An investigation for detection of Breast Cancer using Data Mining Classification Techniques

Sonu Bala Garg[a*], Ajay Kumar Mahajan[b] and T.S.Kamal[c]

[a]PhD Scholar, IKG Punjab Technical University, Jalandhar, Punjab, India,
[b]Associate Professor, Beant College of Engineering and Technology, Gurdaspur, Punjab, India.
[c]Professor Emeritus, Radiant institute of Engineering and Technology, Abohar, Punjab, India.
Email: [a]sonugarg79@yahoo.com,[b] ajaykm_20@yahoo.co.in, [c]tsk1005@gmail.com

Abstract

Breast cancer is one of the curses for women. Breast cancer caused deaths. It is the second most common cause. 1 in 28 women develop breast cancer during her lifetime in India. Urban/Rural ratio in a lifetime of women for the risk of developing breast cancer is 60:22. High risk group in India has the average age of 43-46 years whereas the same in the west is 53-57 years. The main objective of this paper is to investigate the performance of different classification techniques. Here, the breast cancer data available from the Wisconsin dataset from UCI machine learning is analyzed. In this experiment, Comparison of three different classification techniques have been done in Weka software and comparison results shows that Sequential Minimal Optimisation (SMO) has higher prediction accuracy i.e. 95.8512 % than methods Instance based K-Nearest neighbours classifier (IBK) and Best First (BF) Tree method.

Keywords: Breast Cancer, Data Mining, Data Mining classification techniques.

1. INTRODUCTION

Data mining has been one of the important topic in the research used in medical science during the recent years [1]. The amount of data in the world increases day by day and there's no end to it. We are overwhelmed with data. Today, by using computers data can be saved easily. As the quantity of data increases, inexorably, the proportions of it that people understand decreases alarmingly. Lying hidden in all this data is information. In data mining the data is stored automatically [2]. Data is analyzed from different perspectives by using data mining and it is also used for briefing it into useful information. The main aim of data mining is to find out new patterns for the users and understand the data patterns to give important and fundamental

information. Data mining is important for the analysis of healthcare data and treatment to find patterns [3]. To predict the result of some diseases or find out the hereditary behaviour of tumours, the classification of breast cancer data can be helpful. For prediction and classification of breast cancer pattern, several techniques have been developed. In this study, three classification techniques are compared to find out the suitability for direct interpretation of the results.

1.1 Breast Cancer

Cancer starts from cells which are the building blocks that make up all tissues and organs of the body. Usually cells grow in the breast and other parts of the body and divide to form new cells. When normal cells grow, old get damaged, they die, and new cells take their place. Sometimes, this process goes wrong. New cells form when the body doesn't need them, and old or damaged cells don't die as they should. The build-up of extra cells often forms a mass of tissue called a lump, growth, or tumour [1].

Types of tumours benign (not cancer) or malignant (cancer):

Benign tumours

Benign tumours are usually not harmful. They rarely invade the tissues around them. These tissues don't spread to other parts of the body and can be removed and usually don't grow back.

Malignant tumours

Malignant tumours are opposite to that of benign tumours. Malignant tumours are very harmful and may be threat to life and can invade nearby organs and tissues .These tissues can spread to other parts of the body. They can often be removed but sometimes grow back.

1.2 Breast cancer diagnosis model

Fig.1 shows the functional block diagram of the Breast cancer diagnostic model. It consists of two phases namely: training and testing phases. The training phase includes four steps: acquisition, pre-process, feature extraction and feature selection, whereas the testing phase includes the same four steps in the training phase in addition to the classification step. In acquisition step, the sensor data is subject to a feature extraction and selection process to find out the input vector for the succeeding classifier. This makes a conclusion about the class related with this pattern vector. For feature selection or feature extraction, dimensionality reduction is developed. The image is prepared and cleaned; to clear the noise in the pre-processing step and quality of the images are also improved in the pre-processing step. On the other hand, feature extraction considers the entire information content and maps the valuable information content

into a lower dimensional feature space. Feature selection is based on omitting those features from the existing dimensions which do not give to class separability. That is, unnecessary and inappropriate features are unseen. Different classifiers are used to find out the best result of diagnosing and prognosing the tumor in the classification step [4].

