Int. j. inf. tecnol. (October 2021) 13(5):2011–2023
https://doi.org/10.1007/s41870-021-00650-w

ORIGINAL RESEARCH

Cardiovascular risk assessment using data mining inferencing and feature engineering techniques

Aanchal Sahu[1] · Harshvardhan GM[1] · Mahendra Kumar Gourisaria[1] · Siddharth Swarup Rautaray[1] · Manjusha Pandey[1]

Received: 20 July 2020 / Accepted: 24 March 2021 / Published online: 20 April 2021

Abstract With the frequent decline in people's health due to the hectic lifestyle, increased levels of workload and intake of fast food, there has been an unfortunate growth in the number of patients suffering from cardiovascular diseases each year. Around the world, millions of people die each year due to cardiovascular diseases. While the statistics are eye-opening, with the vast amount of data about heart patients in our hands, we can save millions by detecting the risk at an early stage. With the recent advances in soft computing and fuzzy logic, various algorithmic approaches are employed to tackle the issue of cardiovascular risk assessment through machine learning. Using some of the algorithms of machine learning like Logistic Regression (LR), Naïve Bayes (NB), Support vector machine (SVM), and Decision tree (DT), Random Forest (RF) and K-Nearest Neighbours (KNN) classifiers, a model can be built to predict the risk accurately. In this paper, we have analysed each of the above methods normally and through feature engineering techniques like transformation through Principal Component Axes and considering different train-test folds to find the best performing model, which was found to be KNN in terms of all metrics and Logistic Regression in terms of accuracy.

Keywords Machine learning · Data mining · Heart disease prediction · Cardiovascular disease

✉ Mahendra Kumar Gourisaria
mkgourisaria2010@gmail.com

[1] School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha 751024, India

1 Introduction

The heart of a human body works incessantly throughout the whole lifespan. The four chambers of heart, two atria and two ventricles, each play a vital role in the process of pumping blood and maintaining the correct amount of dissolved oxygen in the blood. The division of the chambers ensures the oxygenated and deoxygenated blood do not mix up. The right atrium and ventricle of the heart gather and supply blood to the lungs and pulmonary arteries. The lungs refresh the deoxygenated blood into the oxygenated blood, eliminates the waste products like carbon dioxide, and sends it back to the left portion of the heart. The left portion comprises the left atrium and left ventricle, which supplies the blood with high O_2 concentration to all tissues throughout the body. The heart comprises four valves to direct the blood in the right way. Thus, the proper opening and closing of valves are extremely important for correct flow and no leakage of the blood. The heart follows a continuous loop of contraction and relaxation of the heart muscles. In the systolic process (contraction of cardiac muscles), the ventricles force out blood into the vessels in order to reach the lungs and rest of the body. In the diastolic process (relaxation of cardiac muscles) the ventricles get filled up with the blood coming from upper chambers. These processes keep taking place involuntarily to maintain the proper functioning of the body. Therefore, even the slightest of unease in the heart can cause severe problems in the body. The importance of a healthy heart is often underrated in the day-to-day world. There has been an increase in the number of young heart patients (under 35 years) over the years, which is a great cause of concern. Cardiovascular Disease (CVD) is a broad term that includes many heart-related problems, which include—Coronary (atherosclerotic) heart disease affecting

the arteries, Valvular heart disease regarding the dysfunction of heart valves, Cardiomyopathy regarding the heart muscle contraction, heart rhythm disturbances and heart infections. Since there are many kinds of heart problems, many factors are affecting them too. A disease can have numerous symptoms on the patient. Thus, it is important to look out for the most common ones [1]. The most common problem among cardiovascular diseases is atherosclerosis or coronary heart disease. Looking into the factors that affect the heart, many of them are correctable ones, which include, unhealthy diet, physical inactivity, obesity, etc. High-fat meals, in particular, can lead to an impaired vascular function [2]. It was found that people who were constantly overweight when they already fall into the risk group tend to develop an increased risk of Coronary Heart Disease (CHD) [3]. It is also analysed that the risk of CVD linearly reduces with the increase in physical activity to some extent [4]. This is why the risk prediction at an early stage can prove to be extremely helpful to the patients.

With the advancements in technology, computer science is constantly proving itself effective in various fields, medical science being one of them. Machine learning is being utilized for various medical science scenarios nowadays, may it be medical imaging, cancer prediction, diabetes prediction, or many more. Diseases can be detected, risk factors can be computed and analysed with the help of machine learning or generative modelling [5]. Computer-Aided Diagnosis (CAD) has become a vital part of medical science. Thus, we want to build the best performing model to predict the risk of CVD.

The proposed model will take up information on some most common features that may contribute to developing the condition of CVD. The features being, age, sex, chest pain intensity ranging from 0 to 3, and other heart-related attributes that we describe later in Sect. 3.1 to predict a binary variable which classifies whether the person is at risk or not.

The sections on this paper are as follows—Sect. 2 comprises Literature review, Sect. 3 contains the details of the proposed model and algorithms used in the model. Section 4 analyses and compares all the used techniques. Section 5 concludes the work and Sect. 6 talks about future work.

2 Literature review

With the tremendous amount of data in health-care domain, there is an ever-growing need for more accurate and better performing prediction models. Nayak et al. have contributed a model on prediction of cardiovascular disease using different data mining classifiers like DT (84.91%), SVM (88.68%), KNN (ROC) (58.49%), KNN (Acc.) (62.26%) and NB (96.23%) [6].

Palaniappan et al. [7] worked on development of a system for Intelligent Heart Disease Detection System using DT, NB and Neural Network. They used the cross-industry standard process for data mining (CRISP-DM) methodology, consisting of business and data understanding, date preparation, modelling, evaluation and deployment. They did it using Data Mining Extension (DMX), a language similar to SQL for the purposes of data mining. The model was built on 909 records with 15 medical factors. Classification Matrix and lift chart were used as evaluating methods. The best performing model was found to be NB with an accuracy of 86.12%, followed by Neural Networks (85.68%) and DT (80.4%) [7]. Nayak et al. used frequent item mining for attribute filtration and classification methods like NB, DT, KNN and SVM for heart disease prediction. They obtained the accuracies as—DT (69.81%), SVM (81.13%), NB (88.67%), KNN (ROC) (71.70%), and KNN (ACC) (67.92%) [1]. We can notice that obesity and smoking affects the risk a great deal but were not being taken into account. Thomas et al. managed to take two more attributes, obesity and smoking into consideration for predicting heart disease. They used KNN classifier and ID3 algorithm and found that the accuracy of their model rose up to 80.6% which was previously 40.3% with basic attributes [8].

Buettner et al. [9] built a prediction model using RF classifier for the same cause. The model was tested in 303 patients. They also analysed the difference in accuracy before and after implementing cross validation in Random Forest classifier. The accuracy was initially 82.895% which increased to 84.448% by using tenfold cross validation [9]. Mohan et al. [10] tried a different approach to build a more efficient model. They proposed a hybrid model called as Hybrid Random Forest with Linear Model (HRFLM) and compared the accuracy results. Experimenting with the conventional system of feature selection, they decided to have no restrictions on features. They compared the HRFLM with all the previously existing models and found that the HRFLM yielded enhanced results. The accuracies being, NB(75%), Generalized Linear model (85.1%), LR (82.9%), Deep Learning (87.4%), DT(85%), RF (86.1%), Gradient Boosted Trees (78.3%), SVM (86.1%), VOTE (87.41%) while it got enhanced to 88.4% for the HRFLM [10]. Bashir et al. [11] developed an ensemble model and compared it with other existing models. The ensemble model they proposed was found to be performing better than all the other classifiers. The accuracy being NB (78.79%), DT (72.73%), SVM (75.76%), while that of the proposed ensemble technique model performed 81.82% accurately [11].

