An Efficient Algorithm for the Prediction of Cancer of the Kidney Using Data Analytic Technique

Aranuwa Felix Ola
Aekunle Ajasin University,
Akungba – Akoko, Ondo State, Nigeria
+2347031341911
felix.aranuwa@aaua.edu.ng

Ogundare Olanike
Malaysia University of Science and Technology,
Selangor, Malaysia
+6010212624
ogundareolanike@yahoo.com

Sellappan Palaniappan
Malaysia University of Science and Technology,
Selangor, Malaysia +60192600962
sell@must.edu.my

ABSTRACT
Our focus in this research work is to present an efficient algorithm for apt prediction of cancer of the kidney in which medical practitioners and patients could gain valuable knowledge for early and proactive intervention strategies to save lives from this harmful disease. To achieve these objectives, dataset pertaining to patients of cancer of the kidney were acquired from selected private and public hospitals in south west Nigeria. A two-layered classifier system consisting of Rule Induction (RI) and Decision Tree (DT) classifiers was designed to build the model based on data analytic approach. The classifier system designed was tested successfully using case study data from fifty-two (52) selected Local Governments in South West Nigeria using purposive and selective sampling technique. Ten classification algorithms were used in the modeling. Waikato Environment for Knowledge Analysis was used for the experiment and each model was built in two different ways (10-fold cross validation and percentage split mode). Performance comparison of the various algorithms considered was carried out using standard metrics of accuracy for classification and speed of model building benchmarks. The experimental results show that the J48 decision tree algorithm outperform all other algorithms in all the layers with correctly classified instances of 74.7%, F-Measure of 0.614, TP rate of 0.747, FP rate of 0.135, precision and recall of 0.687 and 0.714 respectively. It took the best algorithm, 0.03 seconds to build the model. This proves that the algorithm is suitable for the research purpose. The results from the system framework when tested with test data shows that the identified attributes, algorithm and the system model performed well and can serve as valuable tool for early detection of the disease in patients.

CCS Concepts
• **Software and its engineering → Software organization and properties → Extra-functional properties → Software performance**

Keywords
Data Analytics, Classification Algorithms, Data Mining, Kidney Cancer

1. INTRODUCTION
In Africa, experimental studies have shown that most cancers are diagnosed at an advanced stage of the disease which usually contributes to its complications and mortality rate. This is due to a limited awareness of the early signs and symptoms of the disease among the public and healthcare providers. According to Lasebikan, Nwadinigwe & Onyegbule, (2014), the mortality rates of this disease is always compounded by the later stage at which the disease is diagnosed, presenting a ticking time bomb of life expectancy and lifestyle changes such as women having fewer children, as well as hormonal intervention such as post-menopausal hormonal therapy [1]. To reduce this harm caused by the disease, an effective way is to detect it early [2]. However, early detection and prognosis requires an accurate information, reliable analytic procedure and efficient algorithm. Therefore, the researcher's direction in this work is to present a reliable analytic procedure and efficient algorithm suitable for the prediction of cancer of the kidney through data analytic approach, in which medical practitioners and patients can gain valuable knowledge and help for proactive intervention strategies in order to save lives from this harmful disease.

Data analytic has proven to be a multi-dimensional discipline that uses descriptive techniques and predictive models to gain valuable knowledge from data warehouses for recommendations and decision making. It is the discovery of patterns and communication of meaningful insight in data [3]. According to Berson, Smith and Thearling (1999), data analytics is the science of examining raw data with the purpose of drawing conclusions from it [9]. It focuses on inference, identify undiscovered patterns and establish hidden relationships[4]. Figure 1 depicts the process of data analytics. The science is generally divided into exploratory data analysis (EDA), where new features in the data are discovered and confirmatory data analysis (CDA) where existing hypotheses are proven true or false. Typically, it is used to describe the technical aspects of data analysis, especially predictive modeling, machine learning techniques. Data Analytics has been commonly apply to business data, marketing mix modeling, web analysis, risk analysis and fraud analysis to communicate insights from data. It is very good in recommending action and guide decision making,

