The Application of Machine Learning Algorithms to Predict Heart Disease

Hadeel Alsolai(✉)

Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
haalsolai@pnu.edu.sa

Abstract. The Early prediction of heart disease at its primary stage, based on symptoms, poses a significant challenge for healthcare organizations worldwide. Machine learning algorithms play a crucial role in predicting and diagnosing heart disease. This paper empirically investigates the application of various types of single machine learning algorithms—Bayes Network, Neural Network, Decision Tree, JRIP—along with one hybrid algorithm (Vote) for heart disease prediction. The study utilizes the heart failure prediction dataset containing 11 features and 918 records. Additionally, two types of feature selection methods— subset and ranking—are employed, each including two techniques: best first, greedy stepwise, Chi-squared, and information gain, respectively. Our empirical findings reveal that the predictive performance of the Vote algorithm surpasses that of other single algorithms. Furthermore, the Vote algorithm outperforms previous studies in most evaluation measures. Concerning feature selection, all methods consistently choose eight attributes, indicating the exclusion of only three attributes—RestingBP, FastingBS, and RestingECG. However, the removal of these attributes does not significantly impact the performance of the prediction algorithms.

Keywords: Heart Disease · Machine Learning · Prediction

1 Introduction

The heart holds a crucial role in the human body by pumping oxygen-rich blood through a network of veins and arteries to various parts of the body. Therefore, it stands as one of the most essential organs, and any form of abnormalities within it is classified as heart disease [1]. Heart diseases continue to be among the primary causes of worldwide fatalities, presenting a significant challenge in the healthcare sector. According to the World Health Organization, heart disease stands as the leading cause of death worldwide, claiming approximately 17 million lives annually, amounting to 30% of global deaths [2]. Additionally, statistics show that between 17% and 45% of patients with heart disease succumb within the first year, while the remaining individuals pass away within five years [3].

Accurately predicting and detecting heart disease in its early stages holds immense significance for successful interventions and improved patient outcomes. Technological advancements, particularly within machine learning algorithms, present promising

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 97–115, 2025.
https://doi.org/10.1007/978-3-031-96997-3_7

avenues to enhance predictive models and aid in swiftly identifying individuals susceptible to such conditions. Scientists and researchers have invested considerable effort and hard work into conducting several studies utilizing machine learning classifiers to predict heart disease [3–10]. These studies rely on various features, notably gender, heart rate, chest pain type, cholesterol levels, age, and several other attributes. Additionally, several factors play a significant role in elevating the risk of heart disease, including cholesterol levels, diabetes, smoking habits, obesity, family history, and high blood pressure [2]. Nevertheless, manually detecting heart disease based on the aforementioned features poses a challenging task to establish.

This research paper investigates and evaluates the effectiveness of machine learning algorithms in predicting heart disease. Four distinct single algorithms—Bayes Network (BN), Neural Network (NN), Decision Tree (DT), JRIP—alongside the hybrid Vote algorithm, which combines these single algorithms, have been developed. Furthermore, these algorithms are applied to a heart failure prediction dataset comprising 11 features and 918 observations. The study integrates five different datasets related to heart disease, namely Cleveland, Stalog Heart, Long Beach, Switzerland, and Hungarian. Additionally, two types of feature selection methods—subset and ranking—are employed, each utilizing two techniques: best first, greedy stepwise, Chi-squared, and information gain, respectively.

The aim of this paper is to clarify the most effective methodologies that facilitate accurate and early identification of individuals at risk of developing heart disease. This is achieved by expanding upon existing knowledge about heart health through the integration of real-world evidence and information derived from comprehensive analysis. The findings are expected to provide valuable insights for developing robust prediction methods, thereby assisting healthcare professionals in early identification and intervention for individuals prone to heart disease.

Essentially, this research strives to bridge complex machine learning techniques with the critical need for precise, early prediction of heart disease. The ultimate objective is to enable more proactive and personalized healthcare approaches.

The goal of this empirical study based on Goal-Question-Metric (GQM) template [11] as follows:

Evaluate: The detecting abilities of heart disease using machine learning classifiers.

Purpose: Detecting heart disease.

Respect: Precision, Recall, F-Measure, Area under carve (AUC) and Accuracy.

Perspective: Machine learning and healthcare researchers.

Context: Heart disease dataset.

In order to achieve the previous goal, we address the following Research Questions (RQ) in our study:

RQ1) Which attributes turn out to be the most important attributes in predicting heart disease?

RQ2) Does the performance of feature selection increase the accuracy of the prediction algorithms of heart disease?

RQ3) Does the performance of hybrid algorithm (Vote) increase the accuracy of the single algorithms of heart disease?

RQ4) Does the performance of hybrid algorithm (Vote) better than the model used by selected previous studies in predicting heart disease?

The remaining sections of this paper are organized as follows: Sect. 2 provides a related work in the heart disease prediction. Section 3 includes a detailed description of the methodology used in this study. Section 4 contains an experiment set up. Section 5 presents results and discussions. Conclusion and future work are conducted in Sect. 6.

