Performance analysis and comparison of machine learning algorithms for predicting heart disease

Neha Bhadu, Jaswinder Singh
Department of Computer Science and Engineering, Faculty of Engineering and Technology, Guru Jambheshwar University of Science and Technology, Hisar, India

Article Info

Article history:

Received Apr 19, 2024
Revised Mar 21, 2025
Accepted Jun 8, 2025

Keywords:

Decision tree
Heart disease
Machine learning
Performance metrics
Random forest
WEKA

ABSTRACT

Heart disease (HD) is a serious medical condition that has an enormous effect on people's quality of life. Early as well as accurate identification is crucial for preventing and treating HD. Traditional methods of diagnosis may not always be reliable. Non-intrusive methods like machine learning (ML) are proficient in distinguishing between patients with HD and those in good health. The prime objective of this study is to find a robust ML technique that can accurately detect the presence of HD. For this purpose, several ML algorithms were chosen based on the relevant literature studied. For this investigation, two different heart datasets the Cleveland and Statlog datasets were downloaded from Kaggle. The analysis was carried out utilizing the Waikato environment for knowledge analysis (WEKA) 3.9.6 software. To assess how well various algorithms predicted HD, the study employed a variety of performance evaluation metrics and error rates. The findings showed that for both the datasets random forest (RF) is a better option for predicting HD with an accuracy and receiver operating characteristic (ROC) values of 94% and 0.984 for the Cleveland dataset and 90% and 0.975 for the Statlog dataset. This work may aid researchers in creating early HD detection models and assist medical practitioners in identifying HD.

Corresponding Author:

Neha Bhadu
Department of Computer Science and Engineering, Faculty of Engineering and Technology
Guru Jambheshwar University of Science and Technology
Hisar-125001, Haryana, India
Email: nehabhadu100@gmail.com

1. INTRODUCTION

People nowadays are facing major health challenges. Utilization of tobacco, unhealthy dietary patterns, and insufficient physical activity are leading to numerous chronic illnesses. Chronic illnesses are the main reasons for death and disability worldwide. As per the US National Centre for Health Statistics, chronic diseases persist for an extended duration, typically exceeding three months. These diseases are neither curable through medication nor preventable through vaccination. Health conditions like heart disease (HD), cancer, arthritis, diabetes, obesity, depression, and others fall under this category of diseases [1]. One of the deadliest chronic illnesses, HD, will be the subject of this investigation. The human heart is in charge of pumping blood, supplying all body organs with nutrition and oxygen, and removing harmful elements like carbon dioxide. Several conditions that affect the structure and function of the heart are collectively referred to as HD. HD is classified as cardiovascular disease (CVD). CVD encompasses a range of heart and blood vessel conditions, such as peripheral arterial disease, heart attacks, strokes, and coronary HD. It is essential to understand that while all HDs are CVDs, not all CVDs are classified as HDs [2]. Several factors, including age, sex, tobacco use, having a family history of HD, high blood pressure, high cholesterol, eating an unhealthy diet, hypertension, being overweight, inactivity, and alcohol consumption, can raise one's chance of developing HD [3]. There exist various forms of HD, such as "coronary HD, angina pectoris, congestive heart failure, cardiomyopathy, congenital HD, arrhythmias, and myocarditis" [4]. The most prevalent condition among these is coronary HD. As a result of this condition, the coronary arteries, which feed the heart with blood rich in oxygen, shrink or block. Common signs of HD include chest discomfort, difficulty breathing, light-headedness, nausea, puffy feet, extreme sweating, and general fatigue.

Timely identification of HD can help reduce the mortality rate and minimize overall consequences. Traditionally, HD is diagnosed by analyzing the patient's medical background, carrying out a thorough physical examination, and assessing the relevant signs by the physician. This traditional diagnosis, however, can be inaccurate and is costly and time-consuming. The use of artificial intelligence (AI) methods, particularly machine learning (ML) algorithms, is one possible approach to overcoming these obstacles. ML, a branch of AI, applies algorithms to data analysis so that computers can recognize, learn, spot patterns, and make informed judgments. ML algorithms operate on a mathematical model that relies on a training dataset to predict outcomes or make decisions without explicit programming [5]. By analyzing medical records, these algorithms can recognize persons who might develop HD, leading to earlier diagnosis and treatment and eventually lowering mortality rates.

Every year, approximately 17.9 million lives are claimed by CVDs, making them the major cause of fatalities globally, according to data from the World Health Organization (WHO) [6]. According to the World Health Federation's (WHF) World Health Report 2023, CVDs claimed the lives of 20.5 million people in 2021, accounting for roughly one-third of global mortality. In 1990, there were 12.1 million deaths from CVD. However, this number has significantly increased. If nothing is done to prevent it, by 2030, the global death toll is expected to reach around 22 million [7]. According to the data, HD is a serious universal health concern, highlighting the need for more study in this area.

Recent developments in ML have greatly enhanced HD prediction through the use of ensemble techniques such as random forest (RF) and extreme gradient boosting (XGB), feature selection methods, integration of feature selection methods with metaheuristic optimization techniques, and the creation of hybrid models that combine traditional ML learning with deep learning. These models perform better than conventional ML techniques by identifying intricate data patterns. But even with these advancements, a thorough evaluation of various ML algorithms is still required to ascertain how well they perform in diverse scenarios. Many currently available research concentrates on specific models without assessing their relative advantages, disadvantages, and effectiveness. The chief purpose of this analytical study is to evaluate and contrast several ML models, offering a systematic performance analysis to choose the most precise, effective, and reliable algorithm by examining research questions (RQs) that will help healthcare institutions as well as hospitals in advancing the knowledge and directing the development of new healthcare applications. The RQs include: i) which ML algorithms are frequently used for predicting HD? and ii) which of these algorithms demonstrate superior performance in HD prediction? To answer these RQs, a thorough examination of relevant literature is required, as elaborated in the following segment.

This work is organized into different sections. An overview of HD, including its types, symptoms, primary risk factors, statistics, current state of the art, and objective of the study is given in section 1. The work of multiple researchers on the early detection of HD using various conventional and hybrid ML models is compiled in section 2. The techniques employed in this investigation for identifying HD are described in section 3. The findings from the experiment and a comprehensive analysis are provided in section 4. Finally, the last section sums up the findings and makes recommendations for additional research and study implications.

2. RELATED WORK

Researchers predicted HD using a range of ML approaches. Extensive research has already been done and is continuing for further enhancements in prediction. Numerous publications covering the years 2018 to 2024 have been compiled from resources like IEEE Xplore, Google Scholar, ResearchGate, and ScienceDirect to address RQ1. This section provides insight into different ML prediction models for predicting HD.

