International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction

Guide: Prof. Divya Bharathi

School of Engineering

Malla Reddy University

S Vinay Kumar Reddy
School of Engineering(AIML)
B Tech
Malla Reddy University

B VINAYAKA DATTA
School of Engineering(AIML)
B Tech
Malla Reddy University

J Vinay Kumar
School of Engineering(AIML)
B Tech
Malla Reddy University

M VINESH GOUD
School of Engineering(AIML)
B Tech
Malla Reddy University

B VINAY
School of Engineering(AIML)
B Tech
Malla Reddy University

A VINOD
School of Engineering(AIML)
B Tech
Malla Reddy University

1 Abstract

One of the more difficult ailments is heart disease, which affected a large number of people worldwide. Heart illness must be promptly and accurately diagnosed in order to be treated, especially in the field of cardiology. In this work, we suggested a machine learning-based approach for diagnosing cardiac disease that is both effective and accurate. The system was created using classification algorithms, which Standard feature selection algorithms like Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator, and Local learning have been used to omit unnecessary and redundant features. Other feature selection algorithms include Support vector machine, Logistic regression, Artificial neural network, K nearest neighbor, Nave bays, and Decision tree. Additionally, we provided a brand-new, quick conditional mutual information feature selection approach to address.

1.1 Machine Learning :

This paper presents a Machine Learning that addresses Comparative Analysis of Heart Disease Prediction Models Heart disease is a major cause of death in the world. Early diagnosis and treatment can help to prevent heart attacks and other complications. However, there is no single test that can definitively diagnose heart disease.

2. Introduction:

Heart disease is a major worldwide health issue that requires precise prediction models in order to enhance patient care and intervene in a timely manner. In the context of heart disease prediction, this research study does a thorough comparative examination of three popular machine learning algorithms: Random Forest, Support Vector Machines (SVM), and Logistic Regression. Each model was created and assessed using common performance measures, drawing from a wide dataset of clinical and lifestyle variables. The study reveals the most trustworthy and accurate method for predicting heart disease, empowering medical practitioners to make well-informed decisions. Moreover, feature significance analysis clarifies the important variables affecting precise predictions. The insights acquired have the potential to impact clinical practice by selecting the best model based on interpretability and performance. In the end, this study advances the development of cardiac disease prediction models, improving patient outcomes and quality of life.

3. Literature Review:

Using 10-fold cross-validation using the Cleveland database, S. Musfiq Ali et al.'s research produced a maximum accuracy of 91.2% for GNB [5].Using the Cleveland dataset, A. Kondababu et al. (2021) performed comparative analysis and discovered that the HRFLM technique—a mix of Random Forest (RM) and Linear Method (LM)—had the greatest accuracy.Using an Isolation Forest preprocessed dataset with thirteen features, Rohit Bharti et al.[7] discovered that the KNeighbors classifier performed the best.After comparing many models for heart disease prediction, Sfruti Sarah et al.[8] discovered that LR had the highest accuracy, at 85.25%.Riyaz Lubana et al. In a comparison of several machine learning algorithms and how well they predicted cardiac illness, Lubana Riyaz et al.[9] discovered that ANN had the greatest average prediction accuracy, at 86.91%, while C4.5 decision trees had the lowest, at 74.0%.The optimal results were obtained by Xiao-Yan Gao et al. [10] using the bagging ensemble learning algorithm with Decision Tree and Principle component analysis feature extraction technique.Based on the RF classifier, Abdullah et al. created a data mining model to improve the accuracy of heart disease prediction [11]. Cleveland Dataset with 303 occurrences and 19 characteristics was utilized by Sonam Nikhar et al. [12] using GNB, DT approach.Additionally, they found that the Decision Tree outperforms the Nave Bayes Classifier in terms of accuracy. For the Cardio Vascular Disease Prediction Survey, Ravindra Yadav et al.[13] used an ML strategy that comprised DT, GNB, Neural Networks, Deep Learning, and SVM. The decision tree's conclusion is produced with the use of J48, ID3, CART Cpercent.0, and CYT.The Cleveland database, which has 303 instances and 14 characteristics, was used by Devansh Shah et al. [14] to predict cardiac disease. The most accurate algorithm was K-NN.