2 Related Work

Extensive researches have been published into the problem of predicting heart disease using a variety of machine learning and features. In this section, we present the summary of the selected studies related to prediction heart disease. We consider a total of eight studies published in the well-known journal and conference proceedings. The selected studies have chosen based on the following criteria: (i) papers targeted to predict heart disease. (ii) papers evaluate and compare machine learning algorithms. (iii) papers used different evaluation measures. (iv) papers published recently (after 2019). The main aim of this selection is to determine main gaps in the previous selected studies.

Also, we aim to compare and evaluate our result with these recent studies.

There is a large volume of published studies compared and evaluated several machine learning algorithms and identified the best algorithm that achieved the highest accuracy to predict heart disease. In 2019, Ali et al. [4] developed a novel hybrid algorithm based on X^2 statical model and deep neural network techniques for the diagnosis of heart disease. The proposed algorithm was implemented to remove irrelevant features using X^2 statical model and diagnose heart disease using DNN.

One year later, Pires et al. [5] performed multiple machine learning algorithms, namely Neural Networks (NN), Decision Tree (DT), K-Nearest Neighbors (KNN), Combined nomenclature (CN), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM). The main findings are that SGD achieved the highest prediction accuracy of heart disease by using 5-fold cross-validation. During the same year, Srivastava and Choubey [6] used well-known machine learning models (i.e., DT, KNN, SVM and Random Forests (RF)) to predict heart disease and compared their results with previous nine studies. They stated that KNN attained the highest accuracy (87%) compared with used models as well as the used model by previous nine studies. Aggrawl and Pal [3] created six different algorithms and found the result obtained by RF, along with applying sequential feature selection is the best that gained (86.67%) in term of accuracy performance. Also, Buettner and Schunter [7] built several algorithms and have the same conclusion, where RF outperformed other algorithms and got (84.44%) in term of accuracy.

In 2021, Rani et al. [8] created multiple single algorithms, namely SVM, Naïve Bayes (NB), Logistic Regression (LG), along with two hybrid algorithms RF and Adaptive Boosting (AdaBoost) for detection heart disease. Among these algorithms, again RF produced the optimum performance in term of accuracy (86.6%). During the same period, Kuruvilla and Balaji [9] predicted the heart disease using four types of machine learning classifiers Multilayer Perceptron (MLP), AdaBoost, NB and SVM. The best result is reported by MLP (84.90%).

A recent study by Pan et al. [10] and Li et al. [12] investigated several single and hybrid machine learning algorithms, and the results obtained in these studies reveal that the hybrid algorithms produced the most outstanding performance.

Table 1 presents a summary of the selected studies related to prediction heart disease. This table illustrates datasets and numbers of features and record used in each study, along with pre-processing, evaluation measures and algorithm used. The information proposed in this table helps us to determine main gaps. Also, this table determines algorithm that achieved the best result to enable compare and evaluate our result with these algorithms.

Table 1. Summary of the selected studies related to prediction heart disease.

Study ID	Author	Ref	Year	Dataset/ Numbers of features and record	Pre-processing	Evaluation measures	Algorithm	Algorithm achieved the best result
S1	Ali et al.	[4]	2019	Cleveland heart disease /13 attributes/ 303 records	X^2 statical model to remove irrelevant features	Accuracy, Sensitivity, Specificity, Matthews correlation coefficient (MCC), Area under curve (AUC) and ROC charts	hybrid model X^2-DNN, ANN and DNN	X^2-DNN (93.33%)
S2	Pires et al.	[5]	2020	Statlog (Heart)/13 attributes / 270 records	-	Accuracy, F1measure, Precision, Recall	NN, DT, KNN, CN, SVM and SGD	SGD (87.69%)
S3	Srivastava and Choubey	[6]	2020	Heart Disease /13 attributes / 303 records	Recover the null values	Accuracy, Sensitivity, Specificity, Precision and F-Measure	DT, KNN, SVM, RF	KN (87%)
S4	Aggrawl and Pal	[3]	2020	Heart Failure Clinical Record Data Set 2020 /13 attributes / 299 records	Sequential Feature Selection	Accuracy, Re-call, Precision, AUC and F-Measure	LDA, RF, GBC, DT, SVM, and KNN	RF (86.67%)

(*continued*)

Table 1. (*continued*)

Study ID	Author	Ref	Year	Dataset/ Numbers of features and record	Pre-processing	Evaluation measures	Algorithm	Algorithm achieved the best result
S5	Buettner and Schunter	[7]	2020	Cleveland heart disease /13 attributes / 303 records	NA	Confusion matrix and accuracy	RF	RF (84.44%)
S6	Rani et al.	[8]	2021	Cleveland heart disease /13	Feature selection (genetic	Accuracy, Sensitivity,	SVM, NB,	RF (86.6%)
				attributes / 303 records	algorithm and recursive feature) and sampling (SMOTE)	Specificity, Precision and FMeasure	LR, RF, and AdaBoost	
S7	Kuruvilla and Balaji	[9]	2021	Framingham heart disease /15 attributes / 4241 records	Feature selection (Correlation Based Feature Selection, Principal Component Analysis)	Accuracy, Precision, Recall, F-Measure and ROC	MLP, AdaBoost, NB and SVM	MLP (84.90%)
S8	Pan et al.	[10]	2022	Cleveland heart disease /13 attributes / 303 records	-	Precision, Recall, F-Measure, Accuracy and AUC	Gradient Boosting, XGBoost, AdaBoost, CatBoost, ANN, RF, SVM, DT and LR	Hybrid (SVM + AdaBoost) 86.88%
S9	Li et al.	[12]	2024	Coronary heart disease/ 7291 records	-	Precision, Recall, F-Measure, Accuracy and AUC	deep neural network, a recurrent neural network model hybrid model	Hybrid model 82.8%

