Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal

Vol. 3 No. 3, September 2017

Development of Data Mining Classification Tool Benchmarked With Weka Classifier (A Case Study of a High Blood Pressure Disease Dataset)

Alawode A. John
Department of Computer Science
The Federal Polytechnic
Ilaro, Nigeria
ademola.alawode@federalpolyilaro.edu.ng

ABSTRACT

The availability of enormous medical data that are not utilize for knowledge discovery is a great concern for researchers. Data mining as a field that extracts interesting patterns for knowledge discovery which enhances better decision making. The major data mining tool used for classification is WEKA, which has demonstrated high level of efficiency and effectiveness. It is therefore pertinent to have another classification model that will contest accuracy and sensitivity of WEKA classifier, and High Blood Pressure Disease dataset are used as case study. This research extracted knowledge from abundance of High Blood Pressure disease patients' record. The acquired dataset was classified on both WEKA classifier and on the developed Model from Artificial Neural Network Algorithm, which was trained with learning vector quantization algorithm. The comparison was done based on their accuracy and It was discovered that WEKAs' accuracy was 0.64% less than that of the developed Model (DMCT), but time taken to build WEKA model was extremely faster, for the same number of instances WEKA spent 0.66 seconds while DMCT spent 4.66 seconds.

Keywords: Development, Data Mining, Classification, Weka Classifier, High Blood Pressure & Disease Dataset

Aims Research Journal Reference Format:
Alawode, A.J. (2017): Development of Data Mining Classification Tool Benchmarked With Weka Classifier (A Case Study of a High Blood Pressure Disease Dataset). Advances in Multidisciplinary & Scientific Research Journal. Vol. 3. No.3, Pp 73-84

1. INTRODUCTION

Data mining is the series of events/actions of discovering interesting knowledge from large volume of data kept in databases, data warehouses, or other information repositories. This field of information technology has attracted a great deal of mental focus in the information industry and in society as a whole in recent time, due to the great availability of enormous amounts of data and the imminent need for turning such data into useful information and knowledge. Databases are rich with hidden facts that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends, which are being referred to as knowledge. The knowledge acquired thus far, has really helped in various areas such as Market Analysis, Fraud Detection, Medical Diagnosis, Production Control, Science Exploration e.t.c. Basic techniques of data classification, such as building decision tree classifiers, Bayesian classifiers, Bayesian belief networks, rule-based classifiers and Back-propagation (a neural network technique) are approach to classification known as support vector machines.

The knowledge acquired from different data mining techniques, such as Artificial Neural Network, Bayesian classifiers and C4.5 decision tree Algorithms, and even WEKA, are wonderful tools in medical diagnosis of diseases by clinicians, Artificial Neural Network is widely used in data mining due to characteristics like parallel performance, self-organising adaptive, robustness, Non-parametric, Non-linearity, Input-Output mapping and fault tolerance.
Artificial Neural Network and WEKA classifier were comparatively analysed to determine their strength and weakness in classification of high blood pressure disease dataset.

High Blood Pressure Disease is serious health condition that can lead to coronary heart disease, heart failure, stroke, kidney failure, and other health problems. Availability of data set for prediction of High Blood Pressure is enormous, which can substantiate and confirm the degree of the disease. In the absence of medical diagnosis evidences, it is difficult for the experts to opine about the grade of disease with affirmation, (Rashedur et al., 2013). In this research, a model was developed, and its' performance accuracy was compared with WEKA classifier and data set of High Blood Pressure was used as a case study.

1.1 Statements of the Problem

Effectiveness and efficiency are great and important features of models used in data mining. WEKA demonstrated these features perfectly, but it is pertinent to develop another model that will contest these features in terms of performance accuracy in order to extract meaningful and interesting patterns from databases especially in medicine which involved accurate descriptive and prescriptive analysis. **Adeyemo et.al,2015**. Compared the performance accuracy of Multilayer perceptron (MLP) Artificial Neural Network and ID3 (Iterative Dichotomiser 3), C4.5 (also known as J48) Decision Trees algorithms in WEKA data mining software in predicting Typhoid fever, but WEKA's accuracy was not contested. **Arpneek Kaur and Abhishek Bhardwaj, 2014** researched on using artificial intelligence methods to diagnose hypertension, but did not take the performance accuracy into consideration. Development of a model using Artificial Neural Network Algorithms which will predict better because it will use historical dataset for knowledge discovery. Similarly, its' performance accuracy will be compared with that of WEKA classifier in terms of effectiveness and efficiency to determine which classifies better. Comparative analysis of the Model and that of WEKA classifier is of great importance to this research work, and this performance accuracy will be analysed with High Blood Pressure dataset obtained from public hospital.

1.2 Aim and Objectives

The aim of this research work is to develop a model, using Artificial Neural Network Algorithm and evaluate its' performance accuracy with WEKA data mining software.

1.3 Objectives

- To discover appropriate knowledge and extract useful patterns from existing stored data of patients so that these knowledge and patterns can be used for decision making.
- To establish new degree of accuracy that can be achieved by Artificial Neural Network Algorithm.
- To save more life.

2. RELATED WORKS

A good number of researches have been reported in literature on diagnosis of different diseases. Today, world-wide increasing death rate of heart disease patients each year and the availability of enormous amount of patients' data from which to extract useful information, researchers are using data mining techniques to help healthcare specialists in the diagnosis and treatment of heart disease.

