ISSN No: - 2456 – 2165

Evaluation of Predictive Ability of Some Data Mining and Statistical Techniques Using Breast Cancer Dataset

H.G. Dikko, Y. Musa, H.B. Kware.
Ahmadu Bello University, Zaria, Kaduna State, Nigeria.
Dept. of Mathematics, UsmanuDanfodiyo University, Sokto, Nigeria
UsmanuDanfodiyo University, Sokoto, Nigeria

Abstract:-There is no single best algorithm since it highly depends on the data any one is working with. Nobody can tell what should use without knowing the data and even then it would be just a guess. This research work focuses on finding the right algorithm that works better on breast cancer data sets. The aim of this study is to perform a comparison experiment between statistical and data mining modeling techniques. These techniques are Data mining Decision Tree (C4.5), Neural Network (MLP), Support vector machine (SMO) and statistical Logistic Regression. The comparison will evaluate the performance of these prediction techniques in terms of measuring the overall prediction accuracy for each technique on the bases of two methods (cross validation and percentage split). Experimental comparison was performed by considering the breast cancer dataset and analyzing them using data mining open source WEKA tool. However, we found out that a C4.5 and MLP algorithm has a much better performance than the other two techniques.

Keywords:-Breast Cancer Survivability, Multi-Layer Perception, Logistic Regression, Data Mining.

I. INTRODUCTION

Data mining (DM) is also popularly known as Knowledge Discovery in Database (KDD). DM, frequently treated as synonymous to KDD, is actually a part of knowledge discovery process and is the process of extracting information including hidden patterns, trends and relationships between variables from a large database in order to make the information understandable and meaningful and then use the information to apply the detected patterns to new subsets of data and make crucial business decisions. The ultimate goal of data mining is prediction. Predicting the outcome of a disease is one of the most interesting and challenging tasks in data mining applications [2].

Data mining is becoming an increasingly important tool to transform these data into information. Data mining can also be referred as knowledge mining or knowledge discovery from data. Many techniques are used in data mining to extract patterns from large amount of database [3]. Classification and Association are the popular techniques used to predict user interest and relationship between those data items, which has been used by users association, preprocessing, transformation, clustering, and pattern evaluation.

Classification and Association are the popular techniques used to predict user interest and relationship between those data items, which has been used by users. Statistical methods alone, on the other hand, might be described as being characterized by the ability to only handle data sets that are small and clean, which permit straightforward answers via intensive analysis of single data sets. Literature shows that a variety of statistical methods and heuristics have been used in the past for the classification task. Decision science literature also shows that numerous data mining techniques have been used to classify and predict data; data mining techniques have been used primarily for pattern recognition purposes in large volumes of data [2].

This research paper aims to analyze the several data mining techniques proposed in recent years for the prediction of breast cancer survivability. Many researchers used data mining techniques in the diagnosis of diseases such as tuberculosis, diabetes, cancer and heart disease in which several data mining techniques are used in the prediction of cancer disease such as KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree, Genetic Algorithm, Naïve Bayes, Decision tree, WAC which are showing accuracy at different levels.

Automated breast cancer prediction can benefit healthcare sector. This automation will save not only cost but also time. This paper presents different data mining techniques, which are deployed in these automated systems. Various data mining techniques can be helpful for medical analysts for accurate breast cancer prediction.

ISSN No: - 2456 – 2165

II. RELATED WORK

Many studies have been done across countries on data mining. Applications of data mining were used in a large number of fields, especially for business and medical purposes.

Prediction techniques performance comparison issues is an interesting topic for many researchers. A comparative study by Lahiri R. [2] compared the performance of three statistical and data mining techniques on Motor Vehicle Traffic Crash dataset, resulted that the data information content and dependent attribute distribution is the most affecting factor in prediction performance. Delen D. et al. [1] targeted data mining methods comparison as a second objective in the study, while the main objective was to build the most accurate prediction model in a critical field, breast cancer survivability. In the same area, Artificial Intelligence in Medicine Bellaachia A. et al. [3] continued the work of [1] and improved the research tools especially the dataset. An important application area that exploited data mining techniques heavily was the network security. Panda M. et al. [4] also performed a comparative study to identify the best data mining technique in predicting network attacks and intrusion detection. Also the data contents and characteristics revealed as an affecting factor on the data mining and prediction algorithms performance. Vikas C. et al. [5] used a diagnosis system for detecting breast cancer based on Reptree, RBF network and simple logistic. The research demonstrated that the simple logistic can be used for reducing the dimension of feature space and proposed Rep tree and RBF network model can be used to obtain fast automatic diagnostic systems for other diseases.

