ISSN 1512-1232

BREAST CANCER DIAGNOSTIC SYSTEM USING DECISION TREE ALGORITHM AND SYNTHETIC SUPPORT VECTOR MACHINE

Taye Oladele Aro[1], Muyiwa Adeolu Olugbebi[2], Samuel Oladeji Omokanye[3], Akande Hakeem Babalola[4]

[1]Department of Computer Science, University of Ilorin, Ilorin, Nigeria.
Email: taiwo774@gmail.com

[2]Department of Mechanical Engineering, University of Ladoke Akintola University Ogbomosho, Oyo, Nigeria.
Email:olugbebimuyiwa@gmail.com

[3]Department of Computer Science, University of Ilorin, Ilorin, Nigeria.
Email: oladejiomokanye@yahoo.com

[4]Department of Computer Science, University of Ilorin, Ilorin, Nigeria.
Email:hakeemobabs@gmail.com

Abstract
Breast cancer is the most common cancer among women in the Africa. Every thirteen minutes a woman dies of breast cancer. These facts have led researchers to continue studying how to diagnose and treat breast cancer in women, especially older women, who are at higher risk. Sonography (ultrasound) has become a great addition to mammography and magnetic resonance imaging (MRI) for imaging techniques dedicated to providing breast cancer screening. Identifying a high classifier algorithm that will help to proffer solutions to medical experts is crucial to the development of medical data expert systems for diagnosis of breast cancer in women. This paper focuses majorly on a study to improve the general low accuracy in classification algorithms by hybridizing Support Vectors Machine and Classification Regression Tree Decision Algorithm (CART) for breast cancer diagnosis. Two cases; Case A and Case B are mentioned, the result of Case B shows higher accuracy of 95.032400% with low mis-classification of 4.9676%, when synthetic support vector machine is used with Decision Tree compared to Case A when synthetic SVM is not applied.

Keywords: *Breast cancer, Sonography, Ultrasound, Mammography, Decision Tree, Support Vector Machine and Classification Regression Decision Tree Model*

1. Introduction

Breast cancer is serious ailment which has been discovered to be second cause of death among women in the society [1]. It is the most common of cancer in females that is affecting approximately 10% of women population at some stage of their life [2]. Among the medical applications, breast cancer diagnosis and prognosis pose a major challenge to researchers in medical [3]. The medical experts are faced with various problems in the prognosis and diagnosis of some diseases in which breast cancer is a good example. Several issues that often occur in breast cancer prediction include: existing predictive system is expensive and time consuming, inadequate understanding of symptoms and risk factor from patient, lack of quality diagnostic measure for patient, no adequate information to properly predict the case of breast cancer in the society.

ISSN 1512-1232

In data mining, classification is one of the most important tasks, it maps data into predefined targets [4]. Numerous data mining classification algorithms such as Decision Tree, Neural Network, Support Vector Machine (SVM) and Naïve Bayes have been proposed for prediction of breast cancer disease [5]. Identifying a high classification methods in data mining that will help to proffer solutions to medical experts is very crucial to the development of medical data expert systems for prediction or diagnosis of breast cancer in woman. The success of Adaboost can be traceable to its ability to enlarge margin, which could improve the generalization ability of this technique [6]. Adaboost classifiers can improve the classification accuracy as well as reduce the processing time and perform reliability better for defect classification [6]. SVM is a robust machine learning algorithm that separates different classes of data by a hyperplane [7], while a Decision Tree is another classification method in data mining that possesses ability to generate an understandable rules[8].

Hence, there is need in the prediction algorithm of breast cancer using data mining techniques to consider properly the effectiveness of machine learning algorithms that will function optimally for prediction. The choice of classifier use in data mining technique may affect the accuracy of a predictive system in healthcare industries that will help the medical experts predict breast cancer. To provide solution to the problem of deciding which classifier will perform better in the prediction task, a study on a comparative analysis of performance capability in data mining classification algorithms must be carried out. This study presented improved diagnostic system for breast cancer disease using Classification Regression Decision Tree Algorithm and Synthetic Support Vector Machine.

2. Overview of Breast Cancer

Breast cancer is a malignant tumour that usually begins with the formation of a small, confined tumor (lump) or as calcium deposits in the breast tissue [9]. It is found mostly in women, but men can get breast cancer too. Normal cells in the breast and other parts of the body grow (reproduce) and divide to form new cells as they are needed. When a person becomes adult, most cells divide only to replace damaged, worn-out or dying cells. Sometimes, cells in a part of the body grow and divide out of control, which creates a mass of tissue called a tumor. If the cells that are growing out of control are normal cells, the tumor is called benign (not cancerous). If, however, the cells that are growing out of control are abnormal and don't function like the body's normal cells, the tumor is called malignant (cancerous). The clear knowledge of breast aids in getting information about the parts of the breasts as shown in figure 1.

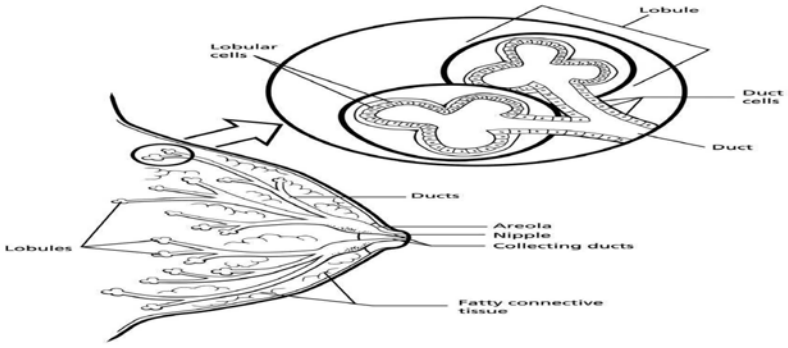


Figure 1. Normal Breast Parts [10]

3. Classification Techniques in Data Mining

Classification is used to find out in which group each data instance is related to within a given dataset [11]. It is used for classifying data into different classes according to some constraints. Classification techniques in data mining are capable of processing a large amount of data. Several

ISSN 1512-1232

major kinds of classification algorithms including C4.5, ID3, k-nearest neighbor classifier, Naive Bayes, SVM, and ANN are used for classification [12].

(a) Support Vector Machine (SVM)

The support vector machine algorithm applies linear models to implement nonlinear class boundaries by transforming the instance space using a nonlinear mapping into a new space, a linear model constructed in the new space can then represent a nonlinear decision boundary in the original space [7]. SVMs are based on an algorithm that finds a special kind of linear model called the maximum-margin hyperplane. The instances that are closest to the maximum-margin hyperplane, the ones with the minimum distance are called support vectors.

(b) Artificial Neural Network (ANN)

Artificial neural network is a computational model which draws its inspiration from biological nervous systems which comprises of neural network [13]. ANN consists of highly interconnected network of an enormous number of neurons, an architecture inspired by the brain. Neural networks learn by examples, they are trained with known examples of the problem that knowledge is to be acquired from, when trained well, the network can be used effectively to solve similar problems of unknown instances.

(c) Naïve Bayes (NB)

Naïve Bayes is based on Bayesian theorem rule and it assumes independence naively. This classification technique analyzes the relationship between each attribute and the class for each instance to derive a conditional probability for the relationship between the attribute values and the class. This technique has been identified to work effectively with actual datasets and when combined with feature selectors which eliminates redundancy and unimportant features.

(d) Decision Tree

Decision tree is one of the classification techniques in data mining method that is used for decision support system and machine learning process [8]. This classification technique plays a significant role in process of data mining and data analysis [14]. Generally, the structure of decision tree allows the applicability to understand the structure of trained knowledge models.

(e) Adaboost

It is a technique in data mining classification method for constructing a strong classifier as linear combination of a weak classifier [6]. Adaboost classifier can be built using fewer features and is considered more appropriate for real time applications. Boosting is one of the most significant developments in classification approach. Boosting works by sequentially applying a classification algorithm to reweighted versions of the training data and then taking a weighted majority vote of the sequence of classifiers thus produced. For many classification algorithms, this simple strategy results in dramatic improvements in performance.

