Iran Journal of Computer Science (2023) 6:397–417
https://doi.org/10.1007/s42044-023-00148-7

RESEARCH

Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison

Md. Imam Hossain[1] · Mehadi Hasan Maruf[1] · Md. Ashikur Rahman Khan[1] · Farida Siddiqi Prity[1] · Sharmin Fatema[1] · Md. Sabbir Ejaz[1] · Md. Ahnaf Sad Khan[2]

Received: 27 January 2023 / Accepted: 18 May 2023 / Published online: 12 June 2023

Abstract
Consolidated efforts have been made to enhance the treatment and diagnosis of heart disease due to its detrimental effects on society. As technology and medical diagnostics become more synergistic, data mining and storing medical information can improve patient management opportunities. Therefore, it is crucial to examine the interdependence of the risk factors in patients' medical histories and comprehend their respective contributions to the prognosis of heart disease. This research aims to analyze the numerous components in patient data for accurate heart disease prediction. The most significant attributes for heart disease prediction have been determined using the Correlation-based Feature Subset Selection Technique with Best First Search. It has been found that the most significant factors for diagnosing heart disease are age, gender, smoking, obesity, diet, physical activity, stress, chest pain type, previous chest pain, blood pressure diastolic, diabetes, troponin, ECG, and target. Distinct artificial intelligence techniques (logistic regression, Naïve Bayes, K-nearest neighbor (K-NN), support vector machine (SVM), decision tree, random forest, and multilayer perceptron (MLP)) are applied and compared for two types of heart disease datasets (all features and selected features). Random forest using selected features has achieved the highest accuracy rate (90%) compared to employing all of the input features and other artificial intelligence techniques. The proposed approach could be utilized as an assistant framework to predict heart disease at an early stage.

Keywords Heart disease · Artificial intelligence · Logistic regression · Naïve Bayes · K-nearest neighbor · Multilayer perceptron

1 Introduction

Heart disease is a broad term for several conditions affecting blood vessels, arteries, and other organs, leading to incorrect heart function. Since the SARS-CoV-2 virus disrupts the human respiratory system and attempts to lower the amount of oxygen in the lungs, it has a significant negative influence on heart health and may even result in heart destruction [1, 2]. The development of atheromatous plaques, aberrant lipid metabolism, and the buildup of lipids and other chemicals in the circulation in the coronary arteries are all indications of heart disease. It can result in luminal narrowing or occlusion, which can cause myocardial ischemia, oxygen deprivation, or necrosis that manifests as chest discomfort, tightness in the chest, myocardial infarction, and other symptoms.

Heart disease is the leading cause of death in the modern world. According to the World Health Organization, heart disorders are expected to cause 12 million deaths worldwide each year [3]. Globally, 10.6 million new cases of coronary heart disease were reported in 2017, and 8.9 million individuals died as a consequence. In 2017, 126.5 million people experienced coronary heart disease [4]. Medical expenses for heart disease are anticipated to rise 41% in the US, from $126.2 billion in 2010 to $177.5 billion in 2040 [5].

Applications of machine learning (ML) are significantly changing the healthcare industry. ML is a branch of artificial intelligence (AI) technology that attempts to enhance the efficiency and precision of medical professionals' work [6]. ML offers an excellent opportunity for countries struggling with an overcrowded healthcare system and physician

✉ Md. Ashikur Rahman Khan
ashik@nstu.edu.bd

[1] Department of Information and Communication Engineering, Noakhali Science and Technology University, Noakhali, Bangladesh

[2] Department of Mechanical Engineering, Military Institute of Science and Technology, Dhaka, Bangladesh

shortage. It is essential to medicine because it can find patterns in massive data sets and facilitate the identification of risk or disease-related diagnostic indicators. In many medical applications, such as medical image analysis [7–10], language processing [11, 12], and tumor or cancer cell detection [13–17], ML methods can assist clinical management, and professionals explore excellent performance. Clinical treatment has started to incorporate ML classification algorithms. Knowledge can be extracted using classification techniques. Through reliable heart disease prediction, practitioners can make better decisions and improve patient outcomes. The early detection of heart disease has made extensive use of Machine Learning-based approaches, including support vector machine (SVM), K-nearest neighbor (K-NN), artificial neural network (ANN), decision tree, logistic regression (LR), AdaBoost (AB), Naive Bayes (NB), and fuzzy logic (FL), etc. [18, 19]. These machine learning-based expert medical decision systems have diminished the ratio of heart disease deaths.

The outcomes of the various methodologies show that various factors can impact the study's findings. Any study's conclusion will be influenced by numerous aspects, which include the method of data collection, the features used, the method used to clean the data, randomization, and standardization of the data. The researchers must understand how the input parameters in a dataset relate to one another and how this affects the final heart disease prediction accuracy.

In prior works, most studies have looked into how classifiers can reliably detect heart disease cases [20–28]. However, studies have yet to make an effort to examine all of the patient's conditions and pinpoint the crucial variables required for heart disease prediction. Studies in similar fields show that selecting the critical features affects how well the classifier framework performs. Instead of utilizing every feature in the feature space, finding the ideal combination of attributes is crucial. Before using a classification method, redundant qualities and completely unrelated attributes to a class should be found and eliminated. It is critical for data mining professionals working in the healthcare industry to understand how the risk factors recorded in a dataset are interdependent and how each affects the precision of heart disease prediction. Therefore, this study aims to explain the numerous risk variables for heart disease prediction comprehensively. Many risk factors in patient records are examined to determine the most crucial elements required for heart disease prediction. Nineteen features (age, gender, occupation, family history, smoking, obesity, diet, physical activity, stress, chest pain type, previous chest pain, edema, blood pressure systolic, blood pressure diastolic, heart rate, diabetes, troponin, ECG, target) have been collected from different hospitals in Bangladesh. This study has carried out the Correlation-based Feature Subset Selection method with the Best First Search (BFS) and a stepwise analysis to select the best collection of features. Finally, the datasets containing all features and chosen features are used to develop seven ML methods. This research has used popular ML techniques such as logistic regression, Naïve Bayes, K-nearest neighbor (K-NN), random forest, support vector machine, decision tree, and multilayer perceptron (MLP) models for heart disease prediction. Logistic regression models are versatile, have a powerful interpretation, and have been used to describe phenomena in diverse areas of medical research [29–31]. Examples of the use of logistic regression in medicine include determining the elements that will decide whether an improvement or no improvement will occur after an intervention [32, 33], the existence or absence of a disease in connection to a variety of factors [34], exploring the impacts of and linkages between many predictors [35, 36], determining which of a range of prospective predictors are essential [37, 38], and determining whether newly examined variables add to the predictive value [39]. Naïve Bayes is well applied in the field of health, performing predictive modeling for different diseases (brain, asthma, prostate, and breast cancer) [40]. The supervised learning algorithm K-NN is primarily employed for classification tasks. It has been extensively applied to disease prediction [41]. Reza et al. have used KNN to classify the liver disease dataset [42]. Random forest (RF) is used to identify the most robust predictors due to its favorable prediction performance in medical sectors [43–45]. SVM has many applications in the medical industry, such as breast cancer [46–48], skin cancer [49, 50], and many other issues relating to disease prognosis. SVM can also attain greater generalization ability in small sample classification assignments. It is also widely utilized in many other domains, including handwritten character recognition, text classification, image classification, and recognition [51–55]. The use of the Decision Tree model aids in the early detection of cancer [56, 57], diagnosing cardiac arrhythmias [58, 59], forecasting stroke outcomes [60–62], and assisting with chronic disease management [63, 64]. MLP has been successfully applied in a variety of medical fields, including disease prediction [65, 66], medical image recognition [67, 68], and gene selection [69, 70]. In this study, all classifiers' performances have been checked on full features and selected features in terms of classification accuracy. The study suggests which dataset is feasible with which classifier for designing high-level intelligent system for heart disease. There are no works that perform heart disease prediction on both datasets (datasets containing all features and selected features) using these seven algorithms and pick the best one based on accuracy. Accuracy, recall, precision, F1-score, and ROC-AUC score metrics are utilized to assess classification performance. The proposed system will assist clinicians in accurately determining whether or not a patient has heart disease. In this proposed system, the dataset of selected features outperforms the dataset of all features.

2 Previous works

Researchers have applied different ML techniques to build a model for predicting heart disease. Rajdhan et al. recommend using machine learning approaches (random forest, Naïve Bayes, logistic regression, and decision trees) to diagnose heart conditions [20]. The UCI machine learning repository served as the source of the database for this investigation. According to this study, the Random Forest technique has produced the best results, with an accuracy rate of 90% compared to other machine learning methods.

The use of machine-learning methods, such as logistic regression and K-NN, to predict and categorize patients with heart disease is recommended by Jindal et al. [21]. This investigation has proven that the K-NN algorithm performs best, with more than 88% accuracy. Sahoo et al. have suggested using a collection of methods, including Naïve Bayes, SVM, decision tree, K-NN, and logistic regression, to assess coronary artery disease data gathered from the UCI repository and comprising 13 crucial properties [22]. Since the SVM technique achieved more than 85% in the data analysis, this study has found that it provides the best accuracy.

Uyar et al. have established a Genetic Algorithm (GA) based on trained recurrent fuzzy neural networks (RFNN) for assessing heart diseases. In that investigation, the UCI heart disease dataset was employed [23]. Of the 297 patient data instances, 257 are used for training, and 45 are selected for testing. The findings showed that the testing set had a 97.78% accuracy rate. A precise hybrid approach for detecting coronary heart disease has been developed by Arabasadi et al. [24]. The strategy can enhance the neural network's performance by approximately 10% by upgrading its initial weights with a genetic algorithm that suggests better consequences for the neural network. Using their approach on the Z-Alizadeh Sani dataset, they have achieved accuracy, sensitivity, and specificity rates of 93.85%, 97.5%, and 92.5%, respectively. Sonawane and Patil have created MLP Neural Network-based prediction method for heart disease [25]. The system's neural network has input from 13 clinical characteristics. With a maximum accuracy of 98%, it is trained to employ the back-propagation algorithm to determine whether the patient has heart disease.

P.K. Anooj employed a weighted fuzzy rule-based system that automatically gathered information from the patient's data to diagnose heart disease [26]. The proposed methodology for heart disease prediction consists of two stages: an automated method for generating weighted fuzzy rules and creating a fuzzy rule-based decision support system. In the first stage, the data mining strategy, attribute selection, and attribute weightage method is used to generate the weighted fuzzy rules. The chosen weighted fuzzy rules and characteristics are then used to construct the fuzzy system. Olaniyi et al. introduced the diagnosis of heart disease using neural networks [27]. The intelligent system has been modelled using SVM and Feed-Forward Multilayer Perceptron. The recognition rates acquired from these models are then compared to choose the best model for the intelligent approach because of the significance of the intelligent system in the medical field. The SVM and Feed-Forward Multilayer Perceptron outputs are 85% and 87.5%, respectively.

Bhatla et al. describe various data mining techniques for assessing heart disease prediction [28]. The results show that neural networks with 15 attributes outperform all other data mining methods. Another study finding is that the Decision Tree can reach high accuracy by employing a GA and feature subset selection. Dehkordi et al. used the data mining technique to create a prediction model based on the prescription [71]. To improve the system's accuracy, they have suggested an algorithm called skating. An ensemble technique called skating is comparable to boosting and bagging. They have contrasted four different classification methods, including decision tree, Naive Bayes, K-NN, and Skating. They have demonstrated that skating is the most precise provided classifier. The accuracy of this classification algorithm is 73.17%. Jan et al. have used the ensemble of five different classification methods, including random forest, neural network, Naive Bayes, classification via regression analysis, and SVM, to build an ensemble data mining strategy employing two benchmark datasets (specifically Cleveland and Hungarian) that are obtained from a UCI repository [72]. Regression approaches are shown to be the least effective algorithm in that investigation, but Random Forest delivered a very high accuracy of 98.136%.

