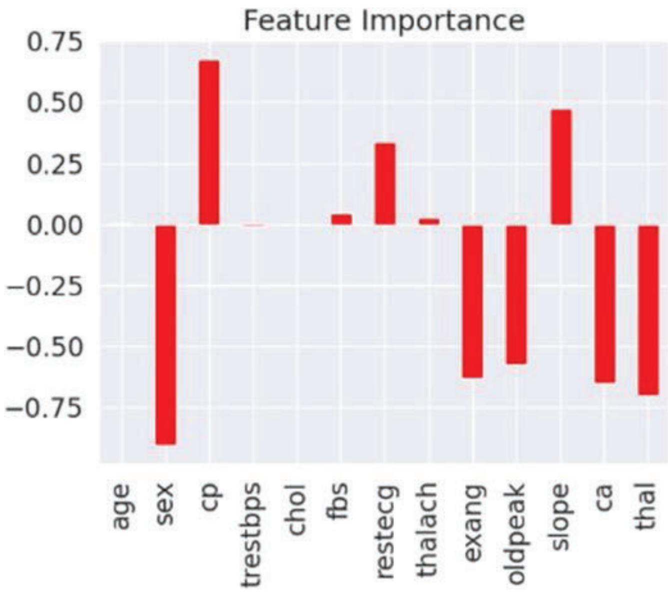


Figure 12. presents the feature importance scores of the predictive model

XXIII. DISCUSSION

The machine learning is a tool to find the accuracy of heart disease and is an effective tool and diagnosis of heart disease and has a unique strength and selection for the model, and the random forest also provides prediction accuracy and traditional classifiers, and the model performance is now strongly influenced by data quality and feature selection and also class balance and also improves the reliability in a prediction models and hyperparameter tuning and cross-validation and also improves the reliability of a model, and also the challenge remains, including a limited data set and real- world deployment and machine learning models in healthcare to achieve more accurate and trustworthy heart disease predictions

XXIV. CONCLUSION

In machine learning, methodologies are used to predict heart disease and to improve the detection of cardiovascular diseases. In it, we use logistic regression, which would be the baseline performance and is straight. KNNK proves that the smaller data set and its performance are highly dependent on the choice of k. Random Forest provides better predictive accuracy, and it will be used as an ensemble learning technique to reduce overfitting. The application of GridSearchCV or RandomizedSearchCV and the hyperparameter also will now improve the model performance and accuracy, which is high in the metrics such as precision, Recall,F1- Score,ROCAUC, and confusion matrix. Machine learning models are both effective in predicting cardiovascular or heart disease, and feature is analysis identifies age, cholesterol level, blood pressure, and maximum heart rate as the most significant clinical factors in disease prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

REFERENCES

[1] World Health Organization, Cardiovascular Diseases (CVDs) Fact Sheet, Geneva, Switzerland, 2023.
[2] World Heart Federation, World Heart Report 2023: Confronting the World's Number One Killer, Geneva, Switzerland, 2023.
[3] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.
[4] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989.
[5] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," IEEE Conference on Information and Communication Technologies, 2013.
[6] M. Akhil Jabbar, B. L. Deekshatulu, and P. Chandra, "Heart Disease Prediction System Using Associative Classification and Genetic Algorithm," International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2015.
[7] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart disease using cardiovascular datasets," BMC Medical Informatics and Decision Making, vol. 20, no. 251, 2020.
[8] American Heart Association, Heart Disease and Stroke Statistics—2023 Update, Dallas, TX, USA, 2023.
[9] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009
[10] I. Kononenko, "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective," Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89–109, 2001.
[11] . Esteva et al., "A Guide to Deep Learning in Healthcare," Nature Medicine, vol. 25, pp. 24–29, 2019.
[12] M. A. Khan, M. Sharif, T. Akram, et al., "An interpretable machine learning framework for heart disease prediction using feature optimization and ensemble learning," IEEE Access, vol. 9, pp. 140363–140375, 2021.
[13] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[14] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.
[15] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
[16] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[17] J. G. Cleland et al., "Heart Failure: Clinical Review," The Lancet, vol. 365, pp. 1877–1889, 2005.
[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2021.
[19] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[20] Bourke,D.Zero to Mastery Machine Learning Repository. GitHub Repository. Available online: https:Dataset(https://www.kaggle.com/ronitf/heart-disease-uci/ (accessed on 25 July 2026)
[21] D. M. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011