Journal of Ilam University of Medical Sciences

P-ISSN: 1563-4728
E-ISSN: 2588-3135

From Data to Treatment: Early Detection of Heart Disease Using Machine Learning Techniques

Mostafa Yousofi Tezerjan [1], **Maryam Mollabagher** [2]✉

[1] Dept of Industry, University of Applied Science and Technology, Tehran, Iran
[2] Dept of Management and Social Services, University of Applied Science and Technology, Tehran, Iran

Article Info

Article type:
Original Article

Article History:
Received: Nov. 30, 2025
Revised: Feb. 06, 2026
Accepted: Feb. 25, 2026
Published Online: Apr. 04, 2026

✉ ***Correspondence to:***
Maryam Mollabagher
Dept of Management and Social Services, University of Applied Science and Technology, Tehran, Iran

Email:
Mollabagher@uast.ac.ir

A B S T R A C T

Introduction: Early detection of cardiovascular diseases (CVDs) is a significant challenge in the medical field. As one of the leading causes of mortality worldwide, CVDs require timely and accurate diagnosis for effective prevention. Delayed diagnoses often result from limitations in traditional methods and the insufficient identification of risk factors. This study aimed to enhance the accuracy of heart disease prediction using machine learning techniques.

Materials & Methods: This research assessed the detection rate and accuracy of cardiovascular disease prediction using machine learning algorithms, including logistic regression, the C5.0 decision tree, neural networks, and mixed models. Data from the Comprehensive Heart Disease Dataset, comprising 1190 samples and 11 clinical features, were analyzed. Performance metrics such as accuracy, sensitivity, and specificity were used to evaluate the models. Data analysis was conducted using SPSS Modeler 18 software.

Results: The evaluated models demonstrated varying accuracies ranging from 85% to 93%. The hybrid model achieved the best performance with an accuracy of 93.44%. Additionally, ST slope, chest pain type, and fasting blood sugar were identified as the most significant risk factors for cardiovascular diseases.

Conclusion: Machine learning algorithms offer significant potential for the early detection of cardiovascular diseases, thereby reducing mortality rates. These techniques enable precise analysis of clinical data, improve medical decision-making, and enhance treatment outcomes.

Keywords: Cardiovascular Diseases, Early Diagnosis, Machine Learning, Risk Factors, Predictive Analytics

➢ **Cite this paper**
Yousofi Tezerjan M, Mollabagher M. From Data to Treatment: Early Detection of Heart Disease Using Machine Learning Techniques. *Journal of Ilam University of Medical Sciences*. 2026; 34(1):72-89.

Publisher: Ilam University of Medical Sciences

Introduction

Cardiovascular diseases are recognized as one of the most important causes of mortality and disability worldwide and are considered a serious threat to public health. According to reports from the World Health Organization, these diseases are not only responsible for a large proportion of deaths but also impose a significant economic and social burden on health systems and societies (1). Risk factors associated with these diseases include high blood pressure, obesity, high cholesterol, smoking, a sedentary lifestyle, and unhealthy diets (2). Evidence shows that up to 80% of deaths from these diseases can be controlled by changing health behaviors and adopting a healthy lifestyle (1).

One of the main challenges in the management of cardiovascular diseases is their timely and accurate diagnosis. Late or incorrect diagnosis often occurs due to limitations in traditional diagnostic methods and lack of sufficient awareness of early symptoms. In this regard, new technologies, especially machine learning algorithms, provide effective tools for analyzing complex data and identifying risk patterns (3). These algorithms can help predict cardiovascular diseases and identify risk factors early with high accuracy and help doctors make better decisions (4). The use of advanced machine learning techniques for modeling and predicting heart diseases has provided new opportunities to improve diagnosis and reduce the burden of these diseases. These methods, by analyzing clinical and demographic data of patients, can lead to the development of more accurate and practical models. As a result, these technologies can have a significant impact on reducing mortality, improving treatment outcomes, and promoting public health (5-6). The idea of clinical economics focuses on having a systematic approach to solving problems in the health system and helps to improve this system by examining the current situation and predicting future conditions (7). Recent developments in machine learning and computational advances have opened up new opportunities for research in the healthcare field. In addition, machine learning can help identify risk factors associated with diseases by analyzing data on lifestyle, medical history, and environmental factors (8). Various researchers have used machine learning algorithms to increase the accuracy of disease prediction, which can analyze complex medical data and identify hidden patterns for more accurate disease prediction and early diagnosis (9-11).

Various methods are used in disease prediction, including advanced algorithms such as random forest and other classifiers, as well as probabilistic principal component analysis, which have significantly increased the accuracy of the models (12-16). In addition, to determine the optimal combination of predictors for heart disease, Gazloglu et al. evaluated 18 machine learning models and three feature selection techniques on the Cleveland dataset consisting of 303 samples and 13 variables (17). Recently, ten classifiers were trained to identify the most effective prediction models for accurate prediction (18). The most suitable features were identified using three feature selection methods, including the correlation-based feature subset estimator, the chi-square feature estimator, and the relief feature estimator. In addition, a hybrid feature selection method aimed at increasing accuracy by combining random forest and linear correlation was proposed by Pavitra et al. (19). Implementing this technique, following the selection of 11 features through a combination of filter, wrapper, and embedding methods, resulted in a 2% increase in the accuracy of the hybrid model. To further increase the accuracy, researchers have used the ensemble technique to combine different algorithms. This ensemble method was developed for heart disease diagnosis by Lathha et al. (20).

