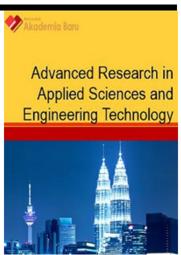
Penerbit Akademia Baru

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage: www.akademiabaru.com/araset.html
ISSN: 2462-1943


Akademia Baru
Advanced Research in Applied Sciences and Engineering Technology

Selecting the best model predicting based data mining classification algorithms for leukemia disease infection

Open

Fahd Sabry Esmail [1,*], Mohamed Badr Senousey [2], Mohamed ragaie sayed [3]

[1] Department of Management Information Systems, Modern Academy for Computer Science & Management Technology, Cairo, Egypt
[2] Department of Computer Sciences and Information Systems, Sadat Academy for Management Sciences, Cairo, Egypt
[3] Department of Computer Sciences and Information Systems, Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt

ARTICLE INFO

Article history:
Received 22 December 2016
Received in revised form 28 January 2017
Accepted 4 April 2017
Available online 21 April 2017

ABSTRACT

Yearly, thousands of people die of leukemia throughout the world due to the nature of Leukemia cells that become out of control and they spread randomly and the most effective way to reduce deaths from this disease is the early discovering, and this requires an accurate diagnosis. DNA microarrays help to discover the diseases, provide accurate medical diagnosis, and help to find the right treatment and cure for many diseases. This work presents a comprehensive comparative analysis of seventeen different classification algorithms with their performance evaluation by using five performance criteria for DNA microarray dataset applied on different machines. This study focused on finding the optimum algorithm for classification of data that can predict the occurrence of leukemia disease infection in earlier stage. The results indicated that the best algorithm based on the leukemia dataset is random tree classifier with an accuracy of 100% and the total time taken to build the model is at 0.01-0.03 seconds.

Keywords:
Data mining, Classification techniques, Leukemia diseases, DNA microarray

1. Introduction

One of the most important applications for predicting the disease infection is data mining. Data mining techniques have become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict disease infection using the historical datasets. Microarray technology offers the capacity to gauge expression levels of thou-sands of genes at the same time. These genes provide very valuable information which can be used to study any disease in depth. Study of genes from a cancer patient helps us diagnose cancer and differentiate between types of cancer. It also helps in separating the healthy people from the patients. Genes contains infinite patterns that cannot be recorded manually

[*] *Corresponding author.*
E-mail address: fahdsabry985@gmail.com (Fahd Sabry)

using a microscope. DNA microarrays are used to study the information obtained. Some of the application areas of DNA microarrays are obtaining the genes values from yeast in various ecological conditions and studying the gene expression values in cancer patients for different cancer types. DNA microarrays have huge potential scientifically as they can be useful in the study of genes interactions and genes regulations [1]. The feature extraction and classification are carried out with combination of the high accuracy of ensemble based algorithms, and comprehensibility of a single decision tree. These allow deriving exact rules by describing gene expression differences among significantly expressed genes in leukemia. It is evident from our results that it is possible to achieve better accuracy in classifying leukemia without sacrificing the level of comprehensibility. There are two common methods for in depth microarray data analysis such as clustering and classification [2]. Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic to the group [3]. Classification is supervised learning and also known as class prediction or discriminate analysis. Generally, classification is a process of learning-from-examples. Given a set of pre-classified examples, the classifier learns to assign an unseen test case to one of the classes.

Leukemia disease is a type of cancer caused by abnormal increase of the white blood cells. There are two main types of leukemia: acute leukemia and chronic leukemia. Acute leukemia can be either lymphocytic (ALL) or myeloid (AML), depending on which cells are under threat. Chronic leukemia can be either lymphocytic (CLL) or myeloid (CML) are categorized as leukemia diseases [4]. In general, leukemia is grouped by how fast it gets worse and what kind of white blood cells it affects [5]. The rest of this paper is organized as the follows. In Section 2, we discuss related works in this domain. In Section 3, we explore the methodologies used in this work. In Section 4, we present experimental results and analysis. In Section 5, we conclude the paper.