Razali & Ali (2009), examined the making of treatment plans for critical upper respiratory infection disease patients using a decision tree. The proposed treatment model gave 94.73% accuracy through giving drugs to patients. The association rules and decision tree to treatment plans are showing satisfactory performance. They also found that the comparison of decision tree technique with other data mining techniques such as naïve bayes, genetic algorithms and neural network still needs further investigation.

The use of neural network is very wide in data mining due to some characteristic like parallel performance, Self-organizing adaptive, robustness and fault tolerance. Data mining models depend on task they accomplish: Association Rules, Clustering, Prediction, and Classification. Neural network is used to find pattern in data. The grouping of neural network model and data mining method can greatly increase the efficiency of data mining methods and it has been broadly used. Different algorithms have been discussed for optimizing the artificial neural network (ANN). ANN combines with other algorithms to find out the high accurate data as compare to traditional algorithm. The role of ANN using data mining techniques is playing an important role in forecasting or prediction about games and weather. This produces high accurate predictions than that of traditional algorithm (**Muhammad et. al. 2015**).

Adeyemo et. al. (2014) explored the accuracy of prediction and classification capabilities of decision tree to achieve data mining in comparison with several other algorithms in different application domains, they took the advantages of large data set in medical records of Heart disease patients that have been gathered over years and due processing was performed on them using decision tree.

Adeyemo & Olurotimi (2012) analyzed the massive data generated from the breast cancer patients from the reputable teaching hospital for several years and these data were transformed in a way that can be accepted to neural network data mining software. The Artificial Neural Network (back propagation learning algorithms on multilayer perceptrons) was used to detect the existence of cancer in a patient. The System was trained using backpropagation learning algorithm (pattern-by-pattern and delta-delta) on dataset acquired from a teaching hospital. The Multi-Layer perceptron and the two learning algorithms were implemented using Java programming language. The implemented algorithms were tested on a real world problem. The result obtained after the application of the learning algorithms were reported and compared.

Sumathi & Santhakumaran (2011) researched, on using ANN to solving the problem of diagnosing hypertension using back – propagation learning algorithm the network was constructed using various factors which are classified into categories, to be trained, tested and validated using the respective dataset. The results generated by using this system have been verified with the physicians and are found to be correct. But, the result of their work were not compared in terms of efficiency with WEKA data mining software.

Arpneek Kaur and Abhishek Bhardway, 2014. Researched extensively on hypertension diagnosis using artificial intelligence methods which is now very popular in medical application due to high reliability. Their work presents an introduction and survey on different artificial intelligence methods used by researchers for the application of diagnosing or predicting Hypertension. It was concluded that artificial neural networks and fuzzy system have been successfully employed by researchers for diagnosing of hypertension risk. However, the performance accuracy of their models were not compared accordingly.

3. RESEARCH METHODOLOGY

This section describes research methodology adopted for this research. The Knowledge Discovery in Database (KDD) methodology consisting of some iterative and interactive steps was adopted to extract significant patterns from a dataset. Using a data mining technique, a Classification Model was developed for classification of High Blood Pressure Disease (HBPD). Artificial Neural Network Algorithm (ANN) was used and it was implemented in WEKA. The performance of the Classification Model and that of Artificial Neural Network algorithm were also compared.

Developed Model
The model has different components that work together to achieve the goal of the system as see in **Fig 1.0**

- The first component describe data to be accepted by the model randomly.
- The entire data set is divided into parts for different purpose. The first part of the data set will be used to train the model, while the other data set is for testing of the model
- Training of the model comprises of four different stages, initialization of the weight for the nodes, computation of the new weight, updating of the learning factors and to make final decision on final weight.
- Classification was made by the developed model, while the other part of the dataset is used to test the model.
- The performance accuracy of the model is recorded to do comparative analysis.

Performance analysis of the Model will be compared with Weka classifier, in terms of accuracy to determine which of the two has better performance for the classification of the disease.

