American Journal of Engineering Research (AJER)
e-ISSN : 2320-0847 p-ISSN : 2320-0936
Volume-4, Issue-3, pp-55-61
www.ajer.org

Research Paper Open Access

Comparison of Different Classification Techniques Using WEKA for Hematological Data

Md. Nurul Amin[1], Md. Ahsan Habib[2]

(Information and Communication Technology Department, Mawlana Bhashani Science and Technology University, Bangladesh)

ABSTRAC : *Medical professionals need a reliable prediction methodology to diagnose hematological data comments. There are large quantities of information about patients and their medical conditions. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Weka is a data mining tools. It contains many machine leaning algorithms. It provides the facility to classify our data through various algorithms. Classification is an important data mining technique with broad applications. It classifies data of various kinds. Classification is used in every field of our life. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. In this paper we are studying the various Classification algorithms. The thesis main aims to show the comparison of different classification algorithms using Waikato Environment for Knowledge Analysis or in short, WEKA and find out which algorithm is most suitable for user working on hematological data. To use propose model, new Doctor or patients can predict hematological data Comment also developed a mobile App that can easily diagnosis hematological data comments. The best algorithm based on the hematological data is J48 classifier with an accuracy of 97.16% and the total time taken to build the model is at 0.03 seconds. Naïve Bayes classifier has the lowest average error at 29.71% compared to others.*

Keywords -*Hematological data, Data Mining, J48 Decision tree, Multilayer Perception, Naïve Bayes.*

I. INTRODUCTION

Data mining technique is a process of discovering pattern of data. The patterns discovered must be meaningful in that they lead to some advantage. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable data in order to aid user decision making [9]. Data mining is being used in several applications like banking, insurance, hospital and Health informatics. In case of health informatics, Data mining plays a vital role in helping physicians to identify effective treatments, and Patients to receive better and more affordable health services. In hematology laboratory, it has become a powerful tool in managing uncountable laboratory information in order to seek knowledge that is underlying or within any given information.

Comparison of Different Classification Techniques Using WEKA for Hematological Data Comment is a challenging and interesting task in medical research area. To find out which classification algorithms is batter it is very difficult to compare different classification algorithms in different dataset. Our dissertation concerns with to make a mobile App, which is capable to Diagnose Hematological data comments. With this purpose to perform a better approach, we divide this problem of Hematology Data comments into three phases: Data Collection, Classification algorithm, and developed mobile App. We proceed in the following ways to achieve our purpose successfully.

- We are going to collect hematological data from oracle 10g database.
- We are going to apply hematological data in WEKA then find three classification algorithms performance.

- Finally developed a mobile App.

We studied various journals and articles regarding performance evaluation of Data Mining algorithms on various different tools, some of them are described here.