Haq *et al*. [8] proposed a hybrid smart ML predictive approach for identifying HD. Seven well-known classifiers logistic regression (LR), artificial neural network (ANN), K-nearest neighbor (KNN), naïve Bayes (NB), support vector machine (SVM), RF, and decision tree (DT) were used to achieve this. Three algorithms were used to find out the most significant features: relief, least absolute shrinkage and selection operator (LASSO), and minimum redundancy maximum relevance (mRMR). The Cleveland dataset was utilized for model assessment, and the outcomes were validated using K-fold cross-validation. The relief algorithm helped achieve an accuracy of 89% with LR using 10-fold cross-validation. Mohan *et al*. [9] merged the benefits of the linear method (LM) along with RF to create the hybrid random forest linear model (HRFLM) hybrid methodology. The model's accuracy score on the Cleveland dataset was 88.7%, indicating improved performance with the use of an R studio rattle. Bashir *et al*. [10] intended to increase the level of accuracy of HD identification by utilizing feature selection methods. They conducted experiments using various ML classifiers namely SVM, LR, NB, DT, and RF on an HD dataset obtained from University of California, Irvine (UCI) using the rapid miner tool. The findings indicated that LR and NB, exhibited improved accuracy. Repaka *et al*. [11], proposed a smart heart disease prediction system (SHDP), by incorporating the NB classifier along with an advanced encryption standard (AES) for predicting HD. The results indicated that this approach outperformed NB, achieving an accuracy rate of 89.77%. Furthermore, AES demonstrated superior security performance when compared to parallel homomorphic encryption algorithm (PHEA). Fitriyani *et al*. [12] proposed HDPM to predict HD. To enhance the accuracy, the model integrated synthetic minority oversampling technique-edited nearest neighbors (SMOTE-ENN), and density-based spatial clustering of applications with noise (DBSCAN) along with XGBoost ML classifier. The training dataset was balanced using SMOTE-ENN. DBSCAN was used for detecting and removing outlier data, and XGBoost was used for generating the predictive model. The model was constructed using the Cleveland and the Statlog datasets. In the evaluation stage, heart disease prediction model (HDPM) outperformed six other ML algorithms, exhibiting a superior accuracy score of 98.40% on the Cleveland and 95.90% on the statlog dataset. Katarya and Meena [13] used the UCI dataset to examine the effectiveness of many ML methods, comprising KNN, LR, NB, SVM, DT, RF, MLP, ANN, and DNN, in predicting HD. RF was identified as the most accurate algorithm of all. Li *et al*. [14], developed an HD prediction model using KNN, SVM, LR, NB, ANN, and DT classifiers of ML. Different methods such as mRMR, relief, local learning, and LASSO were used to eliminate irrelevant and redundant attributes. The cross-validation technique utilized was “leave-one-subject-out”. According to the study, the suggested feature selection method (FCMIM) works well when paired with SVM to create an advanced intelligent system for HD identification. Thakkar *et al*. [15] developed a framework to conduct a comprehensive performance analysis of five ML methods specifically KNN, LR, SVM, NB, and RF. The testing was done using the Cleveland HD dataset. The majority of performance metrics indicated that LR outperformed the other classifiers consistently.

Shah *et al*. [16] applied the Cleveland HD dataset to four ML classification techniques: DT, RF, KNN, and NB. Waikato environment for knowledge analysis (WEKA) was used for carrying out the analysis. The findings revealed that KNN yielded the highest accuracy score. Sharma *et al*. [17] created an ML model using four different classifiers: RF, SVM, NB, and DT. The experiment used an HD dataset from UCI. The results showed that RF attained a 99% accuracy rate in a more efficient prediction timeframe. Hossen *et al*. [18] utilized three ML classifiers namely RF, DT, and LR for predicting HD, and their comparative assessment was done. The experimentation was carried out using the UCI Cleveland database. LR had the highest accuracy score of 92.10%, making it the best performer overall. Bashir *et al*. [19] proposed a voting system using an ensemble approach to accurately predict HD. For testing purposes, four HD datasets sourced from the UCI repository were utilized. Outcomes showed that the ensemble scheme achieved an accuracy of 83%, outperforming other ensemble schemes and individual classifiers. Rani *et al*. [20] created a hybrid approach-based decision support system for HD prediction. For selecting the most relevant features, a hybrid algorithm that integrated recursive feature elimination (RFE) along with a genetic algorithm (GA) was utilized. The Cleveland HD dataset was used for model testing. Pre-processing of the data was done using standard scalar techniques and SMOTE. Missing values were handled by applying the multivariate imputation by chained equations technique. Finally, five ML techniques: LR, SVM, NB, RF, and adaptive boosting (AdaBoost) were used. The hybrid system performed exceptionally well with an accuracy of 86.6%. Ghosh *et al*. [21] developed a hybrid model by combining bagging and boosting techniques with five conventional ML classifiers. Bagging was applied to KNN, DT, and RF resulting in K-nearest neighbors bagging method (KNNBM), decision tree bagging method (DTBM), and random forest bagging method (RFBM) hybrid methods. Boosting was applied to AdaBoost and gradient boosting resulting in AdaBoost boosting method (ABBM) and gradient boosting boosting method (GBBM) hybrid methods. For selecting relevant features LASSO and relief techniques were employed. A comprehensive dataset comprising five benchmark datasets, Cleveland, Statlog, Hungarian, Switzerland, and Long Beach VA for HD, was used to conduct the studies. The findings revealed that RFBM along with relief feature selection outperformed others with an accuracy of 99.05%. Ashri *et al*. [22] proposed an innovative hybrid intelligent framework, integrating five ML methodologies including KNN, SVM, LR, DT, and RF with a majority voting technique. Additionally, a simple genetic algorithm (SGA) was employed for feature selection, improving prediction performance and reducing overall time consumption. Overfitting was addressed by using 10-fold cross-validation. The UCI HD dataset was utilized for the experiments. The outcomes showed that the ensemble technique accomplished a remarkable accuracy of 98.18%. Ali *et al*. [23] carried out a comparative evaluation of various ML classifiers. A feature importance score was computed across all classifiers except for KNN and MLP. This score was used to rate each feature. The HD dataset was obtained Kaggle ML repository. The findings revealed that three classifiers namely DT, RF, and KNN achieved equally outstanding performance with 100% accuracy, sensitivity, and specificity. Ishaq *et al*. [24] employed nine ML classifiers such as LR, SVM, DT, RF, stochastic gradient classifier (SGC), AdaBoost, gradient boosting classifier (GBM), gaussian naive Bayes (GNB), and extra tree classifier (ETC) in this study. The class imbalance issue was addressed with SMOTE. Additionally, the models were trained on top features chosen by RF. The results showed that ETC with SMOTE performed the best, reaching an accuracy of 92.62%. Chang *et al.* [25] created a Python-based application to detect HD with improved precision. The model was constructed using an RF classifier. The application attained a remarkable accuracy rate of 83%.

Abdellatif *et al*. [26] suggested an efficient approach to construct the model by combining SMOTE, extra trees (ET), and hyperband (HB) techniques. SMOTE was used to resolve class inequality, ET was used for classification and HB was used for optimization of hyper-parameters. For predicting the severity level of HD, six distinct ML classifiers, namely LR, SVM, KNN, ET, stochastic gradient descent (SGD), and XGBoost were employed. The experimentation was conducted utilizing the Cleveland and Statlog datasets. The outcomes revealed that the highest accuracy of 99.2% and 98.52% was achieved by SMOTE and ET optimized by HB, respectively. Ahmad *et al.* [27] conducted a performance investigation of various ML classifiers including SVM, KNN, DT, RF, GBC, and linear discriminants analysis (LDA). To select the most significant features, a sequential feature selection technique was used. Employing the K-fold cross-validation technique, verification was completed. The combined (Statlog+Cleveland+Hungary) dataset, together with the individual datasets from Cleveland, Hungary, Switzerland, and Long Beach V, were used to evaluate how well the model performed. With nearly similar findings of 100 and 99.40% for the first dataset and 100 and 99.76% for the second, respectively, the RF sequential feature selection (SFS) and DT SFS showed the greatest accuracy values for both datasets. Ahmad *et al*. [28] utilized GridSearchCV in conjunction with multiple ML methods such as SVM, LR, KNN, and XGBoost for identifying HD. Further, a comparative study was conducted. Fivefold cross-validation was used as a verification approach. The datasets from UCI Kaggle, Long Beach V, Hungary, Switzerland, and Cleveland were utilized to assess the system. The outcomes demonstrated that, when combined, XGBoost and GridSearchCV generated the utmost and approximately equivalent testing as well as training accurateness levels of 100 and 99.03% on both datasets. Abdellatif *et al.* [29] offered a novel strategy that used improved weighted random forest (IWRF) for identifying HD, Bayesian optimization for optimizing IWRF's hyper-parameters and supervised "infinite feature selection (Inf-FSs)" to determine important features. The HD clinical records and the Statlog datasets were used in the model's development and testing. The results demonstrated that, concerning accuracy and F-measure, Inf-FSs-IWRF outperformed other models on both datasets. Cenitta *et al.* [30] designed a novel feature selection technique for ischemic HD namely ischemic heart disease squirrel search optimization (IHDSSO). The model's effectiveness was confirmed by utilizing the UCI HD dataset. The outcomes demonstrated that the IHDSSO model could identify the most significant attributes with an accuracy rate of more than 98.38% by using the RF classifier. Khan *et al.* [31] evaluated the effectiveness of five predictive ML classifiers including LR, SVM, NB, DT, and RF, for patients with CVD. The data was provided by the Khyber Teaching Hospital as well as the Lady Reading Hospital, located in Khyber Province, Pakistan. Upon conducting exploratory analysis, it was revealed that RF had attained the greatest percentages of 85.01, 92.11, and 87.73% for accuracy, sensitivity, and receiver operating characteristic (ROC) curve, respectively. Ullah *et al.* [32] introduced a scalable ML-based framework by integrating sophisticated feature selection techniques including fast correlation-based filter (FCBF), mRMR, relief, and particle swarm optimization (PSO). These methods were applied to extract and identify the most significant features from ECG signals. The refined feature set was then used to train ML classifiers such as ET and RF, which achieved outstanding accuracy rates of 100% on both small and large datasets. Biswas *et al.* [33] used three distinct techniques to choose important features namely analysis of variance (ANOVA), chi-square, and mutual information. Furthermore, six distinct ML methods were utilized, comprising SVM, LR, KNN, NB, DT, and RF. These models were used to determine the most effective model and feature subset. Finally, it was found that when mutual information feature subsets were used, RF had the highest accuracy rate, at 94.51%. Reshan *et al.* [34] developed a new hybrid deep neural network (HDNN) model. The model used convolutional neural networks (CNN), ANN, long short-term memory (LSTM), and an integration of LSTM with CNN over many layers. Further to enhance the quality of data, data imputation techniques were utilized. The model was trained using two datasets, the Cleveland and the combined HD dataset, which includes data from five benchmark datasets. A remarkable accuracy rate of 98.86% was shown by the suggested technique.