(f) K-Nearest Neighhood

In pattern recognition or data mining approach, the KNN algorithm is a method for classifying objects and data based on closest training samples in the feature space [15]. KNN is one of the simplest machine learning algorithms. It is simply based on the idea that objects that are near each other will also have similar characteristics. KNN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. This techniques is also referred to as memory-based classification as the training samples need to be memory at run-time [16]. The KNN is the fundamental and simplest classification technique when there is little or no prior knowledge about the distribution of the data.

ISSN 1512-1232

4. Related Work

Several studies have been conducted using data mining techniques in medical diagnosis of breast cancer. These include:

[5] performed a survey on the importance and usefulness of data mining techniques using of different data mining techniques such as classification, clustering, Decision Tree, Naïve Bayes. Comparison was done on different data mining techniques from clinical dataset with different accuracy. From the several literatures reviewed, it clearly observed that the most existing performance analysis of data mining techniques do not consider the feature selection phase before classification. [17] developed a system which combined K-means clustering algorithm and fuzzy rough feature set and discernibility nearest neighbour classifier for breast cancer diagnosis. The proposed model was compared with previous studies and shown to perform better than others with an accuracy of 98.9%.

[9] designed a diagnostic system of breast cancer using ensemble of data mining classification methods. The study made use of various intelligent techniques including Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN) and also the ensemble of these techniques. Experimental studies were done using SPSS Clementine software and results show that ensemble model outperformed the individual models according to the accuracy of the system, which is the performance evaluation metric In order to increase the efficiency of the models feature selection technique was applied. The models were analysed in term of other error measures like sensitivity and specifically.

[18] presented a novel multi-layered method combining clustering and decision tree technique was used to build a cancer risk prediction system to provide a cost effective and earlier warning to the users. The proposed system predicted lung, breast, oral, cervix, stomach and blood cancers. The study made use of data mining techniques such as classification, clustering and prediction to identify potential cancer patients. The developed predictive system estimated the risk of the breast cancer in the earlier stage and also validated by comparing its predicted results with patient's prior medical information.

[10] proposed a novel approach for breast cancer detection using data mining techniques. The work investigated the performance of different classifier methods. The data breast cancer data with a total 683 rows and 10 columns was proposed to be tested by using classification accuracy. The study analysed the breast cancer data available from the Wisconsin data from UCL machine learning with the aim of developing accurate predictive model for breast cancer using data mining techniques. [13] presented the different data mining classifiers on the database of breast cancer, by using classification accuracy with and without feature selection techniques. Feature selection increases the accuracy of the classifier because it eliminates irrelevant attributes. The experiment result showed that the feature selection enhances the accuracy of all three different classifiers, reduces the Mean Standard Error (MSE) and increase Receiver Operating Characteristics (ROC).

[19] explored the applicability of decision trees in predicting the presence of breast cancer and also analysed some other conventional supervised learning algorithms which are ID3, CART, Random tree, C4.5 and Naïve Bayes. Results presented shows that Random tree provided the highest accuracy. [20] focused on breast cancer diagnosis by combination of fuzzy systems and evolutionary algorithms. Fuzzy rules are desirable because of their interpretability by human experts. Ant colony algorithm is employed as evolutionary algorithm to optimize the obtained set of fuzzy rules. Results on breast cancer diagnosis data set from UCI machine learning repository show that the proposed approach would be capable of classifying cancer instances with high accuracy rate in addition to adequate interpretability of extracted rules. [21] investigated the effect of feature selection on the classification of the type of breast cancer, they used three attribute selectors, rank search, genetic search and greedy step and concluded that feature selection increases performance and reduces the time taken in classifying algorithms. It was also noted that Bayes naive classifier outperformed other algorithms used in the research, the other algorithms are J48, Classification via Regression, Logistic and One.

ISSN 1512-1232

5. Methodology

The developed system for classifying breast cancer disease follows sequential process to achieve diagnostic task as shown in figure 2

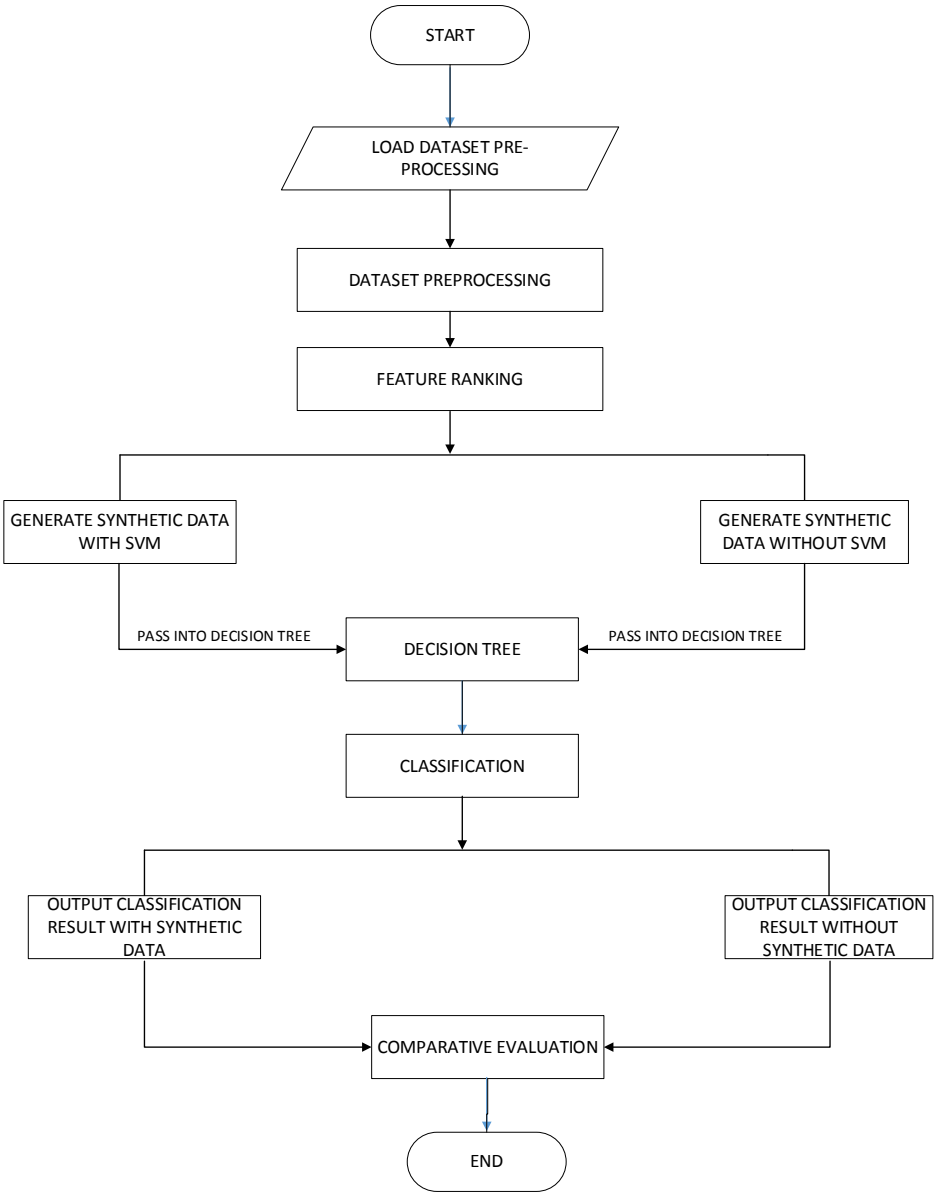


Figure 2. Flowchart of Breast Cancer Diagnostic System

The process of diagnostic system for breast cancer includes:

(i) Data Collection

Dataset was obtained and downloaded from the UCI repository Machine website, Breast Cancer Wisconsin Dataset (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+-(Original). The data set consists of the following attributes; Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Barei Nuclei, Bland Chomatin, Normal Nucleoli, Mitosis, The dataset was divided to the following classes; Class: (2 for benign, 4 for malignant). The data related to the predictors was provided to the network input

ISSN 1512-1232

layer. The data was classified as input and target. The classification was done based on attributes from Breast Cancer Wisconsin Dataset.