- There are related works using data mining techniques to diagnose several types of diseases and phenomena, such as Automated Diagnosis of Thalassemia Based on Data Mining Classifiers, etc. And many other tried to find their own formula. This paper presents an investigation for thalassemia existence by using data mining classifiers depending on CBC. They do that but they say MCV is the main feature. They should need use Hemoglobin is the main feature to classified thalassemia, [11].
- K.Rajesh et al [14] in their paper "Application of Data Mining Methods and Techniques for Diabetes Diagnosis." they provide a comparative analysis of different algorithms. This project aims for mining the relationship in diabetes data for efficient classification. But they need proposed a model that can diagnose diabetes dataset.
- Satish Kumar David et al [15] in his research paper "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics." Studied the performance of Tree Random Forest, J48 decision tree, Bayes Naïve Bayes and Lazy.IBK. In this paper, they compared algorithms based on their accuracy, learning time and error rate. They observed that there is a direct relationship between execution time in building the tree model and the volume of data records, while there is also an indirect relationship between execution time in building the model and the attribute size of the data sets. Through experiment, they conclude that Bayesian algorithms have better classification accuracy over and above compared algorithms.
- Salvitha et al [3] in their article "Evaluating Performance of Data Mining Classification Algorithm in Weka”. They provide performance of different dataset use data mining classification. The main aim of this paper judge the performance of different data mining classification algorithms on various datasets.
- Nookala et al [6] in their article "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification." In this study, they have made a comprehensive comparative analysis of 14 different classification algorithms and their performance has been evaluated by using 3 different cancer data sets. The results indicate that none of the classifiers outperformed all others in terms of the accuracy when applied on all the 3 data sets. Most of the algorithms performed better as the size of the data set is increased. They recommend the users not to stick to a particular classification method and should evaluate different classification algorithms and select the better algorithm.
- Vaithiyanathan et al [1] in their paper "comparison of different classification techniques using different datasets". They used three dataset from benchmark data set (UCI) and they used four classifier algorithms J48, Multilayer Perceptron, Bayes Net, and Naïve Bayes Update. This work has been carried out to make a performance evaluation above algorithms.
- Tiwari et al [7] in their research paper "Performance analysis of Data mining algorithms in Weka”. The aim of their paper is to judge the accuracy of different data mining algorithms on various data sets.
- Bin Othman et al [10] "Comparison of different classification techniques using WEKA for breast cancer". In this paper they present the comparison of different classification techniques using Waikato Environment for Knowledge Analysis or in short, WEKA. The aim of their paper is to investigate the performance of different classification or clustering methods for a set of large data. The algorithm or methods tested are Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm. The best algorithm based on the breast cancer data is Bayes network classifier with an accuracy of 89.71% and the total time taken to build the model is at 0.19 seconds. Bayes network classifier has the lowest average error at 0.2140 compared to others.
- All the previous works tried to makes a model to diagnosis diesis, and most of them just try to use one data mining technique they consider it the best one without any comparison with the other techniques in the domain. In this study, I will used more than one classifier to get most significance one, and make a model that can easily diagnosis hematological data comments.

The main contributions of the thesis are summarized follow:

- J48 based on decision tree algorithm has been achieved to classify different types of hematological data comment.

- Naïve Bayes algorithm has been obtained for high probability of hematological data comment.

- Multilayer perceptron algorithm has been obtained mathematical or computational model for information processing based on a connectionist approach.
- A comparison with different classification techniques has made with optimal features to show which method is appropriate for hematological data.

II. MATERIAL AND METHODS

We have used the popular, open-source data mining tool Weka (version 3.7.11) for this analysis. Two different data sets have been used and the performance of a comprehensive set of classification algorithms (classifiers) has been analyzed. The analysis has been performed on a TOSHIBA Windows 7 Enterprise system with Intel® Core ™ i5 CPU, 2.30 GHz Processor and 3.00 GB RAM. The data sets have been chosen such that they differ in size, mainly in terms of the number of attributes.

The hematological parameter are composed of White blood cell count (WBC), Red blood cell count (RBC), Hemoglobin (Hb), Hematocrit (Hct), Mean corpuscular volume (MCV), Mean corpuscular hemoglobin (MCH), Mean corpuscular hemoglobin concentration (MCHC), Platelet count (PLT), Neutrophil count (NEU), Lymphocyte (LYMP), Monocyte (MONO), Eosinophil (EO), and Basophil (BASO) (SysMex 1000i Sysmex corporation, Kobe, Japan). Hematological data was manually evaluated by medical technologist who has a license certification from the State medical Faculty of Bangladesh. Collected data are assigned to Several labels: Suggestive of anaemia of chronic disorder, Eosinophilia, Microcytic hypochromic anaemia, Normocytic anaemia, Neutrophil leucocytosis, Neutrophilia, Non-specific findings, High ESR.

2.1 Dataset and Preprocessing

The experiment1 dataset consists of 600 samples and experiment2 dataset consists of 298 samples. Its attributes represents the CBC features as in Table 1, some features; such as the sex, age, and some others features which are dropped due the privacy of the blood sample's owner, and finally it contain diagnoses attribute which represent the target label of the sample, it has several labels: Suggestive of anaemia of chronic disorder, Eosinophilia, Microcytic hypochromic anaemia, Normocytic anaemia, Neutrophil leucocytosis, Neutrophilia, Non-specific findings, High ESR, and Other which represented any other hematological data comments.