To enhance classification performance, Jyoti Soni et al. have utilized decision tree with GA; this is contrasted with the other two techniques, such as cluster-based classification and Naïve Bayes [3]. The proposed system's accuracy is found to be 99.2%. The effectiveness of the classification algorithms Random Forest and logistic regression for predicting the risk exposure of cardiovascular patients has been examined by Hend Mansoor et al. [73]. They have demonstrated that the performance of the logistic regression model outperforms that of the random forest classification algorithm. The accuracy of the Logistic Regression model is 89%, while the accuracy of the random forest model is 88%. Regression trees and standard classification trees have been compared for performance by Austin et al. [74]. Traditional logistic regression provides good performance in evaluating the possible existence of heart disease.

3 Methodology

Heart disease is becoming more prevalent for several reasons. Early detection of heart disease is essential for starting treatment. To fulfill this requirement, the strategy mentioned in this study discusses various ML techniques that will enable everyone to become aware of their risk early.

3.1 Overview of prediction model

The main goal of this study is to create a model as soon as feasible that can accurately and automatically predict heart disease. Data collection, pre-processing, feature extraction and selection, different ML algorithms, and performance analysis are the sections that outline the research methodology used to achieve the research goal.

The process begins with collecting patient information. Pre-processing of the data has been carried out to fill in missing values, eliminate ambiguous and redundant data, and standardize data. Feature selection and extraction techniques have been applied to select the most relevant attributes and combine attributes into a new reduced set of features. Two dataset types have been obtained for the proposed system: dataset with all attributes and dataset with selected attributes. Distinct AI techniques have been applied to both datasets to find the best dataset for heart disease prediction. Comparative analysis among all ML techniques has been conducted to discover the more efficient and accurate technique to build an efficient heart disease prediction model that can classify the heart disease correctly based on different performance metrics for both datasets. The fundamental steps in implementing heart disease detection that should be monitored to apply distinct AI techniques for heart disease identification with appropriate confidence are shown in Fig. 1.

3.1.1 Data collection and processing

Data have been collected from hospitals, diagnostic centers, and clinic centers in Bangladesh. Maximum data have been collected from hospital-admitted patients. Patients are questioned, test results are examined, and essential attribute data are collected. The dataset has 59 patients' test results and responses to various questions. Table 1 summarizes the specifications of the 19 attributes (18 independent, while the target attribute is dependent) that comprise the dataset.

The differences in the ranges of variables appear; hence, it is essential to normalize the data. Normalization is accomplished according to the equation as follows:

$$\text{Xnormalized} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Table 1 Dataset Description

Attribute	Type	Value
Age	Nominal	26–90
Gender	Nominal	Male, Female
Occupation	Nominal	Businessman, Driver, Farmer, Worker, Unemployed, Housewife, Service-holder
Family history	Nominal	Yes, No
Smoking	Nominal	Yes, No
Obesity	Nominal	Yes, No
Diet	Nominal	Normal, Abnormal
Physical activity	Nominal	Yes, No
Stress	Nominal	Normal, Abnormal
Chest pain type	Nominal	Typical angina, Atypical angina, Non-angina pain
Previous chest pain	Nominal	Yes, No
Edema	Nominal	Yes, No
Blood pressure systolic (bp_systolic)	Nominal	80–180
Blood pressure diastolic (bp_diastolic)	Nominal	60–120
Heart rate	Nominal	60–120
Diabetes	Nominal	Yes, No
Troponin	Nominal	Positive, Negative
ECG	Nominal	Normal, Abnormal
Target	Nominal	Class 1 (Patient has heart disease) and Class 0 (Patient has no heart disease)

where, Xnormalized is the updated normalized value. The lowest value of each feature is taken as '0', the highest value is considered as '1', and all other values are converted to an integer between '0' and '1'.

3.1.2 Feature extraction and selection

The technique of reducing a tremendous sum of information to a small number of related bits is known as feature extraction. The process of choosing a subset of pertinent characteristics for use in developing statistical and machine-learning models is known as feature selection. Minimizing the information makes it easier for the machine to form the model with less effort and speeds up the machine learning and generalization processes.

All attributes are not equally significant in the data we have gathered. The Proposed model has used the Correlation-based Feature Subset Selection Technique with the Best First

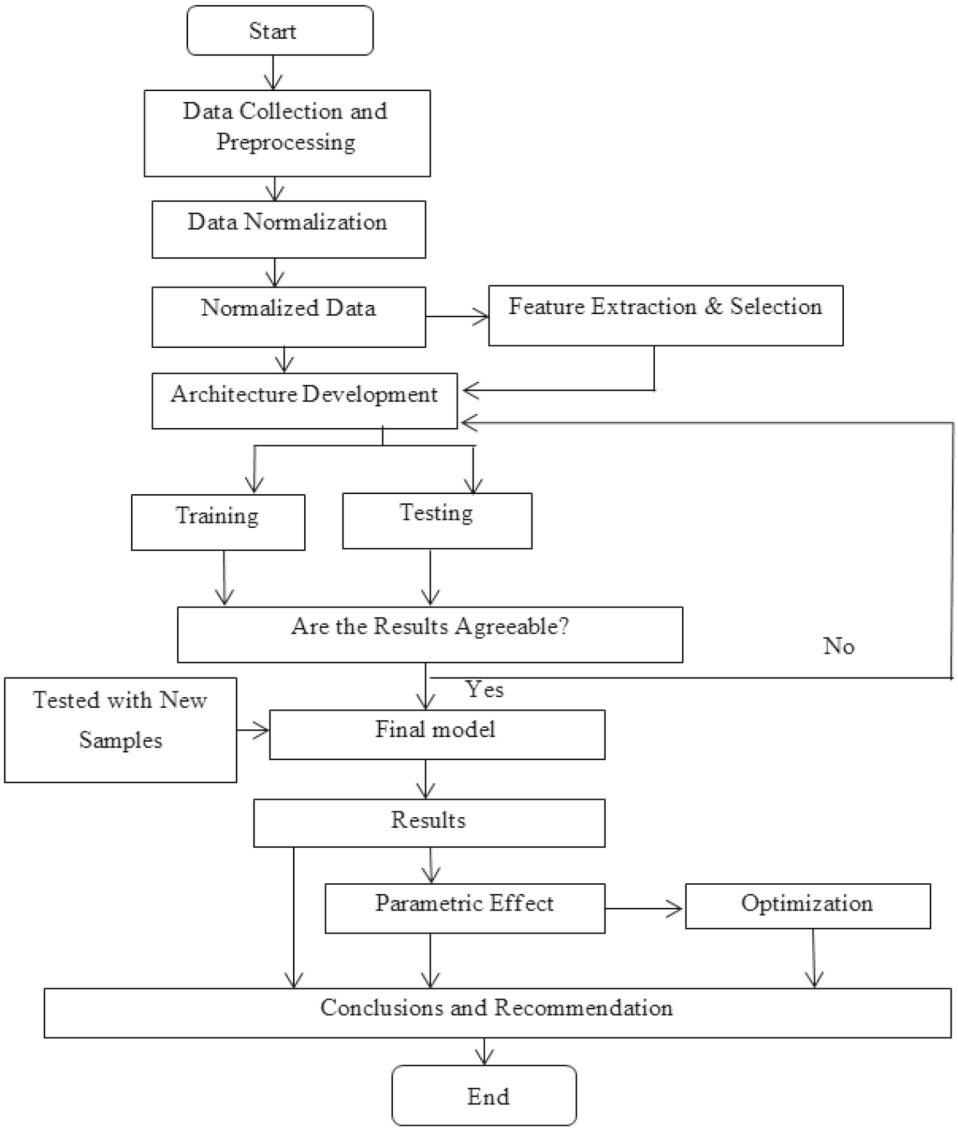


Fig. 1 Workflow diagram of heart disease

Search to extract relevant features from 19 attributes of heart patient records.

Correlation-based feature subset selection technique
The Correlation-based Feature Subset Selection algorithm combines this assessment formula with a suitable correlation measure and a heuristic search method. The correlation between the outside variable and the composite increases as the correlation between the outside variable and the components increases. The correlation between the composite and the external variable is more robust when the components have less inter-correlation. The correlation grows with the number of components in the composite between the external variable and the composite rises.

3.1.3 Machine learning techniques

The sample of a heart disease prediction model is given in Fig. 2, which has been used in this proposed system. It starts with the input parameters of the dataset. After the preprocessing, the dataset is fed to the proposed techniques. This study has used seven ML techniques for comparative analysis.

a. Logistic regression
b. Naïve Bayes
c. K-nearest neighbor (K-NN)
d. Support vector machine (SVM)
e. Decision tree
f. Random forest
g. Multilayer perceptron (MLP)

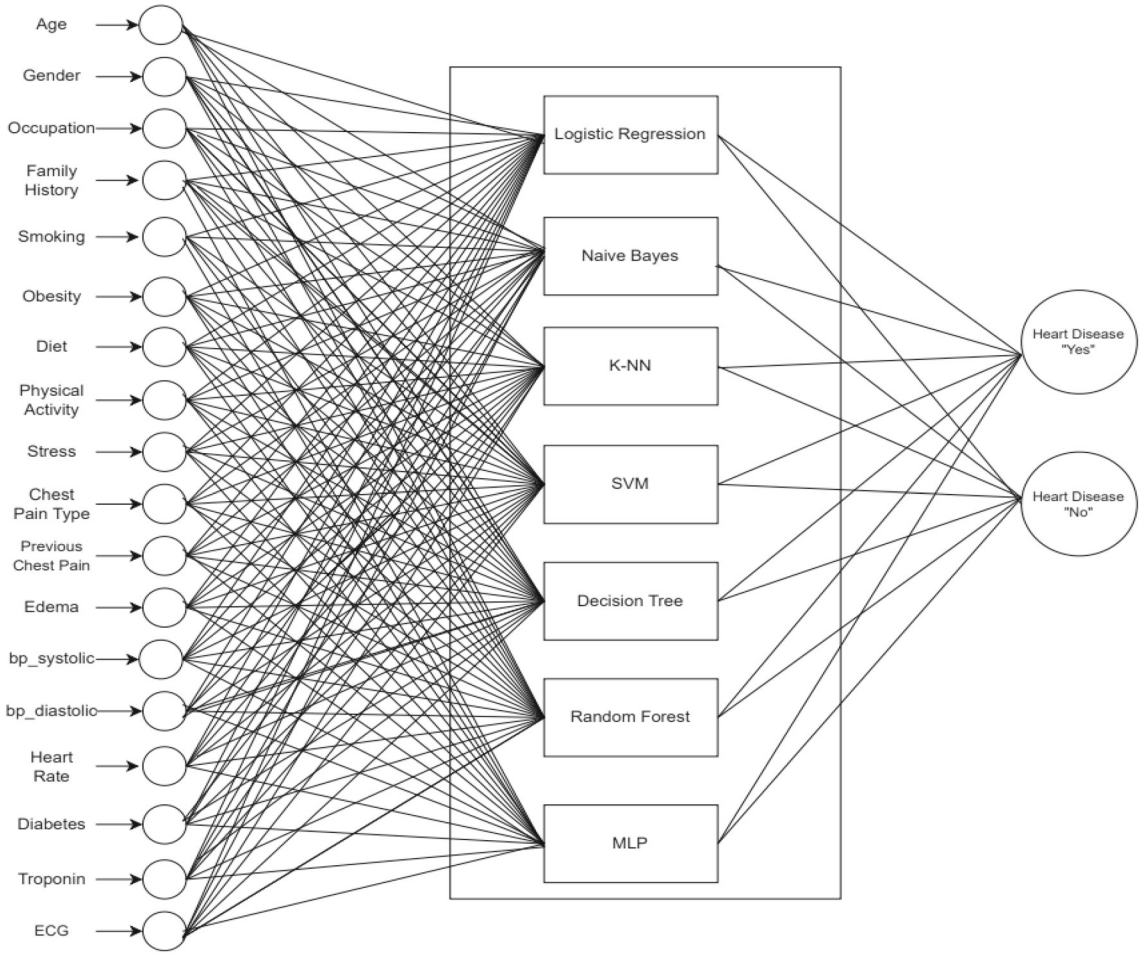


Fig. 2 Heart disease prediction model structure

a. Logistic regression Discrete values (i.e., whether a student passed or failed) are predicted by logistic regression after a transformation function. The transformation function used in logistic regression is known as the logistic function, and its formula is $h(x) = 1/(1 + e^x)$. An S-curve is created as a result. The results of logistic regression are probabilities for the default class (unlike Linear Regression, where the output is directly produced). The result falls between 0 and 1 since it is a probability. The logistic function $h(x) = 1/(1 + e^{-x})$ is used to log-transform the x-value to get this output (y-value). The probability is then forced into a binary categorization using a threshold.