(ii) Data Preprocessing

The Data collected from the domain expert were subjected to pre-processing method for the purpose of data scaling and normalization using the mapminmax method.

(iii) Feature Selection Ranking

The chi-square feature selection was used to rank the breast cancer dataset according to their predictive power.

(iv) Generate Synthetic Data

Synthetic data was generated to improve generalization and the decision accuracy for the dataset by predicting a new class label by the support vector machine.

(v) Classification and prediction

At this phase the data with synthetic information was passed to the decision tree model for classification. Also the selected dataset without passing into the support vector machine for synthetic data was passed into the decision tree model.

6. System Developmental Stage

The overall developed system at run time is showed in the figure 3. The data mining system developed has the following tasks;

i. Loading of Dataset

ii. Feature selection

iii. Portioning of Dataset to training set and testing set

iv. Training of the SVM Model

v. Prediction of new Class label

vi. Generating Rules using decision tree

vii. Testing of unfamiliar test dataset

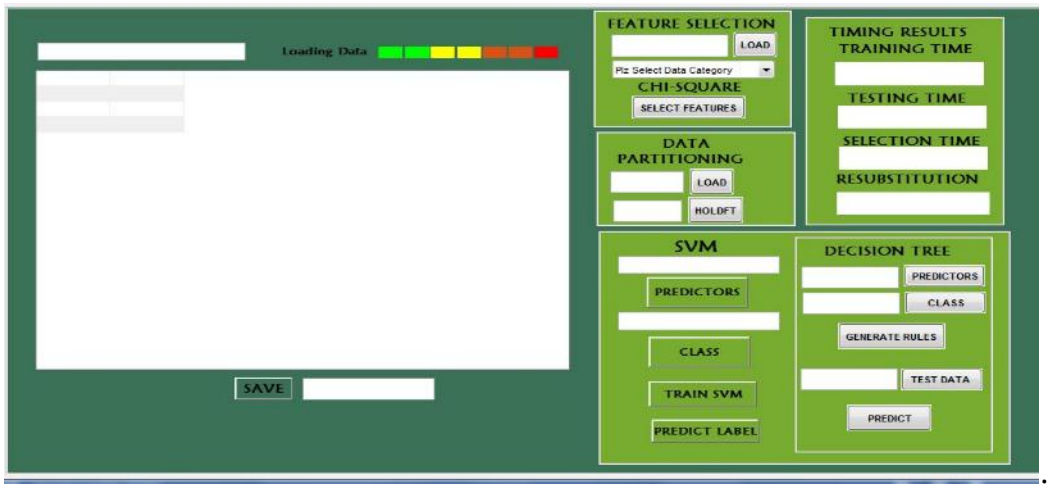


Figure 3. The system start up platform at run time

ISSN 1512-1232

6.1 Experimentation of Data and System Testing

The system was developed to establish a comparative performance evaluation for the improvement of classification regression tree with Support Vector Machine. The system was analyzed based on two experimental observation which are enumerated as follows:

(i) Feature Selection, Decision Tree Evaluation and Prediction of new instances based on the Decision Tree (Case study A)

This phase was evaluated based on the dataset, which was triggered by loading the breast cancer dataset into the platform and then filtered for better data generalization. The next event triggered the feature selection mode by using the chi-square feature selection technique to obtain optimal breast cancer dataset. After selection the data was split into training and testing phase at a proportion of 70% to 30% in ratio of TRAIN: TEST. The data then passed successfully splitting the data into these phases the training samples was passed into the decision tree to generate rules for prediction of new instances of the class label.

(ii) Feature Selection, Generation of Synthetic data, Decision Tree Evaluation and Prediction of new instances based on the Decision Tree (Case study B)

To fulfill the major aim of this study, it is necessary to evaluate the system based on synthetic dataset generation. In this phase the system was evaluated by introducing the dataset at initial loading after it passed through a data filtering, then passed into the feature selection phase, after selecting features with high predictive power the selected features were divided into training and testing phase at a proportion of 70% to 30% in ratio of TRAIN: TEST, then the training samples with its class label are passed into the support vector machine (SVM) to generate a synthetic data by predicting a new class label that represented the class label for the decision tree model. The last stage used training samples and the new predicted class label by the support vector machine (synthetic data) to generate the rules for the decision tree which in turns make predictions for new class label of group factors for benign and malignant.

7. Results and Discussion

Feature selection, decision tree evaluation and prediction of new instances based on the decision tree. Case study A result.

(A) Loading of dataset

The figure 4 shows data loaded into the platform. A total number of nine attributes with one column of class labels (2 and 4) which is the source that divides the data into group with a total number of 699 observation were loaded.

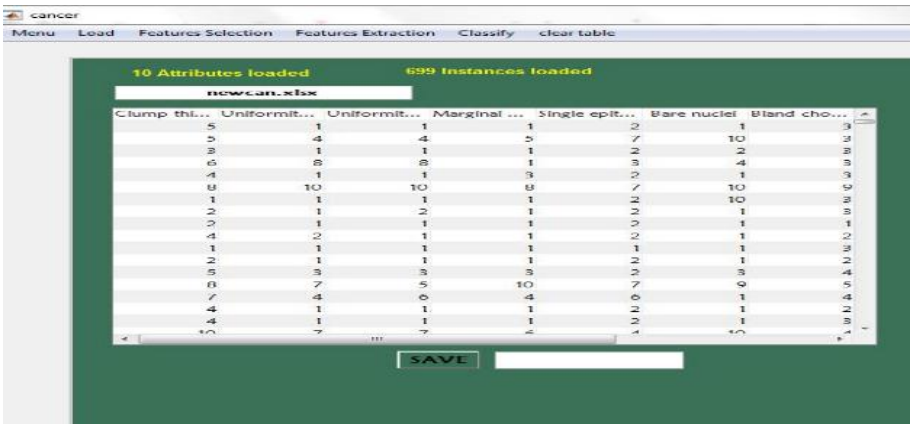


Figure 4. Loading of dataset

ISSN 1512-1232

(B) Feature selection

The window block shows the process of selecting features with the chi-square filter approach by loading the dataset and then selecting features with the full dataset option as shown in figure 5.

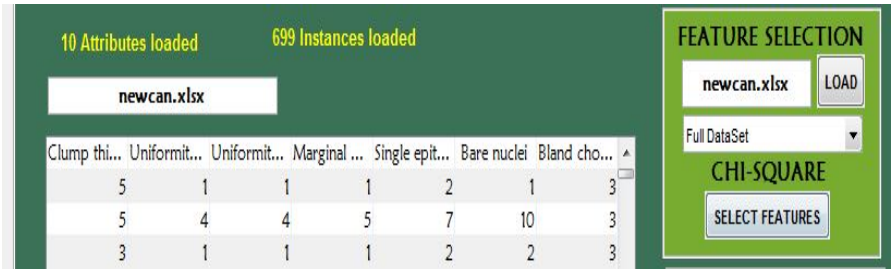


Figure 5. Feature Selection

(C) Attributes Selection

After the selection of features, attributes that were selected with their ranking score represented as shown in figure 6. The attributes selected is illustrated in table 1 and the collation of the selected features is depicted in figure 7.

