



A classification and regression tree algorithm for heart disease modeling and prediction

Mert Ozcan^{*}, Serhat Peker

Department of Management Information Systems, İzmir Bakircay University, İzmir, Turkey

ARTICLE INFO

Keywords:

Machine learning
Data mining
Decision trees
Classification and regression trees
Decision rule
Predictive analytics

ABSTRACT

Heart disease remains the leading cause of death, such that nearly one-third of all deaths worldwide are estimated to be caused by heart-related conditions. Advancing applications of classification-based machine learning to medicine facilitates earlier detection. In this study, the Classification and Regression Tree (CART) algorithm, a supervised machine learning method, has been employed to predict heart disease and extract decision rules in clarifying relationships between input and output variables. In addition, the study's findings rank the features influencing heart disease based on importance. When considering all performance parameters, the 87% accuracy of the prediction validates the model's reliability. On the other hand, extracted decision rules reported in the study can simplify the use of clinical purposes without needing additional knowledge. Overall, the proposed algorithm can support not only healthcare professionals but patients who are subjected to cost and time constraints in the diagnosis and treatment processes of heart disease.

1. Introduction

Heart disease is a particular abnormal condition that covers disorders occurring in the heart and affects the bloodstream. It may vary from blood vessel disease to heart rhythm problems [1]. Healthcare professionals diagnose heart disease from both history of patients and medical examinations such as blood pressure, blood sugar, or cholesterol. Besides, advanced medical analyzes such as electrocardiogram (ECG), exercise stress tests, X-rays, echocardiogram, coronary angiography, radionuclide tests, MRI scans, and CT scans may assist to diagnose whether one has a heart disease [2]. Applying appropriate treatment together with a detection of heart disease plays an important role in preventing the increasing number of deaths and reducing the risk of heart disease.

Machine learning is an increasingly important area in early detection of diseases. It focuses on making inferences in line with new information by detecting patterns hidden in observations [3]. For example, as heart-related diseases take approximately 17.9 million lives per year, they are known as one of the most fatal diseases among all diseases in the world [4]. Since early detection of heart disease may decrease the death rate; many researchers who come from many backgrounds have been investigating the issue by using this approach.

Recent developments in the field of machine learning have led to a renewed interest in determining whether a patient is likely to be diagnosed with heart disease [5]. In order to perform machine learning on heart disease detection in the early phase, as one of the classification techniques, decision trees are widely used [6]. In clinical settings,

decision trees have strengths among other classification techniques; due to their understandability for decision makers and suitability for classifying unseen data [7]. The more new data are produced, the higher different versions of decision trees are successfully applied in this domain [8,9]. Despite the existence of studies in different application domains which focus mainly on prediction by using decision trees, very little research on modeling and extracting rule sets with graphical representations, which may simplify the decision-making; specifically, on heart disease sub-domain.

To make up for aforementioned shortcomings, this study seeks to model and predict heart disease. In this study, a comprehensive data set consisting of five data sets with 1190 observations and eleven features has been used. Firstly, in order to perform the model with more accurately and effectively, the data were pre-processed. Secondly, Classification and Regression Tree (CART) algorithm was proposed and examined for both detecting heart disease with high accuracy and putting forward the graphical representation of rule inferences of the model examined. Different from most of the studies, the results and implications of this study may be quite easily incorporated into clinical decision support systems in the future, thanks to the given rule sets and modeling results. Overall, the main motivation of this study is to support healthcare professionals and researchers working on the diagnosis and treatment of heart disease by a simple way.

The rest of the paper is presented as follows. There is a brief overview of the studies that were conducted in this field in Section 2. The data set, methodology and modeling of decision tree infrastructure

^{*} Corresponding author.

E-mail addresses: ozcanmert87@gmail.com (M. Ozcan), serhat.peker@bakircay.edu.tr (S. Peker).

are described in Section 3. Then, the experimental results are presented in Section 4. Lastly, Section 5 concludes the study and proposes research works for the future.

2. Literature review

Among the deaths worldwide, heart disease-related ones remain at the forefront. Numerous conventional approaches have been used to detect heart disease over the decades. Although Cox Proportional Hazard Ratio, Framingham Risk Score, Thrombolysis in Myocardial Infarction, Systematic Coronary Risk Evaluation, Global Registry of Acute Coronary Events and QResearch Risk are already commonly used conventional risk scoring approaches; they have limitations due to requiring more assumptions, low-dimensionality data and time consuming with larger data sets, compared to machine learning approaches [10, 11]. Furthermore, the aforementioned statistical-based conventional approaches may not be sufficient for studying unstructured or semi-structured data. Hence, it is possible to claim that emerging risk factors used in machine learning approaches are more comprehensive than the ones which are used in traditional risk-scoring methods. Even so, instead to compete with one another; it is significant to consider whereas statistical methods are used to make inferences in samples, machine learning is used to learn the hidden patterns and associations between parameters in the data set and predict the likelihood of subsequent events [12].

There is a large volume of published studies describing the risk factors of heart disease. C.T. Ha et al. investigated emerging risk factors associated with heart-related death [13]. They suggested that age, sex, race, socioeconomic status, diabetes findings and obesity were associated with heart-related deaths. Furthermore, biomarkers including ECG, serum markers and imaging parameters were suggested as good predictors in cases integrated and diverse with machine learning approaches. Alizadehsani et al. investigated the papers published within the three decades that the main decision factors were age, sex, blood pressure, hypertension, ST depression, chest pain type and cholesterol for predicting heart disease with machine learning approaches [14]. Zhao et al. investigated the social determinants of heart disease from the papers that were studied based on a prediction by machine learning algorithms [15]. According to their review, in addition to medical inputs, social determinants such as gender and ethnicity may be considered in the prediction of patients' heart disease risk as well.