b. Naïve Bayes It is a classification method that relies on the independence of predictors in the Bayes theorem. A Naïve Bayes classifier, to put it simply, believes that the existence of one feature in a class is unrelated to the presence of any other feature. Fruit might be categorized as an apple, for instance, if it is red, rounded, and has a diameter of roughly 3 inches. A Naïve Bayes classifier would consider all of these characteristics to individually contribute to the likelihood that this fruit is an apple, even if they depend on one another or the existence of additional attributes. The Naïve Bayesian model is simple to construct and is especially beneficial for massive data sets.

c. K-nearest neighbor It can be applied to situations involving classification and regression. It is, however, more frequently employed in classification issues in the sector. K nearest neighbors is a straightforward algorithm that sorts new instances according to the consensus of its k neighbors and stores all of the existing examples. The K nearest neighbors of the case allocated to the class, as determined by a distance function, share the most instances of it.

Euclidean, Manhattan, Minkowski, and Hamming distances are examples of these distance functions. The fourth function, Hamming, is used for categorical variables, whereas the first three are used for continuous processes. If K is equal to 1, the instance is then merely put into the class of its closest neighbor. It can be difficult at times to choose K when using k-NN modeling.

d. Support vector machine A relatively recent development in supervised machine learning is the support vector machine (SVM). The kernel Adatron technique is used to implement the support vector machine. By isolating those inputs near the data's borders, the kernel Adatron maps inputs to high-dimensional feature space and optimally divides the data into the appropriate classes. As a result, the kernel Adatron is particularly good at separating data sets with complex boundary relationships. SVM cannot be used to approximate functions; it can only be used for classification.

e. Decision tree One of the predictive modeling techniques used in statistics, data mining, and machine learning is decision tree learning. Classification trees are tree models where the target variable can take a discrete range of values. In these tree structures, the leaves correspond to class labels, and the branches to the attributes combine to form those class labels. Regression trees are decision trees when the target variable can take continuous values (usually real numbers).

f. Random forest In AI, ensemble methods combine several learning algorithms to provide more accurate predictions. A machine-learning ensemble usually is far more flexible in its structure than a statistical ensemble, which in statistical mechanics is typically limitless. It only consists of a particular finite collection of different models. The major challenge is finding basic models that make many errors rather than exceedingly accurate ones. When ensembles are employed for classification, for instance, even if the basis classifier accuracy is low, high accuracy can still be achieved if several base models incorrectly categorize different training samples.

g. Multilayer perceptron The multilayer perceptron structure, which has three layers—input, hidden, and output—is used to detect heart disease. As previously stated, the input layer comprises eighteen distinct characteristics. The output's two parameters are the forecast of the prevalence of heart disease (Yes) and the absence of heart disease (No). Since there are 18 attributes, the MLP neural network has 18 neurons in the input layer, seven neurons in the hidden layer, and two neurons in the output layer, as illustrated in Fig. 3. The information and the corresponding signal in MLP NN advance through the hidden layer from the input layer to the output layer.

3.1.4 Training and testing

Typically, the dataset is divided into training and testing data, as illustrated in Table 2. The training set includes a known output, and the model develops on this data to later generalize it to other data. In this study, training is done with 67% of the data. Therefore, 39 cases from the 59-instance dataset are chosen as the training set. In data training, the training accuracy is frequently high, meaning the model performs at a high level of accuracy in the training set but performs poorly when evaluated against the test set. Therefore, tenfold cross-validation is employed to prevent performance errors. The test set is part of the data where the model is examined; it is frequently the data's dependent variable. Twenty instances from the dataset are used to test the dataset in this study, which utilized 33% of the total data. Depending on the model used, cross-validated data will either perform better or worse when tested. Therefore, a method known as parameter tuning is utilized to ensure each model is operating at its best.

Table 2 Proportions of data in each dataset

Data set	Proportion (%)
Training set	67
Test set	33

Table 3 Confusion matrix

	Predicted	
	Negative (Class 0)	Positive (Class 1)
Actual		
Negative (Class 0)	TN	FP
Positive (Class 1)	FN	TP

3.2 Performance analysis

This study has used model construction to evaluate the efficacy and usefulness of several classification algorithms for heart disease prediction. A confusion matrix and all relevant metrics, including the True Positives, True Negatives, False Positives, False Negatives, accuracy, precision, recall, F1-score, and ROC-AUC score, are used to evaluate a model's performance.

Confusion matrix: one of the most straightforward methods for assessing the model's efficacy and accuracy is the confusion matrix. A table containing the two dimensions, "Actual class" and "Predicted class," in each dimension is the confusion matrix. Rows are the actual classifications of heart disease, while columns are the predicted ones. Two classes, Class 0 and Class 1, are present in the dataset. Table 3 is a confusion matrix that has been made for heart disease detection.

True Positives (TP): true positives are the cases when the actual class of the data point is True, and the predicted is also True.

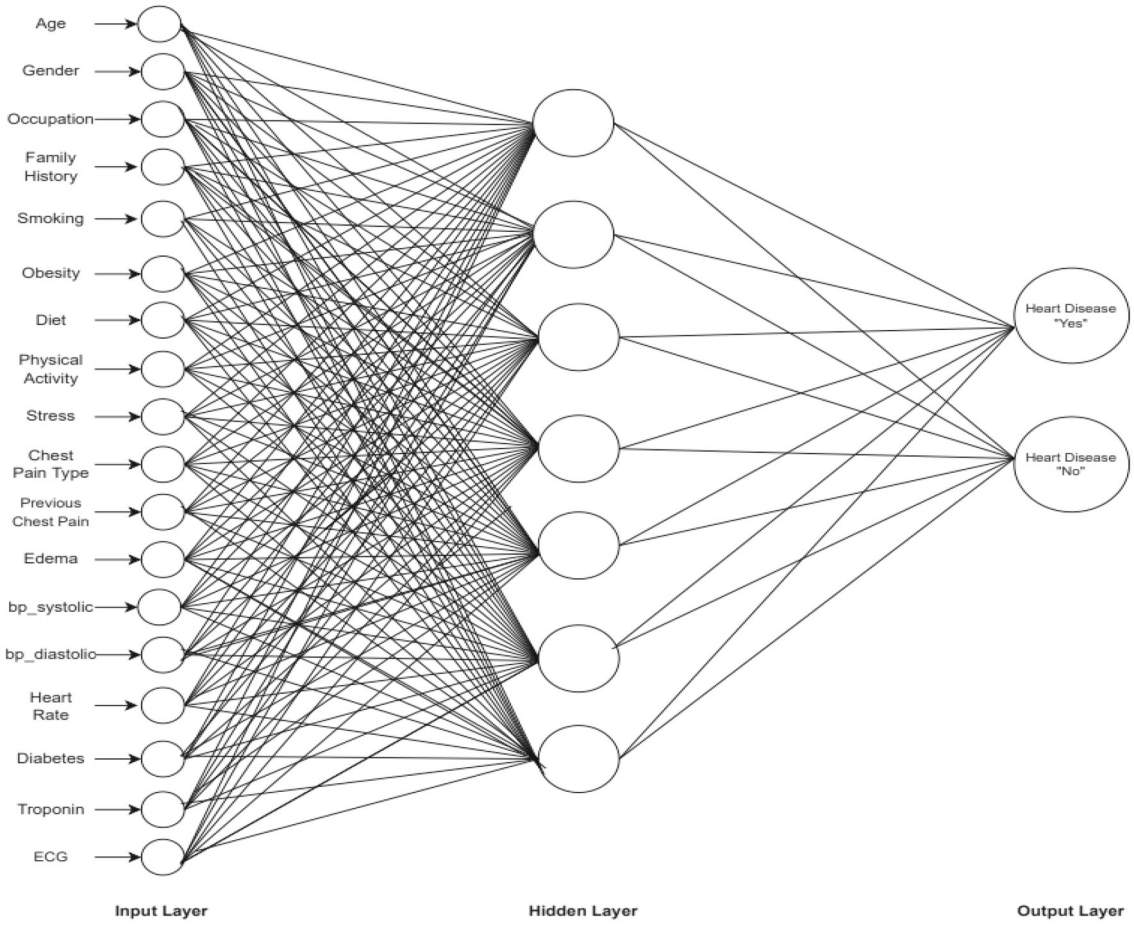


Fig. 3 Multilayer perceptron architecture for heart disease prediction

True Negatives (TN): true negatives are the cases when the actual class of the data point is False, and the predicted is also false.

False Positives (FP): false positives are the cases when the actual class of the data point is False, and the predicted is True.

False Negatives (FN): false negatives are the cases when the actual class of the data point is True, and the predicted is False.

Accuracy: accuracy is calculated as the number of all correct predictions of heart disease divided by the total number of the dataset. Accuracy comparison is based on the performance among the four classification algorithms.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2}$$

Precision: it tells what fraction of predictions of a positive class are actually heart diseases positive. The high precision means the result of the measurements is consistent or the repeated values of the reading are obtained. The low precision means the value of the measurement varies.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Recall: recall refers to a test's ability to designate an individual with heart disease as positive. A highly sensitive test means that there are few false negative results, and thus fewer cases of heart disease are missed. It is also known as the True Positive rate (TPR).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Specificity: it is calculated as the number of correct negative predictions of heart disease divided by the total number of negatives. It is also called True Negative Rate (TNR).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

F1-score: the harmonic mean of precision and recall is called F1-score. The high F1-score indicates perfect precision and recall of the proposed model.

$$\text{F1} - \text{score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (6)$$

AUC-ROC: an efficiency indicator for classification issues is called AUC (Area under Curve)-ROC (Receiver Operating Characteristic). We can learn about the model's capacity for class distinction using the AUC-ROC metric. The model is better when the AUC is higher. It can be produced mathematically by plotting TPR (True Positive Rate) vs. FPR (False Positive Rate) at various threshold values.

