DIAGNOSIS OF CARDIAC PROBLEM USING ROUGH SET THEORY AND MACHINE LEARNING

Subhalaxmi Das

Research Scholar, Dept. of Computer Science and Applications, Utkal University,
Bhubaneswar, Odisha, India
subhalaxmi.das@gmail.com

Sateesh Kumar Pradhan

Former Prof., Dept. of Computer Science and Applications, Utkal University,
Bhubaneswar, Odisha, India
sateesh1960@gmail.com

Sujogya Mishra

Asst. Prof., Department of Mathematics, Odisha University of Technology and Research,
Bhubaneswar, Odisha, India
sujogya123@gmail.com

Sipali Pradhan

Asst. Prof., Department of Computer Science RBVRR Women's College,
Hyderabad, Telangana, India
sipalipradhan06@gmail.com

P. K. Pattnaik

Asst. Prof., Department of Mathematics, Odisha University of Technology and Research,
Bhubaneswar, Odisha, India
papun.pattnaik@gmail.com

Abstract

Cardiac-related problems are responsible for the sudden increase in death rate globally. A quick and accurate diagnosis of the cardiac-related problem is essential to avoid serious consequences. Among the traditional medical methods, angiography is one of the well-known methods to deal with heart problems, but it has several drawbacks. On the contrary, the non-conventional methods, like intelligent learning based on soft computing methods are upright and effective in countering cardiac-related problems. This proposed work is divided into two different parts. Initially, we have used Rough Set theory (RST) and the Correlation method (CM) to find the significant cardiac-related disease i.e., cardiac arrest. Next, we have designed a model for heart disease prediction using machine learning techniques and the stacking method. The research is carried out to diagnose cardiac problems using the Cleveland dataset. Finally, the performance of the model was evaluated in terms of accuracy, precision, recall, and F1 score, and different error functions were also calculated.

Keywords: RST; Soft Computing; Cardiac-related Problems; Angiography; Correlation Methods.

1. Introduction

Cardiac-related problems are very common medical anomalies in recent times because of several intrinsic factors, such as fluctuation in blood pressure, high blood sugar, variation in the cholesterol level, overtiredness, and several factors related to breathing. Quick detection of such diseases is essential for doctors to counter these diseases. Several tools have been developed to help health care providers to detect approximately several early symptoms of cardiac problems. Several tests can be applied on the active patients to take the additional safeguards to decrease the result of having such a disease as discussed by Khourdifi and Bahaj [1], they develop a dependable technique to forecast initial stages of cardiac problems, techniques discussed in their paper, can be vital for saving several lives. Several Machine Learning (ML) algorithms, which include Naïve Bayes, Stochastic Gradient Descents

(SGD), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Adaboost, JRip, Decision tree J48, and several others, were implemented to determine classification and forecast of the cardiac-related dataset, and several interesting outcomes were presented by Mohan *et al.* [2]. Because of the intricate nature of the cardiac problems, suggested tests have to be divided, as suggested by Dai *et al.* [3], and proposed several techniques to precisely and proficiently forecast cardiac-related problems concerning patient-explicit medical past. Durairaj and Revathi [4] developed an algorithm to forecast the presence of cardiac problems using Back Propagation MLP (Multilayer Perceptron), and Gavhane [5] proposed a machine learning technique to develop a method that can forecast the susceptibility of cardiac problems given fundamental indications which includes age, sex, pulse rate, and neural networks which exhibited the precise and dependable algorithm for the projected system. Abdullah [6] discussed a Random Forest classifier technique to further investigate several events concerning cardiac problems. A fusion technique for cardiac-problems forecasting, as discussed by Shehani and Rathnayakc [7], includes risk factors of various cardiac problems, and they classify the risk level of a person using several machine learning algorithms. Kelwade [8] had proposed the prediction time for the cardiac problem by the use of several cardiac-related datasets, and Anju [9] had discussed cardiac-related risk levels using a set of fuzzy rules. Krishnaiah and Chandra [10] predict cardiac-related problems using advanced fuzzy logic. Fatima and Pasha [11] discussed the prediction of various diseases using several machine learning algorithms.

Otoom *et al.* [12] had developed a technique capable of effectively detecting cardiac-related problems. Vembandasamy and Sasipriya [13] developed a model by using naive Bayesian classification to efficiently predict cardiac-related problems. Malav *et al.* [14] discussed cardiac-related problems by fusing K-means and artificial neural network models. Lee *et al.* [15] and Tarle [16] discussed the prediction of cardiac problems using several machine learning algorithms. Saxena and Purushottam [17] had developed a technique to efficiently predict cardiac arrest. Almustafa [18] discussed several classifiers and their sensitivity for predicting cardiac-related anomalies. Karaylan and Kilic [19] had used Neural networks to efficiently predict cardiac problems. Esfahan and Ghazanfari [20] used an ensemble classifier to predict cardiovascular disease by using UCI Laboratory data. RST and the theory of uncertainty were proposed by Pawlak [21] and then it was modified by Pawlak and his co-researchers [22-24]. Yar Muhammad *et al.* [25] developed a technique using an intelligent computer system for the early detection of heart diseases. Bui *et al.* [26] discussed the risk profile and epidemiology of cardiac failure. Alizadehsani *et al.* [27] developed a unique method for diagnosing coronary artery disease using a data mining approach. Vanisree & Singaraju [28] discussed genetic cardiac disease diagnosis using neural networks. Nazir *et al.* [29] had developed a technique that includes m-fuzzy logic-based decision support system for component security evaluation. Gudadhe *et al.* [30] had proposed a model to predict cardiac diseases using SVM. Palaniappan and Awang [31] had proposed a model for accurate and early detection of cardiac diseases using an intelligent computational model. Olaniyi *et al.* [32] had proposed a model using neural network arbitration for the early detection of cardiac arrest. Das *et al.* [33] proposed a model for detecting heart diseases using a neural-network export system. Tomov & Tomov [34] developed a technique using a deep neural network to forecast cardiac arrest. Mohan *et al.* [35] had designed a tool with the help of a hybrid machine-learning algorithm to predict heart diseases. Mishra *et al.* [36] had proposed a set of rules for social science problems using RST. Das *et al.* [37] proposed a model to analyze heart diseases using a soft computing technique. Nayak *et al.* [38] had developed a model for symptom prediction of malaria using RST. Mishra *et al.* [39] had developed a technique using soft computing to predict the symptoms of COVID-19**.** Peng *et al.* [40] had discussed feature selection using mutual information criteria.

The above literature provided a background to build a model that can be used to classify cardiac-related problems, which can be helpful to physicians and patients.

2. Proposed Model

The proposed model, as shown in Fig. 1, integrates many phases like cardiac-related disease data collection, initial approximation using the concept of rough set theory, finding of most significant disease, among others, pre-processing of data, then using machine learning classifiers and stacking methods to classify whether the cardiac-related problem is present or not and finally the overall performance evaluation of the model.

Initially, we collected cardiac-related data from different places in Odisha. Then rough set theory was applied to the collected dataset to find the most significant disease, including upper approximations, lower approximations, and indiscernibility. Then using the concept of reduct and core from the indiscernibility, the most significant disease among all cardiac-related diseases i.e., cardiac arrest found. Further, the complete research is focused on cardiac arrest disease. We have taken Cleveland Dataset, a commonly used dataset by many researchers, for our work. The machine learning techniques were applied to the dataset to predict the disease, and the performances of different techniques were also evaluated.

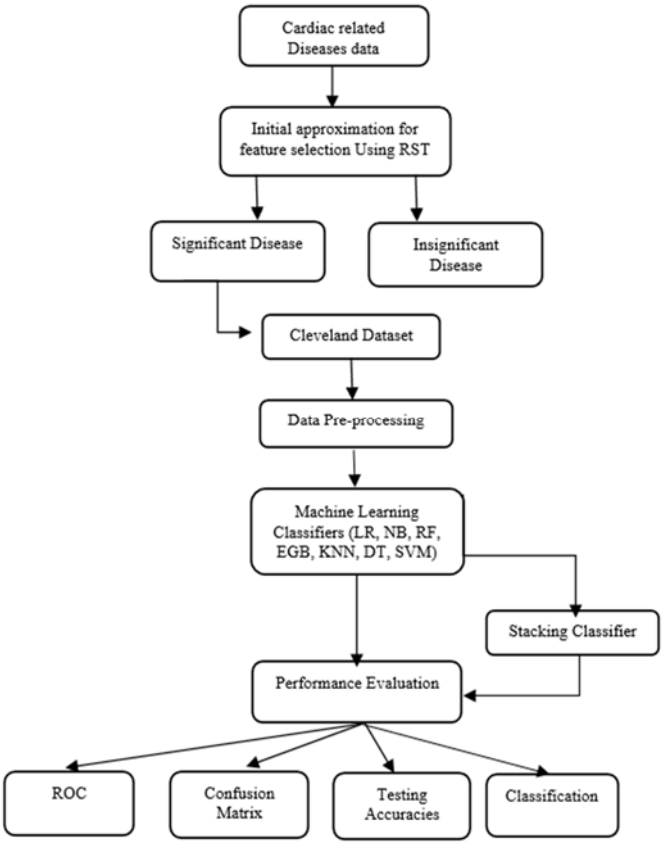

Fig. 1. Proposed Model

For better accuracy, we have used the stacking method, which is slightly different from the above approach. Further, classifiers are chosen based on their performance, and one of them was selected as a meta classifier for the stacking method. Finally, the proposed method's performance is accomplished by standard techniques like confusion matrix, classification report, receiver operating characteristic (ROC) curve, and accuracy.

