Middle-East Journal of Scientific Research
24 (Recent Innovations in Engineering, Technology, Management & Applications): 200-206, 2016
ISSN 1990-9233; © IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.RIETMA133

Analysis of Classification Algorithms for Heart Disease Prediction and its Accuracies

[1]R. Jothikumar and [2]R.V. Sivabalan

[1]Noorul Islam University, Kumaracoil, Tuckalay,
Kanyakumari Dt-629180, Tamil Nadu, India
[2]Department of Master of Computer Application, Noorul Islam University,
Kumaracoil, Tuckalay, Kanyakumari Dt-629180, Tamil Nadu, India

Abstract: Heart disease is the leading cause of death and it is necessary to predict it at earlier stages to save the life of human beings. Many researchers proposed number of data mining algorithms to predict the heart disease. Different algorithms gives various levels of accuracies. Here I am comparing the accuracies of few classification algorithms Random Tree, Naïve Bayes, Decision Tree and Random forest. The Hungarian_csv database with 294 instances and 14 attributes age, sex, cp, trestbps, chol, fbs, restecg, talach, exang, oldpeak, slope, ca, thal and num were used here for the analysis. RapidMiner Software is used to experiment the collected datasets. It is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. The collected datasets are passed as input to the above mentioned classification algorithms and the result obtained is analyzed and different views. It is found that Naïve Bayes gives the best accuracy of 79.25% with next 78.24% of accuracy by Decision Tree. Random tree gives 75.14% accuracy while Random forest stands next with 74.16%. The different measures and results were tabulated and charted.

Key words: Data Mining • RapidMiner • Random Tree • Naïve Bayes • Decision tree and Random Forest

INTRODUCTION

As people have interests in their health recently, development of medical domain application has been one of the most active research areas. One example of the medical domain application is the detection system for heart disease based on computer-aided diagnosis methods, where the data are obtained from some other sources and are evaluated based on computer-based applications [1]. Diagnosis of heart disease is a significant and tedious task in medicine. The term Heart disease encompasses the various diseases that affect the heart [2]. The healthcare environment is still information rich but knowledge poor.

The Proposed concept of the paper is given below in the Figure 1. The test and training data is given as input to the classification algorithms and the accuracy is compared for analysis.

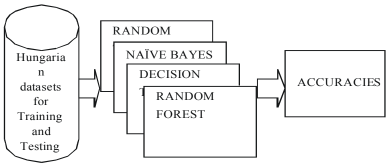

Fig. 1: Proposed Model

There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data [3]. The World Health Organization has estimated that 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths in the United States and other developed countries occur due to cardiovascular diseases [4]. As large amount of data is generated in medical organizations (hospitals, medical centers) but as this data is not properly used. There is a wealth of hidden information present in the datasets [5].

Corresponding Author: R. Jothikumar, Noorul Islam University, Kumaracoil, Tuckalay,
Kanyakumari Dt-629180, Tamil Nadu, India.

Diagnosing of heart disease is one of the important issue and many researchers investigated to develop intelligent medical decision support systems to improve the ability of the physicians [6]. Heart disease is the leading cause of death in the world over the past 10 years. Heart disease is a term that assigns to a large number of medical conditions related to heart [7]. Heart disease is a major health problem and it affects a large number of people. Cardiovascular Disease (CVD) is one such threat. Unless detected and treated at an early stage it will lead to illness and causes death [8]. Cardiovascular disease is the principal source of deaths widespread and the prediction of Heart Disease is significant at an untimely phase. In order to reduce number of deaths from heart diseases there has to be a quick and efficient detection technique [9].

Rapidminer: RapidMiner is one of the world's most widespread and most used open source data mining solutions. The project was born at the University of Dortmund in 2001 and has been developed further by Rapid-I GmbH since 2007. With this academic background, RapidMiner continues to not only address business clients, but also universities and researchers from the most diverse disciplines [10].

Data Mining And Classification: Data mining is used to discover the unknown knowledge from the known information and build predictive models. It is a step to discover knowledge from the data bases. This discovered knowledge can be utilized by the medical practitioners to reduce the time in diagnosis. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining techniques are used for variety of applications. Data mining techniques have been very effective in designing clinical support systems because of their ability to discover hidden patterns and relationships in clinical data. One of the most important applications of such systems is in diagnosis of heart disease [11]. Data Mining is an important extraction of hidden, unknown and potential helpful information about data.