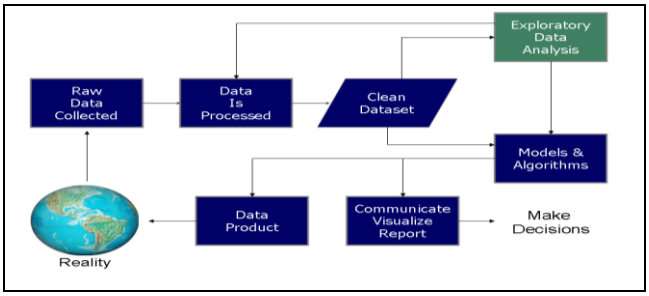

Figure 1: Data Analytics Process

CoRI'16, Sept 7–9, 2016, Ibadan, Nigeria.

7	**Age Group**		
	20-30	38	3.8
	31-40	150	15.0
	41-50	231	23.0
	51-60	240	23.9
	61-70	211	21.0
	70 -80	94	9.13
	81-90	42	4.17
	91-100	0	0

2. METHOD AND MATERIALS

S/N	Variable Name	Variable Format	Variable Type
1	Gender	Male, Female	Categorical
2	Age	25, 30,……..	Numerical
3	Lifestyle	Smoking, Obesity,	Categorical
4	G&H Disorder	Yes, No	Categorical
5	C & I Exposure	Yes No	Categorical
6	Prediction Level	One, Two, Three	Categorical

2.1 Data Collection and Data Format

Dataset pertaining to this research work was collected from selected health centres and hospitals in the south western part of Nigeria using purposive and selective sampling techniques. The researcher collected a sample data totaling, 1,006 records from fifty-two selected health centres in six (6) different states. The data collected was cleaned, normalized and organized in a form suitable for data analytic process. Table 1 shows the data format for the research data collection while Figure 1 and Figure 2 show the visualized information about selected states and health centres respectively.

Table 1 shows the data format for the research data collection

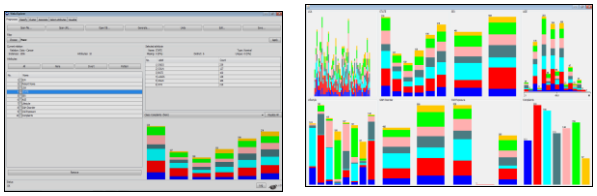

Figure 2: Visualize information about selected health centres in LGAs

2.2 Data Analysis & Interpretation

Statistically, out of the 1,006 patient's data captured, 44.8% were male while the remaining 55.2% are female, (See Table 2). The analysis further revealed that 57.1% of the patients are exposed to chemical and industrial contents while 32.7% of the population as gender and hereditary disorder. The patient's life style data collected also indicated that the people around this region are addicted to smoking and drinking of alcohol, regular use of non-steroidal anti-inflamatory drug (NSAIDs) such as ibuprofen and naproxen, which can double the risk of the disease by 51%. Other factors include obesity; faulty genes; a family history of kidney cancer; having kidney disease that needs dialysis; being infected with hepatitis C; and previous treatment for testicular cancer or cervical cancer. There is an indication also, that High blood pressure is a possible risk factor though still under investigation.

Table 2: Statistical Data for the Selected Attributes

S/N	Attributes	Data	Percentage (%)
1	**Gender**		
	Male	451	44.8
	Female	556	55.2
2	**Lifestyle**		
	Smoking	397	39.5
	Obesity	19	1.9
	Drug Abuse	134	13.3
	HB Pressure	106	10.53
	Water Pills	40	3.98
	Dialysis	8	0.8
	Alcohol	295	29.3
	Radiation	7	0.69
3	**G&H Disorder**		
	Yes	329	32.7
	No	677	67.3
4	**C&I Exposure**		
	Yes	576	57.3
	No	430	42.7
5	**Complaints**		
	Blood in Urine	113	11.23
	Back pain	203	20.17
	Tumor	189	18.8
	Fibroid	131	13.02
6	Stomach Ucher	144	14.31
	Kidney pain	159	15.8
	Abdominal pain	67	6.67

