Cardiovascular Disease Prediction using Novel Hybrid Filter Based Chi-Squared Feature Selection with Backward Elimination and Spider Monkey Optimization-Artificial Neural Network Techniques

Dr.P.Deepika[1], Dr.S.Saranya[2],
[1]Associate Professor, Department of Artificial Intelligence & Machine Learning, Dr.N.G.P. Arts and Science College, Coimbatore, Tamilnadu.
[2] Assistant Professor, Department of Computer Science with Cognitive Systems and Artificial Intelligence and Machine Learning, Hindusthan College of Arts & Science, Coimbatore, Tamilnadu, India.

Abstract

Cardiovascular Disease (CVD) are considered to be the major effect of mortality rate globally. According to the report of World Health Organization in 2019, 17.9 million people are suffering due to this CVD. Among the deaths, 85% are due to respiratory malfunction and stroke. Machine Learning techniques are used to retrieve the information from large amount of data. This research focused to perform the comprehensive analysis of various works done by researchers using different machine learning methods. For this work, Z Alizadeh Sani Heart Disease Dataset is used for interpretation, analysis and correlation of CVD features. A Novel Hybrid Filter Based Chi-Squared Feature Selection with Backward Elimination is used for selecting impactful features from the dataset. The Spider Monkey Optimization with Artificial Neural Network Method (SMO-ANN) is used for predicting cardiovascular disease. The results are compared with the dataset without feature selection and with feature selection. The comparison results showed that impact of novel feature selection is reflecting in the prediction accuracy using SMO-ANN method. This proposed method achieved 89% accuracy in without feature selection and 94% accuracy in with novel proposed feature selection method.

Keywords: ***Machine Learning, Classification, Meta-Heuristic Optimization, Artificial Neural Network***

1. INTRODUCTION

Cardiovascular diseases are the most vital cause of death in all around the world. According to the report of WHO in 2019, there were 17.9 million people suffering due to this illness. This represents the 32% of world population. From this report, it is highlighted that 85% of people are suffered respiratory Failure and stroke. The life style factors are the major impact for this cardiovascular disease. For example, tobacco use, unhealthy eating habits, obesity, actual idleness, lack of physical activity, stress and hurtful utilization of liquor (P.Deepika. D. , 2020). The major requirement is identifying the risk of cardiovascular illness in right time for taking the medical care. Machine learning methods are providing the most promising result for identification and prediction of heart diseases based on features. The healthcare sectors are containing number of records. The patient's details are stored and maintained by the healthcare sectors for further analysis. The cardiovascular diseases indicates the group of disorders which consist heart and blood vessels. The most common type of cardiovascular diseases are heart attack, stroke, heart failure and arrhythmia(Najmul Hasan., 2020). This research work includes the review of medical research based on heart disease and investigating the features involved in cardiovascular diseases.

2. LITERATURE ANALYSIS

AtharvNikam et al (Atharv Nikam., 2020)proposed the prediction of heart disease using methods in machine learning. In this research work, various features are used for the prediction process. The BMI (Body Mass Index) Method is used as an important feature in this heart disease prediction. Various features like age, glucose, height, presence of cardiovascular disease, Weight, lack of Physical activity, Gender,

Alcohol intake, Systolic Blood pressure, Smoking, diastolic blood pressure and cholesterol are used for this prediction. Different classification models are used in this research work such as logistic regression, K-nearest Neighbors algorithm, naïve bayes, neural network, Decision Tree classifier, XGB Classifier and LGBM Classifier with HyperOpt. From the comparison report, the accuracies of these models are compared and highlighted that the XGB model provides high accuracy.

P.Deepika and Dr.S.Sasikala (P.Deepika. S. .., 2020) proposed the cardiovascular disease prediction using classification models. The Particle Swarm Optimization method is used for feature selection process. The feature subset is produced by this PSO. This research work contains UCI dataset with 270 records with 14 attributes. The PSO optimization produced 7 features and the generation value 20. The selected subset features are cp, restecg, thalach, exang, oldpeak, ca and thal. The used 80:20 split ration for training and testing. The decision tree classification model is used for classification and prediction. The performance measure consist sensitivity, specificity, precision, negative Predictive values, False positive Rate, False discovery rate, accuracy and F1 score and Matthews Correlation Coefficient. The accuracy comparison and time comparison was done with PSO model and without PSO model.