4 Implementation and analysis

This study employs various machine learning-based techniques to forecast cardiac disease. These techniques' effectiveness has been assessed, and a comparison between them has also been completed. The following section analyzes the data and presents the findings before moving on to the section on performance evaluation for various techniques.

4.1 Exploratory data analysis

Exploratory data analysis (EDA) is a method for data analysis that uses a range of approaches (mostly graphical) from statistics to linear algebra by employing essential charting tools to comprehend the dataset before moving on to actual machine learning. In a nutshell, EDA is "a first glance at the data," and it is a crucial phase in assessing experiment results. It is used to comprehend and condense the dataset's content to improve features and produce reliable, appropriately interpreted results.

4.1.1 Data visualization

Figure 4 represents the frequency distribution of all features. Age, bp systolic, and bp diastolic appear normally distributed based on the visualization.

4.1.2 Descriptive statistics of dataset

Table 4 contains 59 values for age, 46 for blood pressure (systolic and diastolic), 46 for heart rate, and 59 for the target. The bp_systolic, bp_diastolic, and heart rate attributes for the 59 patients have some missing values. Therefore, this study has used several strategies to fill in the missing variables before creating the model. This research has also observed that the mean age is 57, so most of the patients are old. The minimum and maximum age ranges were 26 and 90 years, respectively. The left-skewed age distribution of the data indicates the absence of populations with low ages. The age range of 57 to 73 years has a standard deviation of 15.87, illustrating the sparseness of the data; this age group visits the doctor the most frequently. The bp_systolic mean is also 128, indicating that most patients have normal bp_systolic values. The mean bp_diastolic value is 77, the minimum value is 50 Hgmm, and the maximum is 120 Hgmm. The minimum bp_systolic value is 75 Hgmm, and the maximum bp_systolic value is 180 Hgmm. Since there are 47 heart rates, it is clear that some data are missing, and those missing values are filled in using the median. The average heart rate is 83 bpm, and the highest recorded heart rate is 165 bpm (bit per minute).

4.1.3 Correlation between features

The correlation of all attributes with the target value is depicted in Fig. 5. The graphic shows there are varying correlations between the target and various attributes. The correlation between the numerical data is seen in Table 5. Figure 6 illustrates the heatmap of the correlation between numerical values. The two variables, bp_systolic and bp_diastolic, are highly correlated. Therefore, we ought to depreciate either bp_systolic or bp_diastolic. But bp_diastolic correlates higher with the target, according to the connection with the target and all other parameters. Consequently, bp_systolic has been removed from the dataset.

4.1.4 Relation of features with heart disease

The collected data are analyzed, and the distribution of features corresponding to heart disease is shown in Fig. 7. Data are arranged according to category—male and female, smoking and non-smoking, obesity and non-obesity, normal diet and abnormal diet, physical activity and non-physical activity, normal stress and abnormal stress, history of chest pain and no history of chest pain, positive troponin and negative troponin, diabetics and non-diabetics and normal ECG and abnormal ECG as shown in Fig. 8 and Table 6.

Age significantly contributes to the decline in circulatory efficiency, which raises the risk of cardiovascular disease (CVD) in older persons. Age-related increases in the prevalence of CVD, such as atherosclerosis, stroke, and myocardial infarction, have been observed in both men and women. According to the American Heart Association (AHA), the incidence of CVD in US men and women is 40% between the ages of 40 and 59, 75% between the ages of 60 and 79, and 86% in individuals over the age of 80 [75]. Figure 7(a) depicted that people older than 50 are particularly susceptible to developing heart disease. So, age is considered to be a key indicator of heart disease.

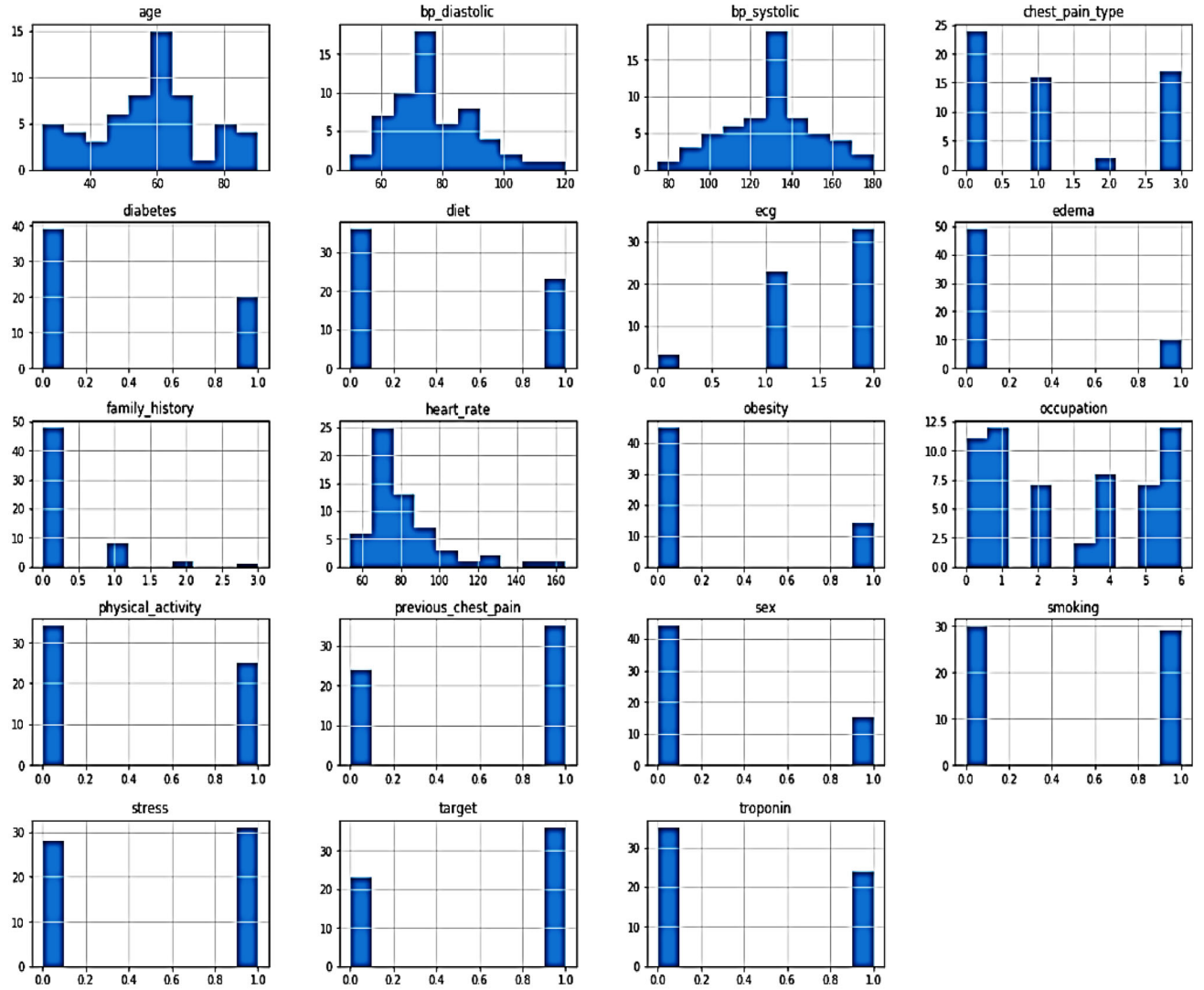


Fig. 4 Frequency distribution of features for all attributes

Table 4 Descriptive statistics of the numeric value

	Age	bp_systolic	bp_diastolic	Heart rate	Target
Count	59.0	46.00	46.00	47.00	59.00
Mean	15.76	128.26	77.86	83.95	0.61
Std	15.87	24.06	15.44	22.70	0.49
Min	26.00	75.00	50.00	54.00	0.00
25%	49.00	110.00	66.25	70.50	0.00
50%	60.00	130.00	75.00	76.00	1.00
75%	65.00	143.75	90.00	90.00	1.00
Max	90.00	180.00	120.00	165.00	1.00

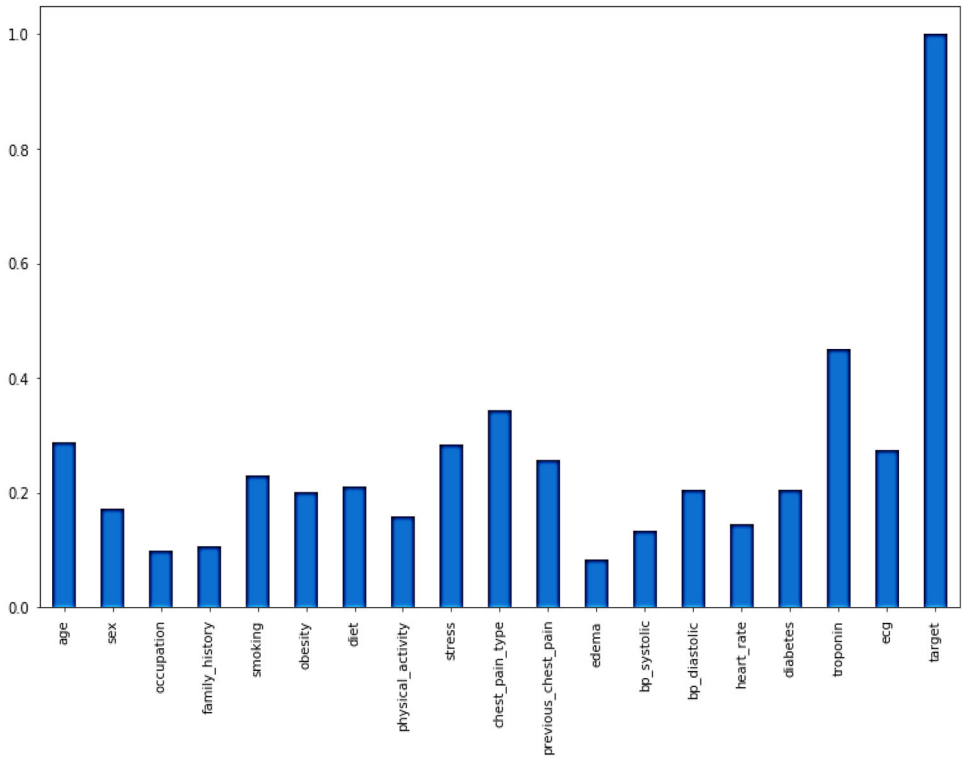


Fig. 5 Correlation of all attributes with target value

Table 5 Correlation among numeric value

	Age	bp_systolic	bp_diastolic	Heart rate	Target
Age	1.000000	0.0199994	0.069676	– 0.030911	0.288249
bp_systolic	0.019994	1.000000	0.805017	0.150733	0.132768
bp_diastolic	0.069676	0.805017	1.000000	0.208148	0.205186
Heart rate	- 0.030911	0.150733	0.208148	1.000000	0.145232
Target	0.288249	0.132768	0.205186	0.145232	1.000000

