International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Ensemble Models for Classification of Coronary Artery Disease using Decision Trees

Pratibha Verma

Abstract*: The foundation of data mining techniques using decision tree methods played a crucial role in the identification and classification of diseases. In the utilization of decision tree classifiers to develop the robust classifier for classification of Coronary Artery Disease data set namely Z-Alizadeh Sani and extension Z-Alizadeh Sani. We have used three decision tree techniques Random Forest (RF), Classification and Regression Tree (CART), J48 (C4.5) and made two ensemble models. These ensemble models have different combining rules like voting and stacking. The Voting Scheme model Vote (J48, RF, CART) and stacking Scheme model Stack (J48, RF, CART) have our proposed model. The findings are compared in individual and ensemble models classifier with 5-Fold Cross-Validation and 10-Fold Cross-Validation. The finding of the proposed ensemble models can be used in the detection and evaluation of Coronary Artery Disease (CAD).

***Keywords*: Random Forest, CART, C4.5, ensemble model, Coronary Artery Disease.**

I. INTRODUCTION

Cardio Vascular diseases (CVD) are among the most common causes of life defeat in the world. A major type of these diseases is Coronary Artery Disease (CAD)[1]. Twenty-five percent of people, who have CAD problem, suddenly dies abruptly beside any previous symptoms [2]. CAD is one of the most life threating diseases affecting the heart and can lead to serious heart problems in patients. Early and timely treatment of disorder symptoms can be useful and can reduce the severity of the side effects of the disease. Medical and healthcare data dealing with the healthcare problem like Coronary artery disease (CAD) may show numerous results when applying the same machine learning technique[3]. Data mining is the method of finding previously unknown patterns and developments in databases and using that information to make predictive models. In healthcare, Data mining is an area of high significance and has become gradually more effective and essential [4]–[6]. The classification performances are based mainly on the efficiency of medical diagnosis and analysis [3]. Data mining provides a group of tools and techniques that can be useful to reach these goals.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Pratibha Verma *, Ph.D. Scholar, Department of Computer Science, Dr. C.V. Raman University, Bilaspur (C.G.), India. Email: bhilai.pratibha@gmail.com

The data mining based the decision tree methods have a crucial task to the identification and classification of diseases. To develop the robust classifier for classification of Coronary Artery Disease (CAD) data set namely Z-Alizadeh Sani and extension Z-Alizadeh Sani, decision tree classifiers has been utilised.

Classification is a technique that allowed comparative opportunities to carry out the treatment of diseases in the healthcare fields. In this study, we have developed a strong model with the help of classification techniques for predicting disease and diagnosis of CAD. The classifier [7] technique such as ensemble is a group of classifiers whose individual decisions are combined to classify new examples differently. In this paper, we have proposed the common style and scheme for classifier combinations. There are many existing models that can be considered as special membrane classification situations where all pattern representations are used together to make decisions[8]. We have derived widely used classifier combination schemes such as voting and stacking under different assumptions and using different approximations. Different combination schemes or ensemble models of classifiers have been experimentally compared. In this study we used three decision tree techniques namely Random Forest (RF), Classification and Regression Tree (CART), J48 (C4.5) tree and their ensemble techniques based on Voting Scheme model Vote (J48,RF,CART) and Stacking Scheme model Stack (J48, RF,CART). Then we compared the proposed models based on the classification accuracy, sensitivity, specificity and F1-Score with 5 Fold and 10 –Fold Cross Validation. The proposed ensemble models are more quickly identifying and screening system that will also assist in early detection of CAD. These models will encourage improved patient caring method with limited resources.

II. LITERATURE REVIEW

The Coronary Artery Disease (CAD) is a high risk disease, several researchers have attempted to classify and predict on it or derive significant risk factors. The researchers used data mining and machine learning approach to work on the CAD risk factors and they obtained significant results. Several researchers have been focusing on data mining and machine learning related to the specific heart and CAD diseases.

The authors [9] worked to use and integrate bioinformatics tools to compare proteomic data from uses of different conditions. A proficient method using the pathological correspondence between Aortic Valve Stenosis and Coronary Artery Disease as a case-study.