Data mining concept was the most appropriate to the study of student retention from sophomore to junior year than the classical statistical methods. This was one main objective of the study addressed by [8] in addition to another objective that identifying the most affecting predictors in a dataset. The statistical and data mining methods used were classification tree, multivariate adaptive regression splines (MARS), and neural network. The results showed that transferred hours, residency, and ethnicity are crucial factors to retention, which differs from previous studies that found high school GPA to be the most crucial contributor to retention. In [8]. Research, the neural network outperformed the other two techniques.

[9]compared the prediction accuracy and error rates for the compressive strength of high performance concrete using MLP neural network, Rnd tree models and CRT regression. The results showed that neural network and Rnd tree achieved the higher prediction accuracy rates and Rep tree outperforms neural network regarding prediction error rates. [7].

III. METHODS

A. *Prediction Models*

We used four different types of classification models: Multi-layer perceptron, C4.5, Support vector machine, Logistic regression and compare their performance measures using two different tasting options: k-fold cross validation and percentage split method. These models were selected for inclusions in this study due to their popularity in the recently published literatures. What follows is a brief description of these classification model types.

B. *Multi Layer Perceptron*

Is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs? As its name suggests, it consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. The architecture of this class of networks, besides having the input and the output layers, also have one or more intermediary layers called the hidden layers. The hidden layer does intermediate computation before directing the input to output layer. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable [11].

C. *C4.5*

This algorithm is a successor to ID3 developed by Quinlan Ross in 1993.It is also known as J48 algorithm. It is also based on Hunt's algorithm. It is serially implemented like ID3.Using this algorithm; pruning can take place that is it replaces the internal node with a leaf node thereby reducing the error rate unlike ID3. C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses gain ratio impurity method to evaluate the splitting attribute that is to build the decision tree. It removes the biasness of information gain when there are many outcome values of an attribute [12].

D. *Logistic Regression*

Logistic regression refers to methods for describing the relationship between a categorical response variable and a set of predictor variables [10]. Logistic regression describes a function of mean (which is a probability) as a function of the exploratory variables. The function of mean it uses is the logit function. It assumes that the relationship between the response and the predictor is a non-linear. It produces linear segmentation of classes.

ISSN No: - 2456 – 2165

E. Support Vector Machine

SVMs are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. SVM is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVM represents a learning technique whichfollows principles of statistical learning theory [13]. Generally, the main idea of SVM comes from binary classification, namely to find a hyper plane as a segmentation of the two classes to minimize the classification error. The SVM finds the hyper plane using support vectors (training tuples) and margins (support vectors).

F. Breast Cancer Data Set

In this study, we will use a newer version of Surveillance, Epidemiology, and End Results (SEER) Cancer Incidence Public-Use Database for the (period of 1973 - 2013 with 700,000 records/cases). The preprocessed dataset consist of 343,285 records. The SEER data files were requested through the SEER web site (http://www.seer.cancer.gov). The SEER Program is a part of the Surveillance Research Program (SRP) at the National Cancer Institute (NCI) and is responsible for collecting incidence and survival data from the participating nine registries, and disseminating these datasets (along with descriptive information of the data itself) to institutions and laboratories for the purpose of conducting analytical research. The data set has 15 attributes; we restricted testing to these same attributes and contain the following variables.

Table 1 shows the summary of attributes or predictor variables used in our analysis.

Nominal variable name	**Number of distinct values**
- Race	19
- Primary site code	9
- Marital status	6
- Histologic type	48
- Grade	5
- Behavior code	2
- Extension of tumor	23
- Radiation	9
- Site specific surgery code	19
- Lymph node involvement	10
- Cancer stage	5

Numeric variable name	**Mean**	**Std. Dev.**	**Range**
- Age	61.67	17.24	10 - 130
- Number of positive nodes	28.88	43.67	00 - 50
- Tumor size	33.18	113.99	00 – 200
- Number of nodes	13.21	10.48	00 - 95

Table 1: Predictor Variables for Survival Modeling

G. Distribution of Dependent Variable

We have adopted three fields in the pre-classification process: survival time recode (STR), Vital Status Recode (VSR) and Cause of Death (COD). The STR field ranges from 0 to 180 months in the SEER database [8]. The pre-classification process is outline as follows:

// Setting the survivability dependent variable for 60 month threshold

If STR ≥ 60 months and VSR is alive then the record is pre-classified as “survived”

else if STR < 60 months and COD is breast cancer, then the record is pre-classified as

"not survived"

else

Ignore the record

end if

In the above approach, the ignored records correspond to those patients that have an STR less than 60 months and are still alive, or those patients that have an STR less than 60 months but the cause of their death is not breast cancer [3].