2.1 Mathematical Background

The entire study deals with two basic concepts i.e., RST and Machine Learning techniques (ML). Pawlak developed RST in the early 80's. The basics of RST include upper approximations, lower approximations, and indiscernibility, further reduct and core of RST derived from the concept of indiscernibility. A detailed definition of the basics of RST was given in the subsequent section.

2.1.1. Basics of RST

- Decision table- The Decision table is the backbone of RST. The decision table consists of three tuples <a,b,c>, a is the set of records or instances is the conditional attributes, and c is the set of decision attributes as shown in Table 1.
- Upper Approximations: $L^1K= \cup\{M \in P/L$, where $M \cap K \neq \varphi\}$ and P is the universal set.
- Lower Approximations: $L^2K= \cup\{M \in P/L$, where M is the subset of K $\}$ where P is the universal set.

Records	Conditional Attributes	Decision attributes
E1	1	P
E2	2	Q
E3	3	Q
E4	4	Q
E5	5	p

Table 1. Information on RST

- Boundary region: The difference between upper and lower approximation.
- Indiscernibility: Indiscernibility in RST is the backbone of RST; for example, the two or more combinations of conditional attributes produce indiscernibility, dropping any attribute(conditional) that affects the decision.
- Reduct: It is the minimal indiscernibility set that leads to a significant result.
- Core: Core is the set of attributes, which is common to all reducts.

2.2. Support vector machine (SVM)

Support vector machine used for classification and regression, in this work we have used SVM for classification. The initial approach of SVMs is to search for a line or hyperplane among data of binary groups. The SVM approach includes input as a dataset and outputs as a hyperplane that divides the dataset into groups if possible. SVM falls under the category of supervised learning, and we have used SVM's Classification Algorithm for our research. f(y): $K^T \rightarrow K^P$, $f(y) \in K^P$, K^P is the feature space, and K^T is the input space, this changed feature space from input feature measured to a changed basis vector f(y). The hyperplane equation is given by $R^T(y)+b$, and our objective is to maximize the minimum distance of the feature points from the hyperplane.

2.3. Data for further studies

We have collected data from various government and non-government hospitals on cardiac-related diseases. The data in the form of discrete numeric values are given in Table 2. For better understanding, we are renaming the conditional attributes <Pneumonia, Asthma, Cardiac Arrest, Atelectasis, Pulmonary edema as <m, n, o, p, q> and their values significant and insignificant as <1,2> decision variable d as notable and pointless renamed as a & b. Using the data from above Table 2, we had six records by studying the scatter plot. From Fig. 2, we have considered 6-records as 6-cluster heads given in the following Table-3.

District	Pneumonia	Asthma	Cardiac Arrest	Atelectasis	Pulmonary edema	Total
Sambalpur	5000	15000	20000	10000	15000	65000
Koraput	15000	15000	10000	5000	5000	50000
Ganjam	15000	10000	5000	15000	5000	50000
Nayagarh	5000	15000	5000	5000	5000	35000
Kandhamal	2500	2500	15000	5000	15000	40000

Table 2. Data Table

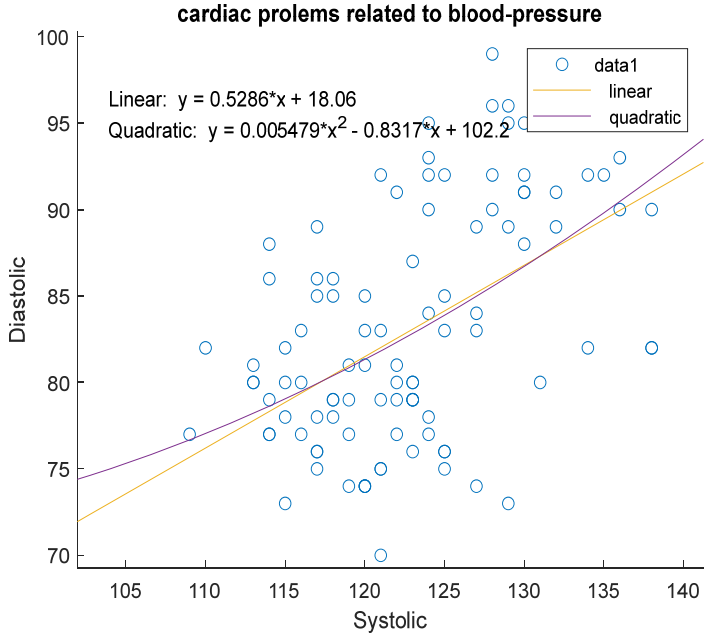

Fig. 2. Cardiac problem related to blood pressure

S	m	n	o	p	q	d
S1	1	1	2	2	1	A
S2	1	1	1	2	1	B
S3	2	2	1	1	1	B
S4	2	2	2	1	1	B
S5	2	2	2	1	2	B
S6	1	2	1	1	2	B

Table 3. Initial information table

This information table may contain many data values having the same features which can be reduced by a representative object named an indiscernible object, for every set of objects with the same features. Further reduct and core are two important concepts, of RTS that are used to reduce unnecessary attributes in the information table. The algorithms to find indiscernibility and reduct are presented in Fig.s 3 and 4 respectively.

Algorithm for Indiscernibility

1. *for i =n down to 1*
2. *check the matching values*
3. *if no matching values are found*
4. *Set name as the unique or distinct value set*
5. *Else*
6. *go to 1*
7. *end if*
8. *end for*

Fig. 3. An algorithm for finding Indiscernibility

Algorithm to find Reduct

1. *Find discernibility sets*
2. *If the discernibility sets have distinct value*
3. *Goto step-6*
4. *Else*
5. *Continue with steps 1&2*
6. *Reduct set found*
7. *Stop*

Fig. 4. An algorithm for finding Reduct

The following sets of Indiscernibility using the above algorithm are shown below.
I(m) = {{1,2,6}, {3,4,5}}, I(n) = {{1,2}, {3,4,5,6}}, I(o) = {{1,4,5}, {2,3,6}}, I(p) = {{1,2}, {3,4,5,6}},
I(q) = {{1,2,3,4}, {5,6}}, I (m, n) = {{1,2}, {3,4,5},6}, I (m, o) = {{1}, {2,6},{3}, {4,5}},
I (m, p) = {{1,2}, {2,6},{3}, {4,5}}, I (m, q) = {{1,2}, {3,4},{5},{6}}, I (n, o) = {{1},{2}, {3,6}, {4,5}},
I (n, p) = {{1,2}, {3,4,5,6}} I (n, q) = {{1,2}, {3,4}, {5,6}}, I (o, p) = {{1},{2}, {3,6}, {4,5}},
I (p, q) = {{1,2}, {3,4}, {5,6}}, I (m, n, o) = {{1},{2},{3}, {4,5},{6}}, I (m, o, p) = {{1},{2},{3}, {4,5},{6}},
I (m, p, q) = {{1,2}, {3,4},{5},{6}}, I (m, n, p) = {{1,2}, {3,4,5},{6}}, I (m, n, q) = {{1,2},{3}, {4,5},{6}},
I (m, n, q) = {{1},{2},{3}, {4,5},{6}}, I (m, p, q) = {{1,2}, {3,4},{5},{6}}, I (n, o, p) = {{1},{2},{3}, {4,5},{6}},
I (n, o, q) = {{1}{2},{3},{4},{5},{6}}, I (n, p, q) = {{1,2}, {3,4}, {5,6}},
I (m, n, o, p) = {{1}{2},{3},{4},{5},{6}}, I (m, n, o, q) = {{1}{2},{3},{4},{5},{6}},
I (m, o, p, q) = {{1}{2},{3},{4},{5},{6}}, I (n, o, p, q) = {{1}{2},{3},{4},{5},{6}}

From the above analysis, we found the following set of reducts as it produced distinct values.

I (m, n, o, p) = {{1}{2},{3},{4},{5},{6}}, I (m, n, o, q) = {{1}{2},{3},{4},{5},{6}},
I (m, o, p, q) = {{1}{2},{3},{4},{5},{6}}, I (n, o, p, q) = {{1}{2},{3},{4},{5},{6}},
I (n, o, q) = {{1}{2},{3},{4},{5},{6}}

Now for finding the core, we have the following result.

Core=⋂Reduct =⋂{(m, n, o, p), (m, n, o, q), (m, o, p, q), (n, o, p, q), (n, o, q)} = o
The symbol o signify for Cardiac Arrest, in the subsequent section we have discussed the prediction of cardiac-arrest symptoms using SVM.

2.4. Prediction of symptoms using SVM

In this study, we are using for classification of symptoms using SVM as discussed by Tabesh *et al.* [41]. To find the distance between test data and the hyperplane, we are using the fundamental concept of SVM for our purpose given in the following equation as follows

$$\text{Max } T(\beta) = \sum_{j=0}^{n} \beta_j - 0.6 \sum_{j=0}^{n} \beta_i z_i z_j f(y_i, y_j) \quad (1)$$

this idea was derived from langrage's multiplier concept Subject to $\sum_{j=1}^{n} z_i \beta_i = 0$ (2)

where $y_i \in R^n$ is the input vectors i=1 to n $0 \leq \beta_i \leq M$ y_i's are the feature derived from the test data given in Table 2. In general, z_i's are associated with binary values -1 to +1, which are the output or the result derived from the data set. The variable β_j is the multiplier used for the dual construction, and M is a user-defined variable for the misclassification penalty of (y_i, y_j) is the function whose domain is the original dataset, and its range is another dataset.

The kernel function, $f(y_i, y_j) = e^{-\frac{\|y_{i-y_j}\|}{2\sigma}}$ (3) where σ is the width of the kernel. In this study, we have used Gaussian kernel.