** DNN: Deep Neural Network, NN: Neural Networks, DT: Decision Tree, KNN: K – Nearest Neighbor, CN: Combined nomenclature, SVM: Support Vector Machine, SGD: Stochastic Gradient Descent, RF: Random Forest, LDA: Linear Discriminant Analysis, GBC: Gradient Boosting Classifier, NB: Naïve Bayes, LR: Logistic Regression, AdaBoost: Adaptive Boosting, MLP: Multilayer Perceptron, ANN: Artificial Neural Network, XGBoost: Extreme Gradient Boosting.

The main gaps from aforementioned table are determined as follows:

- Gap 1: Most of the studies have been used datasets that include few numbers of records (e.g., 303 records).
- Gap 2: A limited number of studies have been performed features selection techniques.
- Gap 3: Relatively few studies have been created hybrid algorithms, whereas several studies have been created single algorithms.

Therefore, this study seeks to formulate methodology which will assist to address these research gaps. The methodology used in this paper will be proposed in the next section.

3 Methodology

The methodology is prepared according to the fundings obtained in the selected studies related to prediction heart disease in the previous section. Consequently, three procedures were performed to resolve the identified gaps in section 2 as follows:

- To resolve **Gap 1:** We integrate five different heart disease datasets, namely Cleveland (303 records), Stalog Heart (270 records), Long Beach (200 records), Switzerland (123 records), and Hungarian (294 records).
- To resolve **Gap 2**: We perform two types of features selection methods (subset and ranking) and each types includes two techniques: Subset: Best first and Greedy Stepwise (forwards) and Ranking: Chi-squared and Information Gain. Therefore, the total of features selection techniques is four types, and we take the *intersection* between them (attributes selected by all four types) (see Fig. 1).
- To resolve **Gap 3**: We create Vote algorithm that considered hybrid algorithm and include four single algorithms from different types: BayesNet, Neural Network, decision tree, JRIP.

The paper aims to increase the prediction accuracy of heart disease by comparing between the following:

1. Models created using feature selection and without feature selection.
2. Single algorithms (i.e., BayesNet, Neural Network, decision tree, JRIP) and hybrid algorithm (i.e., Vote).

Figure 2 illustrates methodology of prediction heart disease that used in this study. The methodology includes the following steps:

1. Integrate five heart disease datasets. The total records become 1190 records.
2. Remove duplicated records. 272 records were removed, so the total records become 918 records.
3. Carry out two datasets: without applying feature selection that include 11 attributes and with applying feature selection that include 8 attributes. Four types of features selection techniques were applied and the intersection between them was performed.
4. Create four single algorithms from different types, and create hybrid algorithm from these four single algorithms.

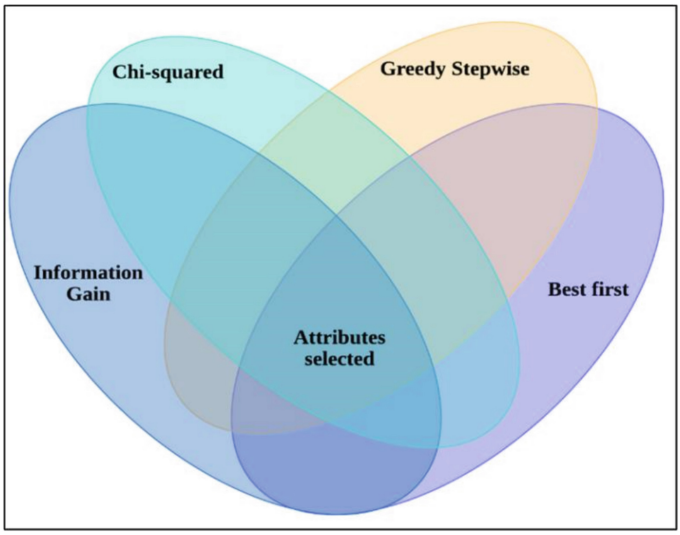

Fig. 1. Method of attributes selected.

4 Experiment Set Up

4.1 Dataset

The heart failure prediction dataset is used in this study and can be obtained from Kaggle from the following link: https://www.kaggle.com/fedesoriano/heart-failure-prediction.

This dataset integrates five different heart disease datasets, namely Cleveland (303 observations), Hungarian (294 observations), Switzerland (123 observations), Long Beach (200 observations), Stalog Heart (270 observations). These datasets Every dataset can be download using the index of heart disease datasets from UCI Machine Learning Repository on the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/.