By combining naive Bayes, random forest, multilayer perceptrons, and Bayesian networks based on majority voting, they achieved an accuracy of 85.48%. It was also used by an ensemble model with five classifiers, including a memory-based learner, a support vector machine, decision tree induction with Bayesian Gini index information augmentation, and decision tree initialization with Gini index (21). Since the dataset in the authors' study only contained relevant features, there was no feature selection. A preprocessing step was performed to remove outliers and missing values from the data. Tama et al. (22) developed an ensemble model for heart disease diagnosis with an accuracy rate of 85.71%. The ensemble model used gradient boosting, random forest, and extreme gradient boosting classifiers. Al-Qahtani et al. (23) developed a set of machine learning and deep learning models for disease prediction with an accuracy rate of 88.70%. The study used a total of six classification algorithms. Trigka et al. (24) developed a stacked ensemble model after applying vector machine, naive Bayes, and nearest neighbor methods with a 10-fold cross-validation of the artificial minority oversampling technique to balance the unbalanced dataset and achieved an accuracy of 90.9%. Another study used stochastic gradient descent, linear regression, and vector machine classifiers to develop a model with an accuracy of 93% using multiple datasets (25).

To further improve the accuracy, Cyriac et al. (26) used seven different machine learning models as well as two group methods (soft voting and hard voting). With this approach, the highest accuracy score was achieved at 94.2%. Another study (27) developed a predictive model approach with mixed classification for better prediction accuracy. Five classifier models were combined with the Cleveland and Hungarian datasets. A total of 590 valid data samples and 13 features were considered. A baseline accuracy of 93% was achieved using the data mining tool Weka. In 2020, Mana Siddhartha created a new dataset by combining five well-known heart disease datasets, including Switzerland, Cleveland, Hungary, Statlog, and Long Beach VA. This new dataset contains all the common features of the five datasets.

Based on the review of previous studies, it is clear that machine learning algorithms are capable of increasing the accuracy of cardiovascular disease prediction and early identification of risk factors. However, many studies have focused on the analysis of specific algorithms or limited datasets, and a comprehensive study of the performance of hybrid models and their comparison in real-world conditions has not yet been fully conducted. The way algorithms are combined can have a significant impact on the accuracy of prediction. The main objective of this study is to analyze and evaluate the performance of different machine learning models in predicting heart disease and identifying important risk factors using the comprehensive Cleveland dataset. By focusing on 11 key clinical features and using decision trees, neural networks, and hybrid models, this study attempts to provide a model with high accuracy and the ability to identify high-risk patients early. This study combines algorithms used in previous research in the field of machine learning to provide a hybrid model with high accuracy.

Materials and methods

This study investigates the prediction of cardiovascular diseases using decision tree algorithms. The aim of this research is to analyze the factors affecting these diseases and predict the probability of their occurrence through existing data. Due to the lack of access to existing indigenous data, the statistical population includes data on 11 indicators from 1190 patients extracted from the Comprehensive

Heart Disease Dataset. This data set provides important information, including demographic and physiological characteristics of patients, which can help develop more accurate and effective diagnostic tools.

To avoid the problem of overfitting and poor predictions in unseen data, the data set is divided into two parts: 75% is selected as training data and 25% as test data. This division allows for a more accurate evaluation of the prediction model. The data were analyzed and modeled using SPSS and SPSS Modeler software. In this study, 11 independent input variables were used to predict heart disease. These variables included characteristics such as gender, type of chest pain, fasting blood sugar, resting electrocardiogram results, exercise angina, ST segment elevation, and numerical characteristics such as resting systolic blood pressure, serum cholesterol level, patient age, and ST segment depression during exercise testing. These characteristics help to analyze the status of patients with heart disease and make more accurate predictions. Table 1 describes the indicators of the dataset.

Table 1. Heart disease dataset indicators along with sample size.

Column	Index Symbol	Variable Name	Data Type	Data Range	Variable Coding & Description	Sample Size
1	Age	Age	Continuous numeric	(28–77)	Patient age	—
2	Sex	Sex	Nominal	0,1	0–Female	281
					1–Male	909
3	Chest Pain Type	Chest pain type	Nominal	0,1,2,3	0–Typical angina	66
					1–Atypical angina	216
					2–Non-anginal pain	283
					3–Asymptomatic	625
4	Resting bps	Resting blood pressure (mm/Hg)	Continuous numeric	(0–200)	Blood pressure measured after hospital admission, in mmHg	
5	Cholesterol	Serum cholesterol	Continuous numeric	(0–603)	Blood cholesterol level measured in mg/dL	
6	Fasting blood sugar	Fasting blood sugar	Binary	0,1	0–Less than 120 mg/dL	936
					1–Greater than 120 mg/dL	254
7	Resting ECG	Resting ECG	Ordinal	0,1,2	0–Normal result	684
					1–ST-T wave abnormality	151
					2–Left ventricular hypertrophy	325
8	Max Heart Rate	Maximum heart rate	Continuous numeric	(60–202)	Maximum heart rate during exercise	
9	Exercise angina	Exercise-induced angina	Binary	0,1	0–No occurrence	729
					1–Occurrence present	461
10	Old Peak	Old peak	Continuous numeric	(-0.2.6–6.2)	ST depression induced by exercise relative to rest in ECG test	
11	ST Slope	ST slope	Ordinal	1,2,3	0–	1
					1–Upsloping	526
					2–Flat	582
					3–Downsloping	81

12	Target	Target	Binary	0,1	0–No heart disease	561
					1–Heart disease	629

Table 1 introduces the indicators used, including the gender of the patients, the type of chest pain, fasting blood sugar, resting ECG results, and other characteristics related to the patient's condition, such as ST slope and exercise angina. The ST slope is one of the important electrocardiographic indicators in exercise testing, which indicates the direction and extent of changes in the ST segment relative to the baseline ECG. This variable is classified into three states: upward, flat, and downward. The flat slope and especially the downward slope of the ST segment indicate a decrease in blood supply to the heart muscle (myocardial ischemia) and are used as one of the most important diagnostic indicators in identifying and predicting coronary artery disease. This information helps researchers and physicians make more accurate clinical decisions. Using these data and decision tree models can help in the early identification of heart diseases and making effective treatment decisions. Data extracted from the heart disease dataset, in addition to being used in this study, can be the basis for further research in this field.

