An Efficient Ensemble Classifier for Heart Disease Diagnosis and early Prediction

S. Brindha, T. Ajisha

Abstract: ***Heart Disease is one of the most significant causes of mortality in the world today. Prediction and Diagnosis of Cardiovascular disease is considered as one of the major challenges in the Medical Field especially for Cardiologists. Artificial Intelligence and Machine learning (ML) was popularly employed for pattern prediction and it was noticed that these Intelligent Mechanisms were used in Medical Feld for better Heart Disease Pattern Prediction. Thus more researchers were focusing Machine Learning based Data Mining Classifiers for Heart Disease Pattern Prediction and Diagnosis in the healthcare Industry especially for Cardiologists. This research work identified the recently proposed Hybrid Random Forest with a Linear Model (HRFLM) Classifier for improving the classification accuracy for the cardiovascular disease patterns prediction well in advance and Diagnosis as well. However, it was noticed that for improving the performances better in terms of Accuracy, Sensitivity, Specificity, Precision, FScore and False Positive Rate FPR, needed an efficient classifier. Thus this work developed and implemented an efficient Classifier ensemble Nu-SVC Classifier and Weighted Random Forest Classifier. From the experimental results, it was noticed that the proposed Ensemble Classifier performs better as compared with that of existing Hybrid Classifier in terms of in terms of Accuracy, Sensitivity, Specificity, Precision, FScore and False Positive Rate FPR***

Keywords **: *Support Vector Classification, Weighted Random Forest, Machine Learning, Artificial Intelligence, and Heart Disease Pattern Prediction.***

I. INTRODUCTION

The prediction of Heart Disease is one of the challenging and toughest tasks as Heart Disease was influenced by different factors such as Cholesterol, Obesity, Depression and Anger, Blood Pressure, Diabetes etc.[1,13,17] There were different Techniques proposed for finding out Heart Diseases. A few are Genetic Algorithm, Decision Tree, K-Nearest Neighbor, Particle Swam Optimization, and Support Vector Machine. The severity of the Heart Disease could be classified by the above mentioned Mining Techniques.

As Heart Diseases are difficult and complex to predict, it is needed to analyze carefully for better Heart Disease prediction. If the system failed to predict Heart Diseases effectively, it may cause death.

Revised Manuscript Received on October 20, 2020.

***** Correspondence Author

S.Brindha*, Computer Science and Applications, St.Peter's Institute of Higher Education and Research, Avadi, Chennai, Tamil Nadu, India. Email: brindha07.future@gmail.com

T.Ajisha, Computer Science and Applications, St.Peter's Institute of Higher Education and Research, Avadi, Chennai, Tamil Nadu, India. Email: ajisha1995@gmail.com

The prime objectives of Heart Diseases Prediction Methods are to predict well in advance so that the premature death can be avoided. Thus, from the literature survey, it was noticed that the Data Mining Techniques particularly Classifiers and Clusters are employed by Researchers for better heart prediction and analysis. The rest of the paper is structured and presented as follows. Related Works and Survey were described in Section II, where, recently introduced a few Classifiers discussed. In Section III, the existing and recently proposed Hybrid Random Forest with Linear Model (HRFLM) was narrated. The major issues of Hybrid Random Forest with Linear Model (HRFLM) were discussed and justified the need of the proposed Ensemble Classifier SVC-WRF in Section IV. The proposed Ensemble Classifier and its detailed Procedure were described in Section V. This work highlights the strength of the proposed model in Section VI and Conclusion was presented in Section VII.

II. RELATED WORKS

Highest Prediction Accuracy in Medical Field is an important demand and to satisfy the same, Artificial Neural Network based Models were proposed. The purposes of these models were for achieving highest classification and prediction accuracy. From the literature survey, it was noticed that the Decision Tree were employed for predicting the possible Heart Disease Pattern and associated factors. Researchers were considered a few Databases which created and maintained readings of Heart Patients to analyze their proposed models. Heart Disease Data Set[9] was downloaded from Machine Learning Repository of University of California, Irvine (UCI). The downloaded dataset is used for analysis of the exiting and our proposed model.

This work considered a few clinical parameters for Heart Disease Diagnosis and Prediction.