2.4.1. Classification based on group properties

There is a certain dataset that contains both group values and continuous values in their properties. Group structures may play diverse roles inaccurate classification. In this approach, we divided the dataset based on the present values of the group properties. Every group property has at least two options for the dataset. We have developed several subsets equal to the total number of exiting values for all group properties. The proposed algorithm was applied to the data set of cardiac arrest [41] as shown in Fig. 5.

Proposed Algorithm

r= index for the group property. r = 1, 2,…, l where l is the number of group properties)
t(r) = index for value present in the group properties r where t(r)= 1, 2, …, m (m is a numeric value considering every possible group property)
n = catalogue for data occurrences. n = 1, 2, …, p (p is the length or total number of occurrences present according to the dataset)
t_{rn}= result of group property r for n^{th} occurrence of data
t_{rt} = t^{th} result of group property r
S_{rt} =subset of occurrences that have t^{th} value for group properties r
for r=1 to l
for t(r) 1 to m
t← t(r)
for n= 1 to p
If $t_{rn} = S_{rt}$ then
Corresponding subset S_{rt} added with the n^{th} row
end for n
end for t(r)
end for r

Fig. 5. An algorithm to find symptoms of cardiac arrest

2.4.2. Numerical Result Analysis

The machine learning source consists of 100 suitable datasets from our state Odisha. Out of these datasets, the cardiac-related dataset includes four types of cardiac arrest. All attributes have represented by numbers. The data were collected from five different places: Sambalpur, Koraput, Ganjam, Nayagarh, and Kandhamal database,

which includes 303 observations, out of which 287 are complete information and 16 observations don't have available information. Initially, there are 86 raw attributes, out of which only 14 attributes are being used for the purpose, with the last one as a goal or output attribute.

2.4.3 Experimental Results

In the initial phase we assign the values 0 for no heart diseases, a- result for heart diseases type-1, b- heart diseases type-2, c- heat diseases type-3, d- heart diseases type-4. Separating type-1, type-2, type-3 and type-4 from type-0 has an accuracy level of 72%.

2.4.4 Experimental Results for various subsets

The data set of Table 2 consists of both continuous and discrete values; the result's accuracy for discrete cases is approximately 83.5%, and for continuous cases, the result has an accuracy of about 60%.

3. Further Classification

In this work, we have used two different approaches to improve the accuracy of classifiers. The first approach was based on combined classifiers and the second one is a two-level stacking approach. In the combined classifier, we have chosen those classifiers that are showing better accuracy than others. So XGB and SVM were selected for combination in case of 80:20 training and testing. Similarly, KNN and SVM were selected for combination in the case of 70:30 training and testing. The accuracy of all classifier's details was given in Tables 8 and 9. As combined classifiers approaches were not showing markable improvement over the Stacking method as shown in Table 4, so further, we have used a two-level stacking-based model to improve the accuracy with different combinations of classifiers for the classification of cardiac disease. The detailed stacking approach in terms of accuracy, precision, recall, and F1 score was demonstrated in Table 8 and 9.

Training:Testing	Classifiers	Combined Classifier Accuracy	Stacking Accuracy
80:20	XGB+SVM	83.6	90.16
70:30	KNN+SVM	80.21	82.41

Table 4. Accuracy of Combined classifiers Vs Stacking

3.1. Stacking-based Classification Model

The proposed model is based on the stacking method, as shown in Fig. 6, consisting of two phases. The first phase is the base level which contains the seven machine learning techniques, namely Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF), Extreme Gradient Boosting (EGB), K-Nearest Neighbor (KNN), Decision Tree (DT) and Support Vector Machine (SVM). The dataset is shuffled and divided into 80% for training and 20% for testing. Similarly, 70:30 and 60:40 for training and testing were also considered. The union set of training and testing generated by these seven different basic models is taken as the output of the base level, and then it is considered the new input feature set for the second phase of the meta-level.
So, the performance of the stacking method depends on the performance of base-level classifiers. For the second phase, we have chosen those classifiers showing the best performance among all classifiers used as a classifier for cardia disease prediction and show its presence and absence in the first phase of our model. Next, among those classifiers, one of the classifiers is selected as a meta classifier after taking different possible combinations to reduce the complexity of the model. The performance of the model is evaluated on the training and testing dataset. Finally, by combining the best classifiers for cardiac disease prediction, the highest accuracy is determined.

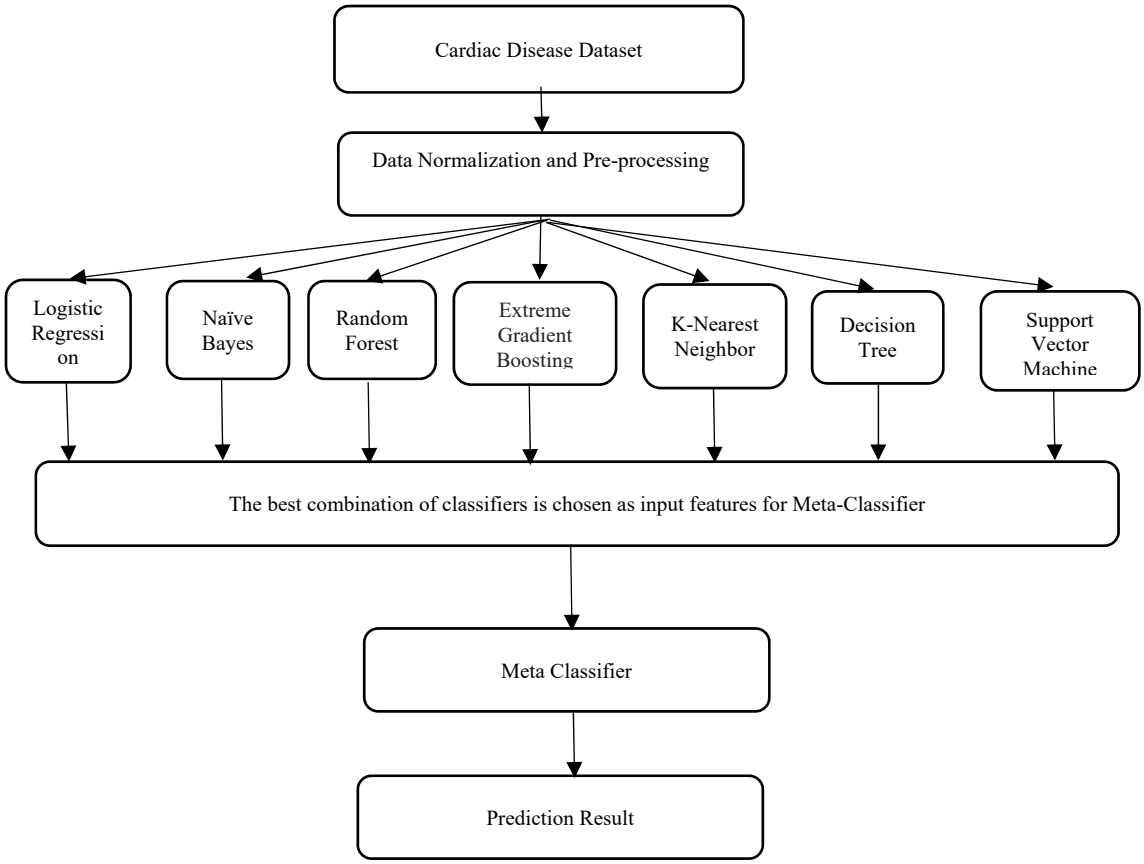

Fig. 6. Process of Stacking method for best classifier combination

3.2. Analysis Using Universal Data Set

We have used a quantitative study. Our objective is to forecast heart diseases. We have applied a numerical technique and the algorithm discussed in the subsequent section on the secondary data to give a precise solution to the recent problem. Our study is based on defining the significance of patients' heart disease. We have used two classification models (both binary and five classes) were designed. The multi-classification (five classes) is designed to analyze the seriousness of heart diseases, and finally, it is converted to binary classification and is used to detect heart diseases.

We have used seven machine learning algorithms (MLA) and then combined them using the stacking method to get a mixed result. We choose the algorithm according to its performance. We have combined these algorithms to develop a model for better accuracy. In general, this idea was adopted. So we randomly verified various combinations to judge the accuracy. Then we randomly split the whole dataset in 80:20 and 70:30 for training and testing. The detailed steps are described as follows.

- Both Standardization and Normalization are applied to the dataset using StandardScaler and MinMaxScaler library in Google Colab.
- In the Pre-processed stage of the data, two different pre-processing approaches were applied separately to that dataset to remove the duplicate and ambiguous values.
- The normalization method scales each input value within the range 0-1, whereas Standardization scales each input variable separately by subtracting the mean (called centering) and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one.
- After Step-1 & 2, the data were grouped into test and train sets
- The dataset is trained based on the seven algorithms
- The dataset is initially imbalanced; we make it balanced by using multiple sampling techniques.
- Then different best classifiers are combined as per their performance. The model that combines the predictions from different classifiers is known as the meta-model, whereas the ensemble members of classifiers are referred to as base models.
- Finally, higher accuracy is determined.