Numerous machine learning approaches mainly consisting of supervised and unsupervised techniques have been investigated in the prediction of heart disease in the literature. As the usage of machine learning in medicine becomes more widespread, it is claimed that certain data sets perform better in the early diagnosis of heart disease. Cleveland, Hungarian, Switzerland and Statlog are both the most known and cited data sets among seventy data sets that are published and include heart diseases related instances within three decades [14,16,17]. The features of all these data sets above include not only numerical data but also categorical data. These data sets include non-medical features such as patients' age and gender, medical features such as cholesterol, blood pressure, blood sugar, electrocardiogram results, maximum heart rate, and many more.

What we know about predicting heart disease with machine learning is largely based on empirical studies that investigate the performance of the algorithm. M.M. Ali et al. applied supervised machine learning algorithms including Multilayer Perceptron, K-Nearest Neighbor (KNN), Adaboost Classifier, Decision Tree, Logistic Regression and Random Forest to compare output performances [18]. Shah et al. benefited from Naïve Bayes, Decision Tree, KNN and Random Forest algorithms for Cleveland Database [19]. Li et al. developed a prediction supported by Support Vector Machines, Logistic Regression, Artificial Neural Network (ANN), KNN, Naïve Bayes and Decision Tree with standard feature selection algorithms [20]. Repaka et al. presented an

application that continuously monitors coronary heart patients who can see their heart disease condition [21]. This application was developed based on the Naïve Bayes algorithm. Rani et al. combined a genetic algorithm with recursive feature elimination for feature selection, then proposed a hybrid system generated from supervised machine learning algorithms [22]. Pathan et al. observed the impact of feature selection on the success of machine learning techniques [23]. They examined certain classification methods on the two data sets, CVD and Framingham. Thus, they put forward that optimizing feature size may decrease time consumption and increase model performance.

Mirmozaffari et al. applied and compared the results of various clustering techniques such as Canopy, Cobweb, EM, Farthest First, Filtered Clusters, Hierarchical Clusterer, Density-Based Clusterer and Simple K-Means [24]. Their research also showed that the accuracy tends to increase in each multilayer filtering preprocess step. Similarly, Kodati et al. discussed which clustering algorithm; including Farthest First, Filtered Cluster, Hierarchical Cluster, OPTICS, and Simple K-Means, might be the best for the selected data set [25]. Ripan et al. benefited from K-Means clustering approach to find out anomalies in their data set. After removing anomalies by clustering; they performed and discussed performances of KNN, Random Forest, Support Vector Machines, Naïve Bayes and Logistic Regression for different clusters [26].

In 2020, a comprehensive data set which is combined with five popular heart disease data sets (Cleveland, Hungarian, Switzerland, Long Beach VA and Statlog) was generated by Siddhartha [27]. The common features in those five data sets have been gathered in the new data set. Up to now, there is a relatively small body of literature that is used this data set [28–31]. Such that, Alalawi et al. used some classification algorithms such as Support Vector Machine, Decision Tree Classifier, Logistic Regression, KNN and Naïve Bayes [28]. In a study conducted by Yilmaz et al. classification outputs of Logistic Regression, Random Forest and Support Vector Machine methods were measured [29]. Doppala et al. proposed an Ensemble Model including Naïve Bayes, Random Forest, Support Vector Machine and XGBoost to improve prediction accuracy [30]. A. Tiwari et al. also examined an Ensemble Model with high accuracy [31]. A body of reviewed studies utilized various machine learning techniques in heart disease prediction, and a summary of previous studies in this field has been presented in Table 1.

Considering all of the evidence, it seems in Table 1 that a wide range of machine learning techniques has been used in previous studies for heart disease prediction. However, there is a lack of research that extracts decision rules by modeling decision tree results beyond performance metrics. In this research, we aim to fill this gap in the literature by proposing an approach to extract decision rule sets and graphical illustration of decision trees for healthcare professionals who henceforth may put it into practice uncomplicatedly.

3. Methodology

3.1. Data set

The data set used in this research was curated by five popular heart disease data sets covering both medical and non-medical features [27]. A total of 1190 patient records were incorporated into this comprehensive data set.

The data set has eleven features including age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise angina, oldpeak and ST slope. Also, it has one target variable including either “normal” or “heart disease”. The data set has not only numerical but also categorical features. Detailed feature descriptions were given in Table 2.

Table 1

Comparison of relevant literature review.

Article	Year	Objective	Data set	Technique
[18]	2021	To identify machine learning classifiers by applying and comparing outputs in heart disease prediction.	Cleveland, Hungary, Switzerland, and Long Beach	Classification
[19]	2020	To anticipate the possibility of patients suffering from heart disease	UCI machine learning repository	Classification
[20]	2020	To diagnose heart disease by classifying and using a feature selection algorithm	Cleveland data set	Classification
[21]	2019	To utilize machine learning for effectively enabling heart disease diagnosis and thereby offering appropriate treatment by offering a clinical decision support system.	UCI machine learning repository	Classification
[22]	2021	To assist clinicians in the early detection of heart disease by proposing a hybrid system	UCI machine learning repository	Classification
[23]	2022	To analyze the impact of heart disease prediction performance by selecting the best features	CVD and Framingham	Classification
[24]	2017	To compare different clustering algorithms to find the most accurate one in heart disease prediction.	A dataset collected from a hospital in Iran, under the supervision of the National Health Ministry.	Clustering
[25]	2019	To compare the performances of different types of clustering algorithms to identify the most appropriate one for the selected heart disease data set.	UCI machine learning repository	Clustering
[26]	2020	To predict heart disease using supervised machine learning algorithms after detecting outliers with K-means clustering algorithm.	UCI machine learning repository	Clustering & Classification
[28]	2021	To identify the best supervised classifier for two data sets through a deep learning network and a few machine learning classification models.	Cardiovascular diseased data set & Heart disease data set (Comprehensive)	Classification
[29]	2022	To measure the performances of machine learning methods for coronary heart disease	Heart disease data set (Comprehensive)	Classification
[30]	2022	To propose a reliable ensemble model to achieve the safety of patients.	Heart disease data set (Comprehensive)	Classification
[32]	2020	To show outputs of applied selected clustering algorithms	Own created data set including sex, age, chest pain, blood pressure, blood sugar, ECG and heart rate.	Clustering
[33]	2021	To compare the performance of K-Means clustering algorithm according to iterations and clusters	UCI heart disease data set	Clustering
[34]	2019	To increase the efficiency of heart disease prediction by integrating supervised and unsupervised algorithms	Framingham and Stat Log	Clustering & Classification

Table 2

Feature description of comprehensive heart disease data set.

