Multiclass Prediction of Heart Disease Patients Using Big Data Analytics

Sarita Mishra, Manjusha Pandey, Siddharth Swarup Rautaray, and Sabyasachi Chakraborty

Abstract The rapidly growing rate of illness and death is the result of many diseases. Another major factor is cardiovascular disease (CVD) due to heart failure. According to statistics from around the world, the highest rate of natural death is caused by heart problems. The number of deaths resulting from this can be controlled by the early detection of heart disease chances in a person. Big data and several machine learning technologies have made it possible to discover the chances of a cardiac issue in a person in much advance. Many data scientists have successfully exploited the big data available for heart disease patients and have developed prediction models using different algorithms that are non-invasive, accurate, and appear to be very effective in analyzing patients' characteristics and detecting the presence or absence of heart disease in them. However, to provide appropriate preventive measures and appropriate treatment to patients, it is not enough to detect the presence of CVD, but the degree of impact the disease has left on a person needs to be measured. In this paper, we have compared the performance of five different machine-based algorithms (Logistic Regression, Support Vector Machine, Random Forest, KNN, and Naïve Bayes) which are used to classify the cardiovascular disease into five different classes. 0–4) with the increasing value from 0. These algorithms are used in their most common ways and in the One-vs-All method with the best performance in the latest scenario. The results of this study showed that the KNN algorithm provided 99.56% best predictive accuracy with a combination of One-vs-all and Principal Component Analysis strategies that surpassed all other algorithms.

Keywords Cardiovascular Disease (CVD) · Multiclass Classification · One-vs-All (OVA) · Principal Component Analysis (PCA)

S. Mishra (✉) · M. Pandey · S. S. Rautaray
School of Computer Engineering, KIIT Deemed to Be University, Bhubaneswar,
Odisha 751024, India
e-mail: 2181058@kiit.ac.in

S. Chakraborty
Chief Technology Officer, Bhusan, South Korea

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big
Data 114, https://doi.org/10.1007/978-981-19-5154-1_11

1 Introduction

The human heart is a vital organ about the size of a big fist and weighed 230 to 340 grams. Its role is to oxidize deoxygenated blood and pump it throughout the body as part of the circulatory system. The heart resides in a two-walled sac, called the pericardium, which protects the heart within the chest. The fluid, known as pericardial fluid, flows between two layers of the pericardium that keep the heart lubricated during various heart movements, the diaphragm, and the lungs. The heart performs its function in two cycles: the circulatory system and the circulatory system. In the pulmonary circulation, the oxygen-deprived blood absorbed by the heart during the process of inhalation reaches the lungs through the pulmonary artery, receives oxygen there, and returns to the heart through the pulmonary artery to the left auricle [12, 14]. In systemic circulation, oxygenated blood from the left auricle descends to the left ventricle and eventually leaves the heart through the aorta which separates and divides into many arteries and capillaries that supply oxygenated blood to all parts of the body. Under Fig. 1 shows the workings of the human heart.

Any blockage in any of these blood vessels blocks the smooth flow of blood and may lead to heart attacks. [13, 15] There are many factors that can lead to such blockages including high-cholesterol diets, diet. excessive fat, physical inactivity, stress at work, sleep disturbances, air pollution, alcohol or tobacco use, etc.

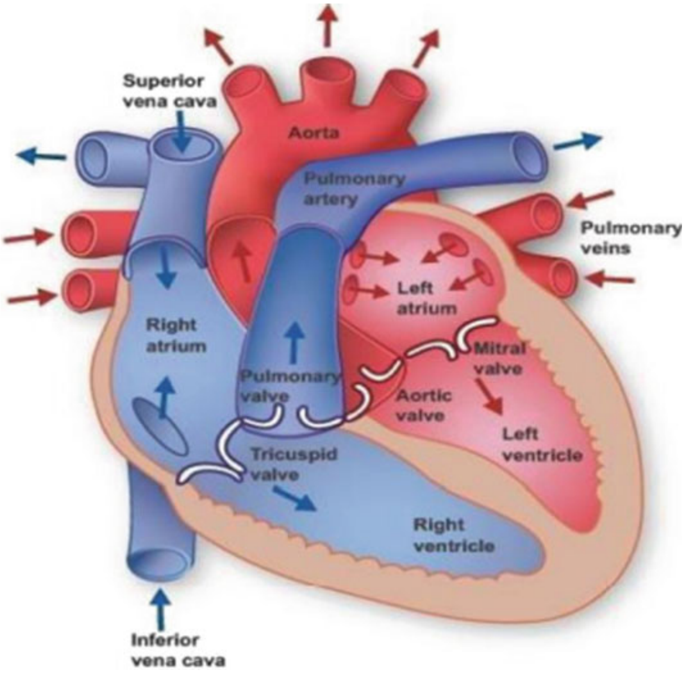


Fig. 1 Working of the human heart

According to the World Health Organization (WHO), approximately 17.9 million people die each year from heart disease [1]. The fast-growing mortality rate can be controlled by predicting the risk of early CVD in a person [2]. Heart disease prognosis can be made with the help of advanced machine-based models that have proven to be very useful for both patients and physicians. Most machine-based predictive models that have been developed so far can differentiate between patients by simply detecting the presence or absence of any heart problem in them. However, in this paper, we have emphasized the division of patients into five categories that reflect the degree of impact of the disease on them and thus provide a deeper understanding of the patient's health status.

In this research work, heart patients are divided into five categories (0 to 4) with a category 0 indicating the absence of the disease, a category 1 showing a minor effect of the disease, and thus an increase in the number of classes showing an increase in disease. criticism of Sect. 4 which means a very critical situation. The algorithms used for this category are Logistic Regression (LR), Vector Support Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Naïve Bayes (NB). Here we have compared the performance of these algorithms with their standard methods and their use of One-vs-All (OVA) in which they create a multi-stage multi-level division instead of a single multi-phase division. The results of this research work showed that the latter provides much better accuracy than the first. OVA-based performance is further enhanced by the Key Component Analysis (PCA) system which has increased the predictive accuracy of algorithms with KNN showing the best performance with 99.56% accuracy.