KeerthiSamhitha et al (B.Keerthi Samhitha., 2020)proposed improving accurateness in prediction of Cardiovascular Disease by using the algorithms of machine learning. In this research work, they used python and pandas activities to explore the results using UCI dataset. The dataset contains 303 records which has 6 records with missing values. The data preprocessing methods are used to handle this missing values. The feature extraction and reduction methods are applied to select the important features. The used dataset contains 13 attributes. The feature selection methods selected 11 attributes for further processing. The K-nearest Neighbour, K-means Clustering methods and Adaboost methods are applied on the dataset. From the comparison results, the k-Neighbour classifier provides high accuracy in the heart disease prediction.

AyonDey et al (Ayon Dey., 2016)proposed the investigation of supervised machine learning processes for the prediction of heart diseases. They used reduced number of attributes by using the Principal Component Analysis. In this research work, the authors collected heart disease dataset from UCI repository. They implemented Naïve Bayes classifier, Decision tree classifier, support vector machine classifier and Principal Component Analysis methods for the heart disease prediction. They used dataset with 50 intervals for implementation.. The principal Component Analysis is used to reduce the number of attributes. From this reduced size, the SVM produced good accuracy when compared with Naïve Bayes and Decision tree.

P.Deepika et al(P.DEEPIKA., 2018) proposed a Noval classification and prediction algorithm for predicting the heart disease. They proposed a new decision tree model with sequential covering techniques. This advanced decision tree provides fast disease classification with high accuracy. This proposed system collecst and mar the score for every label using this ensemble approach. The C4.5 decision tree model is taken for enhancement. They used the dataset with 272 attributes. From this collected dataset, 2 records contains missing tuples. The data pre-processing method is used to solve these missing values in dataset. The training models are having various splitting criterions where the proposed system focus statistical deviance. The comparison results showed that the accuracy produced by the proposed system is better that the existing decision tree models.

3. METHODS

3.1 Statistical Analysis for CVD dataset

The exploratory data analysis is used for interpretation of features to extract the patterns and correlations between attributes in the dataset. For this research work, Z Alizadeh Sani dataset contains 303 records with 54 attributes is used. This research work focused the univariate and Bivariate analysis using different attributes in the dataset. Among the different features, the most influenced features are selected ad involved

in this univariate and bivariate analysis. Those are Cholesterol level, thalassemia and ECG. The old peak against cholesterol is calculated using correlation matrix.

The following formula is used for calculating the linear correlation, covariance and variance.

$$r = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}}$$

$$Covar(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

$$Var(x) = \frac{\Sigma(x-\bar{x})^2}{n}$$

$$Var(y) = \frac{\Sigma(y-\bar{y})^2}{n}$$

r : Linear Correlation

Covar : Covariance

Var : Variance

The correlation coefficient is used to represent the statistical measure which express the extends to which two variables are related linearly. Generally there are three possible outcomes in correlation process such as positive correlation, negative correlation and no correlation. From this results, the linear relationship between two continuous variables are expressed. The covariance term indicates the direction of linear relationship between two variables. The changes in one variable reflect in another represent the covariance. The probability of covariance represents the measure of joint probability for two random variables. The result of covariance is useful to produce the portfolio which highlights the theory to determine what assets is included.

$$\text{Cov}(X,Y) = \frac{\Sigma(X_i - \bar{X})(Y_j - \bar{Y})}{n-1}$$

In this formula, Xi indicates the values of X variable, Yj indicates the values of Y-variable, $\bar{X}$ indicates the mean average of X variables, $\bar{Y}$ indicates the mean average of Y variable and the n variable indicates the number of data points. The strength of relationship occurs between -1 and 1. The negative correlation is represented using -1 and the perfect positive correlation is represented using +1. The value 0 highlights the no correlation between the attributes.