Malav et al. [12] used the K-means clustering method for a change, along with Artificial Neural Network (ANN). They developed a hybrid model which yielded an accuracy of 97% which was higher than both NB (88%) and KNN (93%) [12].

Gadekallu et al. [13] worked on the optimization of heart disease prediction model. They used feature reduction using Cuckoo search (CS) with rough set theory and predicted the disease using fuzzy logic system. They tested it on five different datasets. The accuracy for Cleveland dataset came out to be 91% for the proposed model [13]. Khourdifi et al. [14] tried another approach by implementing optimization techniques on the classifiers. They used KNN, SVM, NB, RF and ANN optimized with the help of Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) techniques. The comparison showed that the best average accuracy without optimization was that of SVM (83.6%) and RF (81.4%).While, after optimization with FCBF, PCO and ACO, the best model turned out to be KNN and RF with 99.7% and 99.6% accuracy [14]. The comparison and analysis of all the related work is given in Table 1.

3 Methods and materials

This section is organized as follows—(3.1) Dataset description, (3.2) Data exploration and pre-processing techniques used in the model, and then we describe theoretically all the classifiers used in our approach as: (3.3) Concept of principal component analysis (PCA), (3.4) Concept of k-fold cross validation, (3.5) Logistic Regression (LR), (3.6) Support Vector Machine (SVM), (3.7) Naïve Bayes (NB), (3.8) K-Nearest Neighbours (KNN), (3.9) Decision Tree (DT), (3.10) Random Forest.

3.1 Dataset description

The dataset consists of age, sex, chest pain intensity ranging from 0 to 3, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0, 1, 2), maximum heart rate achieved by the patient, exercise-related factors, number of major vessels (0–3) coloured by fluoroscopy and.

Thal (3 being normal, 6 being fixed defect and 7 being reversible defect) [14].

3.2 Data exploration and pre-processing techniques

Gaining insights and understanding the data plays a crucial role in developing an efficient model. The better we can interpret the data, the more we can manipulate it to meet our needs. This section is further divided into subsections explaining about the data exploration and pre-processing techniques. It is organized as follows—(3.2.1) Variable identification, (3.2.2) Univariate Analysis, (3.2.3) Bivariate Analysis, (3.2.4) Missing value treatment, (3.2.5) Outlier treatment, (3.2.6) Variable transformation, (3.2.7) Variable creation.

3.2.1 Variable identification

Firstly, we need to identify the input variables and output variables of our dataset. There might be several output variables in a multiple class prediction model, while there is one output variable in a single class prediction model. We observed our data frame and found that we have a single output variable 'Target' which returns '0' if the person is not at risk of cardiovascular disease and '1' if the person might have or develop a cardiovascular disease. The rest of the variables were input variables.

3.2.2 Univariate analysis

Now, since the input variables are identified, we need to check their data one at a time. This can be done using tabular and graphical methods. Tabular methods include calculating the mean, median, and standard deviation of the variable and checking for missing values. Categorical variables can be observed using frequency tables. The graphical method helps to observe the data graphically to know if it is symmetric, left-skewed, right-skewed, or if there are outliers present in the data. Univariate analysis helps to detect anomalies in our dataset, so that it does not affect our prediction results. A boxplot can be used to detect outliers for a continuous variable, while a bar plot can be used to study a categorical variable. A bar plot can be used on the target variable to make sure that the dataset is balanced. Figure 1a shows a bar plot representing the count of 0 and 1 classes of the target variable while Fig. 1b shows a box plot representation.

3.2.3 Bivariate analysis

Bivariate analysis is used for determination of the relation between a couple of attributes at a time. It helps us to eliminate the features which do not have an impact on the prediction or variables which are very similar to each other. To find out the relations between the variables, different methods are used,

- Continuous—continuous variable: to determine the relationship between two continuous variables, correlation can be used. A scatter plot or correlation matrix may be used to visualize the relation. Figure 2 shows

Table 1 Literature review analysis

S. No	Ref. No	Technique	Advantage	Disadvantage/ future work
1	[6]	DT, SVM, KNN and NB	Four classifiers were used with ROC analysis to determine and compare efficacy of different models	No feature selection was used hence resulting in low sensitivity and specificity values for each classifier
2	[7]	CRISP-DM Methodology, Data Mining Extension (DMX), DT, NB and Neural Network	A thorough analysis to obtain useful subsets of data to form a hypotheses for hidden information	Limited dataset, restricted to use of categorical data only, only three data mining techniques were used
3	[1]	Attribute filtration by Frequent Itemset, DT, SVM, NB, KNN (ROC), KNN (Acc)	Detailed performance analysis on the basis of accuracy, sensitivity, specificity, positive predicted value (PPV), negative predicted value (NPV) and AUC	The highest accuracy obtained was 88%, which could be enhanced if ensemble machine learning techniques were used
4	[8]	KNN, ID3 algorithm	Additional parameters like obesity and smoking were taken into account, high increment in accuracy after taking additional attributes	Feature selection was not done, highest accuracy obtained was 80.6%
5	[9]	Cross validation technique, RF	The accuracy increased by 1.5% by using a tenfold cross validation on RF	Dataset doesn't include information about smoking, BMI or family history of the patient, only RF classifier was used
6	[10]	Hybrid Random Forest with Linear Model (HRFLM), DT, Language Model, SVM, RF, NB, KNN, Gradient boosted trees and Neural Networks	HRFLM yielded the best accuracy result juxtaposed with all the other existing methods	No restrictions were put on the features, a feature selection method could be used
7	[11]	Ensemble model using three base classifiers-NB, DT using Gini index and SVM, majority vote technique	Increased accuracy values as compared to other basic classifiers	The intensity of heart disease can be identified using fuzzy learning methods
8	[12]	Hybrid model using K-means clustering and ANN	A very high accuracy of about 97%	The high accuracy may be highly domain-restricted
9	[13]	Feature reduction using Cuckoo search, rough set theory, fuzzy logic system	Tested in five different datasets	Space complexity can be taken into consideration for an improvised model
10	[14]	Particle swarm optimization (PSO), Ant colony optimization (ACO), Fast correlation-based feature selection (FCBF), KNN, SVM, NB, RF and ANN	Optimized accuracy results using PCO and ACO optimization techniques	Work requires an extended duration of study with better processing power of computational resources

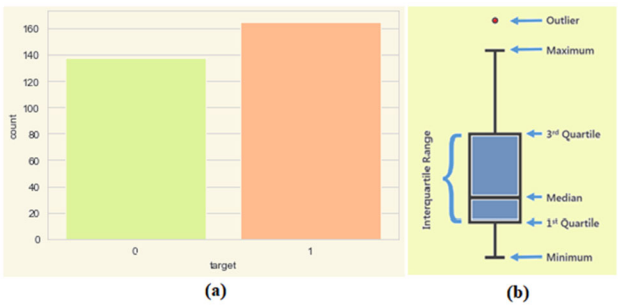

Fig. 1 **a** Bar plot a the 'target' variable. **b** Box plot for outlier detection

the correlation matrix of our dataset. The correlation matrix is coloured according to the correlation of two variables ranging from red representing a highly negative correlation to green representing a correlation of 1. Correlation is defined as,

$$Corr = \frac{Cov(M,N)}{\sigma_m \sigma_n} \tag{1}$$

where, Cov (M, N) = Covariance of M and N, $\sigma_m \sigma_n$ = standard deviations of M and N.

