International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Heart Disease Prediction using Ensemble Methods

Vaishali M Deshmukh

Abstract: ***Nowadays, people are suffering from many health issues. One of them is heart disease among the worldwide population. This causes due to imbalance lifestyle and unhealthy food consumption. The data generated by hospitals is huge and complex by nature which store patients medical and demographic information. Accurate and prompt diagnosis of heart diseases are becoming more challenging task in medical domain due to the complex data. Therefore, the computer aided systems are useful to store this complex and multivariate data to generate useful decisions. Machine learning techniques are used to classify and to predict the diseases. In this study, Majority voting classifier and Bagging ensemble method both have been evaluated. These ensemble methods combined the five base classifiers including DT (Decision Tree), LR (Logistic Regression), ANN (Artificial Neural Network), NB (Naïve Bayesian), and KNN (K-Nearest Neighbour). Bagging ensemble approach is used to combine the multiple classifiers prediction abilities for better performance. Experimental work is performed on Cleveland dataset using 14 attributes which is available online on UCI Repository. The results showed that the Bagging ensemble method is performed better to achieve higher accuracy of 87.78 %.***

Keywords: ***Bagging, Ensemble Method, Machine Learning Methods, Majority Voting Classifier***

I. INTRODUCTION

The heart disease is one of the deadly diseases. Heart failure happens when heart is not pushing the adequate blood to other body parts. Heart disease is growing rapidly among individuals due to the hectic lifestyle [1]. The diagnosis and early detection of heart disease is becoming complex task due to complex data generated by medical industries. Computer aided systems facilitate the physicians for detection of heart disease. It helps to minimize the chances of risk associated with it by prompt and early prediction of heart disease [2]. By using decision making system, medical experts will save lots of time and cost to analyze the patient's past health data, physical laboratory tests report, and related symptoms. All the above mentions tasks cause delayed and sometimes wrong diagnosis due to human interaction which may prone to errors.

Predictive modelling with the help of machine leaning methods are used to smoothen the process of patient's data analysis. Many researchers have investigated various machine learning methods: LR, KNN, ANN, DT, NB, Support Vector Machine, Fuzzy Inference System, Linear Discriminant Analysis, Rough set [4,12,19]. Some studies have evaluated ensemble or hybrid model for predictions [15,16]. Whereas few researchers have worked on clustering methods. Cleveland dataset is analyzed and evaluated in most of the studies to evaluate different classifiers with 14 attributes [8, 18].

This proposed work is carried out to build an ensemble system for heart disease detection. Different machine learning predictive classifiers like LR, KNN, ANN, DT, NB have been used in this study. The proposed model was designed for classification of whether an individual has heart disease, or he/she is healthy (binary classification). Feature importance algorithm is used to choose most corelated features which have high influence on dependent variable. Cleveland dataset for heart disease available on UCI repository is utilized for training and testing. The Work carried out in few steps as follows: 1) Data pre-processing (missing values & Standard-Scalar method) and feature selection (using Extra Tree Classifier) have been performed on dataset. 2) All base classifiers' performances have been cross validated and evaluated in terms of classification accuracy and ROC-AUC values. 3) The majority classifier is used as an ensemble method and evaluated the performance. 4) Bagging method is applied to all base classifier and evaluated the performance of each base learner. e) All Bagged classifiers are used for voting classifier to increase the performance.

Following sections discussed the work carried out in this paper. In section 2, literature review is provided with different machine learning methods evaluated for prediction. It discussed how these classifiers have performed on Cleveland and other datasets. Section 3 described the proposed model, ensemble methods and performance metrics used in this study for evaluation. Results of experiments are presented in section 4 and section 5 provided the conclusion of this work.

II. RELATED WORK

Different machine learning algorithms and ensemble techniques have been investigated for prediction of heart diseases. Some of these studies are analyzed and reviewed in this section.