A few parameters [9,13,17] are namely

- Atrial Brillion (AFIB)
- Atrial Utter (AFL)
- Left Bundle Branch Block (LBBB)
- Normal Sinus Rhythm (NSR)
- Premature Ventricular Contraction (PVC))
- Right Bundle Branch Block (RBBB)
- Second Degree Block (BII) and
- Sinus Bradycardia (SBR)

The Back Propagation Network[14] also employed for better prediction.

From the report and survey, it was noticed that the above mentioned models achieved fair classification accuracy. For experimental analysis, Classification Accuracy, Precision, Sensitivity, Specificity, F-Measure were calculated and analyzed. Later on a few Researchers were proposed Prediction Model with Convolutional Neural Network[2,3] and from the experimental analysis, it was noticed that it performs relatively better.

Mohan and et. al. proposed Genetic Algorithm based Heart Disease Pattern[17]. This is the combination of Association Rule and Genetic Algorithm for predicted fitness function.

The Particle Swarm Optimization (PSO) Model [1] was introduced to predict Hear Disease Patterns which are the effective technique used for predicting Heart Diseases. This is one of the methodologies proposed for improving the prediction of Heart Diseases in a better way. This Research Work also reviewed and surveyed various Mining and Intelligent Methodologies that developed for Heart Disease Diagnosis and Prediction.

This work reviewed a few recently proposed Artificial Intelligent based Technologies and Machine Intelligence Models. The experimental reports have shown that these models were producing better pattern prediction.

The author Senthilkumar Mohan and et. al. proposed an efficient Data Mining Classifier called Hybrid Random Forest with Linear Model (HRFLM).

The prime goal of this is HRFLM [17] is for maximizing the classification and prediction accuracy of Heart Diseases Pattern and Diagnosis. The introduced Hybrid RF Technique considers all features of Heart Disease Pattern without ignoring features of those patterns. This work understood from the experimental report that this model achieves better classification than that of existing models and classifiers.

III. HYBRID RANDOM FOREST WITH LINEAR MODEL (HRFLM)

The Hybrid Classifier was employing three Association Rule Miners namely i. Tertius, ii. Predictive and iii. Apriori. These rule miners were used for finding the possible patterns of heart diseases. The procedure for finding the patterns [17] is shown in the Figure 1.

From the Database Patterns that were available for analysis, it was noticed and revealed that Males have more possibilities and chances of affecting heart disease as compared with chances to Females.

As Cardiologist needed efficient prediction models, the author proposed this Hybrid Model to predict better and early prediction of Heart Diseases than that of existing models. The survey established that the traditional approaches failed to predict pattern with high classification accuracy.

This work is understood that the Hybrid Classifier employed Artificial Neural Network with Back Propagation Network Techniques. It was considered around thirteen features for its input. As more parameters considered for prediction, it was performing well competitively with existing models.

As far as the author Senthilkumar Mohan and et. al. concerned, the Linear Regression and Artificial Neural Network with Back Propagation are selected as better Heart Disease Pattern Predictor and diagnosis Models. These models were employed in the Hybrid Classifier for analysis.

As shown in the Figure 1, the Heart Disease Data Set which downloaded from Machine Learning Repository[9] of University of California, Irvine (UCI) contains various attributes that are shown for the Diagnosis of Heart Diseases Patterns. It shows that 0 was represented for Nil Disease and 1 is represented for noticing Disease Patterns. Depends on the severity of the Disease, the weight value is raised from 1 to 4.

