DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.4870

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

COMPARISON OF CLASSIFICATION TECHNIQUES ON HEART DISEASE DATA SET

K.K.Revathi
Research Scholar,
Dept. of Computer Science, Selvamm Arts & Science
College (Autonomous) Namakkal.
Tamilnadu, India

K.K.Kavitha
Vice Principal, Head of the Department,
Dept. of Computer Science, Selvamm Arts & Science
College (Autonomous) Namakkal.
Tamilnadu, India

Abstract: Today's world in all fields extract useful knowledge from data, Data Mining is an analytic process designed to explore data. Classification analysis is one of the main techniques used in Data Mining. The classification method analyses the data from different perspectives and evaluates their accuracies and detecting a problem in various aspects. In this work, we focus on Heart Disease problem; specifically we took University of California, Irvine (UCI) heart disease dataset and WEKA (Waikato Environment for Knowledge Analysis) data mining tool. Various researches have investigated this dataset for better performance measures. In our paper does a comparative study of commonly used machine learning algorithms to detecting heart diseases. The aim of this research to judge the accuracy of different data mining algorithms such as NaiveBayes, IBK, RandomForest on heart disease dataset and determine the optimum algorithm for detection of heart disease.

Keywords: Data Mining, Classification, Bayesian, lazy, IBK, random forest, WEKA, naive bayes

I. INTRODUCTION

Heart Disease is refers to in a body various types of conditions that not work properly, which can affect heart function. A heart attack symptoms occurs, differ for different ages. It's not come obvious. Most of the people they are not concentrate first signs of their problem. At last, its starts to die [1]. Nowadays many people affected by heart disease, we have only less number of expert doctors in our country. So through this process can find out correct classification algorithm to detect heart disease problem in early.

Data mining is simply to extract useful knowledge from large amount of dataset. It is used in multiple purpose in the field of Information Technology, for instance this data mining techniques are used in many application areas such as, database technique, data visualization, machine learning, pattern identification, information retrieval, statistical analysis, neural networks, knowledge-based systems, artificial intelligence systems and computational performance [2]. Data mining is nothing to extraction of particular knowledge. Three popular techniques are used to analyze the information, these are, association rule mining, classification and clustering. Data mining has attracted a great impression in the information industry and in society as a whole in past decades. Association rule mining, clustering, classifications are doing their unique role accurately. In these three we used classification technique. The classification is the process of finding a model. A model is called as constructing a class label to a set of unclassified classes. In supervised classification, the set of possible classes is known in earlier. Using data mining classification techniques on heart disease data set with help of weka tool, we can detect whether the person attack heart disease or not. To utilize this method we can detect the problem of heart disease, it may help to reduce death rate.

In this paper we described the analysis of three different algorithms (naïve bayes, IBK, random forest) on the basis of various performance parameters. Simple definitions about these three algorithms are,

- The first approach is NaiveBayes classifier. It is one of the techniques of bayes classification. It uses probabilistic feature based on applying bayes theorem with strong assumptions between the features. It is extremely scalable classifier. It mainly used in medical field.
- The second approach is IBK classifier. IBK may refer to: IBK algorithm is implements from the k-nearest neighbor algorithm, In weka it is called as IBK (Instance-Based learning with parameter k) and it is in lazy class folder. In machine learning it sometimes called as memory-based learning. k- Nearest neighbor algorithm, kernel machines and RBF networks are example of IBK classifier.
- The third approach is RandomForest classifier. RandomForest algorithm is comes under tree classification technique. It is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. In machine learning way of saying the RandomForest classifier. RandomForest algorithm can use both for classification and the regression kind of problems.