Manuscript published on 30 September 2019
* Correspondence Author
Vaishali M Deshmukh*, Department of Information Science & Engineering, CMR Institute of Engineering, Bengaluru, India, vaishali.vairale@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Machine learning techniques and methodologies are widely used in healthcare domain to build expert systems. Many researchers have analyzed various machine learning methods using base classifiers, hybrid and ensemble methods to classify and to predict the cardiovascular diseases. The review revealed that various studies are carried out with single base classifiers and few studies had worked on ensemble or hybrid method to achieve the better performance.

The paper presented a novel ensemble model with "Multi-objective Weighted Voting Method" [5]. Five different classifiers: Support Vector Machines, Linear Regression, Instance-based Learner, Quadratic Discriminant Analysis, and NB are used on five different datasets of online repositories available for public use. The ensemble method achieved 84.16% accuracy.

The proposed work designed a framework with voting ensemble method for prediction of heart disease. It included NB, SVM and DT as base classifiers [4]. The heart disease dataset is obtained from repository of UCI for machine learning with 14 features. The study achieved the accuracy of 81.82% for ensemble model.

The work carried out in this study on two datasets namely Cleveland and Statlog [9]. The authors compared DT, SVM, Random forest, NB, and Linear Discriminant Analysis. Decision Tree outperformed on both the dataset with 99 % accuracy on Cleveland data and 98.15% on Stalog data.

Authors developed a prediction system with Random Forest, C4.5, ANN and NB [7]. They used genetic algorithm and chi-square test for attribute selection. These experiments are performed on Stalog with 14 attributes. The system achieved 83.7% accuracy rate with Random Forest compared with other base classifiers.

Authors proposed a hybrid approach with two association rules methods, FP-Growth technique and Apriori method to predict heart diseases [10]. They used OneR, NB, KNN, J48 and ZeroR classifiers on Cleveland dataset with 14 attributes. 99.19% prediction accuracy was achieved using Apriori with KNN hybrid method. However, 97.55% accuracy was obtained for Naive Bayes with FP-Growth algorithm.

Hidden Naive Bayes model was used in this study to predict hear disease [6]. Authors used two pre-processing methods, "Filter discretization" and "Inter-Quartile Range" on Stalog dataset with 14 features. Authors utilized Statlog dataset of heart disease available on UCI repository for experimental work. It consists of 14 attributes with 270 records. The Hidden Naïve Bayes with Inter-Quartile Range pre-processing method outperformed with 100% accuracy.

The study used Random Forest classifier for prediction of the heart disease [14]. The work carried out using Apache Spark on Cleveland dataset with 14 attributes to provide a scalable solution. Random forest outperformed than Naïve Bayes with different number of instances of 200, 400 and 600.

Authors compared Support Vector Machine (SVM), DT, NB, ANN K-NN, "Single Conjunctive Rule Learner" (SCRL), and, Radial Basis Function [17]. The Cleveland data set is used with 14 attributes. SVM achieved 84.16% of accuracy rate compared with other classifiers. SVM obtained same accuracy results after applying bagging. SCRL and Decision boosting performed well. SVM and ANN obtained higher accuracy after applying stacking ensemble method with 85.15% accuracy.

The study evaluated a hybrid approach using NB and Random Forest for diagnosis of heart disease. Cleveland dataset was used with 14 attributes [13]. The experiments were carried out using different set of features, 6,8,10 and 12 on Naïve Bayes and Random Forest. The hybrid method achieved highest accuracy compared with others.

The work was carried out using Random Forest, Sparse Logistic Regression and SVM [21]. To find hidden patients clusters, "Alternating Clustering and Classification" (ACC) and Random Forest methods were investigated. Boston Medical Center (BMC) dataset was used for experimental work using random forest technique. The results demonstrated that Random forest achieved 81.62% whereas ACC obtained 77.06% accuracy rate.

Many ensemble methods were analyzed and evaluated by researchers over past few years. It addresses the area of research to achieve accuracy and reliability at a time on healthcare data. Therefore, many studies designed expert systems with different methodologies till date. The purpose of this current work is to enhance the efficiency of base models by applying ensemble techniques to predict the heart disease data.