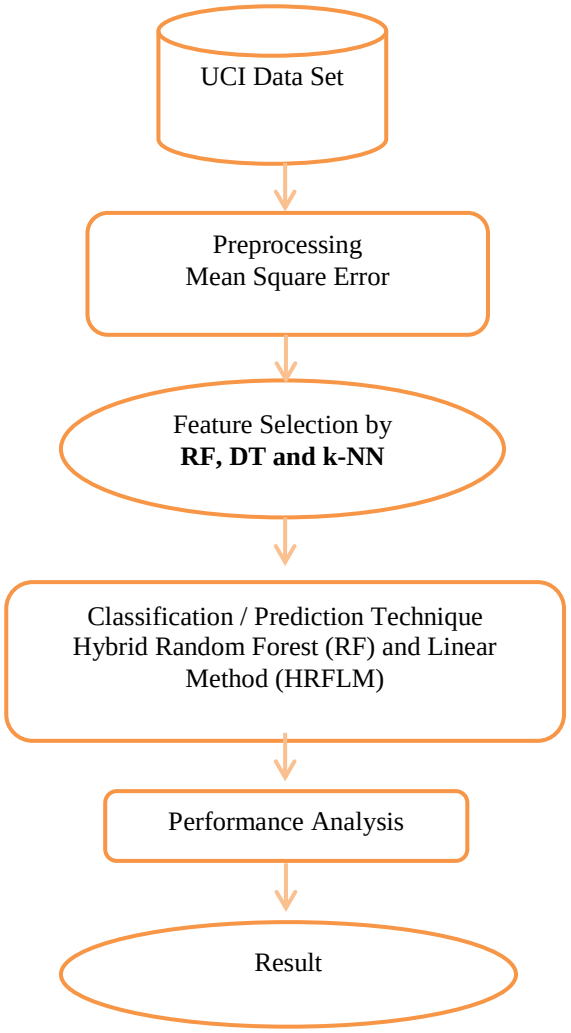

Figure 1. Hybrid Random Forest with a Linear Model (HRFLM)

The Figure 1 represents the sequences of the Hybrid Classifier's Procedure [17] that proposed for Disease Diagnosis and Prediction.

i. Loading UCI Dataset downloaded from Machine Learning Repository of University of California, Irvine (UCI)
ii. The data and its attributes that collected from Dataset were preprocessed
iii. Feature Selection Attributes are Ranked with the help of Decision Tree, Random Forest and k-Nearest Neighbor
iv. Classification was made through Hybrid Classifier
v. Performance Analysis and Report Generation

A. Data Pre-Processing

The preprocessing is performed on Heart Disease Datasets that are done by Senthilkumar Mohan and et. al. this Dataset consists of Patients records.

In the analysis, the disease patterns presence and absence were recorded. As discussed earlier, 0 is assigned if the disease patterns are missed and I to 4 is assigned when disease pattern noticed. The detailed Dataset Information is shown in the Table 1.

Table - I

Attribute	Description
Age	Age of the Patients
Sex	Male : 1, Female = 0
CP	Chest Pain 1 = Typical Angina 2 = Atypical Angina 3 = Non-Anginal Pain 4 = Asymptomatic – No Symptoms
BPRest	Blood Pressure at Rest
Chol	HDL+LDL
FBS	BloodSugarFasting > 120 mg/dl (1 = True; 0 = False)
ECGRest	Electrocardiogram at Rest (0 = Normal; 1 = having ST-T; 2 = Hypertrophy)
THALACH	Maximum Heart Rate Achieved
EXANG	Exercise Induced Angina (1 = Yes; 0 = No)
OLDPEAK	ST Depression induced by Exercise Relative to rest
SLOPE	the Slope of the Peak Exercise ST Segment (1 = Upsloping; 2 = Flat; 3 = Downsloping)
CA	Number of Major Vessels (0-3) Colored by Flourosopy
THAL	3 = Normal; 6 = Fixed Defect; 7 = Reversible Defect
Num	The Predicted Attribute - Diagnosis of Heart Disease (Angiographic Disease Status) (Value 0 = < 50% Diameter Narrowing; Value 1 = > 50% Diameter Narrowing)

B. Feature Selection and Classification Models

The Thirteen Listed Attributes of the UCI Dataset are used for predicting Heart Disease Patterns. The Hybrid Classifier employed Decision Tree, Random Forest and Linear Method for Disease Diagnosis and Prediction.

C. Decision Tree

As we know, the Decision Tree has the methodology to construct tree with the input that has highest entropy on Data Sets D for training samples. This is considered as the fastest methodology which is the top down recursive model and this is the Divide and Conquer Methodology.

The irrelevant samples and noises needed to remove from Dataset D under Tree Pruning Approach as

$$\text{Entropy} = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij} \quad (1)$$

D. Random Forest

The Decision Tree was constructed with the Hybrid Classifier for achieving higher Classification Accuracy.