- Categorical variable—Continuous variable: We use the 2-Sample test to determine the relationship.

$$t = \frac{\bar{Z}_1 - \bar{Z}_2}{\sqrt{\frac{D_1^2}{n_1} + \frac{D_2^2}{n_2}}} \tag{2}$$

In (2), n_1 and n_2 are the sample sizes, with $\bar{Z}_1$ and $\bar{Z}_2$ as the mean or average values and D_1 and D_2 the standard deviations.

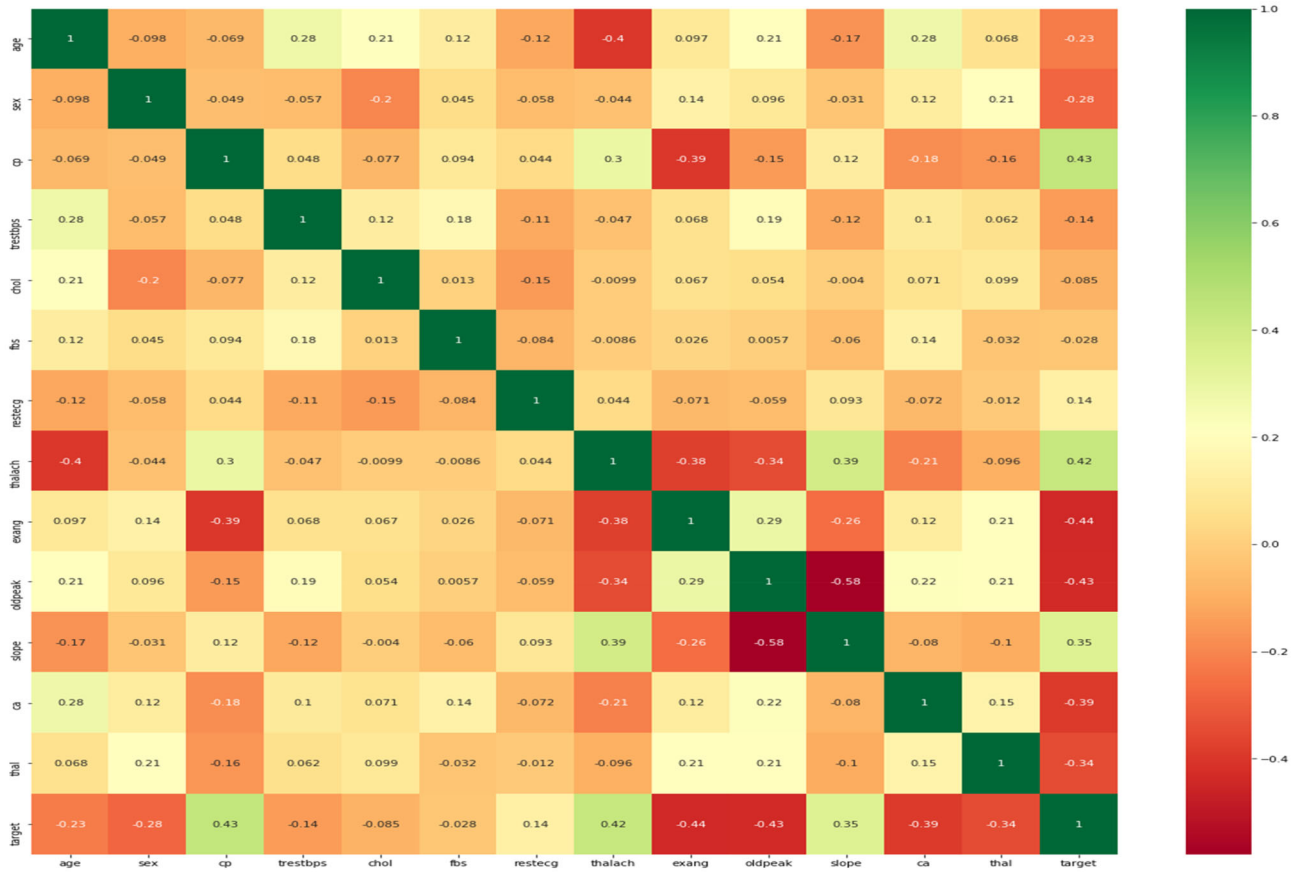

Fig. 2 Correlation matrix of risk factors of heart disease

- Categorical—Categorical variables: Chi-square test is used to compute the relationship.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{3}$$

Where O is observed frequencies and E is expected frequencies

3.2.4 Missing value treatment

Missing or null values can be treated in two ways, by imputation, that is, by replacing the missing data with mean or mode of the variable or by deleting the rows with missing values present in them.

3.2.5 Outlier treatment

Outliers may exist due to data entry, measurement, or processing errors. They can be treated in many ways, by deleting, transforming (binning), imputing using mean and mode or by treating them separately.

3.2.6 Variable transformation

Sometimes the dataset can be skewed, containing non-linear relations or of different scales. To solve this problem, a variable transformation is executed. Some of the most common methods are using logarithm, square root, cube root, and binning.

3.2.7 Variable creation

New variables can be created from the existing ones to explore a hidden relationship between two features.

3.3 Concept of principal component analysis (PCA)

PCA is a technique to get a low dimensional structure out of a potential high dimensional dataset. It includes the extraction of q eigenvectors for q input distribution [15]. It is one of the most used algorithm for dimensionality reduction. The basis vectors are known as principal components. The basic objective of PCA is to find patterns and correlations among the dataset. PCA is usually used to standardize, obtain eigenvalues and eigenvectors and then,

sort the eigenvalues in descending order and make a projection matrix to transform the original dataset.

3.4 Concept of K-fold cross validation

K-fold cross validation is a validation technique to ensure the model performs well on unseen data. The dataset is shuffled randomly and divided into k folds or subgroups. A group is taken as a hold out and the remaining groups are considered as the training set. The model is tested on the hold out set and the evaluation results are stored. This is repeated for each fold. The k-fold cross validation is helpful for avoiding overfitting or underfitting of model. It can also be useful in determining the consistency of the model.