The [10] have analyzed three available data sets: heart disease databases, South African heart disease and Z-Alizadeh Sani data set. They worked on two ways for a predictive analysis based on Decision Trees, Naive Bayes, Support Vector Machines and Neural Networks. Then the descriptive analysis support on rules of association and decision. Finally compared own finding to existing finding and conclude.

The authors [11]have worked on the cardiovascular dataset with various methods Logistic regression RNA, NN, FST like Relief-F, Independent t-test analysis. The outcomes show that the 84.1% NN accuracy obtained with the oversampled dataset and the Relief-f feature selection method. The authors [3]worked on the integration of the results of the machine learning analysis on various data sets targeting CAD disease. They have used fast Decision Tree and prune C 4.5 tree are applied, where the resulting trees are extracted from different data sets and compared to them. The general characteristics among these data sets are extracted and used data for the same disease in any data set. The finding was show that the classification performance his collected datasets is 78.06% is better than different datasets. The authors [12] have analysed of Coronary Artery Disease (CAD). They used three data mining techniques to identify CAD. The algorithms like C 4.5, K - Nearest Neighbor (KNN) and Naive Bayes have been utilized for the classification.

III. METHODOLOGY

Classification algorithms require a learning period on a collection of labeled data, which allow decision-making on the absent class label of a test record. During the learning period, a classification models created for the prediction of the class label of a data record via the values of its features. In this section we have used three decision trees namely Random Forest (RF), Classification and Regression Tree (CART) and C4.5 (J48) and its ensemble model for classification of CAD dataset.

A. J48

C 4.5 or J48 is a classification algorithm. We also know it by other names like j48 Which is a decision tree made by Ross Quinlan's. The C3 4.5 tree Quinlan's is an extension of the ID3 tree. The decision tree made by C. 4.5 can be used for classification, and this reason, C4.5 is often known as a statistical classifier [4], [13].

B. Random Forest

Random forest is a joint learning technique widely used for classification and regression tasks [4], [5]. For each tree in the forest, an initial sample is selected from the original data. The bootstrap sample is obtained by randomly selecting the instances of the original data with the replacement and has the same dimension as the original data set. Next, a decision tree is developed to the maximum possible extent without pruning in the bootstrap example using a modified decision tree learning algorithm [14]. The tree learning algorithm is modified as follows: in each node, the best division is selected by examining a random subset of features rather than the complete set of characteristics [15]. Since the decision of the best division is the most costly aspect of the learning process from the computational point of view, the choice of a subset of features will drastically accelerate tree learning. Once all the trees have been built in this way, the final forecasts are obtained by calculating the average of the individual tree predictions.

C. CART

Classification and Regression Trees (CART), introduced by Breiman Leo (Breiman Leo; Friedman J.H.; Olshen R.A.; Stone, 1987), is a statistical method that can choose from an enormous number of descriptive variables (x) those that are most significant in deciding the response variable (y) to be explained. This is done by growing a tree structure, which partitions the data into mutually exclusive groups (nodes) each as pure or homogeneous as possible concerning their response variable [17]. Such a tree begins with a root node containing every one the objects, which are separated into nodes by recursive binary splitting. Each split is defined by a straightforward rule based on a single explanatory variable [17].

D. Ensemble Model

When we add two or more classification techniques, it is called an ensemble model [18], [19]. If we indicate that certain experts are more qualified than others, weighting the decisions of those qualified experts more heavily may further improve the overall [18].

- **Voting** is a technique for combining multiple classifiers. The voting ensemble technique gives a baseline scheme for combining classifiers by averaging their probability estimates (classification) or numeric predictions (regression). We have used for this study i.e. classified the dataset based on combinations of probability in the classifiers[8], [20].
- **Stacking** is Combines several classifiers using the stacking method. Stacked generalization works by deducing the biases of the generalizer (s) concerning a provided learning set [19]. Combine the classifiers using stacking for both classification and regression problems. Stacking [21] introduces the concept of a Meta-learner, which replaces the voting procedure. The problem with voting is that it's not clear which classifier to trust. Stacking tries to learn which classifiers are the reliable ones, using another learning algorithm— the Meta-learner—to discover how best to combine the output of the base learners. You specify the base classifiers, the Meta-learner, and the number of cross-validation folds.