The distribution of the dependent variable is shown in the table 2.

Class	No. Of instances	Percentage %
0:Not survived	37,117	18.1
1: Survived	167,869	81.9
Total	204,986	100.0

Table 2: Survivability Class Instances

After the preprocessing step, a common analysis would be determining the effect of the attributes on the prediction, or attribute selection.

Table 1: Rank Survivability Attribute

The analysis below highlighted the importance of each attribute individually. It shows that attribute Age impacts output the most, and that it showed the best performance in all of the three tests. Then these attributes follow: number of positive nodes, tumor size, extension of tumor, behavior code, lymph node involvement, number of nodes, marital status, histologic type, radiation, site specific surgery, grade, race, primary site code, stage of cancer. Why these prognostic factors are more important predictors than the other is a question that can only be answered by medical professionals and further clinical studies. Figure 1 shows the importance of each attribute.

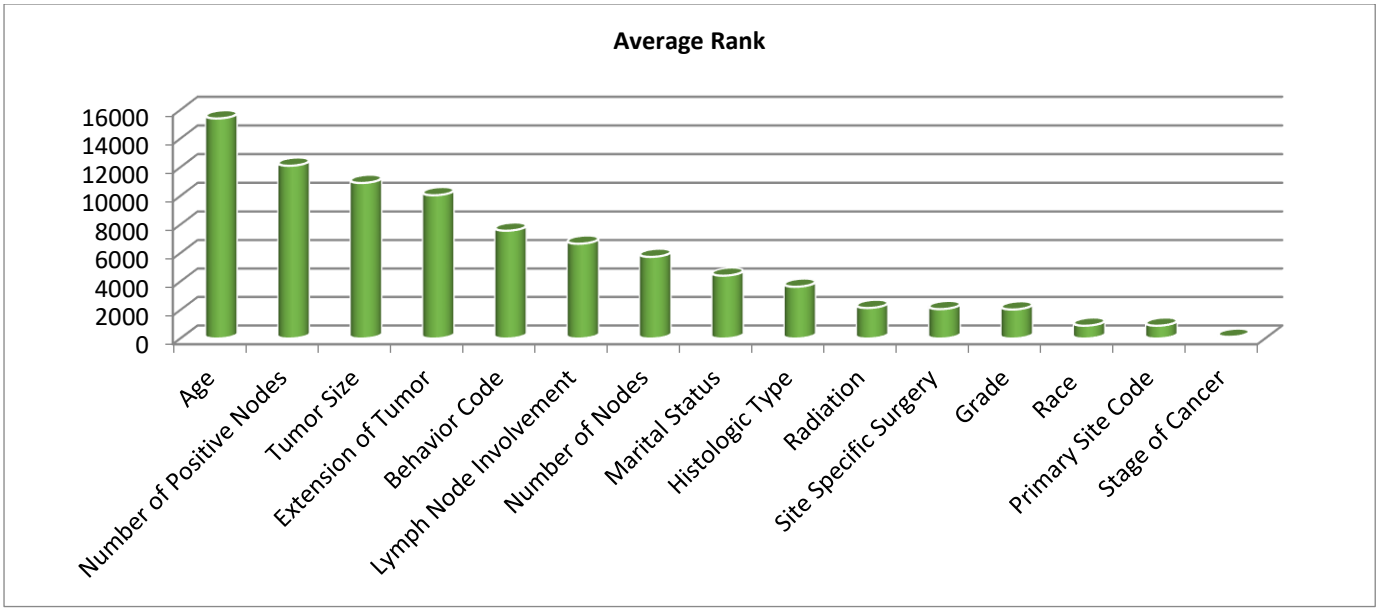

Figure 1: Comparison between Importances's of Attributes.

IV. RESULT AND DISCUSSION

A. *Cross Validation Testing Method*

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In cross validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. The basic form of cross-validation is k-fold cross-validation (kumar*et al*. 2015).

B. *Performance Classifiers*

In this section we have carried out some experiment in order to evaluate the performance of different techniques for predicting breast cancer survivals in order to time and build a model, correctly classified instances versus incorrectly classified instances by algorithms using 10-fold cross validation in the table below.