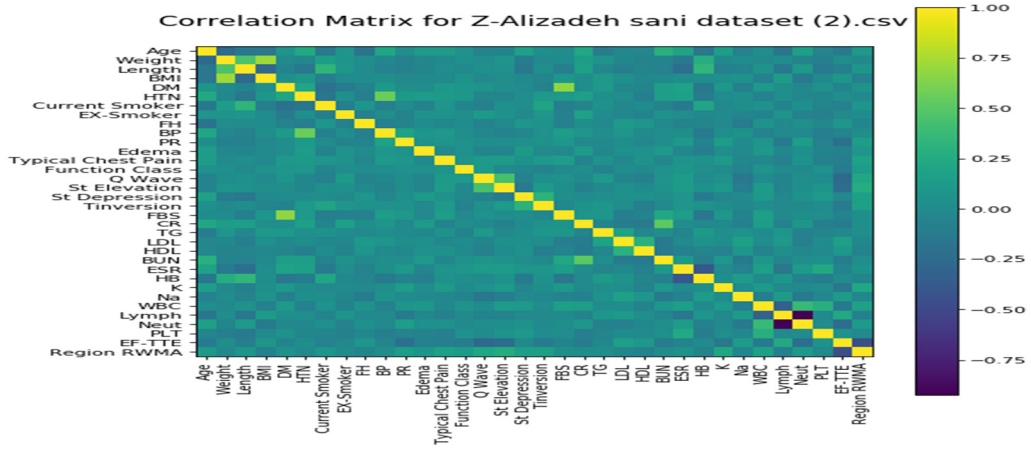

Fig:1 Correlation matrix for Z AlizaDeh Sani dataset

3.2 Classification Models

The widely used classification models and other methods in CVD prediction are considered in this work. The result of discussed model work is compared with these models to prove the efficiency of this proposed model.

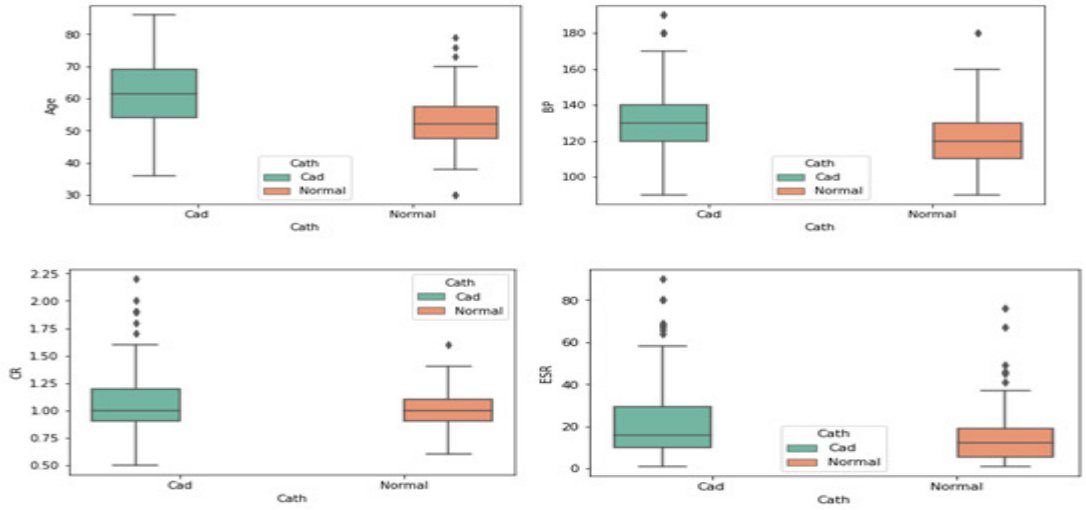

Fig: 2 Class representation from Z Alizadeh Sani dataset

The machine learning methods are used in this work to classify the instances with cardiovascular diseases affected class label. The following diagram provides the classification structure using machine learning methods