Serial number	Attribute	Explanation	Distribution of Data	Mean	Standard Deviation
1	Age	29-77	Continuous	55	9
2	Gender	Male (1), Fe-Male (0)		0.69	0.49
3	Types	Chest pain categories <1 to 4>	Different values are assigned for various categories	0.987	1.05
4	Rbps	Blood pressure chart	Continuous	135	18.5
5	Serum_C	Cholesterol measures	Continuous	246	52.5
6	RElectro	ECG<0,1,2>	(0,2,3)	0.53	0.53
7	FBS	FBS>120 as 1, FBS<=120 as 0	(0,1)	0.15	0.36
8	Max-Min Heart Rate	Heart Rate between (72 to 202)	Continuous	151	23
9	Ex	Problems due to Physical exercise 0 denoted as no and 1 denoted as yes	(0,1)	0.33	0.48
10	Peak Old	Depression of ST due to physical exercise	Continuous	1.05	1.18
11	Slope	Slope of the peak exercise ST segments	(0,1,2)	1.5	0.65
12	Ca	Various significant vessels affected and marked by fluoroscopy	(0,1,2,3)	0.73	1.05
13	Thal	Disorder in blood	(3,6,7)	2.31	0.61
14	Output	0 to 4, categories of heart diseases	<0,1,2,3,4>	0.5	

Table 5. Explanation of features

3.3. Description of Data

We have taken the universal heart-related data on heart diseases from Cleveland UCI. This set consists of 303 instances, 76 characteristics, and the rest are missing attributes. The missing values have been replaced with 0 in the dataset. In general, there are 13 notable characteristics out of 76 characteristics. The results contain five classes, i.e., no presence of cardiac problems signifies 0, and another severity of heart disease has labels 1 to 4. Detail description of the attributes is given in Table 5, along with each attribute's mean values and standard deviations.

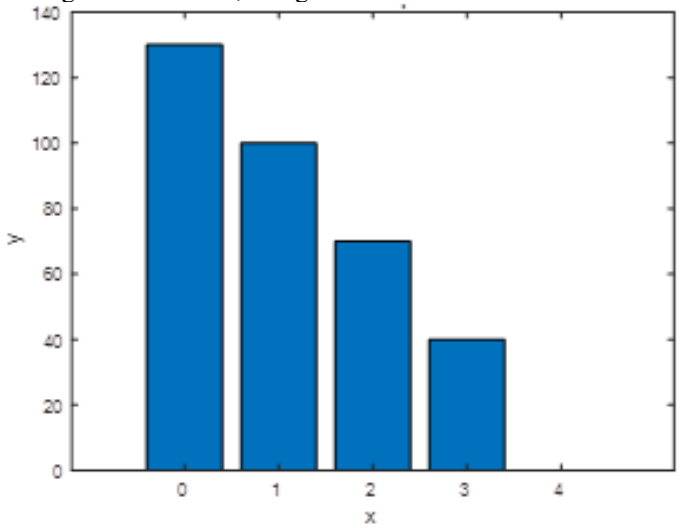

Fig. 7. This is done before processing/pre-processing **(Under Sampling)**

This table also described the data distribution according to its categorical values. The training and testing data set is divided into 70:30 & 80:20; the resulting parameters are estimated in Fig. 7. Techniques like oversampling and under sampling are being used for multiclass class classification. Using under-sampling removes potential valuable data to overcome these anomalies we have used oversampling. Randomly chosen samples from insignificant class examples are added for oversampling to make a balance between significant and insignificant classes described in Fig. 8. Fig. 8 shows the data distribution of the continuous value from the dataset. The output data was balanced for binary classification but unbalanced for multiclass classification. In the first instance, we

resampled the dataset for training using oversampling subsequently used for prediction for the test dataset. We have applied a random over-sampler algorithm to the training dataset to counter the creation of a new test dataset.

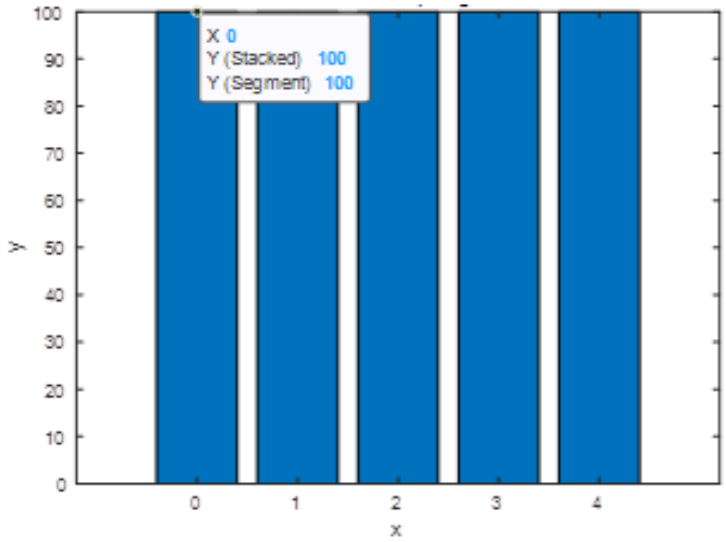

Fig. 8. After processing/after pre-processing **(Over Sampling)**

4. Results and Discussion

To evaluate the general performance of the proposed model, the experiments were executed using Google Colab. The study results are presented as follows: Tables 6 and 7 show the performance of different classifiers using MinMaxScaler taking the consideration of 80:20 and 70:30 training and testing, respectively. As shown in Table 6, XGB, KNN, SVM, RF and NB have better accuracy than other classifiers. So, they have selected baseline models, which was the first layer, that was used to predict the presence or absence of disease. Now, these models will give their predicted values and run in parallel. Next, these predicted values were used as the training set for the meta classifier, which was KNN for our case. The evidence indicates that the stacked classifier has achieved higher accuracy of 91.80 and 86.81, as shown in Table 6 and Table 7, respectively, in both cases using MinMaxScaler normalization to detect cardiac disease.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression (LR)	83.60	0.87 0.82	0.74 0.91	0.80 0.86
Naive Bayes (NB)	85.24	0.88 0.84	0.78 0.91	0.82 0.87
Random Forest (RF)	86.88	0.85 088	0.85 0.88	0.85 0.88
Extreme Gradient Boost (XGB)	90.16	0.89 0.91	0.89 0.91	0.89 0.91
K-Nearest Neighbor (KNN)	85.24	0.88 0.84	0.78 0.91	0.82 0.87
Decision Tree (DT)	81.96	0.77 0.87	0.85 0.79	0.81 0.83
Support Vector Machine (SVM)	88.52	0.92 0.86	0.81 0.94	0.86 0.90
Stacked Classifier (XGB+KNN+SVM+RF+NB)	**91.80**	**0.92** **0.91**	**0.89** **0.94**	**0.91** **0.93**
Stacked Classifier (XGB+SVM+RF+NB)	91.80	0.92 0.91	0.89 0.94	0.91 0.93
Stacked Classifier (XGB+SVM+RF)	91.80	0.92 0.91	0.89 0.94	0.91 0.93
Stacked Classifier (XGB+SVM)	90.16	0.89 0.91	0.89 0.91	0.89 0.91

Table 6 Performance of Different classifiers (80:20 training and testing) using MinMaxScaler

Fig. 9-16 compares the ROC and performance among all solution classifiers. Similarly, again we have taken another standardization method StandardScaler and performed another experiment using all classifiers as previously to train and test the model. Further using, evaluate our model using different evaluation metrics like

accuracy, precision, recall, and F1 score. The evidence indicated that XGB, RF, KNN, SVM, and NB have better accuracy, so they have been taken as base classifiers and KNN as meta classifier for our case. The result shows that the stacked classifier has achieved 93.44 and 83.51 accuracies for both cases 80:20 and 70:30 training/testing, respectively, in Table 8 and Table 9. Fig.s 17-24 show the ROC and performance comparison among the classifiers.

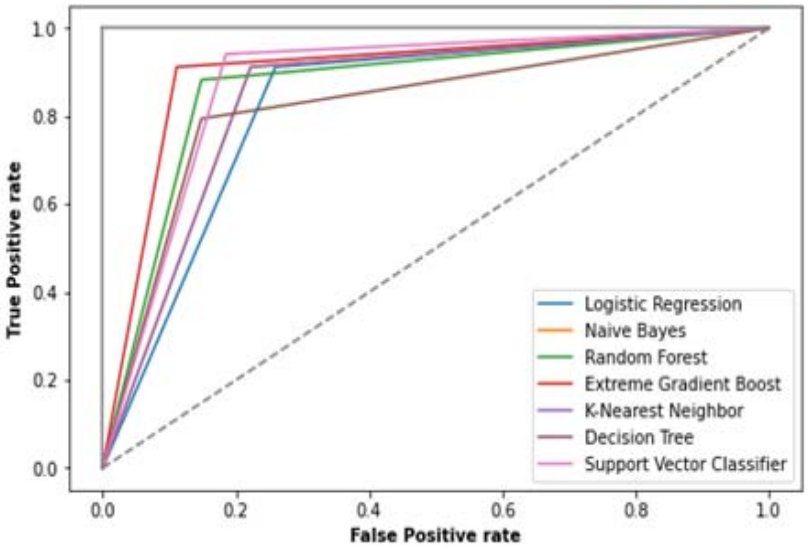

Fig. 9. ROC curve for different classifiers for 80:20 training and testing

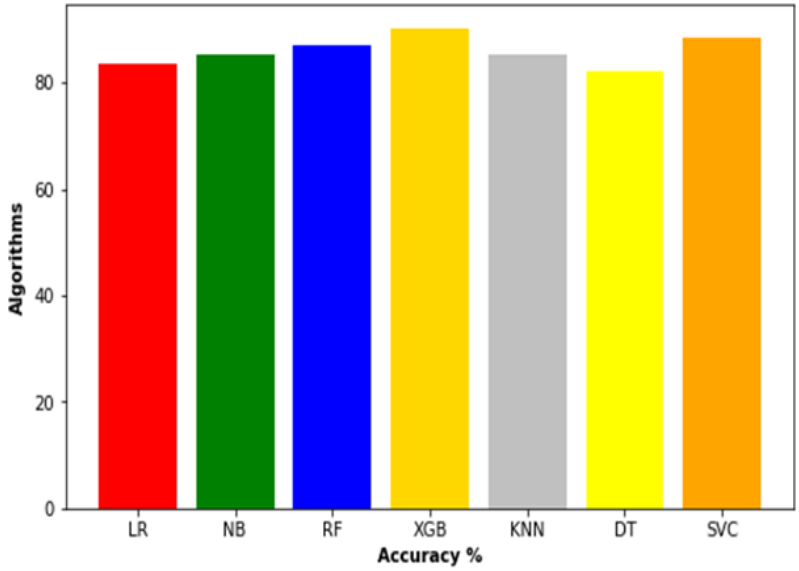

Fig. 10. Accuracy across various Models (80:20 training and testing)