Results

Early diagnosis of heart diseases is one of the main challenges in medicine. Given the high prevalence of these diseases, the use of machine learning algorithms to analyze medical data and predict patient status is important. Before modeling, data preprocessing steps, including missing value management, feature selection using principal component analysis, rough set methods, and data normalization, were performed to improve the performance of the models.

To evaluate the accuracy of machine learning models, several criteria were used, including precision (observation of correct predictions relative to the total number of predictions), F1 score (weighted average of precision and recall: a balanced measure for evaluating classification models), and recall (ratio of identified positive samples to the total number of true positive samples).

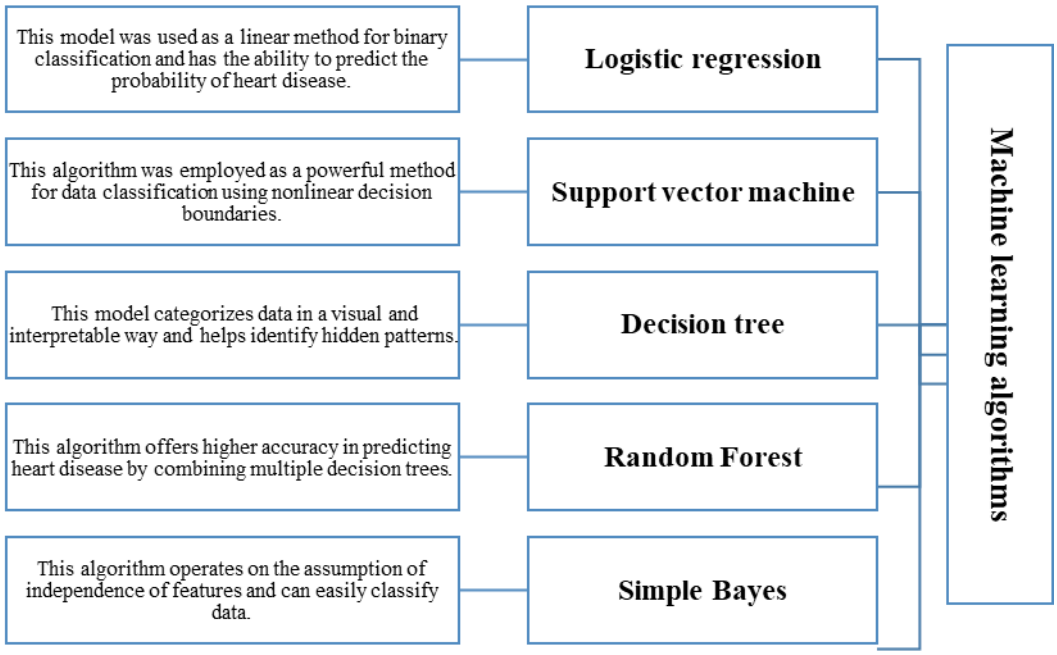

Figure 1. Machine learning algorithms used in this research.

Clustering with anomaly model

In the clustering model, two clusters were identified, and outliers were identified. The first cluster had 736 samples, and the second cluster had 454 samples. Based on the clustering results, Table 2 shows the characteristics of the clusters created. These observations indicate that the second cluster had a higher prevalence of risk factors for heart disease, such as older age, higher type of chest pain, higher resting blood pressure, and higher peak values. In addition, all individuals in the second cluster experienced exercise-induced angina, indicating a higher severity of heart disease compared to the first cluster.

Table 2. Characteristics of the created clusters.

First cluster (736 records)	Second cluster (454 records)
The mean ST slope is 1.433 with a standard deviation of 0.579. The mean age is 52.258 years with a standard deviation of 9.665 years. The mean type of chest pain is 2.914 with a standard deviation of 0.976. The mean cholesterol level is 213.117 mg/dL with a standard deviation of 95.197 mg/dL. The mean exercise angina is 0.01, indicating a low prevalence. The mean fasting blood sugar is 0.194, indicating a relatively low percentage of high blood sugar. The mean maximum heart rate is 147.514 bpm with a standard deviation of 25.073 bpm. The mean old peak is 0.617 with a standard deviation of 0.971.	The mean ST slope is 1.934 with a standard deviation of 0.527, which is higher than the first cluster. The mean age is 56.09 years with a standard deviation of 8.319 years, which is higher than the first cluster. The mean type of chest pain is 3.749 with a standard deviation of 0.566, which is higher than the first cluster. The mean cholesterol level is 901.205 mg/dL with a standard deviation of 734.110 mg/dL, which is lower than the first cluster. The mean exercise angina is 1.0, which indicates that all individuals in this group have experienced exercise angina. The mean fasting blood sugar is 0.244, which is slightly higher than the first cluster.

The mean resting blood pressure is 130.035 mmHg with a standard deviation of 17.635 mmHg. The mean resting ECG is 0.673 with a standard deviation of 0.879. The mean gender is 0.701, indicating a higher proportion of males in this group	The mean maximum heart rate is 127.119 bpm with a standard deviation of 20.766 bpm, which is lower than the first cluster. The mean of the old peak is 1.419 with a standard deviation of 1.081, which is higher than the first cluster. The mean of the resting blood pressure is 135.588 mmHg with a standard deviation of 19.025 mmHg, which is higher than the first cluster. The mean of the resting ECG is 0.74 with a standard deviation of 0.855, which is higher than the first cluster. The mean of the gender is 0.866, which indicates a higher proportion of males in this group compared to the first cluster

Rule Extraction Model

The proposed rule extraction model consists of a set of association rules that examine the relationships between different features in heart disease data. These rules can help doctors and researchers better understand the factors affecting patients' heart conditions.