Qadri *et al.* [35] suggested a new method for feature engineering in principal component heart failure (PCHF), focusing on selecting the top eight features to improve performance. By introducing a novel feature set, PCHF was fine-tuned to achieve optimal accuracy scores. The study utilized nine ML classifiers to conduct thorough analysis and evaluations. The findings indicated that the DT method surpassed other ML models, achieving a remarkable accuracy score of 100%. Patra *et al.* [36] developed a highly effective hybrid voting ensemble approach to accurately identify the risk of HD. The Framingham HD dataset's characteristics were optimized for the model, and their relevance to the result was evaluated. The forward feature selection approach was then used to integrate these ranking features using traditional classifiers to produce meta-models with feature weights. The suggested hybrid model was ultimately formed by selecting the top 5 performing classifiers. The results showed a remarkable accuracy rate of 95.87%. Ahmad and Polat [37] suggested an ML-based intelligent HD diagnostic model. A swarm-based metaheuristic technique called jellyfish optimization was used to choose the optimal features to overcome the overfitting problem brought on by the abundance of characteristics in the Cleveland dataset. The best characteristics from the dataset were then chosen, and four distinct ML algorithms namely SVM, ANN, DT, and AdaBoost were employed for simulation. All ML methods demonstrated higher accuracy rates when using the jellyfish technique. The SVM model in particular had the best accuracy of 98.47%. Noor *et al*. [38] presented PaRSEL, a novel stacking model. The base layer is comprised of the ridge classifier (RC), the passive-aggressive classifier (PAC), XGBoost, and the stochastic gradient descent classifier (SGDC). On the meta layer, LogitBoost was employed. RFE, linear discriminant analysis (LDA), and factor analysis (FA) were the three methods employed to reduce dimensionality. To address the imbalanced nature of the dataset, eight balancing procedures were applied. The outcomes showed that PaRSEL outperformed other stand-alone classifiers, with an accuracy of 97%. Jafar and Lee [39] developed an automatic ML system called HypGB. It used the GB classifier for classification. To choose the best feature subset and eliminate duplicate and noisy attributes, a traditional LASSO technique was employed. The GB model was enhanced using the most recent version of the HyperOpt optimization framework. Experimental results for the Cleveland HD and Kaggle heart failure datasets show that HypGB was able to successfully identify features and obtain outstanding classification accuracies of 97.32 and 97.72%. Chandrasekhar and Peddakrishna [40] tested six ML techniques comprising LR, KNN, NB, RF, GB, and AdaBoost, using the data from Cleveland and IEEE Dataport. To increase model correctness, the study employed GridsearchCV along with five-fold cross-validation. In the Cleveland dataset, LR performed better than the other algorithms with 90.16% accuracy, whereas AdaBoost performed better with 90% accuracy in the IEEE Dataport dataset. The accuracy of the model was further raised to 93.44% and 95% for the Cleveland and IEEE Dataport datasets, correspondingly, by integrating all six approaches with the soft voting ensemble classifier. Hossain *et al.* [41] employed the best first search along with a feature subset selection method based on correlation to discover the best features in the data. Two types of HD datasets one with all features and the other with chosen features were used to test numerous ML approaches. These included SVM, LR, KNN, NB, DT, RF, and MLP. Among these techniques, RF using the selected features demonstrated the highest accuracy of 90%. Jawalkar *et al*. [42] proposed an ML-based approach for identifying HD by employing a loss-optimized decision tree-based random forest (DTRF) classifier. Furthermore, the DTRF classifier was trained utilizing a loss optimization technique called stochastic gradient boosting (SGB). According to the results, the suggested HDP-DTRF approach obtained a 96% accuracy rate on publicly available real-world datasets. Manikandan *et al.* [43] evaluated and contrasted the results of the SVM, LR, and DT algorithms both in conjunction with and without using the feature selection approach named boruta. This investigation was conducted using the Cleveland HD dataset. It was discovered that the Boruta algorithm enhanced the results of the algorithms. Among all, LR achieved the highest accuracy of 88.52%. Alshraideh *et al*. [44] aimed to enhance HD prediction using ML models with the HD dataset obtained from the Jordan University Hospital (JUH). To choose features, several ML classifiers, comprising KNN, SVM, NB, DT, and RF were examined using PSO. The findings showed that SVM combined with PSO showed outstanding performance, indicating its efficiency in classifying patients according to their HD risk, reaching an accuracy of 94.3%.

By reviewing the relevant literature, it is clear that ML methods aid in the early identification of HD. However, these methods also have certain drawbacks and problems. The following research gaps were identified:

- Some models are validated with just one dataset.
- In certain cases, the sample size is very small.
- Some studies used a few performance evaluation metrics to assess their models.
- Some studies have not computed the error rates in prediction.
- Some models are not validated using ROC curve.
- Time complexity is sometimes overlooked by researchers.
- Overfitting has been identified in some studies.
- Certain articles only compared the performance of 2 ML classifiers.

Table 1 showcases a variety of ML algorithms utilized by researchers in detecting HD.

Table 1. ML algorithms for HD prediction along with their reference count

ML algorithm	References	Ref. count
LR	[8], [10], [13]–[15], [18], [20], [22]–[24], [26], [28], [31]–[33], [35], [40], [41], [43]	19
KNN	[8], [13]–[16], [21]–[23], [26]–[28], [32], [33], [35], [36], [44]	17
ANN	[8], [13], [14], [34], [37]	5
SVM	[8], [10], [13]–[15], [17], [19], [20], [22], [24], [26]–[28], [31], [33], [35], [37], [41], [43], [44]	22
NB	[8], [10], [11], [13]–[17], [19], [20], [31], [33], [35], [40], [41], [44]	17
DT	[8], [10], [13], [14], [16]–[19], [21]–[24], [27], [31], [33], [35]–[37], [41]–[44]	22
RF	[8]–[10], [13], [15]–[18], [20]–[25], [27], [29]–[31], [33], [35], [36], [40]–[42], [44]	25
GB	[21], [24], [27], [35], [39], [40]	6
XGBoost	[12], [26], [28], [35], [36], [38]	6
MLP	[13], [19], [23], [35], [41]	5
AdaBoost	[20], [21], [23], [24], [36], [37], [40]	7
CNN	[34]	1
ET	[36]	1
SGB	[42]	1

3. MATERIALS AND METHOD

The research methodology employed for conducting the research is outlined in this section. Figure 1 depicts the several processes associated with predicting HD, including: i) selecting the dataset to be used, ii) processing data, iii) the cross-validation, iv) choosing ML methods, v) performing predictions, and v) evaluating performance. The next sub-section goes into further depth about these stages.

