RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 01 January
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.003

Discovery of Approaches by Various Machine learning Ensemble Model and Features Selection Method in Critical Heart Disease Diagnosis

Gyanendra Kumar Pal[1], *Sanjeev Gangwar*[2]

[1]*Research Scholar, Department of Computer Science & Engineering, VBS Purvanchal University, Jaunpur, Lucknow, India*

[2]*Assistant Professor, Department of Computer Applications, VBS Purvanchal University, Janupur, Lucknow, India*

Emails: gyanpal@gmail.com, gangwar.sanjeev@gmail.com

Article History

Received: 15 December 2022
Accepted: 24 January 2023

Keywords:

Extra tree feature selection methods;
machine learning;
heart disease;
Bagging;
Support Vector Machine;
Multilayer Perception;
Gradient Boost

Abstract

Heart disease is one of the leading killers that are widely recognized throughout the globe. Large volumes of clinical data are stored in a variety of systems and biological equipment at hospitals. It is essential to grasp the facts of heart disease in order to improve forecast accuracy. In this paper, experimental evaluations have been conducted to assess the effectiveness of models created utilizing classification algorithms and relevant attributes selected using Extra Tree feature selection procedures. Several people suffer originated at heart disease globally. It is necessary to use data mining and machine learning techniques to extract new insights originated at this data. Analyzing medical data sets and diagnostic issues, including heart disease, involved the use of a number of categorization approaches. However, these methods were only performed on small, balanced data; then, the features must be derived originated at trial and error. Additionally, several sectors have made substantial use of feature selection techniques to enhance classification performance. This paper aims to propose a comprehensive approach to enhance the prediction of heart disease using several machine learning methods such as Bagging, Support Vector Machine, Multilayer Perception and Gradient Boost with feature selection methods such as extra tree. The experimental results showed improvements of prediction. Bagging received scores in training model on 80% data sample as 99.08, 73.19, 67.20, 69.20 and 80.66 of accuracy, precision, recall, F1-score and roc respectively. In the experiment, we have tested on 20% data sample for each classifier algorithms and find Bagging classifier model perform higher score for accuracy, precision, recall, F1-score and roc 92.62, 48.44, 39.63, 41.89, 66.82 respectively.

1. Introduction:

Advanced statistical techniques may be used to uncover relevant databases utilizing learning techniques, which are relatively new and promising technologies. Many scholars are interested in the relatively young and emerging field of medical data mining and knowledge exploration (Chauhan et al.).

With greater medical data collection, doctors may be able to make more accurate diagnoses. Cardiovascular illnesses have been shown to have the greatest death rates among these conditions in the majority of nations globally (Fu et al.).

Analysist identify illnesses more accurately from disease dataset. Various researchers has been used UCI repository disease dataset. They observed disease features by various features selection techniques and predicted by learning algorithms. Authors used various classifiers for training and testing purpose for gather more information from dataset (Jindal).

The first and most common method for choosing features originated at labeled data is supervised feature selection. The filter, wrapper, and embedding techniques are used in supervised feature selection methods. In the preprocessing step, Extra tree methods are applied regardless of the learning algorithm. This method determines the score based on statistical measures and their dependence on the class label for each characteristic (Gnaneswar and Jebarani).

2. Related Work:

In this research paper, we have studied various previous year research works on different dataset. Various authors used many selected machine learning and deep learning algorithms and test algorithms performance. Some used algorithms are listed in table 1, the literature makes it clear that classifier training using relevant features chosen by various feature selection techniques improves the classifier's accuracy.

3. Methodology:

The ML research community frequently uses the UCI data set. It has 1025 records total, with 14 distinct variables. However, just 14 of the factors [Age. Sex. Chest Pain. Resting Blood Pressure. Cholesterol. Fasting Blood Sugar.

Restecg. Thalach. Exang. Oldpeak. Slope. Ca. Thal. And Class.] Have been found to be significantly associated with heart disease. The descriptions of each variable are shown in table 2.