E. Dataset and Tools

This experiment have used software package Waikato Environment for Knowledge Analysis (WEKA). The data set available on the UC Irvine Machine Learning Repository. The name of data set Z-Alizadeh Sani [22] and extension Z-Alizadeh Sani [23] hold the records of 303 patients, each of which has 54 features (attributes) and 59 features (attributes).

IV. PROPOSED ARCHITECTURE

In figure 1, we have presented the architecture of our proposed model have used to classify Coronary Artery Disease. This study [18], [19] re-examine situation under which ensemble based strategy might be more advantageous than their classifier partner, methods for producing individual segments of the ensemble systems, and a variety of procedures during which the individual classifiers can be joined.

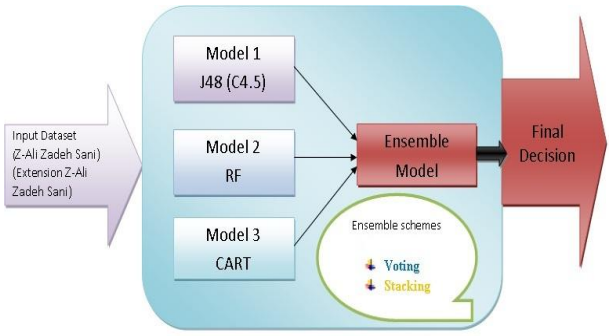

Fig. 1.Proposed Architecture.

Figure 1 presented architecture has been three J48 (C4.5), RF and CART decision tree models have to classify two datasets the Z-Alizadeh Sani and extension Z-Alizadeh Sani. Then combining it or together with the decision tree based on combining schemes. We have proposed two ensemble models Vote (J48, RF, CART) and Stack (J48, RF, CART).

V. RESULT AND DISCUSSION

This part shows the outcomes obtained in the three decision tree (J48, RF, CART) and two proposed different approaches ensemble model (Vote J48, RF, CART and Stack J48, RF, CART). The experiment was conducted in data partitions Cross-Validation techniques. It assists us to make better use of our data, and it provides us a good deal about our algorithm performance. This is sometimes easy in difficult machine learning models, no need to extra concentration and uses the identical data at various segment of the pipeline.

In the result, we have an experimental setup using 5-Fold Cross Validation (5-FCV) and 10 -Folds Cross Validation (10-FCV) technique for the data partition. In this part we calculate the performance of classifiers based on the confusion matrix:

Table- I: Confusion matrix

Actual Vs. Predicted	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Sensitivity}=\frac{TP}{TP+FN} \quad (2)$$

$$\text{Specificity}=\frac{TN}{TN+FP} \quad (3)$$

$$\text{F1-Score}=\frac{2TP}{(2TP+FP+FN)} \quad (4)$$

Table- II: Performance of models using 5-Fold Cross Validation (In %)

Z-Alizadeh Sani dataset				
Algo	**Acc.**	**Sen.**	**Spec.**	**F1-Score**
J48	80.86	87.04	65.52	86.64
RF	82.18	96.3	47.13	88.52
CART	82.51	87.04	71.27	87.65
Vote (J48,RF, CART)	82.51	90.75	62.07	88.09
Stack (J48,RF, CART)	**83.17**	**92.13**	**60.92**	**88.65**

The table II shows that the proposed ensemble model Vote (J48, RF, CART) and Stack (J48, RF, CART) gives higher accuracy compared to individual classifier models J48, RF and CART. The accuracy of our proposed models is high in 5-FCV. The highest accuracy obtained by the ensemble model Stack (J48, RF, CART) has 83.17%. The second highest accuracy obtained by the ensemble model Vote (J48, RF, CART) has 82.51% and individual model RF CART have 82.51%. Similarly, the ensemble model Stack(J48, RF, CART) gives higher F1-Score 88.65% compared to individual classifier models J48, RF and CART. The individual models and ensemble models have a fluctuated sensitivity, specificity.