III. METHODS AND MATERIALS

The proposed model is developed to overcome limitations of base learners by applying an ensemble approach to five base classifiers: DT, LR, ANN, KNN and NB. Experimental setup was built on Cleveland dataset to evaluate and to validate the base classifier's performances. The dataset is publicly available on UCI data repositories to carry the experimental work. Base classifiers are compared with ensemble method to investigate the effectiveness of the proposed model.

The proposed ensemble framework has been designed and developed to classify and predict whether person is suffering from a heart disease or not. The five-machine learning base classifiers, Majority voting and bagging methods are evaluated to diagnosis the heart disease. The system has computed validation testing and evaluated the performance metrics. Fig. 1 showed the proposed model which worked in following phases:

a) Data pre-processing and feature importance

b) Building a model with base classifier, apply Majority Vote classifier and performance evaluation of ensemble majority voting classifier

c) Applying Bagging to each base classifier, performance evaluation of bagged classifiers and applying voting classifier to pool of bagged classifier

d) Comparison of results of bagged ensemble classifier and majority voting classifier.

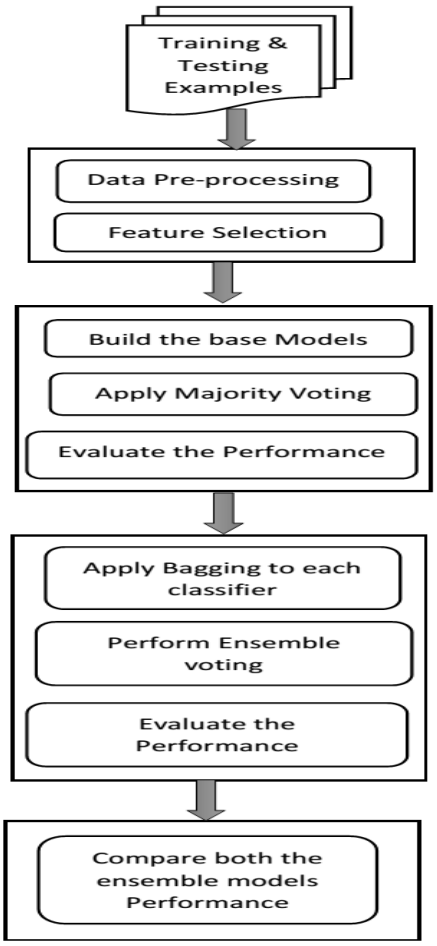


Fig.1 Proposed Ensemble Model

The following subsections had briefly discussed the dataset and methods of the proposed work.

A. Dataset

Many researchers have analyzed Cleveland dataset of heart disease which is available on UCI repository, Irvine. For designing an ensemble model, this dataset was analyzed and evaluated in this research work for detection of presence of heart disease or not. The sample size of this dataset is 303 with 76 attributes with some missing values. The dependent variable (class) has 5 values. In order to prepare this model for binary classification; other than 0 values of target variable were assigned to 1. However, there are 6 missing instances in this dataset. The proposed system analyzed 203 samples and 14 attributes of Cleveland dataset for experimental work after performing data preprocessing tasks. The complete information of 13 features and target label (class) of dataset is described in Table I.

B. Data Pre-processing and Selection of Important Features

Data pre-processing helps to clean the data and feed it to training and testing set to build machine learning models. MinMax scalar method is used to set the data of all features in a range of 0 and 1. Standard scalar technique is used to set variance 1 and mean 0 of each feature. To prepare this task as binary classification, the target label is mapped to 1 for all other values than 0. This current work classified and predicted whether an individual is suffering from heart disease (yes-1) or not (no-0).

Feature selection is an important step for model building which generates correct and reliable model using machine learning classifiers. It identifies the relevant features which improve performance of the classification methods. It also helps to increase the prediction accuracy and to reduce classifier's execution time [3,11]. Forest of trees is used to select the important features with its inter-tree variability.