Data mining gives a set of technique to find hidden or unknown pattern from data. Heart Risk (HR) prediction is important and complicated task that is essential to be executed efficiently and accurately diagnosing the heart problem based on doctor's knowledge and experience [12]. Data mining provides the methodology and technology to transform these heaps of data into useful information for decision making. By using data mining techniques it takes less time for the prediction of the disease with more accuracy. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information, to take decisions effectively, to discover the relations that connect parameters in a database is the subject of data mining [13]. Classification is an important data mining technique with broad applications to classify the various kinds of data used in nearly every field of human life [14].

Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets [15]. Several Data mining tools and Techniques are available to analyze the huge volume of health care data to predict life threatening diseases like Cancer, diabetics, Liver diseases and Heart diseases [16]. Organizations are maintaining history of data for future analysis [17]. Heart disease is one of the life threatening disease overall the globe. [18]

Random Tree: The design view of Random Tree is given in Figure 2 below which includes the UCI Cleveland dataset with 295 instances and 13 attributes, a cross validation Operator with 10 fold cross validation, Apply Model and Performance evaluation operator with necessary parameter values for execution.

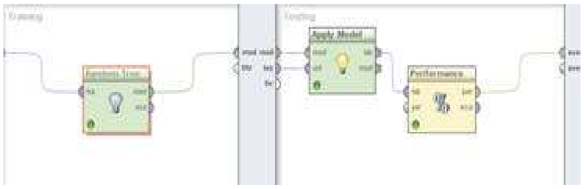

Fig. 2: Design view of Random Tree

The UCI Cleveland Switzerland dataset with 295 instances and 13 attributes is fed as input to the Random Tree in Rapid Miner. It gives the accuracy of 75.14%. The other related measures are as Kappa static 0.421, absolute error of 0.348%, relative error of 34.83% and Root mean squared error of 0.420.

The Confusion Matrix obtained is given below In Table 1 for analysis.

Table 1: Random Tree Confusion Matrix

	true '<50'	true '>50_1'	class precision
pred. '<50'	158	43	78.61%
pred. '>50_1'	30	63	67.74%
class recall	84.04%	59.43%	

The Tree obtained is given below in Figure 3 for reference.