Table 3. Extracted rules.

Consequent	Antecedent	Support %	Confidence%
target	exercise angina and ST slope = 2.0 and chest pain type = 4.0 and sex	18.91	94.67
target	exercise angina and ST slope = 2.0 and resting ecg = 0.0 and chest pain type = 4.0 and sex	10.17	94.21
target	fasting blood sugar and chest pain type = 4.0	12.94	94.16
target	fasting blood sugar and chest pain type = 4.0 and sex	11.43	94.12
target	ST slope = 2.0 and chest pain type = 4.0 and sex	27.39	93.25
target	exercise angina and ST slope = 2.0 and resting ecg = 0.0 and chest pain type = 4.0	12.02	93.01
target	ST slope = 2.0 and resting ecg = 0.0 and chest pain type = 4.0 and sex	15.46	92.39
target	fasting blood sugar and ST slope = 2.0	11.85	92.2
target	exercise angina and ST slope = 2.0 and chest pain type = 4.0	22.18	92.05
target	exercise angina and chest pain type = 4.0 and sex	26.89	91.88
target	exercise angina and resting ecg = 0.0 and chest pain type = 4.0 and sex	13.78	91.46
target	fasting blood sugar and ST slope = 2.0 and sex	10.76	91.41
target	exercise angina and ST slope = 2.0 and sex	24.03	91.26
target	exercise angina and ST slope = 2.0 and resting ecg = 0.0 and sex	12.86	90.85
target	exercise angina and ST slope = 2.0 and resting ecg = 0.0	14.96	89.89
target	ST slope = 2.0 and resting ecg = 0.0 and chest pain type = 4.0	18.24	89.86
target	exercise angina and chest pain type = 4.0	30.84	89.37
target	exercise angina and ST slope = 2.0	27.65	89.36

target	exercise angina and resting ecg = 0.0 and chest pain type = 4.0	16.05	89.01
target	ST slope = 2.0 and chest pain type = 4.0	32.86	88.75
target	ST slope = 2.0 and resting ecg = 0.0 and sex	21.26	87.35
target	exercise angina and sex	33.61	86.5
target	ST slope = 2.0 and sex	39.5	86.38
target	resting ecg = 2.0 and chest pain type = 4.0 and sex	11.34	85.93
target	exercise angina and resting ecg = 0.0 and sex	17.82	83.96
target	exercise angina	38.74	83.08
target	exercise angina and resting ecg = 0.0	20.5	81.97
target	chest pain type = 4.0 and sex	43.7	81.92
target	ST slope = 2.0 and resting ecg = 0.0	26.55	80.38
target	resting ecg = 0.0 and chest pain type = 4.0 and sex	23.95	80

Table 3 presents the extracted rules. The following is an analysis and interpretation of some of these rules, the rules with the highest support and confidence:

Rule 1: (exercise angina and ST slope = 2.0 and chest pain type = 4.0 and sex)

Support 18/91, Confidence: 94/67 This rule shows that the presence of exercise angina, ST slope equal to 2.0 and chest pain type equal to 4.0 are strongly associated with gender.

Rule 3 (fasting blood sugar and chest pain type = 4.0)

Support 12/94, Confidence: 94/16 This rule shows that fasting blood sugar level and chest pain type can be considered as a risk factor for cardiac conditions.

Rule 8 (fasting blood sugar and ST slope = 2.0)

Support: 10/76, Confidence: 91/41 This rule indicates that an ST slope of 2.0 and fasting blood sugar levels can be significantly associated with gender.

These extracted rules indicate significant associations between different features in the data related to heart disease. In general, the presence of exercise-induced angina, the type of chest pain, and the ST slope are among the key factors in determining the cardiac status of patients. Extracting association rules can be used as an effective tool in data analysis and identifying hidden patterns in medical data. This information can help doctors make better treatment decisions and improve the quality of patient care.

Classification models

The classification models are reviewed in Table 4. The performance of different classification models is compared. Criteria such as construction time, potential benefit, overall accuracy, and area under the curve are examined. The build time for all models was less than a minute. In terms of potential gain, the C5 model had the highest gain with 2,803,333 and also provided the highest overall accuracy (93.26%). The AUC metric shows that the KNN algorithm model has a high ability to distinguish between classes with a value of 0.972. The number of features used in the models also varies; most models used 11 features, but models such as CHAID and Decision List used fewer features. In the overall analysis, the C5 model performed better than the other models due to its high accuracy and significant gain. The Nearest Neighbor algorithm also has high quality with a high AUC, but its gain and accuracy are lower than the C5 model. On the other hand, the decision list model showed the weakest

performance because it has lower accuracy and AUC than the other models.

Table 4. Classification models used in the research.

