International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.2 (Apr-2021)

http://dx.doi.org/10.12785/ijcds/100141

Holistic Diagnosis Tool for Early Detection of Breast Cancer

Salim Amour Diwani[1] and Zaipuna Obedi Yonah[2]

[1] *Department of Computation Science and Engineering, Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania.*
[2] *Applied Engineering & ByteWorks (T) Limited, Dar es Salam. Tanzania.*

Received 10 Apr. 2020, Revised 17 Jul. 2020, Accepted 29 Jul. 2020, Published 01 Apr. 2021

Abstract: Globally, of all cancer diseases, breast cancer is the number one killer in women. The diseases commonly occur in high income countries, but recently there is rapid increase of breast cancer in middle and low income countries in Asia, Latin America and Africa. This is due to increase in life expectancy, increased urbanization and adoption of western cultures. Although, some strategies to reduce the risks of occurrence of breast cancer are being implemented in developed countries, the case in middle and low income countries is that majority of breast cancer patients are affected by the disease due to diagnosis at late stages of the disease. Therefore, early detection of breast cancer is needed to overcome this problem. In this paper, a holistic diagnosis tool for early detection of breast cancer is proposed. The tool is software based utilizing a novel prediction model for breast cancer survivability developed by using available data mining (DM) technologies. Specifically, five popular data mining algorithms (logistic regression, decision tree, support vector machine, K nearest neighbors and random forest) were used to develop the prediction tool using Wisconsin breast cancer data set. In the paper, prediction tool training and test set results are reported. Achieved from the reported work of training sets are classification accuracies of 100% (Decision Tree); 99.8046% (Random Forest); 97.46% (Logistic Regression and Support Vector Machine); 97.07% (K Nearest Neighbors) and for testing sets are classification accuracies of 93.5672% (Decision Tree); 92.9% (Random Forest); 92.39% (Logistic Regression, Support Vector Machine and K Nearest Neighbors). These results are much better than those reported in the literature. The results show that the proposed DM disease prediction tool has potential to greatly impact on current patient management, care and future interventions against the breast cancer disease and through customization even against other deadly diseases.

Keywords: Breast Cancer, Data Mining, Machine Learning, Data Mining Algorithms, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest and K-Nearest Neighbors.

1. INTRODUCTION

Medically, healthcare encompasses treatment, management and prevention of diseases based on patient's symptoms to address their health needs and preferences with the help of healthcare professionals using scientific medical knowledge including use of machine learning or data mining (DM) techniques. The complexity of healthcare DM techniques depends on the type of information to be mined [1]. Figure 1 shows different applications of machine learning in healthcare. Impliedly, different healthcare organizations have different databases depending on their needs and requirements. These databases are continually growing as organizations collect data from different sources, which include online transaction processing systems, operational support systems, day to day medical records, disease surveillance systems and medical research. In general, these massive amounts of data collected daily exceed the ability of traditional methods to analyze and extract the hidden knowledge patterns in those data. Therefore, discoveries of new diagnosis tools is a must. These tools will help to discover interesting or useful knowledge patterns or criteria hidden in those data repositories.

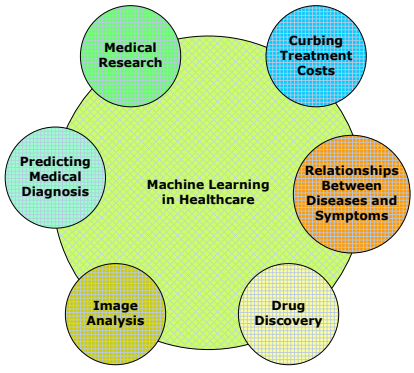

Figure 1. The main applications of machine learning in healthcare.

E-mail: diwania.salim@gmail.com, zaipuna.yonah@aebw.co.tz

In Tanzania, according to Tanzania breast health care assessment of the disease [2], 80% of women suffereing from breast cancer are diagnosed at late stages of the disease (stage III or IV), when treatment is less effective and outcomes are poor. Furthermore, protocols and guidelines for early detection of breast cancer and subsequent diagnosis and treatment are not standardized.

In this paper, an ongoing research work to develop and test a holistic prediction tool for benign and malignant breast cancer cases is reported. The tool is intended to assist healthcare professionals to achieve early detection of the disease and diagnosis. The paper is organized into six sections. Section II covers literature review related to the reported work. Section III describes the Methodology used in the design and testing of the prediction tool; and reports on the data collection component and cleaning algorithms of the prediction tool. Section IV reports about disease learning, classification and prediction components of the diagnosis tool and classifier performance criteria. Section V reports on test results from the breast cancer disease prediction tool. Section VI carries the conclusions.

2. LITERATURE REVIEW

Several researches that focus on breast cancer survivals are reported in the literature. Notably, different studies applied different approaches using the Wisconsin breast cancer data set aiming at achieving higher classification accuracies.

Jadhav *et al.*[3] used Wisconsin breast cancer data sets as training and testing samples. They deployed different machine learning algorithms for creating prediction models like decision tree, logistic regression and random forests, which are applied in the pre-processed data. Amongst all the models, Random Forest(RF) classification gave the best accuracy of 98.6%.

Diwani and Yonah [4] proposed a another prediction tool using best fit DM techniques. They reported on an ongoing research work to develop and test a holistic DM disease prediction (Diagnosis and prognosis) tool, equipped with processes for preprocessing patients' data and a learning procedure for selecting a disease-specific best classifier, for disease prediction and delivery of speedy and cost effective diagnostic interventions and patient follow up in a hospital environment. In their experiment, they used five different data sets for diabetes, breast cancer, heart diseases and liver disorder from University of California Irvin (UCI) data repository and HIV from Amana hospital in Tanzania. Seven algorithms were chosen: Naïve Bayes, J48, RepTree, SMO, LBK, PART and RandomForest. These algorithms were applied singly and in fusion. Test results for breast cancer and HIV data sets were reported. In this work, the authors achieved classification accuracies of 97.0752% (Classifier acting singly); 97.6323% (fusion of three classifiers).

Shravya *et al.* [5] also propose a DM approach for prediction of breast cancer. They used Wisconsin data set from UCI data repository to clsssify benign and malignant breast cancer. They developed models using Logistic Regression, Support Vector Machine and K-Nearest Neighbors. They used different metrics to measure performance such as accuracy, sensitivity, precission, specificity and false positive rate. In their experimental results, SVM gave the best predictive performance with an accuracy of 92.7%.

Sivapriya, *et al*. [6] also propose a breast cancer DM prediction tool. In their experiment, they used Wisconsin breast cancer data set from UCI data repository. They developed a model using SVM, logistic regression, Naïve Bayes and Random Forest classifiers and compared their performances. Based on their experimental results, the random forest (RF) algorithm gave the highest accuracy of 99.76% with the least error rate.

Asri *et al.* [7] proposed another set of DM algorithms for breast cancer risk prediction and diagnosis. In their research, which was conducted using the WEKA DM tool, compared the performance of different machine learning algorithms including: Support Vector Machine (SVM), Decision Tree (C4.5), Naïve Bayes (NB), and K-Nearest Neighbors (K-NN) acting on the Wisconsin breast cancer data sets. The main objectives were to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precission, sensitivity and specificity. Their reported results show that the SVM performed with highest accuracy of 97.13% with lowest error rate.