Table 1: CBC Test Features

Shortcut	Term	Male Normal Value	Female Normal Value
WBC (cmm)	White Blood Cell	4000-11000	
RBC (million/cmm)	Red Blood cell	5.0±0.5	4.3±0.5
HB(g/dl)	Hemoglobin	15.0±2.0	13.5±1.5
HCT(l/l)	Hematocrit	0.45±0.05	0.41±0.05
MCV(ft)	Mean Cellular Volume	92±9	
MCH(pg)	Mean Cellular Hemoglobin	29.5±2.5	
MCHC(g/dl)	Mean Cellular Hemoglobin Concentration	33.0±1.5	
PLT(/Cmm)	Platelet Count	150000-400000	
NEU	Neutrophils(%)	40-75	
LYMP	Lymphocytes(%)	20-40	
MONO	Monocytes(%)	2-10	
EO	Eosinophils(%)	2-6	
BO	Basophils(%)	<1.0	

In the preprocessing of the dataset we eliminate useless attributes, refill the missing values and remove/refill the outlier values on the outlier samples. Table 2 represent the dataset attributes which we used in our investigation.

Table 2: Dataset Attributes

Attribute	Data type	Attribute role
SEX	Binomial	Regular
WBC	Integer	Regular
RBC	Integer	Regular
HB	Integer	Regular
HCT	Integer	Regular
MCV	Integer	Regular
MCH	Integer	Regular
MCHC	Integer	Regular
PLT	Integer	Regular
NEU	Integer	Regular
LYMP	Integer	Regular
MONO	Integer	Regular
EO	Integer	Regular
BO	Integer	Regular
Hematological Comments	Nominal	Label

2.2 Classification Methods

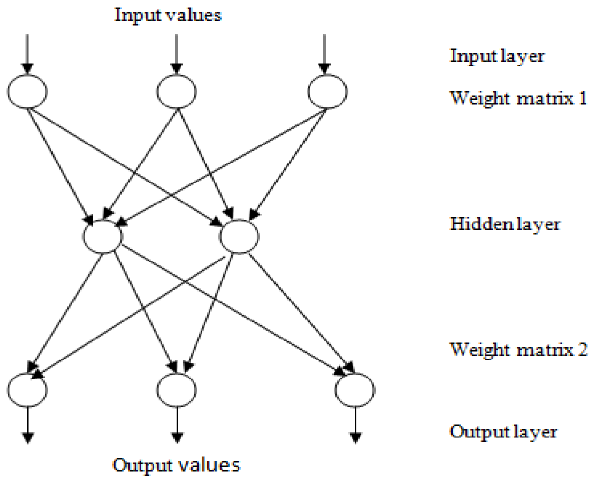
Three candidate classifiers are considered in this study: Decision Tree (J48), Naïve Bayes, and Neural Network (Multilayer Perceptron)

2.2.1 J48 Algorithm

J48 algorithm is called as optimized implementation of the C4.5 or improved version of the C4.5. The output given by J48 is the Decision tree. A Decision tree is same as that of the tree structure having different nodes, such as root node, intermediate nodes and leaf node. Each node in the tree contains a decision and that decision leads to our result as name is decision tree. Decision tree divide the input space of a data set into mutually exclusive areas, where each area having a label, a value or an action to describe or elaborate its data points. Splitting criterion is used in decision tree to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node [1].

2.2.2 Multilayer Perceptron

The single-layer perceptron can only classify linearly separable problems. For non-separable problems it is necessary to use more layers. A Multilayer (feedforward) network has one or more hidden layers whose neurons are called hidden neurons. The Fig.1 illustrates a multilayer network with one input layer, one hidden layer and one output layer.



Figurer 1: multilayer perceptron

2.2.3 Naive Bayes

Naive Bayes implements the probabilistic Naïve Bayes classifier. Naïve Bayes Simple uses the normal distribution to model numeric attributes. Naïve Bayes can use kernel density estimators, which develop performance if the normality assumption if grossly correct; it can also handle numeric attributes using supervised discretization. Naïve Bayes Updateable is an incremental version that processes one request at a time. It can use a kernel estimator but not discretization [13].

