RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 06, Issue 11 November

Check for updates

http://dx.doi.org/10.47392/IRJASH.2024.045

Beyond Boundaries: Achieving 100% Heart Disease Prediction Using Diverse Machine Learning Algorithms

V.Gnanalakshmi[1], P.Haritha[2], K.Gopika[3], P.Kaviya[4]
[1]Assistant professor (Sr. Grade), Dept. of ECE, Mepco Schlenk Engg. College, Viruthunagar, Tamil Nadu, India.
[2,3,4]UG Scholar, Dept. of ECE, Mepco Schlenk Engg. College, Viruthunagar, Tamil Nadu, India.
Emails: *v.gnanalakshmi@mepcoeng.ac.in[1]*

Article history
Received: 22 October 2024
Accepted: 08 November 2024
Published: 23 November 2024

Keywords:
Accuracy, Precision, Recall, F1_score, Heart Disease.

Abstract
Heart disease plays a virtual role in recent years.This study addresses critical need for early detection of heart disease to alleviate its impact .We propose a machine learning architecture for early stage of the art feature extraction utilizing states of the art feature extraction stratagies. The effect of 5 machine learning algorithm-logistic algorithm - k-neighbour - support vector machine - decision tree - random forest was evaluated. Among the 5 algorithm random forest exhibited superior performance .We used 2 heart disease datasets -one 303 instances and another with 1026 instances the larger datasets yields outstanding results with the random forest achieving perfect scores across all metrics - accuracy-1,precision-1,and recall-1 and f1-score-1.This impressive performance underscores the algorithm's effectiveness. The main objective of the project is to resolve and address specific issues

1. Introduction

Heart disease (CVDs) are the top cause of death around the world. In 2019, they caused about 17.9 million deaths making up 32% of all death globally. Most of these deaths were due to heart attacks or strokes additionally of the 17.1 million people who died early (under 70) from diseases that aren't infectious 38% were due to heart disease (CVDs).Often people don't show any signs of these disease until they have a heart attack or stroke which can be the first sign of the problem with the blood vessels. Symptoms includes Pain (or) Discomfort in the center of the chest Pain and Discomfort in the arms, left shoulder, elbows etc.. From 2005-2015 heart related diseases cost -India upto $237 billion. Both the men and women can get these disease. Usually due to unhealthy lifestyle overtime. If we can predict heart disease early it could save many lives and reduce costs. Most deaths from heart disease (CVDs) happen in low and middle income countries where people don't always have access to early detection and treatment. This means many people in these areas get diagnosed too late and die younger than they should think of it. It's important to detect heart disease (CVDs) early so that treatment can start as soon as possible. The dataset from the UIC machine learning respiratory has 303cases and 1026 cases each has same 14 features. Machine learning models can help to predict heart disease (CVDs) Conditions. By testing five different algorithm we can find the most accurate way to predict heart disease and potentially save billions of dollars and many lives. [1-3]