The total of the heart failure prediction dataset is 1190 observations. However, duplicated observations were removed that include 272 observations, so the final dataset includes 918 observations. The fact that worth mention is that the dataset has no missing values.

4.2 Independent Variables

The heart failure prediction dataset includes 11 independent variables, namely Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope and one dependent variables Heart Disease. Table 2 provides the description of independent variables of heart disease dataset.

4.3 Dependent Variable

Heart disease is the dependent variable in the heart failure prediction dataset. The machine learning algorithms used in this paper predict heart disease (if a patient has

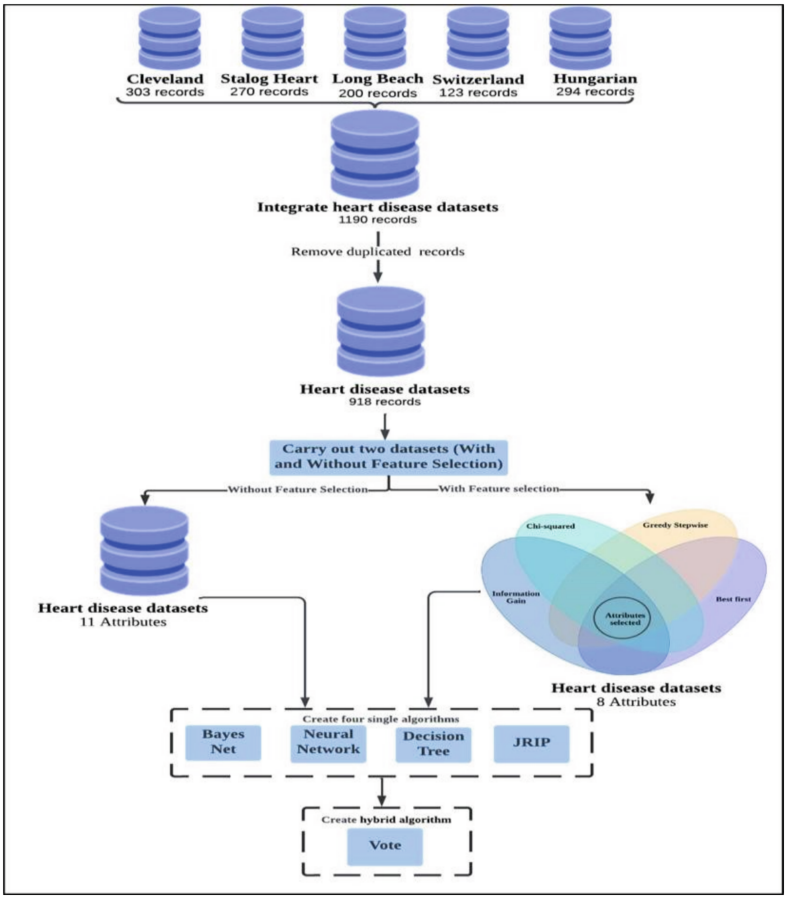

Fig. 2. Methodology of prediction heart disease.

heart disease (True) or not (False). The dataset is considered classification because the dependent variable is Boolean type, and its value class (True) or (False). Figure 3 illustrates the number of values in the dependent variable (heart disease). This figure indicates the dependent variable contain 508 True values and 410 False values. Therefore, the dependent variable is relatively balance and applying sampling techniques to balance datasets is not necessary.

4.4 Feature Selection Methods

Feature selection methods are commonly performed in the dataset to reduce their size and to adjust to the best group for the analysis [13]. Feature selection methods are in general classified into subset and ranking methods. Subset method is identified best suit of feature subset related to the dependent variable [14]. Ranking method is performed in two steps: (i) features are arranged according to a specific statistical measure. (ii) features are selected and sorted according to the highest ranking [15].

- **Best first** can be begun with an empty set of features and identifies the best features for a given dataset by searching forward. In contrast, this method can be begun with a full set of features and identifies the best features for a given dataset by searching backwards. Besides, this method can be begun with any point by searching in both directions to consider all possible features additions and deletions at a given point [16].

Table 2. The description of independent variables of heart disease dataset.

Name	Description	Distinct	Data type	Data values
Age	Age of the patient in terms of years	-	Numeric	Range from 28 to 77
Sex	Sex of the patient that classified into Male (M) and Female (F)	Two	Nominal	M or F
ChestPainType	Chest pain type that classified into Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP) and Asymptomatic (ASY)	Four	Nominal	TA, ATA, NAP or ASY
RestingBP	Resting blood pressure that measures by (mm /Hg)	NA	Numeric	Range from 0 to 200
Cholesterol	Serum cholesterol that measures by mm/dl	NA	Numeric	Range from 0 to 603
FastingBS	Fasting blood sugar that that classified into 1 if FastingBS >120 mg/dl and 0 otherwise	Two	Numeric	1 or 0
RestingECG	Resting electrocardiogram results that classified into Normal (Normal), having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) (ST) and showing probable or definite left ventricular hypertrophy by Estes' criteria (LVH)	Tree	Nominal	Normal, ST or LVH
MaxHR	Maximum heart rate	NA	Numeric	Range from 60 and 202
ExerciseAngina	Exercise-induced angina that classified into Yes (Y) and No (N)	Two	Nominal	Y or N