3.5 Logistic Regression (LR)

Logistic Regression (LR) is a classification method that can be used to classify categorical variables. The model is also known to be a model that operates statistically using a sigmoidal function on a binary dependent variable in order to predict results. To obtain a logistic function, sigmoid function is applied to the formula of linear regression. Equations 4, 5 and 6 represent linear function, sigmoid function and logistic function respectively.

$$y = b_0 + b_1 x \tag{4}$$

$$p = \frac{1}{1 + e^{-y}} \tag{5}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x \tag{6}$$

3.6 Support vector machine (SVM)

Support vector machines (SVM) were initially developed in the 1960s, modified in the 1980s but are becoming a popular choice among researches recently [16–18]. It is because SVMs are very different from the rest of the machine learning algorithms. SVM selects the maximum-margin separating hyper plane (also known as optimal hyper plane) in order to maximize its ability to predict correctly. Due to its uniqueness, SVM sometimes manages to predict way more accurately than other algorithms. SVM is being used largely to get a more accurate prediction, especially in medical science, to predict crucial datasets like cancer classification [19–22].

3.7 Naïve Bayes (NB)

The Naïve Bayes (NB) classifier is a heavily simplified version of the Bayesian probability model. It assumes features as independent classes. Although the independence is a rebellious assumption, Naïve Bayes is often found to compete very well with other sophisticated classifiers. NB is believed to be the best performing when the features are completely independent or when the features are functionally dependent. However, it does not work well when the features fall between these two extremes [23–27]. The Bayesian probability model can be mathematically stated as,

$$P(y|X) = \frac{P(X|y).P(y)}{P(X)} \tag{7}$$

$$P(y|X) = P(x_1|y).P(x_2|y)\ldots P(x_n|y).P(y) \tag{8}$$

The NB classifier can be represented as,

$$y_{NB} = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \tag{9}$$

3.8 K-Nearest Neighbour Classifier (KNN)

Among various other statistical pattern recognition algorithms, KNN often achieves high performance [28]. It is used for a variety of classification problems to yield good results [29–33]. The model involves the training of both positive and negative cases. A new sample gets classified according to its nearest training set. KNN classifies all the similar data points nearby [34]. For classification using the KNN model, the new 'test' instance is first located and its distance from all the train data points is calculated. The distances are sorted in ascending order and first k distances are selected. The right value of k can be calculated using the Elbow method. The distance can be calculated in four ways, being Manhattan distance, Euclidean distance, Minkowski distance, and Hamming distance. The most commonly used is the Euclidean distance. Figure 3 shows a K-value graph for our dataset which represents the number

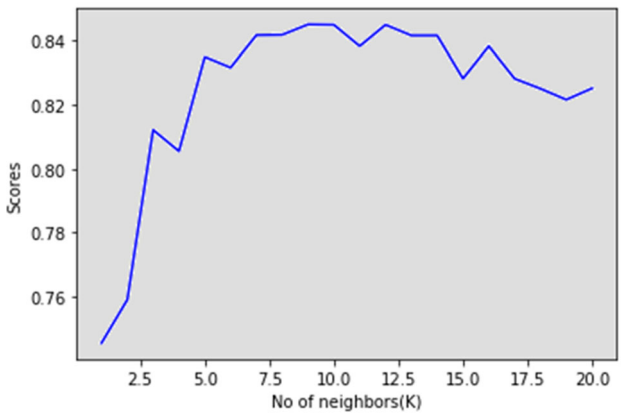

Fig. 3 K-value graph for the heart prediction dataset

of neighbours in the abscissa and their respective scores in the ordinate. Euclidean distance can be represented as,

$$D_E = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{10}$$

3.9 Decision tree (DT)

The DT classifier can be defined a multilevel approach based on the idea of breaking down a complex problem into a union of simpler decisions [35]. It has a single root node which is split further using decision nodes. The decision node which splits up is called the parent node of the successor (or child) nodes. The terminal nodes of the tree are called the leaf nodes. The depth of a tree can be calculated by computing the longest distance from the root to leaves. Figure 4 shows a representation of a typical DT structure. The node split in a decision tree is decided based on its Gini impurity (1-Gini). The lower the Gini impurity, the more homogeneous nodes exist. The DT prediction can be optimized by specifying the lowest number of samples required for split and nodes or by specifying the maximum depth of the tree. The DT classifier proves to be useful and competes well with other classification models [36–38]. This approach is often experimented with to boost the accuracy of the model by using various techniques or making a hybrid decision tree [39–46].

$$GiniImpurity = 1 - Gini \tag{11}$$

$$Gini = (p_1^2 + p_2^2 + \ldots p_n^2) \tag{12}$$

3.10 Random Forest (RF)

The RF classifier which was introduced by Breiman in 2001 [47], consists of multiple tree predictors and each tree depends on the value of randomly chosen vector. For classifying using RF, firstly, 'k' number of random data points are chosen from the training set. Decision trees are built for those 'k' data points, and the number of required trees is chosen. Then, for a new data point, a prediction is made through all of the trees and the classifier takes the majority vote for prediction. Fig. 5 shows a mechanism of the Random Forest classifier with four trees obtaining the final class with the help of majority voting. Random Forest has been proven to be efficient and yields good accuracy in many different fields of research [55–60].

4 Results and comparisons

Predictions were carried out through six classifiers - LR, NB, SVM, DT, KNN, and RF. Some specific parameters were used in order to optimize the accuracy of the models. Logistic Regression was used with parameter of random state specified as 1 to ensure getting the same result after multiple runs. In SVC (support vector classifier), we used the 'linear' kernel and set the probability to true. In KNN and RF models, we specified the cross-validation generator (CV) as 10. In RF model, the no. of trees (n estimators) was set to 100. In DT model, the lowest number of samples needed to be at a leaf node was set to 20. Some of the methods and terminologies related to evaluation of a model are -

- **True positive (TP)—**The values which were predicted positive by the model and were actually positive fall under true positive.
- **False negative (FN)—**Negative predictions of the model which were actually positive.
- **False positive (FP)—**Positive predictions of the model which were actually negative.