Table- III: Performance of models using 10-Fold Cross Validation (In %)

Z-Alizadeh Sani dataset				
Algo	**Acc.**	**Sen.**	**Spec.**	**F1-Score**
J48	79.21	87.96	57.47	85.78
RF	83.17	95.37	52.87	88.98
CART	83.50	90.74	65.52	88.69
Vote (J48,RF, CART)	84.16	92.13	64.37	89.24
Stack (J48,RF, CART)	**84.82**	**93.52**	**63.22**	**89.78**

The table III exhibit that the proposed ensemble models Vote (J48, RF, CART) and Stack (J48, RF, CART) give higher accuracy compared to individual classifier models J48, RF, CART. The accuracy of our proposed models is excessive instances in 10-FCV with Z-Alizadeh Sani dataset. The best possible accuracy obtained by the ensemble model Stack (J48, RF, CART) has 84.82%. The second best possible accuracy acquired via the ensemble model Vote (J48, RF, CART) has 84.16%. Similarly, the ensemble model Stack (J48, RF, CART) has 89.78% and Vote (J48, RF, CART) has 89.24% gives higher F1-Score compared to individual classifier models J48, RF, CART.

The individual models and ensemble models have a fluctuated sensitivity, specificity.

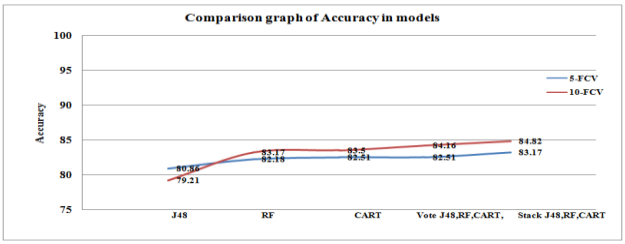

Fig. 2.Accuracy of extension Z – Alizadeh Sani dataset in 5-FCV and 10-FCV (in %).

The figure 2 vertical axis corresponds to the accuracy of models in a 5-FCV and 10-FCV with Z-Alizadeh Sani dataset. The horizontal axis corresponds to the name of models with related accuracies. In case 10-FCV the all individual model and ensemble models are better performed compare to the 5-FCV. In all the cases the proposed ensemble models are better performed compared to the individual models.

Table - IV: Performance of models using 5-Fold Cross Validation (In %)

Extension of Z-Alizadeh Sani dataset				
Algo	**Acc.**	**Sen.**	**Spec.**	**F1-Score**
J48	99.67	100	98.86	99.77
RF	94.39	98.15	85.06	96.15
CART	97.7	99.54	93.11	98.41
Vote (J48,RF,CART)	97.36	99.08	93.11	98.17
Stack (J48,RF,CART)	99.67	99.54	100	99.77

The table IV shows that the Individual model J48 and ensemble model Stack (J48, RF, CART) offers higher accuracy in contrast to Single classifier and ensemble model. Similarly, the Individual model J48 and ensemble model Stack (J48, RF, CART) offers higher F1-Score compared to individual and ensemble models. The single models and ensemble models have a fluctuated sensitivity, specificity.

Table - V: Performance of models using 10-Fold Cross Validation (In %)

Extension of Z-Alizadeh Sani dataset				
Algo	**Acc.**	**Sen.**	**Spec.**	**F1-Score**
J48	99.67	100	98.85	99.77
RF	94.39	98.15	85.06	96.14
CART	99.67	100	98.85	99.77
Vote (J48,RF, CART)	**99.67**	**100**	**98.85**	**99.77**
Stack (J48,RF, CART)	**99.67**	**100**	**98.85**	**99.77**

The table V show that the Individual model J48, CART and ensemble model Vote (J48, RF, CART), Stack (J48, RF, CART) offers higher accuracy in contrast to Single classifier and ensemble model. Similarly, the Individual model J48, CART and ensemble model (Vote J48, RF, CART), (Stack J48, RF, CART) offers higher F1-Score compared to individual and ensemble models. The single models and ensemble models have a fluctuated sensitivity, specificity.

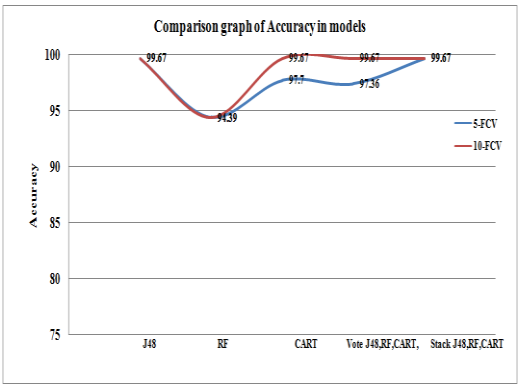

Fig. 3.Accuracy of extension Z – Alizadeh Sani dataset in 5-FCV and 10-FCV (in %).