3. METHODOLOGY AND DATA COLLECTION

In this section, we present the methodology used in the reported work and the data collection process for the purpose of obtaining data for training and testing the prediction tool.

A: Data Collection Process (A-D in Figure 2)

The Wisconsin Breast Cancer dataset was collected from University of California Irvin (UCI) [8] machine learning data repository. Table 1 describes the original Wisconsin breast cancer data set, which has 699 instances or samples and 9 attributes with two classes: *Benign* and *Malignant*. In the data set, 16 instances containing missing values were removed from the data set, leaving 683 instances for experimentation.

Table 1. Training data set description for the experiment.

Attribute Name	*Description*	*Category*	*Range Values*
SCNo	*Simple Code Number*	*Id*	*-*
CT	*Clump Thickness*	*Ordinal*	*1-10*
UCSIZE	*Uniformity Of Cell Size*	*Ordinal*	*1-10*
UCSHAPE	*Uniformity Of Cell Shape*	*Ordinal*	*1-10*
MAF	*Marginal Adhession Fibrous: Fibrous bands tissue that form between two surfaces*	*Ordinal*	*1-10*

ECSIZE	*Epithelial Cell Size: Size of a single cell that forms tissues that lies outside of the body and the passageways that lead to or from the surface*	*Ordinal*	*1-10*
BN	*Bare Nuclei*	*Ordinal*	*1-10*
BC	*Bland Chromatin: Evaluates for the presence of bare bodies*	*Ordinal*	*1-10*
NN	*Normal Nucleoli*	*Ordinal*	*1-10*
M	*Mitosis*	*Ordinal*	*1-10*
Dia	*Diagnosis of tumours*	*Ordinal*	*1-10*

B: Data Processing Stages of Breast Cancer Prediction Tool

The tool predicts whether a patient may be having benign or malignant type of breast cancer based on symptoms. Figures 2a and 2b (Stages A-U) summarize the data processing stages of the breast cancer prediction tool. It is similar to an expert doctor who asks questions to identify symptoms on the patient. Note that stages A-D constitute a data collection process. In stages A1-A10, the tool asks a patient if he/she has experienced any of the symptoms and the responses are collected and saved in the database. Thereafter, the tool classifies and analyzes the symptoms collected from the patient. The tool may ask a patient to take additional or necessary tests based on adequacy of the patient's symptoms. Then the tool selects the patient's data for disease diagnosis.

As illustrated in stages E-F (Fig.2b), the collected data needs to be cleaned or scrubbed by pre-processing in order to remove dirty data, a requirement before doing any data mining task. It is the hardest and very important part in data mining tasks to achieve quality data. Data pre-processing is done by fitting in missing values, removing duplicate values, detecting and removing outliers and extreme values and smoothing noisy data. If data pre-processing is not done correctly, the results from data mining might be unreliable and also may affect decision making in the prediction process.

Sometimes databases can be noisy and therefore irrelevant. That is, some values of attributes in the database can be invalid and attributes present in the database may not be required to perform DM tasks. These invalid attributes if not removed may cause difficulties in discovering relevant information out of selected database.

After obtaining quality data, the same has to be divided into training and testing sets. As a general rule, the data sets were randomly divided such that 75% of the data was for training and 25% for testing. This achieved the independence and smooth generation of samples without any bias of the data sets. The training set was used to create the model, which is already known. After the model was created, it was used to make prediction against the test set, which is also unknown.

Feature or attribute selection is used to improve the accuracy of the model and to have a better understanding of its structure. Hence, this method adds new attributes into existing system. It is also helpful to discover missing information about the relationships among the data attributes that may be useful for knowledge discovery. In feature selection, we used Chi Square, information gain, gain ratio and average rank method to get rid of irrelevant features. Then data reduction was used to obtain a sized data set hence avoid using massive data sets. If data is large, it will slow down the DM process. The main advantage of this technique is that even after reduction of data; integrity of original data is still maintained.

Table 2 shows the feature selection results of the breast cancer attributes. From the table, it can be seen that the best attribute selected was Bare Nuclei, followed by uniformity of cell shape and the least attribute selected is mitoses, followed by the single epithelial cell size. Figure 3 shows the data of Table 2 presented graphically. It is observed from the graph that the most significant change in the graph (the significant slope point) highlights the first eight ranking features located above the slope point. The graph shows the biggest drop just after feature number 8; single epithelial cell size.

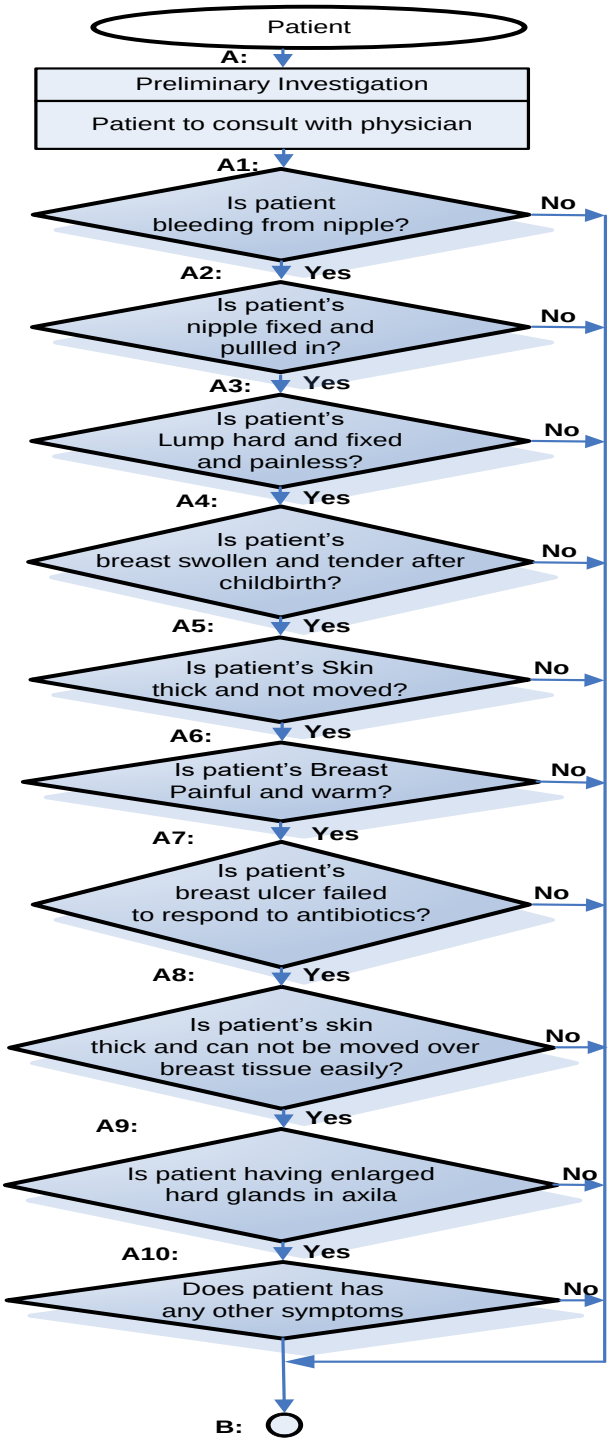

Figure 2(a). Preliminary investigation.