II. LITERATURE SURVEY

Many authors doing research in prediction of heart disease symptoms using support vector machine, logistic regression, neural networks, decision tree, NaïveBayes, C4.5 algorithm methods. Everyone used different type of databases, parameters and algorithms for delved exact result. Mostly data mining methods are used in medical field for predict the problem of heart disease, diabetes, breast cancers, drug causes, AIDS etc.,

The authors told about medicinal data mining methods [3]. Here they used C4.5 algorithm, Maximal Frequent Item set algorithm, K-means clustering algorithm for their analysis. They collected patient's databases from experience specialists. Diagnosis of heart disease is important task in medicinal data mining. Maximal Frequent Item set Algorithm used association rule mining procedure. Item set algorithm is used to classify the data based on cross validations and partitions techniques. C4.5 algorithm is a training set algorithm; it shows about heart attack rank with a tree structure. K-means clustering algorithm is used to cluster database, which will remove positive feedbacks from database. Due to these methods they predict heart disease symptoms successfully.

They used data mining classification techniques [4]. Those are support vector machine (svm), logistic regression and neural networks, to predict the prevalence of heart disease. They used Cleveland dataset for their analysis and evaluate the accuracy. From their research Logistic Regression and SVM gives lesser complex models and give more accurate result. They used F1 score and ROC curves to evaluate measures. Through this process they told about early prediction of heart disease.

They proposed a new method of homogenous data mining technique and hybrid data mining algorithms. It is used to predict the early prognosis of cardiovascular diseases [5]. In this paper they gave detailed account of support vector machine, decision trees, and logistic regression algorithms and rule based approach. Rule based model compare the accuracy of applying rules to individual algorithms on Cleveland heart disease database in order to provide an accurate model of predicting heart disease.

The relative study on another research, these authors analyzed different data mining tools using different data mining techniques [6]. This paper compares performance of different data mining tools like WEKA, XLMiner and KNIME with the methods of classification, clustering and association rule mining. They used Statlog heart disease dataset for analyzing performance of tools and techniques.

They are developed prediction models for heart disease survivability [7]. They used three famous algorithms extracted from a decision tree or rule-based classifier to develop the prediction model using a large set of database. They used 10-fold cross validation methods and rule based classifier on CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and Decision Table (DT) algorithms.

III. METHODOLOGY

A classifier is a supervised function (machine learning tool) where the learned (target) attribute is categorical (nominal) [7]. Methodology incorporate a brief explanation about data set, data mining tool, classification of algorithms etc., The main aim of this paper is to find best classification algorithm for heart disease detection between Bayesian (naïve bayes), Lazy (IBK) and Tree (random forest) classifiers.

A. *Dataset Description*

We have tested the Statlog Heart Disease Dataset (SHDD) on different data mining classification methods such as bayes, lazy, trees. This dataset is taken from the University of California, Irvine (UCI) Machine Learning Database. It contains totally 270 instances and 14 attributes of healthy persons and patients with heart problem [8-10].

1) age
2) sex
3) chest
4) resting_blood_pressure
5) serum_cholestoral
6) fasting_blood_sugar
7) resting_electrocardiographic_results
8) maximum_heart_rate_achieved
9) exercise_induced_angina
10) oldpeak
11) slope
12) number_of_major_vessels
13) thal
14) class

The type of class attribute is nominal and it have two distinct values, it included the label as absent and present regarding the absence and presence of heart disease respectively.

No	Age	sex	chest	resting	serum	fasting	Electro
1	70.0	1.0	4.0	130.0	322.0	0.0	2.0
2	67.0	0.0	3.0	115.0	564.0	0.0	2.0
3	57.0	1.0	2.0	124.0	261.0	0.0	0.0
4	64.0	1.0	4.0	128.0	263.0	0.0	0.0
5	74.0	0.0	2.0	120.0	269.0	0.0	2.0
6	65.0	1.0	4.0	120.0	177.0	0.0	0.0
7	56.0	1.0	3.0	130.0	256.0	1.0	2.0
8	59.0	1.0	4.0	110.0	239.0	0.0	2.0
9	60.0	1.0	4.0	140.0	293.0	0.0	2.0
10	63.0	0.0	4.0	150.0	407.0	0.0	2.0
11	59.0	1.0	4.0	135.0	234.0	0.0	0.0
12	53.0	1.0	4.0	142.0	226.0	0.0	2.0
13	44.0	1.0	3.0	140.0	235.0	0.0	2.0
14	61.0	1.0	1.0	134.0	234.0	0.0	0.0

Figure 1: Data set sample

B. *Tool Description*

WEKA (Waikato Environment for Knowledge Analysis) is open source software issued under the GNU General public License [11]. Named after a flightless New Zealand bird, Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka GUI Chooser is appears on the first screen on the weka tool. The GUI Chooser window encompasses Simple CLI, Explorer, Experimenter, Knowledge Flow methods.