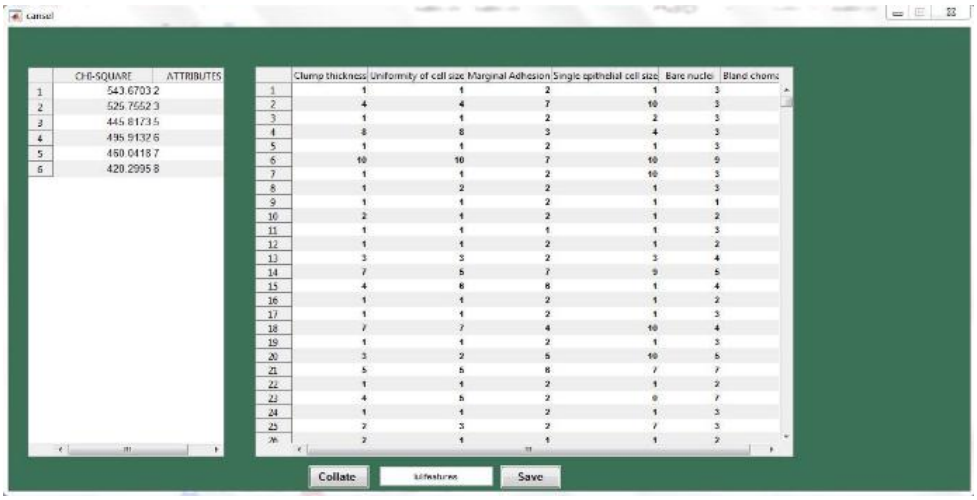
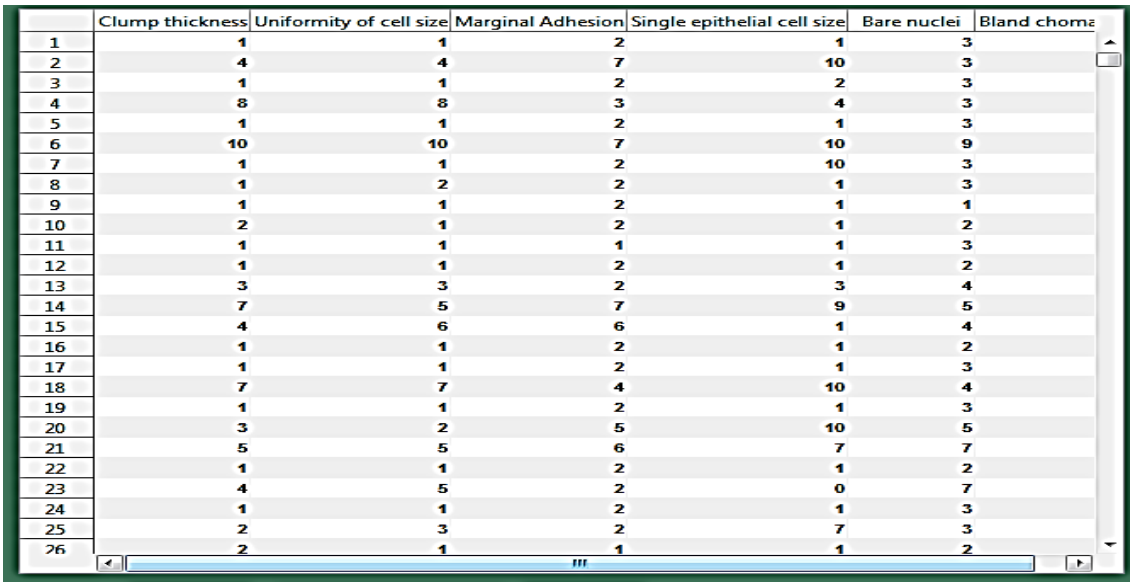


Figure 6. Selected Attributes with Ranking Score

Table 1: Selection of Attributes

Number	Attributes Selected
1	Clump thickness
2	Uniformity of cell size
3	Marginal adhesion
4	Single epithelial cell size
5	Bare nuclei
6	Bland Chromatin

ISSN 1512-1232



	Clump thickness	Uniformity of cell size	Marginal Adhesion	Single epithelial cell size	Bare nuclei	Bland choma
1	1	1	2	1	3	
2	4	4	7	10	3	
3	1	1	2	2	3	
4	8	8	3	4	3	
5	1	1	2	1	3	
6	10	10	7	10	9	
7	1	1	2	10	3	
8	1	2	2	1	3	
9	1	1	2	1	1	
10	2	1	2	1	2	
11	1	1	1	1	3	
12	1	1	2	1	2	
13	3	3	2	3	4	
14	7	5	7	9	5	
15	4	6	6	1	4	
16	1	1	2	1	2	
17	1	1	2	1	3	
18	7	7	4	10	4	
19	1	1	2	1	3	
20	3	2	5	10	5	
21	5	5	6	7	7	
22	1	1	2	1	2	
23	4	5	2	0	7	
24	1	1	2	1	3	
25	2	3	2	7	3	
26	2	1	1	1	2	

Figure 7. Collation of the selected features

(D) Division of the Data into Training and Testing Set

The figure 8 highlights when a hold out value of 0.3 was chosen so as to segment 30% for testing and the remains 70 % for training.

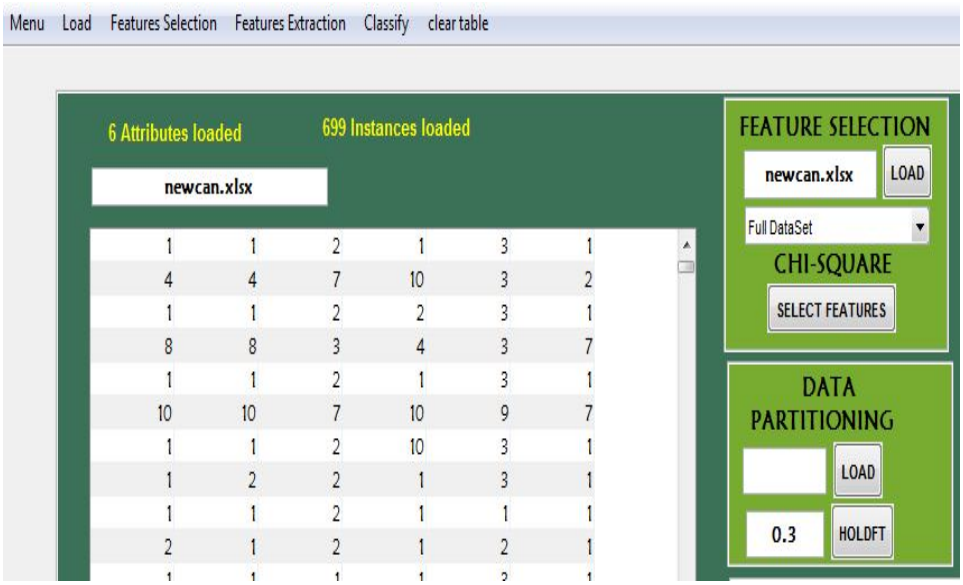


Figure 8. Hold out for Training and Testing

(E) Selection of Optimal Subset for Training and Testing

After selection of optimal subset, the selected subsets were divided into training set and testing set in the ratio of 70%: 30%. A total of 490 observations were occupied by the training samples while a total of 209 samples was occupied by the testing samples based on the ratio sampling as shown in figure 9.

ISSN 1512-1232

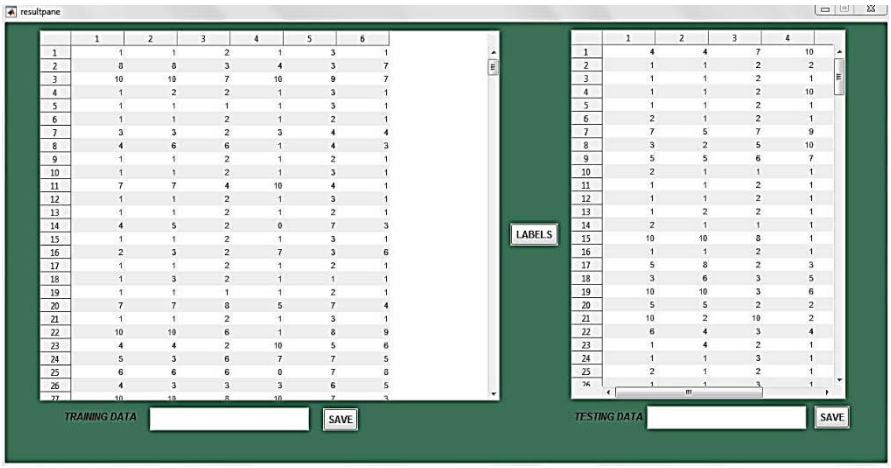


Figure 9. Partitioning of feature subset selection for training and testing samples

(F) Decision Tree Evaluation

The phase illustrates the scenario when the training set and its class label were loaded into the decision tree (CART), after which rules were generated as illustrated in figure 10.