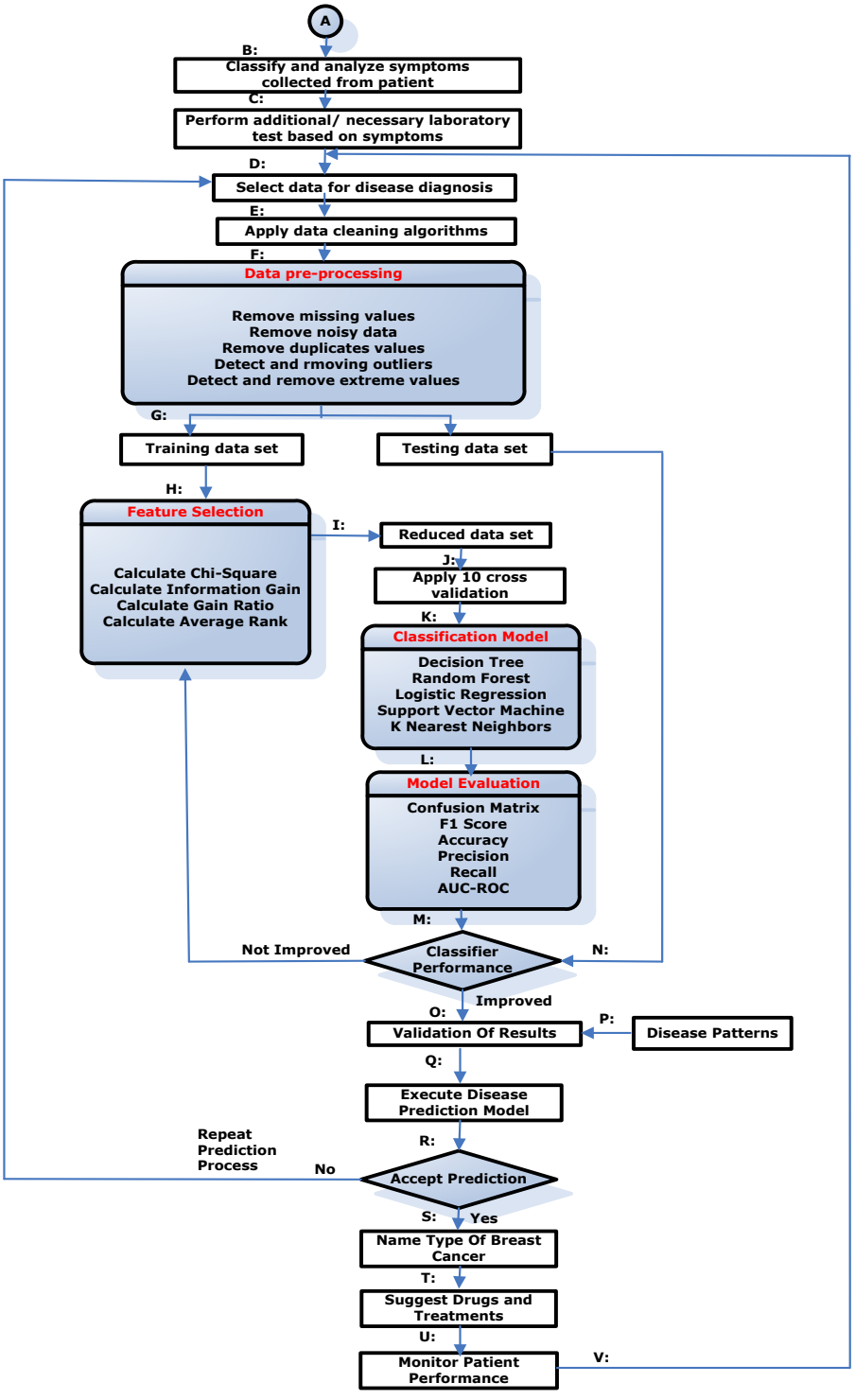

Figure 2(b). Flow chart of proposed breast cancer prediction tool.

Table 2. Feature Selection results for Wisconsin breast cancer dataset.

Variable	Chi-Square	Info Gain	Gain Ratio	Average Rank	Importance
Clump Thickness	315.52	0.405	0.283	105.4023333	5
Cell Size Uniformity	374.317	0.467	0.436	125.073333	3
Cell Shape Uniformity	375.317	0.467	0.436	125.629333	2
Marginal Adhesion	247.246	0.297	0.295	82.6126667	7
Single Epithelial Cell Size	236.134	0.277	0.295	78.902	8
Bare Nuclei	377.087	0.465	0.423	125.991667	1
Bland Chromatin	324.629	0.398	0.381	108.469333	4
Normal Nucleoli	268.28	0.319	0.324	89.641	6
Mitosis	57.204	0.067	0.18	19.1503333	9

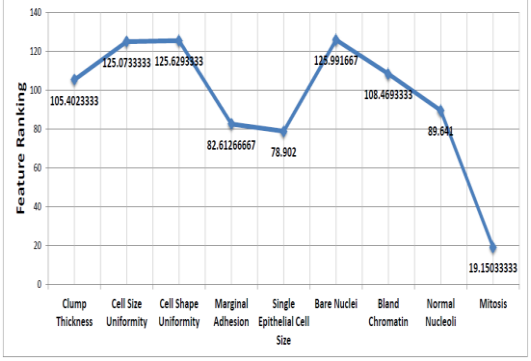

Figure 3. Graph showing the steep slope of change of level of feature significance on the Wisconsin breast cancer dataset.

4. CLASSIFIER TRAINING AND PERFORMANCE EVALUATION

4.1 *Classifier Training*

In this study, supervised learning algorithms were selected due to their popularity and that they are used by many researchers in similar kind of problems. Choosing which algorithms to use is a critical stage. Choice of algorithms depends on the kind of data sets, numbers of features in the data sets, complexity of data sets and different performance metrics (accuracy, sensitivity, specificity, recall, precision and ROC curve) that need to be tested. According to "NO FREE LUNCH THEOREM" [9], there is no single algorithm or model that is best suited for every problem. The assumption of one algorithm works best for one problem does not necessarily mean that the same algorithm will work best for another problem. Therefore, it is common in machine learning to try different algorithms for a particular problem and find out which one works best for that problem.

4.1.1 *Support Vector Machine (SVM)*

Lian *et al*. [10] remark that support vector machine (SVM) is basic yet an intelligent classifier used in machine learning. It has received a considerable attention due to its high performance for training and testing the model. SVM is used for classifying linear as well as non-linear transformation methods (see Fig. 4). Thereafter, the algorithm searches for the best hyper plane to group the transformed data into two different classes. Specifically, the SVM performs the classification process by maximizing the margin of the hyper plane separating the two classes while minimizing the classification errors.

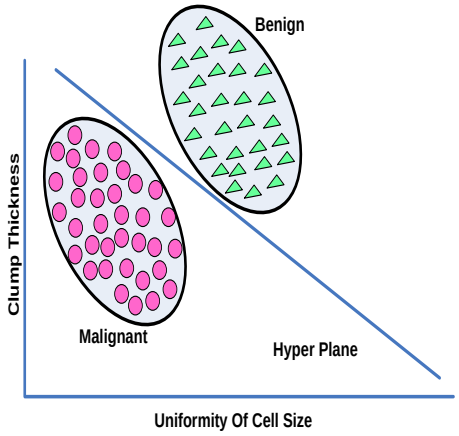

Figure 4. Example of SVM classification.