Model	Model Abbreviation	Build Time (minutes)	Maximum Profit	Interval with Highest Profit	Increase at Highest Percentage	Number of Features	Area Under Curve
K-Nearest Neighbors Algorithm	KNN	<1	2529/247	53	1.892	11	0.972
Chi-square Automatic Interaction Detection	CHAID	<1	2303/939	50	1.824	10	0.935
C5.0 Decision Tree Algorithm	C5	<1	2803/333	53	1.824	11	0.965
Bayesian Network	Bayesian Network	<1	2350	53	1.818	11	0.935
Support Vector Machine	SVM	<1	2450	52	1.818	11	0.949
Neural Network	Neural Net	<1	2370	50	1.802	11	0.937
Decision List	Decision List	<1	1803/139	33	1.801	7	0.789
Classification & Regression Tree	C&R Tree	<1	2253/559	53	1.754	11	0.905

The "Interval where the model made the most profit" column shows which part of the data, or within which percentage of the data, the model made the most profit. This usually refers to the top percentage of the data where the model made the most profit. For example, if the value in this column is 53, it means that the model made the most profit in the top 53% (for example, in the category of patients with the highest probability of heart disease). This information can be very useful for determining important parts of the data that should be targeted. The "Increase in top percent" column shows how much the model improved or was more effective at identifying the best part (for example, the 10% with the highest probability of disease) compared to chance. In other words, it shows how much better the model was at the highest part of interest (such as the highest probability of disease). The "Area under the curve" column shows the overall performance of the model in distinguishing between classes (for example, positive and negative). The closer its value is to 1, the better the model performs. This metric is commonly used to evaluate classification models. Figure 1 shows the AUC graph.

Journal of Ilam University of Medical Sciences, Volume 34, Issue 1, 2026

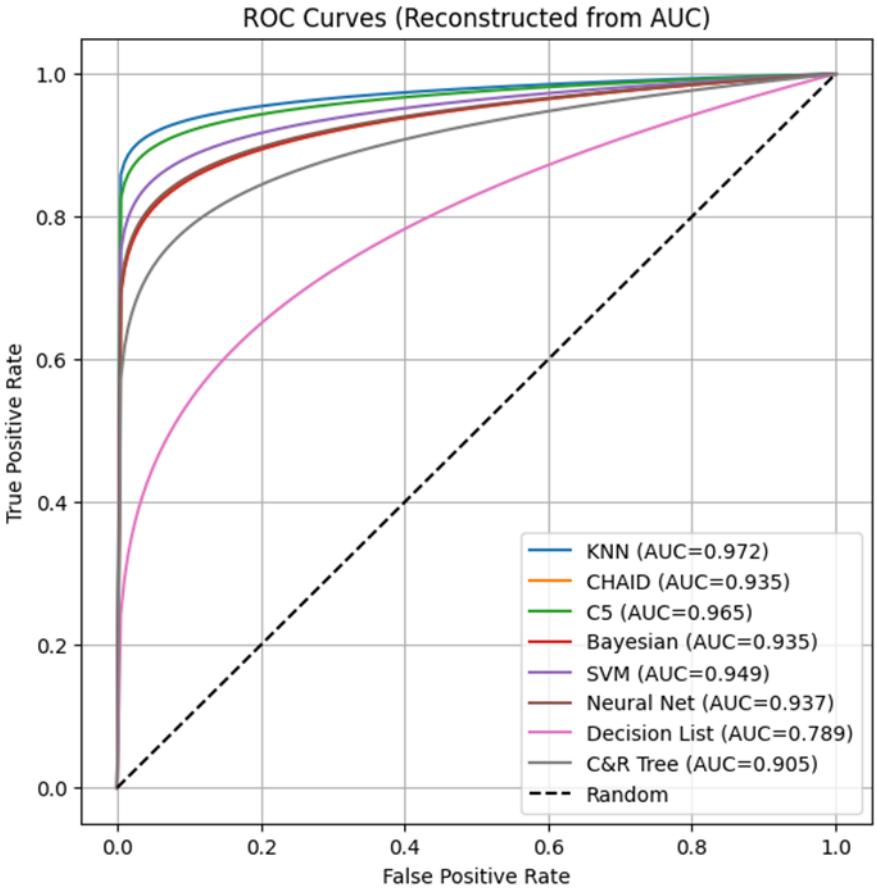

Figure 2. AUC chart of the models used.

Figure 2 compares the discrimination performance of different models based on the area under the ROC curve (AUC). As can be seen, the KNN algorithm with an AUC of 0.972 shows the best ability to discriminate between classes, followed by the C5.0 model and the support vector machine (SVM), which indicates the appropriate predictive power of these models. In contrast, the Decision List model with an AUC of 0.789 has the least efficiency among the models, and its discrimination power is weaker than other methods. In general, the results show that most of the data mining models used have acceptable accuracy with an AUC greater than 0.9 and have a good ability to classify data, although the selection of the best model also depends on the balance between accuracy and complexity of the model.

Neural Network

Table 5 shows the performance of a neural network model in predicting heart disease status. This table contains information about the correct and incorrect predictions of the model in two different categories.

1. Category 1.000 (presence of heart disease): 554 people were correctly identified as having heart disease, 75 people were incorrectly identified as not having heart disease, and 88.1% of the predictions in this category were correct.

2. Category 0.000 (absence of heart disease): 81 people were incorrectly identified as having heart disease, 480 people were correctly identified as not having heart disease, and 85.6% of the predictions in this category were correct.

The neural network model performs well in identifying the presence of heart disease, especially in category 1.000, which has an accuracy of 88.1%. However, the accuracy in the category 0.000 is slightly lower. The overall accuracy of 86.9% indicates that the model was generally successful in predicting the status of heart disease

Table 5. Performance of the neural network model in predicting heart disease status.

Classification	Observed	Predicted 1.000	Predicted 0.000	Correct Percentage
Training Data	1.000	554	75	88.1%
	0.000	81	480	85.6%
	Overall Percentage	53.4%	46.6%	86.9%

The most important variables in the neural network model, in order of priority, are: OLD peak, maximum heart rate, cholesterol, and ST slope

CHAID Model

The CHAID tree is a method used in statistics and machine learning for classification and regression problems. This method automatically identifies and analyzes important interactions between independent and dependent variables based on chi-square statistics. It uses a dataset to build a decision tree.

First, the model evaluates the independent variables using chi-square statistics and selects the variable with the greatest impact on the dependent variable. Then, it splits the data according to the values of the selected independent variable and continues this process until all branches or nodes lead to distinct outcomes or decisions.