3.1. Data Description:

The dataset must be preprocessed in order to effectively reflect the data quality. Preprocessing methods used on the dataset include removing missing values originated at features. Data preparation methods like missing value management are used to make a smooth dataset.

3.2. Proposed Work:

With the use of fewer characteristics in a dataset of heart illness, the suggested study aims to improve classification accuracy. Figure 1 shows the classification scheme for heart disorders. The components of the recommended framework are described in the sections that follow.

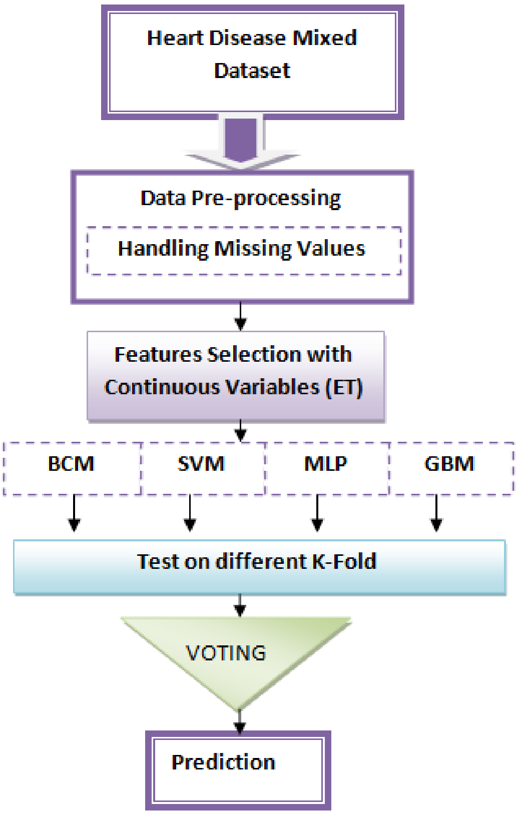

FIGURE 1. Representation of proposed work on heart disease using classifiers

The dataset are organized from UCI repository for training and testing of classifiers. When this dataset was created, there were 14 features and 1025 instances. The label of the output characteristic (num) is separated into two classes, listed in figure 1, The experimental procedure was created to assess how well the search algorithms and strategies worked together when they were applied to the following classification models: Gradient Boost, Multilayer Perception, Support Vector Machine, and Bagging. The results of our application classification models with 10-fold cross validation were given in improvements and decreases in accuracy with respect to epoch values. In the end, we examined the experimentation's outcomes. The primary objective, as previously stated, is to improve the ability to forecast heart disease. But this study also of a helpful manual for picking the optimum feature selection method for various classification models.

TABLE 1. **Representation of previous work on heart disease using machine learning**

Year	Author	Dataset	Techniques	
2019	Ravindhar N et al. (Hasan et al.)	UCI + machine learning	LR. NB. FK-NN. K-Means. and BPNN.	BPNN : 98
2020	Magar R et al. (Ravindhar et al.)	UCI + machine learning	LR. SVM. NB. And DT.	LR : 83
2020	Rajdhan A et al. (Rajdhan et al.)	UCI + machine learning	LR. DT. RF. And NB.	RF : 90
2020	Shah D et al. (Rajdhan et al.)	UCI + machine learning	NB. KNN. RF. And DT.	KNN : 90
2021	Jindal H et al. (Shah, Patel, and Bharti)	UCI + machine learning	KNN. LR. And KNN + LR .	KNN : 89
2021	Pandita A et al. (Pandita et al.)	UCI + machine learning	LR. KNN. SVM. NB. And RF.	KNN : 89
2021	Akella A et al. (A. Akella and S. Akella)	UCI + machine learning	GLM. DT. RF. SVM. NN. And KNN.	87.64% NN : 93
2022	Gupta, C. et al. (Gupta et al.)	UCI + machine learning	RF. DT. And LR.	LR : 92
2022	Truong, V.T et al. (Truong et al.)	UCI + machine learning	AB. ET. LR. MNB. CART. LDA. SVM. RF. And XGM.	AB : 90
2022	Abdalrada, A.S. et al. (Abdalrada et al.)	UCI + machine learning	SVM. NB. And DT.	DT : 90
2022	Singh, N. et al. (N. Singh and Bhatnagar)	UCI + machine learning	KNN. DT. LR. NB. And SVM.	LR : 92