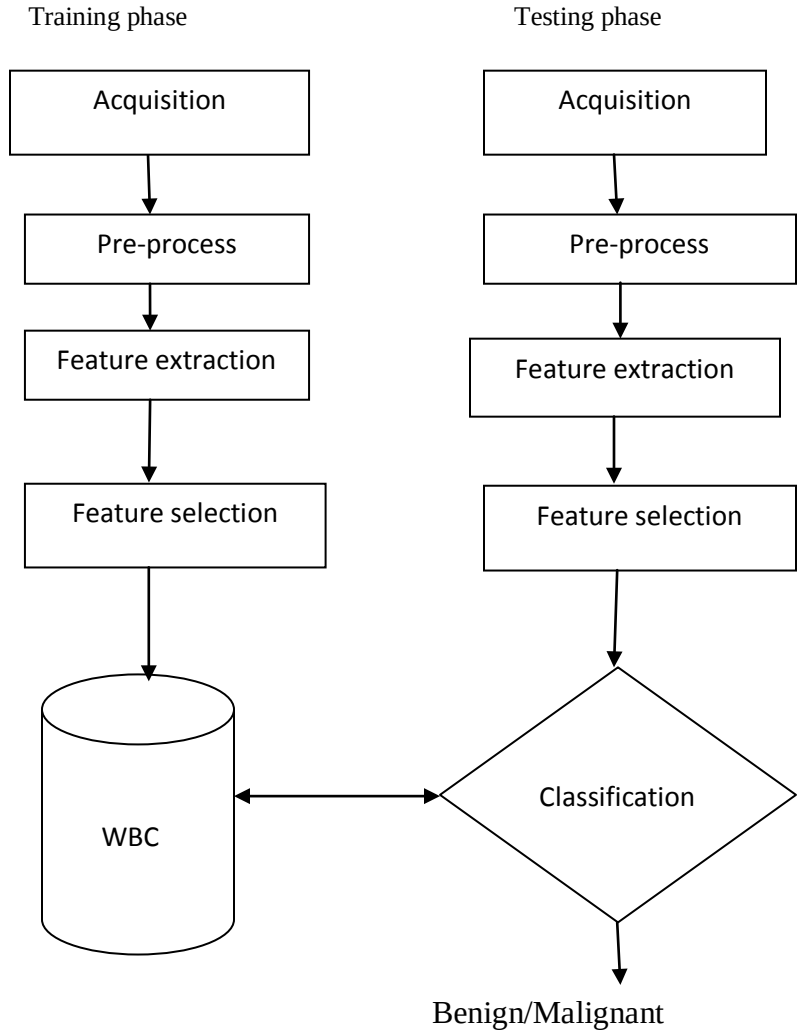

Fig.1: Breast Cancer Diagnosis Model [4]

2. LITERATURE REVIEW

This section overviews the brief contents of classification techniques used in different healthcare datasets in data mining and Knowledge Discovery in Database (KDD).

2.1 KDD and Data Mining

In today's scenario, there is an extensive need for a dominant systematically answer for the mining of the valuable information from the bulk quantity of information composed and stored in an organization's databases or repositories. This has lead to the appearance of KDD which is responsible for transforming low-level data into high-level knowledge for decision making. KDD consists of the list of iterative series of steps of process, and data mining is one of core step in the KDD processes [4]. KDD can be defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [5].

The term KDD and Data Mining are often used inter-changeability. This problem is because of the three different perspectives to look at the data mining but in real, data mining is an important step in KDD process. Data Mining is defined as the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories [6]. Data mining is the process of discovering interesting and expensive information from large data. [7]. Data mining is the essential process of discovering hidden and interesting patterns from massive amount of data where data is stored in data warehouse, OLAP (on line analytical process), databases and other repositories of information [[8]. By studying the various different definitions of data mining. Zhou finally suggested that the database, machine learning and statistics perspectives of data mining put particular emphases on efficiency, effectiveness and validity respectively [9].

2.2 Classification Techniques

Data mining has been useful to a combination of healthcare domains. In healthcare fields, one of the most important challenges is to find out consistent information for the analysis of the data. Here, data mining classification techniques, which are used for the identification of different datasets, have been summarized.

Chaurasia et al. [1] used RepTree, RBF Network and Simple Logistic to predict the survivability for breast cancer patients. From the results it is concluded that Simple Logistic is best based on the patient's data.

Chaurasia et al. [2] used SMO, IBK and BF Tree to predict the survivability for breast cancer patients. SMO is best based on the patient's data.