This method has several advantages, including applicability to both discrete and continuous data, interpretability, and the ability to identify complex interactions. The CHAID tree model clearly illustrates the relationships among different features and can be used as an effective tool for predicting heart disease status.

Key features identified by the CHAID tree model include:

• ST slope: One of the main factors in determining heart disease status.

• Chest pain type: Clearly influential in predicting disease status.

• Cholesterol and exercise-induced angina: These features also play an important role in predicting the presence of heart disease.

C5.0 Decision Tree Model

A decision tree is one of the powerful and common tools for classification and prediction. A decision tree is capable of generating human-understandable descriptions of the relationships in a data set. A decision tree can easily understand rules and does not require much computation. Previous studies show that decision trees can produce highly accurate classifiers. Like any other learning algorithm, this algorithm consists of two steps: learning and prediction. In the learning step, the algorithm tries to create a tree from the training data. In the prediction step, the target variable is predicted based on the characteristics of a new data sample and along a path in the tree.

C5.0 is a prediction model that reaches more complex decisions based on a series of questions. Each node of the decision tree poses a question and is guided to one of the branches of the tree based on the answer to that question. Finally, the decision tree reaches a prediction for the input data. C5 is used with advantages such as

simplicity and understandability, the ability to control overfitting, and usability in large data sets.

The maximum tree depth is an important parameter in decision tree algorithms. This parameter determines how deep the tree can go and determines the number of decision steps in the tree. Whenever the depth of the tree increases, the number of branches and the number of decisions in the tree also increase. This may improve the performance of the algorithm on the training data. However, if the depth of the tree is increased too much, the tree may overfit the training data and reduce the accuracy of predictions on new data. Setting the maximum depth of the tree to an appropriate level helps the accuracy and performance of the algorithm. To determine this parameter, evaluation methods such as cross-validation can be used, and the algorithm's performance on both the training and evaluation data can be examined. Also, automatic parameter selection methods can be used to determine the best maximum depth of the tree. The number of terminal nodes is another important parameter in decision tree algorithms. Terminal nodes, also called "leaves," are the nodes that hold the final decision of the tree and divide the training data into different categories. The number of terminal nodes in the decision tree determines how many different categories the algorithm produces as a result of its prediction. The more terminal nodes, the more complex and detailed the tree becomes, and the more accurate it can be in predicting the training data. For example, if the number of terminal nodes is 2, the tree will only recognize two different classes. But if the number of terminal nodes is 10, the tree can recognize ten different classes. Determining the appropriate number of terminal nodes can improve the performance and accuracy of the algorithm in predicting new data. However, it should be noted that increasing the number of terminal nodes may lead to overfitting of the tree to the training data and reduce the accuracy of prediction on new data. Therefore, determining this parameter can also be done using evaluation methods such as cross-validation or automatic parameter selection.

The C5.0 decision tree model is a machine learning algorithm used for classification problems. This model is divided into a tree based on various features and their values to predict the status of heart disease. Key features extracted from the C5.0 decision tree model are:

• ST slope: It is one of the main factors in determining the status of the heart.

• Chest pain type: The type of chest pain is clearly influential in predicting the status of the disease.

• Cholesterol and Exercise Angina: These features also play an important role in predicting the presence of heart disease.

Bayesian Model

The Bayesian model is a statistical method that analyzes data and predicts outcomes based on Bayesian theory. In this model, conditional probability is used to evaluate relationships between variables. According to the results of the Bayesian model, people with heart disease are more likely to be in the older age groups (47.6 to 57.4). While people without heart disease are more likely to be in the younger age group (<= 37.8).

Logistic Regression Model

The logistic regression model is an algorithm for predicting whether an observation belongs to one of two classes in machine learning. Logistic regression classifiers predict the target class based on the calculated scores. A logistic function is used to convert probabilities into binary values that can be used for prediction (28).

The output information of this model is shown in Table 6.

Table 6. Performance of the neural network model in predicting heart disease status.

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1645.802			
Final	837.057	808.745	16	.000

Pseudo R-Square	
Cox and Snell	.493
Nagelkerke	.658
McFadden	.491

The logistic regression model is used to predict the probability of an event (here, the probability of having heart disease) based on various independent variables. Interpretation of the results of the coefficient (B) is:

• Age: For each year of age, the probability of having heart disease increases by 1.018, but this effect is not statistically significant (p-value = 0.111).

• Blood pressure: The effect of blood pressure is also not significant (p-value = 0.165).

• Cholesterol: Each unit increase in cholesterol level reduces the probability of having heart disease by 0.997 (p-value < 0.001).

• Maximum heart rate: This variable also has no significant effect (p-value = 0.109).

Old peak: Each unit increase in this variable increases the probability of having heart disease by 1.509 (p-value < 0.001).

• Gender: Women (sex = 0) were 79% less likely to develop heart disease than men (sex = 1) (p-value < 0.001).

• Chest pain type: Any type of chest pain (1, 2, 3) was significantly associated with a reduced risk of heart disease (p-value < 0.001).

• Fasting blood sugar: People with normal fasting blood sugar (fasting blood sugar = 0) were 59% less likely to develop heart disease (p-value < 0.001).

• Exercise angina: People without exercise angina (exercise angina = 0) were 57% less likely to develop heart disease (p-value < 0.001).

• ST segment elevation: An ST segment elevation of 2 was significantly associated with an increased risk of heart disease (p-value < 0.001).

The logistic regression model shows that various variables such as cholesterol, old peak, type of chest pain, and exercise-induced angina have a significant effect on the probability of having heart disease. The model fits well: According to Chi-Square and Pseudo R-squared, the model explains the data well.