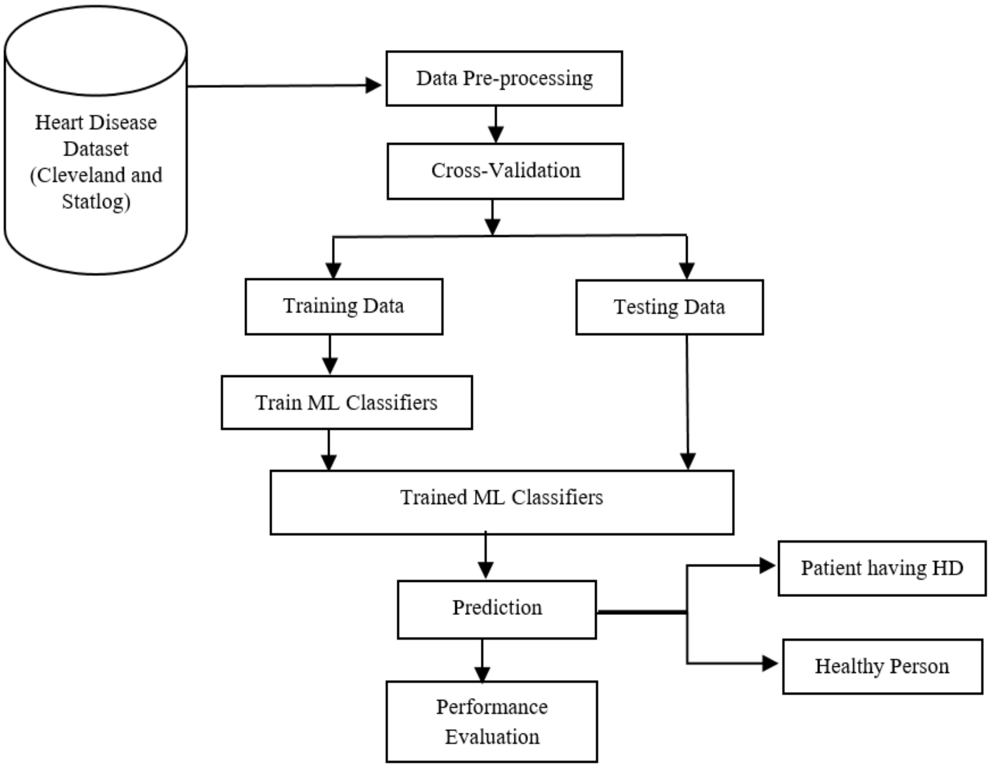

Figure 1. Flow of steps involved in HD prediction

3.1. Dataset

Data is of the utmost importance for ML to produce accurate and reliable results. This analysis used two openly accessible HD datasets from Kaggle: the Cleveland and Statlog (Heart) [45], [46]. These datasets were selected because researchers frequently use them to assess the performance of their HD prediction methods. The Cleveland data has 303 cases, whereas the Statlog dataset includes 270 occurrences.

Each dataset has 14 characteristics, with the initial 13 in a feature type and the last in the target type. Table 2 describes the properties of both datasets, which include the same kind and amount of features.

Table 2. Features information of the Cleveland and the Statlog HD dataset

S.No.	Feature name	Type of data	Explanation	Domain of target attribute
1.	Age	Numeric	Age (years)	29-77
2.	Sex	Categorical	Gender	0: Female 1: Male
3.	Cp	Categorical	Nature of Pain in the Chest	1: Typical angina 2: Atypical angina 3: Non-anginal pain 4: Asymptomatic
4.	Trestbps	Numeric	Resting blood pressure (mm hg)	94-200
5.	Chol	Numeric	Serum cholesterol (mg/dL)	126-564
6.	Fbs	Categorical	Fasting blood sugar > 120 mg/dL	0: False 1: True
7.	Restecg	Categorical	Resting electrocardiogram findings	0: Normal 1: ST-T wave abnormality 2: Probable
8.	Thalach	Numeric	Maximal heart rate	71-202
9.	Exang	Categorical	Exercise-related angina	0: No 1: Yes
10.	Oldpeak	Numeric	Exercise-induced ST depression in comparison to rest	0-6.2
11.	Slope	Categorical	Slope of peak exercise ST segment	1: Upsloping 2: Flat 3: Downsloping
12.	Ca	Categorical	Count of major vessels	1-4
13.	Thal	Categorical	The Thallium imaging	3: Normal 6: Fixed 7: Reversible defect
14.	Target	Categorical	Output variable	0: HD is absent 1: HD is present

3.2. Data pre-processing

The unprocessed data must first be pre-processed before being used with the ML algorithm. Pre-processing transforms less significant information into more relevant data. There are several steps involved in this process, such as gathering data from a database, selecting necessary information, preparing the chosen data, the sampling process, and data conversion. Dealing with missing numbers and eliminating noise and outliers from the data may be necessary to achieve this. It may be challenging for ML algorithms to process incoming data if there are missing values. Consequently, before using any approach, the data must be converted into a structured format. Data preparation is commonly referred to as extract, transform, and load (ETL). The distribution of data is crucial for predictive modeling. Table 3, shows the expected distribution of attribute classes for the two datasets used. This demonstrates that the distribution of the target attribute for both of these datasets is equal, which helps avoid the overfitting issue. In both datasets, there were no missing values found. For the target class, there are five class labels in the original Cleveland dataset, each with an integer value between 0 and 4. The Cleveland dataset mainly attempted to discriminate between the existence of HD with a target possessing values ranging from 1, 2, 3, and 4, and an absence of HD with a value of 0. According to the researchers, the five class features of the target attribute for this dataset can be simplified to two classes i.e. 0 and 1. As a result, the multiclass numbers for its target attribute were transformed into binary numbers by setting every number from 2 to 4 to 1. Thus, the final dataset's diagnostic values are simply 0 and 1, where 0 denotes the absence of HD and 1 denotes its presence. Furthermore, a filtering method known as class balancer was used to ensure every instance in the dataset got equal weight.

Table 3. Distribution of data in both datasets

Dataset (Instances)	Patients having HD (%)	Healthy persons (%)
Cleveland (303)	45.8	54.1
Statlog (270)	44.4	55.5

3.3. Cross-validation

Cross-validation reduces overfitting by evaluating an ML model's performance using unseen data. K-fold cross-validation separates data into k equal-sized folds (in this case, k=10) and uses every single fold as a validation set. The model is trained and evaluated k times, and an unbiased estimate is produced by averaging the performance over all folds. This work splits the dataset into sets for training and testing using a tenfold cross-validation technique.

3.4. Selection of the algorithm

The choice of the algorithm depends on the dataset and prediction type. This study uses Ref. count, a variable tracking the frequency of the algorithms used in previous studies, to select suitable algorithms for analysis. This sub-section examines algorithms with a Ref. count exceeding 6 from Table 1 and discusses them.