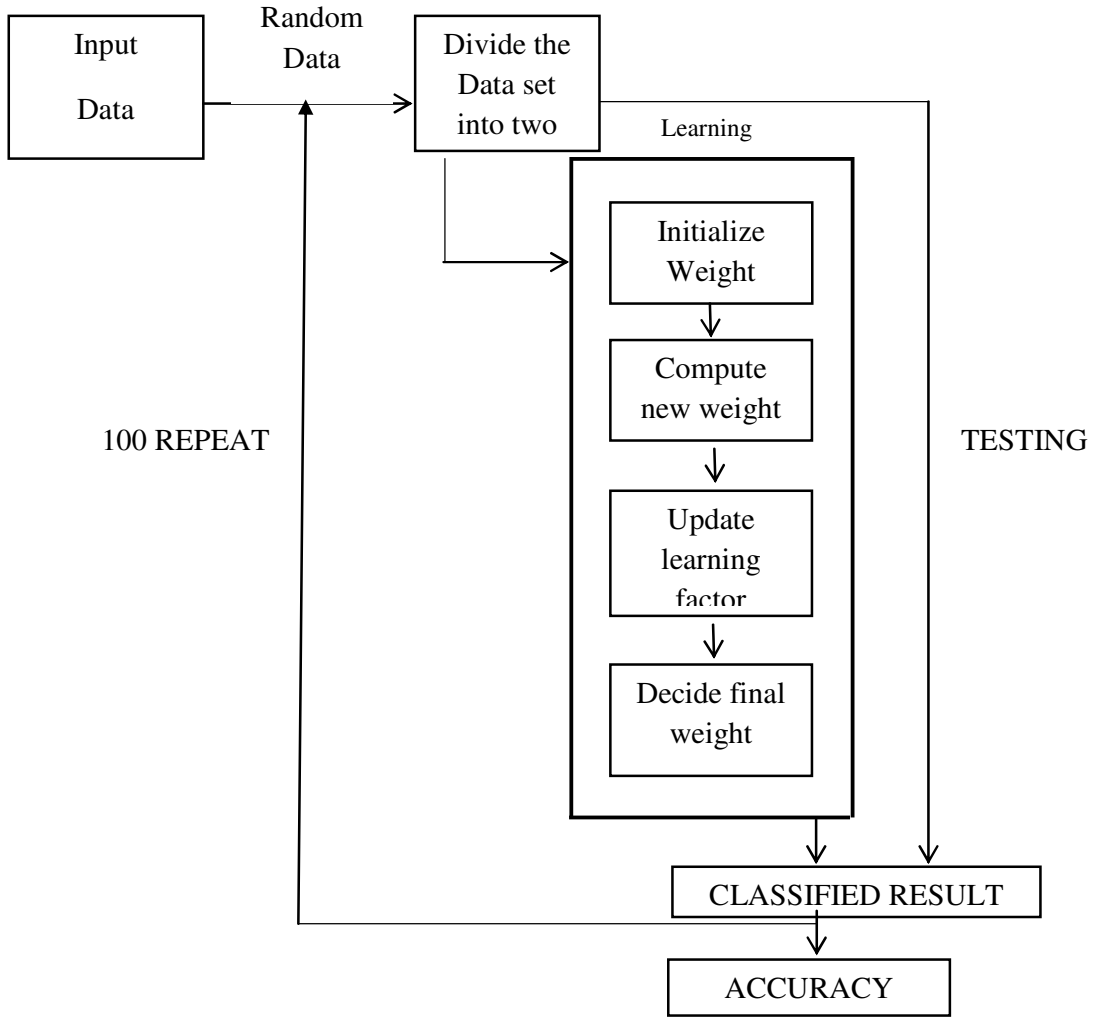

Fig.1: Data Mining Classification Model (DMCM)

The operations on developed model is divided into two, training of the model and testing of the model for prediction of High Blood Pressure Disease. The dataset were divided into two for the operations. The main parameters to note are sensitivity and accuracy of model during prediction.

4. STEPWISE ALGORITHM FOR THE MODEL

The algorithm of the developed model (DMCM) is given below:

STEP 1: Input the data set randomly. (This means that data are supplied into the model without any definite pattern of doing so).
STEP 2: Divide the data set into two. (The data set were separated/divided into to two for different purposes)
STEP 3: Use one part of the data set to train the model. (To teach the model and enable the model to learn the required pattern, so as to be able to predict effectively)
Training of the model
- initialize weight
- Compute new weight
- Update learning factor
- Decide final weight

STEP 4: Make prediction of the disease
STEP 5: Use the second part of the data set for testing the model.
STEP 6: Get the performance accuracy of the model.

5. DATA COLLECTION AND DESCRIPTION

The data used in this research is the dataset of patients with high blood pressure collected from Hospital. The data consists of sex of the patients, their ages, systolic and diastolic of the patients. The data and the attributes that possibly influenced their disease were selected and analysed.

Table 1.: Patient's attributes and their description

S/N	Attribute	Description	Data type
1	Age	Age of the patient in years	Numeric
2	Sex	Sex of the patient (Male / Female	Nominal
3	Weight	Weight of the patient in kg	Numeric
4	Systolic	The upper reading from Sphygmomanometer(High/Low)	Numeric
5	Diastolic	The lower reading from Sphygmomanometer(High/Low)	Numeric
6	Diagnosis	Does the patient have HBP (Yes/No)	Nominal

Each record in the dataset corresponds to a single patient's results collected during the medical examination.

5.1 Training of the Developed Model

The developed model has five input nodes. The algorithm used to train the model is Learning Vector Quantization (LVQ). Learning vector Quantization (LVQ) is a neural net that combines competitive learning with supervision. It can be used for pattern classification, which is a prototype-based supervised classification algorithm. LVQ is known as a special case of an artificial neural network. One of the advantages of LVQ is that it creates prototypes that are easy to read and interpret for experts in the respective application domain. LVQ applies a winner-take-all learning approach that is a precursor to k-Nearest neighbour algorithm (KNN) and Self-organizing maps (SOM). An LVQ system is represented by prototypes **W= (w(i),...,w(n))** which are defined in the feature space of observed data. In winner-take-all training algorithms one determines, for each data point, the prototype which is closest to the input according to a given distance measure. The position of this so-called winner prototype is then adapted, i.e. the winner is moved closer if it correctly classifies the data point or moved away if it classifies the data point incorrectly. A key issue in LVQ is the choice of an appropriate measure of distance or similarity for training and classification. The LVQ is made up of a competitive layer, which includes a competitive subnet, and a linear layer. In the first layer (not counting the input layer), each neuron is assigned to a class. Different neurons in the first layer can be assigned to the same class.