This paper is also organized as follows. The next Sect. 2 discusses a few previous activities regarding heart disease prediction using machine learning models. Section 3 presents the workflow of this study, the data used, the preliminary processing of the data, and the algorithms used. Section 4 discusses the results obtained during this study and the paper ends up in Sect. 5.

2 Literature Review

As mentioned in the previous section, most of the earlier models for heart disease were designed for binary segregation of patients, so our state-of-the-art base has been reduced to a limited number of research activities for many cardiovascular categories. a disease in the health care sector with few binary separation functions.

Kirsi Varpa et al. [3] performed multiple classifications in Otoneurological disorder patients using KNN and SVM classification algorithms. These algorithms are used in their conventional methods as well as using the OVA method where the latter gives the best results with KNN. Anurag Kumar Verma et al. [4], in their paper, classified people suffering from skin diseases into six distinct categories (psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra). This classification is done using six machine learning algorithms and producing their own ensembles using bagging, Adaboost, and gradient boosting. It was noted that the use of ensembles to produce predictive models provides better results than each algorithm.

Hin Wai Lui et al. [5] merged the normal neural network with the convolutional neural network to form a multiclass classification of patients affected by Myocardial Infarction disease. The upgraded model was able to provide 97.2% accuracy with 92.4% sensitivity and can be plugged into portable devices. C. Beulah Christalin Latha et al. [6] performed two classifications of cardiovascular patients in the Cleveland heart disease database. They used six different algorithms: Bayes Net, Naïve Bayes, Random Forest, C4.5, Multilayer perceptron, and Projective Adaptive Resonance Theory (PART) to create their ensembles again. The combining methods used were bagging, voting, and packaging. It has been noted that the use of manufactured merged models offers much higher accuracy than individual weak algorithms, with a multi-voting system showing a higher accuracy with an increase of 7%. This accuracy is also enhanced by the use of the PCA process.

Abderrahmane El.daoudy et al. [1] created a cardiovascular prediction model using Apache Cassandra to store highly generated data and Spark MLlib to make predictions. This model, built using a random forest algorithm was able to handle real-time data and provided 87.5% accuracy and 86.67% sensitivity.

3 Method and Materials

This section describes the flow of work during this research, the dataset used, its analysis and preprocessing, and also explains the various algorithms used.

The steps presented in the flowchart in Fig. 2 are performed for the One-vs-all implementation of each of the machine learning algorithms used, which are, logistic regression, SVM, random forest, KNN, and Naïve Bayes such that the binary classifiers generated from each algorithm give their own accuracies whose mean is estimated to determine the accuracy of that algorithm.

3.1 Dataset Description

The dataset used for this research work is the Cleveland heart disease dataset available in the UCI machine learning repository. This dataset has the target variable which is a multivariate attribute and can take up values between 0 and 4 with class 0 indicating an absence of the disease, class 1 indicating a mild impact of the disease, class 2 indicating a moderate state, class 3 implies a slightly severe state, and class 4 indicates the most critical state.

The other predictor variables that are involved in the classification process are the patient's age, gender, type of chest pain, blood pressure while resting, cholesterol level in the blood, blood sugar level, maximum heartbeat rate, and a few others [7]. The 13 predictors along with the target variable are presented in Table 1.

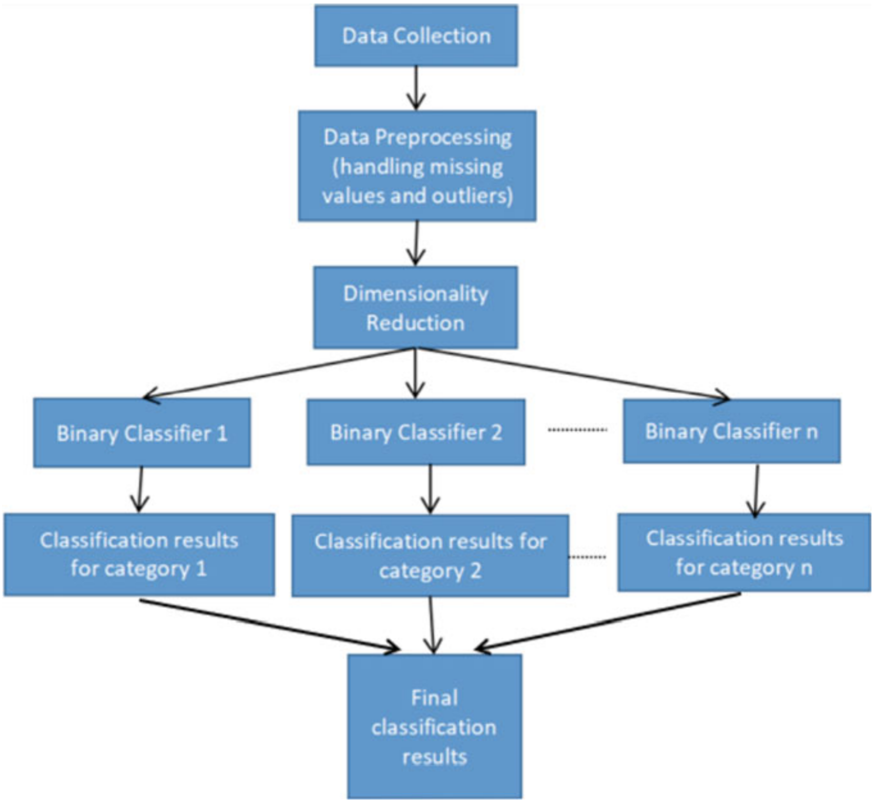

Data Collection
Data Preprocessing (handling missing values and outliers)
Dimensionality Reduction
Binary Classifier 1
Binary Classifier 2
Binary Classifier n
Classification results for category 1
Classification results for category 2
Classification results for category n
Final classification results

Fig. 2 Flow of work

Before applying the various machine learning algorithms to predict the target attribute values, the dataset needs to be preprocessed and analyzed properly to handle the outliers and missing values, and to identify the relevant patterns hidden in it.