Evaluation Criteria	**Classifiers**			
	MLP	C4.5	SVM	LR
Timing to build model (in sec)	1275.66	143.1	26548.54	**14.94**
Correctly classified instances	182578	**183020**	177877	179344
Incorrectly classified instances	22408	21966	**27109**	25642

Table 3: Performance of the Classifiers

From the above table we can conclude that C4.5 is more accurate classifier in comparison of others also it can be easily seen that it has highly classified correct instances 183020 and SVM with greatest number of incorrectly classified instances i.e. 27109. SVM 27109 incorrect instances are very high as compare to number of incorrectly classified instances of other three studied algorithms. It is seen that LR takes the shortest time in building the model compared to others and SVM takes a longer time (see figure 2-4).

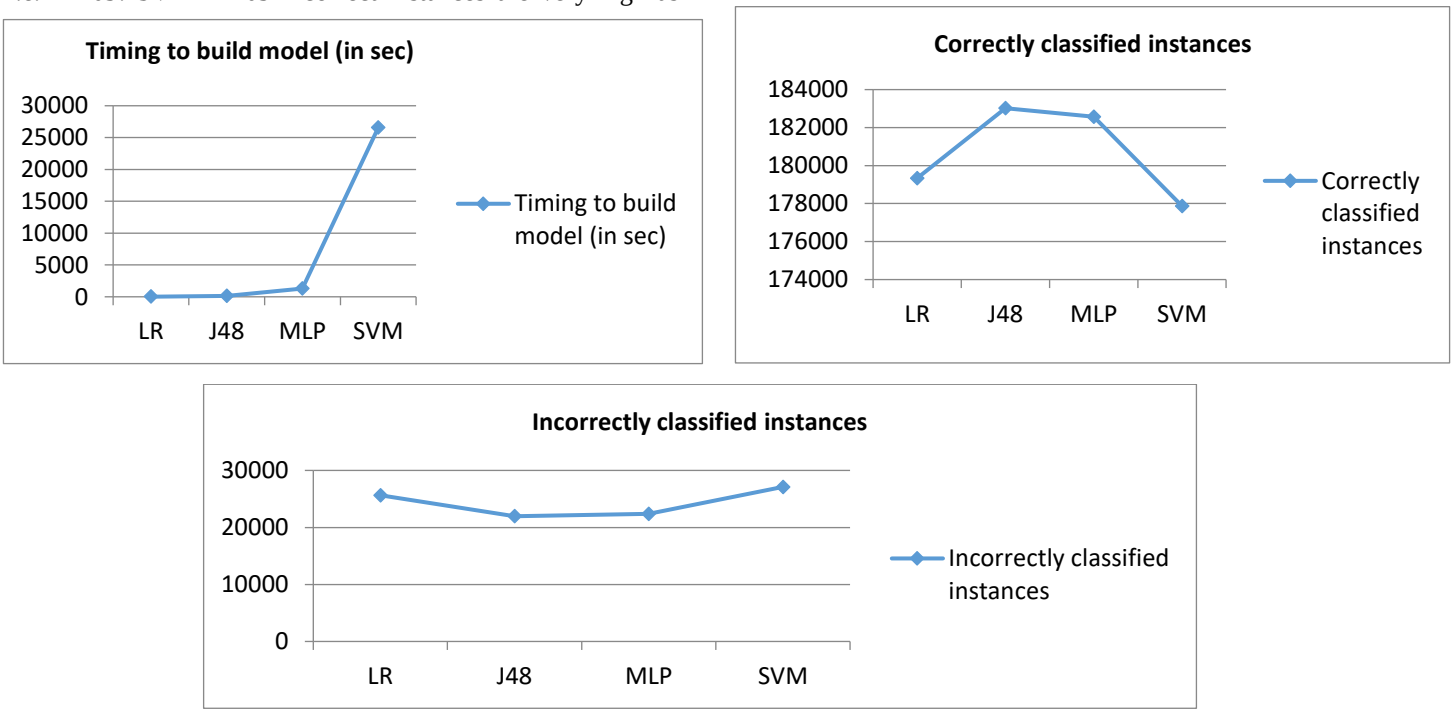

Figure 2: Performance of the Classifiers

ISSN No: - 2456 – 2165

Evaluation Criteria	**Classifiers**			
	MLP	C4.5	SVM	LR
Kappa statistics (KS)	0.5807	**0.593**	0.42	0.4892
Mean absolute error (MAE)	0.1579	0.1619	**0.1322**	0.1907
Root mean square error (RMSE)	**0.2901**	0.2967	0.3637	0.3085
Relative absolute error (RAE)	53.2499 %	54.5871 %	**44.5924 %**	64.3173 %
Root relative squared error (RRSE)	**75.3258 %**	77.0466 %	94.4381 %	80.1085 %

Table 4: Performance Error

The table 4 below shows the values derived for each algorithm based on the performance errors. The kappa statistic value shows that the value of all predictors is above 0.41; this means that our classifiers are moderate according to degree scale proposed by (Landis & Koch, 2015), except that J48 scored the best prediction agreement by retaining the highest value 0.593. It also shows that MLP algorithm had the least value for two parameters i.e. RMSE and RRSE. C4.5 having least value for other two parameters i.e. MAE and RAE.