4.1.2 *K-Nearest Neighbor (KNN)*

Jasmina *et al*. [11], point out that KNN classification algorithm is a non parametric lazy learning classifier. Initially, the KNN classifier stores all available instances and classifies new instances based on similarity functions. Figure 5 illustrates how the KNN classifier works. Essentially, the KNN finds the K training instances that are close to the unseen instances using distance measures such as Euclidean, Manhattan, maximum dimension distance, and others. Then, the algorithm predictively decides on the class for the unseen instance by taking the most commonly occurring class in the nearest K instances.

4.1.3 *Decision Tree (DT)*

The DT classifier is a flow chart like tree structure, where each internal node represents a test on an attribute, each branch represents an outcome of the test, class label is represented by each leaf node (or terminal node) [12]. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 and CART. As its name implies, this classification technique works by recursively constructing branches of the trees based on certain observations or variables. In a DT, the root and each internal node are labeled with a question (See Fig. 6). It is usually built by top down or divide-and-conquer approach. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration. On the other hand, a DT is a classifer with a structure of supervised learning techniques for classification and regression towards creating a model fro predicting the value of the target variable.

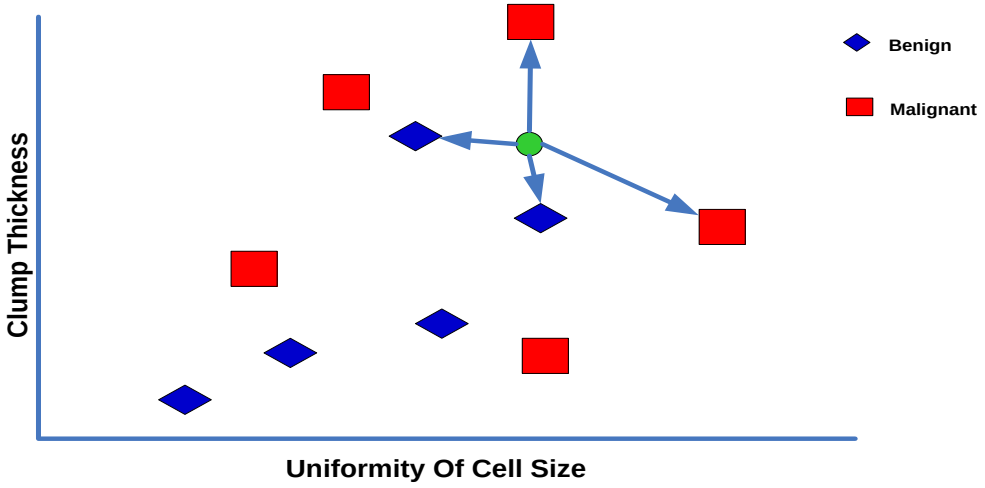

Figure 5. Example of KNN classification.

4.1.4 Logistic Regression (LR)

The LR classifier is a linear method, but the predictions are transformed using the logistic functions [13]. A logistic function is also called sigmoid function, which was developed by statisticians for describing properties of population growth in ecology. Adapted for DM applications, the LR is an S-shaped curve which can take any real valued number and map it into a value between 0 and 1, but never exactly at those limits. The LR uses logit transform to directly predict probabilities that are in turn used to determine the impact of knowln multiple independent variables (presented to the classifer simultaneously) on the process of predicting one or two other dependent variables. Similarly, the LR does not assume a linear relationship between the set of dependent and independent variables used in the prediction process (See Fig. 7), as reported in [14].

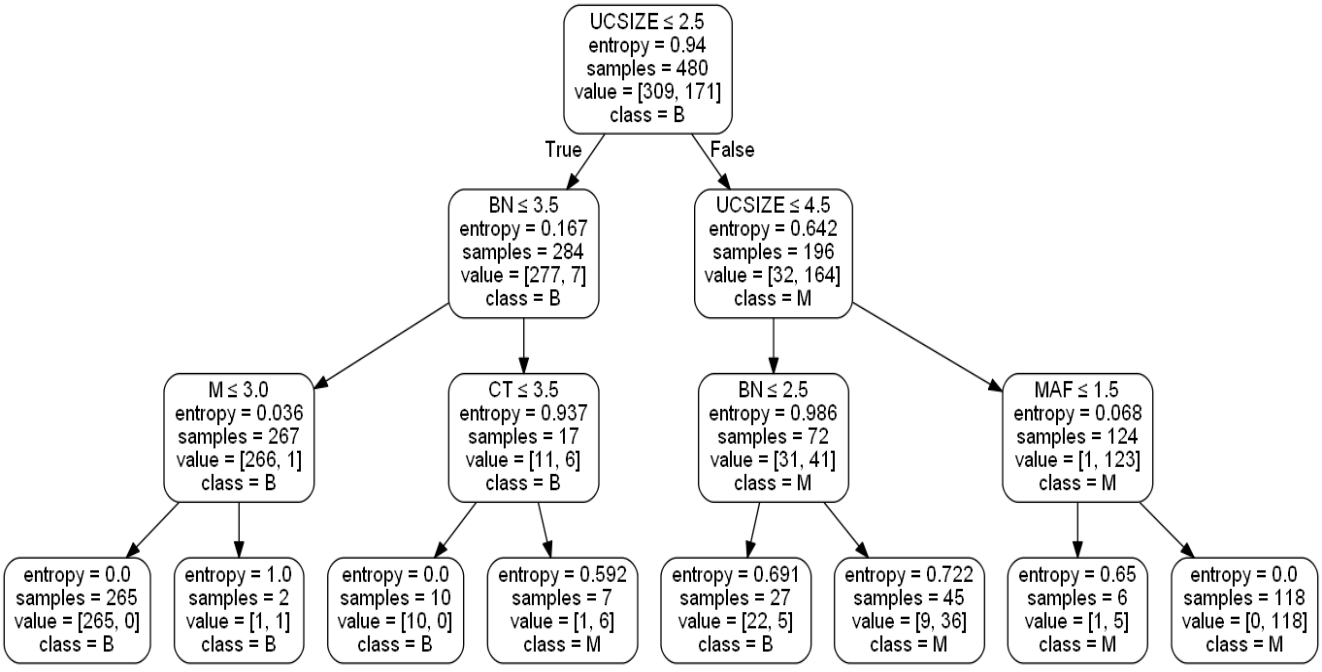

Figure 6. Example. of DT classification.

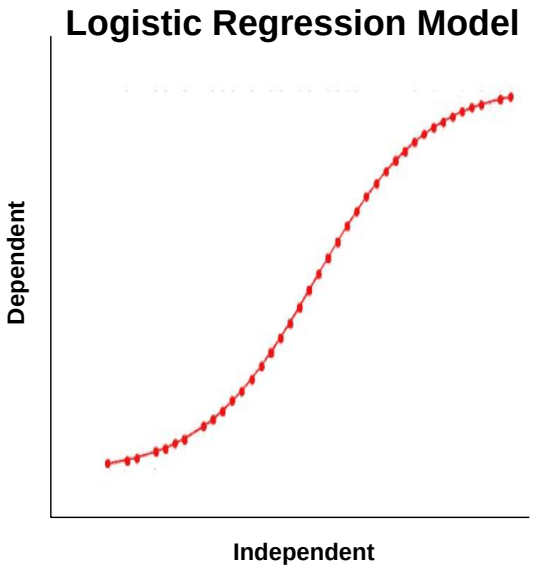

Figure 7. Example of LR classification.