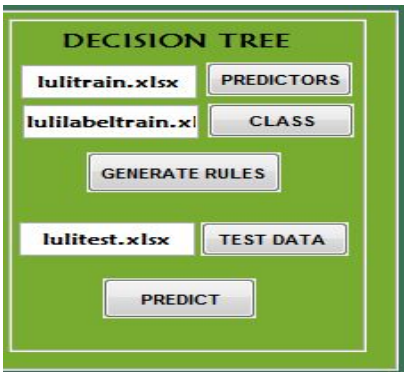


Figure 10. Training set and class label

(G) The rules obtained from the Decision Tree are highlighted with its symbolic representation as follows:

X1 - Clump thickness
X2 - Uniformity of cell size
X3 - Marginal adhesion
X4 - Single epithelial cell size
X5 - Bare nuclei
X6 - Bland chromatin.

(H) Result for Classification Using Decision Tree

The rules generated were passed into Decision Tree for classification. The results are as follows:

1. if x1<2.5 then node 2 elseif x1>=2.5 then node 3 else 2
2. if x4<6 then node 4 elseif x4>=6 then node 5 else 2
3. if x4<1.5 then node 6 elseif x4>=1.5 then node 7 else 4
4. if x6<9 then node 8 elseif x6>=9 then node 9 else 2
5. class = 4
6. if x1<6.5 then node 10 elseif x1>=6.5 then node 11 else 2

ISSN 1512-1232

7. if x1<4.5 then node 12 elseif x1>=4.5 then node 13 else 4
8. class = 2
9. class = 4
10. class = 2
11. class = 4
12. if x4<7.5 then node 14 elseif x4>=7.5 then node 15 else 4
13. class = 4
14. if x3<2.5 then node 16 elseif x3>=2.5 then node 17 else 4
15. class = 4
16. class = 2
17. if x3<3.5 then node 18 elseif x3>=3.5 then node 19 else 4
18. class = 4
19. class = 2

Resubstitution error =0.0245.

The rules generated were used to predict new occurrences in order to determine the accuracy and effectiveness of the predictive system. This was subjected to the 30% of test data that was left aside without passing through training. The figure 11(a) shows tree and figure 11 (b) shows sample scenario of the predicted class label.

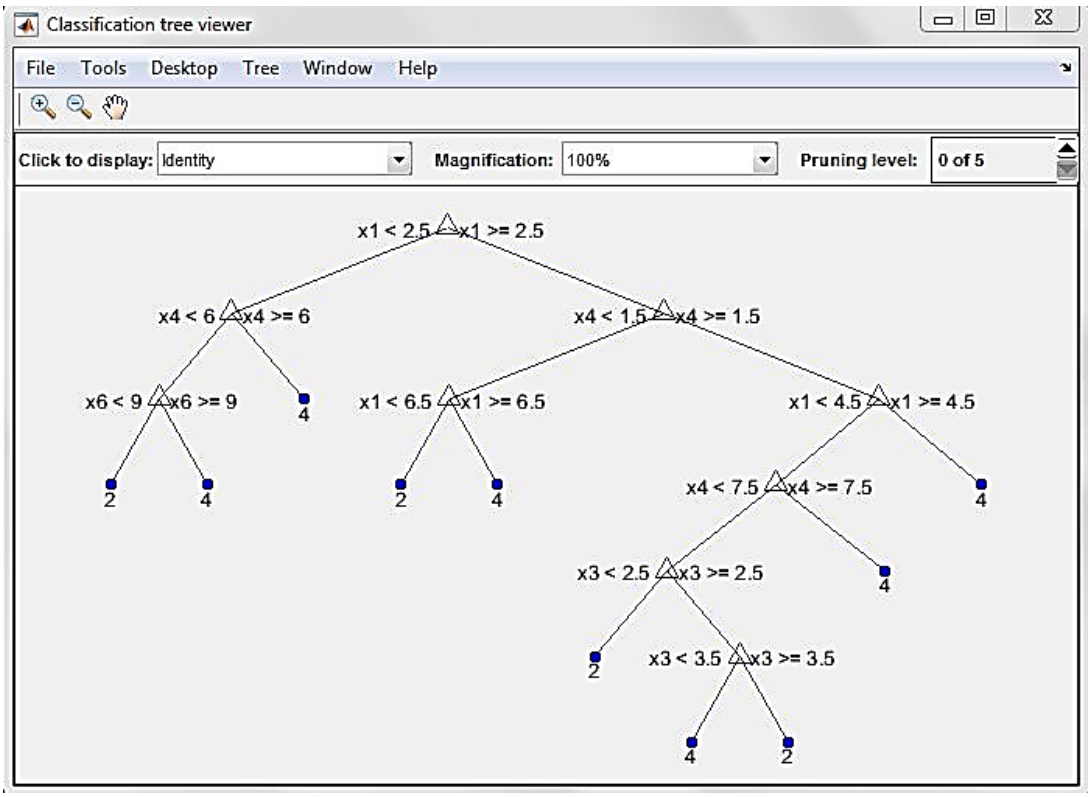


Figure 11 (a). Tree

ISSN 1512-1232

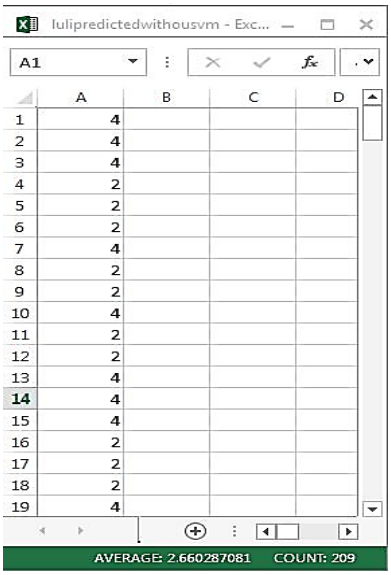


Figure 11 (b). Predicted class label

(I) Performance Evaluation of the Breast Cancer Diagnostic System

To measure the performance of the developed breast cancer predictive system, the predicted data and testing data were compared as shown in figure 12. The dataset consists of total of 209 observations, during the prediction 196 occurrences were predicted as correct while 13 occurrences were predicted as incorrect. The class 2 which represented benign has a total of 137 occurrences but the system predicted 132 as correct and 5 as incorrect. The class 4 which represent malignants has a total of 72 occurrences a total of 64 occurrences were predicted correctly and 8 occurrences were predicted incorrectly as shown in figure 13. The confusion matrix shows a clear overview of the classification accuracy between benign and malignant. The benign shows a classification accuracy of 88.8889% and 11.1111% Mis-classification accuracy. The Malignant shows a classification accuracy of 96.3604% and 3.6496% Mis-classification accuracy as shown in figure 14. Finally, system classification accuracy and misclassification accuracy is shown in table 2.