In the dataset used, the Ca and Thal attributes contained missing values that we replaced with the values 0 and 3 which have the highest count in these attributes, respectively. The outliers are detected by plotting the boxplots of the attributes and it was observed that their presence did not affect the prediction accuracy adversely, hence remained undisturbed.

The following Fig. 3 shows the bar plot for the target attribute, Figs. 4 and 5 show the boxplot representation of the age and chol attributes, respectively.

It can be observed from Fig. 3 that the maximum number of people in the dataset (more than 150) belong to class 0, i.e., they are completely free from CVD and class 4 contains the minimum number of people. Figure 4 indicates that the people belonging to a higher age group are more prone to being in class 4, i.e., they are more likely to reach a critical state with a few outliers where people with an age less than 40 are in a critical state, i.e., they belong to class 4. Figure 5 indicates that the average

Table 1 Features of the Cleveland heart disease dataset

Feature	Description	Range
Age	Age of the patient	29–70
Sex	Gender of the patient	0-Female 1-Male
Cp	Type of chest pain	1-Typical Angina 2-Atypical Angina 3-Non-angina 4-Asymptomatic
Trestbps	Blood pressure of the patient while resting in mm Hg	94–200
Chol	Serum cholesterol in mg/dl	126–564
Fbs	Blood sugar level due to fasting in mg/dl	0, 1
Restecg	ECG result while resting	0, 1, 2
Thalach	Maximum rate of heartbeat	71–202
Exang	Angina induced while exercising	0, 1
Oldpeak	Relative ST depression while exercising and resting	1, 2, 3
Slope	Slope of peak ST segment	1, 2, 3
Ca	Number of major blood vessels that are colored by fluoroscopy	0, 1, 2, 3
Thal	Type of defect	3-Normal defect 6-Fixed defect 7-Reversible defect
Target	Level of impact of the disease on the patient/target class of the patient	0-absences of CVD 1-mild impact of CVD 2-moderate impact of CVD 3-slightly severe impact of CVD 4-highly critical state

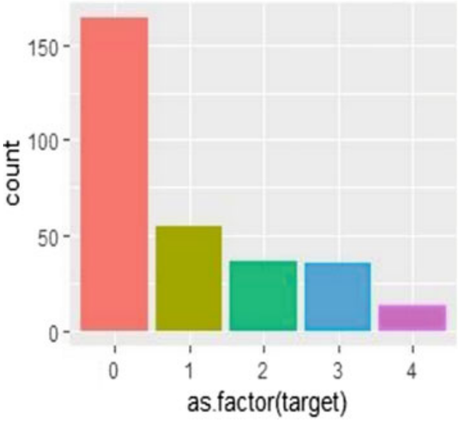


Fig. 3 Bar plot for target variable

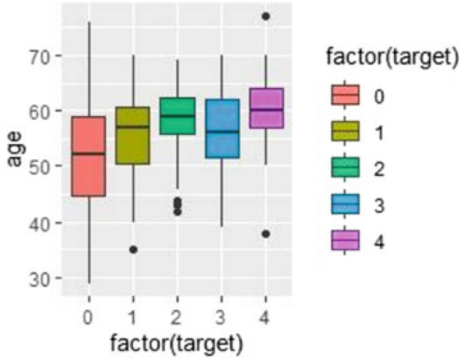


Fig. 4 Boxplot for age attribute

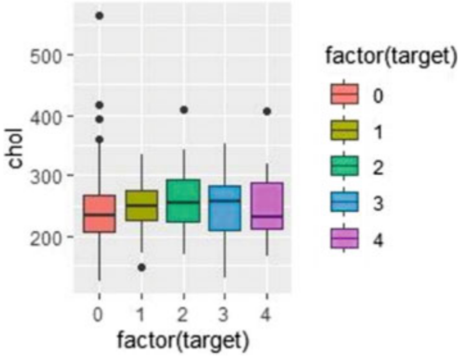


Fig. 5 Boxplot for cholesterol attribute

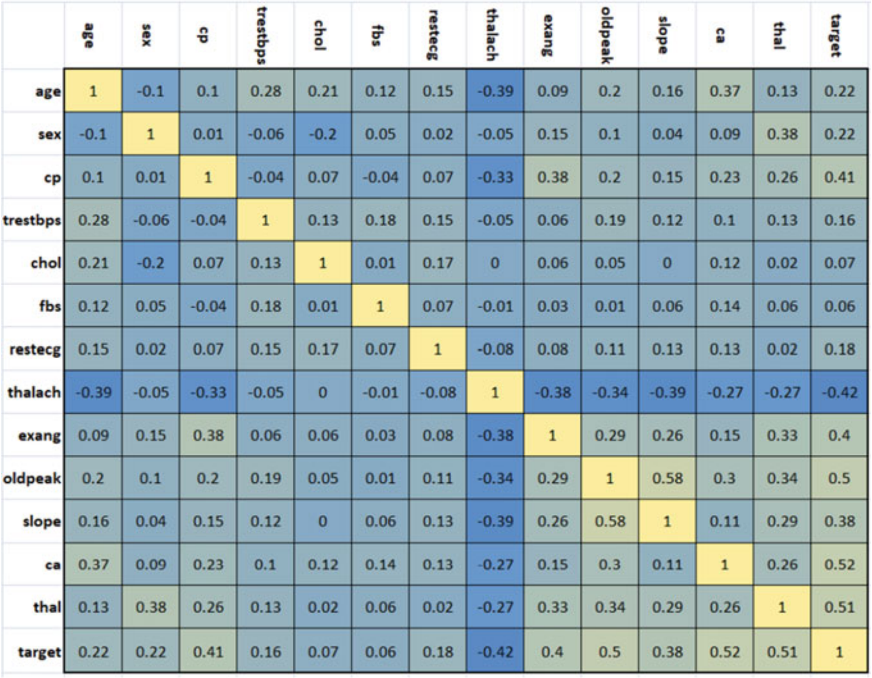
cholesterol of all the people falls in the range of 200–300. Also, it can be seen from Fig. 5 that few people having a cholesterol level close to 500 are completely free from heart disease.