5.2 Software Process Methodology

The software Process adopted for the development of the Data Mining Classification Tool (DMCT) software is PROTOTYPING Model methodology. The Prototype Model are usually not complete systems and many of the details are not built in the prototype. The goal is provide a system with over functionality. In addition, the cost of testing and writing detailed documents are reduced. These factors help to reduce the cost of developing the prototype. This experience helps to reduce the cost of development of the final system and results in a more reliable and better designed systems. However, prototyping is more beneficial in an interactive systems.

5.3 The Conceptualised Diagram Of All The Research

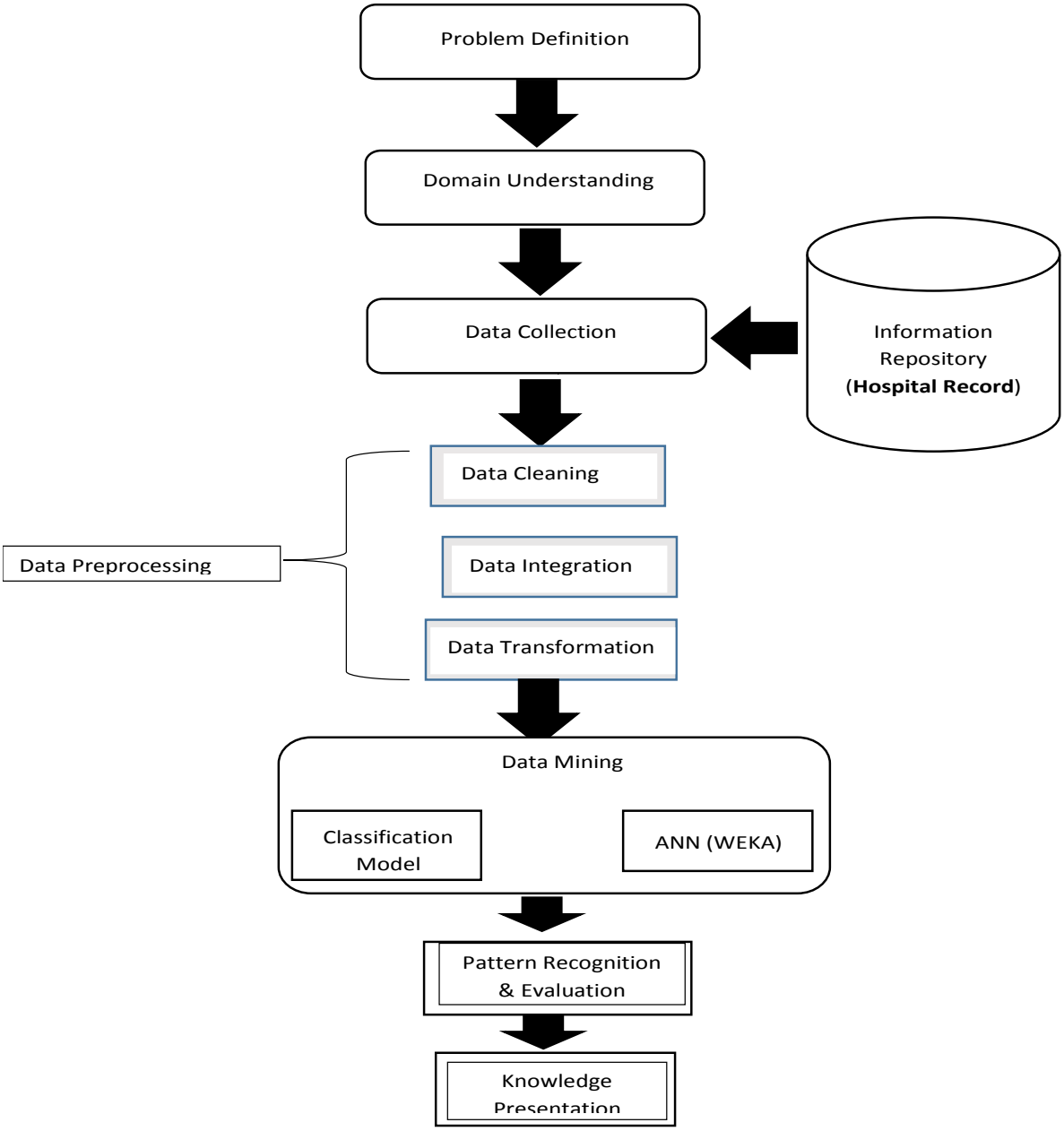

Fig 2: The Conceptualised Diagram of all the Research

5.3.1 Problem Definition
This research aims at developing a model for the detection of High Blood Pressure Disease (HBPD) and also using the data mining Artificial Neural Network algorithm to detect High Blood Pressure Disease (HBPD) and compare the results.

5.3.2 Domain Understanding
Domain understanding involves the process of acquiring wide background knowledge in the research area. The main domain of the research is Data Mining which is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. This is a very crucial stage that determines the overall success of the research. Specifically, Artificial Neural Network and WEKA data mining software were comparatively studied in detection of High Blood Pressure Disease.