4. Problem Statement:

Heart disease is a leading cause of mortality worldwide. Early detection and treatment can aid in the prevention of heart attacks and associated problems. However, no one test can conclusively detect cardiac disease.The issue assertion is that no single test can definitively identify heart disease. As a result, doctors must frequently depend on a combination of testing and clinical judgment to reach a diagnosis. Based on a patient's medical history and other risk factors, machine learning may be used to create models that predict the risk of heart disease. This can assist medical professionals in identifying people who have a high risk of heart disease and in taking preventative measures.The aim of this research is to develop a machine learning model that can reliably diagnose or rule out heart disease in patients. This implies that at least 80% of patients should be accurately classified by the model.

5.Methodology :

1. Data Collection and Preprocessing:

Age, sex, type of chest pain, resting blood pressure, serum cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak (ST depression induced by exercise relative to rest), slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thal (thalassemia type) are the 14 attributes that make up the anonymized patient data used in this study. To protect patient privacy, patient IDs and social security numbers have been substituted with fake values.

2. Data Preprocessing:

Strict preprocessing is applied to the dataset in order to manage missing values and normalize the data. One-hot encoding is used to encode categorical features, such as thal and chest pain kind, into numerical representations that are appropriate for machine learning techniques.

3. Train-Test Split:

To guarantee reliable model assessment, the preprocessed dataset is split into a training set and a test set while keeping a suitable ratio. The test set is set aside for an objective performance evaluation, whereas the training set is utilized to train the models..

4. Feature Scaling:

Age, blood pressure, serum cholesterol, maximal heart rate, and other numerical characteristics are scaled to fall into a similar range in order to improve convergence and training stability for the machine learning algorithms.

5. Model Development:

Standard performance indicators, including as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), are used to assess the trained models. The evaluation's objectives are to ascertain each algorithm's capacity for prediction and to pinpoint the best model for predicting heart disease.

6. Feature Importance Analysis:

An examination of feature significance is carried out for interpretability in order to determine which features have the most influence on the accuracy of heart disease predictions. For this study, methods like the Random Forest model's feature significance scores and permutation importance are used.

7. Comparative Analysis:

The most dependable and efficient method for heart disease prediction is identified by comparing the performance metrics and feature importance findings among the three models: Random Forest, SVM, and Logistic Regression.

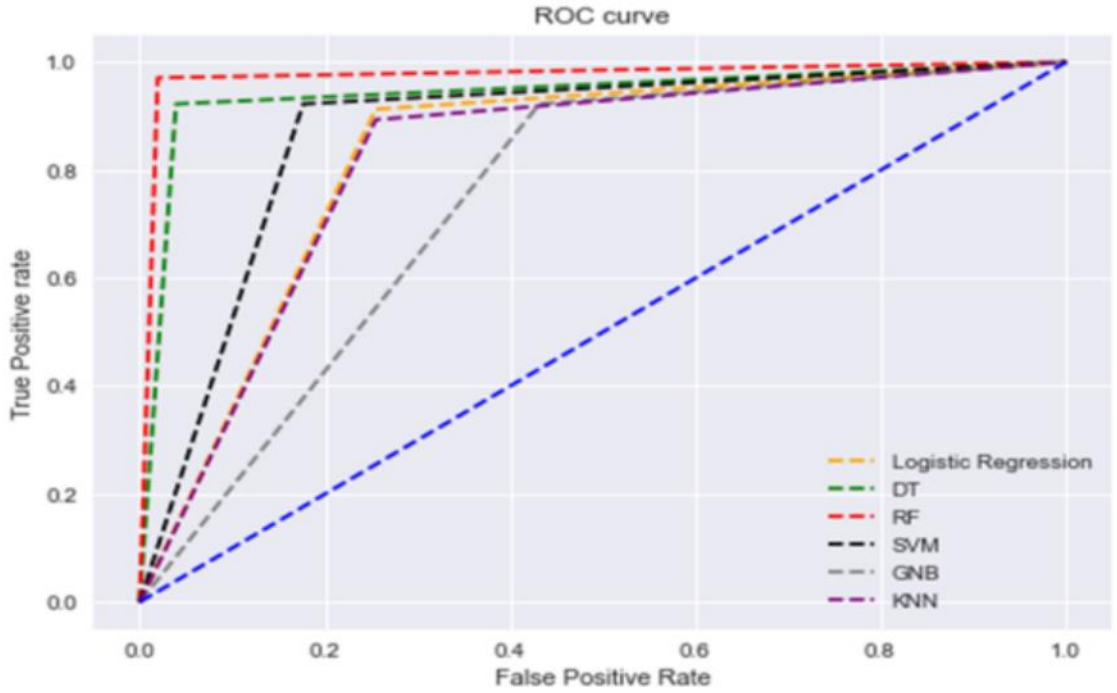
6. Experimental Results:

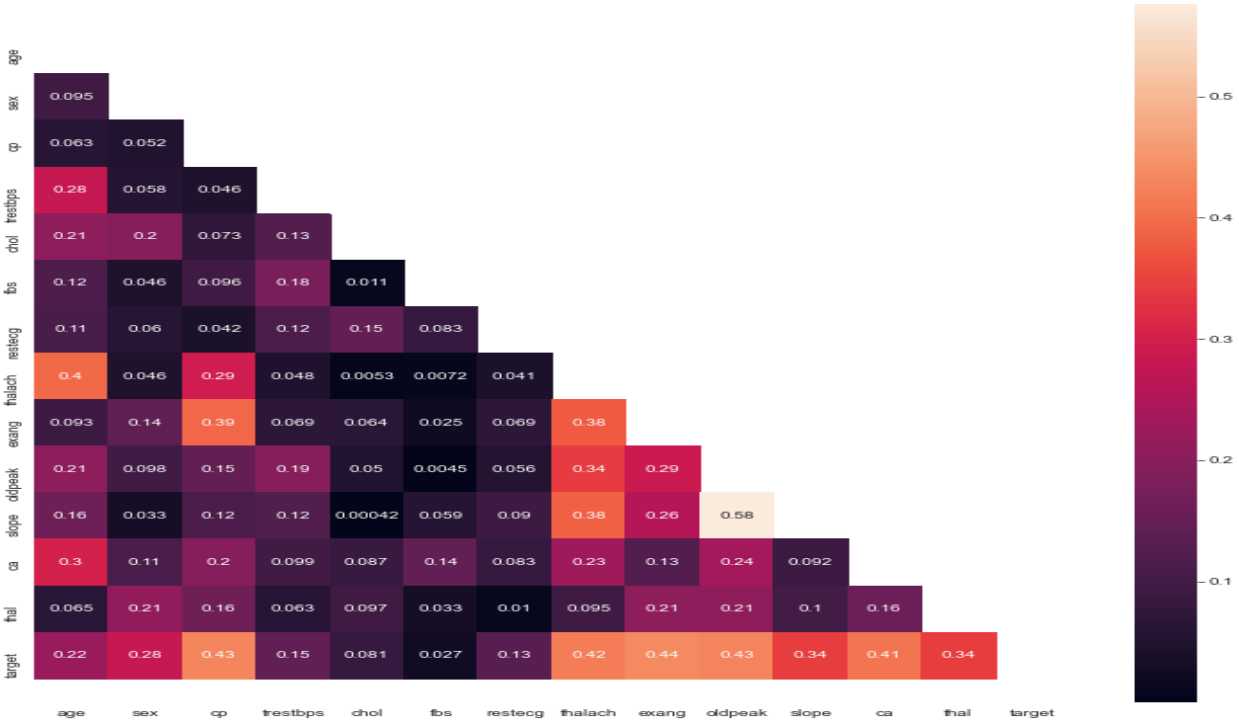
Accuracy, precision, recall/sensitivity, F1 Score, and harmonic mean are among the metrics that have been employed. The Random Forest Classifier has superior performance in terms of F-1 Score, Accuracy, Precision, and Recall.