III. RESULTS AND DISCUSSION

In this investigation, the experiment using the data mining classifiers will be divided into two parts: the experiment with full and reduced features. The results from these two parts and a detailed classification accuracy analysis emphasizing on the classification errors will be presented in following Sections. Three experiments were conducted in each type: the first one is to measure the performance of the decision tree classifier; the second one is to measure the performance of the naïve bayes classifier, the third one to measure the performance of the neural network. The feed-forward back-propagation neural network classifier was adjusted with 500 training cycles, learning rate 0.3, and momentum 0.2.

3.1 Experiments with full features

In these experiments we used the whole records attributes of each sample. The decision tree classifier gives a result with general accuracy of 97.16%, the naïve bayes classifier gives a result with general accuracy of 70.28%, and finally the neural network classifier gives a result with general accuracy of 86.55% as shown in Fig.2,Table3.

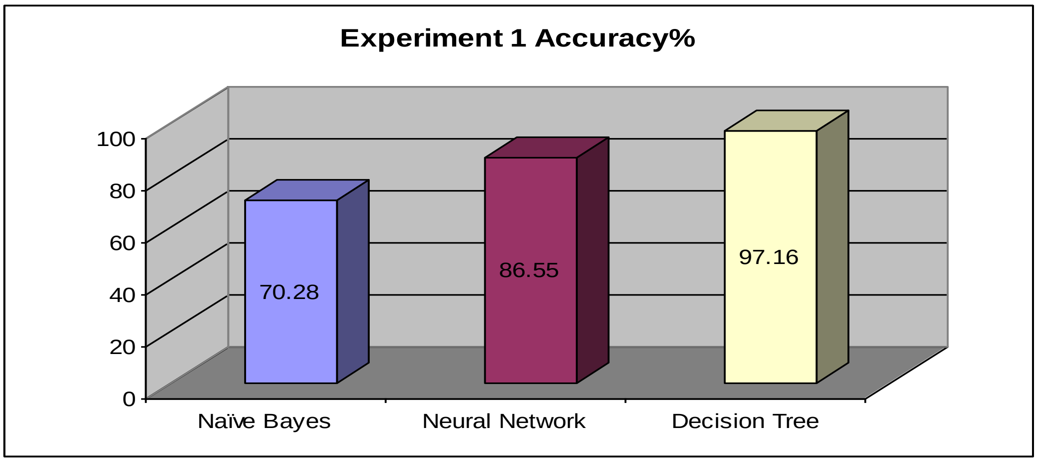


Figure 2: experiment 1 classifiers accuracy values

Table 3: Simulation Result of Each Algorithm for Exprement1

Algorithm (Total instances,425)	Correctly Classified Instances %(Value)	Incorrectly Classified Instances %(Value)	Time Taken (seconds)	Kappa statistic
J48 Decision tree	97.16 %(412)	2.83(12)	0.03	0.9648
Multilayer Perception	86.5566 %(367)	13.4434 %(57)	2.29	0.8346
Naïve Bayes	70.28 %(298)	29.71 %(126)	0.03	0.6329