Kumar et al. [3] used different classification algorithms like C-RT, CS-RT, C4.5, ID3, K-NN, LDA, NAIVE BAYES, PLS-DA, SVM, RND TREE for comparing error rates. In C4.5 algorithm, a classification rate of 91% was obtained.

Salama et al. [4] used different classification algorithms like decision tree (J48), Multi-layer Perceptron (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based K-Nearest neighbour (IBK) on three different datasets of breast cancer. From the results it is concluded that the results are different for different datasets. In WBC dataset show that the fusion between MLP and J48 classifiers with features selection (PCA) is superior to the other classifiers. On the other hand WDBC dataset shows that using single classifiers (SMO) or using fusion of SMO and MLP or SMO and IBK is better than other classifiers. Finally, the fusion of MLP, J48, SMO and IBK is superior to the other classifiers in WPBC dataset.

Kumar et al. [5] used different classification algorithms like C-RT, CS-RT, C4.5, ID3, K-NN, LDA, NAIVE BAYES, PLS-DA, SVM, RND TREE for comparing error rates. In C4.5 algorithm, a classification rate of 91% was obtained.

Gupta et al. [6] performed analysis using different techniques for healthcare data. They found SVM technique showed 96.74% accuracy rate for PIMA Indian Diabetes dataset and 99.25% accuracy rate for Stat Log Heart Disease dataset. Further, C4.5 decision tree showed an accuracy rate of 79.71% for BUPA Liver-disorders dataset whereas for Wisconsin Breast Cancer dataset Bayes Net, SVM, KNN and RBF-NN showed similar results with high accuracy rate of 97.28%.

Tan et al. [10] used C4.5 decision tree, bagged decision tree on seven publically available.

Rajni et al. used Naive Bayes, Multilayer Perceptron, SVM, J48, Random forest and Decision Table to predict lumbar spine diseases. From the result, it is concluded that Multilayer Perceptron is best based on the patient's data [11].

3. EXPERIMENTAL PROCEDURE

3.1 Classification

The choice of accurate and efficient classifier for big datasets is important in data mining and machine learning research. To analyze the data, classification is an initial step for investigating similar group data. The classification helps in raising the understanding and modifying predictions compared to uncertain data. One of the fundamental tasks in data mining is to build an efficient and effective classification method. In literature, different types of classification techniques have been proposed that includes Decision Tree, Naive-Bayesian methods, SMO, IBK and BF Tree etc. [1].

Sequential Minimal Optimization (SMO)

Sequential Minimal Optimization (SMO) is an easy and rapid method. This algorithm is generally applied for solving the optimization problems. It splits the problem into a number of sub-problems, which are then solved logically. It is used for guiding Support Vector Machines

(SVM). In support vector machine, we build hyper planes that split up the data into different classes. Then the hyperplane that is having the largest distance to the nearest training data point is chosen as the best hyperplane [11].

Instance Based K nearest Neighbours Classifier (IBK)

Clustering is most frequently used data mining technique. K-nearest neighbour classification classifies instances based on their similarity. It is one of the most commonly used algorithms for pattern recognition. It is a kind of lazy learning where the function is only approximated locally and all computation is delayed until classification. An item is classified by a majority of its neighbours. K is always a positive integer. The neighbours are chosen from a set of objects for which the correct classification is known. In Weka this classifier is called IBK[12].

Best First Tree (BF)

For selecting the node, Best First trees are developed. To maximize the impurity reduction, among all the existing nodes to split is the main aim. The impurity measure used by this algorithm is the Gini index and information gain.Best –first trees are made in a divide-and –conquer style related to standard depth-first decision tree. The central aim for making the best-first tree is as follows. First, attribute is chosen for placing at the root node. Choose an attribute to place at the root node and based on some criteria make some branches for this attribute. Then, divide training instances into subsets, one for each branch extending from the root node. These constructing procedures continuous until all nodes are pure or a particular number of expansions are reached. The information gain and the Gini index are also used to determine node order. The best-first technique always selects the node for expansion whose consequent best split gives the best information gain or Gini index along with all unexpanded nodes in the tree. [13]

4. DATASET

The data used in this study have been studied from UCI machine learning repository, breast cancer-Wisconsin having 699 instances, 2 class (benign and malignant), as shown in table 1 & 2 below. Weka toolkit has been used for experimentation of different data mining algorithms. Experiments were performed using libraries from Weka machine learning environment. The Weka is an ensemble of tools for data classification, regression, clustering, association rules and visualization. As a data mining tool WEKA version 3.8 was utilized to evaluate the performance and effectiveness of the 3-breast cancer prediction models built from several techniques because WEKA program offers a well defined framework for experimenters and developers to build and evaluate their models. The results show clearly that the proposed method performs well

compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of breast cancer in the patients.