Machine Learning is nothing but a type of artificial intelligence which enables computers to learn the data without help of any explicit programs. Machine learning systems crawl through the data to find the patterns and when these are found, adjust the programs actions accordingly.

Weka data formats: Weka uses the Attributes Relation File Format for data analysis by default. But listed below are some formats that weka supports from where data can be imported:

- CSV
- ARFF
- Database using ODBC

C. *Classification*

Classification is finding a set of model to applying further process of an analysis. The goal of classification is to accurately predict the target class for each case in the data [6]. Different classification algorithms use different techniques for finding relationships [12]. These relationships are summarized in a certain format, to evaluate the result in a different manner to find out accuracy. In this research, we have analyzed three classifiers namely Bayesian, lazy and trees. In Bayesian classifier, we have analyzed the classification algorithm namely naïve bayes, in lazy classifier we have analyzed the classification algorithm IBK, in tree classifier we have analyzed the classification algorithm namely random forest.

D. Bayesian Classification

Bayesian classifiers are statistical classifiers. It's based on bayes theorem. A class is to predict the features value of the class. The Bayes rule is used for a given a set of features for the known class [13]. A bayes classifier includes NaiveBayes, NaiveBayes Multinominal, Naivebayes Simple, NaiveBayes Updateable algorithms and so on. From these NaiveBayes is simple and better to apply large set of databases.

Naïve Bayes Classifier

Naïve Bayesian classifiers assume attribute value on a given class is individual value of the other attributes. This assumption is called class conditional independence. It made to simplify the computations involved and in this sense, it is called as naïve.

Bayes theorem formula: bayes theorem formula is an important method for calculating conditional probabilities. It is used to calculate posterior probabilities. Bayes theorem describes the probability of an event, based on conditions that sometimes might be related to the event [14].

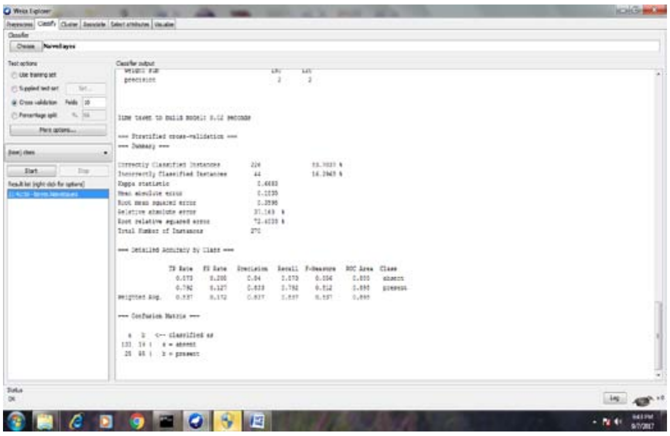
For instance, a patient can observed some symptoms about heart disease, then we can apply patient database to bayes formula to easily find out that person affected by heart disease or not. Because, bayes theorem telling about probabilistic nature for a patient need treatment or not. In simple words, suppose a doctor is interested in whether a person has heart disease, and knows the persons age. If heart disease is related to age, then using bayes theorem, the person's age can be used to access more accurate probability that the patient have heart disease [15].