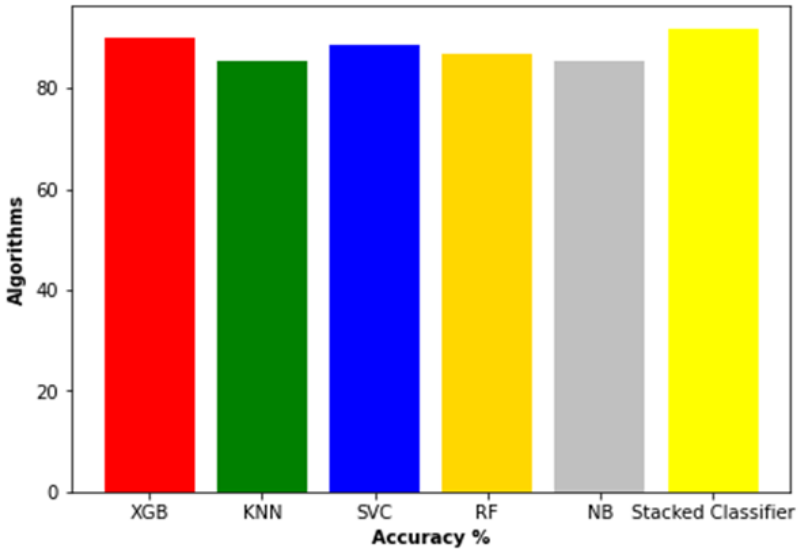

Fig. 11. Accuracy of best 5 classifiers in 80:20 along with Stacking Classifier

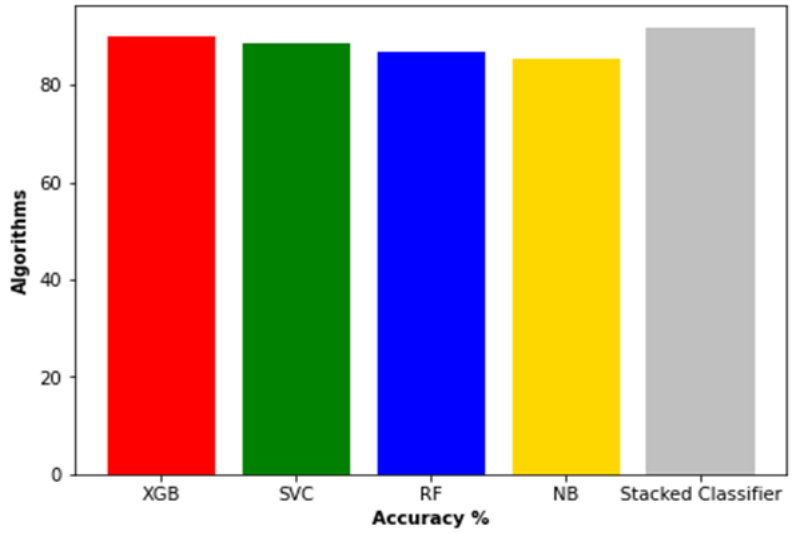

Fig. 12. Accuracy of best four classifiers in 80:20 along with Stacking Classifier

Classifier	Accuracy	Precision	Recall	F1-score
Logistic Regression	79.12	0.84 0.76	0.70 0.87	0.77 0.81
Naive Bayes	80.21	0.84 0.77	0.73 0.87	0.78 0.82
Random Forest Classifier	80.21	0.88 0.75	0.68 0.91	0.77 0.83
Extreme Gradient Boost	80.22	0.86 0.76	0.70 0.89	0.78 0.82
K-Nearest Neighbor	75.82	0.81 0.73	0.66 0.85	0.73 0.78
Decision Tree	80.22	0.82 0.78	0.75 0.85	0.79 0.82
Support Vector Machine	83.51	0.91 0.79	0.73 0.94	0.81 0.85
Stacked Classifier (XGB+DT+SVM+RF+NB)	**86.81**	**0.90** **0.84**	**0.82** **0.91**	**0.86** **0.88**
Stacked Classifier (XGB+RF+NB+SVM)	84.61	0.92 0.80	0.75 0.89	0.83 0.86
Stacked Classifier (XGB+NB+SVM)	85.71	0.88 0.84	0.82 0.89	0.85 0.87
Stacked Classifier (XGB+SVM)	80.21	0.86 0.76	0.70 0.89	0.78 0.82

Table 7. Performance of Different classifiers (70:30 training and testing) using MinMaxScaler

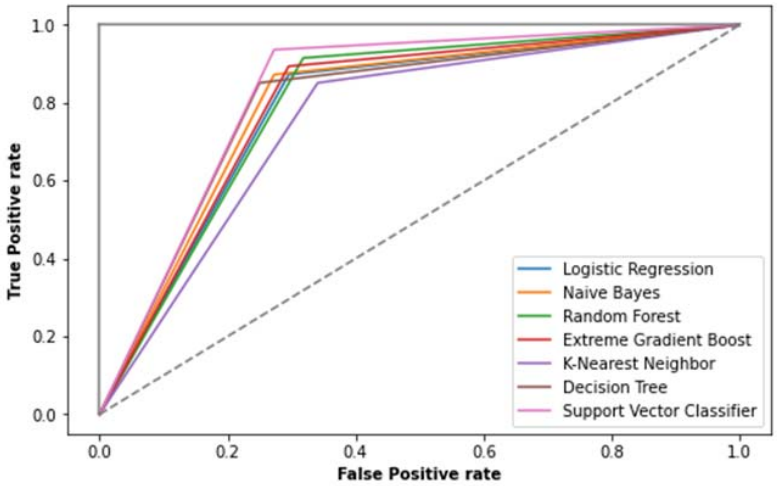

Fig. 13. ROC curve for different classifiers for 70:30 training and testing

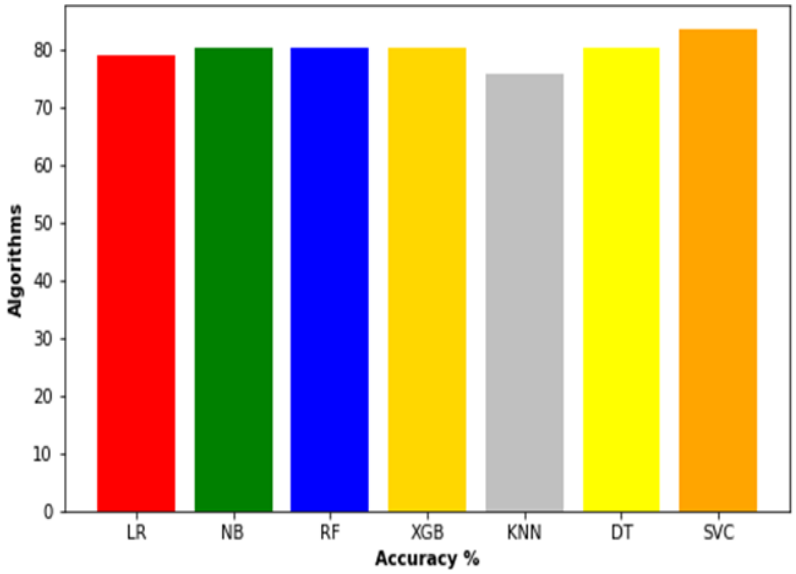

Fig. 14. Accuracy across various Model (70:30 training and testing)