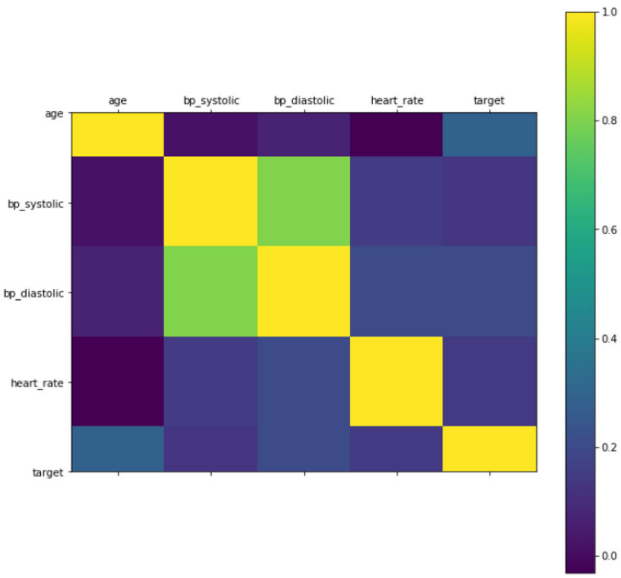


Fig. 6 Correlation numeric value visualization using the heat map

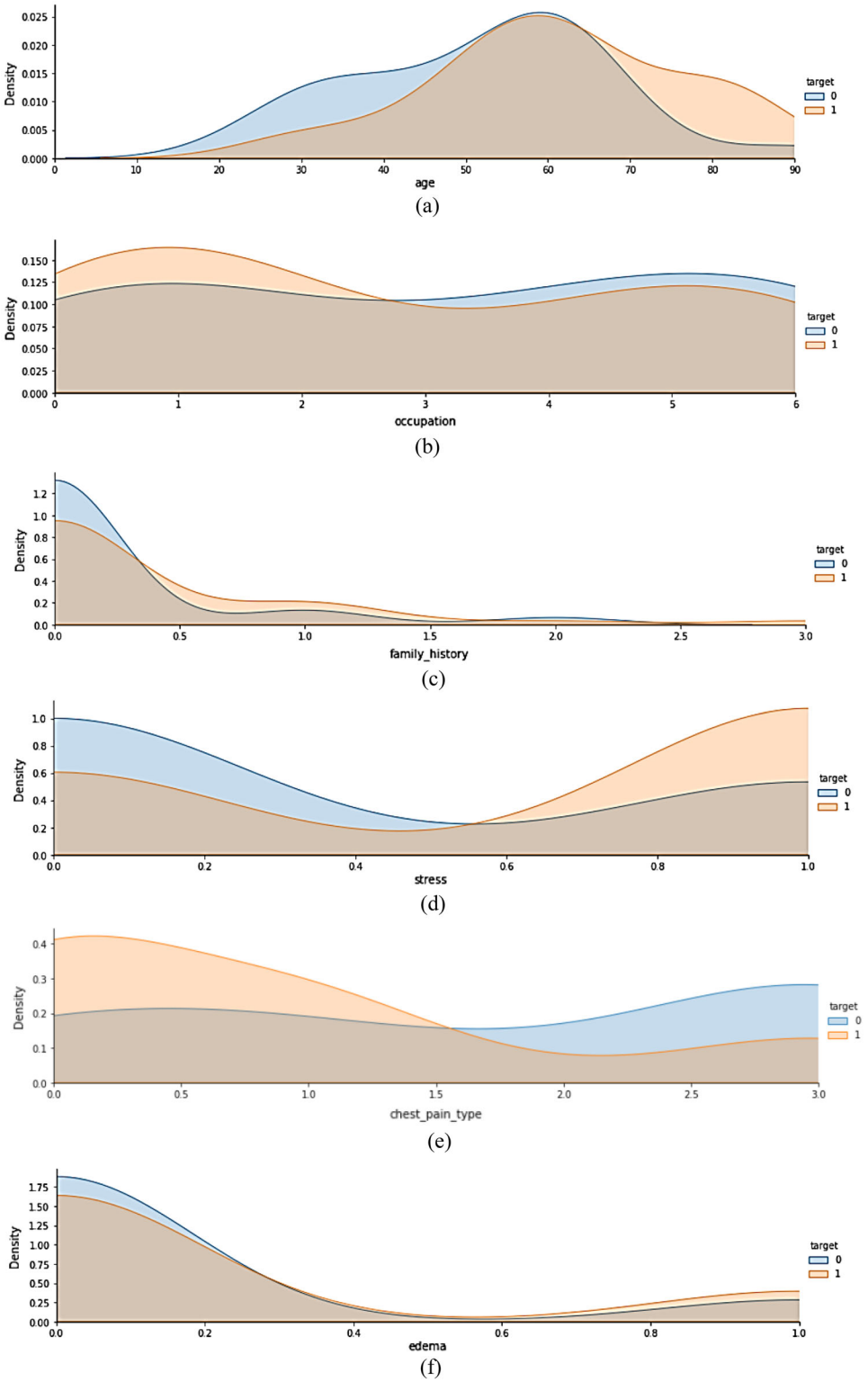
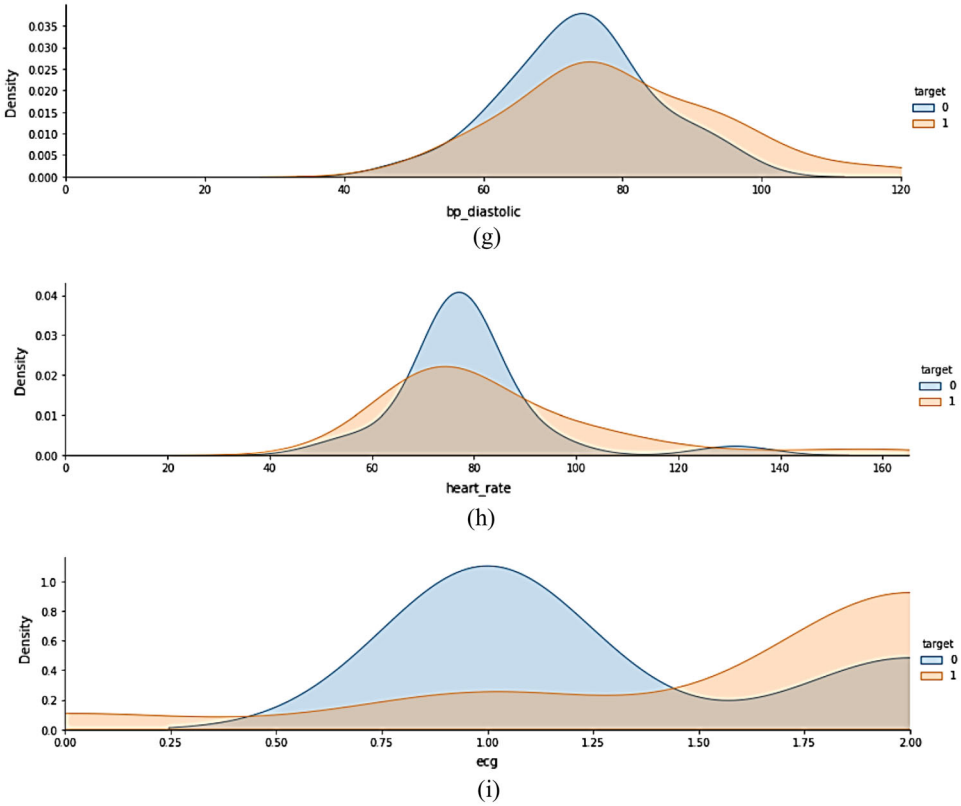


Fig. 7 Distribution of features corresponding to heart disease; **a** Age, **b** Occupation, **c** Family history, **d** Stress, **e** Previous chest pain, **f** Edema, **g** bp_diastolic, **h** Heart rate, **i** ECG

Fig. 7 continued



Regarding the onset and prevalence of CVD, there are regularly documented gender differences among older persons. Figure 8(a) discloses that the affected rates for men and women are 78.38% and 21.62%, respectively. It is evident that the average number of males with heart disease is more than that of females.

Each patient in our dataset is assigned to one of several possible occupations. Figure 7(b) reveals that the connection between occupation and heart disease is not very strong. Family history also has little impact on occurrences of heart disease, according to Fig. 7(c).

Compared to non-smokers, smokers are more likely to get a heart attack. It causes artery lining damage, and the accumulation of a fatty substance termed atheroma, which narrows the arteries and results in heart attacks. The diagnosis rates for smokers and non-smokers are 59.46% and 40.54%, respectively, as shown in Fig. 8(b). Smokers have a higher prevalence of heart disease than non-smokers.

Obesity directly influences the occurrence of cardiovascular risk factors such as dyslipidemia, type 2 diabetes, hypertension, and sleep problems. Independently from other cardiovascular risk factors, obesity increases the risk of developing cardiovascular disease and its mortality. Figure 8(c) shows that obesity has a 67.57% impact on the incidence of heart disease.

The total amount of food consumed by an individual constitutes their diet. One of the main risk factors for many chronic diseases, such as cancer, diabetes, cardiovascular disease, and other disorders associated with obesity, is a poor diet. Additionally, it is advisable to avoid trans-fatty acids and favor unsaturated fats over saturated fats. Figure 4(d) shows that the abnormal diet has a 49.95% impact on heart disease incidence.

Physical activity is any skeletal muscle-driven action of the body that requires net energy expenditure. Physical inactivity (lack of physical exercise) is the fourth most crucial risk factor for mortality worldwide (6% of deaths) [76]. Lack of physical exercise has a 67.57% impact on the occurrence of heart disease, according to Fig. 8(e). Stress is the body's response to dangerous circumstances. Stress has a 78.38% impact on the incidence of heart disease, according to Fig. 8(f).

Angina is a type of chest pain or discomfort brought on by a lack of oxygen-rich blood to the heart muscle. The chest area may experience pressure or squeeze. It typically occurs due to one or more blocked or constricted coronary arteries. The distribution of chest pain corresponding to heart disease is depicted in Fig. 7(d). The history of chest pain is an essential factor in occurring heart disease. People with previous chest pain (70.27%) have been diagnosed with heart disease, according to the analysis of Fig. 8(g).