Table1. Dataset of Breast Cancer

Attribute	**Domain**
1.Sample code number	Id number
2.Clump Thickness	1-10
3.Uniformity of Cell Size	1-10
4.Uniformity of Cell Shape	1-10
5.Marginal Adhesion	1-10
6.Single Epithelial Cell Size	1-10
7.Bare Nuclei	1-10
8.Bland Chromatin	1-10
9.normal Nucleoli	1-10
10.Mitoses	1-10
11.Class	2 for benign, 4 for malignant

4.1 Experimental Results

This section summarizes the results of experiments. Firstly the final data set was described, and then the results of modeling from classification were provided. 10-fold cross validation method has been used for all the classifiers. The visual form of Breast cancer survivals using all attributes have been shown in Fig 2.

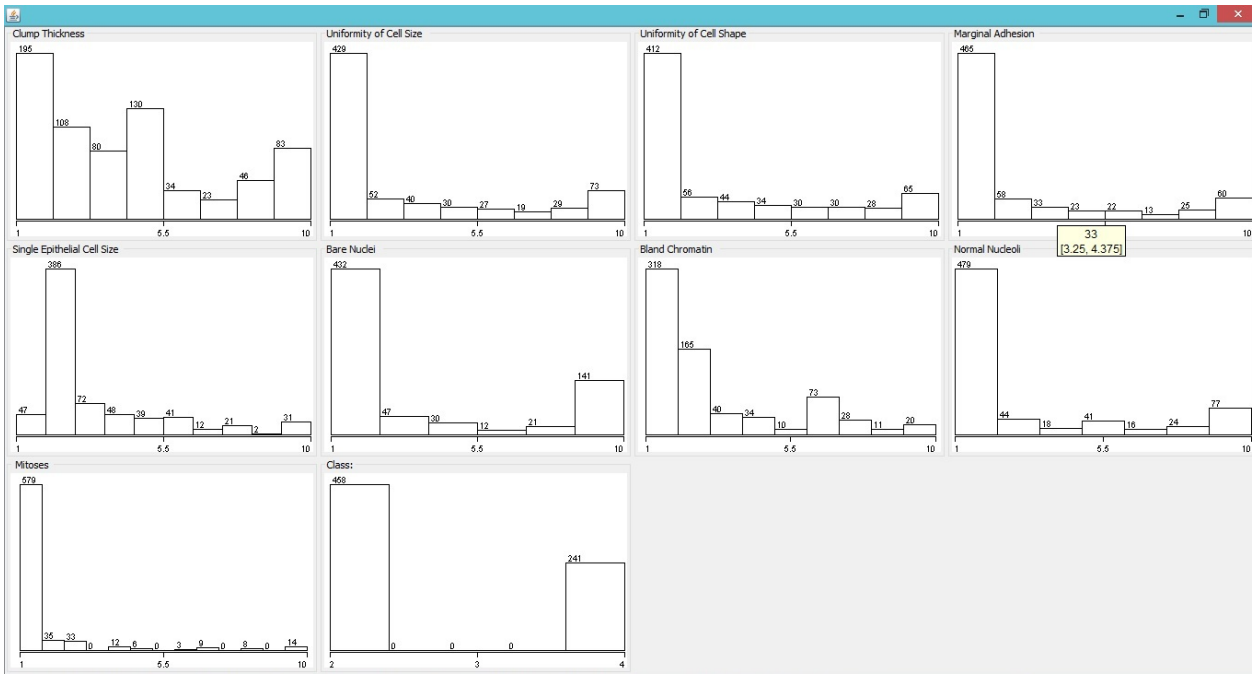

Fig. 2: Graph shows Visual form of breast cancer survivals using all attributes

Table 2 shows the experimental result. Experiments have been carried out in order to evaluate the performance and usefulness of different classification algorithms for predicting breast cancer patients.