4.1.5 Random Forest

The RF classifier is a family of decision tree classifiers [12] consisting of a collection of tree structured classifiers where independent random vectors are distributed identically and each tree cast a unit vote for the most popular class [15]. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In [13], RF consists of a collection of tree like structure classifiers that is an improvement of bagged ensemble learning decision tree classifier. RF changes the algorithm in a way that the sub-trees are learned so that the resulting predictions from all of the sub-trees have less correlation.

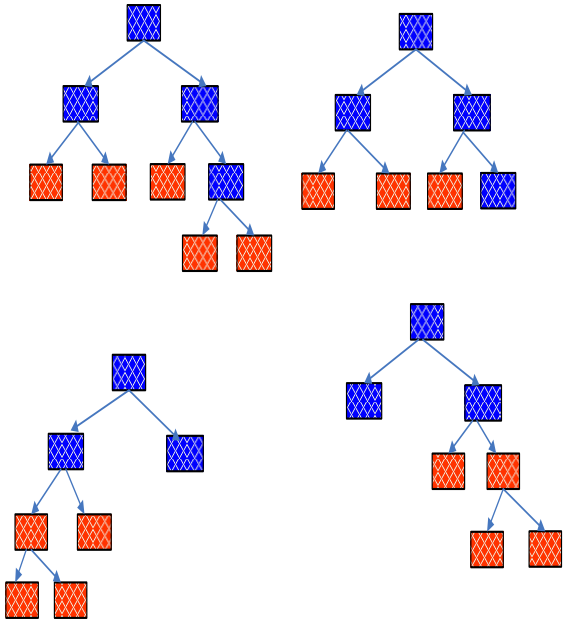

Figure 8. Example of RF classification.

4.2 PERFORMANCE EVALUATION METRICS

Evaluation of the performance of the breast cancer model is based on the count of test data correctly and incorrectly predicted by the model. These counts are tabulated in Table 3 also known as confusion matrix. Each record (d_{11}, d_{10}, d_{01} and d_{00}) represents the number of instances either correctly predicted or incorrectly predicted. For instance, class d_{11} is actual benign and predicted to be benign but in the case of class d_{10}, it is actual benign but predicted to be Malignant, therefore, it is wrongly predicted.

Table 3. Example of confusion matrix.

		Predicted Class	
		Benign	**Malignant**
Actual Class	**Benign**	d_{11}	d_{10}
	Malignant	d_{01}	d_{00}

4.2.1 Accuracy

Accuracy is one of the metrics for measuring the performance of any prediction model. It is the proportion of the total number of correct predictions (See Table 3). The accuracy of the model is calculated as:

$$Accuracy = \frac{Number\ of\ instances\ correctly\ pr}{Total\ number\ of\ instance} = \frac{d11+d00}{d11+d10+d01+d00} \quad \text{..... Eqn. (1)}$$

$$Error\ rate = \frac{Number\ of\ wrong\ predictions}{Total\ number\ of\ predictions} = \frac{d10+d01}{d11+d10+d01+d00} \quad \text{..... Eqn. (2)}$$

4.2.2 Precision

Table 4. Example of confusion matrix for precision and recall.

		Predicted Class	
		+	**-**
Actual Class	**+**	**d ++ (TP)**	**d + - (FN)**
	-	**d -+ (FP)**	**d - - (TN)**

Precision (Positive Predictive Value) is the proportion of positive cases that were correctly identified. The higher the precision the lower the number of positive errors committed by classifiers (See Table 4). Precision of the prediction model is calculated as:

$$Precision = \frac{TP}{TP+FP} \quad \text{.....Eqn. (3)}$$

Where:

- True Positive (TP) or d++ is to the number of positive instances correctly predicted by the classifier.
- False Positive (FP) or d -+ is the number of negative instances wrongly predicted by the classifier.

4.2.3 Recall

Recall, also known as sensitivity, represents the true positive rate of the considered class. The higher the recall the lower the number of positive examples committed by classifiers. Precision of the prediction model is calculated as:

$$Recall = \frac{TP}{TP+FN} \quad \text{..... Eqn. (4)}$$

Where:

False Negative (FN) or d +- is the number of positive instances wrongly predicted as negative instances by the classifier.

5.0 DATA VISUALIZATION

Table 5 and Fig. 9 show the correlation between an two of the selected attributes (features). Correlation refers to the relationship between two variables and how they may or may not change together. Pearson's Correlation Coefficient, which assumes a normal distribution of the attributes involved, was used to calculate the correlation between any two of the attributes. A value of -1 denotes a full negative correlation, a value of 1 denotes a full positive correlation, and a value of 0 denotes there is no correlation at all between two attributes. Some of the machine learning algorithms suffer from poor performances if the attributes are highly correlated. For the breast cancer data sets, the attributes between cell shape and cell size are highly correlated.

Table 5. Correlation between attributes for the Wisconsin breast cancer dataset.

	clump	cell size	cell shape	adhesion	epithelial	nuclei	chromatin	Nucleoli	Mitosis	Class
clump	1.000	0.644	0.655	0.487	0.523	0.592	0.556	0.535	0.351	0.716
cell size	0.644	1.000	0.908	0.705	0.752	0.692	0.756	0.721	0.459	0.820
cell shape	0.655	0.908	1.000	0.684	0.720	0.714	0.736	0.719	0.440	0.821
adhesion	0.487	0.705	0.684	1.000	0.597	0.669	0.667	0.600	0.419	0.703
epithelial	0.523	0.752	0.720	0.597	1.000	0.584	0.617	0.627	0.479	0.686
nuclei	0.592	0.692	0.714	0.669	0.584	1.000	0.681	0.584	0.338	0.821
chromatin	0.556	0.756	0.736	0.667	0.617	0.681	1.000	0.665	0.346	0.759
nucleoli	0.535	0.721	0.719	0.600	0.627	0.584	0.665	1.000	0.429	0.716
mitosis	0.351	0.459	0.440	0.419	0.479	0.338	0.346	0.429	1.000	0.423
class	0.716	0.820	0.821	0.703	0.686	0.821	0.759	0.716	0.423	1.000

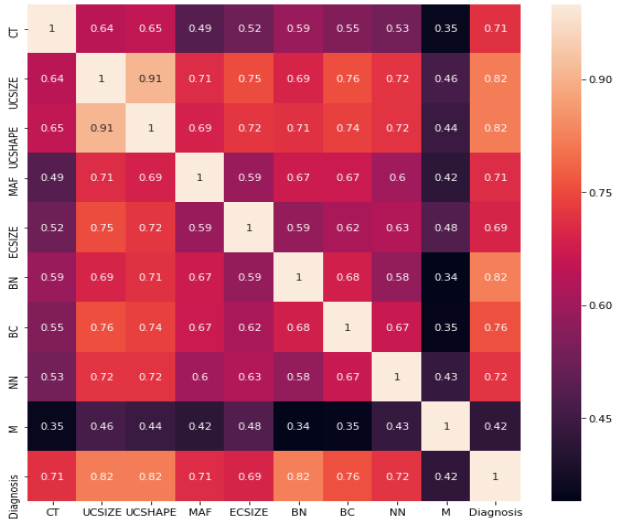

Figure 9. Matrix showing the correlation between attributes of the Wisconsin breast cancer dataset.