Figure 6 represents the correlation matrix for the dataset used where every value less than 0 indicates a negative correlation, every value greater than 0 indicates a positive correlation and a 0 value indicates complete independence between the two associated attributes.

3.2 Algorithms Used

In this research work, we have used the following five algorithms to classify the patients and compared their performances with and without the One-vs-All approach.

Logistic Regression, a supervised classification algorithm, is typically used for binary classification and cannot perform multiclass classification in its ordinary form; however, in this research, we have used it for multiclass classification by implementing it with the One-vs-all approach. LR method can be used on datasets that are



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.1	0.1	0.28	0.21	0.12	0.15	-0.39	0.09	0.2	0.16	0.37	0.13	0.22
sex	-0.1	1	0.01	-0.06	-0.2	0.05	0.02	-0.05	0.15	0.1	0.04	0.09	0.38	0.22
cp	0.1	0.01	1	-0.04	0.07	-0.04	0.07	-0.33	0.38	0.2	0.15	0.23	0.26	0.41
trestbps	0.28	-0.06	-0.04	1	0.13	0.18	0.15	-0.05	0.06	0.19	0.12	0.1	0.13	0.16
chol	0.21	-0.2	0.07	0.13	1	0.01	0.17	0	0.06	0.05	0	0.12	0.02	0.07
fbs	0.12	0.05	-0.04	0.18	0.01	1	0.07	-0.01	0.03	0.01	0.06	0.14	0.06	0.06
restecg	0.15	0.02	0.07	0.15	0.17	0.07	1	-0.08	0.08	0.11	0.13	0.13	0.02	0.18
thalach	-0.39	-0.05	-0.33	-0.05	0	-0.01	-0.08	1	-0.38	-0.34	-0.39	-0.27	-0.27	-0.42
exang	0.09	0.15	0.38	0.06	0.06	0.03	0.08	-0.38	1	0.29	0.26	0.15	0.33	0.4
oldpeak	0.2	0.1	0.2	0.19	0.05	0.01	0.11	-0.34	0.29	1	0.58	0.3	0.34	0.5
slope	0.16	0.04	0.15	0.12	0	0.06	0.13	-0.39	0.26	0.58	1	0.11	0.29	0.38
ca	0.37	0.09	0.23	0.1	0.12	0.14	0.13	-0.27	0.15	0.3	0.11	1	0.26	0.52
thal	0.13	0.38	0.26	0.13	0.02	0.06	0.02	-0.27	0.33	0.34	0.29	0.26	1	0.51
target	0.22	0.22	0.41	0.16	0.07	0.06	0.18	-0.42	0.4	0.5	0.38	0.52	0.51	1

Fig. 6 Correlation matrix for cleveland heart disease dataset attributes

free from missing values. But the dataset used here has missing values in Ca and Thal attributes which we have handled by replacing the blanks with the value having the highest frequency in the respective attribute. The core functionality of this algorithm that is used to estimate the probability of a specific class being applicable on a data point is the sigmoidal function as shown in Eq. 1 [8].

$$y = 1/\left(1 + e^{-x}\right) \tag{1}$$

where ‘e’ is the Euler constant with the value 2.718, x is the linear combination of all the predictors, and y is the probability between 0 and 1 indicating the class to which the new tuple belongs by considering a threshold value between 0 and 1, which is 0.5 by default.

Support Vector Machine is another supervised algorithm, that is applicable for binary as well as multiclass classification. It attempts to generate a separating boundary, known as a hyperplane, depending on the dataset being used that separates the tuples belonging to vivid classes with the maximum margin [9]. Out of all the possible hyperplanes that segregate the data points, the hyperplane that provides the maximum margin is called the Maximal Margin Hyperplane (MMH). Figure 7

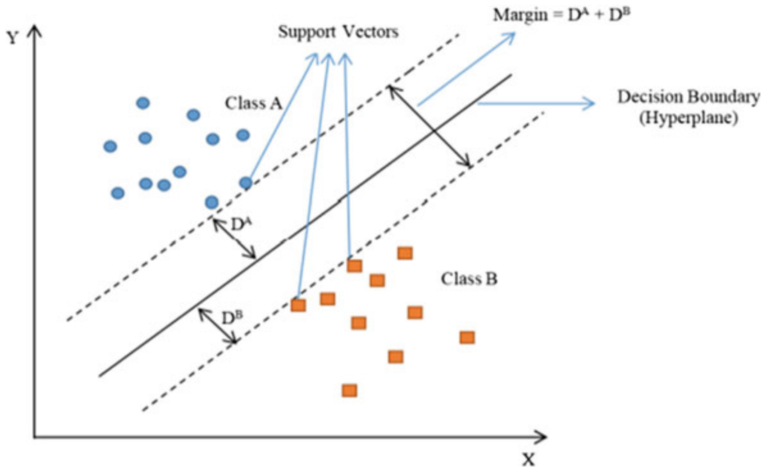
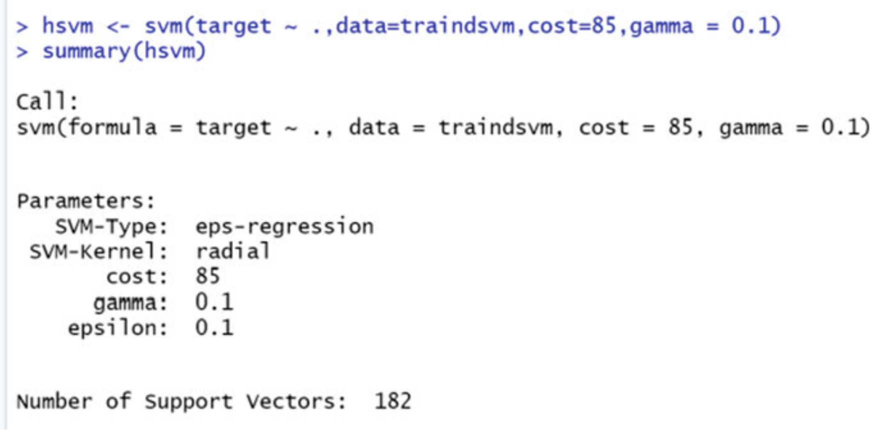