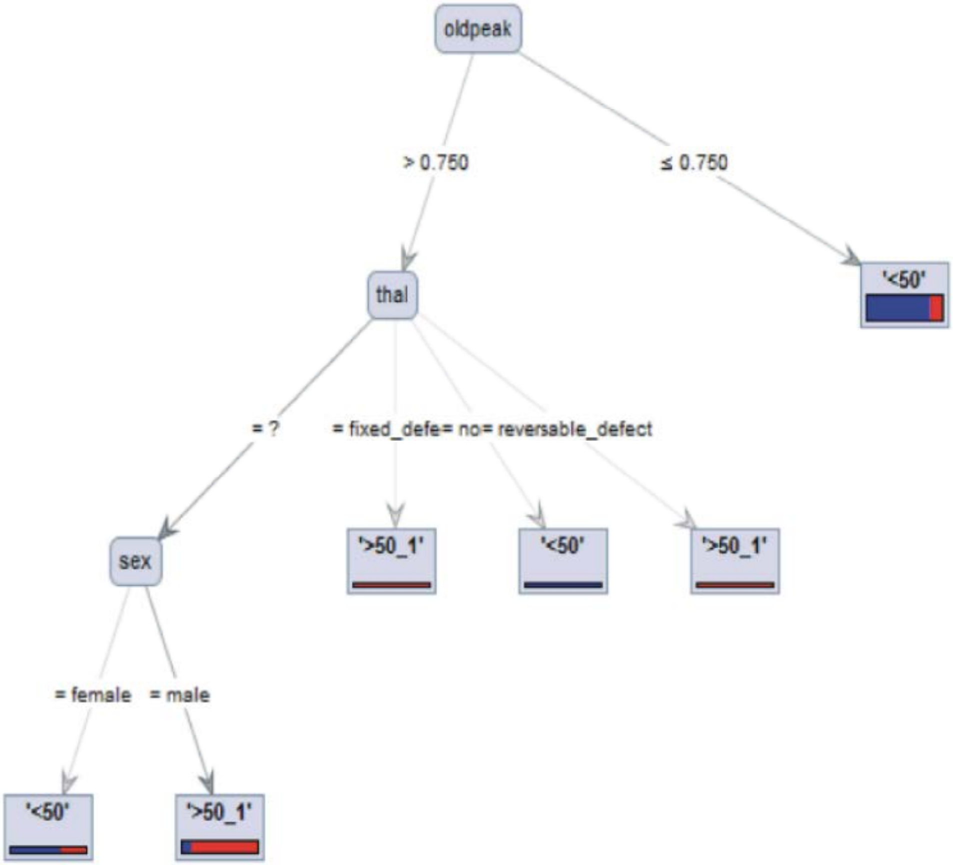

Fig. 3: Random Tree

Decision Tree: The design view of Decision Tree Classifier is given in Figure 4 below which includes the UCI Cleveland dataset with 295 instances and 13 attributes, a cross validation Operator with 10 fold cross validation, Apply Model and Performance evaluation operator with necessary parameter values for execution.

The design view of Random Forest Classifier is given in Figure 6 below which includes the UCI Cleveland dataset with 295 instances and 13 attributes, a cross validation Operator with 10 fold cross validation, Apply Model and Performance evaluation operator with necessary parameter values for execution.

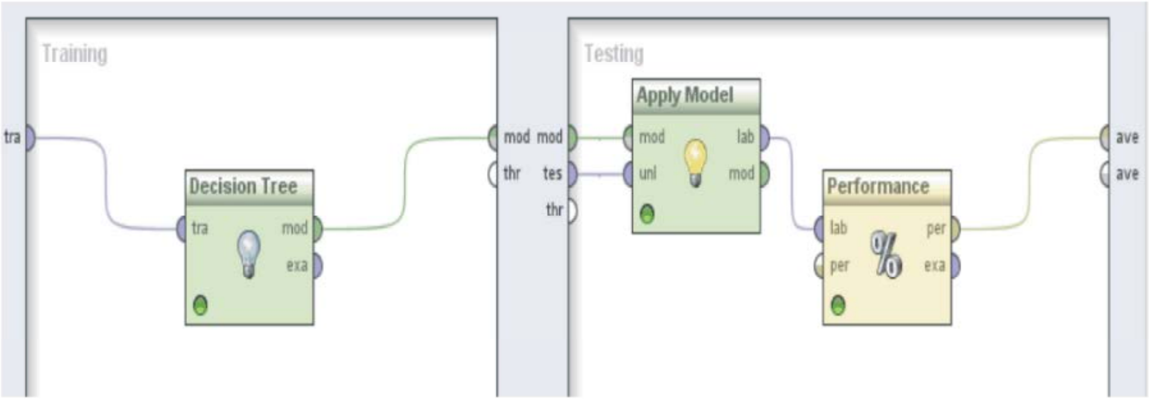

Fig. 4: Design View of Decision Tree

Random Forest: The UCI Cleveland Switzerland dataset with 295 instances and 13 attributes is fed as input to the Naïve Bayes Classifier in Rapid Miner. It gives the accuracy of 78.24%. The other related measures are as Kappa static 0.499, absolute error of 0.247%, relative error of 24.19% and Root mean squared error of 0.378.

The Confusion Matrix obtained is given below in Table 2 for analysis

Table 2: Decision tree Confusion Matrix.

	true '<50'	true '>50_1'	class precision
pred. '<50'	167	43	79.52%
pred. '>50_1'	21	63	75.00%
class recall	88.83%	59.43%	

The Tree obtained is given in Figure 5 below for reference.