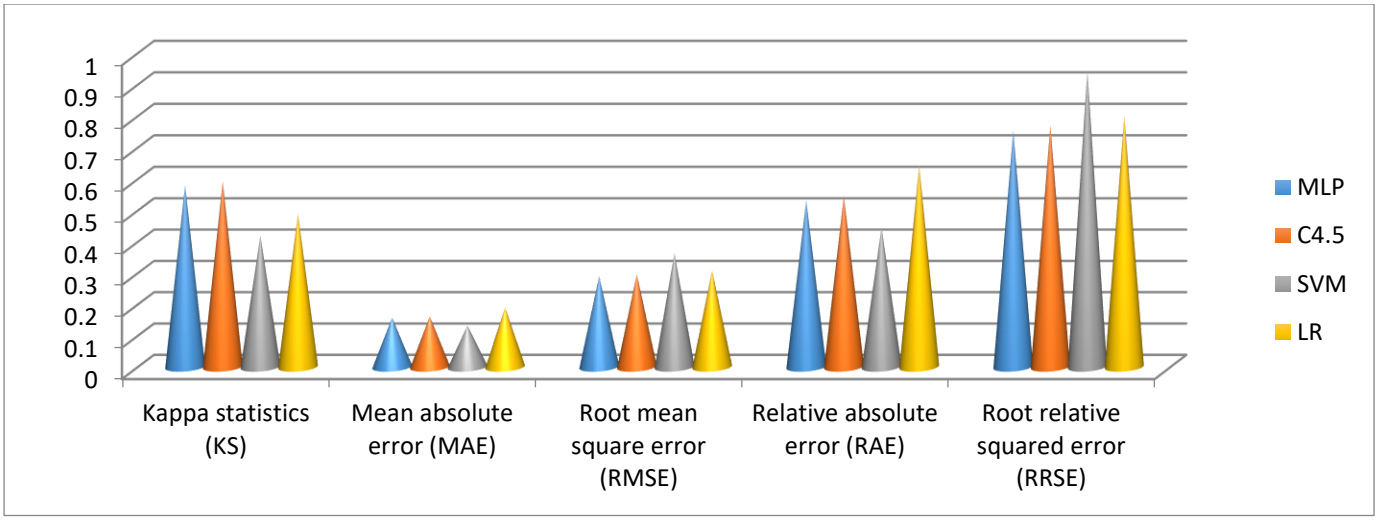

Figure 3: Performance Error

The performance of the learning techniques is highly dependent on the nature of the training data.

C. The Experimental Result for Accuracy, Specificity and Sensitivity

As observed from Table 5, decision tree (C4.5) model had the highest Accuracy value (0.8928) and Specificity (0.5601), which shows best performance in handling breast cancer dataset, followed by neural network (MLP), which has comparable performance with C4.5, while SVM with the lowest accuracy value performed less.

Classification technique	Confusion matrix		Accuracy	Specificity	Sensitivity
Neural Networks (MLP)	162430	5439	0.8907	0.5428	0.9676
	16969	20148			
Decision Tree (C4.5)	162230	5639	**0.8928**	**0.5601**	0.9664
	16327	20790			
Support Vector Machine (SMO)	165331	2538	0.8678	0.3380	**0.9849**
	24571	12546			
Logistic Regression	163348	4521	0.8749	0.4309	0.9731
	21121	15996			

Table 5: The Overall Experimental Result For accuracy, Specificity and Sensitivity of All Model Types

By observing the Sensitivity it appears that Support vector machine has the highest value (0.9849).

D. Percentage Split Method

In percentage split, the database is randomly split in to two disjoint datasets. The first set, which the data mining system tries to extract knowledge from called training set. The extracted knowledge tested against the second set, which is called test set, it is common to randomly split a data set under the mining task in to 2 parts. 66% percentage split is chosen. Objects of the original database are used as a training set and the rest of objects as a test set. Once the tests is carried out using the selected datasets, then using the available classification and 66 % percentage split test mode, results are collected and an overall comparison is conducted.