2. Literature Review

There are several gene selection methods for cancer classification using microarray datasets. However, most of them did not concentrate on identifying minimum number of informative genes with high classification accuracy [6]. Priyanga and Prakasam [7] developed a system called data mining based cancer prediction system. The main aim of this model was to provide the earlier warning to the users, and it was also cost and time benefit to the user. It predicts three specific cancer risks. Specifically, cancer prediction system estimates the risk of the breast, skin, and lung cancers by examining a number of user-provided genetic and non-genetic factors. DursunDelen *et al*. [8] used two popular data mining algorithms (artificial neural networks and decision trees) to develop the prediction models for breast cancer survivability using a large dataset (more than 200,000 cases). The results indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample, artificial neural networks came out to be the second with 91.2% accuracy and the logistic regression models came out to be the worst of the three with 89.2% accuracy. Thangaraju [9] built a model based as a test case on the University of California Irvine (UCI) repository dataset. The experiment had been performed with several data mining classification techniques and it was found that the Naive Bayes algorithm gives a better performance over the supplied data set with the accuracy of 83.4%. Jaya Suji and Rajagopalan [10] proposed work of research, the oral datasets were get form the various diagnostic centers which contained both cancer and non-cancer patients information and collected data had applied many classification algorithms on National Minimum Data Set (NMDS) dataset and the performance of those algorithms had been analyzed. A classification rate of 100% was obtained for C4.5 algorithm and classification rate of 98.7% was obtained for Random tree Algorithm and classification rate of 99.5% was obtained for Multilayer Perception Neural

Network (MPNN). In other study, Chandrasekar *et al*. [11] presented effective classification Techniques. The aim of the research is developing accurate prediction models for breast cancer using data mining techniques. After investigation of different classification Algorithm we have chosen 6 classifiers based on our simulation performance and we have used tree random classifier achieved overall classification accuracy 98%. Pushpalatha and Gupta [12] presented an ensemble model which was constructed to improve classification accuracy by combining the prediction of multiple classifiers. The performance measured gain, accuracy, specificity and sensitivity. From the experimental results concluded the ensemble model (random forest model) with feature selection achieved highest accuracy of 93.84% on test data.

3. Research Methodology

This research uses data mining technique for analysis and evaluation of classification algorithms about leukemia disease dataset through open source WEKA data mining techniques that generate predictive model to classification of leukemia disease infection, evaluate accuracies, and performance of several techniques.

3.1 Data Preprocessing

Preprocessing is one of the important and prerequisite steps in data mining. Feature selection (FS) is a process to select features which are more informative but some features may be redundant, and others may be irrelevant and noisy [13]. When the data set consists of meaningless data that is incomplete (missing), noisy (outliers) and inconsistent data, preprocessing of the dataset is required. Figure 1 shows the preprocessing steps.

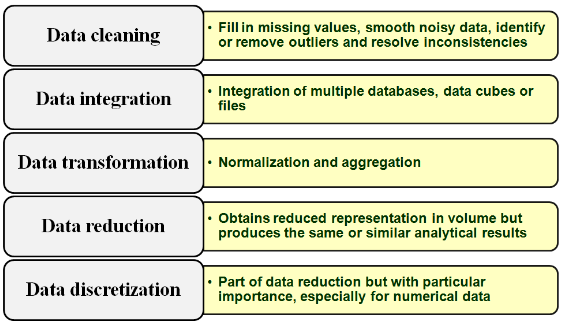

Data cleaning
• Fill in missing values, smooth noisy data, identify or remove outliers and resolve inconsistencies
Data integration
• Integration of multiple databases, data cubes or files
Data transformation
• Normalization and aggregation
Data reduction
• Obtains reduced representation in volume but produces the same or similar analytical results
Data discretization
• Part of data reduction but with particular importance, especially for numerical data

Fig. 1. Preprocessing techniques

3.2 Dataset Description

In this research, the researcher concentrates on the gene expression microarray dataset. To compare these data mining classification techniques and comparison analysis, we need the dataset. This research chooses the leukemia dataset from European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL -EBI) repository. These dataset are available online http://www.ebi.ac.uk/arrayexpress. Table 1 shows the summary of the characteristics of leukemia microarray dataset is described by the following parameters. Genes No.: the number of genes or attributes, Categories: the number of classes, Attributes Type: the type of data, Sample No.: the number of record or instance in the dataset.