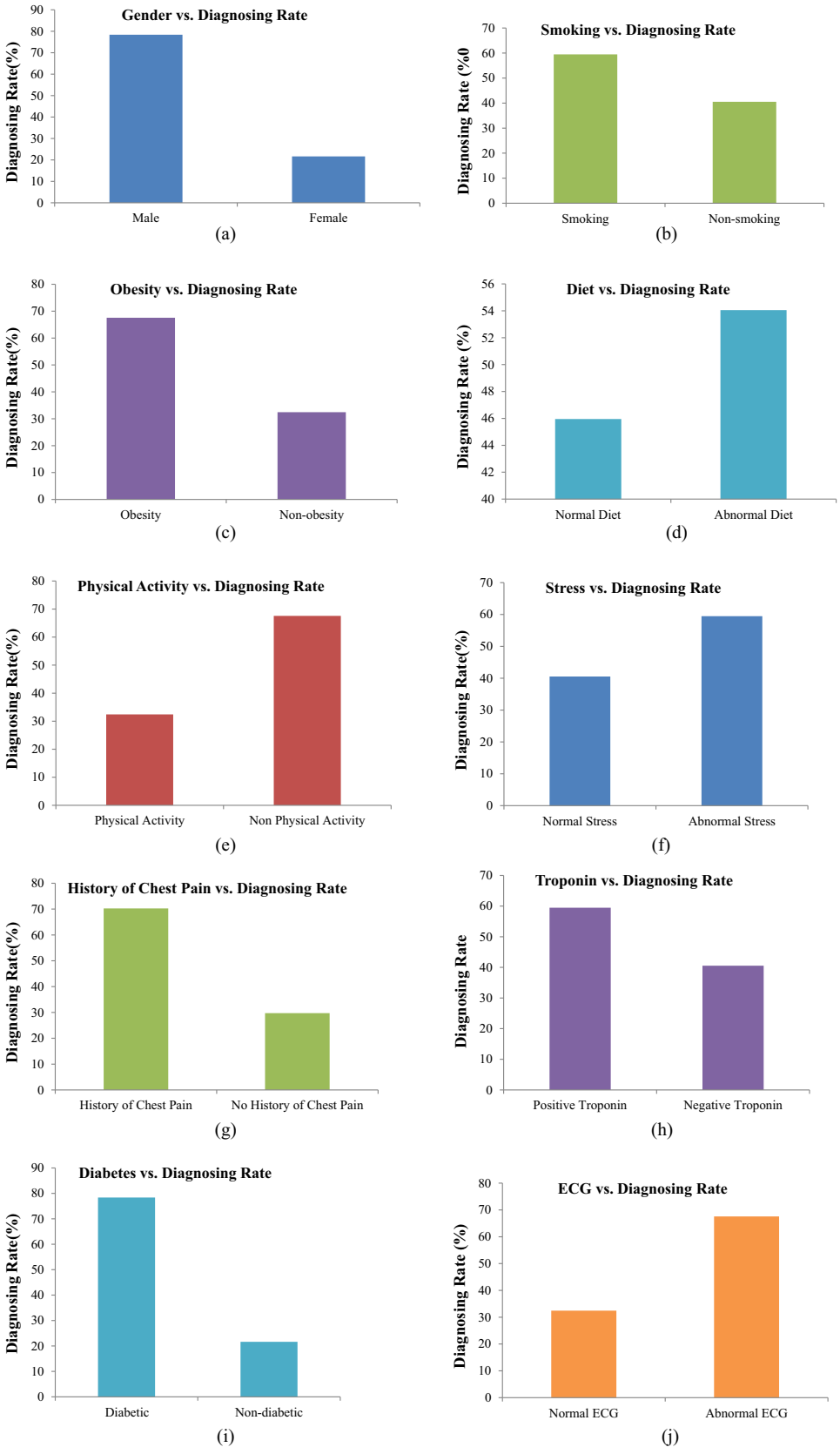


Fig. 8 Relations of features to heart disease; **a** Gender, **b** Smoking, **c** Obesity, **d** Diet, (**e**) Physical activity, **f** Stress, **g** History of chest pain, **h** Troponin, **i** Diabetes, **j** ECG

Table 6 Analysis of the dataset

Category	No. of the diagnosed person	Diagnosing rate
Gender		
Male	29	78.38%
Female	8	21.62%
Smoking		
Yes	22	59.46%
No	15	40.54%
Obesity		
Yes	25	67.57%
No	12	32.43%
Diet		
Normal	17	45.95%
Abnormal	20	54.05%
Physical activity		
Yes	12	32.43%
No	25	67.57%
Stress		
Normal	15	40.54%
Abnormal	22	59.46%
Previous chest pain		
Yes	26	70.27%
No	11	29.73%
Troponin		
Positive	22	59.46%
Negative	15	40.54%
Diabetes		
Yes	29	78.38%
No	8	21.62%
ECG		
Normal	12	32.43%
Abnormal	25	67.57%
Total		
Total no. of sample	59	
Total no. of the diagnosed person	37	

Edema, also known as swelling, happens when an excessive amount of fluid is trapped in the body's tissues, particularly the skin. Medication-induced edema, pregnancy, or an underlying illness—often congestive heart failure, kidney disease, or liver cirrhosis—can all cause it. When the heart becomes weaker and pumps blood less efficiently, fluid can accumulate, leading to edema. The distribution of edema corresponding to heart disease is depicted in Fig. 3(f). Figure 7(f) shows that there is weak correlation between edema and cardiac disease.

Diastolic blood pressure measures the force blood applies to arterial walls when the heart is at rest between beats. The distribution of people's diastolic blood pressure with heart disease is shown in Fig. 7(g). People with normal blood pressure are more likely to be healthy and free of heart disease. There is a chance that a patient has heart disease if there has been any aberrant blood pressure behavior, such as an increase or a decrease in blood pressure. Therefore, we view bp_diastolic as a crucial indicator for predicting heart disease.

The acceleration of heart rate is linked to an increase in cardiovascular risk. It has been shown that the chance of cardiac death increased by at least 20% for every additional ten beats per minute in heart rate [77]. Figure 7(h) illustrates the distribution of People's Heart Rate Corresponding to the target.

Troponins are proteins that are present in the skeletal and cardiac muscles. When the heart is harmed, troponin is released into the blood. To identify a heart attack, doctors check troponin levels in the blood. Additionally, this test can aid medical professionals in quickly determining the best course of action. Figure 8(h) shows that troponins have a 59.46% influence on the prevalence of the cardiac disease.

Major unfavorable cardiac events like stroke, heart attack, and mortality are associated with an increased risk of diabetes. The diagnosis rates for people with diabetics and non-diabetics are 78.38% and 21.62%, respectively, as shown in Fig. 8(i). Heart disease is more common in diabetic patients than in non-diabetics.

An electrocardiogram (ECG) is a test that uses sensors positioned on the skin across the chest to find electrical activity in the heart. It is an easy, rapid test that may be completed in a doctor's office. These tests can identify cardiac issues, including coronary heart disease, mainly when carried out while exercising (such as walking on a treadmill). The abnormal type of ECG has 67.57% impact on the occurrence of heart disease, according to Fig. 8(j).

4.1.5 Feature selection

In this experiment, the correlation-based feature subset selection method and best first search (BFS) extract 14 relevant features from 19 attributes of heart patient records. The 14 features are age, gender, smoking, obesity, diet, physical_activity, stress, chest_pain_type, previous_chest_pain, bp_diastolic, diabetes, troponin, ECG, and target.

4.2 Heart disease prediction

This study aims to evaluate which heart disease classification algorithm performs the best. This section summarizes all the study's findings and introduces the best performer based on various performance indicators. The best classifier

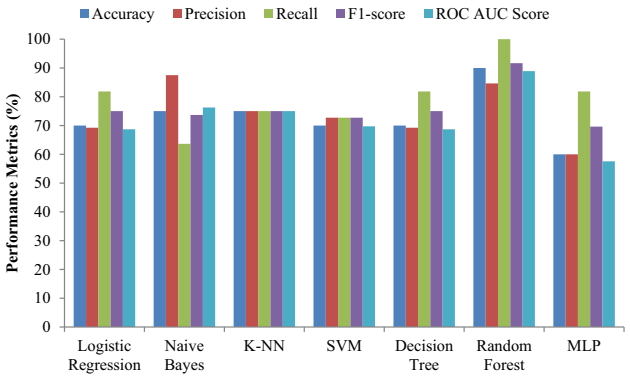


Fig. 9 Performance metrics of different techniques for all attributes

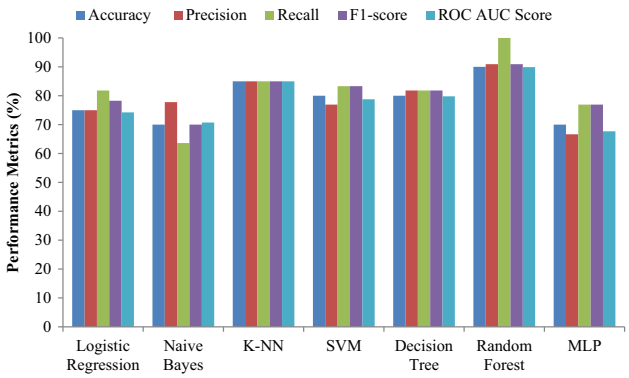


Fig. 10 Performance metrics of different algorithms for selected features

is identified using a dataset with all 19 variables once the performance is obtained. Second, the best classifier for heart disease is found merely by considering 14 parameters that strongly correlate with the target value. Finally, this study has contrasted various performance indicators of multiple ML algorithms for both datasets.

4.2.1 Benchmarking using all features

The dataset contains a total of 19 features for heart disease prediction. Different evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC value have been used to assess the effectiveness of all distinct AI techniques in all features dataset.

In Table 7 and Fig. 9, Random Forest outperforms other models in terms of accuracy, precision, recall, F1-score, and ROC-AUC score. The most outstanding 89.9% accuracy is attained by random forest, with 84% precision, 100% recall, 91% F1-score, and 88% ROC-AUC score. Therefore, it has been concluded that random forest is superior to the other six classification algorithms for the dataset of 19 attributes. K-NN has achieved 75% accuracy, 77.5% precision, 75% recall, 75% F1-score, and 75% ROC-AUC score. Naïve Bayes has 75% accuracy with 87.50% precision, 63.64% recall, 73.68% F1-score, and 76.26% ROC-AUC score. K-NN and Naïve Bayes both perform well, albeit Naïve Bayes has higher precision than K-NN. Naïve Bayes has a lower recall value and F1-score than K-Nearest Neighbor, yet the ROC-AUC Score is more significant than K-NN. Logistic Regression has a 70% accuracy rate. SVM offers 72% F1-score and 70% accuracy. Decision tree provides 75% F1-score and 70% accuracy. Decision tree and logistic regression hence take the same stances regarding race. But when compared to other methods, MLP does not provide good performance. Therefore, it can be said that MLP is not as effective as other classification algorithms, and Random Forest is the best for our entire dataset. The graphical representation of the performance analysis of AI techniques using all features is depicted in Fig. 9.

4.2.2 Benchmarking using selected features

This study has tested performance measures for all AI techniques after reducing the set of attributes to 14 features. Table 8 and Fig. 10 depict the evaluation of algorithms in the test dataset with selected features. Logistic regression shows 75% accuracy, 75% precision, 81.82% recall, 78.26% F1-score, and 74.24% ROC-AUC score. Naïve Bayes achieves 70% accuracy, 77.78% precision, 63.64% recall, 70% F1-score, and 70.71% ROC-AUC score. K-NN performs admirably than logistic regression. MLP shows the lowest 63.64% recall and 70% F1-score than all other algorithms. The value of accuracy, precision, recall, F1-score, and ROC-AUC score of K-NN is 85% for all metrics. Random forest achieves the highest 90% accuracy, 90.91% precision, 100% recall, 90.91% F1 score, and 89.90% ROC-AUC score. SVM and decision tree both have achieved 80% accuracy. SVM and Decision Tree operate nearly equally well in this situation, although decision tree outperforms SVM marginally. With a 76.92% F1-score, logistic regression has 66.67% precision, 76.92% recall, 67.68% ROC-AUC score, and 70% accuracy rate. Compared to other techniques, MLP achieves the lowest precision (66.67%) and F1-score (67.68%). Therefore, we may conclude that Random Forest is the best classification technique for our selected dataset, and MLP is not as successful as other classification algorithms. The graphical representation of the performance analysis of all techniques for selected features is depicted in Fig. 10.