3.3. *Methods Description:*

In this experiment, we have used various classifiers techniques and features selection techniques, describe as below:

3.3.1. *Feature Selection Technique:*

Without bootstrapping, the Extra Tree Classifier (Isabona et al.) creates randomised multiple decision trees with various sub-samples. [0.24 0.03 0.06 0.110.11 0.030.04 0.10 0.03 0.09 0.05 0.06 0.04 0.02]listed in figure 2.

The issue of over fitting is avoided. The effectiveness of data mining techniques is decreased when inappropriate characteristics are included in the dataset. The best feature combinations must first be correctly identified before the best approaches are determined. It is anticipated that accuracy and other performance measures would increase when the optimum feature combination is applied over the methodologies. The process of creating a new set of features originated at the original features is known as feature extraction. In order to lessen the impact of duplication and inconsistency, it integrates the original characteristics (D. Yadav et al.).

3.3.2. *Machine Learning Classifiers:*

Data categorization is still a desirable area in machine learning. The next subsections give a quick introduction to some of the recently suggested algorithms that have been studied in a variety of fields, such as Gradient Boost, Multilayer Perception, and Support Vector Machine Bagging.

Support Vector Machine:: The Support Vector Machine technique is used in this research study to forecast human heart disease. We choose the SVM method for the prediction process because, in comparison to other machine learning algorithms, it will provide a higher level of accuracy. A graphical description of the algorithm's accuracy is presented. The results of the provided data are displayed graphically in the

TABLE 2. Representation of heart disease dataset attributes description

S.No	Variable	Description
1	Age	Age regarding the person in years
2	Sex	Gender 1- Male, 0- Female
3	Chest Pain	1- typical angina, 2- atypical angina, 3- non anginal pain, 4- asympomatic
4	Resting Blood Pressure	Blood Pressure in mm Hg during hospital admission
5	Cholesterol	Serum cholesterol in mg/dl
6	Fasting Blood Sugar (fbs)	If (fbs>120mg/dl) 1 = True, 0 = False
7	Restecg	Electrocardiography 0- Normal, 1- may be some problem, 2- definite problem
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina 1- Yes, 0- No
10	Oldpeak	Induced ST depression due to exercise
11	Slope	Slope regarding the ST segment during peak exercise. 1- Upsloping, 2- flat, 3- downsloping
12	Ca	Number regarding blood vessels coloured by fluoroscopy Values ranges originated at 0 to 3
13	Thallium Scan	It is a method regarding analysing blood flow to heart muscles. 3= normal, 6= fixed defect, 7= reversable defect
14	Class	It is the output or dependable variable 0= No heart disease, 1,2,3 and 4 represents the severity regarding the heart disease

section below (Gangwar and G. K. Pal).

Bagging:: This technique used as ensemble method in experiment. It is basically used to increase accuracy level in machine learning algorithms. Bagging technique reduce the over fitting in analysis and generate parallel way for all selected techniques. In statistical analysis bagging generate average accuracy level with low variance (D. Singh, H. Yadav, and Agrawal).

Multilayer Perception Model:: The suggested multilayer perception model was created with the express intention of lowering the confusion matrix and improving the precision of illness grouping based on severity. Originated at this point forward, the work suggests the multilayer perception model. The suggested MLP is set up so that the input layer is in charge of handling the training inputs, the hidden layers are accessible to consider weight modification, and lastly the output nodes are grouping the results into distinct categories (Isabona et al.).

Gradient Boosting:: A traditional feed forward neural network classifier called Gradient Boost uses the output mistakes to train the network. Three levels of nodes make up Gradient Boost: the input layer, at least one or more hidden layer(s), and the output layer. The hidden layers are connected to the output layer by the input layer. Weighted values are used to process each layer. A Gradient Boost with a single hidden layer is depicted. Gradient Boost is a one-way error propagation method that has been trained and tested using back-propagation techniques (Gangwar and G. Pal).