Let us consider $X = \{x_1, x_{12}, \ldots\ldots\ldots\ldots x_n\}$. the expected responses are $Y = \{x_1, x_{12}, \ldots\ldots\ldots\ldots x_n\}$. This could be repeated for consolidating by considering b = 1 to B

The samples x' which are not selected during possible predictions $\sum_{b=1}^{B} fb(x')$ from the various unique trees which are generated by x'

$$j = \frac{1}{B} \sum_{b=1}^{B} fb(x') \quad (2)$$

Here the Standard Deviation was employed to predict the uncertainty on the above specified tree

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} (fb(x') - f)^2}{B-1}} \quad (3)$$

E. K-Nearest Neighbor

The k-Nearest Neighbors (KNN) is considered as one of the simple and powerful Machine Learning Technique. This is performing as Supervised Learning Model which is employed for Classification.

It is helping to analyses the patterns of input data for classification based on the Euclidean Distance Model.

$$d(x_{i,xi} = \sqrt{(x_{i,1} - x_{j,1})^2} + \ldots\ldots + (x_{i,m} - x_{j,m})^2 \quad (4)$$

IV. IDENTIFIED PROBLEM

The existing Hybrid Classifier was proposed for achieving higher Classification Accuracy to predict Heart Diseases Patterns and Diagnosis. It ensembles Decision Tree, Random Forest and k-Nearest Neighbor.

This research work is noticed that the existing model may not achieve better Classification Accuracy for large Datasets as Decision Tree and k-NN failed to perform well for large Datasets. Further this work initiated to employ better powerful Classifier namely Support Vector Classifier and Weighted Random Forest to achieve and maximize Classification Accuracy.

This research work proposed an Intelligent Framework comprises Support Vector Classifier and Weighted Random Forest.

V. PROPOSED ENSEMBLE CLASSIFIER

The procedure of the proposed Ensemble Classifier is shown in the Figure Fig. 2. ie as shown in the Fig. 2, the Ensemble Classifier consists of the following procedure to perform pattern prediction and diagnosis.

i. Load UCI Dataset)
ii. The data and its attributes that collected from Dataset were preprocessed as
  a. Verify Missing Values
  b. Check Patterns based on Histograms
  c. Standardize the Data Input and Transform with PCA
iii. Splitting and Feature Selection by Attributes are Ranked with the help of Support Vector Classifier and Weighted Random Forest
iv. Classification was made through Ensemble Classifier
v. Measure individual Classification Accuracy and Error
vi. Calculate Weighted Average
vii. Performance Analysis and Report Generation

A. Data Pre-Processing

This proposed work modelled any efficient preprocessing model with the help of PCA and various Histograms. This helps our system well to effectively handle missing data.

The various features and components of PCA were employed for better performances.

This supported to remodel the pattern that eliminated noises and irrelevant patterns as dimensionality reduction. This facilitates our proposed model improve classification accuracy.

B. Data Pre-Processing

This work used Thirteen Listed Attributes of the UCI Dataset for validating our proposed model with the existing model. The proposed Ensemble Classifier employed Support Vector Classifier and Weighted Random Forest for better Disease Diagnosis and early Prediction.

C. Support Vector Classifier

Artificial Intelligence is the Intelligent Tool which is used to design Intelligent Machines. It is facilitating to design Smart Models with the help of Machine Learning which is one of the branches of Artificial Intelligence. It is enabling Intelligent Models to learn the Patterns for prediction.

This Research Work proposed Machine Learning based Support Vector Machine named as Support Vector Classifier and this is proposed for better Classification and Regression too. ie the proposed model is maximizing predictive accuracy. The trade-off is shown in the Figure 2.