Variable name	Values	Data type	Comment
Age	Real	Numeric	Patient's age in years
Sex	1,0	Binary	1: Male, 0: Female
Chest pain type	1,2,3,4	Nominal	1: Typical angina, 2: Atypical angina, 3: Non-anginal pain, 4: Asymptomatic
Resting blood pressure	Real	Numeric	Systolic pressure: Measure of the blood pressure when heart pushes blood out (mm Hg)
Serum cholesterol	Real	Numeric	Amount of cholesterol in patient's blood (Mg/Dl)
Fasting blood sugar	1,0	Binary	1 (True) : >120 mg/dl, 0 (False): ≤120 mg/dl
Resting electrocardiogram result	0,1,2	Nominal	0: Normal, 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria
Maximum heart rate achieved	Real	Numeric	Maximum number of times that heart beats per minute (bpm)
Exercise induced angina	0,1	Binary	1: Yes 0: No
Oldpeak	Real	Numeric	ST depression induced by exercise relative to rest
The slope of the peak exercise ST segment	0,1,2	Nominal	1: Upsloping 2: Flat 3: Downsloping
Target	0,1	Binary	1: Heart disease, 0 = Normal

Table 3

Detailed statistics and distributions of numerical features.

Feature name	Min	Max	Median	Mean	Standard deviation
Age	28	77	54	52.86	9.50
Resting blood pressure	92	200	130	133.0	17.28
Cholesterol	85	603	237	244.7	59.17
Maximum heart rate achieved	69	202	140	140.2	24.54
Oldpeak	0	6.2	0.50	0.90	1.07

Table 4

Detailed statistics for categorical features.

Feature name	Category	Number of observations	Percentage (%)
Sex	Male	563	75.57
	Female	182	24.43
Chest pain type	Asymptomatic	370	49.66
	Non-anginal pain	168	22.55
	Atypical angina	166	22.28
	Typical angina	41	5.50
Fasting blood sugar	No	124	16.64
	Yes	621	83.36
Resting ECG result	0	444	59.60
	1	125	16.78
	2	176	23.62
Exercise induced angina	No	458	61.48
	Yes	287	38.52
The slope of the peak exercise ST segment	Flat	353	47.38
	Upsloping	349	46.85
	Downsloping	43	5.77
Target	Normal	390	52.35
	Heart disease	355	47.65

3.2. Data preprocessing and descriptive statistics

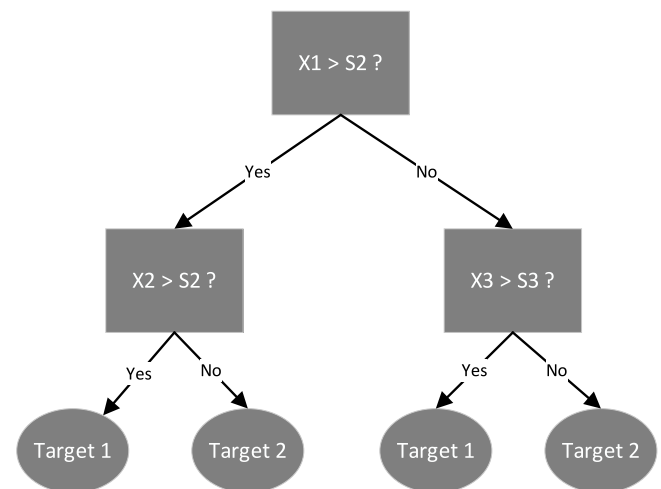
Before implementing the decision tree model, the data set was pre-processed. Specifically, it was focused on duplicated and inconsistent data. Following analysis of data preprocessing, 272 duplicated observations were discovered. Furthermore, it has been analyzed that 172 observations for the cholesterol feature and 1 observation for the ST Slope feature were discovered and labeled as 0. In the follow-up phase of the data preprocessing, all those duplicated and inconsistent records were removed from the data set. To enable the understandability of the model, some numerical values were changed to categorical expressions as illustrated in Table 3. For instance, 1,2 and 3 values were changed to upsloping, flat and down sloping in the ST Slope feature. Eventually, the length of the data was reduced to 745 and some values were adjusted for the decision tree model.

Overall, Tables 3 and 4 summarize the data analyzed in this study. Since the data set has both numerical and categorical variables, the features were presented separately. Whereas numerical features were illustrated based on minimum values, maximum values, median, mean and standard deviation; categorical features were identified for each sub-category.

3.3. Decision tree algorithms and CART

As a part of popular machine learning techniques, decision tree sheds light on dealing with the prediction of target variables based on the given input parameters [35]. Decision trees can be considered in two ways depending on the condition of the variables. When the target variable is numerical, the problem can be solved in the context of regression; otherwise, it is a classification problem [36]. Fig. 1 demonstrates a schematic illustration of a typical decision tree model.