2. Literature Survey

In today's data-driven world, predictive modelling in healthcare has become essential for early diagnosis and decision-making, particularly in heart disease prediction. A growing number of machine learning techniques are being applied tackle this problem, ranging from simple classifiers to complex ensemble methods. Below, we explore several studies that employ logistic regression, decision trees, K-nearest neighbours (KNN), Naïve Bayes, support vector machines (SVM), and hybrid approaches in heart disease prediction .Rahman provide a comparative analysis of logistic regression and decision trees for heart disease prediction. Logistic regression is highlighted for its clarity and interpretability, making it a favoured model in clinical settings where explaining predictions is crucial. Decision trees, on the other hand, excel in datasets with non-linear relationships, often providing higher accuracy. Both models have their strengths—logistic regression is effective forunder standing how variables impact outcomes, while decision trees offer a transparent and visual decision-making process. The study suggests that a hybrid model combining both could a enhance predictive accuracy while retaining interpretability, a crucial balance in clinical applications. Kumar and Gupta's study focuses on KNN, a simple yet powerful algorithm when paired with robust feature selection. The authors emphasize the importance of pre-processing and choosing the right value for K, which determines how many neighboring data points influence classification. KNN is sensitive to irrelevant features, which can degrade performance, particularly in larger datasets. However, with proper feature engineering, KNN proves to be an effective method for predicting heart disease, although its performance is often outpaced by more complex algorithms like decision trees or ensemble methods. Verma and Kumar examine the effectiveness of the Naive Bayes algorithm, particularly its suitability for handling categorical data and small datasets. While Naive Bayes does not always achieve the highest accuracy, it remains competitive due to its computational efficiency and simplicity. This algorithm is particularly advantageous when the dataset aligns well with the independence assumption among features. Pre-processing steps, like data normalization and encoding categorical variables, further enhance its performance, making Naive Bayes a valuable tool in resource constrained environments Sharma and Patil compare decision trees and SVM for heart disease prediction, finding that while decision trees are interpretable, SVM excels in high-dimensional datasets with non-linear relationships. SVM often requires more tuning, including the selection of appropriate kernel functions and regularization parameters, but when properly configured, it can outperform decision trees in terms of accuracy. The study highlights the potential of hybrid models to combine the interpretability of decision trees with the accuracy of SVM. Logistic regression and Decision trees remain useful, the review points to a growing trend toward ensemble methods, such as random forests and gradient boosting, which aggregate multiple models to improve accuracy. Feature selection and pre-processing play a pivotal role in optimizing these algorithms, and hybrid models show promise in enhancing predictive outcomes in clinical applications. Jacob and Nair evaluate various classifiers, including logistic regression, KNN, Naive Bayes, SVM, and decision trees. They find that while decision trees and logistic regression offer interpretability, SVM consistently delivers higher accuracy, particularly for complex datasets. The study underscores the importance of feature selection and pre-processing in achieving optimal performance, particularly for algorithms like KNN, which suffer in larger, more complex datasets. [4-6]

3. Proposed Methods

This paper appears the investigation of different machine learning algorithms, the calculations that are utilized in this paper are Arbitrary forest, Logistic regression ,Decision tree, K closest neighbors and Support vector machine(SVM).The system works by collecting information and choosing the critical features. The information is at that point part into two parts: training information and test data. The system's precision is measured by testing it utilizing the test data. This show employments 14 therapeutic parameters such as age, sex, cholesterol, blood pressure, etc. Pre-processing and information stacking was carried out utilizing the gotten data. This models are assessed utilizing accuracy, precision, recall and F1 score.

3.1 Logistic Regression

Logistic Regression is a essential classification Calculation utilized for foreseeing double outcomes. It models the likelihood of occurrence based on a direct combination of input features.

Before being tried with test data, the isolated information is prepared utilizing calculated regression. The exact comes about are gotten utilizing calculated regression. The to begin with step is to select an appropriate dataset for preparing and testing the model. The target variable is a twofold outcome. Pre-processing is basic to guarantee the quality and consistency of the data some time recently applying the Calculated relapse model. Once the information is pre-processed, the calculated relapse show can be trained. After preparing, the demonstrate can be utilized to anticipate whether the patients in the test set have heart disease. [7-9]

3.2 Naive Bayes

This proposed approach employments the Naïve Bayes method based on the Bayes hypothesis to select the best subset of highlights for the another classification stage, moreover to handle the tall dimensionality issue by maintaining a strategic distance from pointless highlights and select as it were the vital ones in an endeavor to move forward the proficiency and exactness of classifiers. This strategy is able to decrease the number of highlights from 13 to 6 which are (age, sex, blood weight, fasting blood sugar, cholesterol, work out actuate motor) by deciding the reliance between a set of qualities. [10]

3.3 K Nearest Neighbour

KNN is a non-parametric calculation that classifies Information focuses based on the larger part lesson of their course of their k-nearest neighbours. The KNN show was prepared utilizing the preparing set. For each occurrence in the test data, the calculation identified the K closest neighbor from the preparing data. The performance of the KNN show was assessed utilizing the test set like accuracy, confusion matrix, and performance metrics.

3.4 Random Forest:

Random timberland is an outfit learning strategy that leverages different choice tree to upgrade prescient exactness and generalization by combining the yield of person tree, random timberland mitigates over fitting and progresses robustness. The Irregular timberland was chosen for its robustness, ability to handle non-linear relationships, and viability in reducing over fitting. At each part in the trees, a subset of highlights was arbitrarily chosen to decide the best Part, improving the model's capacity to generalize. Each tree in the timberland was built independently, and expectations were made based on the larger part vote of all trees.