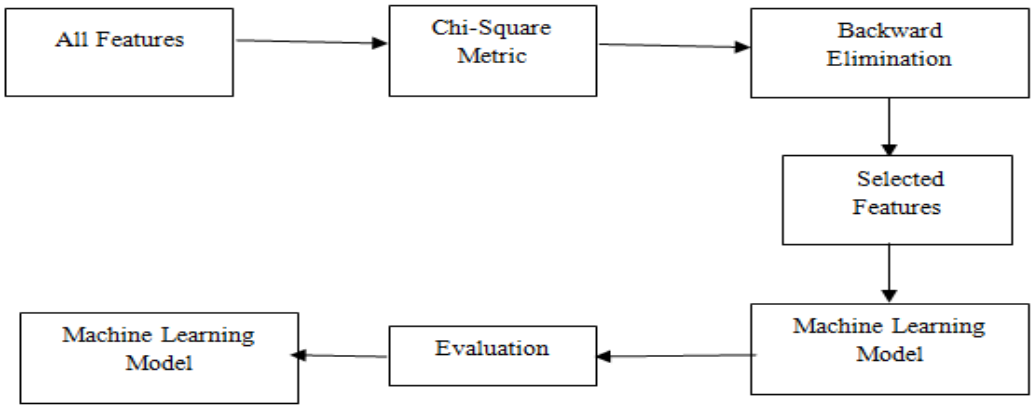

Fig: 3 Proposed Novel Hybrid filter Based Chi-Squared feature selection with backward elimination working scenario

Wrapper based techniques referring to the extensive search techniques of all possible subsets of the features in feature space is the logic behind the wrapper methods. Feature selection rely on the machine learning algorithm. It refers the possible combinations of features on to the evaluation condition. This research deploys the chi squared filter based technique in backward feature elimination. Initially consider all the features and build the model and based on the performance metrics, remove the features and

reconstruct the model. Chi-square value can be calculated for each feature in accordance with the target value. As a result, higher chi-square value features can be selected for the further machine learning aspects.

3.3 Statistical Significance of The Feature In CVD Using Chi-Squared Feature Selection

Chi square distribution can be defined as sum of squared standard normal variables which is derived for a random variable, X.

$$X^2 = \sum z_i^2$$

It has certain parameters considered and they are listed below:

- Degrees of freedom (DF): It can be referred as total count of observations - count of independent constraints supposed on the observations various degrees of freedom we can have various
  DF = N-1, where N is the number of samples
- Chi square distribution: it is noticed that when degrees of freedom increases then x square distribution it approximates to normal distribution. Chi-square test can be deployed for the feature selection and it is utilized in statistics to test independence of two events. Consider two variable one is observed count, O and another one is expected count, E. We can derive chi square using the following formula:

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The correlation between predictor and response has to be observed carefully. the association between Independent and dependent features present in the data set which is termed as predictor was response has to be monitor. In the case of independent features the observed account will be closed with expected count as a consequence we will have minimum size square value. If the features present in the data set process a high dependency then it has to be opted for model preparation.

3.4 Performance Analysis:

In this part, the Z Alizadeh Sani dataset is used for classification by using machine learning methods. The collected dataset contains class label with CVD and No CVD representation. The supervised learning methods are trained using the dataset with appropriate class label. This training consist 80% of instances from the dataset. This training can provide detailed knowledge for the classification methods to find the accurate class for new instances. The attributes and the values are focused by the classification methods for decision making.

In this work, the enhanced fitness function is used in association with spider monkey optimization. The results of this method is compared with 5 existing methods namely MLT Random forest, k-nearest Neighbour, Regression Tree, Naïve Bayes and Support Vector Machine method(Wei Liu., 2019). The main benefit is using this SMO-ANN is to find the weak learners in this process. The use of this method enhances the performance of used model which can reach the optimum solution.

4.SPIDER MONKEY OPTIMIZATION-ARTIFICIAL NEURAL NETWORK FOR CLASSIFICATION

Spider Monkey Optimization (SMO) is a worldwide streamlining calculation roused by Fission-Fusion social (FFS) design of monkeys during their rummaging conduct. SMO meticulously portrays two

principal ideas of swarm intelligence: self-association and division of work. SMO has acquired prominence lately as a multitude insight based calculation and is being applied to many designing improvement issues. SMO adopts the fission-fusion strategy which will help the task of meticulously analyzing the data of CVD in divide and conquer method in rapidly adopting phase for the local and global leader. The self-association quality makes them to coordinate the key insights identified and work split scenario helps to get optimized result in fast manner. Similarly while selecting the leader and updating the values the ANN agent is deployed for effective feedback utilization. This algorithm is the basic work flow of the SMO. Here the main tasks involved can be simply summarized as follows:

- Fitness calculation
- Local leader selection
- Global leader selection
- Compare old and new positions using greedy selection process
- Fitness updating
- Local leader updating
- Global leader updating

Further the ANN agent incorporates the knowledge of the learning into the leader updating and comparison of the old, new positions. The proposed algorithm combines the spider monkey optimization with the artificial neural networks learning.