Although there are many decision tree algorithms in the literature for classification problems; ID3, C4.5, Classification and Regression Trees (CART) and Chi-Square Automatic Interaction Detection (CHAID) algorithms are among the most frequently referenced decision tree

**Fig. 1.** A typical CART structure.

Source: Adapted from Ref. [37]

algorithms [38,39]. Related researchers claim that the CART algorithm proves its reliability among all classification techniques [40]. Similar to other decision tree algorithms, the CART algorithm extracts decision rules from features and builds a model to predict the target values [31]. The features may include either numerical or categorical values. Overall, the CART algorithm was adopted to ensure that the results of the study would help researchers and healthcare professionals in early diagnosis and treatment plans.

3.4. Implementation of the CART algorithm

To ensure that the heart disease prediction method is implemented accurately, the data set was randomly divided into two sets; training and testing sets. The original data was partitioned at a ratio of 8:2; 80%

Table 5
Confusion matrix of heart disease prediction.

		Predicted	
		Heart disease	Normal
Actual	Heart disease	True Positive (TP): number of observations actually is heart disease and predicted as heart disease	False Negative (FN): number of observations actually have heart disease but are predicted as normal
	Normal	False Positive (FP): number of observations actually is normal but predicted as heart disease	True Negative (TN): number of observations actually is normal and predicted as normal

of the data (training set) was used for training to recognize the patterns in the data set and 20% of the data (testing set) was used to verify the classification accuracy of the CART model. The principal findings obtained from both training and testing data sets were presented in the results section.

Data from the training part (80% of the total) consisting of 284 heart disease and 312 normal observations were included in the analysis. All analyses were carried out using R programming language and R Studio which is an open source software tool specialized for statistical analysis, modeling, machine learning, graphics representation and reporting purposes [41]. In the modeling phase, the “rpart” library was used for the implementation of the CART algorithm.

To mitigate overfitting, the decision tree model was pruned by minimizing the complexity parameter which controls the size of the tree and reduces classification errors [42–44]. Hence, better performance scores were intended to obtain based on the CART algorithm.

3.5. Evaluation measures

In case the target of the classification problem is related to finding out binary variables such as True/False, Yes/No; the accuracy parameter which defines the ratio of correct classifications mostly sheds light on the performance of the classifier [38,40]. As such, the confusion matrix helps to evaluate the performance of the classification model by calculating parameters as shown in Table 5 [45]. From these parameters, specific qualities are also obtained; such as accuracy, sensitivity, specificity and precision [46]. Sensitivity, or recall, refers to the probability of observations that are predicted true to the number of all observations that are actually classified as true [47]. Similarly, specificity refers to the probability of observations that are predicted false to the number of all observations that are actually classified as false. Lastly, precision displays how many of the observations that are predicted as true are actually true.

$$\text{Accuracy} : (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Sensitivity} : TP / (TP + FN) \quad (2)$$

$$\text{Specificity} : TN / (TN + FP) \quad (3)$$

$$\text{Precision} : TP / (TP + FP) \quad (4)$$

By extracting rules, the outputs of the decision trees transform into logical rule forms in “If” and “Then” structures. To do this, the leaves are associated with the related way from the root to the leaf [48]. All nodes are connected from antecedent (If) to rule consequent (Then) [49]. Thus, decision tree rules enable the model more interpretable, and it provides the model in the most detailed and transparent form possible.

4. Results

It can be inferred from the descriptive statistics in Tables 3 and 4, a great majority of the patients are male. For this data set, males tended

to be diagnosed with heart disease more than females. Approximately half of the patients are found to have asymptomatic chest pain type. Accordingly, most patients who have asymptomatic chest pain type are also diagnosed with heart disease. Most patients do not have high fasting blood sugar, and also there is no significant difference between the number of normal and heart disease patients, regardless of fasting blood sugar. While most of the patients without exercise induced angina are normal, the vast majority of them who have exercise induced angina is diagnosed with heart disease. Lastly, the slope of the peak exercise ST segment is also a significant indicator of heart disease. Such that, nearly eight-tenths of patients who are diagnosed with heart disease are recorded as flat. On the contrary, nearly eight tenth of patients who are diagnosed with normal are recorded as upsloping.

Based on the descriptive analyses mentioned above, the decision tree implemented by the CART algorithm was illustrated in Fig. 2. Heart disease and normal observation percentages were obtained for each node illustrated in Fig. 2.

The target feature of the model in the root node is either heart disease or normal observation. Additionally, the first level attribute is found to be ST Slope; the second level attribute is chest pain type; the third level attributes are resting blood pressure and exercise angina; the fourth level attributes are sex, maximum heart rate and oldpeak; and the fifth level attribute is cholesterol. The rule-based illustration of the model was listed in Table 6.

In Table 6, the rules that are obtained from the CART model were described line by line, by following the nodes and branches. The table includes 18 rules based on five levels of the tree. The tree level, the percentage of normal and heart disease observations, the total number of observations, the percentage of total observations (frequency), and the information on whether the rule expresses the end node or not were given for each specific rule. Thus, understanding the rules makes the model more efficient and understandable. For instance, one can understand from rule 1 that if the ST slope of a patient is either flat or downsloping then heart disease would be expected at about 78%. Similarly, rule 16 means that if ST slope is either upsloping; chest pain type is asymptomatic; exercise angina is yes and oldpeak is lower than 0,05 then the expected heart disease is 33%. One of the differences between the given rules is that the second rule is an end node which means it shows the outcome of a decision path. Overall, it is possible to clarify the whole tree with this interpreting logic.

The importance of each feature involved in the implementation of the CART approach was investigated. As illustrated in Fig. 3, the ST slope is the most important feature for the data set. The importance of the SP slope is 28% according to the CART algorithm which means that the ST slope will probably larger effect on the CART model. Oldpeak, chest pain type, exercise angina and maximum heart rate follow ST Slope with 18%, 16%, 15% and 11% respectively. Oppositely, resting ECG is the least important feature among the used features in the data set which means it would probably smaller effect on the model. Oldpeak, chest pain type, exercise angina, maximum heart rate, age, resting blood pressure and cholesterol are the other features that affect the model based on importance.