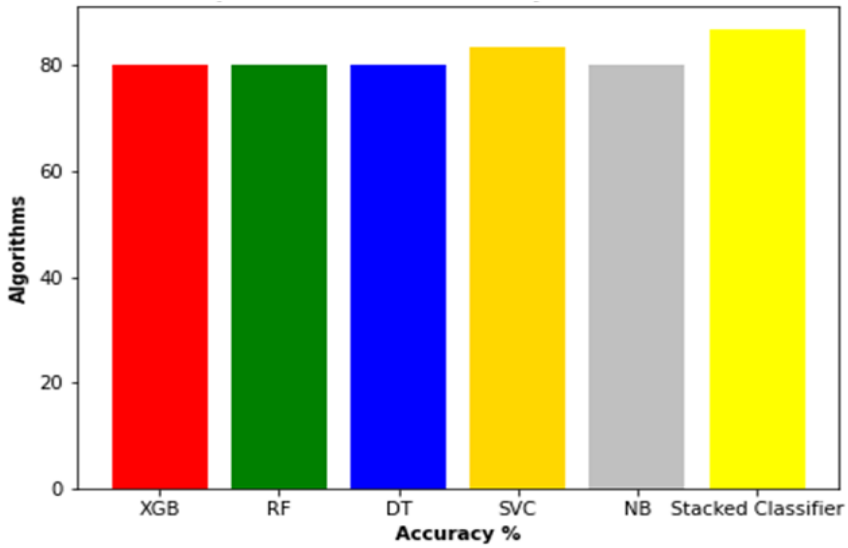

Fig. 15. Accuracy of best 5 classifiers in 70:30 along with Stacking Classifier

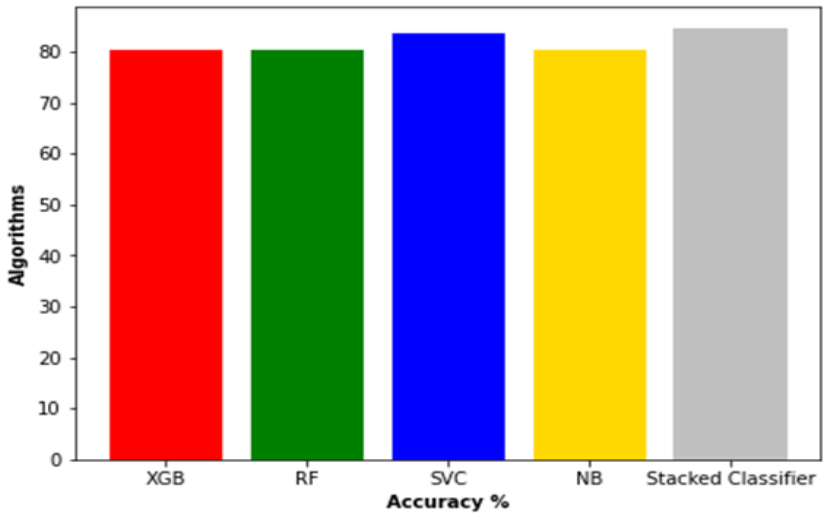

Fig. 16. Accuracy of best 4 classifiers in 70:30 along with Stacking Classifier

Classifier	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	85.24	0.88 0.84	0.78 0.91	0.82 0.87
Naive Bayes	85.24	0.88 0.84	0.78 0.91	0.82 0.87
Random Forest Classifier	85.24	0.82 0.88	0.85 0.85	0.84 0.87
Extreme Gradient Boost	90.16	0.89 0.91	0.89 0.91	0.89 0.91
K-Nearest Neighbor	88.52	0.86 0.91	0.89 0.88	0.87 0.90
Decision Tree	81.96	0.77 0.87	0.85 0.79	0.81 0.83
Support Vector Machine	88.52	0.88 0.89	0.85 0.91	0.87 0.90
Stacked Classifier (XGB+RF+KNN+SVM+NB)	**93.44**	**0.93** **0.94**	**0.93** **0.94**	**0.93** **0.94**
Stacked Classifier (XGB+SVM+KNN+LR)	91.80	0.92 0.91	0.89 0.94	0.91 0.93
Stacked Classifier (XGB+SVM+KNN)	91.80	0.92 0.91	0.89 0.94	0.91 0.93
Stacked Classifier (XGB+SVM)	90.16	0.89 0.91	0.89 0.91	0.89 0.91

Table 8. Performance of Different classifiers (80:20 training and testing) using StandardScaler

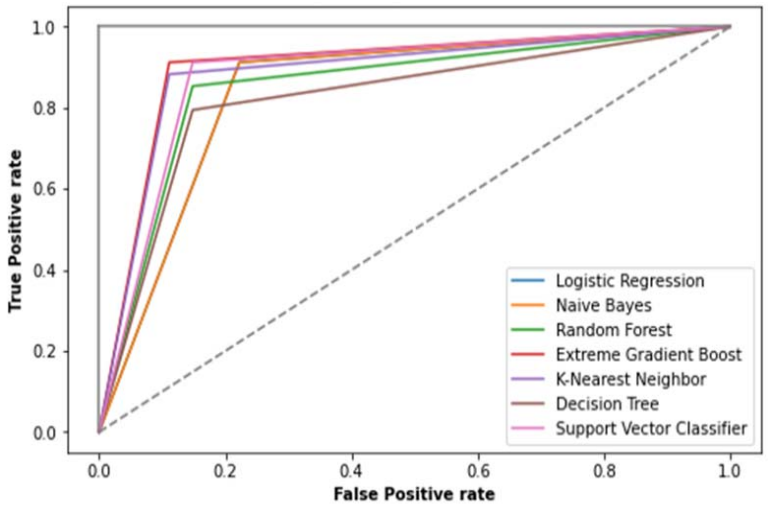

Fig. 17. ROC curve for different classifiers for 80:20 training and testing

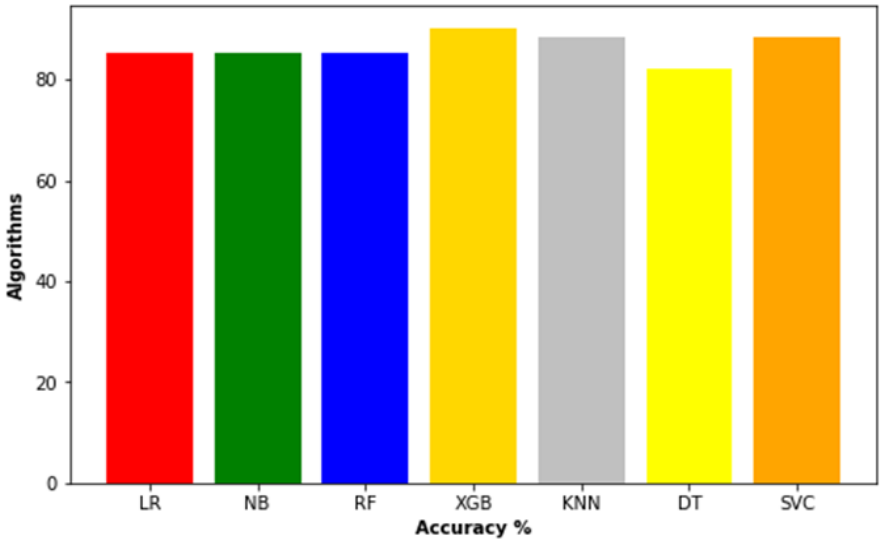

Fig. 18. Accuracy across various Model (80:20 training and testing)

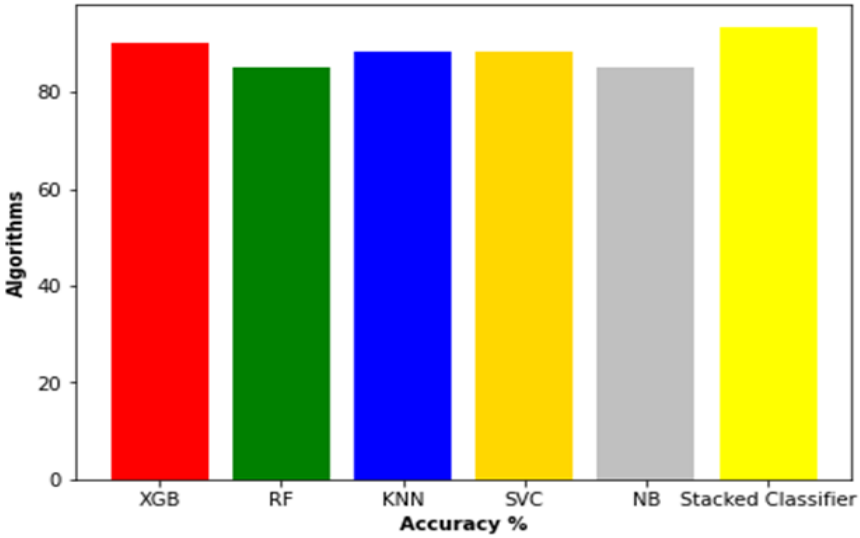

Fig. 19. Accuracy of best 5 classifiers and Stacking Classifier (80:20 training and testing)

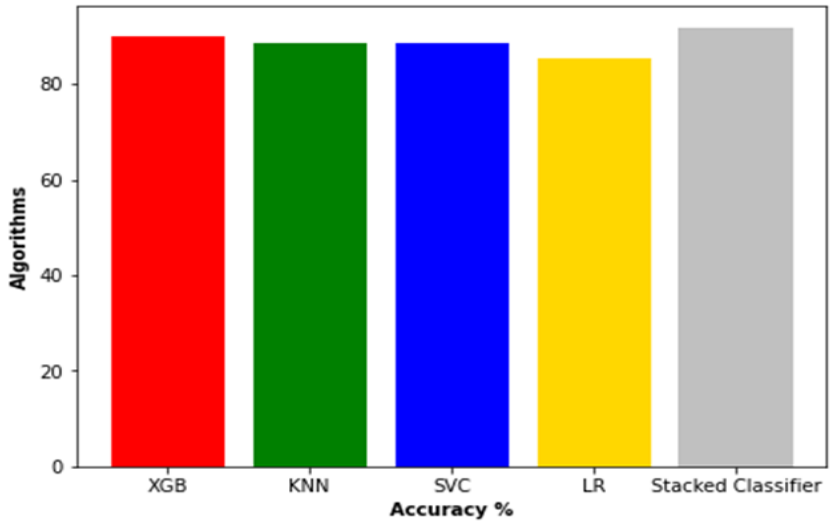

Fig. 20. Accuracy of best 4 classifiers and Stacking Classifier (80:20 training and testing)

Classifier	Accuracy	Precision	Recall	F1-score
Logistic Regression	81.31	0.86 0.78	0.73 0.89	0.79 0.83
Naive Bayes	80.21	0.84 0.77	0.73 0.87	0.78 0.82
Random Forest Classifier	81.31	0.89 0.77	0.70 0.91	0.78 0.83
Extreme Gradient Boost	80.22	0.86 0.76	0.70 0.89	0.78 0.82
K-Nearest Neighbor	82.42	0.87 0.79	0.75 0.89	0.80 0.84
Decision Tree	80.22	0.82 0.78	0.75 0.85	0.79 0.82
Support Vector Machine	84.61	0.89 0.81	0.77 0.91	0.83 0.86
Stacked Classifier (XGB+RF+KNN+SVM+LR)	**83.51**	**0.89** **0.80**	**0.75** **0.91**	**0.81** **0.85**
Stacked Classifier (RM+KNN+SVM+LR)	82.41	0.87 0.79	0.75 0.89	0.80 0.84
Stacked Classifier (RM+KNN+SVM)	82.41	0.87 0.79	0.75 0.89	0.80 0.84
Stacked Classifier (SVM+KNN)	82.41	0.87 0.79	0.75 0.89	0.80 0.84