5.RESULTS AND DISCUSSION

The effective performance of proposed Spider Monkey optimization-Artificial Neural Network is discussed in this part. The performance evaluation is considering different terms like instances that are classified correctly (CCI) , instances which are incorrectly classified (ICI), True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F-Score, Matthews correlation coefficient and time spend for prediction of that model. All these values are considered to find the efficient model for CVD prediction and classification.

5.1 Classification results without Feature Selection and with feature Selection

The following table is shown the comparative results of five different existing models with the proposed model. The feature selection methods are used to find the impactful features from the dataset(Päivi Riihimaa., 2020). The weightage of each attribute is considered in this selection process. From 54 different attributes, only 32 attributes are selected by using the feature selection methods. The results of classification methods with this selected features are compared and the results are shown in the table 1.

	CCI	ICI	TPR	FPR	PRE	Recall	F-Score	MCC	ACC	Time (Sec)
Results without Feature Selection										
Z Alizadeh Sani										
RF	85.14	14.85	0.85	0.3	0.85	0.85	0.85	0.61	0.89	0.39
SVM	87.78	12.21	0.87	0.17	0.87	0.87	0.87	0.7	0.85	0.03
KNN	72.27	27.72	0.72	0.33	0.74	0.72	0.73	0.36	0.75	0
RT	82.83	17.16	0.82	0.24	0.82	0.82	0.82	0.58	0.75	0.03
NB	79.86	20.13	0.79	0.21	0.81	0.79	0.8	0.54	0.77	0.03
Proposed SMOANN	**83.4**	**16.5**	**0.83**	**0.17**	**0.83**	**0.8**	**0.8**	**0.61**	**0.89**	**0.34**
Results with Feature Selection										
Z Alizadeh Sani										
RF	88.11	11.88	0.88	0.2	0.87	0.88	0.87	0.7	0.9	0.28
SVM	87.12	12.87	0.87	0.17	0.87	0.87	0.87	0.68	0.84	0.02
KNN	81.18	18.81	0.81	0.23	0.82	0.81	0.81	0.55	0.86	0
RT	84.15	15.84	0.84	0.24	0.83	0.84	0.84	0.6	0.82	0.02
NB	82.17	17.82	0.82	0.18	0.84	0.82	0.82	0.6	0.89	0.01
Proposed SMOANN	**89.14**	**11.85**	**0.89**	**0.3**	**0.88**	**0.87**	**0.89**	**0.72**	**0.94**	**0.01**

Table:1 Comparison of Classification models using without and with feature selection process

5.2 Discussion

The comparison of diverse models from machine learning in CVD prediction with proposed Spider Monkey optimization with Artificial Neural network is shown in the chart. This comparison includes with feature selection process and without feature selection process for CVD prediction. The fast processing time and the prediction time is also compared with this chart.

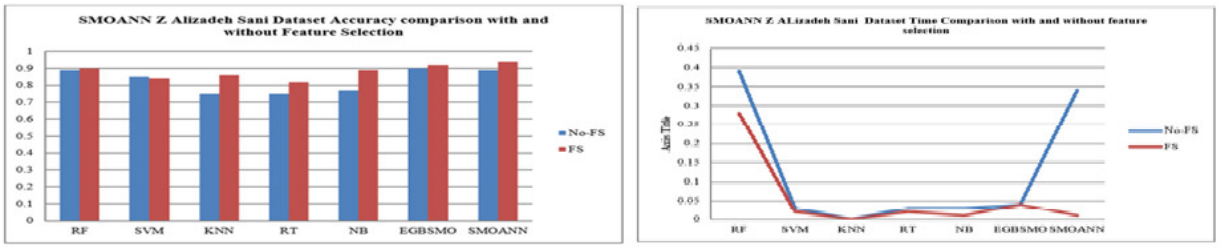

Fig:5 Performance and Time comparison of machine learning methods with Spider Monkey Optimization-Artificial Neural Network

From this chart, the results showed that the proposed Spider Monkey Optimization- Artificial Neural network method achieved good accuracy in classification and CVD prediction. This method produced high accuracy when using feature selection methods and not using feature selection methods. This method is produced prediction results with less time when compared with other models.