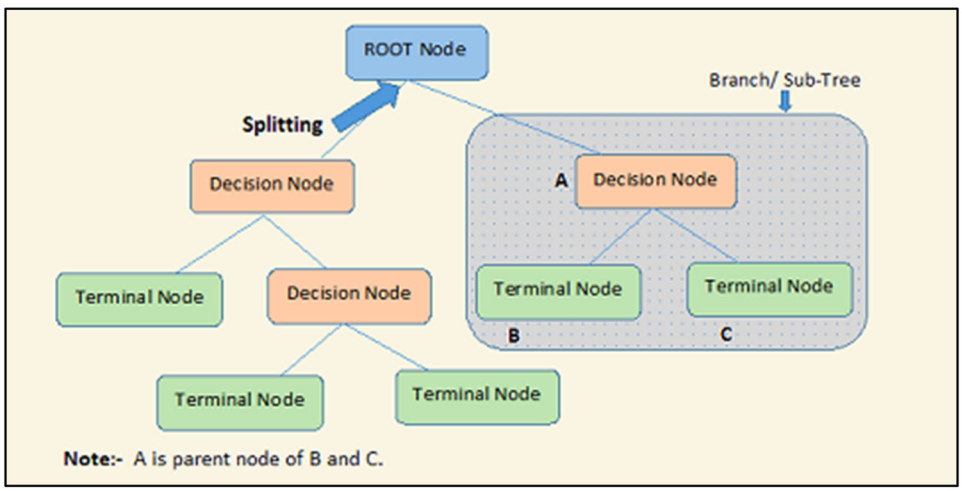

Fig. 4 Decision tree representation

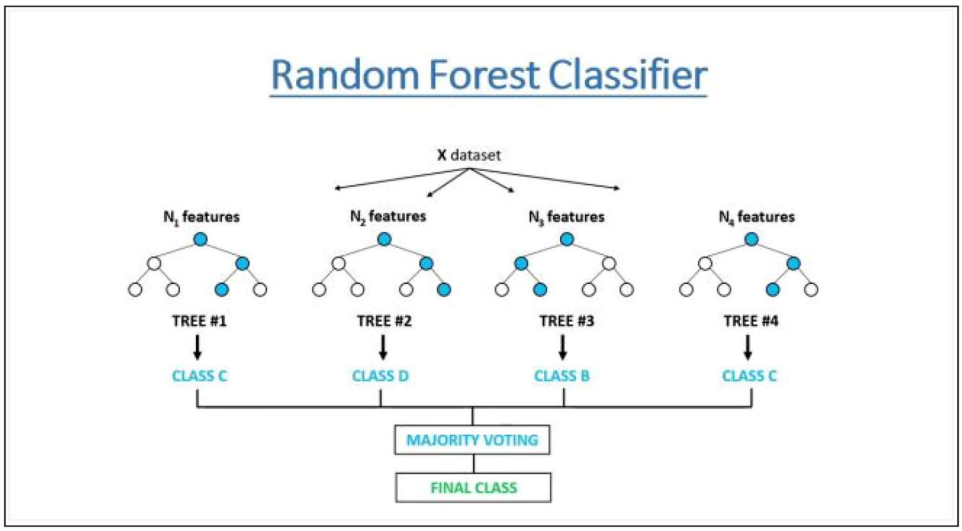

Fig. 5 Random Forest representation

- **True Negative (TN)—**Negative predictions of the model which were actually also negative.
- **Confusion matrix—**An $n \times n$ matrix where n is the number of classes in the model. It contains the values TP, FN, FP, and TN. It is used to get clarity of a classification model's performance. Figure 6a– f show the confusion matrices for LR, NB, SVM, KNN, DT and RF respectively.
- **Precision—**Precision can be defined as the ratio of predictions actually positive and the total predicted positive. It can be represented as,

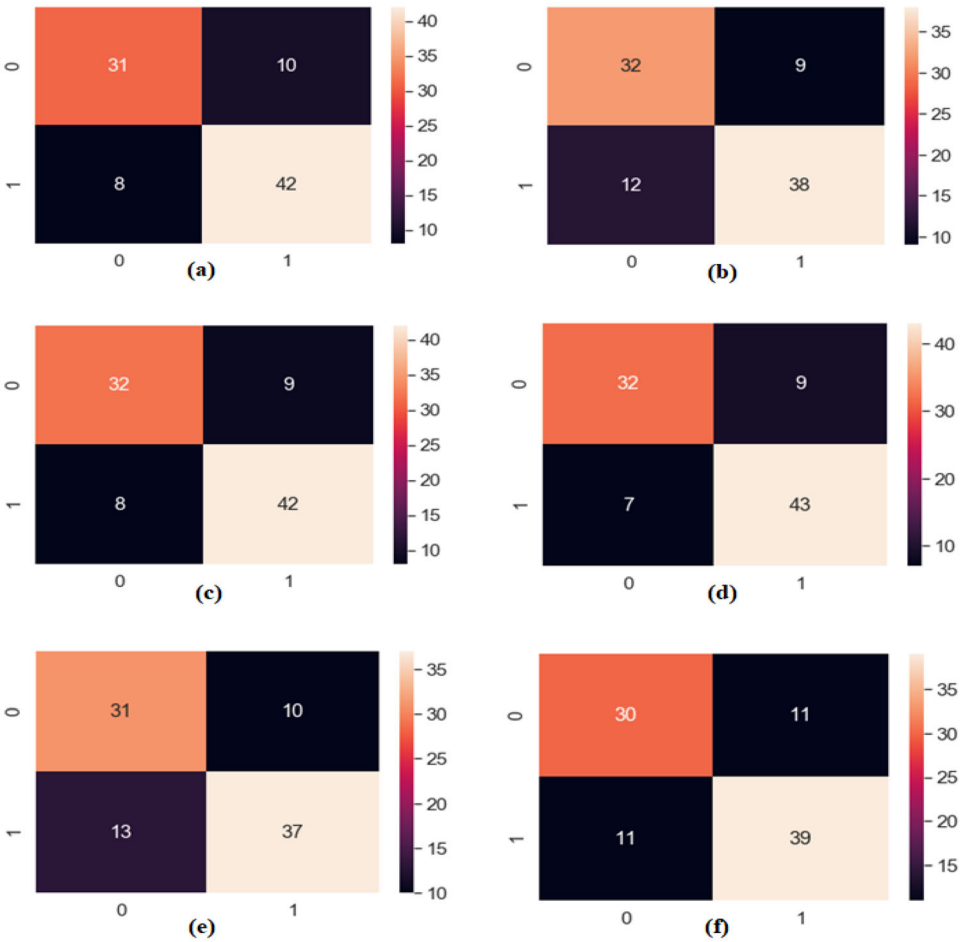

Fig. 6 **a** Confusion matrix (CM) for Logistic Regression. **b** CM for Naïve Bayes. **c** CM for SVM. **d** CM for KNN. **e** CM for Decision Tree. **f** CM for Random Forest

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

- **Recall (Sensitivity)—**Recall can be divided as the ratio of predictions actually positive and the total actual positive. It can be represented as,

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

- **Specificity—**Specificity can be represented as,

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

- **F1-score—**F1-score also known as balanced F-score or F-measure can be computed by taking the weighted average of precision and recall. It can be represented as,

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \tag{16}$$

K-fold cross validation and PCA were applied to the dataset after feature selection through correlation matrix. Table 2 shows the performance analysis of all the classifiers in terms of accuracy, precision, recall, F1-score and specificity. It shows that the LR classifier gave the best accuracy result while KNN proved to be a better model when taking into consideration precision, recall and specificity. Figure 7 shows the line graph with exact accuracy values for all the classifiers with the name of classifiers in the abscissa and accuracy percentage in the ordinate. Figures 8, 9, 10, 11 show a bar graph comparison of the precision values, recall values, F1-scores and specificity values of all the classifiers for normal dataset accuracy, cross validation accuracy fivefold, tenfold, 15-fold and PCA accuracy respectively. Further, in Tables 3, 4, 5, 6 we plot the metrics obtained for cross validations of fivefold, tenfold, 15-fold, and when the data is orthogonally transformed through PCA. The bold values in Tables 2, 3, 4, 5, 6 correspond to the highest in the column (or highest in the metric). We discuss, compare and contrast the results achieved by different methodologies further.

- **Accuracy—**It measures the extent to which the classifier correctly predicts data samples on the test set. The accuracy values were 85.11% for LR, 83.13% for SVM, 80.18% for NB, 84.48% for KNN, 74.72% for DT and 83.48% for RF for normal approach.