E. Performance of the Classifiers

In this section we have carried out some experiment in order to evaluate the performance of different techniques for predicting breast cancer survivals in order to time and build a model, correctly classified instances versus incorrectly classified instances by algorithms using 66% percentage split method in the table below.

Evaluation Criteria	Classifiers			
	MLP	C4.5	SVM	LR
Timing to build model (in sec)	537.97	169.9	37126.45	**15.48**
Correctly classified instances	62095	**62207**	60689	60864
Incorrectly classified instances	7533	7421	**9147**	8764

Table 6: Performance of the Classifiers

Table 6 shows time taken to build model, correctly classified instances, incorrectly classified instances for four algorithms. C4.5 and MLP algorithm had the highest number of classified instances. SVM algorithm has highest number of incorrectly classified instances.

ISSN No: - 2456 – 2165

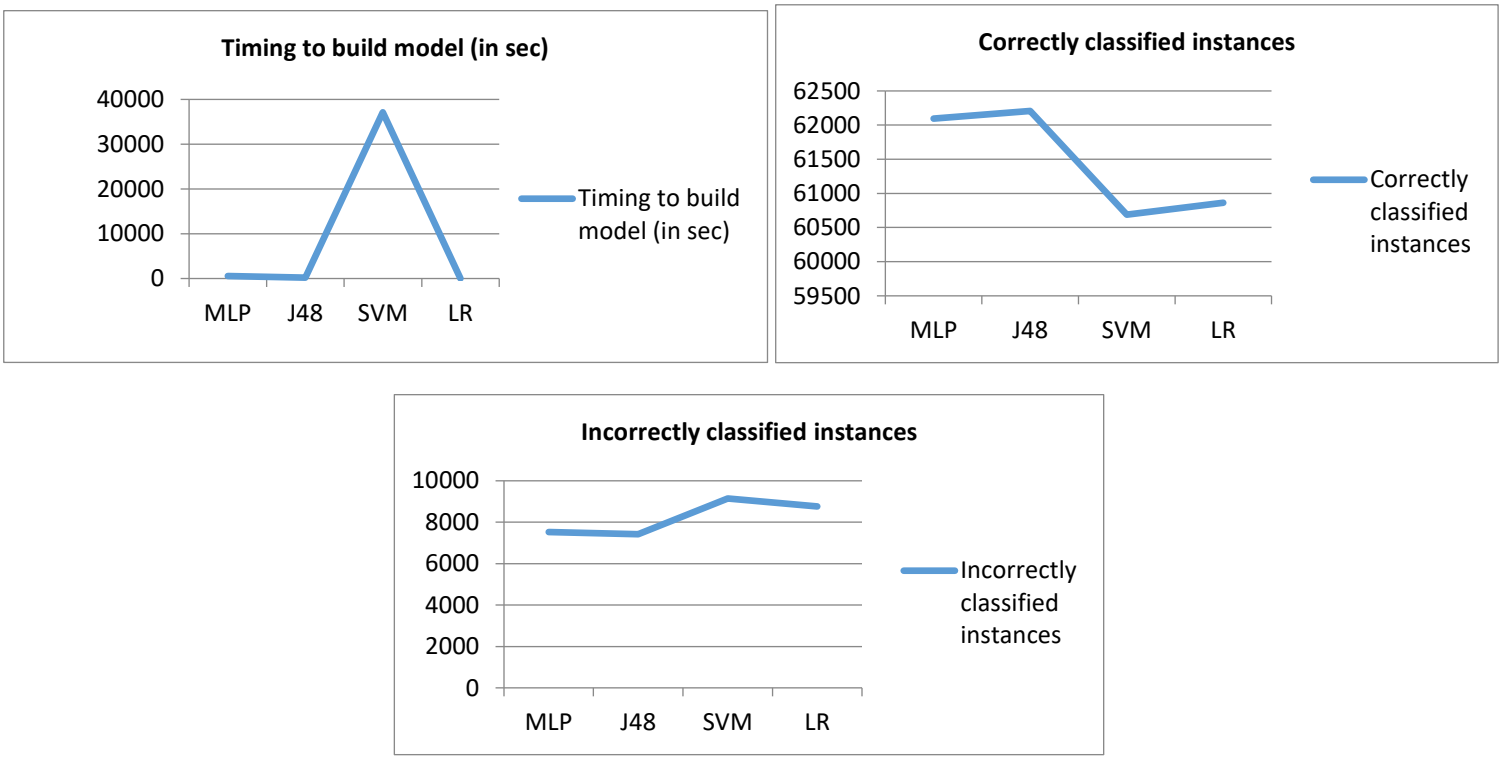

Figure 4: Performance of the Classifiers

From the table 6 and figure 4 it is evident that from overall evaluation C4.5 and MLP algorithm performed well in terms of accuracy. LR shows average performance and proved to be the fastest while SVM shows poor accuracy performance for all parameters.