The results show that approximately 96% of the patients who are predicted as heart disease has ST slope of flat or downsloping. Additionally, asymptomatic chest pain type is a significant indicator, such that, approximately 81% of all patients who suffer from heart disease

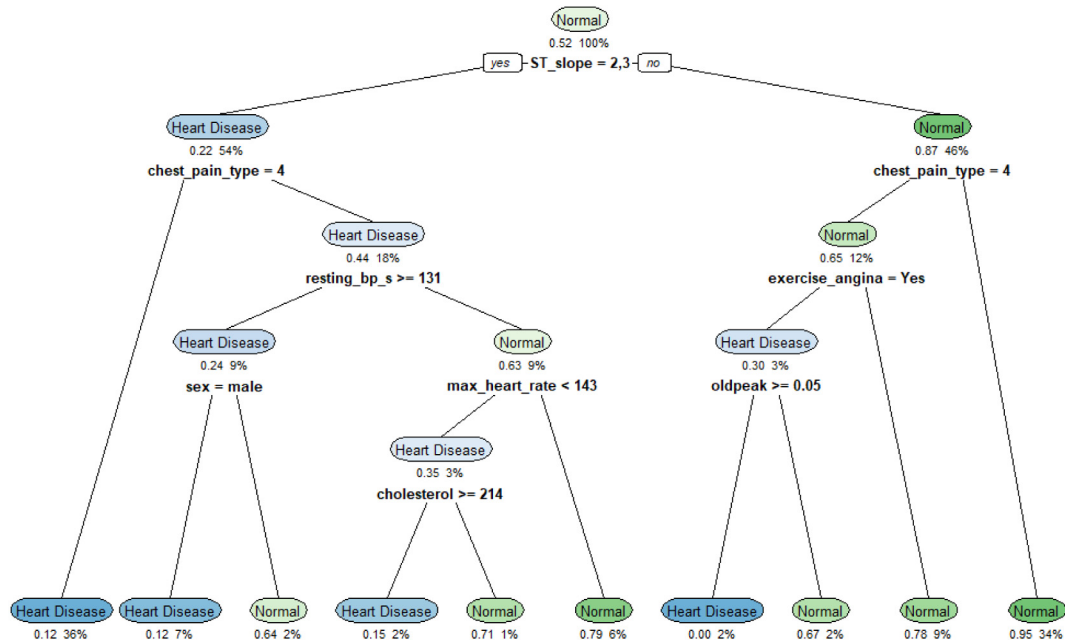


Fig. 2. Graphical representation of the CART approach.

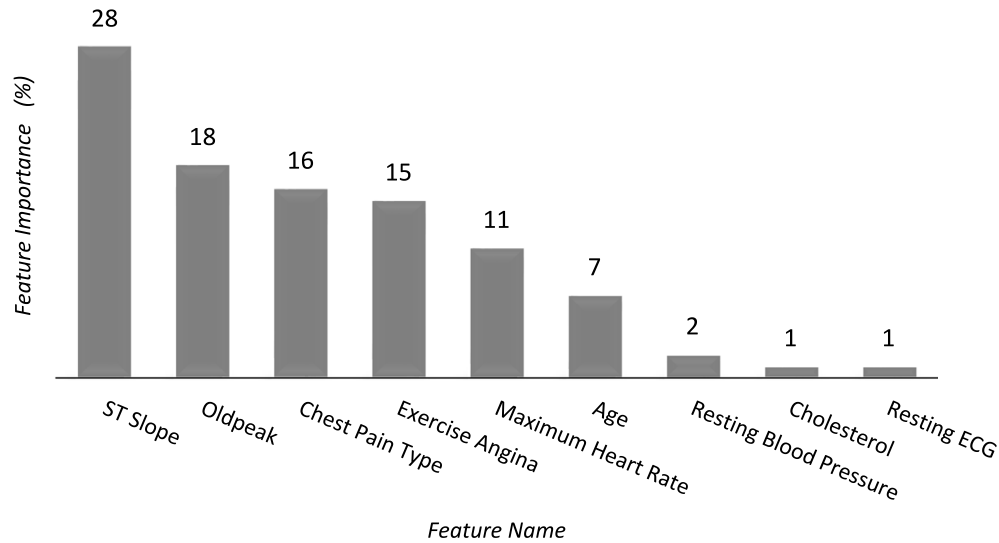


Fig. 3. Importance of features implemented by the CART model.

have asymptomatic chest pain type. These findings also indicate that early diagnosis of heart disease may be challenged where chest pain type and ST slope values are not analyzed. On the other hand, even if cholesterol is one of the features which has a low impact on the model, it can be considered as an input to increase the total accuracy of the prediction.

In Table 7, the testing results for the CART model were presented. It summarizes performance parameters reported based on accuracy, sensitivity, specificity, and precision. As can be seen from the results, the CART model has a good prediction accuracy with 87%. The other performance parameters (sensitivity, specificity, and precision) are 85%, 90% and 88%, respectively. The accuracy of our CART model is higher than other studies which are aimed to predict heart disease by using decision trees [19,20,30,31]. For example, Doppala et al. reached up to 82% for accuracy, 79% for sensitivity, 85% for specificity and 83% for precision [30]. In addition, the difference between training and testing results is relatively low and comparable. For example, Shah et al. found a difference of about 6% for accuracy [19]. In our

study, accuracy, sensitivity, specificity and precision are 88%, 87%, 89% and 88%, respectively; which means our average difference is approximately 1%. As both of the training set and testing set have high accuracy performances and it shows low error rate, the model does not signal overfitting and it can be generalized to unseen data [50]. Based on these results, we can conclude that our CART model is successfully validated and performs reasonably well.