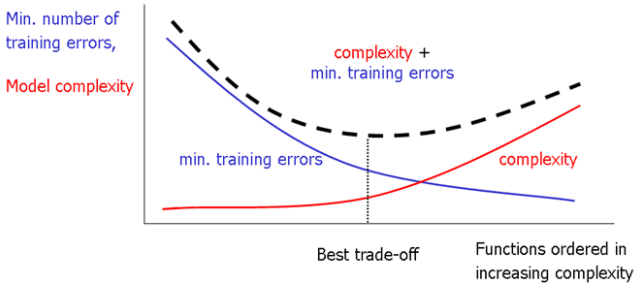

Fig. 2. Support Vector Classifier

Representation of Support Vector Classifier

The Support Vector Classifier formulation is shown below.

$$\min_{f,\xi_i} \|f\|_K^2 + C\sum_{i=1}^{l}\xi_i$$

$$y_i f(x_i) \geq 1 - \xi_i, \text{ for all } i \quad \xi_i \geq 0 \qquad (4)$$

The dual formulation of Support Vector Classifier is

$$\min_{\alpha_i} \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)$$

$$0 \leq \alpha_i \leq C, \text{ for all } i; \qquad (5)$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0 \qquad (6)$$

Here ξ_i is represented as slack variables, which is used to measure the error between points (x_i, y_i) as shown in the Figure 2.

D. Weighted Random Forest Classifier

Let us consider the data comprises different responses like 0 and 1. Consider a predictor p which introduced for validating N samples.

As we know that the traditional Random forest fails to achieve better Classification Accuracy, we considered the Weighted Random Forest. This WRF Technique is building tree of Forest with aggregate weights and construct tree aggregation.

As aggregation provides votes for each tree of the forest, this model yields better performance in term of Classification Accuracy.

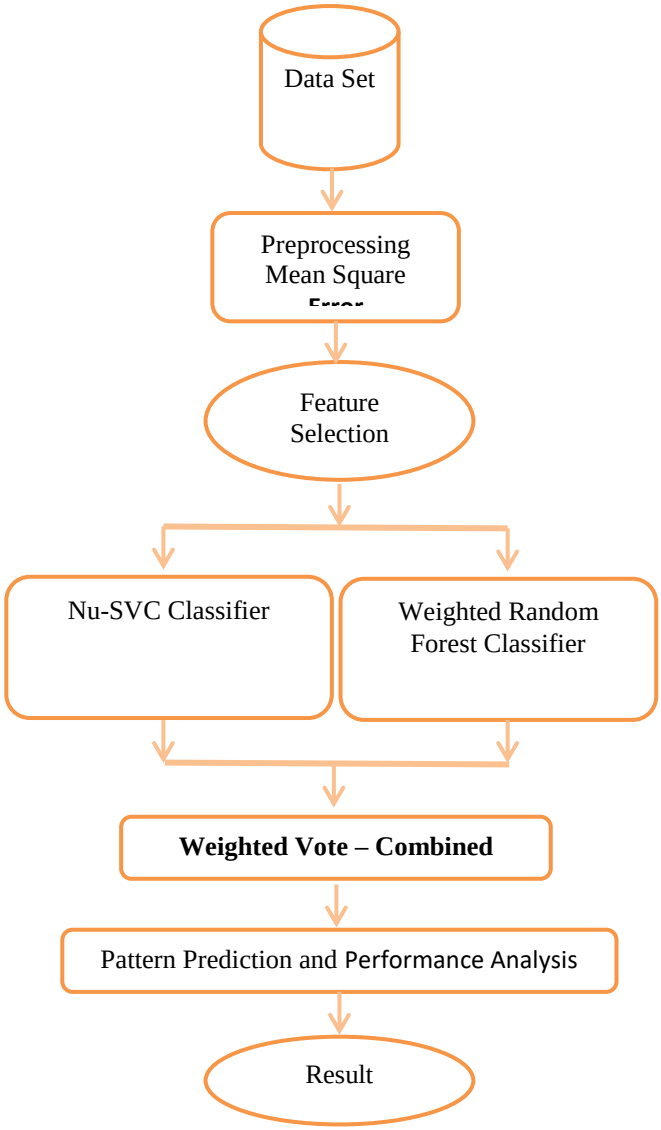

Fig. 3.
Proposed Ensemble Classifier comprises of Nu-SVC Classifier and Weighted Random Forest Classifier

This work applied the calculated weights to the UCI Dataset and analyzed Classification Error. As to minimize the classification error, it has the following steps