Thomas Bayes introduced Bayes theorem, who first provided an equation that allows new evidence to update beliefs.

$$P\,(Q|P) = P\,(P|Q)\;P\,(Q)\,/\,P\,(P)$$

$$P\,(P|Q) = P\,(Q|P)\;P\,(P)/\,P\,(Q)$$

- P (P|Q) is the probability of P and Q has occurred and is known as possibility.
- P(Q) is notable as prior probability and P (Q|P) is notable as posterior probability.



```
 weight sum                              150      120
 precision                                 2        2


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         226               83.7037 %
Incorrectly Classified Instances        44               16.2963 %
Kappa statistic                          0.6683
Mean absolute error                      0.1839
Root mean squared error                  0.3598
Relative absolute error                 37.163  %
Root relative squared error             72.4033 %
Total Number of Instances              270

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.873     0.208      0.84      0.873     0.856      0.899    absent
                 0.792     0.127      0.833     0.792     0.812      0.899    present
Weighted Avg.    0.837     0.172      0.837     0.837     0.837      0.899

=== Confusion Matrix ===

   a   b   <-- classified as
 131  19 |   a = absent
  25  95 |   b = present
```

Figure 2: Implementation of Naïve Bayes Classifier

E. Lazy Classification

Lazy is a classification technique, it includes IBK, KStar, LWL methods to classify the database. Lazy method doesn't do anything until last minute. They always use any other model for classification purpose. The lazy learners use the same dataset as both training set and testing set. Before doing construction model in order to classify a given test set [12]. That is when a given training set, a lazy learner first stores training set and waits until a test set. When it sees the test set it will do generalization in order to classify the training set based on its similar property, at last the classifier store the training set. Unlike eager learning methods, lazy learners do less work. Based on this unique nature lazy learners store the training set before classification. Lazy learning solve multiple problems consecutively and deal the problem area in successful.

IBK Classifier

In weka it's called IBK (instance based learning with parameter k) and it's the lazy class folder. The term learning is indicating store all training instances. K- Nearest Neighbor (KNN) is an example of IBK Classifier [9]. KNN algorithm is used to specify the number of nearest neighbors to use when classifying a test instances and the outcome is determined by majority vote. Opening of IBK classifier have following steps. The first step to choose weka Explorer initially, then choose dataset, and choose classify tap to get options from IBK implementation. It has the cross-validation option that can help by choosing the best value automatically. weka uses cross-validation to select the best value for KNN. Example of instance-based learning algorithm is k-nearest neighbor algorithm, weighted regression, case-based reasoning, kernel machines and RBF networks.

K -Nearest neighbor

Features

All instances correspond to data point in an n-dimensional Euclidean space. Classification is postponed until a new instance arrives. Classification is done by compare feature vector of different data points. The end function may be real valued or discrete function.

The arbitrary instance is represented by,
(x1(a),x2(a),x3(a),….,xn(a))

- Let xi(a) denotes instances

Euclidean distance between two instances s(ai,aj)=sqrt (sum for r=1 to n(xr(ai)-xr(aj))2) sometimes, target function is real valued or discrete function. If it is continuous valued target function means,

- The mean value of the k nearest training examples.

F. Tree Classification

Decision trees are easy to use and easy to understand supervised learning algorithms. Tree classifier incorporate many algorithms to make different kinds of analysis. The algorithms are, ADTree, J48, RandomForest, SimpleCart and so on. These are used successfully for the following types of problems: loan applications, medical diagnosis, movie preferences, spam filters. The objective is to attain faultless classification with minimum number of decision option, even though it is not always possible in noise or inconsistent data [12]. Decision tree classifier performs glowing with large datasets and can handle numerical as well as categorical data.

Random Forest classifier

The most popular classification algorithm is the random forest algorithm. In machine learning, the way of saying random forest classifier [16-17]. Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. Random forest algorithm can use both for classification and the regression kind of problems.

IV. Performance of the Classifiers

The classifiers performance is evaluating under the following factors,

(1) The time to building a model
(2) Classifying correctly and incorrectly instances depends on their attributes
(3) The accuracy of the classifiers

Table I: Performance of the classifiers

Evaluation Criteria	**Classifiers**		
	Naïve Bayes	**IBK**	**Random Forest**
Timing to build model (in Sec)	0.09	0	0.48
Correctly classified instances	226	203	220
Incorrectly classified instances	44	67	50
Accuracy (%)	83.70%	75.18%	81.48%

Correctly classified instances are taken as accuracy in our research. NaïveBayes is better algorithm for heart disease detection compare to other classification algorithms. NaïveBayes is a part of bayes approach. It gave 83.70 percentages accuracy, and 226 correctly classified instances also take less time to build a model. On the other hand, other two algorithms are IBK and RandomForest, these are comes under lazy and tree approach, these algorithms are took incorrectly classified instances is high and process time also high. The accuracy of IBK classifier is 75.18 percentages and the accuracy of 81.48 percentages for random forest. From these process IBK and RandomForest algorithms are not suitable for better heart disease detection. NaïveBayes is a best technique in heart disease database to get an accurate result.