3. DESIGN OF EXPERIMENT AND RESULTS

3.1 Research Experimental Platform

Waikato Environment for Knowledge Analysis (WEKA) platform was used for the data analytic experiment. It is a powerful data mining tool that has a GUI Chooser from which any of the four major WEKA application environments (*Explorer, Experimenter, KnowledgeFlow and Simple CLI*) can be selected. The Explorer Application is selected for this experiment because it has a workbench that contains a collection of visualization tools, data processing, attribute ranking and predictive modeling with graphical user interface (GUI) for easy access to this

functionalities, which are very important to the research work. WEKA is a collection of machine learning algorithms for data mining tasks. Algorithms implemented in WEKA include: Bayesian classifiers, Decision Trees, Rules, Artificial Neural Network (Functions), Lazy classifiers and miscellaneous classifiers. But for the purpose of this work Rule Induction and Decision Tree classifiers was considered. These families of classifiers have been selected because of their performances in various domains. They have both been successfully applied to a variety of real-world classification tasks in industry, business, science and education with good performances [10]. The classifier system designed for the data modeling as shown in Figure 3 is of two layers: Layer 1 consists of JRiP, PART and Decision Table of the family of Rules Induction and Layer 2 consists of J48, LAD Tree, Decision Stump, Random Forest, Rep Tree, BF Tree, and LMT from the family of Decision Tree. The Decision Tree also known as "white box" classification model can provide explanation for their models, and could be used directly for decision making [5], while the Rule Induction is one of the fundamental tools of data mining, in which formal rules are extracted from a set of observations. The rules extracted represent a full scientific model of the data [6]. According to Kapil et al., (2013), rule induction is a popular and well researched method for discovering interesting relations between variables in large database. These abilities and aptitudes of rule induction are suited and of good requirement for any effective and efficient intelligent system. A major paradigm of the Rule Induction is the Association Rules [7].

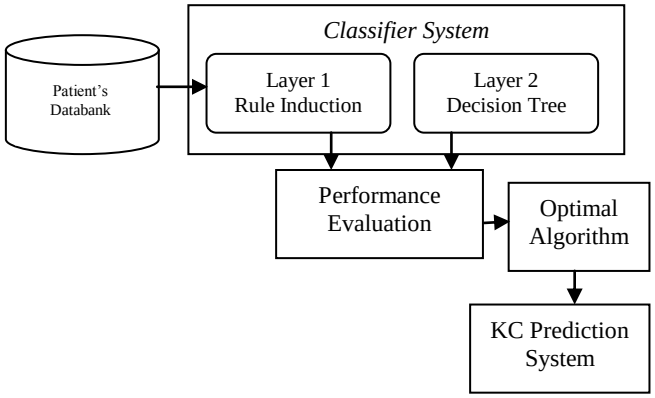

Figure 3: Designed Classifier System

As shown in Figure 3, the patient's databank component is responsible for the data collection, updating and storing patient's data from different sources. The classifier system component is responsible for the data modeling based on the algorithms in the layers. The performance evaluation component is responsible for the evaluation of the performance of the algorithms considered in the layers using standard metric to produce the best (optimal) algorithm. The rule generated from this algorithm is to be incorporated into the prediction system. Since the objective of the research work is to present a suitable algorithm for the cancer of the kidney prediction system, which the work has achieved. Hence the prediction system processes is not discussed in the work, but will be discussed in the future work of this research.

3.2 Experimental Results

Ten (10) classification algorithms from the family of classifiers implemented in this work were used to model the patient's dataset. The datasets for the experiment was first divided into two, which includes the training and testing datasets. 66% of the datasets was devoted to training while the remaining 34% was used for testing of randomly selected data. JRip, PART and Decision Table in layer 1 of the classifier system were first used to model the patient's data and later the Decision Tree classifiers. The 10-fold cross validation test and percentage split modes were also considered in the modeling. Since they are from different classifiers family, they yielded different models that classify differently on some inputs. The algorithms were tested on the datasets in order to determine that which best models the data with best predictive accuracy.