Proposed Hybrid Model

In the proposed model, the data is first divided into homogeneous groups using an appropriate clustering algorithm. This step is performed in order to reduce the computational complexity and increase the accuracy of the classification process. After clustering, each of the 8 proposed models presented in Table 4 is run separately in each cluster. The output of each model determines the classification status of each sample. In the next step, the results of these

models are aggregated, and the final decision is made based on the majority rule. Specifically, if at least 5 out of 8 models classify a sample into a particular class, that class is selected as the final output of the hybrid model. In cases where there is no agreement between the models, i.e., 4 models classify a sample into one class and 4 models into another class, a secondary decision-making mechanism based on the KNN model is used. This algorithm performs the final classification by considering the nearest neighbors of the sample in question. This hybrid approach not only increases the classification accuracy but also improves the reliability of the predictions by reducing the dependence on the results of a single model. To evaluate the performance of the proposed model, in Table 7, this model is compared with other heart disease prediction models. The results of this comparison show that the proposed model performs better than the existing models in terms of accuracy, sensitivity, and specificity. This performance improvement results from the intelligent combination of clustering methods, the use of several diverse models, and the application of majority-based and KNN decision-making mechanisms.

Table 7. Comparison of the proposed system with existing heart disease prediction systems in the heart disease dataset.

Reference	Year	Classifiers Used	Methodology Applied	Maximum Accuracy (%)
(19)	2019	NB Net, C 4.5, MLP, PART, Bagging, Boosting, Majority Voting, Stacking	Ensemble techniques such as Bagging and Boosting are used to improve prediction accuracy.	85.48
(18)	2020	Logistic Regression, Support Vector Machine, Nearest Neighbor	Feature normalization and dimensionality reduction using Principal Component Analysis (PCA)	87.00
(12)	2020	Logistic Regression, Decision Tree, Naive Bayes, Gaussian	Dimensionality reduction was performed via Singular Value Decomposition (SVD).	82.75
(25)	2022	Stochastic Gradient Descent classifiers, Logistic Regression, Support Vector Methods, Naive Bayes, ConvSGLV, and Ensemble	Majority voting, CNN used for feature extraction; flatten layer transforms 3D data into 1D so ML models operate on 1D data.	93.00
(24)	2023	Logistic Regression, Nearest Neighbor, Decision Tree, Gaussian, Support Vector Machine, Random Forest	Hyperparameter tuning using GridSearchCV	87.91
Present Study	–	Nearest Neighbor, CHAID Tree, C5.0 Tree, Bayesian Network, Support Vector Machine, Neural Network, Decision List, and Logistic Regression	Hybrid method	93.44

Discussion

Early diagnosis of heart diseases, as one of the fundamental challenges in the field of medicine and public health, can have a significant impact on improving the quality of life of patients and reducing mortality from these diseases.

In this study, using machine learning algorithms, including logistic regression, support vector machine, decision tree, random forest, and Naive Bayes, medical data were analyzed, and the presence of heart diseases was predicted. The results obtained showed the high efficiency of these algorithms in early detection of diseases. In particular, the logistic regression, nearest neighbor, decision tree, Gaussian, vector machine, random forest, and logistic regression models showed better performance with an accuracy of 93.44% compared to other models. Including the study by Latha et al. (19), who achieved an accuracy of 85.48% using combined techniques. This difference could be due to the use of an appropriate combination of features and algorithms.

Several studies have pointed out the importance of features such as ST segment elevation, chest pain type, and fasting blood glucose in predicting heart disease (17, 20, 22). The results of this study also confirmed these features as key factors. This agreement indicates that the proposed hybrid model performed well not only in terms of accuracy but also in terms of clinical interpretability. One of the highlights of this study was the high performance of the hybrid model (93.44%) compared to the individual algorithms. Combining algorithms likely compensated for the weaknesses of each algorithm and exploited their strengths. For example, while simpler models such as logistic regression are suitable for linear data, more complex models such as neural networks are capable of identifying nonlinear patterns. Combining these two approaches increased the overall accuracy of the model. In addition, careful data preprocessing, including normalization and feature selection, played an effective role in improving model performance. For example, using features such as fasting blood sugar and chest pain type, which are strongly associated with heart disease, enabled the model to identify key patterns.

Conclusion

The results of this study showed that the appropriate selection of algorithms and optimization of data preprocessing processes can significantly increase the accuracy of heart disease prediction. Also, the use of feature selection methods such as principal component analysis and rough set-based methods reduced the data dimensions and increased the efficiency of the model. The findings showed that machine learning models have a high ability in the early diagnosis of heart disease and can be used as a practical tool in clinical settings. Due to their high accuracy and appropriate processing speed, these models can reduce the time and cost of diagnosis, especially in resource-limited medical centers. In addition, identifying key features helps doctors make more accurate decisions and focus on the main risk factors. Finally, the development of larger datasets, the use of deep learning methods, the investigation of new risk factors, and the expansion of interdisciplinary collaborations can further improve the accuracy and efficiency of heart disease prediction models.

Funding

The authors did not receive any financial support or funding for the research, authorship, and/or publication of this article.

CRediT authorship contribution statement

Conceptualization, Methodology, Validation, Formal Analysis, Investigation: MYT, Resources, Software, Data Curation, Writing–Original Draft Preparation, Writing–Review & Editing, Visualization, Supervision, Project Administration: MM.

Conflict of interest

The authors declare no conflict of interest.

Ethical considerations

Given that this study used data from global databases and did not involve any contact with patients or national data, a code of ethics was not adopted. The authors avoided data fabrication, falsification, plagiarism, and misconduct.

Acknowledgements

The authors thank the all of participants and persons who help in this project.