Table 9. Performance of Different classifier (70:30 training and testing) using StandardScaler

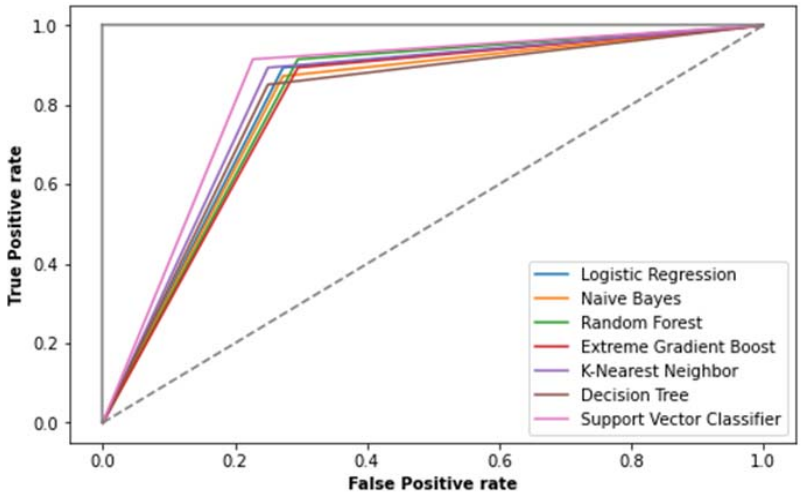

Fig. 21. ROC curve for different classifier for 70:30 training and testing

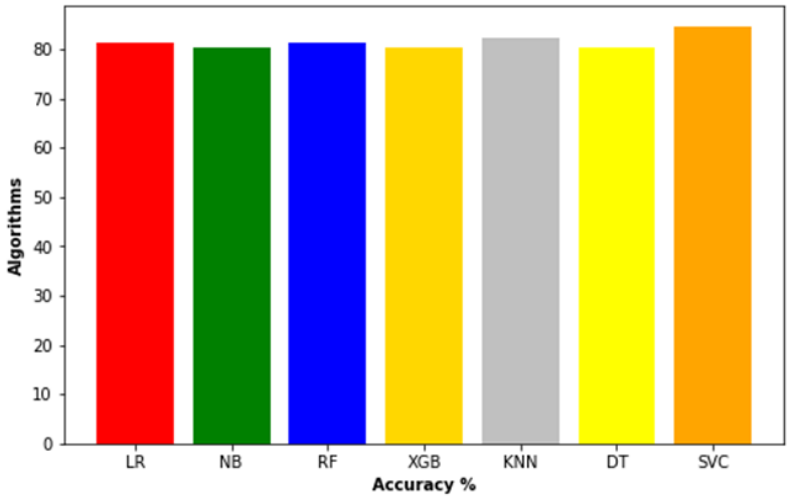

Fig. 22. Accuracy across various Models (70:30 training and testing)

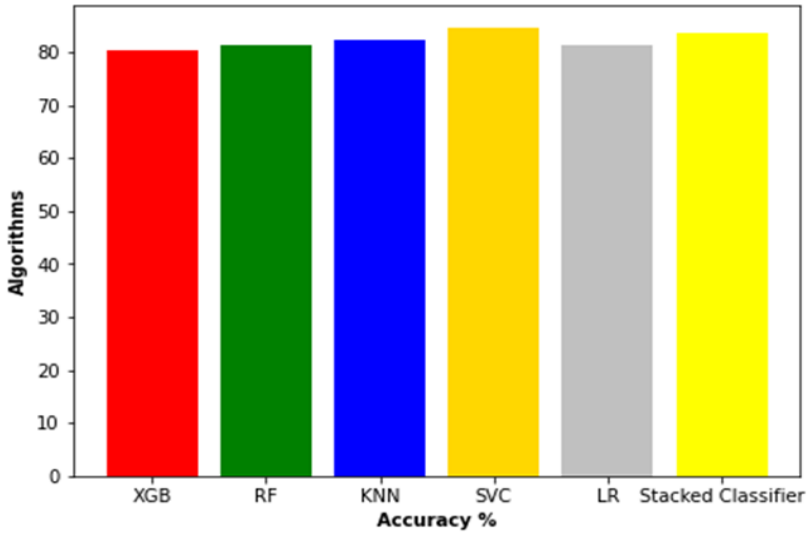

Fig. 23. Accuracy of best 5 classifiers in 70:30 along with Stacking Classifier

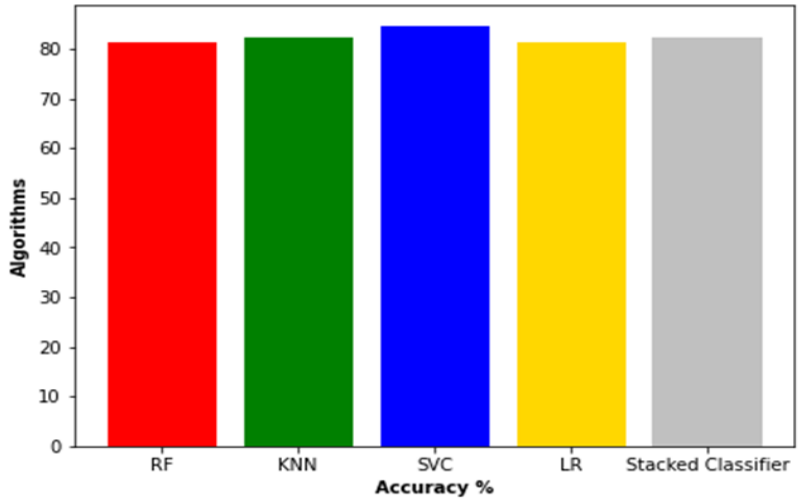

Fig. 24. Accuracy of best 4 classifiers in 70:30 along with Stacking Classifier

Training: Testing	**Classifiers**	**MAE**	**MSE**	**RMSE**	**MSLE**
80:20	**Stacked Classifier (XGB+KNN+SVM+RF+NB)**	**0.081**	**0.081**	**0.286**	**0.039**
	Stacked Classifier (XGB+SVM+RF+NB)	0.081	0.081	0.286	0.039
	Stacked Classifier (XGB+SVM+RF)	0.081	0.081	0.286	0.039
70:30	**Stacked Classifier (XGB+DT+SVM+RF+NB)**	**0.131**	**0.131**	**0.363**	**0.063**
	Stacked Classifier (XGB+RF+NB+SVM)	0.153	0.153	0.392	0.073
	Stacked Classifier (XGB+NB+SVM)	0.142	0.142	0.377	0.068

Table 10. Performance of Different Stacking classifiers using Loss function under MinMaxScaler

Further different error calculation function has been taken into consideration to evaluate the performance of stacked classifiers. We have considered mean absolute error (MAE), mean squared error (MSE), root mean squared (RMSE) and mean squared logarithmic (MSLE) as different loss functions. The desired lower value for error confirms the better accuracy of the model. The detailed performance evaluation of the stacked classifier is given in Tables 10 and 11.

Training: Testing	Classifiers	MAE	MSE	RMSE	MSLE
80:20	**Stacked Classifier (XGB+RF+KNN+SVM+NB)**	**0.065**	**0.065**	**0.256**	**0.031**
	Stacked Classifier (XGB+SVM+KNN+LR)	0.081	0.081	0.286	0.039
	Stacked Classifier (XGB+SVM+KNN)	0.081	0.081	0.286	0.039
70:30	**Stacked Classifier (XGB+RF+KNN+SVM+LR)**	**0.164**	**0.164**	**0.405**	**0.079**
	Stacked Classifier (RM+KNN+SVM+LR)	0.175	0.175	0.419	0.084
	Stacked Classifier (RM+KNN+SVM)	0.175	0.175	0.419	0.084

Table 11 Performance of Different Stacking classifiers using Loss function under StandardScaler

5. Conclusion

This work focused on the seriousness of the cardiac disease and its prediction. Our work has included RST to find the significant disease among all cardiac diseases like pneumonia, asthma, cardiac arrest, atelectasis, and pulmonary edema. We found that cardiac arrest was the most significant disease among all cardiac-related diseases. So further, we have focused on cardiac arrest. In our proposed method, machine learning and stacking with all the best combinations of classifiers are used. The technique is based on improving the accuracy of prediction for cardiac disease diagnosis. In our work, we have used the Cleveland dataset and normalized it with two different methods named min-max normalization and StandardScalar standardization (Z scores). Applying different machine learning techniques, we can predict whether the disease is present or absent. Finally using, various combinations of classifiers we are able to predict cardiac arrest. The maximum accuracy we obtained by using this technique is 93.44 as our z-score and 91.80 as our normalization min-max respectively and error was also calculated.