The comparison of the performance of the various algorithms in layer 1 and layer 2 based on the output from the percentage split (hold-out) and 10-fold cross validation modes was carried out. The results of the models from the two modes and the performance evaluations are presented in Table 3. The 10-fold cross-validation test mode was considered good since it produced the best model both in layer 1 and 2 of the classifier system. Moreover, the 10-fold cross validation mode have been widely used, and it is described a better option to determine the performance of a classifier [8]. Table 4 shows the standard metric accuracy details from the 10-fold cross validation mode considered for all the algorithms in the experiment. Figure 4 and Figure 5 show the graphs of predictive accuracy and time taken to build the models by the classifiers respectively.

Table 3: Classification Accuracy Comparison between Hold-out and 10-fold Cross Validations in Layer 1 and Layer 2

		10-fold Cross Validation		Hold-out (Percentage Split)	
S/N	Classifiers	Correctly Classified Instances	Time taken to build model	Correctly Classified Instances	Time taken to build model
1	J48 Decision Tree	74.7	0.03	74.5	0.02
2	LMT	74.6	29.25	73.7	29.03
3	LAD Tree	72.6	0.92	73.1	0.91
4	RepTree	71.6	2.54	72.4	2.4
5	JRiP Rules	70.9	0.03	70.1	0.03
6	PART	70.8	0.02	71.8	0.03
7	Decision Table	70.2	0.03	70.3	0.03
8	Random Forest	69.6	0.13	70.7	0.11
9	Decision Stump	64.7	0.01	64.9	0.01
10	BF Tree	57.9	2.54	60.8	2.55

Table 4: Compared standard metric accuracy details for all the Classification Algorithms

S/N	Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Built Time(s)	Correctly classified %
1	J48 Decision Tree	0.747	0.135	0.687	0.714	0.614	0.78	0.03	74.7
2	LMT	0.746	0.239	0.73	0.746	0.733	0.863	29.25	74.6
3	LAD Tree	0.731	0.292	0.714	0.731	0.702	0.85	0.91	73.1
4	RepTree	0.716	0.548	0.536	0.658	0.533	0.571	0.03	71.6
5	JRiP	0.709	0.274	0.728	0.749	0.731	0.754	0.06	70.9
6	PART	0.718	0.294	0.694	0.718	0.695	0.814	0.03	71.8
7	Decision Table	0.704	0.238	0.716	0.704	0.702	0.816	0.05	70.4
8	Decision Stump	0.649	0.36	0.579	0.647	0.612	0.669	0.02	64.9
9	Random Forest	0.643	0.327	0.622	0.643	0.629	0.74	0.08	64.3
10	BF Tree	0.579	0.223	0.718	0.716	0.717	0.748	2.54	57.9

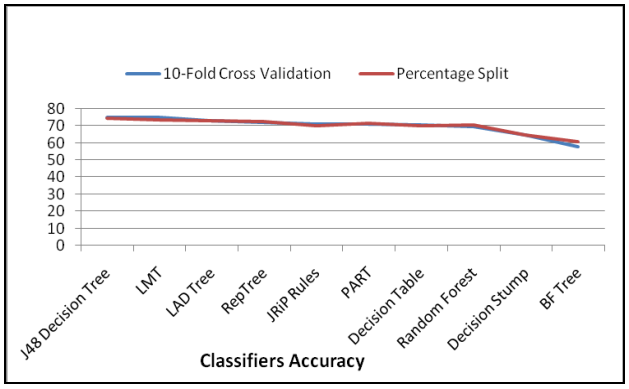

Figure 4 Predictive Accuracy of Classifiers in Layers 1 and 2 for both 10-fold and Hold-out (Percentage Split) Validations