- LR: LR is an approach to supervised learning that can be utilized for classification and regression. It is commonly employed in binary classification problems where the outcome variable can be 0 or 1. LR analyses the connection between independent variables and categorizes them into distinct classes using the logistic function, often referred to as the sigmoid function.
- SVM: SVM is a robust supervised learning approach that performs well in both regression and classification applications. The primary objective is to identify the optimum hyperplane in a space with N dimensions that can efficiently divide data points into different classes. The hyperplane's purpose is to maximize the distance amongst points that are closest in each class.
- NB: NB classifiers are probabilistic classifiers that use Bayes' theorem. It is assumed that the presence of a specific attribute in the class does not affect the presence of another attribute in a similar class. It computes the likelihood of an input relating to a given class, assuming feature independence [47].
- KNN: KNN is a non-parametric approach to supervised learning that can be employed for both classification and regression problems. It works by comparing data points to find similarities. The label associated with new data is predicted by evaluating the labeling of the K closest neighbors in the training set. The distance amongst data points is determined utilizing Euclidean, Manhattan, or Minkowski distances.
- DT: DT is a non-parametric supervised learning method used for regression and classification. It uses a hierarchical tree structure with leaf nodes, internal nodes, branches, and a root node. Decisions are made using branches, internal nodes describe dataset properties and leaf nodes display desired outcomes. DT uses a greedy search and divide-and-conquer strategy to find optimal split locations, repeating the top-down dividing process until most records are categorized under specific class labels.
- RF: RF is an ML strategy used for regression and classification. It creates DT during training, each evaluating a random sample of features. This randomization prevents overfitting and improves prediction accuracy. During prediction, the algorithm combines the outputs of all trees through voting or averaging, repeating recursively until most records are categorized under specific class labels [48].
- AdaBoost: AdaBoost involves combining several weak classifiers into one ensemble method to produce a stronger classifier. This algorithm trains and deploys a sequence of trees, implementing boosting. Each classifier improves the classification of samples incorrectly classified by its predecessor. By combining weak classifiers, boosting effectively generates a powerful classifier that categorizes records under specific class labels [49].

3.5. Prediction

AdaBoost, DT, RF, KNN, NB, LR, and SVM are the ML algorithms selected from Table 1. Predictions are generated using these classifiers on both datasets. The target variable with value 0 indicates an absence of HD and value 1 indicates its presence. Each classifier's efficacy is then evaluated using several performance metrics.

3.6. Performance evaluation

To determine how effectively a model operates, it is necessary to employ several evaluation standards that provide a comprehensive picture of its performance. The effectiveness of the chosen classifiers is evaluated using several evaluation measures, comprising MCC, Kappa value, F-measure, ROC area, accuracy, precision, and recall. The metrics are computed utilizing the confusion matrix as a base. The confusion matrix in Table 4 shows both the actual as well as predicted classifications generated by a two-class classifier. This matrix provides insights into the performance of classification systems by investigating the data it contains.

Table 4. The confusion matrix

	Predicted HD patients	Predicted healthy individuals
Actual HD patients	True positive (TP)	False negative (FN)
Actual healthy individuals	False positive (FP)	True negative (TN)

Here, TP denotes the total number of cases accurately identified with HD. FN signifies the total number of individuals having HD who are incorrectly categorized as healthy. TN signifies the number of accurately classified healthy patients. Finally, FP signifies the number of healthy instances that are incorrectly identified with HD. Table 5 provides an overview of the evaluation metrics and their mathematical formulas [50]. These formulas are useful for measuring the performance of ML algorithms in predicting HD.

Table 5. Performance metrics and their mathematical formula

Performance metric	Formula	Description
Accuracy	$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}$	It represents the proportion of accurate predictions amongst all predictions made.
Precision	$Precision = \frac{TP}{TP+FP}$	It measures the accuracy of positive predictions.
Recall or Sensitivity	$Recall = \frac{TP}{TP+FN}$	The accuracy of the model in identifying positive cases among all of the actual positive instances in the dataset.
Specificity	$Specificity = \frac{TN}{TN+FP}$	The accuracy of the model in identifying negative cases among all of the actual negative instances in the dataset.
FP rate	$FP\ Rate = \frac{FP}{FP+TN}$	It reflects the number of cases in the dataset that are incorrectly categorized as positive when they are negative.
F-measure	$F - measure = 2 \times \frac{Recall \times Precision}{Recall + Precision}$	It is a measure of statistical significance that uses a weighted average to combine recall and precision.
MCC	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$	It measures the predictive capacity of a classifier and is represented by values between -1 and +1.
Kappa statistic	$\frac{2 \times ((TP \times TN) - (FP \times FN))}{(TP+FP) \times (TN+FP) + (TP+FN) \times (TN+FN)}$	It is a measure that compares the observed accuracy to the expected accuracy, which is based on random chance.
AUC	$\frac{1}{2}(\frac{FP}{FP+TN} + \frac{TN}{TN+FP})$	It graphically depicts the ratio of true positives vs false positives, with the region located under the ROC curve.

Further, the performance of the classifiers is checked using error rate analysis. For computing the prediction errors, different error rates like mean absolute error (MAE), relative absolute error (RAE), root mean square error (RMSE), and root relative square error (RRSE) are calculated [51]. Table 6 outlines different error rates along with their description.

Table 6. Error rate metrics and their description

Error rate metric	Description
MAE	It is defined as the mean of a dataset's estimated and actual values.
RMSE	It is the basic statistical metric calculated by taking the square root of the average squared difference between expected and observed target values in a dataset.
RAE	It is a ratio-based statistic used to evaluate the efficiency of a model in making predictions.
RRSE	It is defined as the square root of a predictive model's total squared errors normalized by the total squared errors of the basic model.

3.7. Software used

The WEKA, is a publicly accessible ML software application. This platform compromises a Java programming language API that incorporates pre-built algorithms from a certain area and makes the execution of different data analysis methods simpler. It has features for association, rule mining, clustering, regression, classification, feature selection, and data visualization [52]. In this study, WEKA v3.9.6 was employed on an 11th generation "Intel(R) Core(TM) i5-1135G7 @ 2.40 GHz 2.42 GHz" CPU with RAM of 8.00 GB, operating on a 64-bit version of Windows 11.

4. RESULTS AND DISCUSSION

This analytical study introduced two RQs to thoroughly and impartially evaluate the ML algorithms in predicting HD. To address RQ1, a comprehensive examination of various ML predictive algorithms is carried out. To answer RQ2, a framework is presented to determine the most effective ML algorithm out of the chosen algorithms from RQ1. Further, the selected algorithms are applied to two identical structured HD datasets and then each algorithm undergoes a performance evaluation phase.

The study compared the performance of multiple classifiers in predicting HD, unlike some previous studies that compared only two ML classifiers. For experimentation, two balanced and identical HD datasets are used, whereas some earlier studies have used only one dataset. Previous research revealed overfitting issues, but this study utilized cross-validation and balanced datasets to prevent this issue. Some earlier studies used few performance metrics for evaluation and did not compute the error rates. While accuracy is crucial, it's also vital to take into account other crucial metrics into consideration. This study employed several metrics including MCC, kappa value, F-measure, ROC area, accuracy, precision, recall, and different error rates like MAE, RAE, RMSE, and RRSE. This study validates models using the ROC curve, comparing it to some previous studies that did not. This study calculates the time taken in prediction, unlike previous studies which did not consider time complexity. The performance evaluation findings for the ML classifiers are shown in Tables 7 and 8 on the respective datasets. The highlighted text indicates the best outcomes.