6.0 EXPERIMENTAL RESULTS

The selected five learning DM algorithms (Classifiers) were tested twice. First: in a learning experiment and second – in a prediction experiment. Conforming to best DM best practice, the experimental data set was divided randomly into 75% subset for training and 25% subset for testing (see Section 3B).

6.1 Learning experiment using training data set

Figure 10 shows the classifier training performance on the breast cancer data set. After investigating the performance of the selected classifiers, Decision Tree was found to perform better than the other classifiers with classification accuracy of 100%; followed by Random Forest, Logistic Regression, Support Vector Machine and K Nearest Neighbors with accuracies of 99.8%, 97.4%, 97.4% and 97.1%, respectively; all of which are much better than those reported in the literature ([3] –[7]). These references do not report any results on testing their classifiers on unknown cases.

Figure 10. Classifier training performance.

6.2 Prediction experiment using test data set

The test data used had 171 *actual* breast cancer instances, out of which 107 instances were *actual benign* cases and 64 instances were *actual malignant* cases. Figure 11-14 show the results of testing the performance of the selected classifiers on the data sets using confusion matrix of Table 3.

For example, Fig. 11A shows the prediction results for the case of using the LR classifier. It can be observed that the classifier correctly predicted 103 instances as benign (out of the 107 actual cases) and classified 4 instances incorrectly). It can also be observed that the classifier correctly predicted 55 instances as malignant (out of the 64 actual instances), and classified 9 instances incorrectly. These results, as per Eqn. (1) gives an overall prediction accuracy of 92.4%. Similary for cases 11B through 11E, the corresponding classifiers performed with accuracies of: 93.6%(DT), 92.9%(RF), 92.4%(SVM) and 92.4%(KNN). As expected from the classifier training results, the DT classifier outperformed the others with an accuracy of 93.6%; which is a new contribution into the literature.

Figure 12 shows the results for the five different evaluation metrics, namely; accuracy, precision, recall, F1-Score and Support. These results are graphically presented in Fig.13. After investigating the performance of the selected classifiers, DT was found to perform better than other classifiers, with classification accuracy of 93.5%, followed by 92.9%(RF), 92.4%(LR), 92.4%(SVM) and 92.4%(KNN).

Figure 14 shows the receiver operating characteristic (ROC) curve for three best classifiers. The ROC curve reveals the tradeoff between true positive rate (TPR) and false positive rate(FPR) for classifiers. In the ROC curve, the TPR is plotted along y axis and FPR along x-axis. A good classification model should be located as close as possible to the upper left corner of the graph (0,0 to 0,1 vertically and 1,1 horizontally), while a model that predicts using random guesses is expected to reside along the main diagonal. Therefore, in this case as shown in the graphs, the RF classifier is a better model outperforming the other classifiers.

In the case of Area Under the Curve (AUC), RF performs better at 99.9% accuracy, followed by LR and DT with 98.9% and 95.8% accuracies, respectively.

6.3 Prediction Logical Values

Figure 15 shows the prediction of breast cancer from the 25% of the 683 instances as test data (171 instances) using the five different classifiers. The results show both the actual value and predicted value for each patient: *1 means the patient has cancer and zero means patient does not have cancer*. Observable from the figures is that there are cases where some actual values show patient has cancer but the corresponding predicted values show the patient has no cancer. These are the erroneous cases. Also notably, Random Forest and Logistic Regression has less prediction errors. It means the classifiers have fewer incorrectly predicted instances compared to Support Vector Machine, Decision Tree and K Nearest Neighbors.

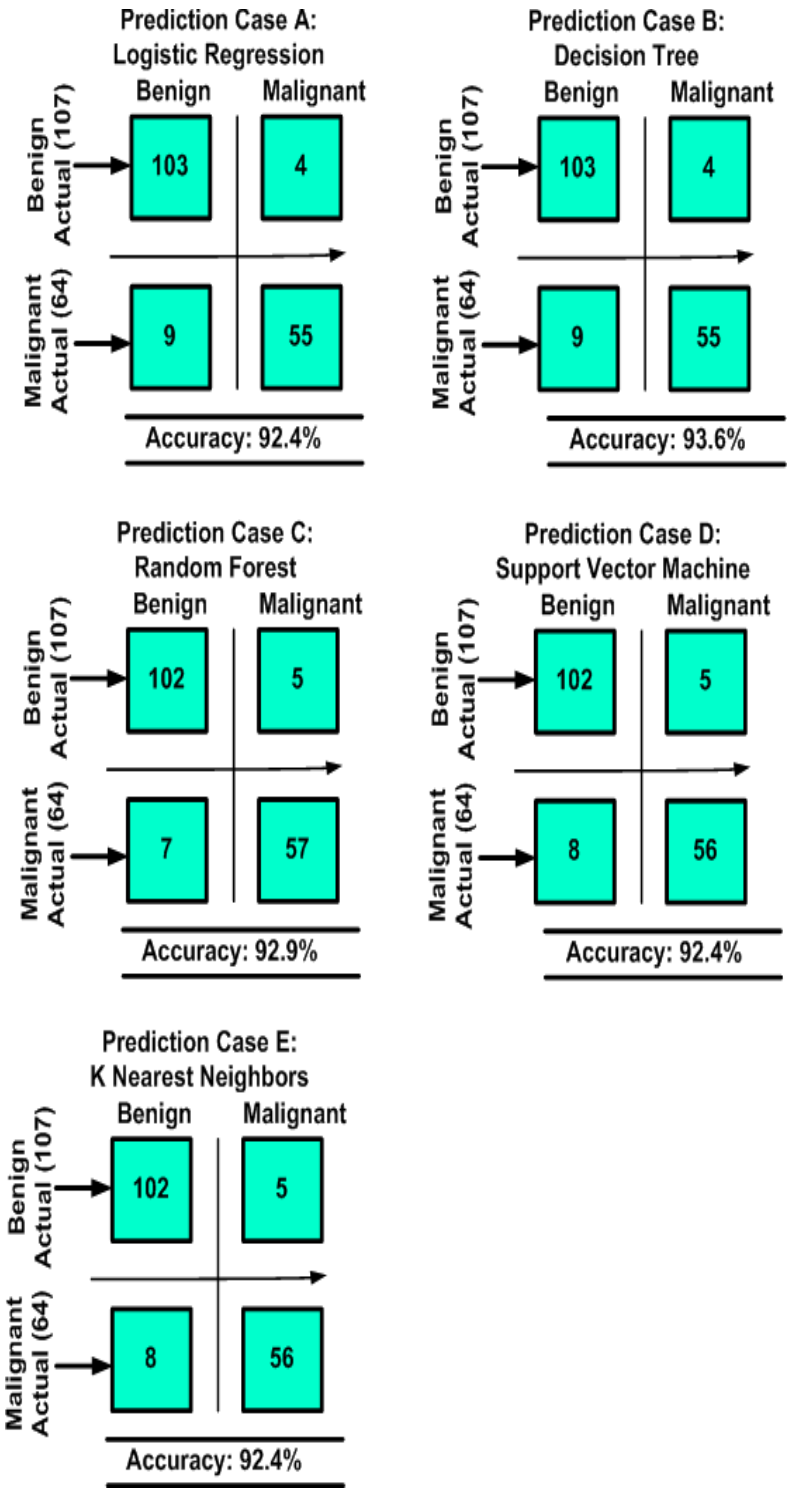

Figure11. Confusion matrix for different classifiers

Prediction Case A: Logistic Regression					
	Model	Precision	Recall	F1-Score	Support
	Benign	94%	95%	95%	107
	Malignant	92%	91%	91%	64
Accuracy				94%	171
Macro Avg		93%	93%	93%	171
Weighted Avg		94%	94%	94%	171