(*continued*)

Table 2. (*continued*)

Name	Description	Distinct	Data type	Data values
Oldpeak	ST segment to measure the depression	NA	Numeric	Range from −2.6 to 6.2
ST_Slope	Slope of the peak exercise ST segment that classified into upsloping (Up), flat (Flat) and down sloping (Down)	Tree	Nominal	Up, Flat and Down

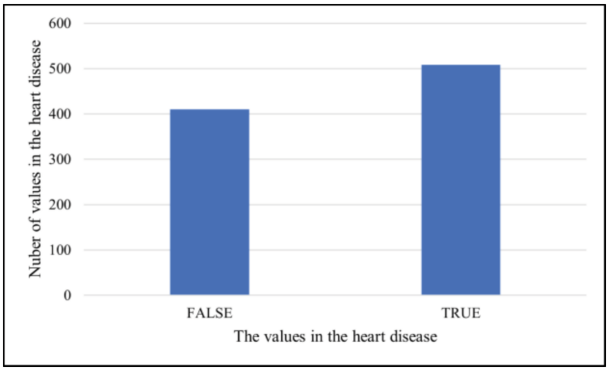

Fig. 3. The number of values in the heart disease.

- **Greedy stepwise** can be started with an empty set of features and identifies the best features for a given dataset by searching forward or backwards. Moreover, this method can be started with a full set of features by doing the same search (i.e. forward or backwards). The search terminates when the addition or deletion of any remaining features effect negatively on the overall performance [17]. The main difference between greedy stepwise and best first is that greedy stepwise can be searched either forward or backward, whereas the best first can be searched in both directions.
- **Chi-squared** aims to select attribute that is highly effect on the dependent variable. It is calculated by testing the independence of two attributes, which are each dependent variable and independent variable. Chi-squared is used only for categorical attributes to determine whether the association between these attributes of the sample would impact their real association in the population [18].
- **Information gain** compares and evaluates the gain of each attribute with the dependent variable. It is calculated how much the amount of information produced by the attributes by measuring the reduction in the surprise. This is performed by dividing dataset according to a specific value of a random variable [19].

4.5 Machine Learning Algorithms

To classify predict if a person has heart disease or not, we create four single well-known algorithms, namely Bayes Net, Neural Network, J48, JRIP, along with one hybrid algorithm (i.e., Vote) that integrate these single algorithms. This section briefly discusses the machine learning algorithms that perform.

- **Bayes Network** performs several quality measures and search algorithms. It provides a collection of random variables and their conditional dependence structure by using a directed acyclic graph. Bayes Network includes information about probability relationships between variables and base class that represents various data structures, including conditional probability, network structure, distributions. Bayesian network is a probabilistic graphical algorithm that based on the Bayes theorem as the following equation [20]:

$$(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

- **Neural Network** has various elements, namely *Input Layers*, which the input variables that provide into the network. *Hidden Layers,* which are the parameters layers that train and *Output Layer*, which are the output of the model. The structure of neural network contains activation function that determine which neurons will be activated and which variables passed to the next layers. Also, the structure of neural network includes cost function that to compute loss of the true and predict results, where neural network aims to reduce this loss [21].
- **Decision tree** provides several decision trees and sets of rules. This performs by dividing the training data into many subsets and each subset has one class. One of the main advantages of the decision is pruned that improve the prediction accuracy and optimize the size of the trees. Pruned replaces a near-leaf node with its child leaves or by adding a smaller sub-tree. Furthermore, the pruned is very beneficial techniques to deal with overfitting dataset or complex rules by decreasing the number of nodes in the tree as well as decreasing the prediction error [22].
- **JRIP** was proposed by William W. Cohen as a new version of IREP, which is abbreviation for Incremental Reduced Error Pruning. is considered rule machine learning algorithm that builds a propositional rule learner. It creates repeated incremental pruning to generate error reduction. JRIP perform better than C4.5 in term of the prediction accuracy due to the running more effectively. Also, another advantage is allowing to increase scalability with large datasets [23].
- **Vote** integrates multiple different single algorithms to produce single result and take advantage from this combination. The combination rule used in Vote is majority voting, which only works with classification problem. Each single algorithm predicts one class in the dependent variable in the test sample. The class which was predicted by the most of algorithms will then be chosen as the result of the vote algorithm [24].

4.6 Performance Measures

In this study, a variety of different performance measures are applied to evaluate and compare the machine learning algorithms in heart disease prediction. We used popular

and well-known measures that are suitable for classification problem and calculated from the confusion matrix as shown in Table 3.