4.2.3 Comparative analysis

When comparing the performance metrics from the results of the entire dataset and the selected dataset, the selected dataset table reveals significant variations. Logistic Regression, K-NN, SVM, Decision Tree, and MLP have improved precision for the selected dataset. However, precision declines for Naïve Bayes just as accuracy does. The value of recall has improved for K-NN, SVM, Decision Tree, and MLP. The

Table 7 Evaluation of algorithms in test dataset with all features

	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC Score (%)
Logistic Regression	70.0	69.23	81.82	75.00	68.69
Naïve Bayes	75.0	87.50	63.64	73.68	76.26
K-NN	75.0	75.00	75.00	75.00	75.00
SVM	70.0	72.73	72.73	72.73	69.70
Decision Tree	70.0	69.23	81.82	75.00	68.69
Random Forest	89.9	84.62	100.00	91.67	88.89
Multilayer Perceptron	60.0	60.00	81.82	69.63	57.58

Table 8 Evaluation of algorithms in test dataset with selected features

	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC AUC Score (%)
Logistic regression	75.0	75.00	81.82	78.26	74.24
Naïve Bayes	70.0	77.78	63.64	70.00	70.71
K-NN	85.0	85.00	85.00	85.00	85.00
SVM	80.0	76.92	83.33	83.33	78.79
Decision tree	80.0	81.82	81.82	81.82	79.80
Random forest	90.0	90.91	100	90.91	89.90
Multilayer perceptron	70.0	66.67	76.92	76.92	67.68

recall for any algorithms does not fall this time, though. F1-scores have increased for MLP, K-NN, SVM, Decision Tree, and Logistic Regression. The F1-score of Naïve Bayes has fallen. For all algorithms except Naïve Bayes, the ROC-AUC score has increased. Therefore, it can be concluded from the above data that Random Forest performs best for both datasets. Naïve Bayes reduces its performance in all sectors than the dataset of all features. It is possible to improve Naïve Bayes' performance by expanding the dataset. Additionally, MLP improves performance over the all attributes dataset. It is evident from Fig. 11 that the selected attributes dataset shows better performance than the dataset of all features.

5 Conclusion

This research has provided a comprehensive study of patient characteristics for heart disease prediction. Correlation-based Feature Subset Selection method with the Best First Search has been carried out to select the most significant features. It has been discovered that all of the features are not strongly connected and that a combination of just 14 features (age, gender, smoking, obesity, diet, physical activity, stress, chest pain type, previous chest pain, blood pressure diastolic, diabetes, troponin, ECG, and target) significantly contribute to the prediction of heart disease. Finally, the datasets containing all features and selected features are used to develop seven AI (logistic regression, Naïve Bayes, K-NN, SVM, decision tree, random forest, and MLP) methods. The accuracy rate of Random Forest utilizing selected attributes is 90%, coupled with 90.91% precision, 100% recall, 90.91% F1-score, and 89.90% ROC-AUC score, which is the highest performance rate when compared to other AI techniques. The dataset of selected features outperforms the dataset of all features, excluding Naïve Bayes. The lack of extra discriminatory feature sets and additional datasets has drastically decreased the performance of the Naïve Bayes model. It has been noticed that most of the dataset's features are strongly associated with one another. The clinicians will be helped in proficiently archiving the records by systematically studying the efficiency of the various features. The data management team can archive only the features crucial for predicting heart disease, as opposed to recording and preserving all the features. As part of our next effort, we want to validate our suggested methodology externally.

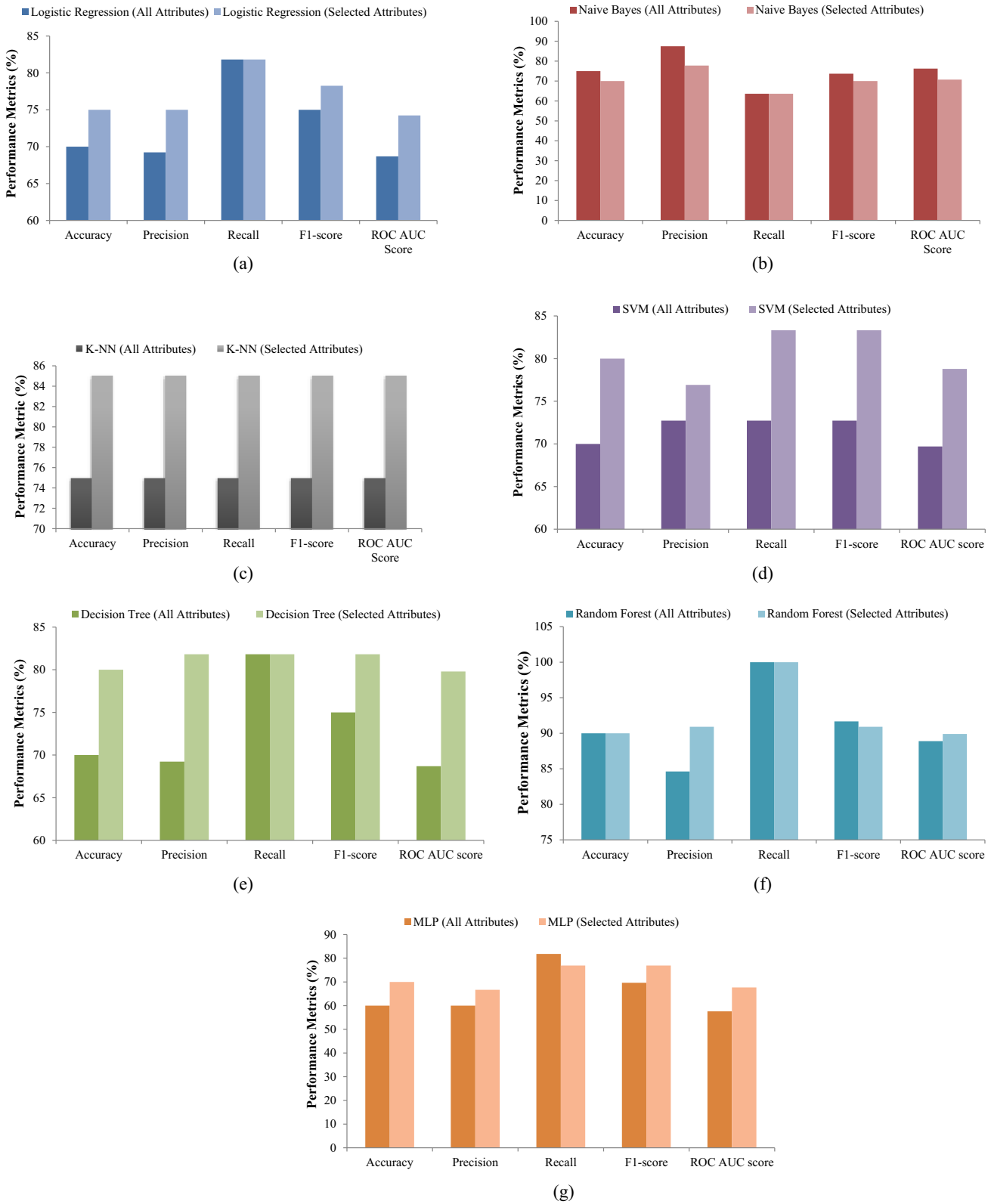


Fig. 11 Comparative analysis of both datasets for all classifiers; **a** Logistic regression, **b** Naïve Bayes, **c** K-NN, **d** Decision tree, **e** SVM, **f** Random forest, **g** Multilayer perceptron

Author contributions This work was carried out in collaboration between all authors. MIH and MHM designed the study and wrote the first draft of the manuscript. MARK and FSP managed the analyses of the study. SF, MSE, and MASK managed the literature searches. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no conflict of interest to disclose.

Ethical approval This article does not contain any studies with human participants and animals performed by any of the authors.

References

1. Mijwil MM., Al-Mistarehi AH., Aggarwal K: The effectiveness of utilising modern artificial intelligence techniques and initiatives to combat COVID-19 in South Korea: a narrative review. Asian J. Appl. Sci. **9**(5) (2021). (**ISSN: 2321-0893**)
2. Madjid, M., Safavi-Naeini, P., Solomon, S.D., Vardeny, O.: Potential effects of coronaviruses on the cardiovascular system: a review. JAMA Cardiol. **5**(7), 831–840 (2020)
3. Soni, J., Ansari, U., Sharma, D., Soni, S.: Predictive data mining for medical diagnosis: an overview of heart disease prediction . Int. J. Comput. Appl. **17**(8), 43–48 (2011)
4. Dai, H., Much, A.A., Maor, E., Asher, E., Younis, A., Xu, Y., Lu, Y., Liu, X., Shu, J., Bragazzi, N.L.: Global, regional, and national burden of ischaemic heart disease and its attributable risk factors, 1990–2017: results from the global burden of disease study 2017. Eur. Heart J.Qual. Care Clin. Outcomes **8**(1), 50–60 (2022)
5. Odden, M.C., Coxson, P.G., Moran, A., Lightwood, J.M., Goldman, L., BibbinsDomingo, K.: The impact of the aging population on coronary heart disease in the United States. Am. J. Med. **124**(9), 827–833 (2011)
6. Koh, H.C., Tan, G.: Data mining applications in healthcare. J. Healthcare Inform. Manag. **19**(2), 65. 35 (2011)
7. Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S.: Artificial intelligence and machine learning for medical imaging: a technology review. Physica Med. **83**, 242–256 (2021)
8. Taleb A., Lippert C., Klein T., Nabi M: Multimodal self-supervised learning for medical image analysis. In International conference on information processing in medical imaging. 661–673 (2021)
9. De Bruijne, M.: Machine learning approaches in medical image analysis: from detection to diagnosis. Med. Image Anal. **33**, 94–97 (2016)
10. Criminisi, A.: Machine learning for medical images analysis. Med. Image Anal. **33**, 91–93 (2016)
11. Le Glaz, A., Haralambous, Y., Kim-Dufor, D.H., Lenca, P., Billot, R., Ryan, T.C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., Lemey, C.: Machine learning and natural language processing in mental health: systematic review. J. Med. Internet Res. **23**(5), e15708 (2021)
12. Khanbhai, M., Anyadi, P., Symons, J., Flott, K., Darzi, A., Mayer, E.: Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform. **28**(1), e100262 (2021)
13. Manhas, J., Gupta, R.K.: Roy PP: A review on automated cancer detection in medical images using machine learning and deep learning based computational techniques: challenges and opportunities. Arch. Comput. Methods Eng. **29**, 2893–2933 (2021)
14. Allugunti, V.R.: Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. Int. J. Eng. Comput. Sci. **4**(1), 49–56 (2022)
15. Alanazi, S.A., Kamruzzaman, M.M., Islam Sarker, M.N., Alruwaili, M., Alhwaiti, Y., Alshammari, N., Siddiqi, M.H.: Boosting breast cancer detection using convolutional neural network. J. Healthcare Eng. (2021). https://doi.org/10.1155/2021/5528622
16. Abdullah, D.M., Ahmed, N.S.: A review of most recent lung cancer detection techniques using machine learning. Int. J. Sci. Bus. **5**(3), 159–173 (2021)
17. Bhise S., Gadekar S., Gaur AS., Bepari S., Deepmala Kale DSA.: Breast cancer detection using machine learning techniques. Int. J. Eng. Res. Technol. **10**(7) (2021). (**ISSN: 2278-0181**)
18. Nazir, S., Shahzad, S., Mahfooz, S., Nazir, M.: Fuzzy logic based decision support system for component security evaluation. Int. Arab J. Inf. Technol. **15**(2), 224–231 (2018)
19. Haq, A.U., Li, J.P., Memon, M.H., Nazir, S., Sun, R.: A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Inform. Syst.. **2018**, 1–21 (2018)
20. Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., Ghuli, P.: Heart disease prediction using machine learning. Int. J. Res. Technol. **9**(04), 659–662 (2020)
21. Jindal, H., Agrawal, S., Khera, R., Jain, R., Nagrath, P.: Heart disease prediction using machine learning algorithms. IOP Conf. Ser.: Mater. Sci. Eng.. **1022**(1), 012072 (2021)
22. Sahoo PK., Jeripothula P: Heart failure prediction using machine learning techniques. Available at SSRN 3759562. (2020)
23. Uyar, K., İlhan, A.: Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia Comput. Sci. **120**, 588–593 (2017)
24. Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., Yarifard, A.A.: Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm. Comput. Methods Programs Biomed. **141**, 19–26 (2017)
25. Sonawane, J.S., Patil, D.R.: Prediction of heart disease using multilayer perceptron neural network. In: International conference on information communication and embedded systems (ICICES2014), pp. 1–6. IEEE (2014)
26. Anooj, P.K.: Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. J. King Saud Univ. Comput. Inform. Sci. **24**(1), 27–40 (2012)
27. Olaniyi, E.O., Oyedotun, O.K., Adnan, K.: Heart diseases diagnosis using neural networks arbitration. Int. J. Intell. Syst. Appl. **7**(12), 72–79 (2015)
28. Bhatla, N., Jyoti, K.: An analysis of heart disease prediction using different data mining techniques. Int. J. Eng. **1**(8), 1–4 (2012)
29. Srivastava, N.: A logistic regression model for predicting the occurrence of intense geomagnetic storms. Ann. Geophys. **23**, 2969–2974 (2005). https://doi.org/10.5194/angeo-23-2969-2005
30. Jiang X., El-Kareh R., Ohno-Machado L: Improving predictions in imbalanced data using pairwise expanded logistic regression. AMIA Annual Symposium Proceedings: 625–634. (2011). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243279/
31. Reed, P., Wu, Y.: Logistic regression for risk factor modelling in stuttering research. J. Fluency Disord. **38**, 88–101 (2013). https://doi.org/10.1016/j.jfludis.2012.09.003
32. Khan, K.S., Chien, P.F., Dwarakanath, L.S.: Logistic regression models in obstetrics and gynecology literature. Obstet. Gynecol.