Fig. 7 Support vector machine architecture in a 2-d space

shows a 2-dimensional space containing data points of two different classes A and B separated by a hyperplane.

In the case of non-linearly separable data, that is when the hyperplane required to separate the data points is not a straight line, a kernel function is used that takes as input the low-dimensional feature space and converts it into a high-dimensional feature space and generates a curve or a plane as a hyperplane to separate the data elements belonging to distinct classes. While implementing SVM for our dataset, without the One-vs-all approach, 182 support vectors were obtained as shown in Fig. 8.



```
> hsvm <- svm(target ~ .,data=traindsvm,cost=85,gamma = 0.1)
> summary(hsvm)

Call:
svm(formula = target ~ ., data = traindsvm, cost = 85, gamma = 0.1)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  85
      gamma:  0.1
    epsilon:  0.1


Number of Support Vectors:  182
```

Fig. 8 SVM model built

Random Forest is a machine learning classifier based on the ensemble technique and uses the concept of decision trees in a randomized manner [10]. Each decision tree generated takes up its own set of tuples and attributes, known as a bootstrap, from the original dataset and comes up with its prediction result for the newly introduced data point. These individual prediction outputs are aggregated using the majority voting technique, i.e., the prediction value that is generated by the maximum number of trees as their output is considered to be the final prediction result of the random forest model [11].

Nearest Neighbors is another supervised machine learning algorithm used to solve classification and regression problems. To classify a newly introduced data point into one of the classes in the dataset, 'K' data points that are closest to the new data point are identified. These 'K' nearest data points are detected by measuring the distance of all the existing data points from the new data point. The class or category to which majority of these 'K' neighbors belong is the desired class to which the new data point should be assigned. The distance between the new data point and existing data points can be calculated using several measures [10]. Few commonly used techniques are given as follows.

Euclidean distance given by

$$d = \left(\sum \left(x_i - y_i\right)^2\right)^{1/2} \tag{2}$$

Manhattan distance given by

$$d = \sum \left(x_i - y_i\right) \tag{3}$$

Minkowski distance given by

$$d = \left(\sum \left(x_i - y_i\right) s^r\right)^{1/r} \tag{4}$$

The error in prediction varies with the value chosen for the variable 'K'. In our research, we iterated over 1 to 10 to choose the value of 'K' which will give the minimum error rate. The following Figs. 9, 10 and 11 show the error rate versus K plots for the KNN algorithm implemented in its ordinary form, with One-vs-all, and with OVA combined with the principal component analysis technique.

It can be observed from the above figures that the error rate is minimum for $K = 10$ for ordinary KNN implementation, reaches a minimum at $K = 2$, and remains same till $K = 10$ when implemented with OVA, and the error rate becomes 0 for $K = 1$ to 10 when KNN is implemented with OVA combined with PCA.

Naïve Bayes is also a supervised learning algorithm that is also a probabilistic classifier. This algorithm assumes complete independence among the features of the dataset, i.e., occurrence of one attribute does not depend on the occurrence of any other attribute [10]. It works on the principle of the Bayes theorem which is given as follows:

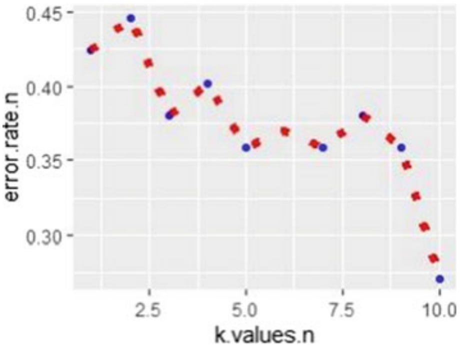


Fig. 9 Ordinary KNN

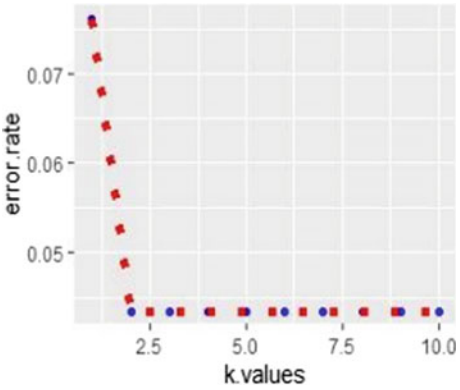


Fig. 10 One-vs-all

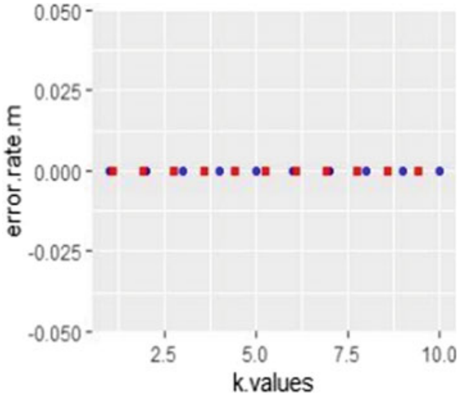


Fig. 11 One-vs-all & PCA

$$P(A|B) = (P(B|A) * P(A))/P(B) \tag{5}$$

where P(A|B) is the probability of hypothesis A when event B has occurred, P(B|A) is the likelihood of occurrence of an event given that A is true, P(A) denotes the prior probability of the hypothesis before the event has occurred, P(B) denotes the evidence, i.e., probability of occurrence of event B.