Table I. Attributes of heart disease dataset

Feature #	Attribute Name	Value	Description of Feature
1	AGE	Discrete	Age (years)
2	SEX	Male-1 Female-0	Gender of patient
3	CP	Typical angina-1 atypical angina-2 non-anginal pain-3 asymptomatic-4	Type of Chest Pain
4	TRESBPS	94–200 (mmHg)	Resting Blood Pressure
5	CHOL	120–564 (mg/dl)	Serum cholesterol value
6	FBS	1- true (> 120 mg/dl) 0- false (< 120 mg/dl)	Fasting Blood Sugar
7	RESTECG	Normal-0 having ST-T -1 showing probable or definite left ventricular-2	Resting Electrocardiographic output
8	THALACH	71–202	Maximum Heart Rate
9	EXANG	Yes-1 No-0	Exercise which induced angina
10	OLDPEAK = ST	0 – 6.2	Depression imposed by exercise with relative to rest
11	SLOPE	Upsloping-1 Flat-2 down sloping-3	Slope of peak exercise-ST segment
12	CA	0-3	Number of Major Vessels (0-3) colored by fluoroscopy
13	THAL	Normal-3 fixed defect-6 reversable defect-7	Thallium scan
14	Class	Yes-1 No-0	Detection of heart disease

C. Ensemble Methods

The ensemble method is used to combine various base classifiers. It creates a meta-classifier to achieve better accuracy as compared to the performance of the single base classifier. In this study; two ensemble methods, Majority voting classifier and Bagging were analyzed. In Majority Voting classifier, same training sample set is given to the single base classifier to build the base models. Whereas, in bagging, random samples of training data with replacement are provided to each base classifier.

1. Majority Voting Classifier

In this method, target label is predicted with Majority Voting classifier by aggregating majority of predictions made by base classifiers.

The selection of test sample target label is decided with number of votes given by classifiers; specifically, more than 50%. Five classifiers were used in this study for ensemble method.

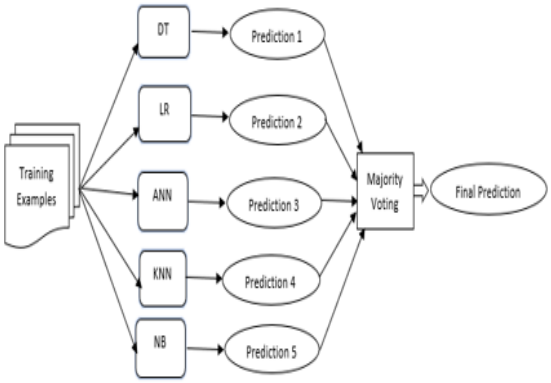


Fig. 2 Majority Voting Classifier

Fig.2 showed that each base classifier is trained and tested with full data samples. The prediction obtained by each base classifier is passed to Majority voting classifier to draw the final prediction.

2. Bagging Ensemble Method

This technique aggregates the predicted outcome of n-base classifies. K-bootstrap set of instances are passed to each base classifier to build the model [20]. Bagging is simplest ensemble method than boosting and stacking. Bagging algorithm works as follows:

Input: D- set of instances, d- one single instance, k-Bootstrap sample sets

Procedure for bagging:

- For each i from 1 to k with training sample set D_i with d instances from original dataset is sampled along with replacement.
- Classifier C_i build the model with training bootstrap sample set D_i.
- Every classifier is predicted a label for test instance and do the voting for the same.
- Voting Classifier collects all the votes provided by bagged classifier and perform majority voting to make the final prediction.

Output: predict heart disease; 0 for absence and 1 for presence

In this work, bagging is performed on five classifiers as shown in Fig.3. Classifiers: DT, LR, ANN, KNN, and NB had used the bootstrap samples of original dataset. All bagged classifiers had predicted the target label for testing data instance. Ensemble voting classifier had performed majority voting to draw a final prediction.