Table2.Performance of the classifiers

Technique	SMO	IBK	BF Tree
Correctly Classified Instances	670	668	654
Incorrectly Classified Instances	29	31	45

From above table it can be concluded that SMO has highly classified correct instances as well as incorrectly classified instance than other two classifiers IBK and BF Tree as shown in figure 3.

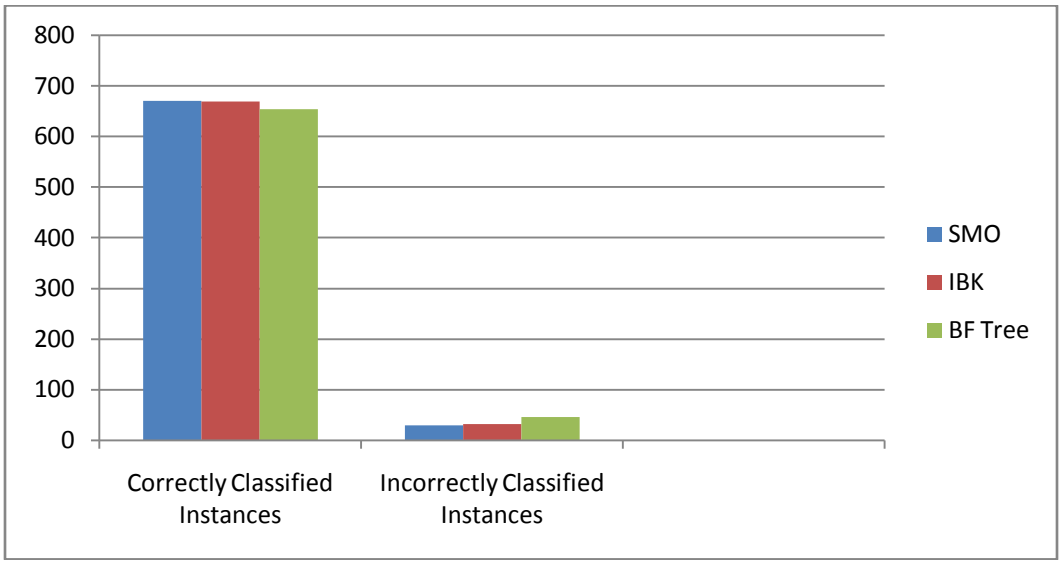

Fig. 3: Graph of different classifiers showing correctly classified instances and incorrectly classified instances

Table 3: Accuracy of the classifiers

Technique	SMO	IBK	BF Tree
Accuracy	95.8512%	95.5651%	93.5622%

From above table it can be concluded that the accuracy of SMO is more in comparison to BFTree and IBK.