References

[1] Khourdifi Y, Bahaj M. (2019): Heart Disease Prediction and Classification Using Machin Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization, International Journal of Intelligent Engineering and Systems, **12**(1):242–52.
[2] Mohan S, Srivastava G, Thirumalai C. (2019): Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. IEEE Access, **7**, pp. 81542-81554.
[3] Dai W, Brisimi T.S., Adams W.G., Mela T, Saligrama V, Paschalidis I. (2015): Prediction of hospitalization due to heart diseases by supervised learning methods, International Journal of Medical Informatics, **84**(3), pp. 189–197.
[4] Durairaj M., Revathi, V. (2015): Prediction of Heart Disease Using Back Propagation MLP Algorithm, **4**(08), pp. 235–239..
[5] Gavhane A. (2018): Prediction of Heart Disease Using Machine Learning, Second International Conference on Electronics, Communication, and Aerospace Technology (ICECA), pp. 1275–1278.
[6] Abdullah A. S. (2012): A Data mining Model for predicting the coronary heart disease using Random Forest Classifier, Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls, pp. 22–25.
[7] Rathnayakc B.S.S. and Ganegoda G.U. (2018): Heart diseases prediction with Data Mining and Neural Network Techniques, 3rd Int Conference Convergence Technology (I2CT), pp.1–6.
[8] Kelwade J.P. (2016): Radial basis function Neural Network for Prediction of Cardiac Arrhythmias based on Heart rate time series, Conference. IEEE First Int Conf Control Measurement Instrument (CMI), pp. 454–8.
[9] Anooj P.K. (2012): Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University- Computer and Information Sciences, **24**(1), pp. 27–40.
[10] Krishnaiah V, Chandra, N.S. (2016): Heart disease prediction system using data mining techniques and intelligent fuzzy approach: a review, International Journal of Computer Applications, **136**(2), pp. 43–51.
[11] Fatima M, Pasha M. (2017): Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl., **9**(01):1.
[12] Otoom AF, Abdallah EE, Kilani Y, Kefaye A, Ashour M. (2015): Effective diagnosis and monitoring of heart disease. Int J Software Eng Appl., **9**(1), pp. 143–56.
[13] Vembandasamy K, Sasipriya R, Deepa E. (2015): Heart diseases detection using naive Bayes algorithm, IJISET- International Journal of Innovative Science, Engineering & Technology, **2**, pp. 441–4.
[14] Malav A, Kadam K, Kamat P. (2017): Prediction of heart disease using k-means and artificial neural network as a hybrid approach to improve accuracy, Int J Eng Technol., **9**(4).
[15] Lee HG, Noh KY, Ryu KH. (2007): Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV, Pacific-Asia Conference on Knowledge Discovery and Data Mining Emerging Technologies in Knowledge Discovery and Data Mining, pp. 218–28.
[16] Tarle B. (2017): An artificial neural network based pattern classification algorithm for diagnosis of heart disease, Int Conf Comput Commun Control Automation (ICCUBEA), pp. 1–4.
[17] Saxena K, Purushottam, Sharma R(2016): Efficient Heart Disease Prediction System, Procedia Computer Science, **85**, pp. 962–969.
[18] Khaled M. A. (2020): Prediction of heart disease and classifiers sensitivity analysis, BMC Bioinformatics, **21** (278).

[19] Karaylan T, Kilic O. (2017): Prediction of heart disease using neural network, Int Conf Computer Sci Eng (UBMK) Antalya, pp. 719–23.
[20] Esfahani HA, Ghazanfari M. (2017): Cardiovascular disease detection using a new ensemble classifier, IEEE 4th international conference on knowledge-based engineering and innovation (KBEI), Tehran, **2017**, pp.1011–1014.
[21] Pawlak Z. (1982): Rough Sets, International Journal of Computer & Information Science, **11**, pp. 341–356.
[22] Konrad E., Orłowska E., and Pawlak Z. (1981): *An approximate concept learning* ,Berlin, Bericht, pp. 81–87.
[23] Marek W. and Pawlak Z. (1981): Rough sets and information systems, *ICS PAS Reports* (441).
[24] Michalski R., S. (1971): Pattern Recognition as Role-Guided Inductive Interference, *IEEE Transaction on Pattern Analysis and Machine Intelligence,* **2**, pp.179–187.
[25] Muhammad, Y., Tahir, M., Hayat, M. *et al*. (2020): Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Sci Rep* **10,** 19747.
[26] Bui, A. L., Horwich, T. B. & Fonarow, G. C. (2011): Epidemiology and risk profile of heart failure. *Nat. Rev. Cardiol.* **8**, 30, 2011.
[27] Alizadehsani R., Habibi J., Hosseini M. J., Mashayekhi H., Boghrati R., Ghandehariouna A., Bahadorian B., Sani Z.A. (2013): A data mining approach for diagnosis of coronary artery disease, Computer methods, and programs in biomedicine ,**111**,issue-1.
[28] Vanisree, K. & Singaraju, J. (2015): Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. *Int. J. Comput. Appl.,* **19**, pp. 6–12.
[29] Nazir, S., Shahzad, S., Mahfooz, S. & Nazir, M. (2018): Fuzzy logic-based decision support system for component security evaluation. Int.Arab J. Inf. Technol, **15**, pp. 224–231.
[30] Gudadhe, M., Wankhade, K. & Dongre, S. (2010): Decision support system for heart disease based on support vector machine and Artificial Neural Network In 2010 International Conference on Computer and Communication Technology (ICCCT), pp. 741–745.
[31] Palaniappan S, Awang R(2008): Intelligent heart disease prediction system using data mining techniques, In IEEE/ACS international conference on computer systems and applications, 2008. AICCSA 2008, IEEE, pp 108–115.
[32] Olaniyi, E. O., Oyedotun, O. K. & Adnan, K. (2015): Heart diseases diagnosis using neural networks arbitration, Int. J. Intel. Syst. Appl. , **7**(12).
[33] Das R., Turkoglu, I. & Sengur, A. (2011): Effective diagnosis of heart disease through neural networks ensembles. *Expert Syst. Appl.* **36**,7675–7680.
[34] Tomov, N. S. & Tomov S. (2018): On deep neural networks for detecting heart disease. arXiv :1808.07168.
[35] Mohan S., Thirumalai C. & Srivastava G. (2019): Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* **7**, 81542–81554.
[36] Mishra S, Mohanty S.P., Pradhan S.K. (2016): Reasons for employees need justice from legal bodies: A rough set approach, 3rd International Conference on Computing for Sustainable Global Development (INDIACom), Conference Paper | Publisher: IEEE, PP.2018-2022.
[37] Das S, Pradhan S. K., Mishra S, Pradhan S, Pattnaik P. K. (2021): Analysis of Heart Diseases Using Soft Computing Technique, (OCIT) IEEE-Conference.
[38] Nayak S. k., Pradhan S.K., Mishra S, Pradhan, Pattnaik P.K. (2021): Rough Set Technique to Predict Symptoms for Malaria, 8th International Conference on Computing for Sustainable Global Development (INDIACom) IEEE.
[39] Mishra, S., Mohamed, A., Pattnaik, P.K., Muduli, K., Ahmad, T.S.T. (2022): Soft Computing Techniques to Identify the Symptoms for COVID-19, Advances in Data Science and Management, Lecture Notes on Data Engineering and Communications Technologies, vol 86. Springer, Singapore.
[40] Peng, H., Long, F. & Ding, C. (2005): Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238.
[41] UCI Machine learning Repository, Centre for Machine Learning and Intelligent Systems Date of Donation 1.7.1988, Area Life.

Authors Profile

Subhalaxmi Das is a Ph.D. Research Scholar in the Department of Computer Science and Applications, Utkal University, Bhubaneswar, Odisha, India. She received B.Tech in Computer Science and Engineering from Biju Patnaik Universityof Technology, Odisha, India and M.Tech in Computer Science & Engineering from KIIT University, India. She has 12 years of teaching experience. She has published some reseach papers in International Journals and Conferences.Her special fields of interest included Machine Learning, Soft Computing, Image Processing and Data Mining.

Sujogya Mishra received Ph. D. in Computer Science from Utkal University, India. He is currently working as an Assistant Professor in the department of Mathematics and Computing in Odisha University of Technology and Research, Bhubaneswar, Odisha, India. He has published more than 26 publications in reputed international Scopus based high reputed Journals in the areas of Soft Computing, Rough Set Theory, Machine Learning and Bioinformatics which are the subjects of his research interest. He has over 25 years of teaching and research experiences.

Pradyumna Kumar Pattnaik is currently working as an Assistant Professor in the department of Mathematics and Computing in Odisha University of Technology and Research, Bhubaneswar, Odisha, India. He has more than 46 publications in reputed international Scopus based high reputed Journals. He has over 22 years of teaching and research experiences.

Dr. Sipali Pradhan completed her Ph.D. in Computer Science & IT from North Orissa University, Baripada , Odisha, India in the year 2019. The title of her Ph. D. thesis is “Rea-Time HealthCare Using Internet of Things”. Dr. Sipali Pradhan completed her M. Tech in Computer Science & Engineering from Utkal University, Bhubaneswar, India in the year 2015 before joining North Orissa University for Ph.D. programme. She is current working as Assistant Professor, Department of Computer Science, RBVRR Women’s College (Osmania University), Hyderabad, India. There are more than ten Journal & Conference Publications to her credit.

Professor Sateesh Kumar Pradhan completed his Ph.D. in Computer Science from Berhampur University, Berhampur, Odisha, India. He was the Former Professor & Head, P.G. Department of Computer Science, and Former Dean Faculty of Engineering, Utkal University, Bhubaneswar, Odisha, India. Twenty-two research scholars have been awarded Ph.D. degree in the area of E-Commerce, Mobile Ad hoc Networks, Intrusion Detection, Computer Forensic, Web Services, Parallel Task Scheduling, Medical Data Mining, Software Testing, Internet of Things, Brain Image Classification, Optical Character Recognition of Odia Document etc. under his supervision and currently eight scholars are continuing for Ph. D. degree under his supervision and guidance. There are more than forty Journal Publications and more than Sixty Conference Publications to his credit. He also served as Senior Faculty in Computer Engineering, King Khalid University, Abha, Saudi Arabia (2006-2011) and Faculty, Computer Science, Berhampur University, Odisha, India (1987-2000).