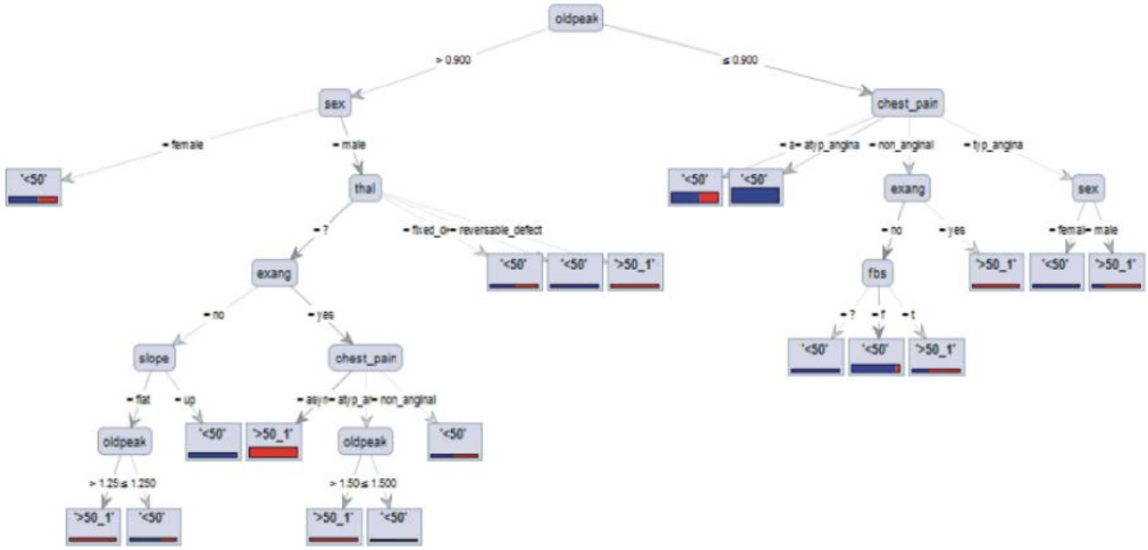

Fig. 5: Decision tree Generated

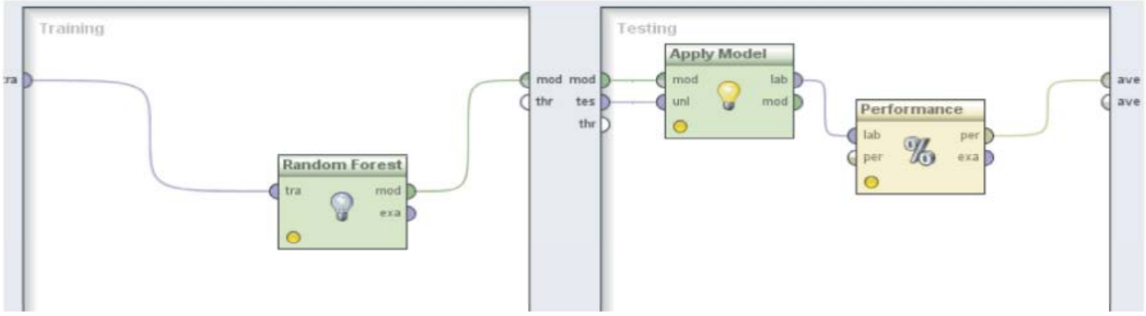

Fig. 6: Design view of Random Forest

The UCI Cleveland Switzerland dataset with 295 instances and 13 attributes is fed as input to the Random Forest Classifier in Rapid Miner. It gives the accuracy of 74.16%. The other related measures are as Kappa static 0.338, absolute error of 0.271%, relative error of 27.06% and Root mean squared error of 0.418.

The Confusion Matrix obtained is given in Table 3 below for analysis.

Table 3: Random Forest Confusion matrix

	true '<50'	true '>50_1'	class precision
pred. '<50'	181	69	72.40%
pred. '>50_1'	7	37	84.09%
class recall	96.28%	34.91%	

The Tree obtained is given in Figure 7 below for reference.