Table 7. Performance analysis of Cleveland dataset

ML Algorithm	Accuracy (%)	FP rate	Precision	Recall	F-measure	MCC	ROC area	Kappa value
LR	88.7	0.150	0.888	0.888	0.888	0.738	0.956	0.7378
KNN	87.7	0.154	0.879	0.878	0.878	0.717	0.925	0.7172
SVM	89.4	0.129	0.896	0.894	0.895	0.756	0.882	0.7561
NB	87.4	0.162	0.875	0.875	0.875	0.709	0.946	0.7087
DT	93.7	0.079	0.938	0.937	0.938	0.855	0.967	0.8548
RF	94.0	0.075	0.941	0.941	0.941	0.861	0.984	0.8612
AdaBoost	85.4	0.136	0.869	0.855	0.858	0.687	0.918	0.6795

Table 8. Performance analysis of Statlog dataset

ML Algorithm	Accuracy (%)	FP rate	Precision	Recall	F-measure	MCC	ROC area	Kappa value
LR	88.1	0.143	0.885	0.881	0.883	0.725	0.955	0.7237
KNN	84.0	0.195	0.846	0.841	0.843	0.631	0.866	0.6299
SVM	89.2	0.152	0.892	0.893	0.892	0.743	0.870	0.7434
NB	85.9	0.216	0.857	0.859	0.858	0.659	0.943	0.6577
DT	91.8	0.103	0.919	0.919	0.919	0.806	0.953	0.806
RF	90	0.149	0.899	0.900	0.899	0.760	0.975	0.7594
AdaBoost	85.9	0.113	0.885	0.859	0.864	0.709	0.907	0.6931

The study discovered that for the Cleveland dataset, RF exceeds other classifiers with an accuracy score of 94.0% in Table 7 and its experimental results on WEKAv3.9.6 are shown in Figure 2. With almost the same accuracy of 93.7%, DT performs better after RF. Therefore, it can be concluded that, in terms of accuracy, considering the Cleveland dataset, RF and DT are better choices illustrated in Figure 3. For the statlog dataset, the outcomes revealed that DT exceeds other classifiers, with an accuracy score of 91.8% in Table 8. With an almost identical accuracy of 90% as DT, RF works better after it. The fundamental and practical evaluation metric is accuracy; however, it might not be enough in datasets that are imbalanced and have a predominance of one class over the other. Since both of the datasets used in this research are evenly distributed and balanced, therefore, DT, and RF can be considered as appropriate classifiers in terms of accuracy metrics for both of the datasets. In situations where minimizing false positives is of utmost importance, such as in HD prediction, precision plays a vital role. False positives might cause worry or unneeded medical procedures. A higher level of precision signifies a reduced occurrence of false positives. For the Cleveland dataset, RF has achieved the highest precision of 0.941, followed by DT with a precision of 0.938. With a precision of 0.919 for the statlog dataset, DT offers the highest precision, followed by RF with 0.899. In prediction, sensitivity (recall) plays a critical role in minimizing false negatives to ensure that individuals with HD are accurately identified. In the Cleveland dataset, RF demonstrated the highest sensitivity of 0.941, while DT followed closely behind with a sensitivity of 0.937. Conversely, in the statlog dataset, DT exhibited the highest sensitivity of 0.919, with RF trailing slightly at a sensitivity of 0.900. MCC examines the relationship between actual and predicted values. A strong correlation leads to accurate predictions. The MCC value of a perfect prediction is +1, whereas the MCC value of a completely wrong prediction is -1. Random predictions are implied by a value close to 0. RF had the highest MCC score for the Cleveland dataset, at 0.861, which was followed by DT, which had 0.855. With an MCC value of 0.806, DT had the highest value for the statlog dataset, followed by RF at 0.760. Upon analysis of the Kappa values of the two datasets, it can be observed that RF performed well on the Cleveland dataset (Kappa value: 0.8612) and DT did well on its Statlog dataset (kappa value: 0.806). AUC values that are near to 1 signify an ideal model. A higher AUC value denotes better model performance. An investigation of the ROC levels of both datasets demonstrated that RF does better in comparison to other classifiers, with ROC values of 0.984 and 0.975 for the Cleveland and Statlog datasets, correspondingly shown in Figure 4. Both, DT and RF are shown to have good performance in the performance evaluation stage on both datasets and therefore can be classified as effective classifiers for HD prediction.

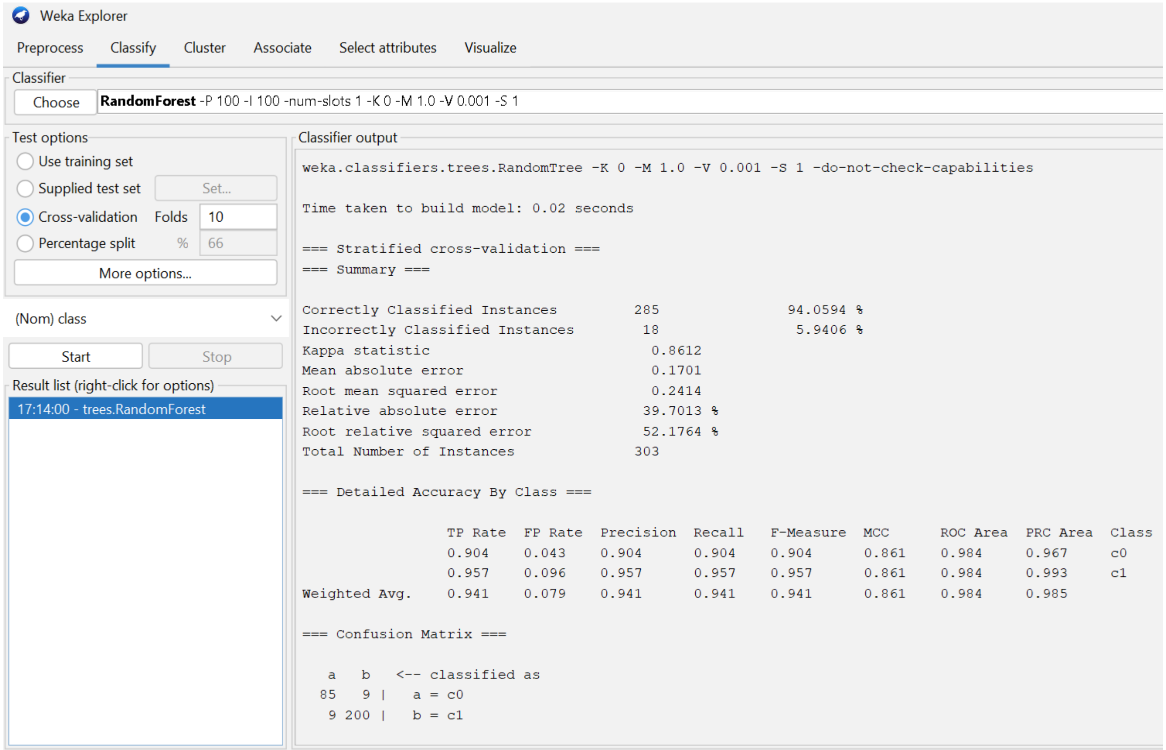

Figure 2. Experimental results of the RF classifier in WEKA v3.9.6 on the Cleveland dataset

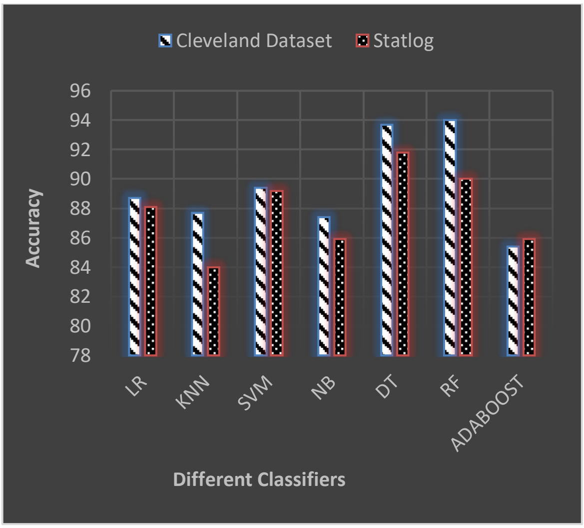

Figure 3. Accuracy of selected classifiers on Cleveland and Statlog dataset

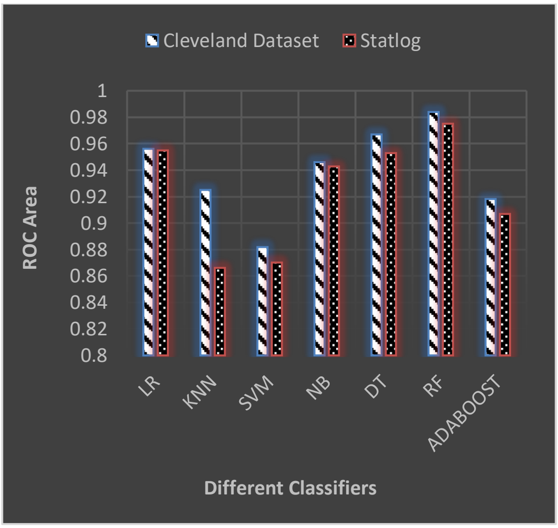

Figure 4. ROC area of selected classifiers on Cleveland and Statlog dataset

Tables 9 and 10 analyzed the error rates associated with each classifier on both datasets. A stronger prediction model is typically indicated by a reduced error rate. To achieve optimal results, the error rate should be minimized. In terms of the MAE values, DT performed the best followed by the SVM classifier for both datasets. As for the RMSE values, both DT and RF classifiers demonstrated similar and comparable minimum values. DT also exhibited a lower RAE percentage on both datasets. In the Cleveland dataset, DT had the lowest RRSE values followed by RF. However, in the statlog dataset, RF produced the lowest RRSE values followed by DT. It is worth noting that the time consumption for each classifier, as shown in Tables 9 and 10, is less than 1, which is a positive indication.