- Split the Dataset into Training and Testing
- With Bootstrap Patterns, estimate predictive ability and assess Tree Weight and measure performance. Once Tree Weight was calculated for the given training dataset, let us consider ntree will be the observed vote achieved for test data

- The aggregate vote w_j . Let us consider $V_{test,ij}$ will be the vote for the j tree and the i subject. The Weighted predicted can be measured as

$$wP_i = \sum_{j=1}^{ntree} w_j v_{test,ij} \quad (7)$$

VI. EXPERIMENTAL STUDY AND PERFORMANCE ANALYSIS

The proposed Ensemble Classifier named SVC-WRF is implemented in VC++ Programming Language and interfaced with R Programming for Feature Selection, Matching and prediction of Heart Diseases Pattern Diagnosis and Early Prediction. To evaluate the proposed model, the Dataset was downloaded from Machine Learning Repository of University of California, Irvine (UCI)[] that contains various attributes that are shown for the Diagnosis of Heart Diseases Patterns. It shows that 0 was represented for Nil Disease and 1 is represented for noticing Disease Patterns. Depends on the severity of the Disease, the weight value is raised from 1 to 4. The details of the considered attributes are shown in the Table 1. The proposed and the existing Classifiers were implemented and analyzed. The confusion matric were computed and from the matric, the following metrices were calculated and analysed the performance efficiencies of the proposed Ensemble Classifier in terms of Accuracy, Sensitivity, Specificity, Precision, FScore and False Positive Rate FPR.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$\text{Sensitivity} = \text{Recall} = \text{TPR} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{Specificity} = \text{FPR} = \frac{TN}{TN+FN} \quad (10)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

$$\text{FScore} = \frac{2(\text{Recall} * \text{Precision})}{\text{Recall}+\text{Precision}} \quad (12)$$

From the Table 2 and Figure 4, it is revealed that the proposed Ensemble Classifier is performing well as compared with that of the existing Hybrid Classifier.

Table – II Performance Analysis of the Proposed SVC-WRF

Performance Matrices	Classifiers	
	Ensemble Classifier SVC-WRF	Hybrid Classifier HRFLM
Accuracy	**0.962**	0.936
Sensitivity	**0.980**	0.973
Specificity	**0.953**	0.939
Precision	**0.980**	0.973
FScore	**0.980**	0.973
FPR	**0.078**	0.143

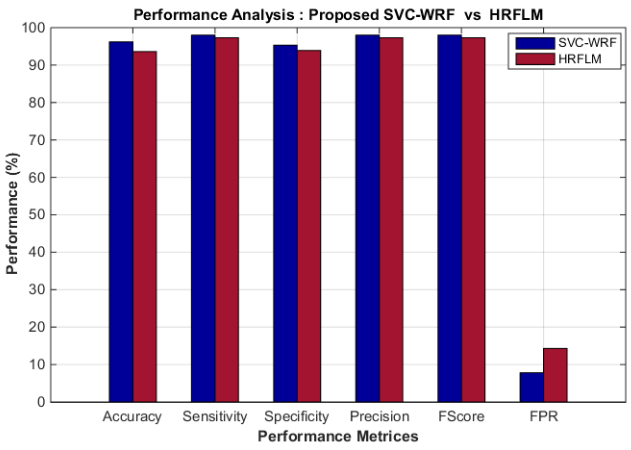

Figure 4. Performance Analysis of the proposed Ensemble Classifier

VII. CONCLUSION

This research work identified the recently proposed Hybrid Random Forest with a Linear Model (HRFLM) Classifier for improving the classification accuracy for the cardiovascular disease patterns prediction well in advance and Diagnosis as well. However, it was noticed that for improving the performances better in terms of Accuracy, Sensitivity, Specificity, Precision, FScore and False Positive Rate FPR, needed an efficient classifier. Thus this work developed and implemented an efficient Classifier ensemble Nu-SVC Classifier and Weighted Random Forest Classifier. From the experimental results, it was noticed that the proposed Ensemble Classifier performs better as compared with that of existing Hybrid Classifier in terms of in terms of Accuracy, Sensitivity, Specificity, Precision, FScore and False Positive Rate FPR