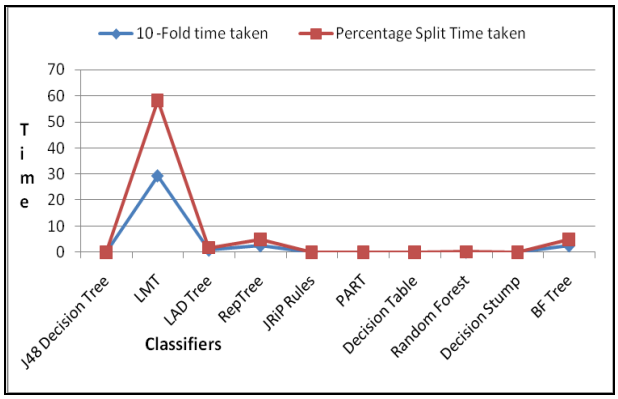

Figure 5: Time Taken by the Classifiers to build Models in Layers 1 and 2 for both 10-fold cross validations and percentage Split (hold- out)

From the experimental results and analysis, it shows that the J48 decision tree and LMT rules outperform all other algorithms in the layers. However, J48 decision tree was chosen as the best algorithm in this work because it has the correctly classified instances of 74.7%, ROC Area of 0.78 and recall of 0.714 respectively. It has a lower FP rate of 0.153, F-Measure of 0.614 and took lesser time of 0.03 seconds to build the model compared to LMT and other classifiers as shown in Table 4. Additionally, J48 decision tree algorithms generally have this ability that can produce a simple tree structure with high accuracy in term of classification rate, even with huge volume of data [9]. Pruning methods have been introduced to reduce the complexity of tree structure without any decrease in classification accuracy. The J48 decision tree structure and rules as generated by WEKA are presented in Figure 6.

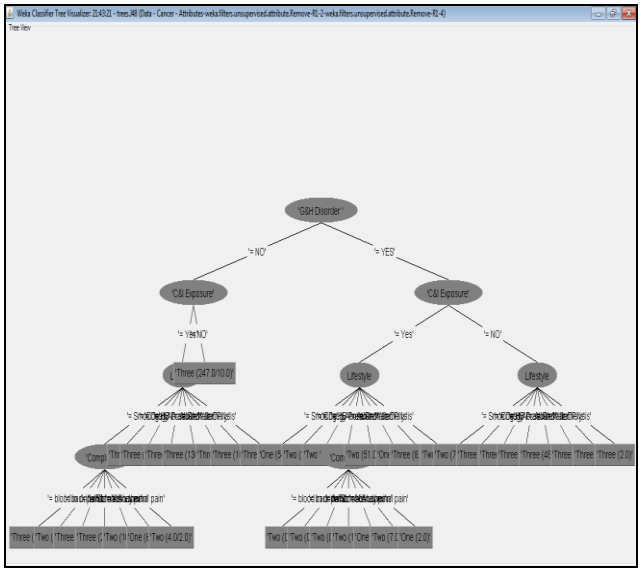

Figure 6: J48 Decision Tree Structure as presented by WEKA

The rules generated from the best algorithm (J48 pruned decision tree) are as stated in rules 1 to 20. The rules were tested in a prediction system framework and their prediction levels are classified as follows: (PL) – One, Two and Three. This show the status of patients and by interpretation: Level One and Two indicates a risk level or status of the disease manifestation in the patients that needs to be attended to urgently. While, level Three indicates that the patient is not manifesting any symptoms of kidney cancer disease, but may suffer from other diseases. A back-end for updating the rules as the situation arises will be incorporated into the system to match other conditions.

Rule 1: IF (G&H Disorder = NO) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = blood in urine: PL = One

Rule 2: IF (G&H Disorder = NO) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = back pain: PL = Two

Rule 3: IF (G&H Disorder = NO) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = tumor: PL = Three

Rule 4: IF (G&H Disorder = NO) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = Fibroids: PL = Three

Rule 5: IF (G&H Disorder = NO) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = Stomach ucher : PL = Two

Rule 6: IF (G&H Disorder = NO) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = Kidney pain: One

Rule 7 IF (G&H Disorder = NO) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = Abdominal pain: Two

Rule 8 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = blood in urine: PL = One

Rule 9 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Obesity) AND Complaints = blood in urine: PL = Two

Rule 10 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = HB Pressure) AND Complaints = blood in urine: PL = Two