Table 9. Error rate analysis of Cleveland dataset

ML Algorithm	MAE	RMSE	RAE (%)	RRSE (%)	Time (Sec)
LR	0.1412	0.2855	32.9453	61.7197	0.02
KNN	0.1413	0.3041	32.9701	65.7313	0.00
SVM	0.1056	0.325	24.6449	70.2456	0.05
NB	0.211	0.3057	49.2385	66.0695	0.00
DT	0.0783	0.2314	18.274	50.0109	0.00
RF	0.1701	0.2414	39.7013	52.1764	0.08
AdaBoost	0.2041	0.3222	47.6198	69.6514	0.00

Table 10. Error rate analysis of Statlog dataset

ML Algorithm	MAE	RMSE	RAE (%)	RRSE (%)	Time (Sec)
LR	0.1398	0.282	33.2362	61.537	0.02
KNN	0.1857	0.3633	44.1475	79.2813	0.00
SVM	0.1074	0.3277	25.5318	71.5114	0.02
NB	0.2171	0.3109	51.6178	67.8311	0.00
DT	0.0967	0.2628	22.9935	57.3326	0.00
RF	0.1819	0.2595	43.2368	56.6246	0.00
AdaBoost	0.2059	0.3119	48.9476	68.0488	0.00

The study found that DT and RF performed well in assessing the effectiveness as well as the rate of error of the selected classifiers over both datasets, indicating that they are robust classifiers in HD prediction. DT and RF both obtained the highest and almost identical accuracies on both datasets. However, RF has been shown to have a greater ROC value than DT for both datasets. In general, ROC is chosen over accuracy because it is a far better predictor of model performance. This is because ROC takes into account the model's true and false positive rates at various cut-off values. Based on both the ROC curve and accuracy, it is clear from the evaluation and comparison of classifier performance that RF is the better option for classification in Figures 3 and 4. As a result, RF can effectively predict HD on both datasets.

Even with the encouraging outcomes, it's important to acknowledge certain limitations in the research. First, the study mentioned several hybrid models but no tests were carried out using them. Furthermore, the study considered every feature found in the dataset for prediction i.e. no feature selection technique is employed. Lastly, the results are not validated using large and real-world datasets. It would be beneficial to carry out additional research to overcome these issues and get an improved understanding of the potential of ML classifiers for HD prediction in light of these constraints. Hence, to make the models more reliable and universal, and make sure they function well throughout a range of people and situations, future studies will concentrate on creating hybrid models incorporating feature selection and optimization techniques, and further assessing their efficacy using more diverse and large datasets.

5. CONCLUSION

Early diagnosis of HD is critical since it may result in various problems. To automate the identification process, ML predictive algorithms are the best approach. This study examined several ML predictive techniques, chosen based on previous research, including SVM, LR, NB, KNN, DT, RF, and AdaBoost. The experiment was carried out utilizing the Cleveland and Statlog HD datasets provided by Kaggle and implemented using WEKA software. Out of all the classifiers tested, RF performed better for the Cleveland dataset in measures of MCC, ROC area, accuracy, precision, sensitivity, and kappa value. But when it comes to the statlog dataset, RF performed better regarding the ROC area, while DT shows superior accuracy, precision, sensitivity, MCC, and Kappa value. The study additionally examined the error rates related to the selected classifiers. Since ROC is a better predictor for a model’s performance, therefore, it can be concluded that for both the datasets, RF appears to be a more effective classifier for diagnosing HD with an accuracy and ROC values of 94% and 0.984 for Cleveland and 90% and 0.975 for Statlog dataset respectively. Several hybrid models are mentioned in this article, but no tests are carried out using them. Therefore, Future studies will concentrate on building hybrid models employing some feature selection techniques and evaluating their effectiveness with both these datasets, real-world datasets, and models in previous studies for a more comprehensive understanding of the model's performance. This study would aid researchers in developing more robust and generalized HD prediction models and help medical facilities identify HD early on, saving their time as well as effort.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Neha Bhadu	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Jaswinder Singh		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		

C : **C**onceptualization	I : **I**nvestigation	Vi : **Vi**sualization
M : **M**ethodology	R : **R**esources	Su : **Su**pervision
So : **So**ftware	D : **D**ata Curation	P : **P**roject administration
Va : **Va**lidation	O : Writing - **O**riginal Draft	Fu : **Fu**nding acquisition
Fo : **Fo**rmal analysis	E : Writing - Review & **E**diting	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from Kaggle. Restrictions apply to the availability of these data, which were used under license for this study. Data are available https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland with the permission of Cleveland Heart Disease Dataset and https://www.kaggle.com/datasets/ritwikb3/heart-disease-statlog with the permission of Statlog Heart Disease Dataset.

REFERENCES

[1] R. Alanazi, "Identification and prediction of chronic diseases using machine learning approach," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/2826127.

[2] NHLBI, "Know the difference: cardiovascular disease, heart disease, coronary heart disease," *National Heart, Lung, and Blood Institute*, 2021.

[3] R. R. Sarra, A. M. Dinar, and M. A. Mohammed, "Enhanced accuracy for heart disease prediction using artificial neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 375–383, Jan. 2022, doi: 10.11591/ijeecs.v29.i1.pp375-383.

[4] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, 2019, doi: 10.1016/j.imu.2019.100203.

[5] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, "Latest trends on heart disease prediction using machine learning and image fusion," *Materials Today: Proceedings*, vol. 37, pp. 3213–3218, 2021, doi: 10.1016/j.matpr.2020.09.078.

[6] WHO, "Cardiovascular diseases," *World Health Organization*. 2024. Accessed: Mar. 08, 2024. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[7] M. Di Cesare *et al.*, *World heart report 2023: Confronting the world's number one killer*. Geneva, Switzerland: World Heart Federation, 2023.

[8] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, pp. 1–21, Dec. 2018, doi: 10.1155/2018/3860146.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[10] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum, and K. Bashir, "Improving heart disease prediction using feature selection approaches," in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, Jan. 2019, pp. 619–623, doi: 10.1109/IBCAST.2019.8667106.

[11] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives Bayesian," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 292–297, doi: 10.1109/ICOEI.2019.8862604.

[12] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[13] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," *Health and Technology*, vol. 11, no. 1, pp. 87–97, Jan. 2021, doi: 10.1007/s12553-020-00505-7.

[14] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[15] H. K. Thakkar, H. Shukla, and S. Patil, "A comparative analysis of machine learning classifiers for robust heart disease prediction," in *2020 IEEE 17th India Council International Conference (INDICON)*, 2020, pp. 1–6, doi: 10.1109/INDICON49873.2020.9342444.

[16] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, 2020, doi: 10.1007/s42979-020-00365-y.

[17] V. Sharma, S. Yadav, and M. Gupta, "Heart disease prediction using machine learning techniques," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 177–181, doi: 10.1109/ICACCCN51052.2020.9362842.

[18] M. D. A. Hossen *et al.*, "Supervised machine learning-based cardiovascular disease analysis and prediction," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–10, Dec. 2021, doi: 10.1155/2021/1792201.