Table 1
Description of Leukemia microarry dataset

Categories	Attribute Type	Genes No.	Sample No.
ALL AML MLL	Numeric	12582	15

3.3 Classification Algorithms

Classification is a data mining (machine learning) technique used to predict group membership for data instances [14]. It is the problem of finding the model for class attribute as a function of the values of other attributes and predicting accurate class assignment for test data. It can be divided in two types: supervised and unsupervised. Supervised is further divided in probabilistic and geometric. Probabilistic is further divided in parametric and nonparametric type. Classification is a two-step process: first is model construction i.e. describing a set of predetermined classes and second is using that model for prediction i.e. classifying future or unknown objects. For the classification in WEKA, the researcher has supervised and unsupervised categories of classifiers. All the classifiers like lazy, tree, rules and naïve comes under these categories only. The present research proposes a comprehensive analysis of different classification algorithms, and performance of evaluate by applying leukemia micro-array data set. A classifier model and other classification parameters will be obtained for the training dataset. Now this classifier model can be used for the test dataset to evaluate the model. The prediction about the test data set can be summarized on the basis of various performance criteria's.

3.4 Performance Factor Evaluation

This work adopted precision, recall and lift as the performance metrics for estimating the accuracy of a given classification model. Each of these was used where appropriate in the analysis of the performances. Apart from the major performance criteria mentioned, the work will also measure the response time of the classifiers.

3.4.1 Accuracy

Accuracy is the proportion of the total number of predictions that were correct. It is determined using the equation (1):

$$\text{Accuracy} = \frac{\text{tp+tn}}{\text{tp+tn+fp}} * 100 \; [15] \tag{1}$$

where, True Positive (TP) denotes the correct classifications of positive examples, true negatives (TN) are the correct classification of negative examples, false positive (FP) represents the incorrect classification of negative examples in correctly classified into the negative class.

3.4.2 Precision

Precision is the proportion of the predicted positive cases that were correct, as calculated using the equation (2) [15]:

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}} * 100 \quad (2)$$

3.4.3 Recall

Recall or Sensitivity or True Positive Rate (TPR) It is the proportion of positive cases that were correctly identified, as calculated using the equation (3) [15]:

$$\text{Recall} = \frac{\text{tp}}{\text{tp+fn}} * 100 \quad (3)$$

4. Experimental Results

In this section, the researcher conducted an experiment using WEKA (The Waikato Environment for Knowledge Analysis) application. WEKA is a comprehensive suite of Java class libraries that perform many advanced machine learning and data mining algorithms [16]. We analyze and compare the performance of decision tree algorithms namely Decicion Stump, forecast tree (FT), J48(C4.5), logical analysis of data(LADTtree), REPTree, logistic model tree(LMT), Naïve Bayes(NBTree), Calcification and regression tree (CART), Random Forest and Random Tree, and compare the performance of Rule classifier algorithms namely Java repeated incremental pruning (JRip), Nearest-Neighbor-like algorithm (NNge), One Rule, PART, Ripple Down Rule learner (Ridor), Zero Rule [17].

4.1 Measuring Accuracy

This approach has been implemented on two different machines (M1 and M2), as shown in Table 2. The simulation results are partitioned into several sub items for easier analysis and evaluation. Different performance matrix like accuracy, time taken to build model (Seconds), true positive rate, false positive rate, precision, recall are presented in numeric value during training and testing phase. The summary of those results by running the techniques in WEKA is reported in Tables 3-6

Table 2
Description of Machines

Machine Name	Specification
M1	Intel Core 2 Due 2.13 GHz Processor 4 GB RAM
M2	Intel 3.00 GHz Processor 2 GB RAM

Table 3
Accuracy Measure for Classification Rule Algorithms (M1)

Methods	Accuracy %	Recall	Precision	FP Rate	TP Rate	Time Taken to Build Model (Seconds)
JRIP	80	0.8	0.665	0.229	0.8	0.55
NNge	100	1	1	0	1	0.72
One R	80	0.8	0.655	0.229	0.8	0.12
PART	100	1	1	0	1	0.27
Ridor	80	0.8	0.665	0.229	0.8	0.39
Zero R	53.33	0.533	0.284	0.533	0.533	0

Table 4
Accuracy Measure for Classification Tree Algorithms (M1)