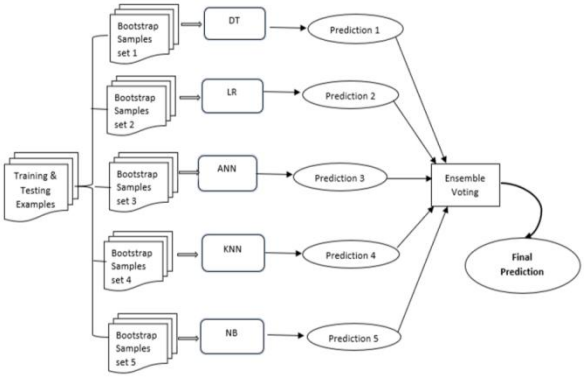


Fig.3 Bagging Ensemble Method

As shown in Fig.3, a bootstrap sample set was assigned to individual classifier to predict a target label. Then, the prediction of each base learner was involved in the voting scheme. The final label was predicted with aggregation of votes [22].

***D.* Performance Evaluation Metrics**

In this experimental setup, classifiers' performance has been checked with various performance evaluation metrics. Classification report was used to provide Precision, Recall and F1-score of base classifiers, Majority Voting and Bagging classifier. Accuracy of all classifiers was evaluated by Accuracy score matrix. ROC-AUC score was used to find the probability of classifier.

There are four following terms which are used in performance metrics:

- True Positives: It is score when actual and predicted both outputs are YES.
- True Negatives: It is a score when actual and predicted both outputs are NO.
- False Positives**:** It is the score when predicted output is YES; but actual output is NO.
- False Negatives: It is a score when predicted is NO; but actual output is YES.

1. Classification Accuracy

Accuracy is measured in terms of ratio of number of correctly predicted instances to total instances in a given dataset.

2. Area Under Curve

AUC-Area Under Curve is widely used evaluation metric. AUC is an aggregate score of performance among given classification thresholds values. AUC measures the probability where model ranks higher random positive instance verses random negative instance.

3. ROC curve

ROC-Receiver Operating Characteristic curve is used to represent a graph which shows performance of classifiers at all its classification threshold values. ROC curve is plotted between two values; true positive and false positive rate values.

4. Precision

It is a score of ratio of correctly identified positive outputs to the all positive predicted outcomes by the classifier.

5. Recall

It is a score of ratio with number of correctly predicted positive outputs to the total number of all related samples (all instances which are identified as positive samples).

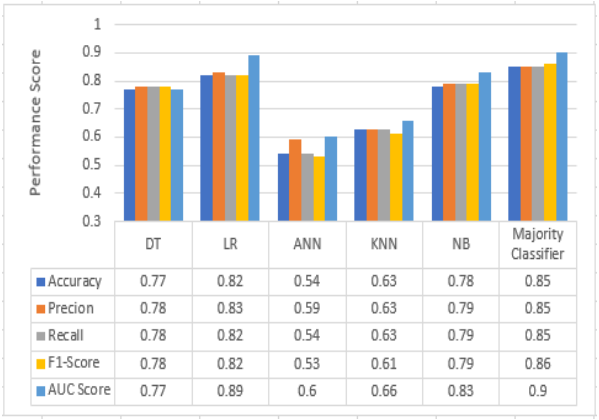
6. F1 Score

It is a harmonic mean between precision score and recall score. It's value ranges between 0 to 1. If F1 score is more, better is the model's performance.

IV. RESULT AND DISCUSSIONS

This section discussed the performance evaluation of base classifiers and ensemble algorithms used in this paper. First, the performance of different base machine learning algorithms DT, LR, ANN, KNN and NB was checked on Cleveland dataset of heart disease with important features. Normalized and standardized features were used for classification. Performance of classifiers was cross validated by using K-fold method.

As shown in Fig.4, LR performed well among other four classifiers with high accuracy score, precision , recall and ROC-AUC score. The predictive accuracy of LR was 82%, and for NB, it was 78%, and next DT with 77%. It was observed that ANN had poor performance than other classifiers. After applying Majority Voting Classifier, the accurcy was 85%; increased by 3% than LR. Results showed that encemble Majority voting classifer achieved better performace as compared with individual base classifiers.