In future we have desired to develop web mining and text mining methods to get a better prediction of heart disease symptoms.

System architecture of this research work,

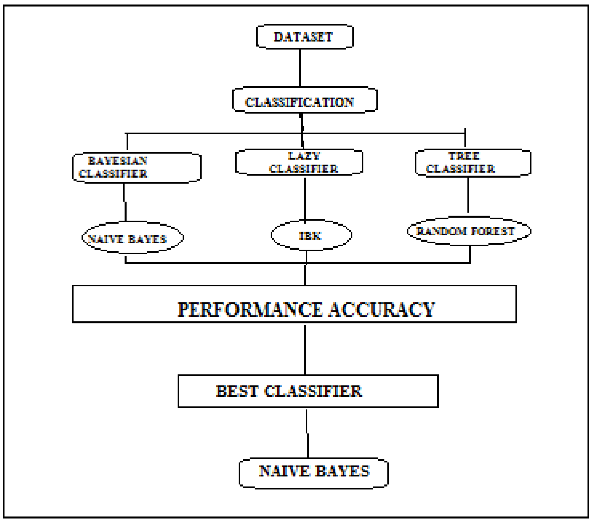


Figure 3: System Architecture of Classification Algorithms

V. Result and Discussion

The data set consist of 270 patient's record. Among them, 44 or 16.29% are reported to have Heart disease while the remaining 226 or 83.70% are not having any symptoms in NaiveBayes approach. The other two methods are classified as 75% patient's are not affected by heart disease, rest of 24% are possible to heart attack and RandomForest give 81% guarantee to healthy records but it classify 18% as possible to heart attack in future. In order to verify the prediction results comparison of the three popular data mining algorithms in the help of 10-fold crossover validation. The k-fold crossover validation is usually used to reduce the error resulted from random sampling in the comparison of the accuracies of a number of prediction models. The entire set of data is randomly divided into k folds with the same number of cases in each fold. The training and testing are performed for k times and one fold is selected for further testing while the rest are selected for further training.

The present study divided the data into 10 folds. The 1 fold is testing and 9 folds are training for the 10-fold crossover validation. NaïveBayes Approach checking the probability model based on Bayes theorem, IBK and RandomForest checking like 1 nearest neighbor and Bagging with 100 iterations, base learners simultaneously.

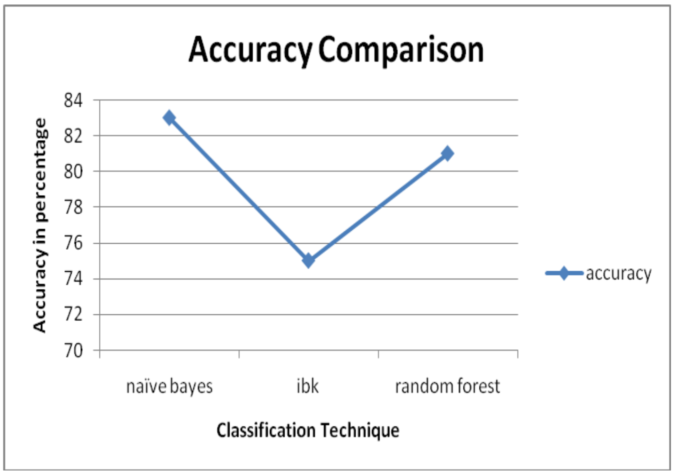


Figure 4: Accuracy comparison of the Classifiers

VI. Conclusion and Futurework

This paper initially gives brief introduction about Data Mining, Weka tool and heart disease dataset. Weka is a powerful tool used in data mining techniques. The dataset contains 270 instances. The classification algorithms are classifying the data into correctly; incorrectly instances and also it classify the database in different manner. Through this effort we can find out Naïve bayes algorithm give better result compare to IBK and random forest algorithm. These algorithms are used in many areas like banking, medical and stock market analysis to predict good result.