References

1. World Health Organization. Cardiovascular diseases, key facts [Internet]. 2021. Available from: https://www.who.int/news-room/factsheets/detail/cardiovascular-diseases-(cvds).
2. Choudhury RP, Akbar N. beyond Diabetes: A Relationship between Cardiovascular Outcomes and Glycaemic Index. Cardiovasc Res. 2021; 117: 97-98. doi: 10.1093/cvr/cvab162.
3. Ordonez C. Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction. IEEE Tran INF Technol Biomed. 2006; 10:334- 43. doi: 10.1109/titb.2006.864475.
4. Magesh G, Swarnalatha P. Optimal Feature Selection through a Cluster-Based DT Learning (CDTL) in Heart Disease Prediction. Evol Intell. 2021; 14:583-593. doi: 10.1007/s12065-019-00336-0.
5. Chowdary KR, Bhargav P, Nikhil N, Varun K, Jayanthi D. Early Heart Disease Prediction Using Ensemble Learning Techniques. J Phys Conf Ser. 2022; 2325:012051. doi: 10.1088/1742-6596/2325/1/012051.
6. Liu J, Dong X, Zhao H, Tian Y. Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion. Processes. 2022; 10:749. doi: 10.3390/pr10040749.
7. Sheikhi Chaman MR, Barati O, Hamidi H, Abdoli Z. The role of clinical economics in the governance of the health system. Med Purif. 2022; 31:81-85. [Persian].
8. Uddin S, Khan A, Hossain ME, Moni MA. Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. BMC Med Inform Decis Mak. 2019; 19:281. doi: 10.1186/s12911-019-1004-8.
9. Patro SP, Nayak GS, Padhy N. Heart Disease Prediction by Using Novel Optimization Algorithm: A Supervised Learning Prospective. Inform Med Unlocked. 2021; 26:100696. doi: 10.1016/j.imu.2021.100696.
10. Song Q, Zheng YJ, Yang J. Effects of Food Contamination on Gastrointestinal Morbidity: Comparison of Different Machine Learning Methods. Int J Environ Res Public Health. 2019; 16:838. doi: 10.3390/ijerph16050838.
11. Pasha SJ, Mohamed ES. Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction. IEEE Access. 2020; 8:184087-184108. doi: 10.1109/ACCESS.2020.3028714.
12. Ananey-Obiri D, Sarku E. Predicting the Presence of Heart Diseases Using Comparative Data Mining and Machine Learning Algorithms. Int J Comput Appl. 2020; 176:17-21. doi: 10.5120/ijca2020920034.
13. Mohan S, Thirumalai C, Srivastava G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access. 2019; 7:81542-81554. doi: 10.1109/ACCESS.2019.2923707.
14. Kodati S, Vivekanandam R. Analysis of Heart Disease Using Data Mining Tools Orange and Weka. Glob J Comput Sci Technol C. 2018; 18:17-21.
15. Shah SMS, Batool S, Khan I, Ashraf MU, Abbas SH, Hussain SA. Feature Extraction through Parallel Probabilistic Principal Component Analysis for Heart Disease Diagnosis. Physica A: Stat Mech Appl. 2017; 482:796-807. doi: 10.1016/j.physa.2017.04.113.
16. Perumal R. Early Prediction of Coronary Heart Disease from Cleveland Dataset Using Machine Learning Techniques. Int J Adv Sci Technol. 2020; 29:4225-34.
17. Gazelog˘lu C. Prediction of Heart Disease by Classifying with Feature Selection and Machine Learning Methods. Prog Nutr. 2020; 22:660-670.
18. Reddy KVV, Elamvazuthi I, Aziz AA, Paramasivam S, Chua HN, Pranavanand S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. Appl Sci. 2021; 11:8352. doi: 10.3390/app11188352.
19. Pavithra V, Jayalakshmi V. Hybrid Feature Selection Technique for Prediction of Cardiovascular Diseases. Mater Today Proc. 2022.; 20: 11: 4871-8. doi: 10.14704/NQ.2022.20.11. NQ66495.
20. Latha CBC, Jeeva SC. Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques. Inform Med Unlocked. 2020; 16:100203.
21. Bashir S, Qamar U, Khan FH, Javed MY. MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble. Arab J Sci Eng. 2014; 39:7771-83. doi: 10.1007/S13369-014-1315-0.
22. Tama BA, Im S, Lee S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. Biomed Res Int. 2020; 2020:9816142. doi: 10.1155/2020/9816142.
23. Alqahtani A, Alsubai S, Sha M, Vilcekova L, Javed T. Cardiovascular Disease Detection Using Ensemble Learning. Comput Intell Neurosci. 2022; 5267498. doi: 10.1155/2022/5267498.
24. Trigka M, Dritsas E. Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models. Sensors. 2023; 23: 1193. doi: 10.3390/s23031193.
25. Rustam F, Ishaq A, Munir K, Almutairi M, Aslam N, Ashraf I. Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction. Diagnostics. 2022; 12:1474. doi: 10.3390/diagnostics12061474.
26. Cyriac S, Sivakumar R, Raju N, Woon Kim Y. Heart Disease Prediction Using Ensemble Voting Methods in Machine Learning. In: Proceedings of the 2022 13th International Conference on Information and Communication Technology Convergence (ICTC);

2022 Oct 19–21; Jeju Island, Republic of Korea; 2022:1326-31. doi: 10.3390/pr11041210.
27. Jan M, Awan AA, Khalid MS, Nisar S. Ensemble Approach for Developing a Smart Heart Disease Prediction System Using Classification Algorithms. Res Rep Clin Cardiol. 2018; 9: 33-45. doi: 10.2147/RRCC.S172035.
28. Li C, Chang W. Cardiovascular Disease Prediction Using Machine Learning: A Review. Int J Comput Appl. 2020; 176:15-22. doi: 10.1038/s41598-020-72685-1.