One-vs-All is a technique used to implement machine learning algorithms to perform multiclass classification much more efficiently compared to their performance without it. In this approach, 'n' binary classifiers are built based on a chosen algorithm instead of a single multiclass classifier, where 'n' is the number of classes in the dataset. Each binary classifier is committed toward a single class, i.e., each binary classifier gives the accuracy in prediction for a single class by considering that class as class 1 and all other classes as class 0. The accuracies obtained in making predictions for each class as provided by their associated binary classifiers are averaged to find the overall accuracy of that algorithm. Figure 12 shows the architecture of the One-vs-all technique.

Principal Component Analysis is an unsupervised machine learning algorithm that is used to reduce the dimensionality of the dataset thus allowing the model to predict the target variable values for a reduced dataset thus reducing the chances of overfitting [10]. PCA tries to find attributes, known as Principal Components, that provide maximum variance in the higher dimensional space and project those onto a smaller dimensional space retaining only the relevant information. Using these principal components to make predictions not only reduces the burden on the models in making predictions but also improves the accuracy of the prediction by the model.

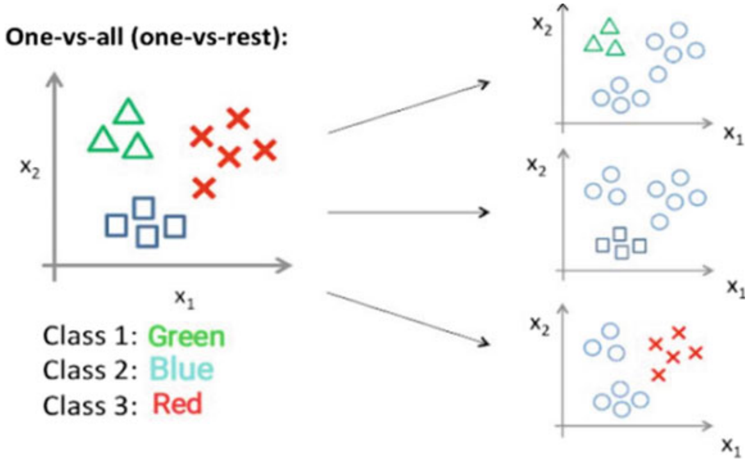


Fig. 12 Schematic diagram of one-vs-all technique

4 Experimental Results

The research outcome shows a comparison among the heart disease prediction accuracies provided by the five algorithms: LR, SVM, RF, KNN, and NB when implemented on the Cleveland heart disease dataset. These algorithms are implemented in three ways: in their ordinary form, with the One-vs-all approach, and One-vs-all and principal component analysis combined. The accuracies obtained in each case are compared and it is observed that using the OVA approach significantly increases the accuracy in prediction for each algorithm. Also, using the OVA technique on the dataset reduced by PCA further enhances the performance of the models.

The accuracy of each algorithm is computed using the confusion matrices generated. Confusion matrix refers to an nxn matrix where n is the number of distinct classes available in the dataset. In a confusion matrix, each column sum indicates the number of people that actually belong to that class and each row sum indicates the number of people who have been categorized into that class. Hence, it can be concluded that the diagonal elements of the confusion matrix indicate the number of patients who have been correctly classified. The Eq. 6 is used to compute the accuracy in classification from the confusion matrix generated.

$$\text{Accuracy } = \left\{\sum(\text{diagonal elements})/\sum(\text{all elements})\right\} * 100 \tag{6}$$

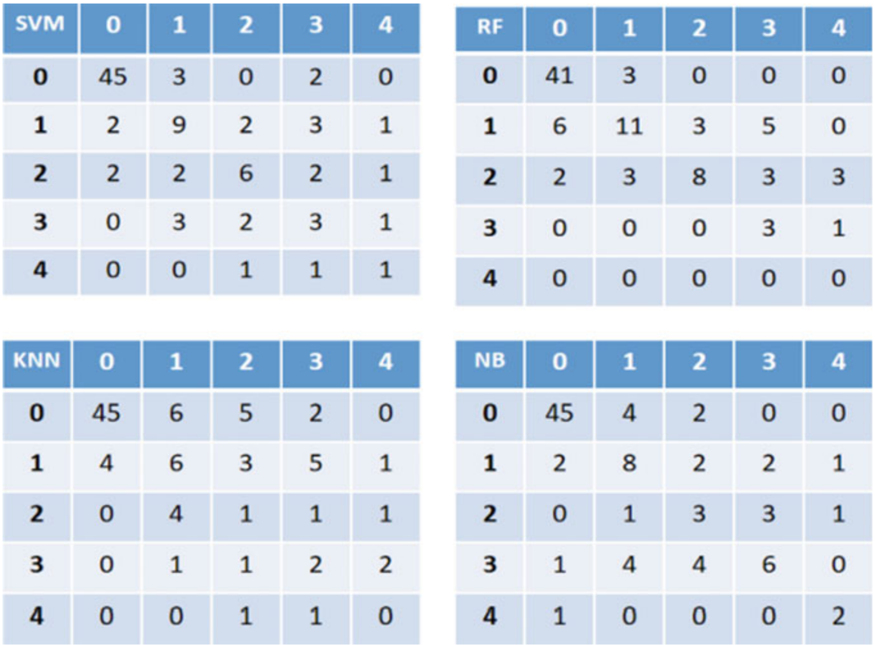
The following Figs. 13 and 14 depict the generated confusion matrices by all the above-mentioned algorithms without OVA and with OVA, respectively. In the OVA approach, as already mentioned earlier, instead of a single multiclass classifier, 5 binary classifiers are built for each algorithm that generates their own 2x2 confusion matrices. The accuracies of all 5 confusion matrices are averaged to compute the overall accuracy provided by that algorithm when implemented with the OVA approach.