6. CONCLUSION

The feature selection methods are used to find the weighted features from the dataset. The weightages are calculated and the correlation of each attribute is considered. The classification methods are trained to predict the class label using Z Alizadeh Sani dataset. The performances of the classification methods are compared with four different metrics like accuracy of the model with no feature selection process, accuracy of the model with feature selection process, Fastness of classification without feature selection and fastness of classifier with feature selection. The results are compared with the dataset without feature selection and with feature selection. The comparison results showed that impact of novel feature selection is reflecting in the prediction accuracy using SMO-ANN method. This proposed method achieved 89% accuracy in without feature selection and 94% accuracy in with novel proposed feature selection method.

References:

[1] Atharv Nikam., Sanket Bhandari., Aditya Mhaske., Shamla Mantri.,. "Cardiovascular Diseases Prediction Using Machine Learning Models." *IEEE Pune Section International Conference*, 2020.
[2] Ayla Gulcu., Zeki Kus.,. "Hyper-Parameter Selection in Convolutional Neural Networks Using Microcanonical Optimization Algorithm." *IEEE Access*, 2020: Vol-8.
[3] Ayon Dey., Jyoti Singh., Neeta Singh.,. "Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis." *International Journal of Computer Applications* , 2016: Volume 140 – No.2.
[4] B.Keerthi Samhitha., Sarika Priya.M.R., Sanjana.C., Suja Cherukullapurath Mana., Jithina Jose.,. "Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms ." *International Conference on Communication and Signal Processing*, 2020.
[5] Devansh Shah., Samir Patel., Santosh Kumar Bharti.,. "Heart Disease Prediction using Machine Learning Techniques." *SN Computer Science*, 2020: Vol-1, issue-6.
[6] Najmul Hasan., Yukun Bao.,. "Comparing different feature selection algorithms for cardiovascular disease prediction." *Health and Technology*, 2020: Vol-11, issue-1.

[7] P Deepika, S Saranya, S Sasikala, Dr S Jansi, A Kiruthika. "Anticipating Heart Disease using C4. 5 Classification Augmented with Feature Selection." *International Journal for Research & Development in Technology*, 2016: Vol-6, .
[8] P.DEEPIKA., DR. S. SASIKALA., S.SARANYA., A.KIRUTHIKA.,. "A Noval Classification And Prediction Algorithm For Heart." *International Journal of Engineering Science Invention*, 2018: Volume 7 Issue 2.
[9] P.Deepika., Dr.S.Sasikala.,. "Effectiveness of Feature Selection methods in Cardiovascular Disease Prediction Using Classification." *Journal of Maharaja Sayajirao University of Baroda*, 2020: Vol-54, issue-2, pp 153-158.
[10] P.Deepika., S.Sasikala .,. "Enhanced Model for Prediction and Classification of Cardiovascular Disease using Decision Tree with Particle Swarm Optimization ." *Fourth International Conference on Electronics, Communication and Aerospace Technology- IEEE*, 2020.
[11] Päivi Riihimaa. "Impact of machine learning and feature selection on type 2 diabetes risk prediction." *Journal of Medical Artificial Intelligence*, 2020: Vol-3.
[12] Priya R. L., S. Vinila Jinny., Yash Vijay Mate.,. "Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques." *Health and Technology*, 2020.
[13] Pulugu Dileep., Kunjam Nageswara Rao., Prajna Bodapati.,. "Enhancing Heart Disease Prediction Models with Feature Selection and Ensemble Methods." *Journal of Advanced Research in Dynamical and Control Systems*, 2019: Vol-11, issue-11.
[14] P.Deepika., Dr.S.Sasikala., " Data Classification Pertaining to Heart Disease Using Hybrid Chicken Swarm Optimization With Artificial Neural Network (HCSOANN)" International journal of Mechanical Engineering, Vol-7, Issue-1, 2022.
[15] K.Thenmozhi., P.Deepika., " Heart Disease Prediction using Classification with Different Decision Tree Techniques" International Journal of Engineering Research and General Science, Vol-2, Issue-6, 2014.
[16] S.Sasikala., S.Saranya., P.Deepika., S.Jansi., A.Kiruthika., " Anticipating the Chronic kidney Disorder using performance optimization in Adaboost and Multilayer Perceptron" Imperial Journal of Interdisciplinary Research, Vol-3, 2017.