The figure 3 vertical axis corresponds to the accuracy of models in a 5-FCV and 10-FCV with extension Z-Alizadeh Sani dataset. The horizontal axis corresponds to the name of models with related accuracies. In case 10-FCV the all individual model and ensemble models are better performed compare to the 5-FCV. In all the cases the proposed ensemble models are better performed compared to the individual models.

VI. COMPARISON OF PROPOSED MODEL WITH EXISTING MODELS

Table VI shows that the comparative analysis of the proposed model and existing model in terms of accuracy. The outcome demonstrates that the proposed model significantly improves CAD classification accuracy. The highest accuracy was obtained by the proposed ensemble models.

Table - VI: Accuracy of the proposed model compared with existing models

Name of author and year	Dataset	Algorithms	Outcomes
Our Proposed model	Extension Z-Alizadeh Sani	J48, RF, CART, Vote (J48, RF, CART), Stack (J48, RF, CART,) with 5-FCV and 10-FCV	J48, Ensemble model Vote (J48,RF,CART) and Stack (J48,RF,CART) **(99.67%)** with 10-FCV
Our Proposed model	Z-Alizadeh Sani	J48, RF, CART, Vote (J48, RF, CART), Stack (J48, RF, CART,) with 5-FCV and 10-FCV	Ensemble model Vote J48,RF,CART**(84.16%)**, Stack J48,RF,CART **(84.82%)**with 10-FCV
Amutha, Padmajavalli, & Prabhakar (2018) [24]	clinical records of 1000	Decision Tree algorithm, another adjusted variant of K-Nearest Neighbor (KNN) algorithm, K-Sorting and Searching (KSS)	The framework utilized six clinical attributes. AN mobile app "TMT Predict", enhanced to 84%.
Bektas, Ibrikci, & Ozcan (2017) [11]	Cardiovascular dataset	Logistic regression RNA, NN, FST like Relief-F, independent t-test analysis,	84.1% NN with the oversampled dataset and the Relief-f feature selection method
Alizadehsani et al., (2013) [2]	Z-Alizadeh Sani dataset	Bagging and C4.5 classification	79.54 (C4.5)
Alizadehsani et al., (2012) [12]	Z-Alizadeh Sani dataset	SMO, Naïve Bayes, C4.5 and AdaBoost.	79.86 SMO

VII. CONCLUSION

The proposed two ensemble models Vote (J48, RF, CART) and Stack (J48, RF, CART) which are created by adopting two different types of ensemble learning method. It is an advanced ensemble learning method has helpful for identification technique which detects CAD problem quickly. The outstanding result obtained the experiment of the ensemble learning rule in both data partition techniques 5-FCV, 10-FCV and both Z-Ali Zadeh Sani, extension Z-Ali Zadeh Sani datasets. This ensemble models successfully classified the Coronary Artery Disease (Z-Ali Zadeh Sani Dataset) with accuracies of 84.82% using the ensemble model Stack (J48, RF, CART) in cases of 10-FCV. Similarly, the ensemble models effectively classified the Coronary Artery Disease (extension Z-Alizadeh Sani Dataset) with accuracies of 99.67% using the ensemble model Vote (J48, RF, CART) and 99.67% using the ensemble model Stack (J48, RF, CART) in cases of 10-FCV.The accuracies of 99.67% using the ensemble model (Stack J48, RF, CART) in case 5-FCV. Finally, we have two ensemble techniques such as Voting and stacking were compared experimentally.