3.5 Support Vector Machine (SVM)

SVM is a capable classification calculation that works by finding a hyperplane that maximally seperates classes in the include space. SVM is viable in taking care of complex choice boundaries and is well-suited for datasets with tall dimensionality. In The dataset ought to be part into preparing and test sets to assess the model's performance. SVM models handle both straight and non-linear choice boundaries. After preparing the SVM model, it can be utilized to make expectations on the testing dataset. The execution of this demonstrate can be assessed utilizing classification measurements like accuracy, precision, recall and F1score. Figure 3 shows Accuracy Benchmarking for 1026 Instances. Figure 1 shows the Flow chart. Table 1 shows Heart Disease 13 Features Dataset.

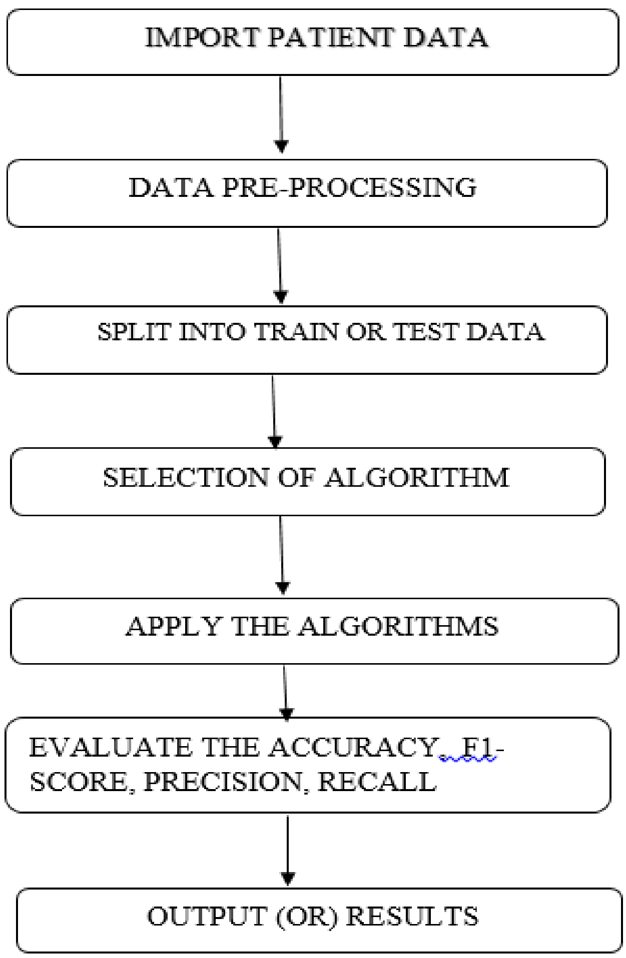


Figure 1 Flow Chart

Table 1 Heart Disease 13 Features Dataset

Attributes	Description
Age	Age in years
Sex	Male or Female
Cp	Chest pain type
Thestbps	Resting blood pressure
Chol	Serum cholesterol
Restecg	Resting electrographic results
Fbs	Fasting blood sugar
Thalach	Max. heart rate achieved
exang	Exercise induced angina
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment
Ca	No. of major vessels colored
Thal	Defect type

4. Experimental Results

The prediction models are developed using 13 Features and the accuracy is calculated for modelling techniques. The given below table compares the accuracy, precision, F-Score, Recall for 2 heart disease dataset (303, 1026 instances) .The highest accuracy is achieved by Random forest classification method in comparison with existing methods. The table 2 presents the metrics for various algorithms evaluated on 1026 instances. The table 3, presents the metrics for various algorithms evaluated on 303 instances. The table 4 presents the accuracy for random forest algorithm evaluated on 303 and 1026 instances. Table 5 shows Accuracy Comparison of Supervise Machine Learning Algorithms for 1026 instances. Table 6 shows Accuracy Comparison of Supervise Machine Learning Algorithms for 303 instances. Figure 2 shows Accuracy Benchmarking for 303 instances.