Evaluation Criteria	**Classifiers**			
	MLP	J48	SVM	LR
Kappa statistics (KS)	**0.5971**	0.5944	0.4231	0.4902
Mean absolute error (MAE)	0.1673	0.1623	**0.131**	0.1912
Root mean square error (RMSE)	**0.2881**	0.2974	0.3619	0.309
Relative absolute error (RAE)	56.3061 %	54.6307%	**44.2227 %**	64.3749 %
Root relative squared error (RRSE)	**74.5569 %**	76.957 %	94.2424 %	79.953 %

Table 7: Performance Error

Table 7 shows four basic error rate parameters and kappa statistics for the evaluation of five classification algorithms. MLP had the least value for RMSE 0.2881, RRSE 74.5569 % and highest value for KS 0.5971**.** SVM had least value for MAE and RAE.

ISSN No: - 2456 – 2165

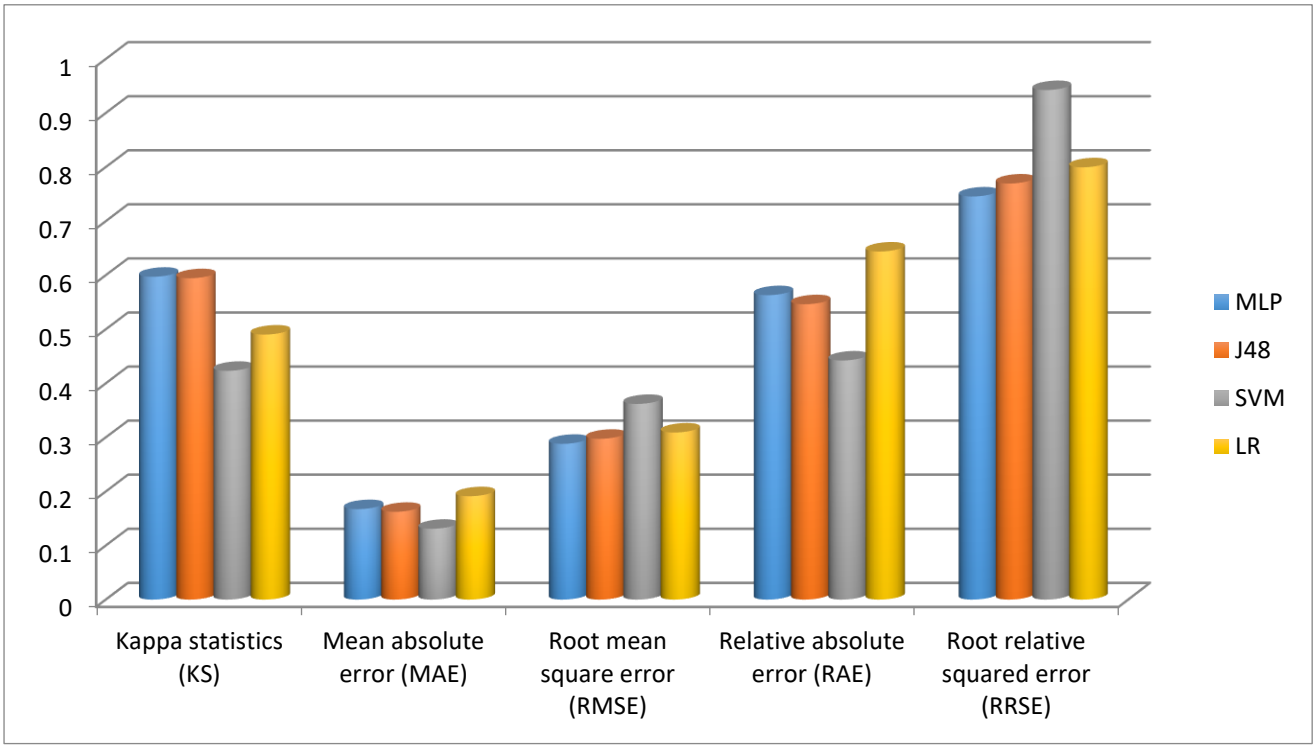

Figure 5: Performance Error

From table 7 figure 5 it is evident that MLP algorithm has the best performance when compares to other techniques. SVM has minimum error followed by C4.5 algorithm. LR has very high error rate and show poor performance.