5.3.2 Data Preprocessing
Data in the training dataset were pre-processed before evaluation by the algorithms. In this study, pre-processing was done based on Artificial Neural Network algorithm implemented. Some missing data were removed from the dataset to improve the classification performance and also some of the dataset were filtered and all attributes were made nominal, so that it could be accepted by WEKA.

5.3.3 Data Cleaning
Data collected tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. At this phase, to improve the classification performance, irrelevant data are removed from the data set.

5.3.4 Data Integration
This involves bringing together all the needed data from different data store into a single storage so that it can be used for the data mining process. Data from different sources are combined into a coherent data store. The data collected from the hospital's patient medical record was collated using Microsoft office excel 2007.

5.3.5 Data Transformation
This phase involves converting the data set into a format that is acceptable by WEKA. After the data was collated using Microsoft office excel 2007 spreadsheet, it was saved in a CSV (Comma Separated value) format. This is because WEKA only accepts Microsoft excel data in CSV format. After the data has been imported, it was then subsequently saved as an ARFF (Attribute-Relation File Format) format in WEKA.

5.3.6 Data Mining
The mining was carried out after the data was successful imported into WEKA, the data set was processed and Artificial Neural Network Algorithm was used to discover the hidden patterns in the data set. Classification rules for prediction were generated.

5.3.7 Pattern Recognition And Evaluation
This is the phase where interesting and previously undetected patterns are searched for, and when such patterns are identified, they are extracted for knowledge representation.

5.3.8 Data Training
The data set is divided into two, the training set and the test set. The data mining model for predicting High Blood Pressure Disease is built on the training set, but its quality is estimated on the test set. The training set is used to train the algorithm. It is ensured that the algorithm understands the pattern in the training set and the algorithm does not memorize the pattern. This will allow the algorithm to successfully predict accurate patterns for unfamiliar data. That will enhance proper predictions

6. RESULTS PRESENTATION AND DISCUSSION

This section is in two phases, the first phase implements WEKA to predict High Blood Pressure Disease and its' performance accuracy were recorded. Secondly, the developed model from Artificial Neural Network was implemented using C# programming language and its performance accuracy were compared with that of WEKA classifier.

This Fig 3. shows classification of WEKA with **99.34%** of accuracy.