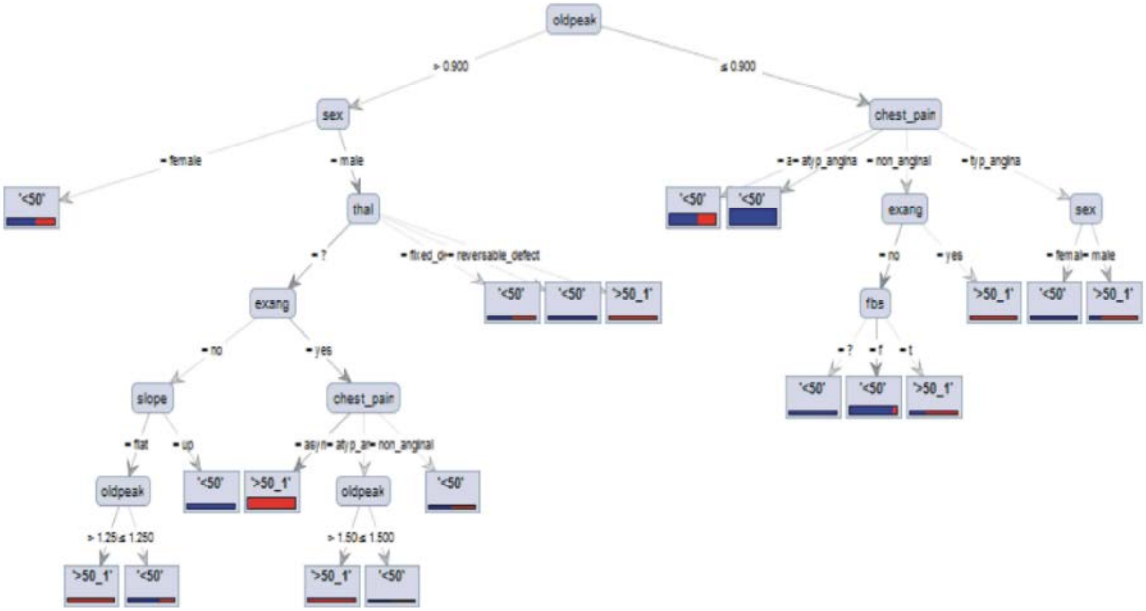

Fig. 7: Random Tree Generated

Naïve Bayes: The design view of Naïve Bayes Classifier is given below in Figure 8 which includes the UCI Cleveland dataset with 295 instances and 13 attributes, a cross validation Operator with 10 fold cross validation, Apply Model and Performance evaluation operator with necessary parameter values for execution.

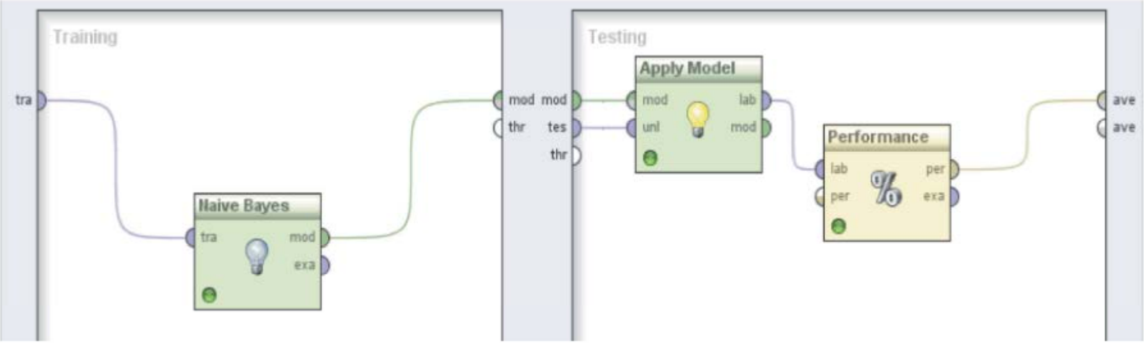

Fig. 8: Design View of Naïve Bayes

The UCI Cleveland Switzerland dataset with 295 instances and 13 attributes is fed as input to the Naïve Bayes Classifier in Rapid Miner. It gives the accuracy of 79.25%. The other related measures are as Kappa static 0.530, absolute error of 0.217%, relative error of 21.75% and Root mean squared error of 0.424.

The Confusion Matrix obtained is given below in Table 4 for analysis.

The accuracies obtained with different classifier is tabulated in the Table 5 below.

The accuracies of different classifier obtained is charted below in Figure 9 for analysis.

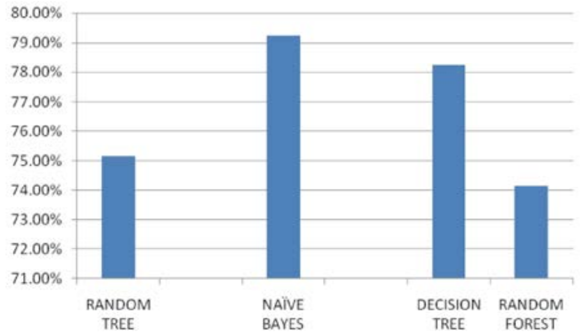

Fig. 9: Accuracies of different classifiers

Table 4: Naïve Bayes Confusion matrix

	true '<50'	true '>50_1'	class precision
pred. '<50'	167	40	80.68%
pred. '>50_1'	21	66	75.86%
class recall	88.83%	62.26%	