Rule 11 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = Drug Abuse OR Tumor OR Fibroids: PL = Two

Rule 12 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = Abdominal pain: PL = Two

Rule 13 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = Kidney pain: PL = One

Rule 14 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Smoking) AND Complaints = stomach ucher: PL = One

Rule 15 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Alcohol OR Dialysis) AND Complaints = stomach ucher: PL = Two

Rule 16 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Radiation) AND Complaints = stomach ucher OR blood in urine: PL = One

Rule 17 IF (G&H Disorder = YES) AND (C&I Exposure = Yes) AND (Lifestyle = Water pills) AND Complaints = stomach ucher: PL = Three **Rule 18** IF (G&H Disorder = YES) AND (C&I Exposure = NO) AND (Lifestyle = Smoking) AND Complaints = stomach ucher OR kidney pain: PL = One

Rule 19 IF (G&H Disorder = YES) AND (C&I Exposure = NO) AND (Lifestyle = Smoking) AND Complaints = stomach ucher: PL = Two

Rule 20 IF (G&H Disorder = YES) AND (C&I Exposure = NO) AND (Lifestyle = Smoking OR Obesity OR Drug Abuse OR Radiation OR Water Pills OR Dialysis) AND Complaints = stomach ucher: PL = Three

4. CONCLUSIONS

The research work was focused at presenting an efficient algorithm suitable for predicting the status of kidney cancer in patients. To achieve the objectives of the research work: (i). Dataset pertaining to patient was acquired from fifty LGA (52) selected Health Centres in the south western region of Nigeria using purposive and selective sampling techniques. (ii) the researcher developed a two-layered classifier system consists of Rule Induction and Decision Trees implemented on Waikato Environment for Knowledge Analysis (WEKA) to build the data model using data analytic approach, and (iii) different machine learning algorithms were used in search for the algorithm that produced the best model with predictive accuracy. In the experiment, ten (10) classification model algorithms from different classifier family were implemented on the patients'dataset. Since they are from different classifiers family, they yielded different models that classify differently on some inputs. The comparison of the performance of the various algorithms in layer 1 and layer 2, and the standard metrics of accuracy, precision, recall and f-measure for the best classifier considered in this work was carried out as shown in Table 3 and Table 4 respectively. The results show that the J48 decision tree outperform all other algorithms in the layers with predictive accuracy of correctly classified instances of 74.7 % in 0.03 seconds, ROC Area of 0.78, FP rate of 0.153, TP rate of 0.714, precision and recall of 0.614.

.

5. REFERENCES

[1] Lasebikan OA, Nwadinigwe CU, Onyegbule EC Pattern of bone tumours seen in a regional orthopaedic hospital in Nigeria.

[2] Kushi LH, Doyle C, McCullough M, et al. (2012). "American Cancer Society Guidelines on nutrition and physical activity for cancer prevention: reducing

[3] Kohavi, R., Rothleder, N. J', & Simoudis, A.P (2002): Emerging Trends in Business Analytics Published by ACM Volume 45 Issue 8, Pages 45-48 August 2002.

[4] Berson, Smith ad Thearling ((199)

[5] Romero, C., Olmo, J. L & Ventura, S (2013): A meta-learning approach for recommending a subset of white-box classification algorithms for Moodle datasets. Department of Computer Science, University of Cordoba, Spain.

[6] Grzymala-Busse, J. W (2013). Rule Induction - University of Kansas. Extracted 20-06-2013.

[7] Kapil, S., Sheveta, V., Heena, S., Richa, D & Jasreena, K. B (2013). A Hybrid Approach Based On Association Rule Mining and Rule Induction in Data Mining International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-1, March 2013 146.

[8] WEKA,(2011): WEKA Tutorial. The University of Waikato (2011). Available at: http://www.cs.waikato.ac.nz/ml/weka/, (Accessed 20 July, 2013).

[9] Mohamed, W. Nor Haizan W, Mohd N. S, & Abdul H. O (2012). A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms. IEEE International Conference on Control System, Computing and Engineering, 23 - 25 Nov. 2012, Penang, Malaysia