Prediction Case B: Decision Tree					
	Model	Precision	Recall	F1-Score	Support
	Benign	94%	95%	94%	107
	Malignant	92%	91%	91%	64
Accuracy				94%	171
Macro Avg		93%	93%	93%	171
Weighted Avg		94%	94%	94%	171

Prediction Case C: Random Forest					
	Model	Precision	Recall	F1-Score	Support
	Benign	94%	95%	94%	107
	Malignant	92%	89%	90%	64
Accuracy				93%	171
Macro Avg		93%	92%	92%	171
Weighted Avg		93%	93%	93%	171

Prediction Case D: Support Vector Machine					
	Model	Precision	Recall	F1-Score	Support
	Benign	93%	95%	94%	107
	Malignant	92%	88%	90%	64
Accuracy				92%	171
Macro Avg		92%	91%	92%	171
Weighted Avg		92%	92%	92%	171

Prediction Case E: K Nearest Neighbors					
	Model	Precision	Recall	F1-Score	Support
	Benign	93%	95%	94%	107
	Malignant	92%	88%	90%	64
Accuracy				92%	171
Macro Avg		92%	91%	92%	171
Weighted Avg		92%	92%	92%	171

Figure 12. Performance metrics for different classifiers.

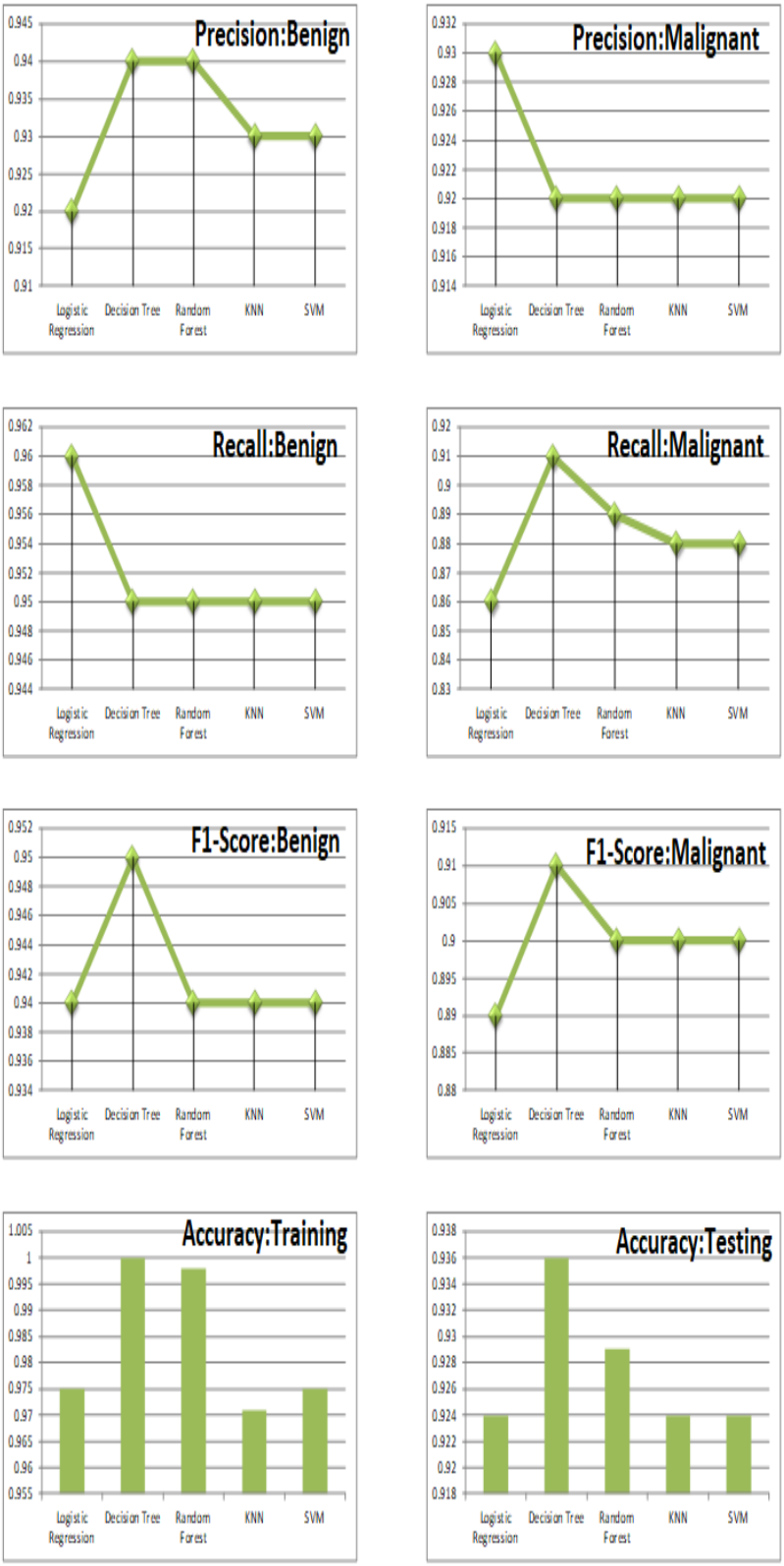

Figure 13. Comparative graphs of different evaluation metrics.