Methods	Accuracy %	Recall	Precision	FP Rate	TP Rate	Time Taken to Build Model (Seconds)
BF Tree	53.33	0.533	0.284	0.533	0.533	2.69
Decision Stump	80	0.8	0.686	0.073	0.8	0.14
FT	53.33	0.533	0.284	0.533	0.533	1.1
J48(C4.5)	100	1	1	0	1	0.16
LADTree	100	1	1	0	1	4.84
REP Tree	73.33	0.733	0.587	0.201	0.733	0.21
LMT	100	1	1	0	1	7.56
NBTree	100	1	1	0	1	1.84
CART	53.33	0.533	0.284	0.533	0.533	2.19
Random Forest	100	1	1	0	1	0.28
Random Tree	100	1	1	0	1	0.01

Table 5
Accuracy Measure for Classification Rule Algorithms (M2)

Methods	Accuracy %	Recall	Precision	FP Rate	TP Rate	Time Taken to Build Model (Seconds)
JRIP	80	0.8	0.665	0.229	0.8	1.16
NNge	100	1	1	0	1	1.28
One R	80	0.8	0.655	0.229	0.8	0.19
PART	100	1	1	0	1	0.66
Ridor	80	0.8	0.665	0.229	0.8	1.11
Zero R	53.33	0.533	0.284	0.533	0.533	0

Table 6
Accuracy Measure for Classification Tree Algorithms (M2)

Methods	Accuracy %	Recall	Precision	FP Rate	TP Rate	Time Taken to Build Model (Seconds)
BF Tree	53.33	0.533	0.284	0.533	0.533	4.77
Decision Stump	80	0.8	0.686	0.073	0.8	0.34
FT	53.33	0.533	0.284	0.533	0.533	1.8
J48(C4.5)	100	1	1	0	1	0.3
LADTree	100	1	1	0	1	5.34
REP Tree	73.33	0.733	0.587	0.201	0.733	0.45
LMT	100	1	1	0	1	9.09
NBTree	100	1	1	0	1	4.59
CART	53.33	0.533	0.284	0.533	0.533	8.47
Random Forest	100	1	1	0	1	1.35
Random Tree	100	1	1	0	1	0.03

Figures 2 and 3 shows the comparison based about the accuracy by each learning algorithm.

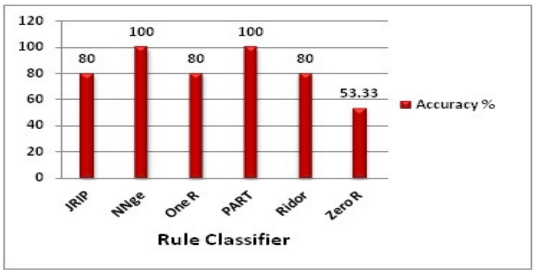


Fig. 2. Accuracy % of Rule Classifiers

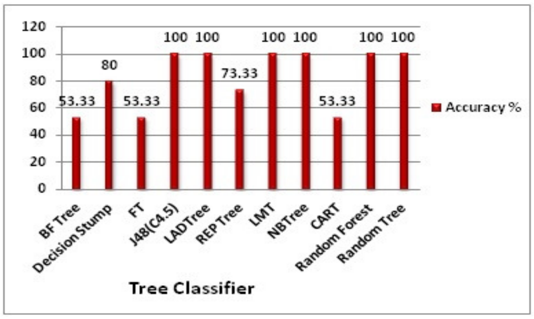


Fig. 3. Accuracy % of Tree Classifiers

Based on Figures 2 and 3, we can clearly see that the highest accuracy is 100% and the lowest is 53.33%. In fact, the highest accuracy belongs to the NNge and PART from rule classifier and J48, LAD tree, LMT, NBtree, random forest and random tree from tree classifier. The total time required to build the model is also a crucial parameter in comparing the classification algorithm. In this simple experiment, from Tables 3-6, we can say that a Zero R from rule classifier requires the shortest time which is around 0 seconds consecutive with compared to random tree from tree classifier which requires the longest model building time which is around 0.01-0.03 seconds.

4.2 Comparisons between Accuracy and Response Time on Different Machines

The total time required to build the model is also a crucial parameter in comparing the classification algorithm. In Tables 7 and 8, the researcher has summarized two main measures of evaluation for each algorithm such as time taken to build the model and accuracy.