Futurework

Future research should concentrate on collecting data from a more recent time period and real symptoms record of the patients. Our dataset contains only 270 instances. This is very small work of a research. This work is extended up to large set of database and identifies the result in the help of two or three machine learning algorithms and then takes better approach from that performance. In future we have compare tools using classification techniques which tool execute accurately.

VII. Acknowledgment

The authors thank to Selvamm Arts and Science College for giving an opportunity to publish the paper. The authors also thank University of Waikato for WEKA tool availability as an open source.

VIII. References

[1] Rovina Dbritto, Anuradha and Vincy Joseph, “Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods”, International Journal of Applied Information Systems(IJAIS)-ISSN:2249-0868, Vol-11 No 2, July 2016

[2] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques, Third Edition”

[3] V.Manikantan & S.Latha, “Predicting the Analysis of Heart Disease Symptoms using Medicinal Data Mining Methods”, International Journal on Advanced Computer Theory and Engineering (IJACTE)-ISSN (Print):2319-2526, Volume-2, Issue-2,2013.

[4] Divyansh Khanna, Rohan Sahu, Veeky Baths and Bharat Deshpande, “Comparative Study of Classification Techniques(SVM, Logistic Regression and Neural Networks) to predict the prevalence of Heart Disease”, International Journal of Machine Learning and Computing, Vol-5, No.5, October 2015.

[5] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu, “A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)”, International Journal of Computer Applications (0975 – 8887) Volume 68–No.16, April 2013

[6] Pritam H.Patil, Suvarna Thube, Bhakti Ratnaparkhi and K.Rajeswari, “Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining”, International Journal of Computer Applications(0975-8887), Volume 93-No.8, May 2014.

[7] Vikas Chaurasia & Saurabh pal, “Early Prediction of Heart Disease Using Data Mining Techniques”, Caribbean Journal of Science and Technology, Carib.j.SciTech,2013, Vol.1,208-217.

[8] http://repository.seasr.org/Datasets/UCI/arff.

[9] Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen, “ Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method”, IJSETR Volume 2, Issue 10, October 2013

[10] M.Akhil jabbar, B.L Deekshatulu and Priti Chandra, “ Classification of Heart Disease using K-Nearest Neighbor and Genetic Algorithm”, International Conference on (CIMTA) 2013.

[11] http://www.cs.waikato.ac.nz/ml/weka.

[12] Daljeet Kaur A and Aman Paul , “ Performance Analysis of Different Data mining Techniques over Heart Disease dataset”, International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 - 5161 ©2014 INPRESSCO

[13] K.Vembandasamy R.Sasipriya and E.Deepa, “Heart Diseases Detection Using Naive Bayes Algorithm”, IJISET Vol. 2 Issue 9, September 2015. ISSN 2348 – 7968

[14] Ruchika Rana, Jyoti Pruthi, “Heart Disease Prediction using Naive Bayes Classification in Data Mining”, IJSRD - International Journal for Scientific Research & Development| Vol. 2, Issue 05, 2014 | ISSN (online): 2321-0613

[15] Prerana T H M, Shivaprakash N C and Swetha N, “ Prediction of Heart Disease Using Machine Learning Algorithms- Naive Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS”, International Journal of Science and Engineering Vol-3, Number 2-2015

[16] Dr. N. Venkatesan and Mrs.G. Priya, “A Study of Random Forest Algorithm with Implementation using Weka”, International journal of Innovative Research in Computer Science and Engineering (IJIRCSE) ISSN: 2394-6364, Vol-1, Issue-6, May 2015.

[17] Ajay Kumar Mishra and Bikram Kesari Ratha, “Study of Random Forest Data Mining Algorithms for Microarray Data Analysis”, International Journal on Advanced Electrical and Computer Engineering”, (IJAECE)- ISSN : 2349-9338, vol-3, Issue-4, 2016.