93, 1014–1020 (1999). https://doi.org/10.1097/00006250-199906000-00024
33. Kim, Y., Kwon, S., Song, S.H.: Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. Comput. Stat. Data Anal. **51**, 1643–1655 (2006). https://doi.org/10.1016/j.csda.2006.06.007
34. Howell, P., Davis, S.: Predicting persistence of and recovery from stuttering by the teenage years based on information gathered at age 8 years. J. Dev. Behav. Pediatr. **32**, 196–205 (2011). https://doi.org/10.1097/DBP.0b013e31820fd4a9
35. Jones, S.R., McEwen, M.K.: A conceptual model of multiple dimensions of identity. J. Coll. Stud. Dev. **41**, 405–414 (2000)
36. Vollmer, R.T.: Multivariate statistical analysis for pathologists: part I, the logistic model. Am. J. Clin. Pathol. **105**, 115–126 (1996). https://doi.org/10.1093/ajcp/105.1.115
37. Holland, A.L., Greenhouse, J.B., Fromm, D., Swindell, C.S.: Predictors of language restitution following stroke: a multivariate analysis. J. Speech Lang. Hear. Res. **32**, 232–238 (1989)
38. Fleck, M.P.D.A., Simon, G., Herrman, H., Bushnell, D., Martin, M., Patrick, D.: Major depression and its correlates in primary care settings in six countries: 9-month follow-up study. Br. J. Psychiatry **186**, 41–47 (2005). https://doi.org/10.1192/bjp.186.1.41
39. Mahdizah, M., Zamanzade, E.: Efficient body fat estimation using multistage pair ranked set sampling. Stat. Methods Med. Res. **28**, 223–234 (2019). https://doi.org/10.1177/0962280217720473
40. Langarizadeh, M., Moghbeli, F.: Applying naive bayesian networks to disease prediction: a systematic review. Acta Informatica Medica **24**(5), 364 (2016)
41. Chandel, K., Kunwar, V., Sabitha, S., Choudhury, T., Mukherjee, S.: A comparative study on thyroid disease detection using K-nearest neighbor and naive bayes classification techniques. CSI Trans. on ICT **4**, 313–319 (2016)
42. Reza, M., Hossain, G., Goyal, A., Tiwari, S., Tripathi, A., Bhan, A., Dash, P.: Automatic diabetes and liver disease diagnosis and prediction through SVM and KNN algorithms. In: Emerging technologies in data mining and information security, pp. 589–599. Springer, Singapore (2021)
43. Seo, H., Brand, L., Barco, L.S., Wang, H.: Scaling multi-instance support vector machine to breast cancer detection on the BreakHis dataset. Bioinformatics **38**, i92–i100 (2022)
44. Badr, E., Almotairi, S., Salam, M.A., Ahmed, H.: New sequential and parallel support vector machine with grey wolf optimizer for breast cancer diagnosis. Alex. Eng. J. **61**(3), 2520–2534 (2022)
45. Alyami, J., Sadad, T., Rehman, A., Almutairi, F., Saba, T., Bahaj, S.A., Alkhurim, A.: Cloud computing-based framework for breast tumor image classification using fusion of AlexNet and GLCM texture features with ensemble multi-kernel support vector machine (MK-SVM). Comput. Intell. Neurosci. (2022). https://doi.org/10.1155/2022/7403302
46. Mishra, R., Meher, S., Kustha, N., Pradhan, T.: A skin cancer image detection interface tool using vlf support vector machine classification. In: Computational intelligence in pattern recognition, pp. 49–63. Springer, Singapore (2022)
47. Sethy, P.K., Behera, S.K., Kannan, N.: Categorization of common pigmented skin lesions (CPSL) using multi-deep features and support vector machine. J. Digit. Imaging. **35**(5), 1207–1216 (2022)
48. Rustam, Z., Angie, N.: Prostate cancer classification using random forest and support vector machines. J. Phys.: Conf. Ser. **1752**(1), 012043 (2021)
49. Aszhari, F.R., Rustam, Z., Subroto, F., Semendawai, A.S.: Classification of thalassemia data using random forest algorithm. J. Phys.: Conf. Ser. **1490**(1), 012050 (2020)
50. Yekkala, I., Dixit, S.: Prediction of heart disease using random forest and rough set based feature selection. Int. J. Big Data Anal. Healthcare (IJBDAH) **3**(1), 1–12 (2018)
51. Routray, S., Ray, A.K., Mishra, C., Palai, G.: Efficient hybrid image denoising scheme based on SVM classification. Optik **157**, 503–511 (2018)
52. Barghout, L.: Spatial-taxon information granules as used in iterative fuzzy-decision-making for image segmentation. In: Granular computing and decision-making, pp. 285–318. Springer, Cham (2015)
53. DeCoste, D., Schölkopf, B.: Training invariant support vector machines. Mach. Learn. **46**(1), 161–190 (2002)
54. Le, N.Q.K., Yapp, E.K.Y., Ho, Q.T., Nagasundaram, N., Ou, Y.Y., Yeh, H.Y.: iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. Anal. Biochem. **571**, 53–61 (2019)
55. Do, D.T., Le, N.Q.K.: A sequence-based approach for identifying recombination spots in Saccharomyces cerevisiae by using hyperparameter optimization in fasttext and support vector machine. Chemom. Intell. Lab. Syst. **194**, 103855 (2019)
56. Afolayan, J.O., Adebiyi, M.O., Arowolo, M.O., Chakraborty, C., Adebiyi, A.A.: Breast Cancer Detection Using Particle Swarm Optimization and Decision Tree Machine Learning Technique. In: Intelligent Healthcare, pp. 61–83. Springer, Singapore (2022)
57. Nasser, F.K., Behadili SF.: Breast cancer detection using decision tree and K-nearest neighbour classifiers. Iraqi J. Sci. **63**(11), 4987–5003 (2022)
58. Sahoo, S., Subudhi, A., Dash, M., Sabut, S.: Automatic classification of cardiac arrhythmias based on hybrid features and decision tree algorithm. Int. J. Autom. Comput. **17**(4), 551–561 (2020)
59. Behadada, O., Chikh, M.A.: An interpretable classifier for detection of cardiac arrhythmias by using the fuzzy decision tree. Artif. Intell. Res **2**(3), 45–58 (2013)
60. Santos, L.I., Camargos, M.O., D'Angelo, M.F.S.V., Mendes, J.B., de Medeiros, E.E.C., Guimarães, A.L.S., Palhares, R.M.: Decision tree and artificial immune systems for stroke prediction in imbalanced data. Expert Syst. Appl. **191**, 116221 (2022)
61. Imura, T., Iwamoto, Y., Inagawa, T., Imada, N., Tanaka, R., Toda, H., Inoue, Y., Araki, H., Araki, O.: Decision tree algorithm identifies stroke patients likely discharge home after rehabilitation using functional and environmental predictors. J. Stroke Cerebrovasc. Dis. **30**(4), 105636 (2021)
62. Qiu, X., Miao, J., Lan, Y., Sun, W., Li, G., Pan, C., Wang, Y., Zhao, X., Zhu, Z., Zhu, S.: Artificial neural network and decision tree models of post-stroke depression at 3 months after stroke in patients with BMI≥ 24. J. Psychosom. Res. **150**, 110632 (2021)
63. Mishra, S., Mallick, P.K., Tripathy, H.K., Bhoi, A.K., González-Briones, A.: Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. Appl. Sci. **10**(22), 8137 (2020)
64. Chaudhuri, A.K., Sinha, D., Banerjee, D.K., Das, A.: A novel enhanced decision tree model for detecting chronic kidney disease. Netw. Model. Anal. Health Inform. Bioinform. **10**(1), 1–22 (2021)
65. Selwal, A., Raoof, I.: A multilayer perceptron based intelligent thyroid disease prediction system. Indones. J. Electr. Eng. Comput. Sci. **17**(1), 524–533 (2020)
66. Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., Nawaz, R.: An expert system for diabetes prediction using auto tuned multilayer perceptron. In: Intelligent systems conference (IntelliSys), pp. 722–728. IEEE (2017)
67. Lai, Z., Deng, H.: Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron. Comput. Intell. Neurosci. (2018). https://doi.org/10.1155/2018/2061516

68. Xing, W., Zhu, Z., Hou, D., Yue, Y., Dai, F., Li, Y., Tong, L., Song, Y., Ta, D.: CM-SegNet: a deep learning-based automatic segmentation approach for medical images by combining convolution and multilayer perceptron. Comput. Biol. Med. **147**, 105797 (2022)
69. Seo, H., Cho, D.H.: Cancer-related gene signature selection based on boosted regression for multilayer perceptron. IEEE Access **8**, 64992–65004 (2020)
70. Ram PK., Kuila P.: Dynamic scaling factor based differential evolution with multilayer perceptron for gene selection from pathway information of microarray data. Multimed. Tools Appl. 1–26 (2022)
71. Dehkordi, S.K., Sajedi, H.: Prediction of disease based on prescription using data mining methods. Heal. Technol. **9**, 37–44 (2019)
72. Jan, M., Awan, A.A., Khalid, M.S., Nisar, S.: Ensemble approach for developing a smart heart disease prediction system using classification algorithms. Res. Reports Clin. Cardiol.. **9**, 33–45 (2018)
73. Mansoor, H., Elgendy, I.Y., Segal, R., Bavry, A.A., Bian, J.: Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach. Heart Lung **46**(6), 405–411 (2017)
74. Austin, P.C., Tu, J.V., Ho, J.E., Levy, D., Lee, D.S.: Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J. Clin. Epidemiol. **66**(4), 398–407 (2013)
75. Rodgers, J.L., Jones, J., Bolleddu, S.I., Vanthenapalli, S., Rodgers, L.E., Shah, K., Karia, K., Panguluri, S.K.: Cardiovascular risks associated with gender and aging. J. Cardiovasc. Dev. Dis. **6**(2), 19 (2019)
76. World Health Organization T (2010) Global recommendations on physical activity for health. World Health Organization.
77. Perret-Guillaume, C., Joly, L., Benetos, A.: Heart rate as a risk factor for cardiovascular disease. Prog. Cardiovasc. Dis. **52**(1), 6–10 (2009)