5. Conclusion

Heart disease is one of the most fatal diseases around the world. It is highly significant especially both for patients who are risk of heart attack, or even death; and for the healthcare professionals who are responsible for diagnosing it as promptly as practicable. To deal with these issues, it is essential to effectively detect the patients in advance. In this manner, this paper proposes the CART model to gain deeper and reasonable insight to healthcare professionals about the patients and presents a logical rule set of the model to use it in practical cases

Table 6

The rules for the degree of heart disease resulted from the CART algorithm.

Rule number	Tree level	Rule	Degree of heart disease (%)		Observations		Is leaf ?
			Heart disease	Normal	Number of cases	Frequency (%)	
1	1	ST slope is either flat or downsloping =>	78	22	320	54	No
2	5	ST slope is downsloping & chest pain type is asymptomatic =>	87	13	276	46	Yes
3	2	ST slope is either flat or downsloping & chest pain type is not typical angina =>	56	44	105	18	No
4	3	ST slope is either flat or downsloping & chest pain type is not typical angina & resting blood pressure ≥ 131 =>	76	24	51	9	No
5	5	ST slope is either flat or downsloping & chest pain type is not typical angina & resting blood pressure ≥ 131 & sex is male =>	88	12	40	7	Yes
6	5	ST slope is either flat or downsloping & chest pain type is not typical angina & resting blood pressure ≥ 131 & sex is female =>	36	64	11	2	Yes
7	3	ST slope is either flat or downsloping & chest pain type is not typical angina & resting blood pressure < 131	37	63	54	9	No
8	4	ST slope is either flat or downsloping & chest pain type is not typical angina & resting blood pressure ≥ 131 & maximum heart rate < 143 =>	65	35	20	3	No
9	5	ST slope is either flat or downsloping & chest pain type is not typical angina & resting blood pressure ≥ 131 & maximum heart rate < 143 & cholesterol ≥ 214 =>	85	15	13	2	Yes
10	5	ST slope is either flat or downsloping & chest pain type is not typical angina & resting blood pressure ≥ 131 & maximum heart rate < 143 & cholesterol < 214 =>	29	71	7	1	Yes
11	5	ST slope is either flat or downsloping & chest pain type is not typical angina & resting blood pressure ≥ 131 & maximum heart rate ≥ 143 =>	21	79	34	6	Yes
12	1	ST slope is either upsloping =>	13	87	276	46	No
13	2	ST slope is either upsloping & chest pain type is asymptomatic =>	35	65	74	12	No
14	3	ST slope is either upsloping & chest pain type is asymptomatic & exercise angina is yes =>	70	30	6	3	No
15	5	ST slope is either upsloping & chest pain type is asymptomatic & exercise angina is yes & oldpeak $\geq 0,05$ =>	100	0	11	2	Yes
16	5	ST slope is either upsloping & chest pain type is asymptomatic & exercise angina is yes & oldpeak $< 0,05$ =>	33	67	9	2	Yes
17	5	ST slope is either upsloping & chest pain type is asymptomatic & exercise angina is no =>	22	78	12	2	Yes
18	5	ST slope is either upsloping & chest pain type is not asymptomatic =>	5	95	202	34	Yes

Table 7

Performance parameters of the CART algorithm for the testing data set.

		Predicted		Total	ACC	SE	SP	PR
		1	0					
Actual	1	60	8	68	87.25%	84.51%	89.74%	88.24%
	0	11	70	81				
	Total	71	78	149				

oversimply. In this context, the model was built by using electronic health records of 1190 patients. After the original data set was preprocessed and analyzed, the model was performed. The model performed in this study proved to be sufficiently accurate and validated. One of the significant findings to emerge from this study is that our model performs reasonably well in predicting heart disease. The second major finding is that, according to the importance of features in the CART model, the features relating electrocardiogram test including ST Slope and Oldpeak are more important features to detect heart disease.

This study may contribute to prior literature by extracting a new type of rule set that makes the decision-making process in heart disease diagnosis more understandable. This study has also significant practical directions and a broad clinical implication for healthcare professionals who can benefit from the model to diagnose different patient groups based upon the decision tree model and the rule set, oversimply. Moreover, by integrating with the proposed model, the development of clinical decision support systems could be greatly facilitated. Thus, the healthcare professionals can ultimately improve the quality of

healthcare delivery by combining their knowledge with the suggestions provided by the clinical support systems.