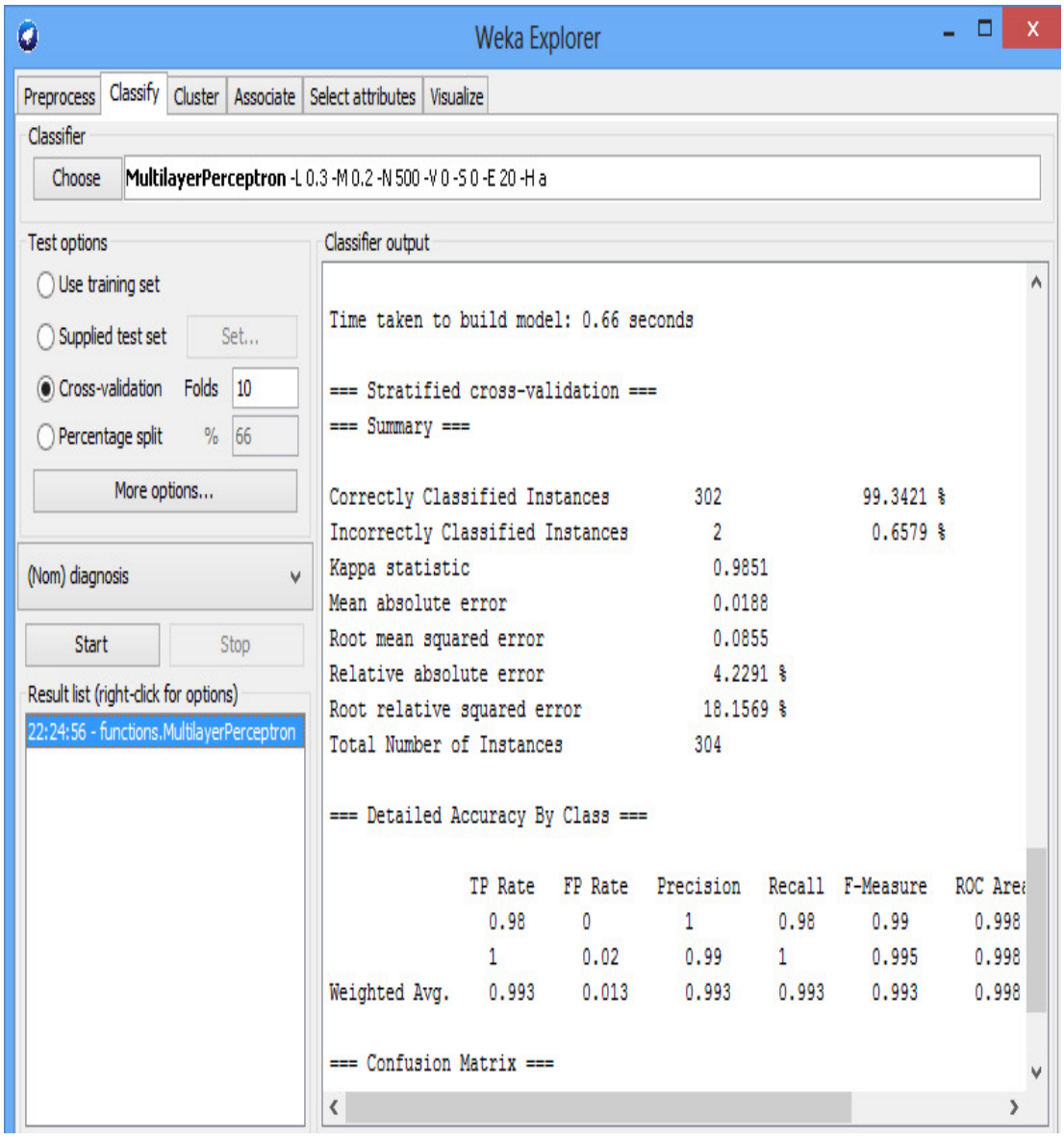

Fig 3.0 Weka Interface for classification

6.1 Implementation of (DMCT) MODEL

The programming language used in the implementation is C#

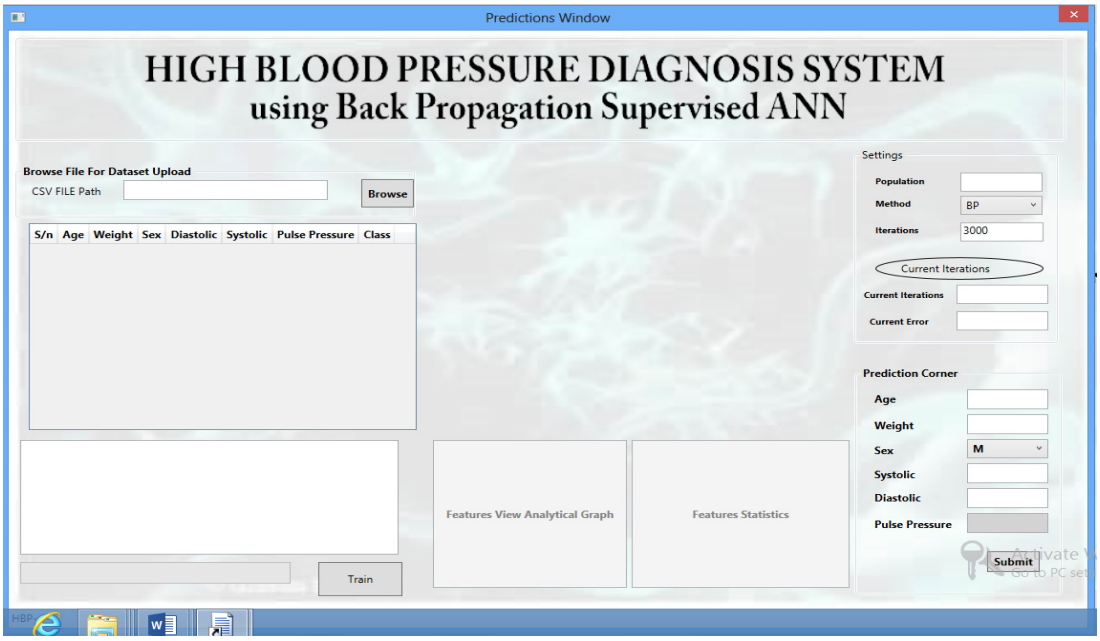

Fig 4. Developed Data Mining Classification Tools' Interface.

The model has the following features:

Up-loading field: this is used to fetch dataset to be analyzed. The dataset must be in the format of CSV.

Data area: This is where the up-loaded dataset can be visualized before initialization training process.

Training Button: This where the training of the model will be initialized.
Graphical representation of the analysis and statistical results of the prediction done can be visualized with two big Button beside Training Button.

Setting group and **prediction corner** are at the right hand side of the model interface.

Setting group has five fields: Population, Method, Iterations, Current Iteration and Current Error.
Prediction Corner has six fields which represent attributes considered in this research work. That is Age, Weight, Sex, Systolic, Diastolic and pulse pressure.