Table 2 Metrics for Various Algorithms Evaluated on 1026 Instances

ALGORITHM	ACCURACY	PRECISION	RECALL	F1SCORE
NAÏVE BAYES	0.82	0.86	0.80	0.83
RANDOMFOREST	1.0	1.0	1.0	1.0
SVM	0.70	0.75	0.69	0.72
LOGISTIC REGRESSION	0.85	0.92	0.81	0.86
K-NEIGHBOUR	0.89	0.89	0.89	0.89

Table 3 The Metrics for Various Algorithms Evaluated on 303 Instances

ALGORITHM	ACCURACY	PRECISION	RECALL	F1SCORE
NAÏVE BAYES	0.85	0.89	0.84	0.86
RANDOMFOREST	1.0	1.0	1.0	1.0
SVM	0.68	0.83	0.66	0.74
LOGISTIC REGRESSION	0.85	0.91	0.83	0.87
K-NEIGHBOUR	0.77	0.83	0.76	0.79

Table 4 The Accuracy for Random Forest Algorithm Evaluated on 303 and 1026 Instances

DATASET	ACCURACY	PRECISION	RECALL	F1SCORE
303 SAMPLES	1.0	1.0	1.0	1.0
1026 SAMPLES	1.0	1.0	1.0	1.0

Table 5 The Accuracy for Random Forest Algorithm Evaluated on 303 and 1026 Instances

NAÏVE BAYES	RANDOM FOREST	SVM	LOGISTIC REGRESSION	K-NEIGHBOUR
0.82	1.0	0.70	0.85	0.89

Table 6 Accuracy Comparison of Supervise Machine Learning Algorithms for 303 Instances

NAÏVE BAYES	RANDOM FOREST	SVM	LOGISTIC REGRESSION	K-NEIGHBOUR
0.85	1.0	0.68	0.85	0.77

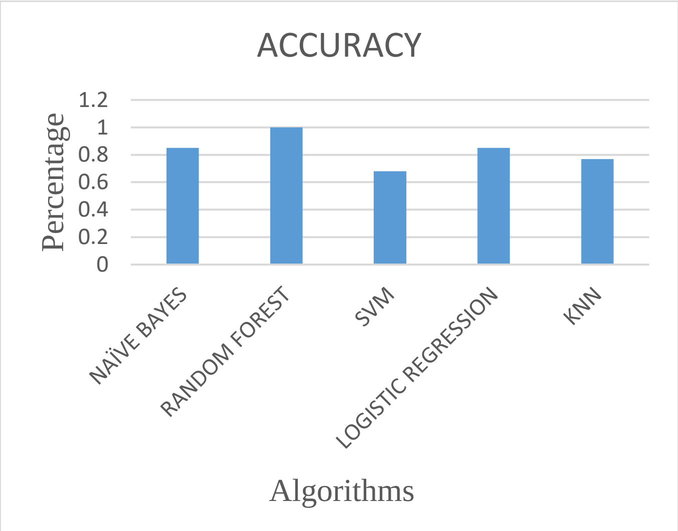


Figure 2 Accuracy Benchmarking for 303 Instances

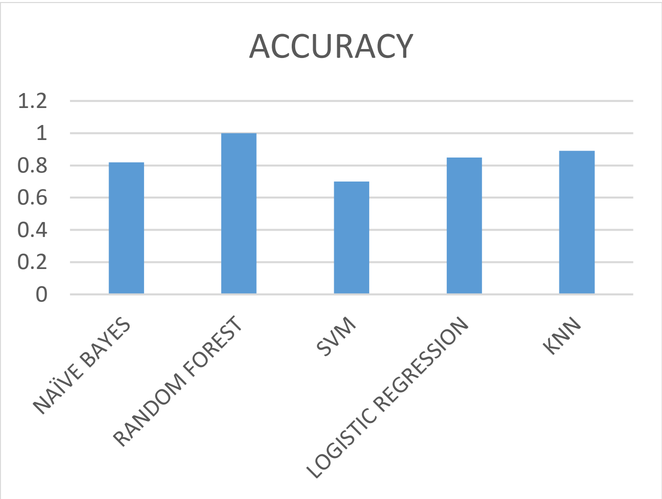
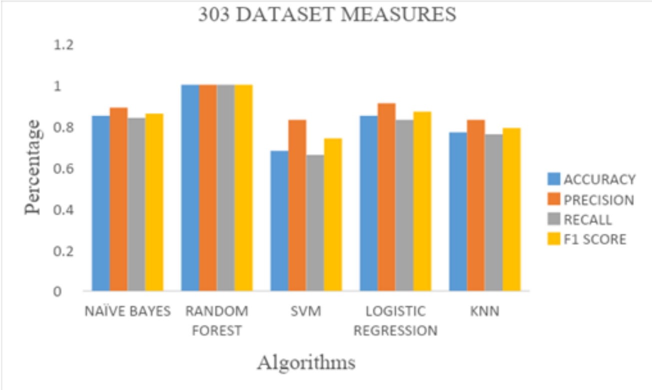