[19] S. Bashir, A. A. Almazroi, S. Ashfaq, A. A. Almazroi, and F. H. Khan, "A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction," *IEEE Access*, vol. 9, pp. 130805–130822, 2021, doi: 10.1109/ACCESS.2021.3110604.

[20] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, Sep. 2021, doi: 10.1007/s40860-021-00133-6.

[21] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[22] S. E. A. Ashri, M. M. El-Gayar, and E. M. El-Daydamony, "HDPF: Heart disease prediction framework based on hybrid classifiers and genetic algorithm," *IEEE Access*, vol. 9, pp. 146797–146809, 2021, doi: 10.1109/ACCESS.2021.3122789.

[23] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[24] A. Ishaq *et al.*, "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[25] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthcare Analytics*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100016.

[26] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods," *IEEE Access*, vol. 10, pp. 79974–79985, 2022, doi: 10.1109/ACCESS.2022.3191669.

[27] G. N. Ahmad, S. Ullah, A. Algethami, H. Fatima, and S. M. H. Akhter, "Comparative study of optimum medical diagnosis of human heart disease using machine learning technique with and without sequential feature selection," *IEEE Access*, vol. 10, pp. 23808–23828, 2022, doi: 10.1109/ACCESS.2022.3153047.

[28] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi, and Imdadullah, "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151–80173, 2022, doi: 10.1109/ACCESS.2022.3165792.

[29] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "Improving the heart disease detection and patients' survival using supervised infinite feature selection and improved weighted random forest," *IEEE Access*, vol. 10, pp. 67363–67372, 2022, doi: 10.1109/ACCESS.2022.3185129.

[30] D. Cenitta, R. Vijaya Arjunan, and K. V. Prema, "Ischemic heart disease prediction using optimized squirrel search feature selection algorithm," *IEEE Access*, vol. 10, pp. 122995–123006, 2022, doi: 10.1109/ACCESS.2022.3223429.

[31] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A novel study on machine learning algorithm-based cardiovascular disease prediction," *Health & Social Care in the Community*, vol. 2023, pp. 1–10, Feb. 2023, doi: 10.1155/2023/1406060.

[32] T. Ullah *et al.*, "Machine learning-based cardiovascular disease detection using optimal feature selection," *IEEE Access*, vol. 12, pp. 16431–16446, 2024, doi: 10.1109/ACCESS.2024.3359910.

[33] N. Biswas *et al.*, "Machine learning-based model to predict heart disease in early stage employing different feature selection techniques," *BioMed Research International*, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/6864343.

[34] M. S. Al Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, and A. Shaikh, "A robust heart disease prediction system using hybrid deep neural networks," *IEEE Access*, vol. 11, pp. 121574–121591, 2023, doi: 10.1109/ACCESS.2023.3328909.

[35] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective feature engineering technique for heart disease prediction with machine learning," *IEEE Access*, vol. 11, pp. 56214–56224, 2023, doi: 10.1109/ACCESS.2023.3281484.

[36] S. C. Patra, B. U. Maheswari, and P. B. Pati, "Forecasting coronary heart disease risk with a 2-step hybrid ensemble learning method and forward feature selection algorithm," *IEEE Access*, vol. 11, pp. 136758–136769, 2023, doi: 10.1109/ACCESS.2023.3338369.

[37] A. A. Ahmad and H. Polat, "Prediction of heart disease based on machine learning using jellyfish optimization algorithm," *Diagnostics*, vol. 13, no. 14, Jul. 2023, doi: 10.3390/diagnostics13142392.

[38] A. Noor, N. Javaid, N. Alrajeh, B. Mansoor, A. Khaqan, and S. H. Bouk, "Heart disease prediction using stacking model with balancing techniques and dimensionality reduction," *IEEE Access*, vol. 11, pp. 116026–116045, 2023, doi: 10.1109/ACCESS.2023.3325681.

[39] A. Jafar and M. Lee, "HypGB: High accuracy GB classifier for predicting heart disease with HyperOpt HPO framework and LASSO FS method," *IEEE Access*, vol. 11, pp. 138201–138214, 2023, doi: 10.1109/ACCESS.2023.3339225.

[40] N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and

optimization," *Processes*, vol. 11, no. 4, 2023, doi: 10.3390/pr11041210.

[41] M. I. Hossain *et al.*, "Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison," *Iran Journal of Computer Science*, vol. 6, no. 4, pp. 397–417, 2023, doi: 10.1007/s42044-023-00148-7.

[42] A. P. Jawalkar *et al.*, "Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting," *Journal of Engineering and Applied Science*, vol. 70, no. 1, Dec. 2023, doi: 10.1186/s44147-023-00280-y.

[43] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. L. Karthikeyan, R. Manikandan, and A. H. Gandomi, "Classification models combined with Boruta feature selection for heart disease prediction," *Informatics in Medicine Unlocked*, vol. 44, 2024, doi: 10.1016/j.imu.2023.101442.

[44] M. Alshraideh, N. Alshraideh, A. Alshraideh, Y. Alkayed, Y. Al Trabsheh, and B. Alshraideh, "Enhancing heart attack prediction with machine learning: a study at Jordan University Hospital," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/5080332.

[45] Ritwik, "Heart disease cleveland," *Kaggle*. 2023. Accessed: Mar. 19, 2024. [Online]. Available: https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland

[46] Ritwik, "Heart disease statlog," *Kaggle*. 2023. Accessed: Mar. 19, 2024. [Online]. Available: https://www.kaggle.com/datasets/ritwikb3/heart-disease-statlog

[47] S. Krishnan, "Machine learning for biomedical signal analysis," in *Biomedical Signal Analysis for Connected Healthcare*, Elsevier, 2021, pp. 223–264, doi: 10.1016/B978-0-12-813086-5.00006-2.

[48] P. Gupta and D. Seth, "Comparative analysis and feature importance of machine learning and deep learning for heart disease prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 451–459, 2022, doi: 10.11591/ijeecs.v29.i1.pp451-459.

[49] S. S. Devi, V. K. Solanki, and R. H. Laskar, "Recent advances on big data analysis for malaria prediction and various diagnosis methodologies," in *Handbook of Data Science Approaches for Biomedical Engineering*, Elsevier, 2020, pp. 153–184, doi: 10.1016/B978-0-12-818318-2.00006-4.

[50] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges," *Algorithms*, vol. 17, no. 2, Feb. 2024, doi: 10.3390/a17020078.

[51] R. Naseem *et al.*, "Empirical assessment of machine learning techniques for software requirements risk prediction," *Electronics*, vol. 10, no. 2, Jan. 2021, doi: 10.3390/electronics10020168.

[52] WEKA, "Weka 3 - Data mining with open source machine learning software in Java," *Waikato*. 2025. Accessed: Mar. 29, 2024. [Online]. Available: https://ml.cms.waikato.ac.nz/weka

BIOGRAPHIES OF AUTHORS

Neha Bhadu is pursuing her Ph.D. from Guru Jambheshwar University of Science and Technology, Hisar, Haryana in Computer Science and Engineering. She has completed her B.Tech. and M.Tech. in Computer Science and Engineering from Mody University of Science and Technology, Laxmangarh, Rajasthan. Her areas of research include artificial intelligence, machine learning, and wireless sensor networks. She can be contacted at email: nehabhadu100@gmail.com.

Jaswinder Singh is working as a Professor in the Department of Computer Science and Engineering at Guru Jambheshwar University of Science and Technology, Hisar, Haryana. He has teaching experience of more than 20 years and he has published more than 30 research papers in international journals and conferences. He has completed his Ph.D. in Computer Science and Engineering from Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonepat, Haryana, and completed his M.Tech. in Computer Science and Engineering from Kurukshetra University, Kurukshetra, Haryana. His areas of research include machine learning, opinion mining, web information retrieval, search engine optimization, web mining, information processing, information systems, and social network analysis. He can be contacted at email: Jaswinder_singh_2k@rediffmail.com.