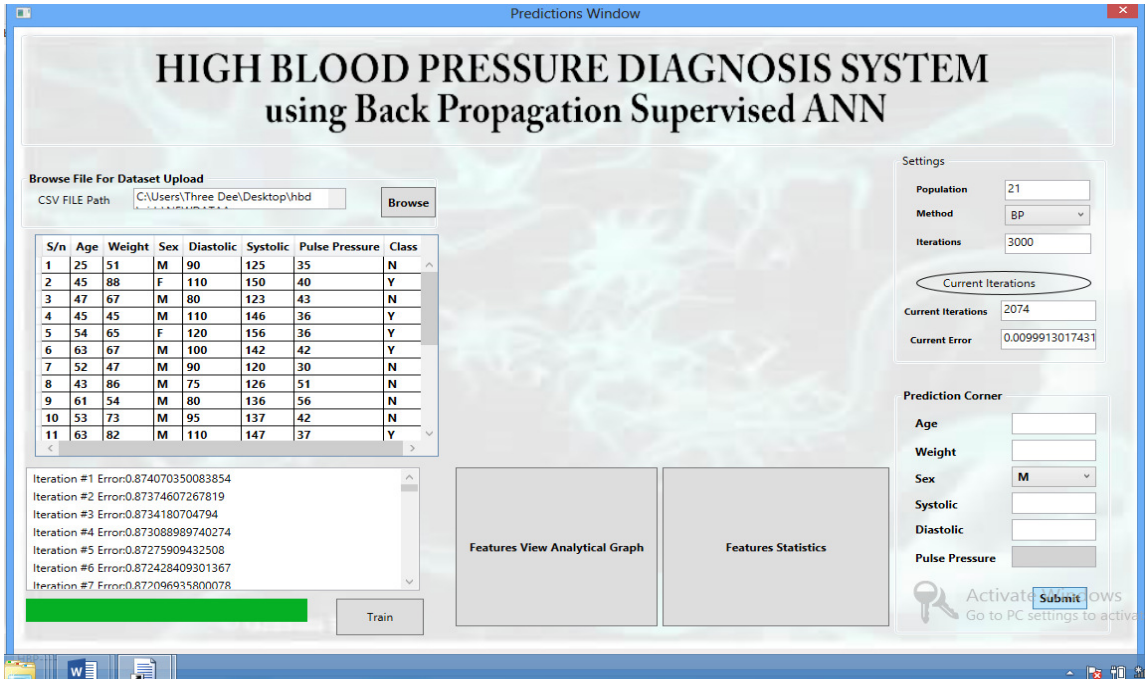

Fig 5.0 Training of DMCT for classification

Training of the Developed Model

Training dataset were up-loaded from Excel, the dataset was initially stored in csv format and it appears in Excel format inside the **data area** of the model. If the model is not trained after the dataset have being up-loaded. Any attempt to make new prediction will generate an error message displayed in **fig 5.0**

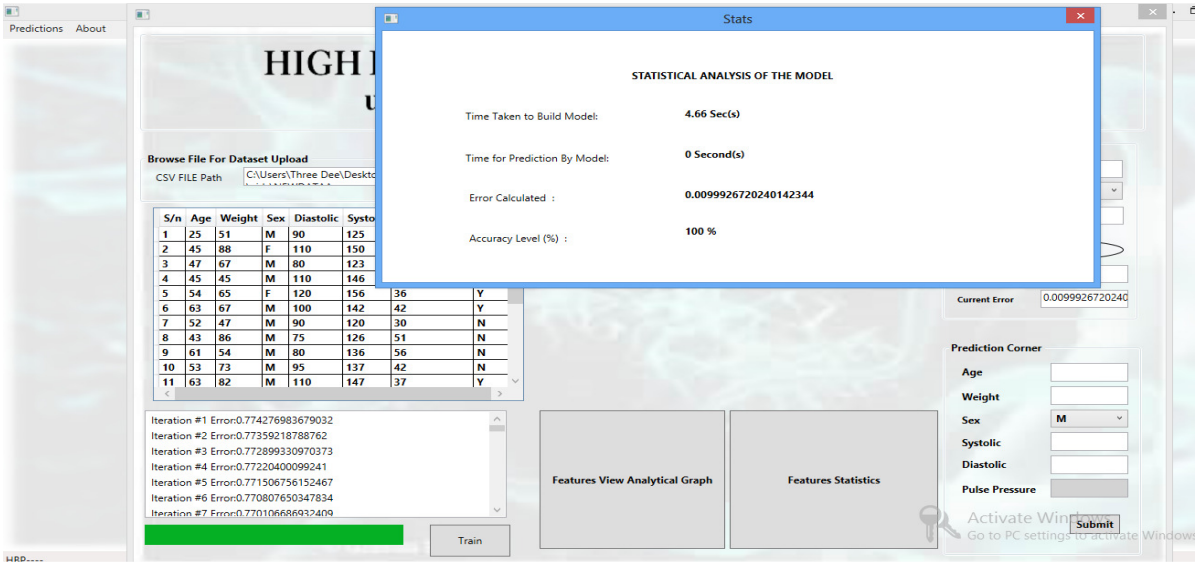

Fig 6. Performance Statistics of (DMCT)

Table 2: Comparison table between WEKA classifier and the Model

	Accuracy %	Precision%	Recall %	Time taken to build model (S)
Dev. Model	99.98	0.955	1	4.66
WEKA Classifier	99.34	0.990	1	0.66
% Difference	0.64	0.005	0.00	4.00