4. Result:

In this experiment, we have used training sets 80% and testing sets 20%. The 10-fold build models are used for evaluating the accuracy, precision, recall, and F-score. The experimental setup used learning techniques in tables Gradient Boost, Multilayer Perception, Bagging, and Support Vector Machine.

5. Discussion:

shows the performance of all classifiers on different k-folds as k=6, 8 & 10 after applying Extra Tree features selection method. In comparison to the other classifiers, the findings indicated that Bagging had the best performance across all assessment metrics. In the experiment, we find on the k=10, each classi-

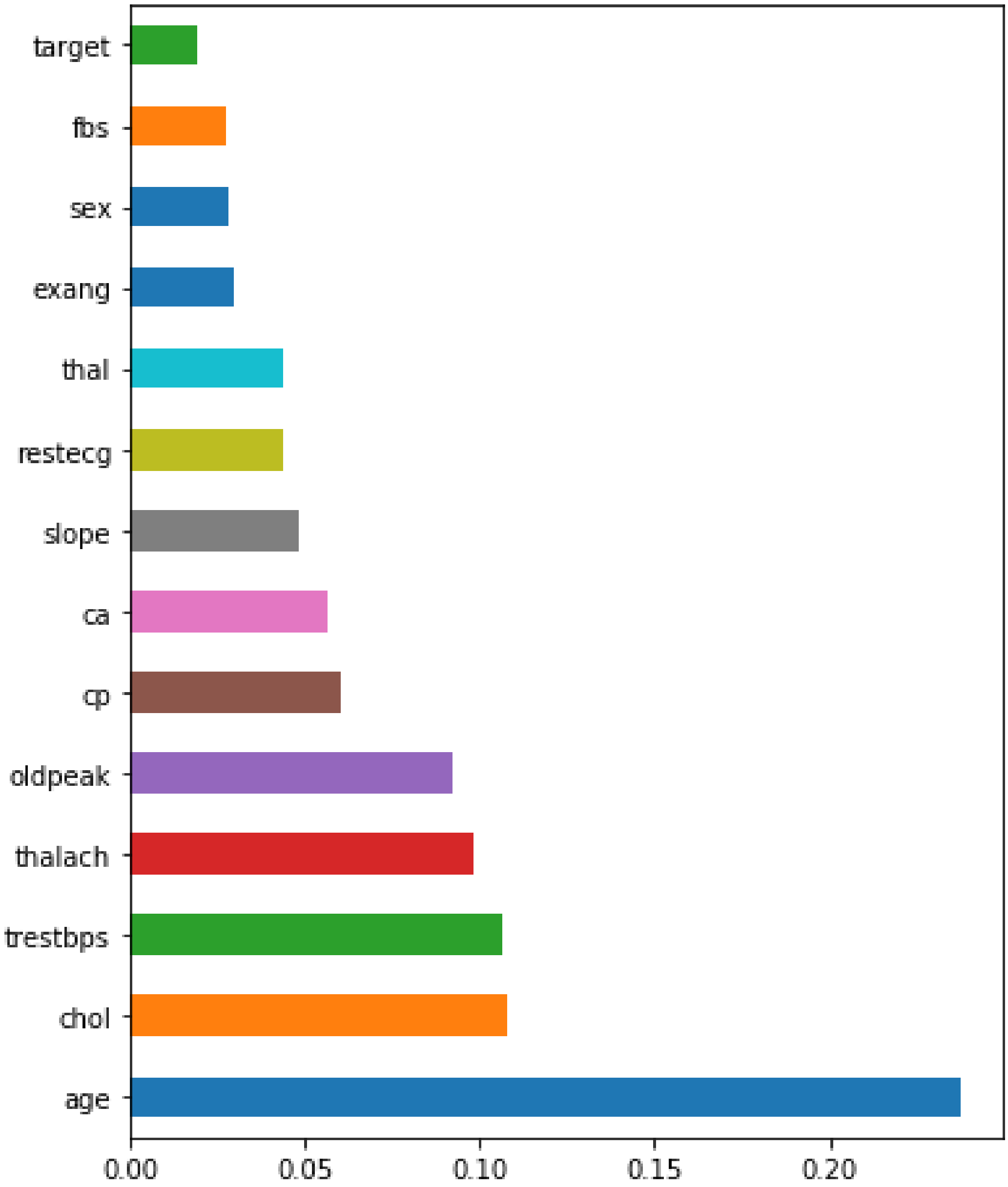

FIGURE 2. **Representation of Extra Tree classifier work on heart disease**

fiers perform better for accuracy, precision, recall, F1-score and roc. The Bagging classifier model (BCM) perform better compare to all other classifiers. Bagging received scores in training model on 80% data sample as 99.08, 73.19, 67.20, 69.20 and 80.66 of accuracy, precision, recall, F1-score and roc respectively, listed in table 3.

In the experiment, we have tested on 20% data sample for each classifier algorithms and find Bagging classifier model perform higher score for accuracy, precision, recall, F1-score and roc 92.62, 48.44, 39.63, 41.89, 66.82 respectively.

Our suggested methodology provides doctors with basic diagnosis for additional medical care. We draw the conclusion that the methodology can increase the rate at which individuals with heart disease are identified. As indicated in table 3, a comparison of the top approaches and past research on predicting heart disease was done for this work using the dataset. The comparative findings demonstrated that the Bagging classifier connected with an additional tree feature selection approach as Extra Tree using Bagging as the basis classifier produced the best results.

6. Conclusions:

In this research we have used heart disease dataset with 14 attributes and 1025 instances. In order to

TABLE 3. **Representation of prediction model using heart disease**

Predicted training model on **80 %** sample dataset size of heart disease

Folds	Classifiers	Accuracy	Precision	Recall	F1-Score	ROC_AUC
K=10	GBM	91.57	61.58	37.76	45.22	72.35
	SVM	93.59	59.69	49.21	52.95	62.94
	MLP	92.49	66.65	64.90	67.21	77.73
	BCM	99.08	73.19	67.20	69.20	80.66
K=8	GBM	90.17	61.10	40.58	48.09	64.15
	SVM	89.86	70.96	71.11	70.36	82.19
	MLP	93.82	62.78	34.82	44.09	63.60
	BCM	98.68	71.85	68.85	69.65	81.56
K=6	GBM	87.70	51.86	39.63	43.69	67.28
	SVM	91.53	60.18	51.51	53.22	69.63
	MLP	87.28	41.54	41.51	40.52	60.88
	BCM	93.25	58.30	30.88	37.31	63.91
Predicted testing model on **20 %** sample dataset size of heart disease						
K=10	GBM	86.57	61.58	37.76	47.22	72.35
	SVM	82.72	41.48	36.51	37.79	62.11
	MLP	88.85	64.33	22.13	31.66	70.31
	BCM	92.62	48.44	39.63	41.89	66.82

improve the diagnosis of heart's disease, this article evaluated the performance of a number of classifiers using additional Extra Tree feature choices. Accuracy, precision, recall, and F-score were some of the assessment criteria that were employed. The accuracy of heart disease prediction increased to 99% as a consequence of the additional tree feature selection approach using Bagging, according to the data. More machine learning and deep learning techniques may be used in the future in conjunction with these feature selection method combinations. To enhance the accuracy of Heart's disease prediction, further characteristics selection techniques may be researched.

References

Abdalrada, Ahmad Shaker, et al. "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study". *Journal of Diabetes & Metabolic Disorders* 21.1 (): 251–261.

Akella, Aravind and Sudheer Akella. "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution". *Future Science OA* 7.6 (2021). 10.2144/fsoa-2020-0206.