Despite its advantages and contributions, generalization of these results has certain limitations. Firstly, the proposed model does not totally include patients' medical information and other social determinants that may lead to heart disease, such as socioeconomic status, current smoking status, etc. In this context, further research might extend this study by adding new clinical, demographic and social determinants to validate the findings and increase the quality of the results. Secondly, the unstructured data such as text or image were not considered in this study. For future studies, our model may be enhanced by adding unstructured data such as text files/documents, sensor data, images, website/application logs and/or social media data. Due to large amount of data related to heart disease, it is recommended that big data analytics could bring some benefits to the field. Therefore, it may be possible to develop better practices for clinical support systems supporting by big data applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- [1] Heart Disease Overview, Mayo Clinic, 2022, <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>.
- [2] Coronary Heart Disease - Diagnosis, NHS, 2022, <https://www.nhs.uk/conditions/coronary-heart-disease/diagnosis/>.
- [3] R. Williams, T. Shongwe, A.N. Hasan, V. Rameshar, Heart disease prediction using machine learning techniques, in: 2021 International Conference on Data Analytics for Business and Industry, ICDABI 2021, 2021, pp. 118–123, <http://dx.doi.org/10.1109/ICDABI53623.2021.9655783>.
- [4] Cardiovascular Diseases, WHO, 2022, <https://www.who.int/health-topics/cardiovascular-diseases#tab=tab.1>.
- [5] V. Chang, V.R. Bhavani, A.Q. Xu, M. Hossain, An artificial intelligence model for heart disease detection using machine learning algorithms, *Healthc. Anal.* 2 (2022) <http://dx.doi.org/10.1016/j.health.2022.100016>.
- [6] M.M. Ahsan, Z. Siddique, Machine learning-based heart disease diagnosis: A systematic literature review, *Artif. Intell. Med.* 128 (2022) 102289, <http://dx.doi.org/10.1016/j.artmed.2022.102289>.
- [7] A. Alanazi, Using machine learning for healthcare challenges and opportunities, *Inform. Med. Unlocked* 30 (100924) (2022) <http://dx.doi.org/10.1016/j.imu.2022.100924>.
- [8] L. Yahaya, N. David Oye, E. Joshua Garba, A comprehensive review on heart disease prediction using data mining and machine learning techniques, *Am. J. Artif. Intell.* 4 (1) (2020) 20, <http://dx.doi.org/10.11648/j.ajai.20200401.12>.
- [9] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, N. Kumar, Latest trends on heart disease prediction using machine learning and image fusion, *Mater. Today Proc.* 37 (2) (2020) 3213–3218, <http://dx.doi.org/10.1016/j.matpr.2020.09.078>.
- [10] A.S. Mohd Faizal, T.M. Thevarajah, S.M. Khor, S.W. Chang, A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach, *Comput. Methods Programs Biomed.* 207 (2021) <http://dx.doi.org/10.1016/j.cmpb.2021.106190>.
- [11] N. Biswas, K. Mohammad, M. Uddin, S. Tasmin, Healthcare analytics a comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach, *Healthc. Anal.* 2 (2022) <http://dx.doi.org/10.1016/j.health.2022.100116>.
- [12] D. Mpanya, T. Celik, E. Klug, H. Ntsinjana, Machine learning and statistical methods for predicting mortality in heart failure, *Heart Fail. Rev.* 26 (3) (2021) 545–552, <http://dx.doi.org/10.1007/s10741-020-10052-y>.
- [13] A.C.T. Ha, B.S. Doumouras, C. (Nancy) Wang, J. Tranmer, D.S. Lee, Prediction of sudden cardiac arrest in the general population: Review of traditional and emerging risk factors, *Can. J. Cardiol.* 38 (4) (2022) 465–478, <http://dx.doi.org/10.1016/j.cjca.2022.01.007>.
- [14] R. Alizadehsani, et al., Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020, *Comput. Biol. Med.* 128 (2021) <http://dx.doi.org/10.1016/j.combiomed.2020.104095>.
- [15] Y. Zhao, E.P. Wood, N. Mirin, S.H. Cook, R. Chunara, Social determinants in machine learning cardiovascular disease prediction models: A systematic review, *Am. J. Prev. Med.* 61 (4) (2021) 596–605, <http://dx.doi.org/10.1016/j.amepre.2021.04.016>.
- [16] D. Yewale, S.P. Vijayaragavan, Comprehensive review on machine learning approach for heart disease prediction: Current status and future prospects comprehensive review on machine learning approach for heart disease prediction: Current status and future prospects, 2022, <http://dx.doi.org/10.1063/5.0080483>.
- [17] R. Porto, J.M. Molina, A. Berlanga, M.A. Patricio, Minimum relevant features to obtain explainable systems for predicting cardiovascular disease using the statlog data set, *Appl. Sci.* 11 (3) (2021) 1–18, <http://dx.doi.org/10.3390/app11031285>.
- [18] M.M. Ali, B.K. Paul, K. Ahmed, F.M. Bui, J.M.W. Quinn, M.A. Moni, Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, *Comput. Biol. Med.* 136 (2021) <http://dx.doi.org/10.1016/j.combiomed.2021.104672>.
- [19] D. Shah, S. Patel, S.K. Bharti, Heart disease prediction using machine learning techniques, *SN Comput. Sci.* 1 (2020) 345, <http://dx.doi.org/10.1109/ICEECCOT52851.2021.9707966>.
- [20] J.P. Li, A.U. Haq, S.U. Din, J. Khan, A. Khan, A. Saboor, Heart disease identification method using machine learning classification in E-healthcare, *IEEE Access* 8 (2020) 107562–107582, <http://dx.doi.org/10.1109/ACCESS.2020.3001149>.
- [21] A.N. Repaka, S.D. Ravikanti, R.G. Franklin, Design and implementing heart disease prediction using naive Bayesian, in: Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019, 2019, pp. 292–297, <http://dx.doi.org/10.1109/icoei.2019.8862604>.
- [22] P. Rani, R. Kumar, N.M.O.S. Ahmed, A. Jain, A decision support system for heart disease prediction based upon machine learning, *J. Reliab. Intell. Environ.* 7 (3) (2021) 263–275, <http://dx.doi.org/10.1007/s40860-021-00133-6>.
- [23] M.S. Pathan, A. Nag, M.M. Pathan, S. Dev, Analyzing the impact of feature selection on the accuracy of heart disease prediction, *Healthc. Anal.* 2 (2022) <http://dx.doi.org/10.1016/j.health.2022.100060>.
- [24] M. Mirmozaffari, A. Alinezhad, A. Gilanpour, Heart disease prediction with data mining clustering algorithms, *Int. J. Comput. Commun. Instrum. Eng.* 4 (1) (2017) <http://dx.doi.org/10.15242/ijccie.dir1116009>.
- [25] S. Kodati, R. Vivekanandam, G. Ravi, Comparative Analysis of Clustering Algorithms with Heart Disease Datasets using Data Mining Weka Tool, Springer Singapore, 2019, <http://dx.doi.org/10.1007/978-981-13-3600-3.11>.
- [26] R.C. Ripan, et al., A data-driven heart disease prediction model through K-means clustering-based anomaly detection, *SN Comput. Sci.* 2 (2) (2021) 1–12, <http://dx.doi.org/10.1007/s42979-021-00518-7>.
- [27] M. Siddhartha, Heart Disease Dataset (Comprehensive), IEEE Dataport, 2020, <https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive#files>.
- [28] H.H. Alalawi, M.S. Alsuwat, Detection of cardiovascular disease using machine learning classification models, *Int. J. Eng. Res. Technol.* 10 (07) (2021) 151–157.
- [29] R. Yilmaz, F.H. Yağın, Early detection of coronary heart disease based on machine learning methods, *Med. Rec. Med. J.* 4 (1) (2022) 1–6, <http://dx.doi.org/10.37990/medr.1011924>.
- [30] B.P. Doppala, D. Bhattacharyya, M. Janarthanan, N. Baik, A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques, *J. Healthc. Eng.* 2022 (2022) <http://dx.doi.org/10.1155/2022/2585235>.
- [31] A. Tiwari, A. Chugh, A. Sharma, Ensemble framework for cardiovascular disease prediction, *Comput. Biol. Med.* 146 (2022) <http://dx.doi.org/10.1016/j.combiomed.2022.105624>.
- [32] S.K. Pandey, B. Mishra, S.S. Gautam, Cluster based mining for prediction of heart disease, *Int. J. Comput. Sci. Mob. Comput.* 9 (7) (2020) 136–143.
- [33] A.Z. Jebakumar, R. Ramanan, A deductive learning of heart disease dataset by using K means clustering, *Ann. Rom. Soc. Cell Biol.* 25 (6) (2021) 4269–4289.
- [34] J.D. Bodapati, V.R. Krishna Sajja, N.B. Mundukur, N. Veeranjanyulu, Robust cluster-then-label (RCTL) approach for heart disease prediction, *Ing. Des. Syst. d'Inf.* 24 (3) (2019) 255–260, <http://dx.doi.org/10.18280/isi.240305>.
- [35] R. Sann, P.C. Lai, S.Y. Liaw, C.T. Chen, Predicting online complaining behavior in the hospitality industry: Application of big data analytics to online reviews, *Sustainability* 14 (3) (2022) <http://dx.doi.org/10.3390/su14031800>.
- [36] J. Abellán, G. López, J. De Oña, Analysis of traffic accident severity using decision rules via decision trees, *Expert Syst. Appl.* 40 (15) (2013) 6047–6054, <http://dx.doi.org/10.1016/j.eswa.2013.05.027>.
- [37] S.S. Moon, S.Y. Kang, W. Jitpitakert, S.B. Kim, Decision tree models for characterizing smoking patterns of older adults, *Expert Syst. Appl.* 39 (1) (2012) 445–451, <http://dx.doi.org/10.1016/j.eswa.2011.07.035>.
- [38] M.M. Ghiasi, S. Zendeheboudi, Decision tree-based methodology to select a proper approach for wart treatment, *Comput. Biol. Med.* 108 (April) (2019) 400–409, <http://dx.doi.org/10.1016/j.combiomed.2019.04.001>.
- [39] M. Batra, R. Agrawal, Comparative analysis of decision tree algorithms, *Adv. Intell. Syst. Comput.* 652 (2018) 31–36, http://dx.doi.org/10.1007/978-981-10-6747-1_4.