Table 7
Comparison of Rule Classifier Methods

Methods	Accuracy %	Time Taken to Build Model(M2) (Seconds)	Time Taken to Build Model(M1) (Seconds)
JRIP	80	1.16	0.55
NNge	100	1.28	0.72
One R	80	0.19	0.12
PART	100	0.66	0.27
Ridor	80	1.11	0.39
Zero R	53.33	0	0

Tables 7 and 8 show that NNge from rule classifier take maximum amount of time to build the model i.e. is around 0.72-1.28 seconds. Next highest LMT is around 7.56-9.09 and LADtree 4.84-5.34 seconds to build the model from tree classifier. In terms of second measure of evaluation, Random tree has the highest percentage of accuracy is 100% and has the longest model building time which is around 0.01-0.03 seconds and the best Measure among all and the Next highest accuracy is 100% belongs to PART and take time to build model is around 0.27-0.66 seconds that also lowest time compare to others. Hence, we conclude that random tree has performed better than all the other classifiers in the analysis by two machines of our dataset because it has the potential to significantly improve the conventional classification methods to be used in the medical field or in general, bioinformatics field.

Table 8
Comparison of Tree Classifier Methods

Methods	Accuracy %	Time Taken to Build Model (M2) (Seconds)	Time Taken to Build Model(M1) (Seconds)
BF Tree	53.33	4.77	2.69
Decision Stump	80	0.34	0.14
FT	53.33	1.8	1.1
J48(C4.5)	100	0.3	0.16
LADTree	100	5.34	4.84
REP Tree	73.33	0.45	0.21
LMT	100	9.09	7.56
NBTree	100	4.59	1.84
CART	53.33	8.47	2.19
Random Forest	100	1.35	0.28
Random Tree	100	0.03	0.01

4.3 Analysis of Classification Algorithms Results

This study, has examined the performance of different classification methods that could generate accuracy and predict best model to disease infection diagnosis the data set. According to Figures 2 and 3; Tables 3-8, we can clearly see the highest accuracy is 100% belongs to random tree, J48, LAD tree, LMT, NBtree, random forest, NNge and PART classifier and lowest accuracy is 53.33% that belongs to Zero R. The total time required to build the model is also a crucial parameter in comparing the classification algorithm. Based on Tables 3-8, we can compare time taken to build model among different classifiers in WEKA. We clearly find out that Random tree is the optimum; second best is the J48 and PART method is third best. An algorithm which has highest accuracy and lowest time to build model will be preferred as it has more powerful classification capability and ability in terms of bioinformatics fields.

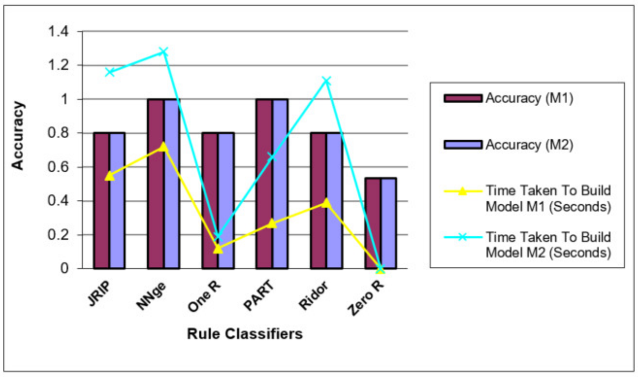


Fig. 4. Accuracy, T1 and T2 of Rule Classifiers

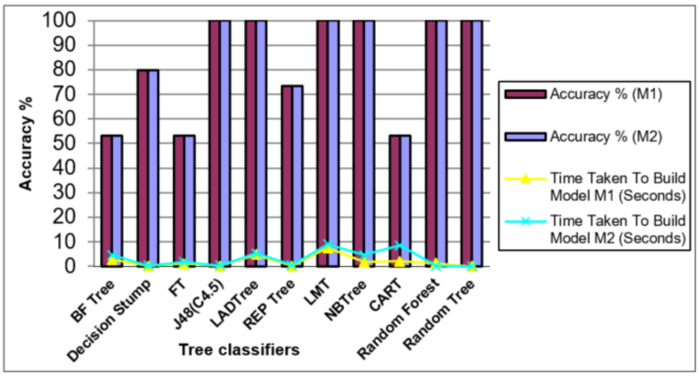


Fig. 5. Accuracy, T1 and T2 of Tree Classifiers

Based on Figures 4 and 5, it can be concluded that the PART methods is best comparatively others rule classifiers cause 100% accuracy achieved by PART and take the time to build model is around 0.27-0.66 seconds that also lowest time compare to others. In fact, the highest accuracy belongs to the decision tree classifier by Random tree has the highest percentage of accuracy is 100% and has the longest model building time which is around 0.01- 0.03 seconds and the best Measure among all.