Table 3. Confusion matrix

	Predicted Class
actual class	heart disease (yes) heart disease (No)
heart disease (Yes)	True Positive (TP) False Negative (FN)
heart disease (NO)	False Positive(FP) True Negative (TN)

The following are the performance measures used in this study and computed from the confusion matrix proposed in Table 3:

Precision is the proportion of positive observations that are accurately corrected as shown in Eq. 2:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}). \tag{2}$$

Recall is the proportion of the positive observations that are accurately corrected to all the observations in the real class-yes as shown in Eq. 3:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}). \tag{3}$$

F-Measure is calculated by using Precision and Recall. Therefore, this measure uses both false positives and false negatives as shown in Eq. 4:

$$\text{F} - \text{Measure} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}). \tag{4}$$

AUC is a probability curve that plots the TP rate, which is a synonym for recall as shown in Eq. 5:

$$\text{TP rate} = \text{TP}/(\text{TP} + \text{FN}). \tag{5}$$

against FP rate at different threshold values as shown in Eq. 6:

$$\text{FP rate} = \text{FP}/(\text{FP} + \text{TN}). \tag{6}$$

It is computed by dividing the area between the two points, and then these areas are integrated to gain the total area.

Accuracy is the percentage of correct classified observations divided by the total number of the observations in the confusion matrix as shown in Eq. 7:

$$\text{Accuracy} = (\text{TN} + \text{TP})/(\text{TN} + \text{FN} + \text{FN} + \text{FP}). \tag{7}$$

The higher of these performance measures (i.e., Precision, Recall, F-Measure, AUC and Accuracy) indicate the better the performance of the machine learning algorithms.

5 Results and Discussion

RQ 1 Which attributes turn out to be the most important attributes in detecting heart disease?

Four types of features selection techniques were performed, which are Best first and Greedy Stepwise (forwards) techniques that belong to subset method, along with Chisquared and Information Gain techniques that belong to ranking method. After that, intersection was applied to determine attributes that selected by all four techniques.

Table 4 illustrates the results of feature selection using subset method that include two techniques: Best first and Greedy Stepwise (forwards). The overall results of two techniques demonstrate that eight attributes were selected and only three attributes were removed from the selection, namely RestingBP, FastingBS and RestingECG.

Table 4. Results of feature selection using subset method.

Attribute	Best first	Greedy Stepwise (forwards)
Age	✓	✓
Sex	✓	✓
ChestPainType	✓	✓
RestingBP	×	×
Cholesterol	✓	✓
FastingBS	×	×
RestingECG	×	×
MaxHR	✓	✓
ExerciseAngina	✓	✓
Oldpeak	✓	✓
ST_Slope	✓	✓

Table 5 shows the results of feature selection using ranking method that include two techniques: Chi-squared and Information Gain. It is apparent from this table that all the attributes were selected except three attributes FastingBS, RestingBP and RestingECG. Data from this table can be compared with the data in Table 3 which shows that the results of feature selection using either subset method or ranking method are the same. To answer **RQ1,** eight attributes turn out to be the most important attributes in detecting heart disease, which are ST_Slope, ChestPainType, ExerciseAngina, Oldpeak, MaxHR, Cholesterol, Age and Sex, whereas three attributes turn out to be the least important attributes in detecting heart disease (i.e., FastingBS, RestingBP and RestingECG). Therefore, these three were removed from of heart disease dataset.

RQ2) Does the performance of feature selection increase the accuracy of the prediction algorithms of heart disease?

Figure 4 presents the results of algorithms with applying feature selection methods (after removing three attributes), along without applying feature selection methods

Table 5. Results of feature selection using ranking method.

Rank Number	Attribute	Chi-squared	Information Gain	Selected Attribute
1	ST_Slope	355.9184	0.29932	✓
2	ChestPainType	268.0672	0.22504	✓
3	ExerciseAngina	224.2809	0.18997	✓
4	Oldpeak	190.1125	0.16143	✓
5	MaxHR	155.8043	0.12748	✓
6	Cholesterol	94.2807	0.08337	✓
7	Age	87.0136	0.0696	✓
8	Sex	85.6463	0.06849	✓
9	FastingBS	65.5861	0.05488	×
10	RestingBP	19.3133	0.01552	×

(including whole eleven attributes). The results indicate that applying feature selection methods on the algorithms produced either worsen performance or the similar performance. To answer **RQ2**, applying feature selection does not increase the accuracy of the prediction algorithms of heart disease.



Fig. 4. Results of algorithms with and without feature selection methods.

RQ3) Does the performance of hybrid algorithm (Vote) increase the accuracy of the single algorithms of heart disease?

Figure 5 and Fig. 6 provide the results of hybrid algorithm (Vote) that integrates four single algorithms (i.e., BayesNet, Neural Network, decision tree, JRIP) with applying feature selection and without applying feature selection; respectively. In both figures, the performances of single algorithms (i.e., BayesNet, Neural Network, decision tree, JRIP) were compared against hybrid algorithm (Vote). In order to response RQ3, taken together, these results show that hybrid algorithm (Vote) increase the accuracy of the single algorithms of heart disease across all evaluation measures (i.e., precision, recall, F-measure, AUC and accuracy). Therefore, this study recommends using hybrid algorithm (Vote) to predict heart disease.