- [40] M.M. Ghiasi, S. Zendejboudi, A.A. Mohsenipour, Decision tree-based diagnosis of coronary artery disease: CART model, *Comput. Methods Programs Biomed.* 192 (2020) <http://dx.doi.org/10.1016/j.cmpb.2020.105400>.
- [41] What is R? The R Foundation, 2022, <https://www.r-project.org/about.html>.
- [42] I. Yoo, et al., Data mining in healthcare and biomedicine: A survey of the literature, *J. Med. Syst.* 36 (4) (2012) 2431–2448, <http://dx.doi.org/10.1007/s10916-011-9710-5>.
- [43] T. Xie, R. Li, X. Zhang, B. Zhou, Z. Wang, Research on heartbeat classification algorithm based on cart decision tree, in: 8th International Symposium on Next Generation Electronics, ISNE 2019, 2019, pp. 27–29, <http://dx.doi.org/10.1109/ISNE.2019.8896650>.
- [44] G. Mistikoglu, I.H. Gerek, E. Erdis, P.E. Mumtaz Usmen, H. Cakan, E.E. Kazan, Decision tree analysis of construction fall accidents involving roofers, *Expert Syst. Appl.* 42 (4) (2015) 2256–2263, <http://dx.doi.org/10.1016/j.eswa.2014.10.009>.
- [45] A.K. Dwivedi, Performance evaluation of different machine learning techniques for prediction of heart disease, *Neural Comput. Appl.* 29 (10) (2018) 685–693, <http://dx.doi.org/10.1007/s00521-016-2604-1>.
- [46] S.I. Ayon, M. Islam, R. Hossain, Coronary artery heart disease prediction : A comparative study of computational intelligence techniques coronary artery heart disease prediction : A comparative study of, *IETE J. Res.* (2020) 1–20, <http://dx.doi.org/10.1080/03772063.2020.1713916>.
- [47] S. Peker, A. Kocyigit, P.E. Eren, A hybrid approach for predicting customers' individual purchase behavior, *Kybernetes* 46 (10) (2017) 1614–1631, <http://dx.doi.org/10.1108/K-05-2017-0164>.
- [48] K. Khalili-Damghani, F. Abdi, S. Abolmakarem, Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries, *Appl. Soft Comput.* 73 (2018) 816–828, <http://dx.doi.org/10.1016/j.asoc.2018.09.001>.
- [49] J. De Oña, G. López, J. Abellán, Extracting decision rules from police accident reports through decision trees, *Accid. Anal. Prev.* 50 (2013) 1151–1160, <http://dx.doi.org/10.1016/j.aap.2012.09.006>.
- [50] X. Ying, An overview of overfitting and its solutions, *J. Phys. Conf. Ser.* 1168 (2) (2019) <http://dx.doi.org/10.1088/1742-6596/1168/2/022022>.