	DT	LR	ANN	KNN	NB	Majority Classifier
Accuracy	0.77	0.82	0.54	0.63	0.78	0.85
Precion	0.78	0.83	0.59	0.63	0.79	0.85
Recall	0.78	0.82	0.54	0.63	0.79	0.85
F1-Score	0.78	0.82	0.53	0.61	0.79	0.86
AUC Score	0.77	0.89	0.6	0.66	0.83	0.9

Fig 4. Performance Evaluation of Base & Majority Voting Classifiers

AUC values and ROC curve of different classifiers were shown in Fig.5. AUC score for LR was more than other base classifiers. Majority Voting Classifiers improved the AUC score as compared with other base classifiers.

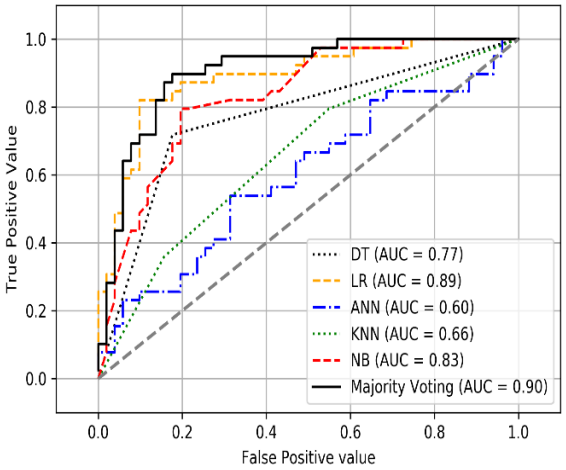
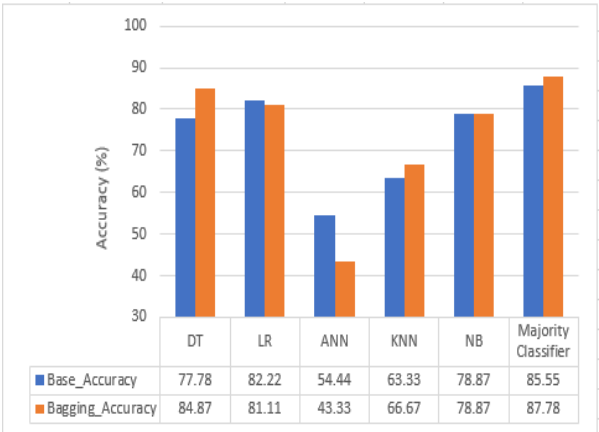


Fig. 5 ROC-AUC Curve for Base & Majority Voting Classifies

The classification accuracy of each base method after applying bagging classifier was given in Fig.6. DT had achieved 85% accuracy score; where as KNN performed well with bagging classifier. LR and ANN had shown decline performance in terms of accuracy with bagging classifiers. The accuracy score of Majority Voting Classifier was increased after combining all bagged base classifiers.



	DT	LR	ANN	KNN	NB	Majority Classifier
Base_Accuracy	77.78	82.22	54.44	63.33	78.87	85.55
Bagging_Accuracy	84.87	81.11	43.33	66.67	78.87	87.78

Fig. 6 Accuracy Comparison between Bagging and Majority Classifier

V. CONCLUSION

The proposed work compared the Majority Voting classifier's performance with base learners and bagging method. Base classifiers prepared the models with full dataset and then Majority Voting classifier was applied for final prediction. However, in bagging, bootstrap samples of dataset were provided to each base classifier. Voting scheme was used by aggregating the votes of all base bagged classifiers. The performances of both were evaluated with performance metrics like accuracy and AUC score. Cleveland dataset for heart disease was tested on five base classifiers namely DT, LR, ANN, KNN and NB.

The other important aspect of this work was feature selection using extra tress classifier. Ensemble methods had improved the performance of the base classifiers using Bagging and Majority Voting classifiers. Performance metrics like accuracy, AUC, Precision, Recall and F1 score are evaluated on dataset with base and ensemble method. The system achieved 87.78% accuracy as compared with others. Bagged classifier with majority voting obtained the better results. The further enhancement for this work will be provided by using optimization methods and feature extraction techniques.