REFERENCES

1. A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ``Using PSO algorithm for producing best rules in diagnosis of heart disease," in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306-311.
2. C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ``Analysis of neural networks based heart disease prediction system," in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233-239.
3. C.-A. Cheng and H.-W. Chiu, ``An articial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database," in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566-2569.
4. F. Dammak, L. Baccour, and A. M. Alimi, ``The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains," in Proc. IEEE Int. Conf. Fuzzy Syst.(FUZZ-IEEE), vol. 9, Aug. 2015, pp. 1-8.
5. G. Sun, C. Jiang, P. Cheng, Y. Liu, X. Wang, Y. Fu, and Y. He, ``Shortterm wind power forecasts by a synthetical similar time series data mining method," Renew. Energy, vol. 115, pp. 575_584, Jan. 2018.
6. G. Sun, Y. Cong, and X. Xu, ``Active lifelong learning with `watchdog,'" in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 4107_4114.
7. H. A. Esfahani and M. Ghazanfari, ``Cardiovascular disease detection using a new ensemble classier," in Proc. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011-1014.
8. H.-H. Yang, M.-L. Huang, C.-M. Lai, and J.-R. Jin, ``An approach combining data mining and control charts-based model for fault detection in wind turbines," Renew. Energy, vol. 115, pp. 808_816, Jan. 2018.

9. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
10. L. Baccour, ``Amended fused TOPSIS-VIKOR for classication (ATOVIC) applied to some UCI data sets," Expert Syst. Appl., vol. 99,pp. 115-125, Jun. 2018.
11. M. Abdar and N. Y. Yen, ``Design of a universal user model for dynamic crowd preference sensing and decision-making behavior analysis," IEEE Access, vol. 5, pp. 24842_24852, 2017.
12. M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, ``Performance analysis of classi_cation algorithms on early detection of liver disease," Expert Syst. Appl., vol. 67, pp. 239_251, Jan. 2017.
13. Moloud Abdar, U. Rajendra Acharya, Nizal Sarrafzadegan, and Vladimir Makarenkov, "NE-Nu-SVC: A New Nested Ensemble Clinical Decision Support System for Effective Diagnosis of Coronary Artery Disease," IEEE Access, pp. 167605-167620, 2019.
14. N. Al-milli, ``Backpropogation neural network for prediction of heart disease," J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131-135, 2013.
15. P. K. Anooj, ``Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," J. King Saud Univ.-Comput. Inf. Sci., vol. 24, no. 1, pp. 27-40, Jan. 2012.
16. S. Abdullah and R. R. Rajalaxmi, ``A data mining model for predicting the coronary heart disease using random forest classier," in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22-25.
17. Senthilkumar Mohan , Chandrasegar Thirumalai, and Gautam Srivastavaa, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," Smart Caching, Communications, Computing and Cybersecurity For Information-Centric Internet Of Things, pp. 81542- 81554, Vol. 7, 2019.
18. Y. Cong, G. Sun, J. Liu, H. Yu, and J. Luo, ``User attribute discovery with missing labels," Pattern Recognit., vol. 73, pp. 33_46, Jan. 2018.
19. Y.Wang, L. Kung,W. Y. C.Wang, and C. G. Cegielski, ``An integrated big data analytics-enabled transformation model: Application to health care," Inf. Manage., vol. 55, no. 1, pp. 64_79, 2018.

AUTHORS PROFILE

S.Brindha MCA, M.Phil P.G Diploma in Bioinformatics, is currently working as Assistant Professor of Computer Science and Applications Department at the St.Peter's Institute of Higher Education and Research, Avadi, Chennai. She has more than 20 years of teaching experience in various Engineering Colleges and University. She has published more than 100 technical papers, in reputed Journals and Conferences. Her research is centered on Development of an Energy Efficient Data Centre Network(DCN) Architecture for Cloud Applications, Bioinformatics, Wireless Sensor Networks. She is the life Member of ISTE and CSI..

T. Ajisha. **MCA, M.Phil,** is currently working as Assistant Professor of Computer Science and Applications Department at the St.Peter's Institute of Higher Education and Research, Avadi, Chennai. She has 6 months of teaching experience. Her research work is cloud computing and cryptography. She has published 2 papers in reputed Journals and conference.