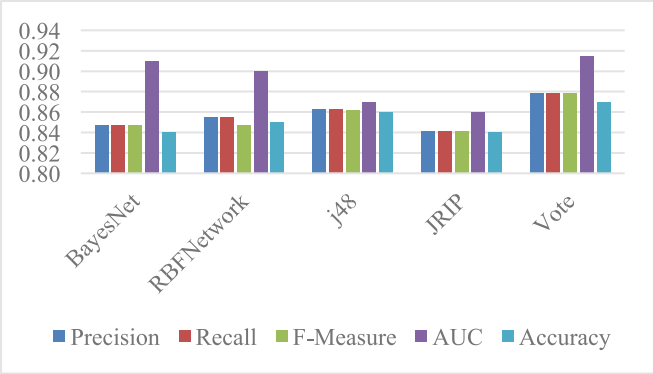

Fig. 5. Results of hybrid algorithm (Vote) with feature selection.



Fig. 6. Results of hybrid algorithm (Vote) without feature selection.

RQ4) Does the performance of hybrid model (Vote) better than the algorithm used by selected previous studies in predicting heart disease?

Five evaluation measures, namely accuracy, AUC, precision, recall and F1-measure were used to evaluate different machine learning algorithms in the current study. We extract these evaluation measures from each selected previous study to compare the performance of hybrid algorithm (Vote) used in the current study with the best algorithm in the selected previous studies. However, some studies did not provide the results of all these measures, such as S1 and S5.

In term of accuracy, the current study reached to 87%, which is considered better than all studies except two studies (S2 and S8) that reported a little bit higher or the same result; respectively. However, there are much higher difference between the accuracy in the current study and S1 that reported 91.57%, which is (4.57%) higher than current study. However, the current study achieved better performance than S1 in term of AUC. Also, the fact that worth mention is that S1 used deep learning algorithm that is much more time-consuming than simple algorithm and it is not recommended for small datasets.

Regarding AUC, the current study and S8 had the best AUC value (92%), where AUC is considered more reliable and informative measure for the prediction problem. In term of the remaining measures, the current study performed better performance than other studies. Furthermore, the current study achieved superior results compared to recent studies S9 published in 2024, across all evaluation measures. To answer **RQ4**, the performance of the current study is better than the previous selected studies in most evaluation measures (See .Table 6).

Table 6. Results of algorithm in the selected previous studies and current study.

Study ID	Ref	Algorithm	Best Algorithm	Accuracy	AUC	Precision	Recall / Sensitivity	F1Measure
S1	[4]	hybrid model X^2-DNN, ANN and DNN	X^2-DNN	**91.57%**	88%	--	87.80%	-
S2	[5]	NN, DT, KNN, CN, SVM and SGD	SGD with 5-fold cross validation	87.69%	-	87.67%	87.69%	86.78%
S3	[6]	DT, KN, SVM, RF	KN	87%	88.5%	83%	73%	77%
S4	[3]	LDA, RF, GBC, DT, SVM, and KN	RF	86.67%	73%	82%	82%	81%
S5	[7]	RF	RF	84.44%	-	-	-	-
S6	[8]	SVM, NB, LR, RF, and adaboost	RF	86.6%	-	85.71%	84.14%	81.50%
S7	[9]	MLP, Ada-Boost, NB and SVM	MLP	84.90%	62%	78%	79%	66%

(*continued*)

Table 6. (*continued*)

Study ID	Ref	Algorithm	Best Algorithm	Accuracy	AUC	Precision	Recall / Sensitivity	F1Measure
S8	[10]	Gradient Boosting, XGBoost, AdaBoost, CatBoost, ANN, RF, SVM, DT and LR	Hybrid (SVM + AdaBoost)	86.88%	**92%**	85.18%	85.18%	85.18%
S9	[12]	deep neural network, a recurrent neural network model hybrid model	Hybrid model	82.8%	80%	87.08%	88.57%	87.82%
Current study		BayesNet, Neural Network, decision tree, JRIP and hybrid model (Vote)	Hybrid model (Vote)	87%	**92%**	**88%**	**88%**	**88%**

6 Conclusion

This paper has empirically investigated the application of four single algorithms, along with namely Bayes Network, Neural Network, Decision Tree, JRIP, along with one hybrid algorithm (Vote) in predicting heart disease. These algorithms have performed on the heart failure prediction dataset that include 11 features and 918 records. Furthermore, the application of features selection methods has explored including subset and ranking methods and each types includes two techniques, best first, greedy stepwise, Chi-squared and information gain; respectively.

This study presents interesting finding and empirical evidence for both Machine learning and healthcare researchers. The result of our empirical study shows that the prediction performance of Vote algorithm is better than other single algorithms. Furthermore, it was observed that the performance of Vote algorithm outperforms the previous

selected studies in most evaluation measures. Regarding the application of feature selection, eight attributes have selected by all methods, which means only three attributes have removed from the selection, namely RestingBP, FastingBS and RestingECG. The deleting of these attribute does not really affect on the performance of the prediction algorithms.