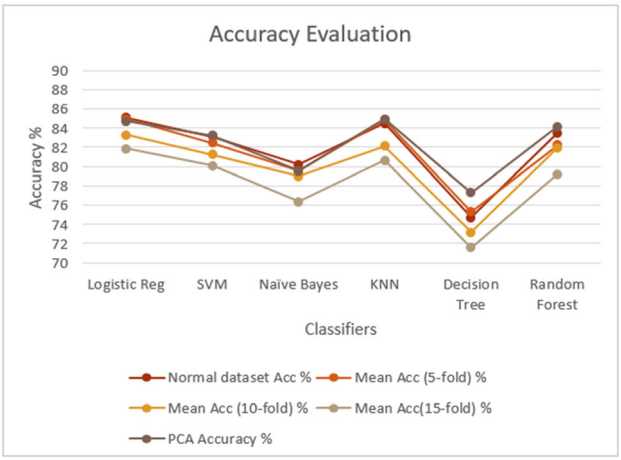

Fig. 7 Accuracy line graph comparison of normal dataset accuracy, fivefold cross validation, tenfold cross validation, 15-fold cross validation and PCA accuracy

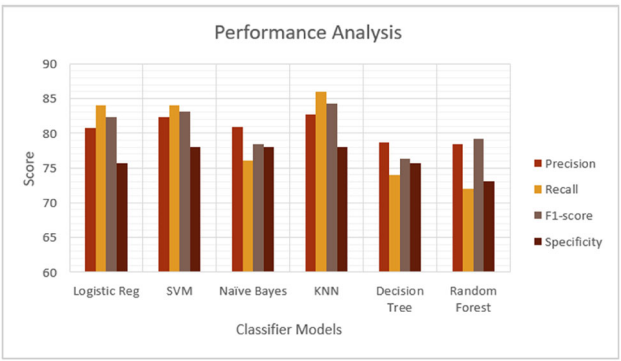

Fig. 8 Precision, recall, F1-score and specificity analysis through bar graph

- **Receiver operating characteristics (ROC) curve—**ROC curve is a graph of TP rate vs. FP rate on the Cartesian plane. AUC is the abbreviated term for area under the curve of the ROC. The AUC score is often used for analysing the performance of a model. Figure 12 shows the ROC curves for all the classifiers used in our model giving the best ROC curves for KNN and SVM while lowest ROC curve for DT classifier.

Table 2 Accuracy, precision, recall, F1-score and specificity comparison table

Models	Accuracy %	Precision %	Recall %	F1-score %	Specificity %
Logistic Regression	**85.118**	80.769	84	82.352	75.6
SVM	83.139	82.352	84	83.168	78.0
Naïve Bayes	80.182	80.851	76	78.35	78.0
KNN	84.483	**82.692**	**86**	**84.313**	**78.0**
Decision tree	74.725	78.723	74	76.288	75.6
Random Forest	83.483	78.431	72	79.207	73.1

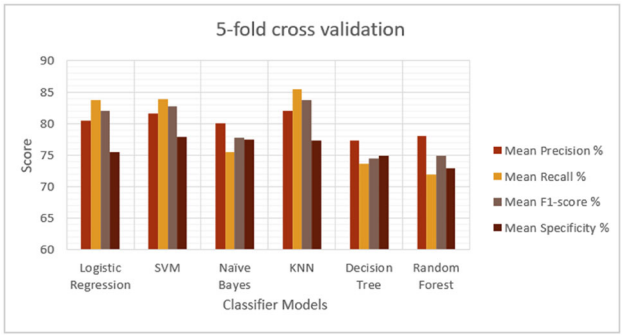

Fig. 9 Precision, recall , F1-score and specificity after fivefold cross validation

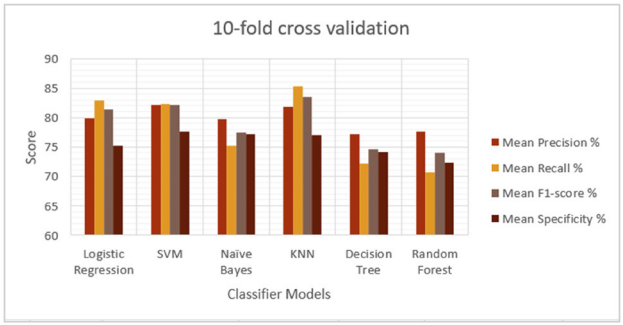

Fig. 10 Precision, recall, F1-score and specificity after tenfold cross validation

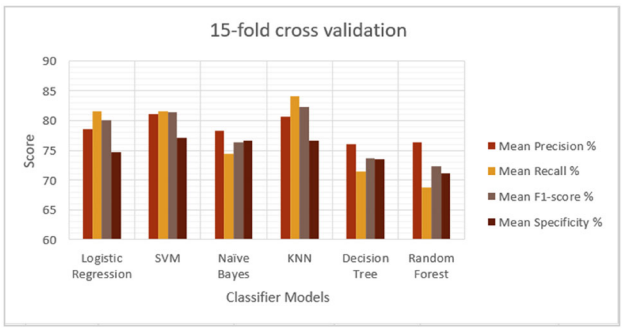

Fig. 11 Precision, recall, F1-score and specificity after 15-fold cross validation

From the evaluation process mentioned above, we come to notice that LR performed the best in terms of accuracy in Table 2, while KNN having 0.635% less accuracy than LR, outperforms when precision, recall and specificity is taken into consideration. SVM also competes with LR and KNN when precision, recall, specificity or AUC score is observed. So, depending on the problem here, we know that detecting a patient with CVD risk is more important than other factors. That means, our main focus is minimizing the false negative (FP). In simpler words, we need a model with the highest recall value. So, analysing all the conditions, while LR, SVM and KNN all performed good in different aspects, KNN might prove to be the best model to detect the risk of a CVD. Moreover, from Tables 3, 4, 5, we notice a diminishing trend in the mean of all metrics when the number of folds are increased, this is due to the averaging effect of different metrics attained over different numbers of folds (5, 10 and 15). It is evident that k-fold cross validation is not a technique that increases performance of models, but rather gives a proper intuitive approximation of the true performance of a model. We notice from Table 6 that application of PCA on the data has both positive and negative impact, depending upon the model. For SVM, KNN, DT and RF, performance is slightly improved. The reason for this improvement could be attributed to the fact that orthogonal rotation of data points maximizes variance which renders it easier for models such as SVM, KNN, DT and RF to demarcate regions and construct more distinctive decision boundaries for effective classification. Over all these variations, KNN still manages to achieve the most number of maximum values in terms of the different metrics used.