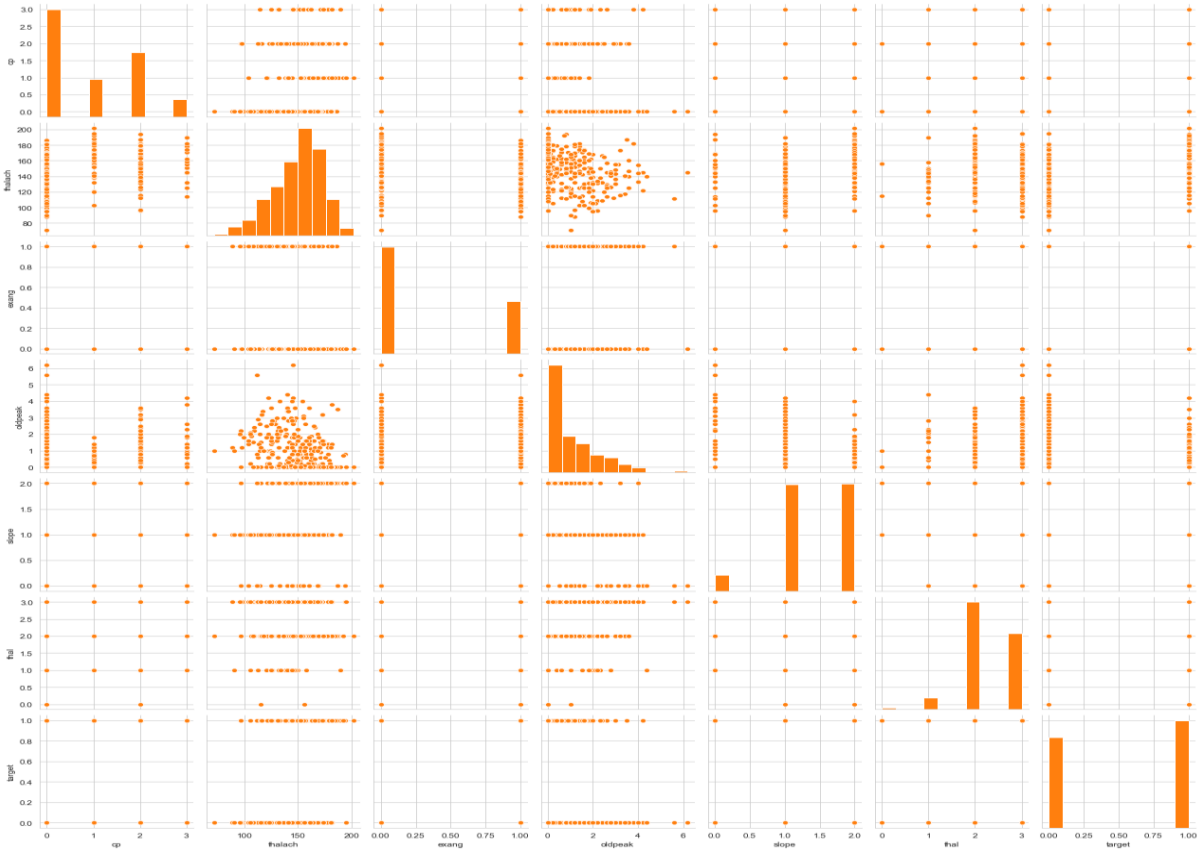
Out of all four measures, Naïve Bayes has shown the weakest performance. One-hot encoding was used for pre-processing the data, which greatly improves its usability, expressiveness, and ease of rescaling.

The ROC-AUC curve for each classifier is shown in Fig 1. It is preferable the larger the area under the curve. The value is a number between 0 and 1, with a value closer to 1 indicating higher algorithmic performance. The graph shows that the random forest, represented by the red dotted line, has the maximum

S. No.	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)
1	Logistic Regression	82.92	78.33	91.26	84.30
2	Decision Tree	95.12	94.28	96.11	95.19
3	Random Forest	98.53	100	97.08	98.52
4	Support Vector Machine	87.31	84.07	92.23	87.96
5	Gaussian Naïve Bayes	74.63	68.34	92.23	78.51
6	K- Nearest Neighbors	81.95	77.96	89.32	83.25