Figure 3 Accuracy Benchmarking for 1026 Instances

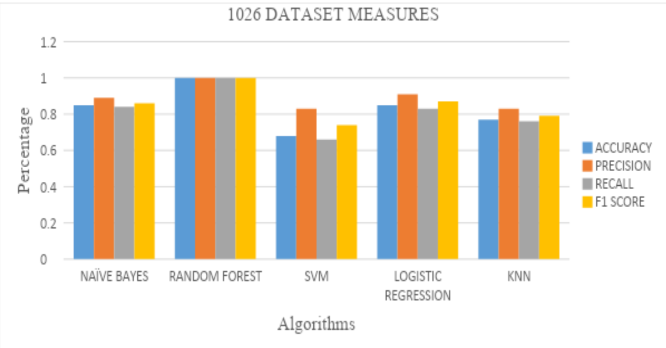
The given below graph predicts the accuracy of different algorithms for 303 instances. (Refer Figure 4)

Figure 4 The Accuracy of Different Algorithms for 303 Instances



The given below graph predicts the accuracy of different algorithms for 1026 instances. . (Refer Figure 5)

Figure 5 The Accuracy of Different Algorithms for 1026 Instances



Conclusion

In conclusion, early detection of heart disease is vital for saving lives and reducing healthcare costs. This study shows that machine learning, especially the Random Forest algorithm, can significantly improve predictions of heart disease. By using these advanced methods, we can help ensure that people, especially in underserved areas, receive timely care. Our work aims to make a positive impact on heart disease management and enhance patient outcomes around the world.

References

[1]. Heart Disease Prediction Using Logistic Regressio Algorithm Based on Features Selection A. K. Rana, S. Bansal International Journal of Engineering and Advanced Technology (IJEAT), 2020.

[2]. A Comparison of Machine Learning Techniques for Heart Disease Prediction C. Soni, A. Jain, U. Kumar Advances in Computational Sciences and Technology,2017.

[3]. Prediction of Heart Disease Using Machine Learning Algorithms: Logistic Regression, Decision Tree, and Random Forest S. Jain, M. Gupta International Journal of Research in Engineering, Science and Management (IJRESM), 2021.

[4]. HeartDiseasePredictionUsingMachineLear ningTechniques: A Case Study of Logistic Regression and Random Forest R. Shukla, N. Patel International Journal of Advanced Computer Science and Applications (IJACSA) 2019.

[5]. Classification of Cardiovascular Disease Using K-Nearest Neighbor Algorithm A. Kumar, V. Kumar International References to papers accepted for publication but not yet published should show the journal name, the probable year of publication (if known), and they should state "in press." Journal of Innovative Research in Computer and Communication Engineering, 2021.

[6]. Using Support Vector Machine to Predict Heart Disease M. Refaeilzadeh, L. Tang International Journal of Computer Science and MobileComputing (IJCSMC), 2020.

[7]. Comparative Analysis of Heart Disease Prediction Using Data Mining Techniques P. Patel, A. Patel International Journal of Emerging Technology and Advanced Engineering (IJETAE), 2018.

[8]. Heart Disease Diagnosis Using Decision Tree and Random Forest Algorithms J. Lee, S. Yang IEEE Transactions on Computational Biology and Bioinformatics, 2019.

[9]. Machine Learning-Based Cardiovascular Disease Prediction Models Using K-Nearest Neighbor and Logistic Regression A. Roy, S. Mazumdar International Journal of Scientific & Technology

[10]. Heart Disease Prediction Using Support Vector Machinesand Feature Extraction Techniques"M. Ahmed, R. Farah International Journal of Advanced Research in Artificial Intelligence (IJARAI), 2020.