The confusion matrices generated by the application of OVA approach on the dataset reduced by PCA have slightly better values than the confusion matrices without PCA, thus providing slightly better accuracy.

The following Table 2 states the accuracy provided by each of the algorithms that are implemented without OVA method, with the OVA approach, and OVA on the PCA reduced dataset.

Figure 15 shows the bar graph representation of the accuracy values acquired by all the algorithms for the three types of implementation. It can be observed from Table 2 and Fig. 15 that no accuracy value exists for the ordinary implementation of logistic regression as it is a binary classification algorithm and can perform multiclass classification only with the One-vs-all approach.

Also, it can be seen that the KNN algorithm has outperformed all other algorithms by providing the highest accuracy of 99.56%.



SVM	0	1	2	3	4
0	45	3	0	2	0
1	2	9	2	3	1
2	2	2	6	2	1
3	0	3	2	3	1
4	0	0	1	1	1

RF	0	1	2	3	4
0	41	3	0	0	0
1	6	11	3	5	0
2	2	3	8	3	3
3	0	0	0	3	1
4	0	0	0	0	0

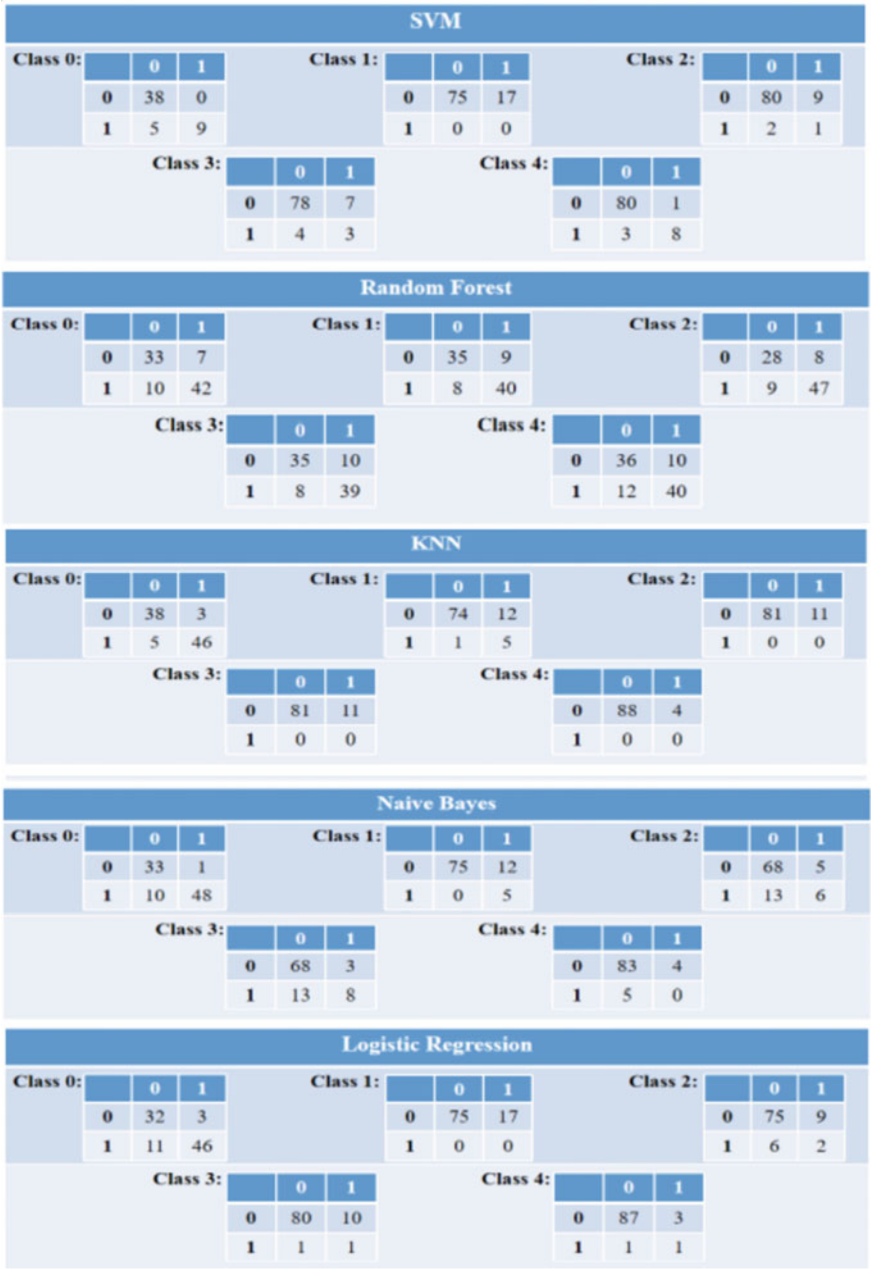
KNN	0	1	2	3	4
0	45	6	5	2	0
1	4	6	3	5	1
2	0	4	1	1	1
3	0	1	1	2	2
4	0	0	1	1	0

NB	0	1	2	3	4
0	45	4	2	0	0
1	2	8	2	2	1
2	0	1	3	3	1
3	1	4	4	6	0
4	1	0	0	0	2

Fig. 13 Confusion matrices without one-vs-all

5 Conclusion

This paper has emphasized the classification of CVD patients into more than two classes that will be more helpful for the physicians in providing the best possible treatment to their patients with more precision instead of simply discovering the sign of any cardiac problem in them. We have recommended the exploitation of this One-vs-all method with different machine learning algorithms to perform a multiclass classification of the patients. The obtained accuracy is further improved by implementing these algorithms with OVA on the dataset reduced by PCA. Out of the five algorithms implemented during this research, KNN has shown the best performance with an accuracy of 99.56% with a combination of PCA and One-vs-All techniques.