REFERENCES

1. J. López-Sendón, "The heart failure epidemic", Medicographia, vol. 33, pp. 363–369, (2011).
2. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," Nature Reviews Cardiology, vol. 8, no. 1, pp. 30–41, (2011).
3. T. John Peter and K. Somasundaram, "Study and Development of Novel Feature Selection Framework for Heart Disease Prediction", International Journal of Scientific and Research Publications, Volume 2, Issue 10, (2012).
4. A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early heart disease prediction using data mining techniques," in *Proceedings of Computer Science & Information Technology (CCSIT-2014)*, vol. 24, pp. 53–59, Sydney, NSW, Australia, (2014).
5. S. Basir, U. Qamar U, F.H. Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Australas Phys Eng Sci Med. 2015 Jun;38(2):305-23, (2015).
6. M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden Naive Bayes classifier," International Conference on Circuits, Controls, Communications and Computing (I4C), pp. 1-5. (2016).
7. M.A. Jabbar, B.L.Deekshatulu and Priti Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing, Volume 4, pp.175-184, (2016).
8. S. Radhimeenakshi, "Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network," 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 3107-3111, (2016).
9. I.A. Zriqat, A. M. Altamimi, M. Azzeh, " A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods", International Journal of Computer Science and Information Security (IJCSIS), Vol. 1, (2016).
10. J. Singh, A. Kamra and H. Singh, "Prediction of heart diseases using associative classification," 5th International Conference on Wireless Networks and Embedded Systems (WECON), Rajpura, 2016, pp. 1-7.
11. S. A. Sabab, M. A. R. Munshi, A. I. Pritom and Shihabuzzaman, "Cardiovascular disease prognosis using effective classification and feature selection technique," International Conference onMedical Engineering, Health Informatics and Technology (MediTec), pp. 1-6, (2016).
12. T. Karayılan and Ö. Kılıç, "Prediction of heart disease using neural network," International Conference on Computer Science and Engineering (UBMK), pp. 719-723, (2017).
13. K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features," 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), pp. 500-504 (2017).
14. R. G. Saboji, "A scalable solution for heart disease prediction using classification mining technique," International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 1780-1785, (2017).
15. J. Zhang et al., "Coupling a Fast Fourier Transformation with a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment," in IEEE Access, vol. 5, pp. 10674-10685, (2017).
16. R. Li, W. Liu, Y. Lin, H. Zhao, C. Zhang, "An Ensemble Multilabel Classification for Disease Risk Prediction" Journal of Healthcare Engineering Volume 2017, Hindawi, Article ID 8051673, (2017).
17. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," IEEE Symposium on Computers and Communications (ISCC), pp. 204-207, (2017).
18. O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," Expert Systems with Applications, vol. 68, pp. 163–172, (2017).
19. S. Nazir, S. Shahzad, and L. Septem Riza, "Birthmark-based software classification using rough sets," Arabian Journal for Science and Engineering, vol. 42, no. 2, pp. 859–871, (2017).
20. I. Yekkala S. Dixit ; M. A. Jabbar , "Prediction Of Heart Disease Using Ensemble Learning And Particle Swarm Optimization", **:** International Conference On Smart Technologies For Smart Nation (SmartTechCon), IEEE Explorer, (2017).
21. T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams and I. C. Paschalidis, "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach," in Proceedings of the IEEE, vol. 106, no. 4, pp. 690-707, (2018).

AUTHORS PROFILE

Vaishali M Deshmukh is currently working with Department of Information Science & Engineering, CMRIT, Bengaluru, India. She has completed her master's degree in Computer engineering from Pune university. She is pursuing her PhD in Computer Science & Engineering from CHRIST (Deemed to be University), Bengaluru, India. Her research interest areas are Data Science, Machine Learning, Soft Computing and Big Data Analytics.