Several important limitations need to be considered. First, the dataset comprising "11 features and 918 records", may not fully represent predicting heart disease, suggesting that larger datasets could strengthen the algorithms. Second, the algorithms were used without parameter tuning. Applying structured hyperparameter optimization in future studies could improve algorithm performance.

Acknowledgment. The authors gratefully acknowledge the financial support provided by the Graduate Studies and Scientific Research Vice-Rectorate at Princess Nourah bint Abdulrahman University, through the Management of Conferences and Seminars. Thanks are also extended to the Artificial Intelligence in Healthcare Researcher Chair at Princess Nourah bint Abdulrahman University.

References

1. Kumar, R., Rani, P.: Comparative analysis of decision support system for heart disease. Adv. Math. Sci. J. **9**(6), 3349–3356 (2020)
2. Latha, C.B.C., Jeeva, S.C.: Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked **16**, 100203 (2019). https://doi.org/10.1016/j.imu.2019.100203
3. Aggrawal, R., Pal, S.: Sequential feature selection and machine learning algorithm based patient's death events prediction and diagnosis in heart disease. SN Comput. Sci. **1**(6), 1–16 (2020)
4. Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., Khan, J.A.: An automated diagnostic system for heart disease prediction based on X^2 statistical model and optimally configured deep neural network. IEEE Access **7**, 34938–34945 (2019). https://doi.org/10.1109/ACCESS.2019.2904800
5. Pires, I.M., Marques, G., Garcia, N.M., Ponciano, V.: Machine learning for the evaluation of the presence of heart disease. Procedia Comput. Sci. **177**, 432437 (2020)
6. Keshav Srivastava, D.K.C.: Heart disease prediction using machine learning and data mining. Int. J. Recent Technol. Eng. (IJRTE) **9**(1), 212–219 (2020)
7. Buettner, R., Schunter, M.: Efficient machine learning based detection of heart disease. In: 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom), 14–16 Oct. 2019, pp. 1–6 (2019). https://doi.org/10.1109/HealthCom46333.2019.9009429
8. Rani, P., Kumar, R., Ahmed, N.M., Jain, A.: A decision support system for heart disease prediction based upon machine learning. J. Reliab. Intell. Environ. **7**(3), 263–275 (2021)
9. Kuruvilla, A.M., Balaji, N.: Heart disease prediction system using correlation based feature selection with multilayer perceptron approach. IOP Conf. Ser. Mater. Sci. Eng. **1085**(1), 012028 (2021)
10. Pan, C., Poddar, A., Mukherjee, R., Ray, A.K.: Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction. Biomed. Signal Process. Control **76**, 103666 (2022)

11. Basili, V.R., Rombach, H.D.: The TAME project: Towards improvement oriented software environments. IEEE Trans. Software Eng. **14**(6), 758–773 (1988)
12. Li, F., Chen, Y., Xu, H.: Coronary heart disease prediction based on hybrid deep learning. Rev. Sci. Instrum. **95**(1) (2024). https://doi.org/10.1063/5.0172368
13. Jović, A., Brkić, K., Bogunović, N.: A review of feature selection methods with applications. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1200–1205. IEEE (2015)
14. Shahana, A.H., Preeja, V.: Survey on feature subset selection for high dimensional data. In: 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 18–19 March 2016, pp. 1–4 (2016). https://doi.org/10.1109/ICCPCT.2016.7530147
15. Cilia, N.D., De Stefano, C., Fontanella, F., Scotto di Freca, A.: A ranking-based feature selection approach for handwritten character recognition. In: Pattern Recognition Letters, vol. 121, pp. 77–86 (2019). https://doi.org/10.1016/j.patrec.2018.04.007
16. Lei, X., Pingfan, Y., Tong, C.: Best first strategy for feature selection. In: [1988 Proceedings] 9th International Conference on Pattern Recognition, 14 May–17 Nov. 1988, vol. 2, pp. 706–708 (1988). https://doi.org/10.1109/ICPR.1988.28334
17. Caruana, R., Freitag, D.: Greedy attribute selection. In: Machine Learning Proceedings 1994, pp. 28–36. Elsevier (1994)
18. Haryanto, A.W., Mawardi, E.K., Muljono: Influence of word normalization and chi-squared feature selection on support vector machine (SVM) text classification. In: 2018 International Seminar on Application for Technology of Information and Communication, 21-22 Sept. 2018, pp. 229–233 (2018). https://doi.org/10.1109/ISEMANTIC.2018.8549748
19. Gao, K., Khoshgoftaar, T.M., Wald, R.: Combining feature selection and ensemble learning for software quality estimation. In: International Florida Artificial Intelligence Research Society Conference, pp. 47–52 (2014)
20. Yang, X.-S.: Introduction to Algorithms for Data Mining and Machine Learning. Academic press (2019)
21. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. **79**(8), 2554–2558 (1982)
22. Drazin, S., Montag, M.: Decision tree analysis using WEKA. In: Machine Learning Project II, University of Miami, vol. 1, no. 3 (2012)
23. Cohen, W.W.: Fast effective rule induction. In: Machine Learning Proceedings 1995, pp. 115–123. Elsevier (1995)
24. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley (2014)