5.0 Conclusion

In this study, we compared the classification accuracy and response time of different classification algorithms. The algorithms performance has been evaluated by using leukemia dataset applied on different machines. This study focused on finding the right algorithm for classification of data that works better on dataset that predict the leukemia disease infection in earlier stage. The results indicate that random tree of the classifiers outperformed all others in terms of the accuracy when applied to the data because gave the best accuracy, recall and precision on two machines. However, it is observed that the accuracies of the tools vary depending on the dataset used but also the response time taken to build model varies according to the machines used.

References

[1] Armananzas, Rubén, Borja Calvo, Inaki Inza, Marcos López-Hoyos, Víctor Martínez-Taboada, Eduardo Ucar, Irantzu Bernales, Asier Fullaondo, Pedro Larranaga, and Ana M. Zubiaga. "Microarray analysis of autoimmune diseases by machine learning procedures." *IEEE Transactions on Information Technology in Biomedicine* 13, no. 3 (2009): 341-350.

[2] Mutch, David M., Alvin Berger, Robert Mansourian, Andreas Rytz, and Matthew-Alan Roberts. "Microarray data analysis: a practical approach for selecting differentially expressed genes." *Genome biology* 2, no. 12 (2001): preprint0009-1.

[3] Rajeswari, B., and Aruchamy Rajini. "Survey On Data Mining Algorithms to Predict Leukemia Types." (2010).

[4] Dash, Sujata, Bichitrananda Patra, and B. K. Tripathy. "A hybrid data mining technique for improving the classification accuracy of microarray data set." *International Journal of Information Engineering and Electronic Business* 4, no. 2 (2012): 43.

[5] Madhukar, Monica, Sos Agaian, and Anthony T. Chronopoulos. "Deterministic Model for Acute Myelogenous Leukemia Classification." In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pp. 433-438. IEEE, 2012.

[6] Oprea, Cristina. "Performance evaluation of the data mining classification methods." *Annals of the "Constantin Brâncuşi" University of Târgu Jiu, Economy Series, Spec* 4 (2014): 249-253.

[7] Priyanga, A., and S. Prakasam. "Effectiveness of Data Mining-based Cancer Prediction System (DMBCPS)." *International Journal of Computer Applications* 83, no. 10 (2013).

[8] Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." *Artificial intelligence in medicine* 34, no. 2 (2005): 113-127.

[9] Thangaraju, P., G. Barkavi, and T. Karthikeyan. "Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques." *International Journal of Advanced Research in Computer and Communication Engineering Vol* 3 (2014).

[10] Suji, R. Jaya, and S. P. Rajagopalan. "An automatic oral cancer classification using data mining techniques." *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)* 2, no. 10 (2013): 3579-3765.

[11] Chandrasekar, R. M., and V. Palaniammal. "Performance and Evaluation of Data Mining Techniques in Cancer Diagnosis." *IOSR Journal of Computer Engineering* 15, no. 5: 39-44.

[12] Pujari, Pushpalata, and Jyoti Bala Gupta. "Improving classification accuracy by using feature selection and ensemble model." *International Journal of Soft Computing and Engineering* 2, no. 2 (2012): 380-386.

[13] Velayutham, C., and K. Thangavel. "Unsupervised quick reduct algorithm using rough set theory." *journal of electronic science and technology* 9, no. 3 (2011): 193-201.

[14] Phyu, Thair Nu. "Survey of classification techniques in data mining." In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, pp. 18-20. 2009.

[15] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

[16] Landwehr, Niels, Mark Hall, and Eibe Frank. "Logistic model trees." *Machine Learning* 59, no. 1-2 (2005): 161-205.

[17] Ashish Kumar Dogra, Tanuj Wala, "A Review Paper on Data Mining Techniques and Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 5, May (2015).