SVM

Class 0:

	0	1
0	38	0
1	5	9

Class 1:

	0	1
0	75	17
1	0	0

Class 2:

	0	1
0	80	9
1	2	1

Class 3:

	0	1
0	78	7
1	4	3

Class 4:

	0	1
0	80	1
1	3	8

Random Forest

Class 0:

	0	1
0	33	7
1	10	42

Class 1:

	0	1
0	35	9
1	8	40

Class 2:

	0	1
0	28	8
1	9	47

Class 3:

	0	1
0	35	10
1	8	39

Class 4:

	0	1
0	36	10
1	12	40

KNN

Class 0:

	0	1
0	38	3
1	5	46

Class 1:

	0	1
0	74	12
1	1	5

Class 2:

	0	1
0	81	11
1	0	0

Class 3:

	0	1
0	81	11
1	0	0

Class 4:

	0	1
0	88	4
1	0	0

Naive Bayes

Class 0:

	0	1
0	33	1
1	10	48

Class 1:

	0	1
0	75	12
1	0	5

Class 2:

	0	1
0	68	5
1	13	6

Class 3:

	0	1
0	68	3
1	13	8

Class 4:

	0	1
0	83	4
1	5	0

Logistic Regression

Class 0:

	0	1
0	32	3
1	11	46

Class 1:

	0	1
0	75	17
1	0	0

Class 2:

	0	1
0	75	9
1	6	2

Class 3:

	0	1
0	80	10
1	1	1

Class 4:

	0	1
0	87	3
1	1	1

Fig. 14 Confusion matrices with one-vs-All

Table 2 Accuracies obtained by each algorithm in all 3 cases

Algorithm	Accuracy without OVA (%)	Accuracy with OVA (%)	Accuracy with OVA and PCA (%)
SVM	69.57	89.57	89.57
Random forest	68.48	89.13	92.53
KNN	7.83	89.78	99.56
Naïve Bayes	69.57	85.65	85.71
Logistic		86.74	93.41

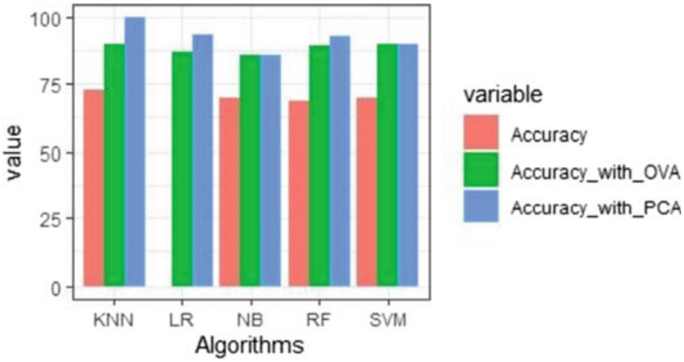


Fig. 15 Bar plot for accuracies obtained

References

1. Ed-Daoudi A, Maalmi, K.: Real-time machine learning for early detection of heart disease using big data approach. In: 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1–5. https://doi.org/10.1109/WITS.2019.8723839
2. Baitharu, T.R., Pani, S.K: Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. In: International Conference On Computational Modelling And Security (CMS 2016). Procedia Compuer Science 85(2016), pp. 862–870
3. Varpa, K., Joutsijoki, H., Iltanen, K., Juhola, M.: Applying one-vs-one and one-vs-all classifiers in k-nearest neighbors method and support vector machines to an otoneurological multi-class problem. In: Article in Studies in Health Technology and Informatics-January 2011. 10.3233|978-1-60750-806-9-579
4. Verma, A.K., Pal, S., Kumar, S.: Comparison of Skin Disease Prediction by Feature Selection Using Ensemble Data Mining Techniques. Informatics In Medicine Unlocked, vol. 16, p. 100202 (2019). 10/1016/j.imu.2019.1002.02
5. Lui, H.W., Chow, K.L.: Multiclass classification of myocardial infarction with convolutional and recurrent neural networks for portable ECG devices. Inform. Med. Unlocked **13**, 26–33 (2018)
6. Latha, C.B.C., Jeeva, S.C.: Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked **16**, 100203. https://doi.org/10.1016/j.imu.2019.100203

7. Saxena, K., Sharma, R.: Efficient heart disease prediction system. Procedia Comput. Sci. **85**, 962–969 (2016). https://doi.org/10.1016/j.procs.2016.05.288
8. Bagley, S.C., White, H., Golomb, B.A.: Logistic regression in the medical literature::standards for use and reporting with particular attention to medical domain **54**(10), 979–985 (2001). https://doi.org/10.1016/s0895-4356(01)00372-9
9. Mishra, S., Pandey, M., Rautaray, S.S., Gourisaria, M.K.: A survey on big data analytical tools & techniques in healthcare sector. Int. J. Emerg. Technol. **11**(3), 554–560
10. Lebedev, A.V., Wesman, E., Van Westen, G.J.P.: Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness **6**, 115-125 (2014). https://doi.org/10.1016/j.nicl.2014.08.023
11. Kamran, M., Javed, A.: A survey of recommender systems and their application in healthcare. In: Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan, vol. 20 No. IV-2015
12. Alarsan, F.I., Younes, M.: Analysis and classifcation of heart diseases using heartbeat features and machine learning algorithms. Alarsan and Younes J Big Data **6**, 81 (2019). https://doi.org/10.1186/s40537-019-0244-x
13. Salma Banu, N.K, Swamy, S.: Prediction of heart disease at early stage using data mining and big data analytics: a survey. In: 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEEECCOT). 978-1-5090-4697-3/16/
14. Saboji, R.G., Ramesh, P.K.: A scalable solution for heart disease prediction using classification mining technique. In: International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017). 978-1-5386-1887-5/17/
15. Karthick, D., Priyadharshini, B.: Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using Classification Techniques within fifty years of age. In: Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018) IEEE Xplore Compliant - Part Number:CFP18J06-ART. ISBN:978-1-5386-0807-4; DVD Part Number:CFP18J06DVD, ISBN:978-1-5386-0806-7