Chauhan, Aakash, et al. "Heart Disease Prediction using Evolutionary Rule Learning". *2018 4th International Conference on Computational Intelligence & Communication Technology (CICT)* (2018). 10.1109/ciact.2018.8480271.

Fu, J, et al. "Spark-a big data processing platform for ma-chine learning." (Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)". *IEEE* (2016). 10.1109/iciicii.2016.0023.

Gangwar, S and G Pal. "Heart Disease Priediction by Stacking Ensemble Model on Multiple Classifiers by Applying Feature Selection Methods ". *JTAIT 2022* (2022): 100–123.

Gangwar, S and G K Pal. "Heart Disease Prediction by Stacking Ensemble Model on Multiple Classifiers by Applying Feature Selection Methods ". *Prediction of Cardiovascular Disease Using Feature Selection Techniques, IJCTE 2022* (): 97–103. 10.7763/ijcte.2022.v14.1316.

Gnaneswar, B and M R Ebenezar Jebarani. “A review on prediction and diagnosis of heart failure”. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (2017). 10.1109/iciiecs.2017.8276033.

Gupta, Chiradeep, et al. “Cardiac Disease Prediction using Supervised Machine Learning Techniques.” *Journal of Physics: Conference Series* 2161.1 (2022): 012013–012013.

Hasan, S, et al. “Comparative analysis of classification approaches for heart disease prediction”. *Material and Electronic Engineering (IC4ME2)* (2018): 1–1. 10.1109/ic4me2.2018.8465594.

Isabona, Joseph, et al. “Development of a Multilayer Perceptron Neural Network for Optimal Predictive Modeling in Urban Microcellular Radio Environments”. *Applied Sciences* 12.11 (2022): 5713–5713. 10.3390/app12115713.

Jindal, Harshit. “IOP Conf. Ser.: Mater. Sci. Eng”. 1022 (2021): 12072–12072.

Pandita, Aadar, et al. “Prediction of Heart Disease using Machine Learning Algorithms”. *International Journal for Research in Applied Science and Engineering Technology* 9.VI (): 2422–2429.

Rajdhan, Apurb, et al. “Heart Disease Prediction using Machine Learning”. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)* 09.04 (2020).

Ravindhar, N V, et al. “Intelligent Diagnosis of Cardiac Disease Prediction using Machine Learning”. *International Journal of Innovative Technology and Exploring Engineering* 8.11 (2019): 1417–1421. 10.35940/ijitee.J9765.0981119.

Shah, D, S Patel, and S K Bharti. “Heart Disease Prediction using Machine Learning Techniques”. *SN COMPUT. SCI.* 1 (2020): 345–345. 10.1007/s42979-020-00365-.

Singh, Deepti, Himanshu Yadav, and Chetan Agrawal. “Enumerable Learning-Based Machine Learning Techniques for Sentiment Analysis”. *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)* (2022): 270–275. 10.1109/csnt54456.2022.9787596.

Singh, Nirupma and Sonika Bhatnagar. “Machine Learning for Prediction of Drug Targets in Microbe Associated Cardiovascular Diseases by Incorporating Host-pathogen Interaction Network Parameters”. *Molecular Informatics* 41.3 (2022): 2100115–2100115. 10.1002/minf.202100115.

Truong, V T, et al. “Application of machine learning in screening for congenital heart diseases using fetal echocardiography”. *Int. J. Cardiovasc. Imaging* 2022 (): 1007–1015. 10.1007/s10554-022-02566-3.

Yadav, Dhyan, et al. “Prediction of thyroid disease using decision tree ensemble method. Human-Intelligent Systems Integration”. 2 (2020): 89–95. 10.1007/s42454-020-00006-y.

Embargo period: The article has no embargo period.

To cite this Article: Kumar Pal, Gyanendra, and Sanjeev Gangwar. “**Discovery of Approaches by Various Machine learning Ensemble Model and Features Selection Method in Critical Heart Disease Diagnosis**.” International Research Journal on Advanced Science Hub 05.01 January (2023): 15–21. http://dx.doi.org/10.47392/irjash.2023.003