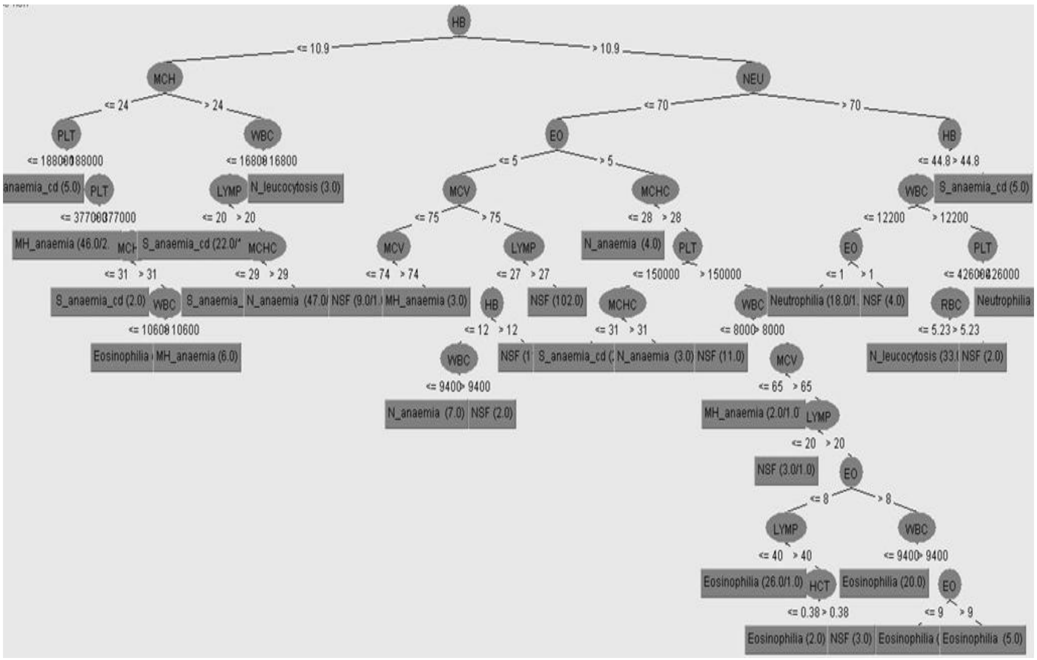


Figure 3: decision tree form experiment 1

3.2 Experiment with reduced features

In our experiments we used the whole record's attributes of each sample as in Table 4.The Decision Tree classifier gives a result with general accuracy: 94.27%, while the Naïve Bayes classifier gives a result with general accuracy: 70.03% and the Neural Network classifier give a result with general accuracy: 78.45% as shown in Fig.4, Table 4.

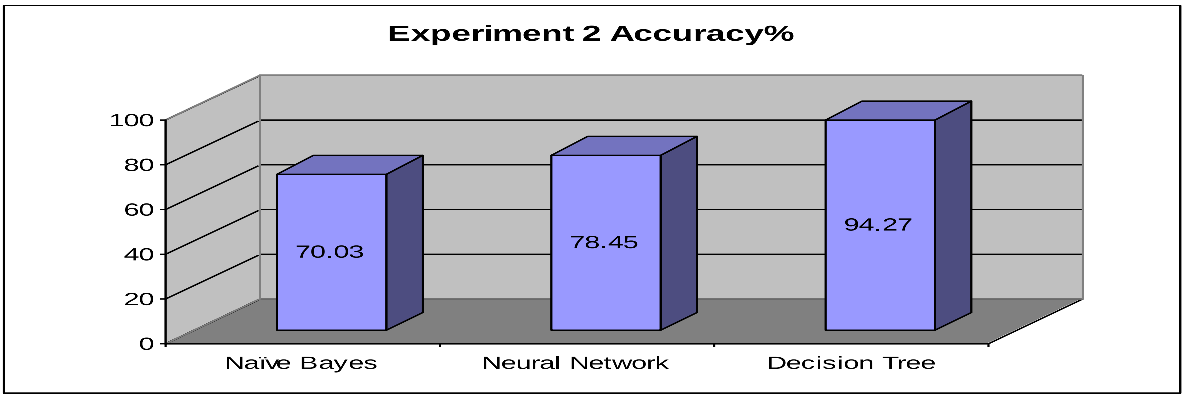


Figure 4: experiment 2 classifiers accuracy values

Table 4: Simulation Result of Each Algorithm for Exprement2

Algorithm (Total instances,298)	Correctly Classified Instances %(Value)	Incorrectly Classified Instances %(Value)	Time Taken (seconds)	Kappa statistic
J48 Decision tree	94.2761 %(280)	5.7239 %(17)	0.03	0.9258
Multilayer Perception	78.4512 %(233)	21.5488 %(64)	1.76	0.7137
Naïve Bayes	70.0337 %(208)	29.9663 %(89)	0.01	0.5981

Table 5: Comparison on Various Datasets Depend Accuracy and Classifiers.

Name of the classifier	Experiment 1	Experiment 2
J48 Decision tree	97.16	94.2761
Multilayer Perception	86.5566	78.4512
Naïve Bayes	70.28	70.0337

Based on the above Fig.2, 4 and Table 5, we can clearly see that the highest accuracy is 97.16% and the lowest accuracy is 70.03%. We can say that J48 Decision tree is batter.