ROC curve
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
DT
RF
SVM
GNB
KNN


age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
target
0.095
0.063 0.052
0.28 0.058 0.046
0.21 0.2 0.073 0.13
0.12 0.046 0.096 0.18 0.011
0.11 0.06 0.042 0.12 0.15 0.083
0.4 0.046 0.29 0.048 0.0053 0.0072 0.041
0.093 0.14 0.39 0.069 0.064 0.025 0.069 0.38
0.21 0.098 0.15 0.19 0.05 0.0045 0.056 0.34 0.29
0.16 0.033 0.12 0.12 0.00042 0.059 0.09 0.38 0.26 0.58
0.3 0.11 0.2 0.099 0.087 0.14 0.083 0.23 0.13 0.24 0.092
0.065 0.21 0.16 0.063 0.097 0.033 0.01 0.095 0.21 0.21 0.1 0.16
0.22 0.28 0.43 0.15 0.081 0.027 0.13 0.42 0.44 0.43 0.34 0.41 0.34
0.5
0.4
0.3
0.2
0.1



6 conclusion:

This study looked at how well several supervised machine learning algorithms performed in heart disease prediction. It included "Decision Tree" (DT), "Logistic Regression" (LR), "Random Forest" (RF), and "Support Vector.""k-Nearest Neighbor" (kNN), "Gaussian Naïve Bayes" (GNB), and "SVM" (SVM).Numerous earlier studies on the same subject were examined. One-hot encoding was used to preprocess the data before it was divided into training and testing sets. A variety of metrics were developed and models were trained. According to this study, the random forest algorithm worked the best, with an accuracy of 98.53%, while the GNB algorithm performed the poorest, with an accuracy of 74.63%. One-hot encoding played a major role in enabling Random Forest to make more intelligent selections. The amount of work that can be done in this field is growing.the accuracy through feature selection, ensemble techniques, and hyper-parameter adjustment.

References:

[1] Seckeler MD, Hoke TR. Acute rheumatic fever and rheumatic heart disease: a global epidemiology. 2011;3:67; Clinical Epidemiology.

[2] Zhang J, Wang Y, and Liu Y. Random forest is a new machine learning algorithm. InGlobal

September 14, 2012, Conference on Information Computing and Applications (pp. 246-252). Heidelberg, Berlin: Springer.

[3] Padalia N, Naidu A, Mukherji D, Mythili T. A model for predicting heart disease using SVM-decision trees-logistic regression (SDL). Jan. 1, 2013; 68(16): International Journal of Computer Applications.

[4] Bharti SK, Patel S, Shah D. Predicting heart disease via machine learning methods. Nov. 2020;1(6):1-6; SN Computer Science.

[5] Jeeva, S. C., and Latha, C. B. C. (2019). enhancing the precision of ensemble classification-based heart disease risk prediction. Unlocked: Informatics in Medicine, 16, 100203.

[6] Jyoti, K., and N. Bhattala (2012). an examination of several data mining methods for the prediction of heart disease. 1(8), 1-4, International Journal of Engineering.

[7] Khan, M. S., Hassan, C. A. U., & Shah, M. A. (September 2018). Comparison of data classification machine learning algorithms. Volume 24, Issue 1, Pages 1-6, 24th International Conference on Automation and Computing (ICAC), 2018. IEEE.

[8] Mohammad, R. M. A., Singh, G., Thabtah, F., and Gonçalves, A. H. (2019, July). Experimental investigation of machine learning for coronary heart disease prediction. In the Third International Conference on Deep Learning Technologies (2019), proceedings (pp. 51–56).

[9] Riyaz L, Butt MA, Zaman M, Ayob O. Machine Learning Techniques for Predicting Heart Disease: A Quantitative Review. 81–94) at the 2022 International Conference on Innovative Computing and Communications. Singaporean Springer.

[10] Shaban Hassan H, Anwar EM, Gao XY, and Amin Ali A. increasing the precision of the ensemble method-based analysis of heart disease prediction. Complexity. Feb 10, 2021; 2021.

[11] AS Abdullah, R Rajalaxmi. a random forest classifier-based data mining model for coronary heart disease prediction. in 2012 April; pp. 22–25) in International Conference in Recent Trends in Computational Methods, Communication, and Controls.