Classification technique	**Confusion matrix**		**Accuracy**	**Specificity**	**Sensitivity**
Neural Networks (MLP)	54794	2112	0.8918	**0.5739**	0.9629
	5421	7301			
Decision Tree (C4.5)	55164	1742	**0.8934**	0.5536	0.9694
	5679	7043			
Support Vector Machine (SMO)	56409	871	0.8690	0.3409	**0.9848**
	8276	4280			
Logistic Regression (LR)	55363	1543	0.8741	0.4324	0.9729
	7221	5501			

Table 8: The Experimental Result For 66% Percentage Split of All Model Types

From table 8 it shows that C4.5 had the highest accuracy 0.8934 values.While MLP has the highest number of specificity **0.5739.**SVM with the maximum number of sensitivity 0.9848. LR has poor performance on both of the techniques. From all of the above performance measurement parameters it is evident that MLP is the best techniques for the analysis of breast cancer data set.

V. CONCLUSION

In this research work, different techniques are studied and the experiments are conducted to find the best classifier for predicting the patient of breast cancer. Four classifiers such as C4.5, MLP, SVM and LR were used for diagnosis of patients with breast cancer under two different testing methods: 10-fold cross-validation and 66% percentage split. The classification algorithms experimentally compared base on Time taken to build the model, Correctly classified versus Incorrectly classified instances, kappa statistics (KS), Mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE), Root relative squared error (RRSE), Accuracy, Specificity and Sensitivity. After considering and comparing all the tables and graphs under different taste options in our study we found that C4.5 and MLP are best algorithms for classification of Breast cancer dataset. Therefore, they are recommended among all four-classification algorithms.

We also shows that the most important attributes for breast cancer survivals are Age, positive nodes, tumor size, extension of tumor, behavior code, lymph node involvement, and number of nodes, marital status, histologic type, radiation, site-specific surgery, grade, race, primary site code and stage of cancer. These attribute were found using three tests for the assessment of input variables: Chi-square, info Gain test and Gain ratio test.

REFERENCE

[1]. Delen D., Walker G., and Kadam A., Predicting breast cancerSurvivability:a comparison of Three data mining methods, Artificial Intelligence in Medicine. 2005 Jun; 34(2): 113-27

[2]. Lahiri R., Comparison of Data Mining and Statistical Techniques for ClassificationModel, A Thesis submitted to the graduate faculty of the Louisiana State University In partial fulfillment of the requirements for the degree of Master of Science in The Department of Information Systems & Decision Sciences. (December 2006).

[3]. Bellaachia A. And Guven E., Predicting Breast Cancer Survivability Using Data Mining Techniques, Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006).

[4]. M. Panda and M. R. Patra, A comparative study of data mining algorithms for network intrusion detection, proc. of ICETET,India, 2008.pp.504-507. IEEE Xplore.

[5]. vikaschaurasia, Saurabhpal,international journal of computer science and mobile computing. Vol. 3 Issue 1, January – 2014, pg 10-22.

[6]. Frédéric Santos, The Kappa Cohen: a tool to measure the inter-rater agreement on qualitative characters, 2015 Available at http://www.pacea.u-bordeaux1.fr/IMG/pdf/Kappa_Cohen.pdf.

[7]. Kumari M. and Godara S., Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction,

[8]. Ho Yu C., DiGangi S., Jannasch-Pennell A., and Kaprolet C., A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year, Journal of Data Science 8(2010), 307-325.

[9]. C. Deepa, K. SathiyaKumari, and V. PreamSudha, A Tree Based Model for High Performance Concrete Mix Design, International Journal of Engineering Science and Technology Vol. 2(9), 2010, 4640-4646.

[10]. O..O. Adeyemo, T. .O Adeyeye& D. Ogunbiyi (2015). Ccomparative Study of ID3/C4.5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever.Afr J. of Comp & ICTs. Vol 8, No. 1. Pp 103-112.

[11]. cybanko, G. 1989. Approximation by super positions of a sigmoidal function mathematical control, signals and systems, 29(4), 303-314.

[12]. quinlan J.R. (1986). Induction of decision tree, machine learniung, pp81-106.

[13]. Cortes, C.; Vapnik, V. (1995). "Support-vector networks". *Machine Learning*. **20** (3): 273–297.doi:10.1007/BF00994018

[14]. JyotiRohilla , PreetiGulia , International Journal of Advanced Research in Computer Science and Software Engineering 5(7), July- 2015, pp. 696-700