IV. CONCLUSION

As a conclusion, we have met our objective which is to evaluate and investigate three selected classification algorithms based on Weka. The best algorithm based on the hematological data is J48 classifier with an accuracy of 97.16% and the total time taken to build the model is at 0.03 seconds. Naïve Bayes classifier has the lowest average error at 29.71% compared to others. These results suggest that among the machine learning algorithm tested, Naïve Bayes classifier has the potential to significantly improve the conventional classification methods for use in medical or in general, bioinformatics field.

We would like to develop web based software for performance evaluation of various classifiers where the users can just submit their data set and evaluate the results.

V. ACKNOWLEDGEMENTS

The author is extremely grateful to the Ibn Sina diagnostic and imaging center, IBN SINA Trust, for their collaborative and providing us with dataset.

REFERENCES

[1] Vaithiyanathan, V., K. Rajeswari, Kapil Tajane, and Rahul Pitale. *"Comparison of Different Classification Techniques Using Different Datasets."Vol.6, no. 2 (2013).*

[2] Sharma, Narendra, Aman Bajpai, and Mr Ratnesh Litoriya. *"Comparison the various clustering algorithms of weka tools."Volume2, no.5 (2012).*

[3] Salvithal, Nikhil N., and R. B. Kulkarni. *"Evaluating Performance of Data Mining Classification Algorithm in Weka." Vol 2., no. 10 (2013).*

[4] Khan, S. A., J. H. Epstein, K. J. Olival, M. M. Hassan, M. B. Hossain, K. B. M. A. Rahman, M. F. Elahi et al. *"Hematology and serum chemistry reference values of stray dogs in Bangladesh." ." Vol. 1: 13-20 (2011).*

[5] Zhang, Wenjing, Donglai Ma, and Wei Yao. *"Medical Diagnosis Data Mining Based on Improved Apriori Algorithm." Journal of Networks 9, no. 5 (2014): 1339-1345.*

[6] Nookala, Gopala Krishna Murthy, Bharath Kumar Pottumuthu, Nagaraju Orsu, and Suresh B. Mudunuri. *"Performance analysis and evaluation of different data mining algorithms used for cancer classification." International Journal of Advanced Research in Artificial Intelligence (IJARAI) 2, no. 5 (2013).*

[7] Tiwari, Mahendra, Manu Bhai Jha, and OmPrakash Yadav. *"Performance analysis of Data Mining algorithms in Weka." IOSR Journal of Computer Engineering (IOSRJCE) ISSN (2012): 2278-0661, Vol.6, Iss.3.*

[8] Kaushik H. Raviya, Biren Gajjar *"Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA"Vol. 2, Issue. 1. (2013).*

[9] Saichanma, Sarawut, Sucha Chulsomlee, Nonthaya Thangrua, Pornsuri Pongsuchart, and Duangmanee Sanmun. *"The Observation Report of Red Blood Cell Morphology in Thailand Teenager by Using Data Mining Technique." Advances in hematology 2014 (2014).*

[10] bin Othman, Mohd Fauzi, and Thomas Moh Shan Yau. "*Comparison of different classification techniques using WEKA for breast cancer." 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006. Springer Berlin Heidelberg, 2007.*

[11] Elshami, E. H., & Alhalees, A. M. (2012). *Automated Diagnosis of Thalassemia Based on DataMining Classifiers. In The International Conference on Informatics and Applications (ICIA2012) (pp. 440-445). The Society of Digital Information and Wireless Communication*.

[12] Pankaj saxena & sushma lehri, International Journal of Computer & Communication Technology ISSN (PRINT): *"Analysis of various clustering algorithms of data mining on health informatics"Vol. 4, Issue. 2. (2013),*

[13] Ms S. Vijayarani , Ms M. Muthulakshmi, International Journal of Advanced Research in Computer and Communication Engineering: *"Comparative Analysis of Bayes and Lazy Classification Algorithms". Vol.2, Issue. 8, (2013).*

[14] Rajesh, K., and V. Sangeetha. *"Application of data mining methods and techniques for diabetes diagnosis." International Journal of Engineering and Innovative Technology (IJEIT) Volume. 2, Issue. 3 (2012).*

[15] David, Satish Kumar, Amr TM Saeb, and Khalid Al Rubeaan. *"Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics." Computer Engineering and Intelligent Systems 4, no. 13 (2013): 28-38.*