REFERENCES

1. M. M. Al-nozha, H. M. Ismail, and O. M. Al Nozha, "Coronary artery disease and diabetes mellitus," J. Taibah Univ. Med. Sci., vol. 11, no. 4, pp. 330–338, 2016.
2. R. Alizadehsani, J. Habibi, M. Javad, B. Bahadorian, and Z. Alizadeh, "A data mining approach for diagnosis of coronary," Comput. Methods Programs Biomed., pp. 1–10, 2013.
3. R. El-bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets," Procedia - Procedia Comput. Sci., vol. 65, no. Iccmit, pp. 459–468, 2015.
4. J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, Third. Elsevier, 2012.
5. A. Pujari, Data mining techniques, Third. University press, 2013.
6. N. Bhatla and K. Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques," vol. 1, no. 8, pp. 1–4, 2012.
7. K. R. Lakshmi and S. P. Kumar, "Utilization of Data Mining Techniques for Prediction and Diagnosis of Major Life Threatening Diseases Survivability-Review," Int. J. Sci. Eng. Res., vol. 4, no. 6, pp. 923–932, 2013.
8. J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," vol. 20, no. 3, pp. 226–239, 1998.
9. F. Trindade, R. Ferreira, B. Magalhães, A. Leite-moreira, I. Falcão-pires, and R. Vitorino, "How to use and integrate bioinformatics tools to compare proteomic data from distinct conditions? A tutorial using the pathological similarities between Aortic Valve Stenosis and Coronary Artery Disease as a case-study," J. Proteomics, 2018.
10. F. Babič, J. Olejár, Z. Vantová, and J. Paralič, "Predictive and Descriptive Analysis for Heart Disease Diagnosis," no. September, pp. 155–163, 2017.
11. J. Bektas, T. Ibrikci, and I. T. Ozcan, "Classification of Real Imbalanced Cardiovascular Data Using Feature Selection and Sampling Methods: A Case Study with Neural Networks and Logistic Regression," Int. J. Artif. Intell. Tools, vol. 26, no. 06, p. 1750019, 2017.
12. R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, and A. Ghandeharioun, "Diagnosis of Coronary Arteries Stenosis Using Data Mining," vol. 2, no. July, pp. 57–65, 2012.
13. S. E. E. Profile and S. E. E. Profile, "ForEx ++ : A New Framework for Knowledge Discovery from Decision Forests ForEx ++ : A New Framework for Knowledge Discovery from Decision Forests," no. September, 2017.
14. R. Nagalla, P. Pothuganti, and D. S. Pawar, "Analyzing Gap Acceptance Behavior at Unsignalized Intersections Using Support Vector Machines, Decision Tree and Random Forests," Procedia Comput. Sci., vol. 109, no. 2016, pp. 474–481, 2017.
15. S. Yang, W. Wang, Y. Jiang, J. Wu, and S. Zhang, "What contributes to driving behavior prediction at unsignalized intersections ?," Transp. Res. Part C, vol. 108, no. February 2018, pp. 100–114, 2019.
16. C. J. Breiman Leo; Friedman J.H.; Olshen R.A.; Stone, Classification and Regression Tree. Wadsworth and brooks Cole advanced books and software, 1987.
17. F. Questier, R. Put, D. Coomans, B. Walczak, and Y. Vander Heyden, "The use of CART and multivariate regression trees for supervised and unsupervised feature selection," Chemom. Intell. Lab. Syst., vol. 76, no. 1, pp. 45–54, 2005.
18. R. Polikar, "Ennsemble Based Systems," IEEE Circuits Syst. Mag., vol. 6, no. 3, pp. 21–45, 2006.
19. D. Wolpert, "Stacked Generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
20. Kuncheva Ludmila I., Combining Pattern Classifiers, vol. 47, no. 4. 2005.
21. I. H. Witten, E. Frank, and M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," ACM SIGMOD Rec., no. May, 2011.
22. "Z-Alizadeh Sani dataset," 2016. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/extention+of+Z-Alizadeh+sani+dataset.
23. "extension Z-Alizadeh Sani," 2016. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani#.
24. A. J. Amutha, R. Padmajavalli, and D. Prabhakar, "A novel approach for the prediction of treadmill test in cardiology using data mining algorithms implemented as a mobile application," Indian Heart J., vol. 70, no. 4, pp. 511–518, 2018.
25. R. Alizadehsani et al., "PT," Knowledge-Based Syst., 2016.

AUTHORS PROFILE

Pratibha Verma is currently a Ph.D. candidate in Department of Information Technology at Dr. C.V. Raman University, Bilaspur (Chhattisgarh), India. She received her master's Degree in Computer Application from Chhattisgarh Swami Vivekananda Technical University Bhilai (Chhattisgarh), India in 2012. Her research interests include machine learning and data mining.