5 Conclusion

The objective of this paper was to build a cardiovascular disease prediction model using six classifiers—Logistic Regression, SVM, Naïve Bayes, KNN, Random Forest, and decision tree and find the best one. Cross validation technique was used to better evaluate our model's accuracy. PCA was also applied on the data to analyse differences in performance of the models. Their predictions were compared with each other using confusion matrices and ROC curves. Logistic regression was observed to be have the

Table 3 Accuracy, precision, recall, F1-score and specificity after fivefold cross validation

Models	Mean accuracy %	Mean precision %	Mean recall %	Mean F1-score %	Mean specificity %
Logistic Regression	**84.928**	80.486	83.72	82.071	75.445
SVM	82.439	81.558	83.88	82.702	77.878
Naïve Bayes	79.566	80.112	75.54	77.758	77.427
KNN	84.883	**82.042**	**85.52**	**83.744**	77.366
Decision tree	75.332	77.313	73.69	74.458	**74.882**
Random Forest	82.232	78.102	71.86	74.851	72.978

Table 4 Accuracy, precision, recall, F1-score and specificity after tenfold cross validation

Models	Mean accuracy %	Mean precision %	Mean recall %	Mean F1-score %	Mean specificity %
Logistic Regression	**83.282**	79.925	82.95	81.409	75.221
SVM	81.233	**82.112**	82.32	82.215	**77.587**
Naïve Bayes	78.989	79.791	75.14	77.395	77.152
KNN	82.124	81.887	**85.34**	**83.577**	77.032
Decision tree	73.163	77.169	72.25	74.628	74.217
Random Forest	81.989	77.552	70.66	73.945	72.378

Table 5 Accuracy, precision, recall, F1-score and specificity after 15-fold cross validation

Models	Mean accuracy %	Mean precision %	Mean recall %	Mean F1-score %	Mean specificity %
Logistic Regression	**81.853**	78.579	81.56	80.041	74.662
SVM	80.089	**81.155**	81.58	81.366	**77.151**
Naïve Bayes	76.354	78.257	74.39	76.274	76.672
KNN	80.669	80.658	**84.13**	**82.357**	76.677
Decision tree	71.572	75.993	71.42	73.635	73.451
Random Forest	79.227	76.313	68.8	72.362	71.166

Table 6 Accuracy, precision, recall, F1-score and specificity comparison table by PCA feature selection

Models	Accuracy %	Precision %	Recall %	F1-score %	Specificity %
Logistic Regression	84.689	80.229	84	82.071	75.2
SVM	83.255	82.586	85	83.775	78.1
Naïve Bayes	79.524	80.197	75	77.551	77.8
KNN	**84.925**	**83.298**	**87**	**85.108**	**78.2**
Decision tree	77.319	80.021	76	77.958	76.2
Random Forest	84.167	79.952	74	76.860	73.7

highest accuracy of 85.11% on the normal approach. But according to the needs of this particular problem, which is to focus on correctly detecting the heart disease, KNN was observed to be better performing with the highest recall of 86%. In the current scenario, there is no standard method to diagnose a cardiovascular disease, partly because this disease does not show any major symptoms until complications occur [48]. Thus, prediction approaches using machine learning techniques can prove to be extremely helpful to medical science in future.

6 Future work

Prediction models can be optimized by experimenting with different hybrid classifiers or using ensemble modelling to achieve more accurate results to help the patients to take the necessary preventive measures on time. Our work is novel owing to extensive experimentation of various models on different folds with the application of PCA, however, more methods such as linear discriminant

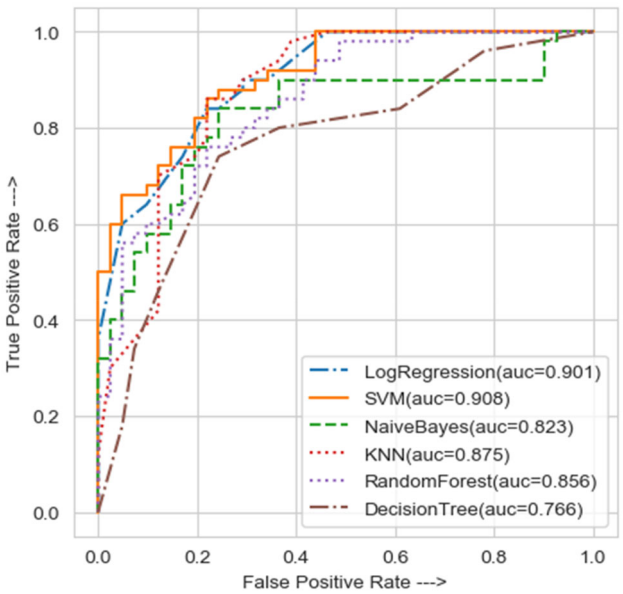

Fig. 12 ROC curve comparison for all the classifiers for normal approach

analysis (LDA) can be applied. Optimization techniques such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Genetic Algorithm (GA) can also be implemented for better results. Grid search can be used for gaining optimal hyperparameters. Features may be extracted using deep learning architectures such as Convolutional Neural Networks (CNN) and Boltzmann Machines for better feature vector representations and higher results.

Acknowledgements To co-authors for their support and guidance throughout the research work.

Funding Not applicable.

Data availability The dataset was taken from UCI repository of machine learning.

Code availability Not applicable.

Declarations

Conflict of interest There is no conflict of interest.

References

1. Nayak S, Gourisaria MK, Pandey M, Rautaray SS (2019) Prediction of Heart Disease by Mining Frequent Items and Classification Techniques. 2019 International conference on intelligent computing and control systems (ICCS). Doi: https://doi.org/10.1109/iccs45141.2019.9065805
2. Devaraj S, Wang-Polagruto J, Polagruto J, Keen CL, Jialal I (2008) High-fat, energy-dense, fast-food–style breakfast results in an increase in oxidative stress in metabolic syndrome. Metabolism 57:867–870. https://doi.org/10.1016/j.metabol.2008.02.016
3. Freedman DS, Khan LK, Dietz WH, Srinivasan SR, Berenson GS (2001) Relationship of childhood obesity to coronary heart disease risk factors in adulthood: the bogalusa heart study. Paediatrics 108:712–718. https://doi.org/10.1542/peds.108.3.712
4. Williams PT (2001) Physical fitness and activity as separate heart disease risk factors: a meta-analysis. Med Sci Sports Exerc 33:754–761
5. Gm H, Gourisaria MK, Pandey M, Rautaray SS (2020) A comprehensive survey and analysis of generative models in machine learning. Comput Sci Rev 38:100285. https://doi.org/10.1016/j.cosrev.2020.100285
6. Nayak S, Gourisaria MK, Pandey M, Rautaray SS (2020) Comparative analysis of heart disease classification algorithms using big data analytical tool. In Second international conference on computer networks and communication technologies (ICCNCT 2019) 44:582–588. Doi: https://doi.org/10.1007/978-3-030-37051-0.
7. Palaniappan S, Awang R (2008) Intelligent heart disease prediction system using data mining techniques. 2008 IEEE/ACS international conference on computer systems and applications. doi:https://doi.org/10.1109/aiccsa.2008.4493524
8. Thomas J, Princy RT (2016) Human heart disease prediction system using data mining techniques. 2016 International conference on circuit, power and computing technologies (ICCPCT). Doi: https://doi.org/10.1109/iccpct.2016.7530265
9. Buettner R, Schunter M (2019) Efficient machine learning based detection of heart disease. 2019 IEEE International conference on e-health networking, application and services (HealthCom). Doi: https://doi.org/10.1109/healthcom46333.2019.9009429
10. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. https://doi.org/10.1109/access.2019.2923707
11. Bashir S, Qamar U, Younus Javed M (2014) An ensemble based decision support framework for intelligent heart disease diagnosis. International conference on information society (i-Society 2014). Doi: https://doi.org/10.1109/i-society.2014.7009056
12. Malav A, Kadam K, Kamat P (2017) Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy. Int J Eng Technol 9(4):3081–3085
13. Gadekallu TR, Khare N (2017) Cuckoo search optimized reduction and fuzzy logic classifier for heart disease and diabetes prediction. Int J Fuzzy Syst Appl 6:25–42. https://doi.org/10.4018/ijfsa.2017040102
14. Khourdifi Y, Bahaj M (2019) Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. Int J Intell Eng Syst 12(1):242–252
15. Kim KI, Jung K, Kim HJ (2002) Face recognition using kernel principal component analysis. IEEE Signal Process Lett 9(2):40–42. https://doi.org/10.1109/97.991133
16. Niu XX, Suen CY (2012) A novel hybrid CNN–SVM classifier for recognizing handwritten digits. Pattern Recogn 45(4):1318–1325
17. Dadi HS, Pillutla GM (2016) Improved face recognition rate using HOG features and SVM classifier. IOSR J Electr Commun Eng 11(4):34–44
18. Alkan A, Günay M (2012) Identification of EMG signals using discriminant analysis and SVM classifier. Expert Syst Appl 39(1):44–47
19. Noble WS (2006) What is a support vector machine? Nat Biotechnol 24:1565–1567. https://doi.org/10.1038/nbt1206-1565
20. Rejani Y, Selvi ST (2009) Early detection of breast cancer using SVM classifier technique. arXiv preprint. arXiv:0912.2314
21. Vijayarajeswari R, Parthasarathy P, Vivekanandan S, Basha AA (2019) Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. Measurement 146:800–805
22. Polat K, Güneş S (2007) Breast cancer diagnosis using least square support vector machine. Digit Signal Process 17(4):694–701
23. Rish I (2001) An empirical study of the Naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence 3:41–46
24. Mukherjee S, Sharma N (2012) Intrusion detection using naive bayes classifier with feature reduction. Proced Technol 4:119–128. https://doi.org/10.1016/j.protcy.2012.05.017
25. Granik M, Mesyura V (2017) Fake news detection using naive Bayes classifier. In 2017 IEEE First Ukraine conference on electrical and computer engineering (UKRCON), pp 900–903. IEEE
26. Sebe N, Lew MS, Cohen I, Garg A, Huang TS (2002) Emotion recognition using a cauchy naive bayes classifier. In Object recognition supported by user interaction for service robots (Vol. 1, pp. 17–20). IEEE
27. Liu B, Blasch E, Chen Y, Shen D, Chen G (2013) Scalable sentiment classification for big data analysis using naive bayes classifier. In 2013 IEEE international conference on big data, pp 99–104
28. Suguna N, Thanushkodi K (2010) An improved k-nearest neighbor classification using genetic algorithm. Int J Comput Sci Issues 7(2):18–21