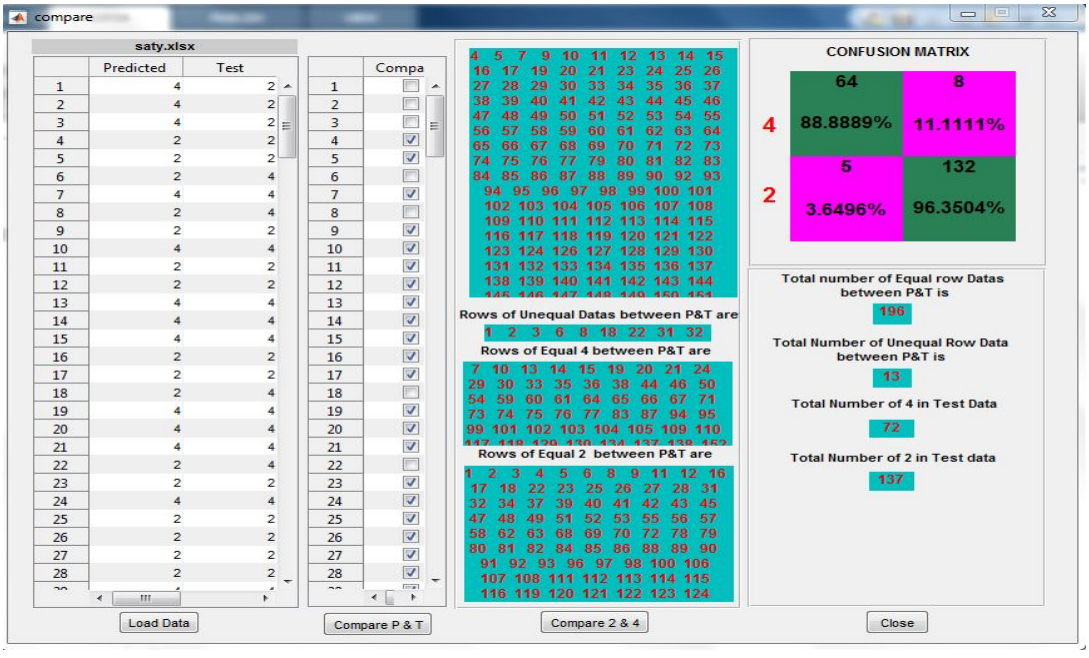


Figure 12. Performance Evaluator Model

ISSN 1512-1232

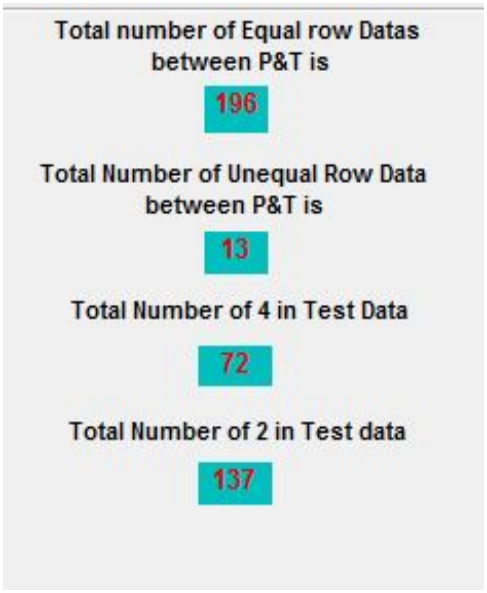


Figure 13. Performance Model

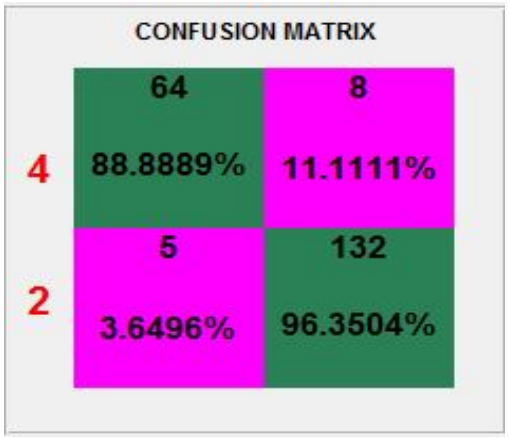


Figure 14. Confusion Matrix

Table 2. System Classification Accuracy and misclassification accuracy

System Classification Accuracy	System Misclassification Accuracy
92.16196%	7.3804%

6. 1. Result Analysis for feature selection, generation of synthetic data, decision tree evaluation and prediction of new instances based on the decision tree. CASE B.

(i) The second phase follows a systematic approach, the first phase followed the same steps that was carried out for feature selection and data division. The difference in this stage is the passing of dataset into the support vector machine to predict label in order to obtain the synthetic data as represented in figure 15. The training time for support vector machine is shown figure 16, the result of the predicted SVM class is shown figure 17.

ISSN 1512-1232

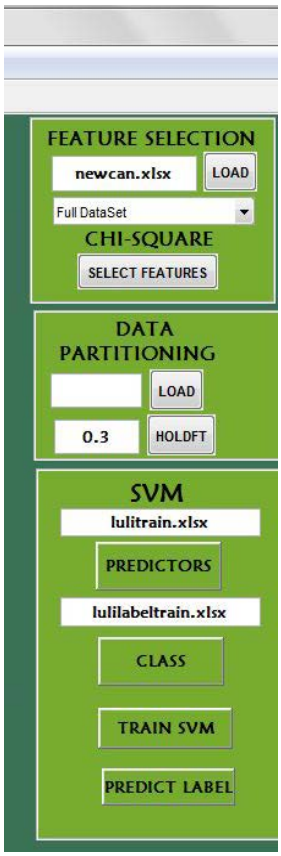


Figure 15. Predicted label for synthetic data

Figure 16. Training Time

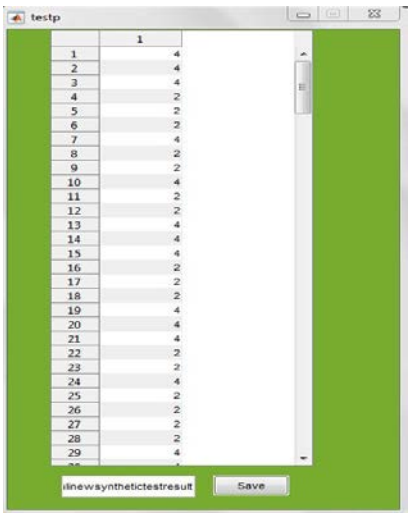


Figure 17. Result of Predicted SVM

ISSN 1512-1232

(ii) To evaluate the predicted class label by the support vector machine the predicted class label was passed into the decision tree model as shown in figure 18, which indicates the passing of the new predicted class label to the classification regression tree model. The predictors button loads the training data into the decision tree, the class button load the new predicted class label by the support vector into the decision tree while the generate rules is used to generate the rules. While the test data will evaluate the decision tree model with predicted class label on the testing dataset.

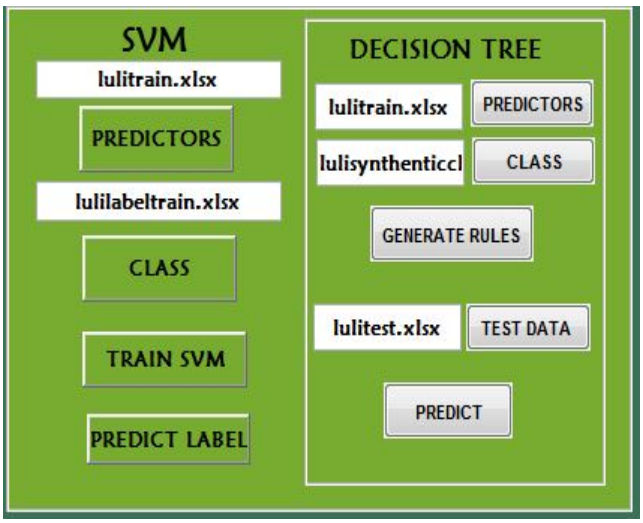


Figure 18**.** New predicted class label of classification regression tree model

(iii) Rules Generated from Decision Tree

Decision tree for classification where the label represents
X1-Clump thickness
X2-Uniformity of cell size
X3-Marginal adhesion
X3-Single epithelial cell size
X4-Bare nuclei
X5-Bland chromatin
1 if x1<2.5 then node 2 elseif x1>=2.5 then node 3 else 2
2 if x4<6 then node 4 elseif x4>=6 then node 5 else 2
3 if x2<2.5 then node 6 elseif x2>=2.5 then node 7 else 4
4 if x6<9 then node 8 elseif x6>=9 then node 9 else 2
5 class = 4
6 if x5<3.5 then node 10 elseif x5>=3.5 then node 11 else 2
7 if x4<3.5 then node 12 elseif x4>=3.5 then node 13 else 4
8 class = 2
9 class = 4
10 class = 2
11 class = 4
12 if x2<3.5 then node 14 elseif x2>=3.5 then node 15 else 4
13 class = 4
14 class = 2
15 class = 4
Resubstitution error = 0.0143
The predicted results is shown in the figure 19(a) and figure 19 (b) shows the predicted results by decision tree:

ISSN 1512-1232

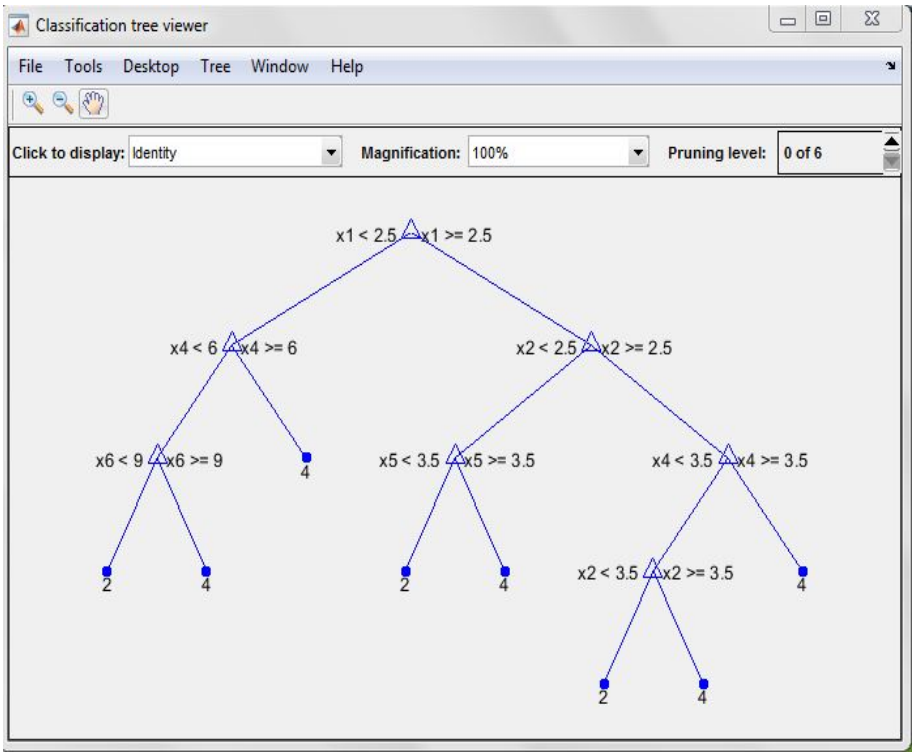


Figure 19 (a). Tree Model

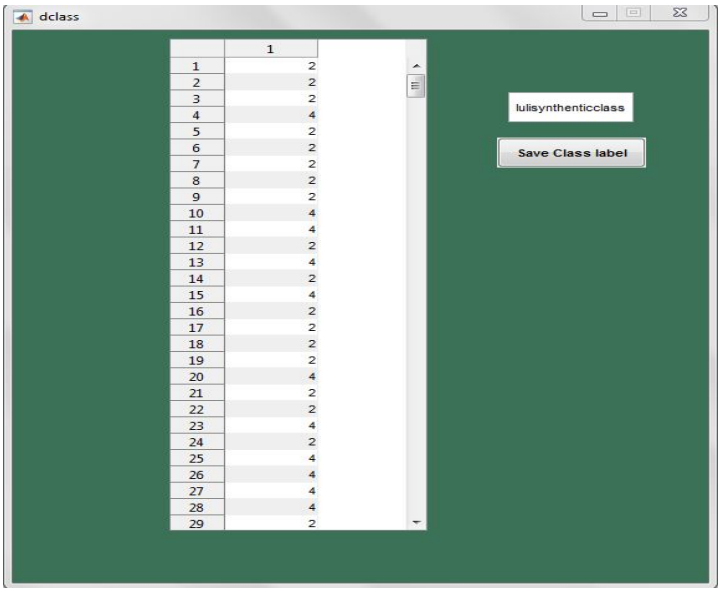


Figure 19 (b). Predicted results by decision tree.

(iv) The results is now passed to evaluation mode for a further validation check as illustrated in figure 20, then translated the results when the synthetic data was loaded to the evaluation platform and it validated the need of getting a synthetic dataset which gave a better accuracy more than when synthetic data is not obtained with support vector machine. The results obtained are better presented in the figure 21 which showed that out of the 209 observations 199 instances were classified correctly and 10 were not classified incorrectly. The confusion matrix gives a more elaborate result of the system with class representing benign and class 4 representing malignant as shown in figure 22**.** The malignant shows a total classification accuracy of 88.8889% and misclassification accuracy of 11.1111%. The benign shows a classification accuracy of 96.3504% and misclassification accuracy 3.64 96%.

ISSN 1512-1232

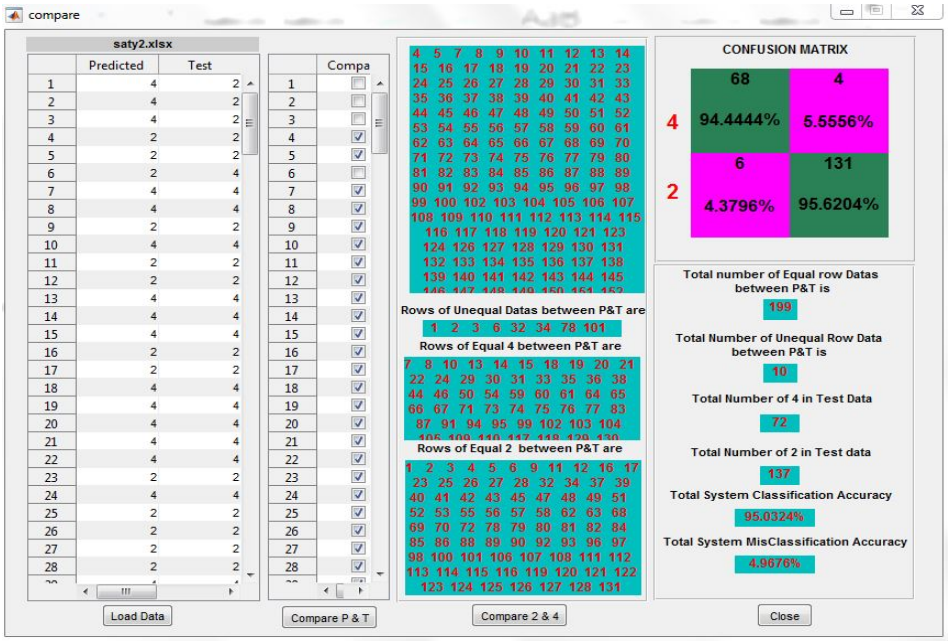


Figure 20. Performance Evaluator model

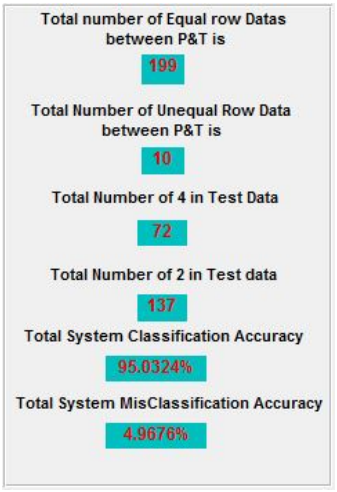


Figure 21. Result Panel

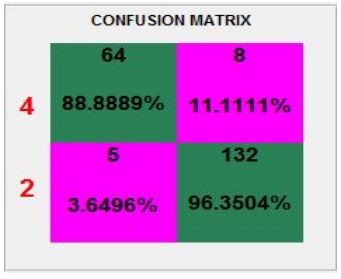


Figure 22. Confusion matrix

(v) The total classification accuracy and Mis-classification of the system is shown in table 3.