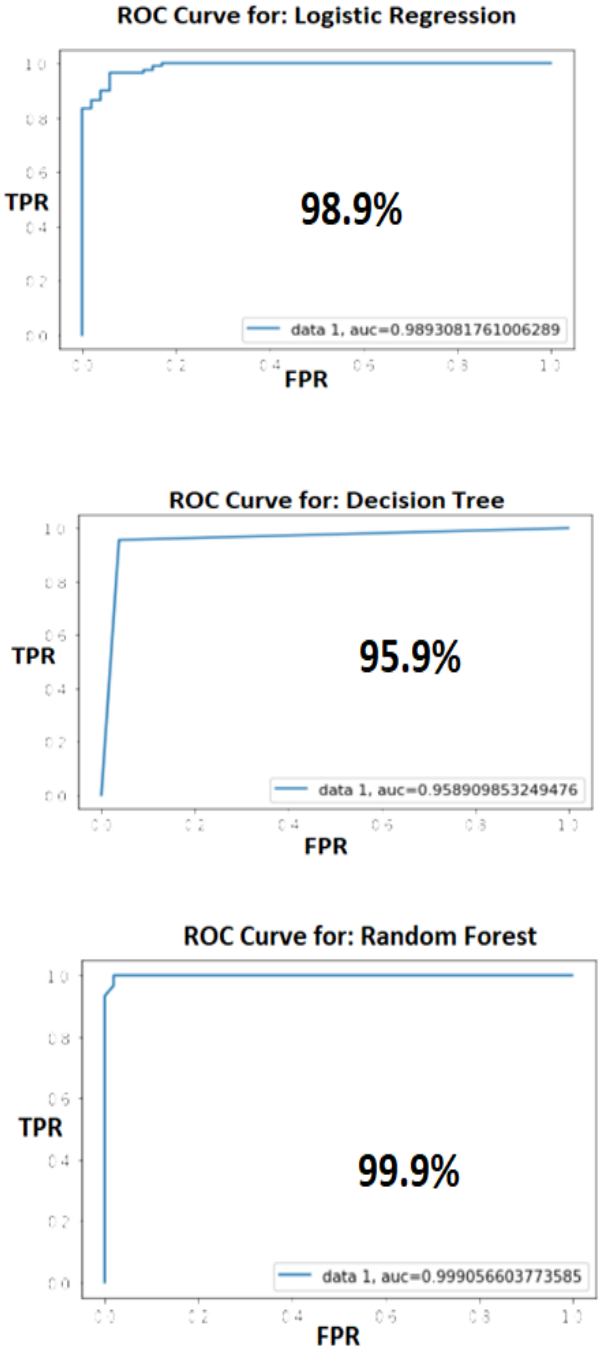

Figure 14. ROC curveS for best classifiers.

Figure 15. Cases (A) to (E) of actual prediction values.

7. CONCLUSSION

Disease prediction and its treatment methods is a major concern that needs much attention these days. This paper supports the fact that machine learning can be of big help when it comes to medical diagnosis and prognosis. Presented in this paper is an approach for prediction and detection of breast cancer using machine learning techniques. The presented tool can assist physician either new or experienced in diagnosis and prognosis of breast cancer at benign and malignant stages. Training and test results for breast cancer data sets are reported. Achieved from the reported work of training sets are classification accuracies of 100% (Decision Tree); 99.8046% (Random Forest); 97.46% (Logistic Regression and Support Vector Machine);97.07% (K Nearest Neighbors) and for testing sets are classification accuracies of 93.5672% (Decision Tree); 92.9% (Random Forest); 92.39%(Logistic Regression, Support Vector Machine and K Nearest Neighbors). These results are much better than those reported in the literature. The results show that the proposed DM disease prediction tool has potential to greatly impact on current patient management, care and future interventions against the breast cancer disease and through customization even against other deadly diseases.

REFERENCES

[1] Siri Krishan Wasan ,Vasudha Bhatnagar and Harleen Kaur., "The Impact of Data mining Techniques on Medical Diagnostics". Data Science Journal, Volume 5, 19 October 2006, 119.

[2] Tanzania Breast Health Care Assessment., " An Assessment Of Breast Cancer Early Detection". Diagnosis and Treatment in Tanzania, 2017. Accessed4th December 2019, (https://ww5.komen.org/uploadedFiles/_Komen/Content/Grants_Central/International_Grants/Grantee_Resources/Full_Tanzania_Assessment_report.pdf)

[3] Mamta Jadhav, Zeel Thakar, Pramila Chawan., "Breast Cancer Prediction Using Suppervised Machine Learning Algorithms". International Research Journal Of Engineering and Technology (IRJET) Vol. 6, No. 10, Oct 2019.

[4] Salim A. Diwani and Zaipuna O. Yonah., "A Novel Holistic Disease Prediction Tool Using Best Fit Data Mining Techniques". International Journal of Computing and Digital Systems, ISSN (2210-142X), Int. J. Com. Dig. Sys. 6, No.2 (Mar-2017).

[5] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction Of Breast Cancer Using Supervised Machine Learning Techniques". International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol.8, No.6, April 2019.

[6] Sivapriya J, Aravind Kumar, Siddarth Sai, Sriram S "Breast Cancer Prediction Using Machine Learning". International Journal Of Recent Technology and Engineering (IJRTE) , Volume 8, No. 4, November 2019.

[7] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis". The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016)

[8] Wisconsin Diagnostic Breast Cancer (WDBC) Dataset and Wisconsin Prognostic Breast Cancer (WPBC) Dataset. http://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/ (Accessed March 2020)

[9] Adam S.P., Alexandropoulos SA.N., Pardalos P.M., Vrahatis M.N. (2019) No Free Lunch Theorem: A Review. In: Demetriou I., Pardalos P. (eds) Approximation and Optimization. Springer Optimization and Its Applications, vol 145. Springer, Cham.

[10] Meng-Jia Lian , Chih-Ling Huang , Tzer-Min Lee, "Construction of an Oral Cancer Auto Classfy System Based on Machine Learning for Artificial Intelligence" Computer Science and Information Technology (CS&IT) Computer Science Conference Proceedings (CSCP), pp 33-39, 2019.

[11] Jasmina Novacovic, Alompije Veljovic, Sinisa Ilic, Milos Papic., "Experimental Study of Using K-Nearest Neighbor Classifier with Filter Methods". Conference Computer Science and Technology, Varna Bulgaria, June 2016.

[12] Himani Sharma , Sunil Kumar., "A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research (IJSR), April 2016.

[13] Jason Brownlee, "Master Machine Learning Algorithms" Jason Brownlee, 2016, http://MachineLearningMastery.com

[14] David W. Hosmer, Stanley Lemeshow (2005). "Applied Logistic Regression", Second Edition. New York, N.Y. : Wiley, 2000. Wiley series in probability and statistics Texts and references section.

[15] Himani Sharma , Sunil Kumar., "A Survey on Decision Tree Algorithms of Classification in Data Mining". International Journal of Science and Research (IJSR), April 2016.

Salim Amour Diwani received his BS and M.Sc. degrees in computer science from Jamia Hamdard University, New Delhi, India in 2006 and in 2008, respectively. He is currently a PhD scholar in Information communication Science and Engineering at Nelson Mandela African Institution of Science and Technology in Arusha, Tanzania. His primary research interests are in the areas of Big Data, Data Mining, Machine Learning and Database Management Systems.

Zaipuna O. Yonah -MIET, SMIEEE - holds a B.Sc. degree (with Honors - 1985) in Electrical Engineering from University of Dar es Salaam - Tanzania; and M.Sc. (1986) and PhD (1994) Degrees in Computer-based Instrumentation and Control Engineering from the University of Saskatchewan, Saskatoon-Canada. In Tanzania, he is a Registered Consulting Engineer in ICTs.

Yonah has over 35 years of practice. His work spans the academia, industry and policy making fields. He is currently associated with The Nelson Mandela Institute of Science and Technology, Applied Engineering & ByteWorks (T) Ltd and the Institute of Electrical and Electronics (IEEE) Inc. He is one of the mentors and pioneers driving the national broadband agenda in Tanzania and EAC, SADC regions. He believes that ICTs, as tools for development, promise so much: *interactivity*, *permanent availability*, *global reach*, *reduced per unit transaction costs*, *creates increased productivity and value*, *jobs and wealth*, *multiple source of information and knowledge*. Armed with such a belief, his current work aims at creating and delivering value through ICT-enabled services in the shortest times possible. His research interests include: ICT4D, Mobile and Web applications, Big Data, Data Mining, High Performance Computing, high-capacity Broadband networks, Intelligent Instrumentation and Control Engineering; ICT enabled 21st Century Education delivery (ICT4E): *Personalized, Facilitated, and Connected Learning* and Poverty Alleviation Initiatives – ICT4PA