29. Liu CL, Lee CH, Lin PM (2010) A fall detection system using k-nearest neighbor classifier. Expert Syst Appl 37(10):7174–7181
30. Arif M, Malagore IA, Afsar FA (2012) Detection and localization of myocardial infarction using k-nearest neighbor classifier. J Med Syst 36(1):279–289
31. Shen H, Chou KC (2005) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334(1):288–292
32. Yazdani A, Ebrahimi T, Hoffmann U (2009) Classification of EEG signals using Dempster Shafer theory and a k-nearest neighbor classifier. In 2009 4th International IEEE/EMBS conference on neural engineering, pp 327–330
33. Liao Y, Vemuri VR (2002) Use of k-nearest neighbor classifier for intrusion detection. Comput Secur 21(5):439–448
34. Islam MJ, Wu QMJ, Ahmadi M, Sid-Ahmed MA (2007) Investigating the performance of Naïve Bayes classifiers and K-Nearest Neighbor Classifiers. 2007 International conference on convergence information technology (ICCIT 2007). Doi: https://doi.org/10.1109/iccit.2007.148
35. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21:660–674. https://doi.org/10.1109/21.97458
36. Stein G, Chen B, Wu AS, Hua KA (2005) Decision tree classifier for network intrusion detection with GA-based feature selection. In Proceedings of the 43rd annual Southeast regional conference, 2, 136–141
37. Friedl MA, Brodley CE (1997) Decision tree classification of land cover from remotely sensed data. Remote Sens Environ 61(3):399–409
38. Cramer GM, Ford RA, Hall RL (1976) Estimation of toxic hazard—a decision tree approach. Food Cosmet Toxicol 16(3):255–276
39. Freund Y, Mason L (1999) The alternating decision tree learning algorithm. In icml (Vol. 99, pp. 124–133)
40. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Liu TY (2017) Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 3146–3154
41. Kohavi R (1996) Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. Kdd 96:202–207
42. Utgoff PE, Berkman NC, Clouse JA (1997) Decision tree induction based on efficient tree restructuring. Mach Learn 29(1):5–44
43. Turney PD (1994) Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. J Artif Intell Res 2:369–409
44. Polat K, Güneş S (2007) Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. Appl Math Comput 187(2):1017–1026
45. Farid DM, Zhang L, Rahman CM, Hossain MA, Strachan R (2014) Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Syst Appl 41(4):1937–1946
46. Carvalho DR, Freitas AA (2004) A hybrid decision tree/genetic algorithm method for data mining. Inf Sci 163(1–3):13–35
47. Scornet E (2016) On the asymptotics of random forests. J Multivar Anal 146:72–83. https://doi.org/10.1016/j.jmva.2015.06.009
48. Grainger DJ (2006) Metabolic profiling in heart disease. Heart Metab 32:22–25
49. Blake CL, Merz CJ (1998) UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html. Accessed 9th July 2020
50. Shinde R, Arjun S, Patil P, Waghmare J (2015) An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. Int J Comput Sci Inf Technol (IJCSIT) 6:637–639
51. Gourisaria MK, Das S, Sharma R, Rautaray SS, Pandey MA (2020) Deep learning model for malaria disease detection and analysis using deep convolutional neural networks. Int J Emerg Technol 11:699–704
52. Das S, Sharma R, Gourisaria MK, Rautaray SS, Pandey M (2020) Heart diseases heart disease detection using core machine learning and deep learning techniques: a comparative study. Int J Emerg Technol 11:531–538
53. Rautaray SS, Dey S, Pandey M, Gourisaria MK (2020) Nuclei segmentation in cell images using fully convolutional neural networks. Int J Emerg Technol 11:731–737
54. Nayak S, Gourisaria MK, Pandey M, Rautaray SS (2019) Prediction of heart disease by mining frequent items and classification techniques. 3rd International conference on intelligent computing and control systems, pp. 607–611. Doi: https://doi.org/10.1109/ICCS45141.2019.9065805
55. Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217–222
56. Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 67:93–104
57. Masetic Z, Subasi A (2016) Congestive heart failure detection using random forest classifier. Comput Methods Programs Biomed 130:54–64
58. Azar AT, Elshazly HI, Hassanien AE, Elkorany AM (2014) A random forest classifier for lymph diseases. Comput Methods Programs Biomed 113(2):465–473
59. Nguyen C, Wang Y, Nguyen HN (2013) Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic
60. Zabihi M, Rad AB, Katsaggelos AK, Kiranyaz S, Narkilahti S, Gabbouj M (2017) Detection of atrial fibrillation in ECG hand-held devices using a random forest classifier. In 2017 computing in cardiology (CinC) (pp. 1–4). IEEE