Table 5: Accuracies of classifiers

No	Classifier	Accuracy
1	Random Tree	75.14%
2	Naïve Bayes	79.25%
3	Decision Tree	78.24%
4	Random Forest	74.16%

CONCLUSION

The accuracies of different classifiers with UCI Cleveland dataset is experimented with the support of Rapid Miner software. The test and Training Datasets were passed as input to the Random tree, Naive Bayes, Decision tree and Random Forest. It is found that Naïve Bayes better accuracy of 79.25%, Decision Tree with 78.24%, Random Tree with 75.14% and Random Forest with 74.16%. In future, the accuracies of these algorithms can be further improved by preprocessing the datasets as the datasets may subject to noisy, inconsistent, missing and outdated values. These improved results can be used by healthcare professionals to predict the heart disease earlier.

REFERENCES

1. Anooj, P.K., 2012. "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules." Journal of King Saud University-Computer and Information Sciences, 24(1): 27-40.
2. Manikantan, V. and S. Latha, 2013. "Predicting the analysis of heart disease symptoms using medicinal data mining methods." International Journal of Advanced Computer Theory and Engineering, 2: 46-51.
3. Soni, Jyoti, *et al*., 2011. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications, 17(8): 43-48.
4. Soni, Jyoti, *et al*., 2011. "Intelligent and effective heart disease prediction system using weighted associative classifiers." International Journal on Computer Science and Engineering, 3(6): 2385-2392.
5. Medhekar, Dhanashree S., Mayur P. Bote and Shruti D. Deshmukh, 2013. "Heart Disease Prediction System using Naive Bayes." Heart Disease, 2(3).
6. Al-Milli, Nabeel, 2013. "Backpropagation Neural Network for Prediction of Heart Disease." Journal of Theoretical and Applied Information Technology, 56(1): 131-135.
7. Thenmozhi, K. and P. Deepika, "Heart Disease Prediction Using Classification with Different Decision Tree Techniques."
8. Venkatalakshmi, B. and M. Shivsankar, 2014. "Heart Disease Diagnosis Using Predictive Data mining." 2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14), Tamil Nadu, India.
9. Chandna, Deepali, 2014. "Diagnosis of heart disease using data mining algorithm." Int. J. Comput. Sci. Inf. Technol.(IJCSIT), 5(2): 1678-1680.
10. Goyal, Vishnu Kumar, "A Comparative Study of Classification Methods in Data Mining using RapidMiner Studio." IJIRSE) International Journal of Innovative Research in Science & Engineering, ISSN (Online), pp: 2347-3207.
11. Christopher, T., 2014. "Enhanced Heart Disease Analysis and Prediction System [EHDAPS] Using Data Mining." International Journal of Emerging Trends in Science and Technology, 1.09.
12. Radhimeenakshi, S. and G.M. Nasira, 2015. "Remote Heart Risk Monitoring System based on Efficient Neural Network and Evolutionary Algorithm." Indian Journal of Science and Technology, 8(14).
13. Kaur, Beant and Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques."
14. Patil, R.R., 2014. "Heart disease prediction system using Naive Bayes and Jelinek-mercer smoothing." Int. J. Adv. Res. Comput. Commun. Eng., (2014).

15. Jothikumar, R. and R.V. Sivabalan, 2015. Performance Analysis on Accuracies of Heart Disease Prediction System Using Weka by Classification Techniques. AJBAS, 9(7): 741-749.
16. Jothikumar, R., R.V. Sivabalan and A.S. Kumarasen, 2015. Data Cleaning Using Weka For Effective Data Mining In Health Care Industries. International Journal of Applied Engineering Research, 10(30).
17. Jothikumar, R., R.V. Sivabalan, Efficient Data Pre-Processing For Data Mining Using Neural Networks. Int. Journal of Scientific Research and Management Studies, 1(4): 118-123.
18. Jothikumar, R., R.V. Sivabalan and E. Sivarajan, 2015. Accuracies of j48 weka classifier with different supervised weka filters for predicting heart diseases, ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, 10(17): 7788-7793.