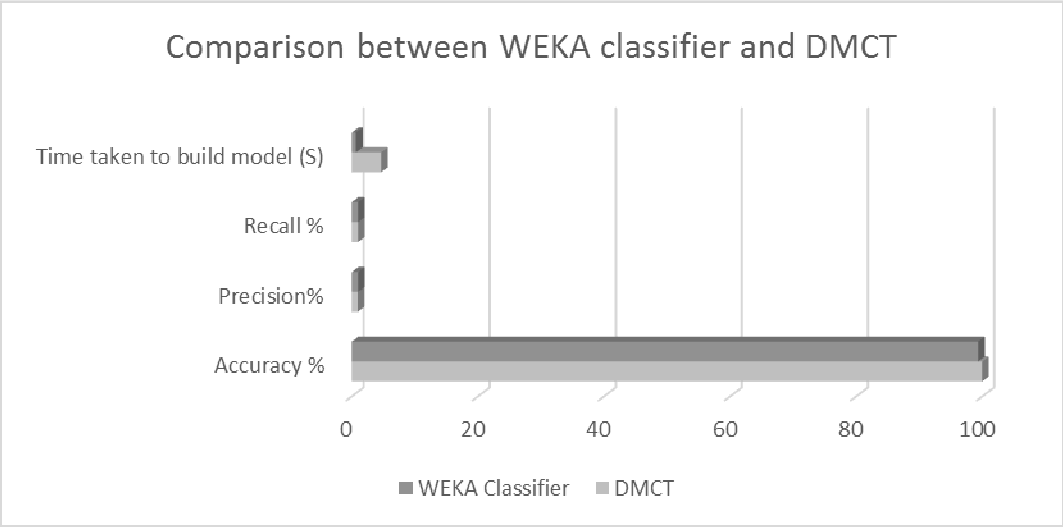

Fig 7. Comparison between WEKA Classifier and DMCT

7. DISCUSSION OF THE RESULTS

Table 2.0 clearly shows the outcome of the comparative analysis of Data Mining Classification Tool (DMCT) and WEKA classifier in classification of High Blood Pressure Disease dataset. The Accuracy of WEKA is **99.34%,** which is **0.64%** less than that of the Developed Artificial Neural Network Algorithm model. Developed model is slightly more accurate than WEKA Classifier. The implication of this is that, the developed model is more effective in predicting High Blood Pressure Disease in term of accuracy. The second significant difference is the Time taken to build the models. WEKA is extremely faster than the developed model. From the analysis, under the same circumstances. WEKA was confirmed to be efficient in term of timing. It was established that WEKA was faster by **4.00** seconds.

8. CONCLUSION AND RECOMMENDATION.

This research covers the use of WEKA classifier to classify the dataset of High Blood Pressure Disease patients. It also contains the implementation of the developed Data Mining Classification tool, using Artificial Neural Network Algorithm. It was discovered that, data mining Artificial Neural Network algorithm is an effective tool in the classification of High Blood Pressure Disease in term of accuracy. WEKA is faster in classification with slight error that reduces its' accuracy by **0.64%.** Therefore, speed of WEKA can be traded off for accuracy of the developed Data Mining Classification Tools' accuracy. Clinician will prefer an accurate system that is slower to faster system that is not very accurate, because life is involved. In the future, further research can be conducted to detect other data mining algorithms that can also be implemented for medical diagnosis. Some of these algorithms include K-means Algorithm, Support Vector Machines and Naive Bayes. Also, a larger dataset can also be used, to get a better model that is robust.

References

1. Abdullah A. Aljumah and Mohammad Khubeb Siddiqui. (2014), Hypertension Interventions using Classification Based Data Mining: Research Journal of Applied Sciences, Engineering and Technology 7(17): 3593-3602, 2014 ISSN: 2040-7459
2. Adesesan B. Adeyemo, Olufemi Oriola, Olamide Olaniyan and Endurance Awokola. (2010), Practical Challenges of Setting up Data Mining Projects in the Nigerian Environment: Proceedings of the International Conference on Software Engineering and Intelligent Systems 2010, July 5th-9th, Ota, Nigeria
3. Adeyemo omowunmi, adewale P., Ogunbiyi D. and Samson O. (2014) Prediction and classification capabilities of decision tree algorithms in modelling: Transition from Observation to knowledge to intelligent. Pp 2-29.
4. Adeyemo O.O, Adeyeye T.O and D. Ogunbiyi, (2015), Comparative Study of ID3/C4.5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever, African Journal of Computing & ICT, Vol 8. No. 1 – March, 2015, pg 103-112
5. Alao D. & Adeyemo A. B. (2012), Analyzing Employee Attrition Using Decision Tree Algorithms
6. Bharati, M. Ramageri, (2010), Data Mining Techniques and Applications, Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305
7. Computing, Information Systems & Development Informatics Vol. 4 No. 1 March, 2013
8. Jiawei Han and Michelin Kamber, (2006). Data Mining: Concepts and Techniques, Second Edition. Publisher: Morgan Kaufmann, New York.
9. Mahendra Tiwari, Manu Bhai Jha, OmPrakash Yadav,(2012). Performance analysis of data mining algorithms in weka. IOSR Journal of Computer Engineering, ISSN: 2278-0661, ISBN: 2278-8727 Vol. 6, Issue 3. Pg 32-41
10. Michael J. A. Berry, Gordon S. Linoff. (2004), Data mining techniques second edition by publisher: Wiley Inc, Indiana 2004.
11. Muhammad Arif, Khubaib Amjad Alam and Mehdi Hussain. Application of Data Mining Using Artificial Neural Network: Survey International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.245-270
12. P. Gaur, “Neural Networks in Data Mining”, International Journal of Electronics and Computer Science Engineering (IJECSE, ISSN: 2277-1956), vol. 1, (2012), pp. 1449-1453.
13. Razali, A.M. and S. Ali, Generating Treatment Plan in Medicine: A Data Mining Approach. American Journal of Applied Sciences,2009. 6 (2): 345-351.
14. Thirunavukkarasu, K. S. and Sugumaran S, (2013). Analysis of classification techniques in data mining . Int. J. Eng. Sci. Res. Technol. 3740-3746.