ISSN 1512-1232

Table 3: Total Classification Accuracy and Mis-classification

System Classification Accuracy	System Misclassification Accuracy
95.0324%	4.9676%

The results above shows a higher accuracy than the CASE A without SVM synthetic data generation.

(vi) Comparative Evaluation Between CASE A and CASE B

The classification accuracy used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. The accuracy is the proportion of true results (true positives and true negatives) among the total number of cases examined, the Resubstitution error measures the error on the model training state. It could be seen from the result that there were higher errors on the **CASE A** than the **CASE B** which has a lower Resubstitution error rate.

CASE A: Feature Selection, Decision Tree Evaluation and Prediction of new instances based on the Decision Tree.

CASE B: Feature Selection, Generation of Synthetic Data, Decision Tree Evaluation and Prediction of new instances based on the Decision Tree.

The table 4 and figure 23 and figure 24 present a comparative evaluation of the two cases carried out in this work. It is quite clear from the results obtained that the CASE B has a better performance that the CASE A due to higher value of its Classification Accuracy and lower value of its misclassification accuracy in comparison to the **CASE A**.

Table 4. Comparative Evaluation of Two Cases

Case Study	**Classification Rate**	**Mis classification rate**	**Resubstitution Error**
CASE A	92.161969%	7.3504%	0.0245
CASE B	95.032400%	4.9676%	0.0143

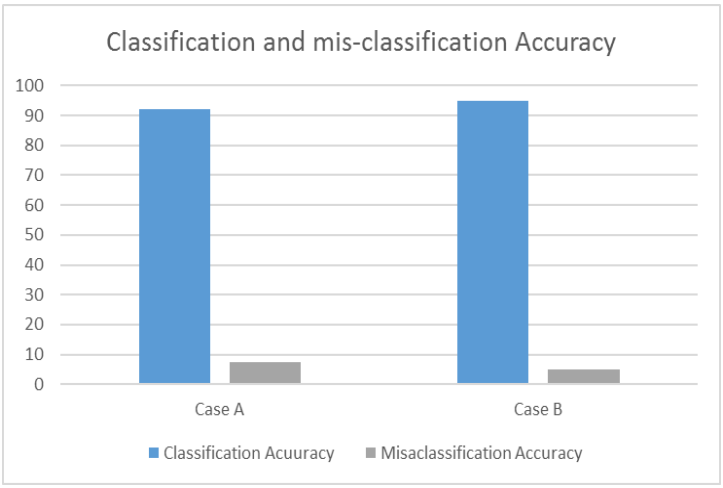


Figure 23. Classification and Mis-Classification Accuracy

ISSN 1512-1232

Figure 24. Re-Substitution Error

7. Conclusion

Cancer is one of life threaten or deadly diseases in the society today. Earlier stage detection of cancer is curable. Among other types of cancer, breast cancer is increasing daily and the mortality rate in women is not encouraging. The prediction of serious disease like breast cancer is a very challenging problem and it requires the development of robust system for diagnosis or prediction. This study discusses the development of an effective breast cancer diagnostic system by hybridizing Support Vector Machine with Decision Tree in order to improve the low accuracy in classification algorithms.

References

[1] I. O. Kehinde, W., Peter, A. I., Jeremiah, A. B. & Adeniran, "Breast Cancer Risk Prediction Using Data Mining," *Trans. Networks Commun.*, vol. 3, no. 2, pp. 1–12, 2015.

[2] J. Majali, R. Niranjan, V. Phatak, and O. Tadakhe, "Data Mining Techniques For Diagnosis And Prognosis Of Cancer," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 3, pp. 613–616, 2015.

[3] B. Padmapriya, "A survey on breast cancer analysis using data mining techniques," *IEEE Int. Conf. Comput. Intell. Compuing Res.*, pp. 4–5, 2014.

[4] H. Karim and K. Z and, "A Comparative Survey on Data Mining Techniques for Breast Cancer Diagnosis and Prediction," *Indian J. Fundam. Appl. Life Sci.*, vol. 5, no. 1, pp. 4330–4339, 2015.

[5] S. Shukla, D. L. Gupta, and B. R. Prasad, "Comparative Study of Recent Trends on Cancer Disease Prediction using Data Mining Techniques," *Int. J. Database Theory Appl.*, vol. 9, no. 9, pp. 107–118, 2016.

[6] E. R. Kaur and V. Chopra, "Implementing Adaboost and Enhanced Adaboost Algorithm in Web Mining," *Int. J. Adanced Res. Comput. Commun. Eng.*, vol. 4, no. 7, pp. 306–311, 2015.

[7] A. Karatzoglou, D. Meyer, and K. Hornik, "Support Vector Machines in R," *J. Stat. Softw.*, vol. 21, no. 9, pp. 1–26, 2005.

[8] Seema, M. Rathi, and Mamta, "Decision Tree: Data Mining Techniques," *Int. J. Latest Trends Eng. Technol.*, vol. 1, no. 3, pp. 150–155, 2012.

[9] G. Zorluoglu and M. Agaoglu, "Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods," *Int. J. Bioinforma. Biomed. Eng.*, vol. 1, no. 3, pp. 318–322, 2015.

[10] V. Chaurasia and S. Pal, "Data Mining Techniques : To Predict and Resolve Breast Cancer

ISSN 1512-1232

Survivability," *Int. J. Comput. Sci. Mob. Comput.*, vol. 3, no. 1, pp. 10–22, 2014.
[11] G. Porkodi, R & Suganya, "A Comparative Study of Different Deployment Models in a Cloud," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 5, pp. 512–515, 2015.
[12] S. Gupta, D. Kumar, and A. Sharma, "DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR," *Indian J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 188–195, 2011.
[13] A. Lebbe, S. Saabith, E. Sundararajan, and A. A. Bakar, "Comparative Study on Different Classification Techniques for Breast Cancer Datsset," *Int. J. Comput. Sci. Mob. Comput.*, vol. 3, no. 10, pp. 185–191, 2014.
[14] D. Singh, H. Naveen, and J. Samota, "Analysis of Data Mining Classification with Decision Tree Technique," *Glob. J. Comput. Sci. Technol.*, vol. 13, no. 13, pp. 1–6, 2013.
[15] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background," *Int. J. Eng. Res. Appl.*, vol. 3, no. 5, pp. 605–610, 2013.
[16] H. S. Khamis, K. W. Cheruiyot, and S. Kimani, "Application of k-Nearest Neighbour Classification in Medical Data Mining Application of k- Nearest Neighbour Classification in Medical Data Mining," *Int. J. Infromation Commun. Technol. Res.*, vol. 4, no. 4, pp. 121–128, 2014.
[17] I. M. El-hasnony, H. M. El-bakry, and A. A. Saleh, "Classification of Breast Cancer Using Softcomputing Techniques," *Int. J. Electron. Inf. Eng.*, vol. 4, no. 1, pp. 45–54, 2016.
[18] K. Arutchelvan and R. Periyasamy, "Cancer Prediction System Using Data mining Techniques," *Int. Res. J. Eng. Technol.*, vol. 2, no. 8, pp. 1179–1183, 2015.
[19] S. S. Shajahaan, S. Shanthi, and V. Manochitra, "Application of Data Mining Techniques to Model Breast Cancer Data," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 11, pp. 1–8, 2013.
[20] A. Einipour, "A Fuzzy-ACO Method for Detect Breast Cancer," *Glob. J. Health Sci.*, vol. 3, no. 2, pp. 195–199, 2011.
[21] G. Devi, "Breast Cancer Prediction System using Feature Selection and Data Mining Methods," *Int. J. Adv. Res. Comput. Sci.*, vol. 2, no. 976, pp. 81–87, 2011.

Article received: 2017-10-03