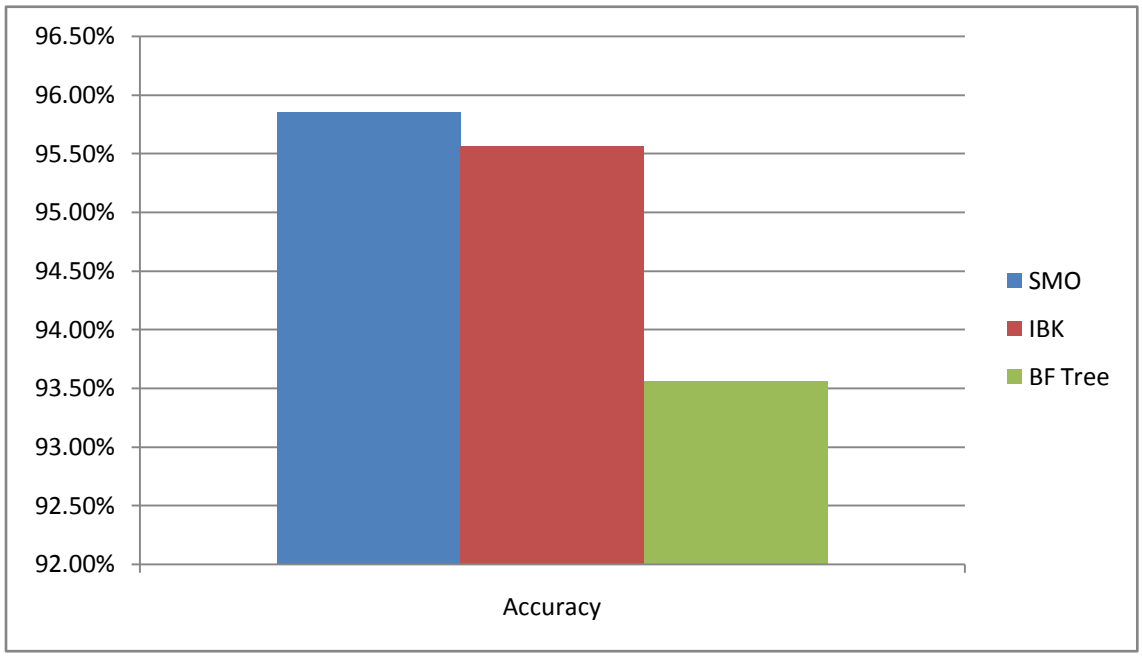

Fig. 4: Graph of different classifiers showing accuracy

Table 4: Error Rate of the classifiers

Technique	SMO	IBK	BF Tree
Error Rate	4.1488%	4.4349%	6.4378%

From table 4, it has been found that SMO has less error than IBK and BF Tree.

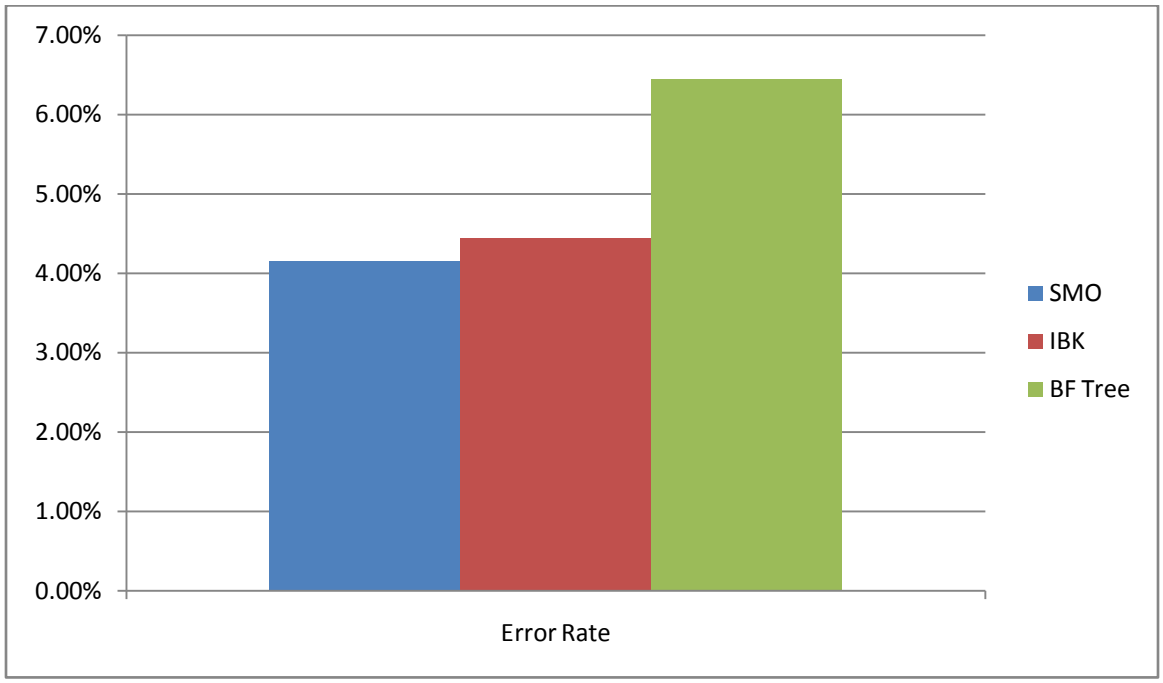

Fig. 5: Graph of different classifiers showing Error Rate

In table 5, the values of parameters i.e. Kappa statistic, mean absolute error and root mean squared error have been given in numeric value only whereas the values of relative absolute error and root relative squared error have been given in percentage for references and evaluation. The results of the simulation are shown in Tables 5 and Figure 6 shows the graphical representations of the comparison between parameters.

Table 5: Training and simulation error

Technique	SMO	IBK	BF Tree
Kappa Statistics	0.9083	0.901	0.8579
MAE	0.0415	0.0468	0.0872
RMSE	0.2037	0.1939	0.2459
RAE	9.1793%	10.3647%	19.301%
RRSE	42.854%	40.7975%	51.7441%

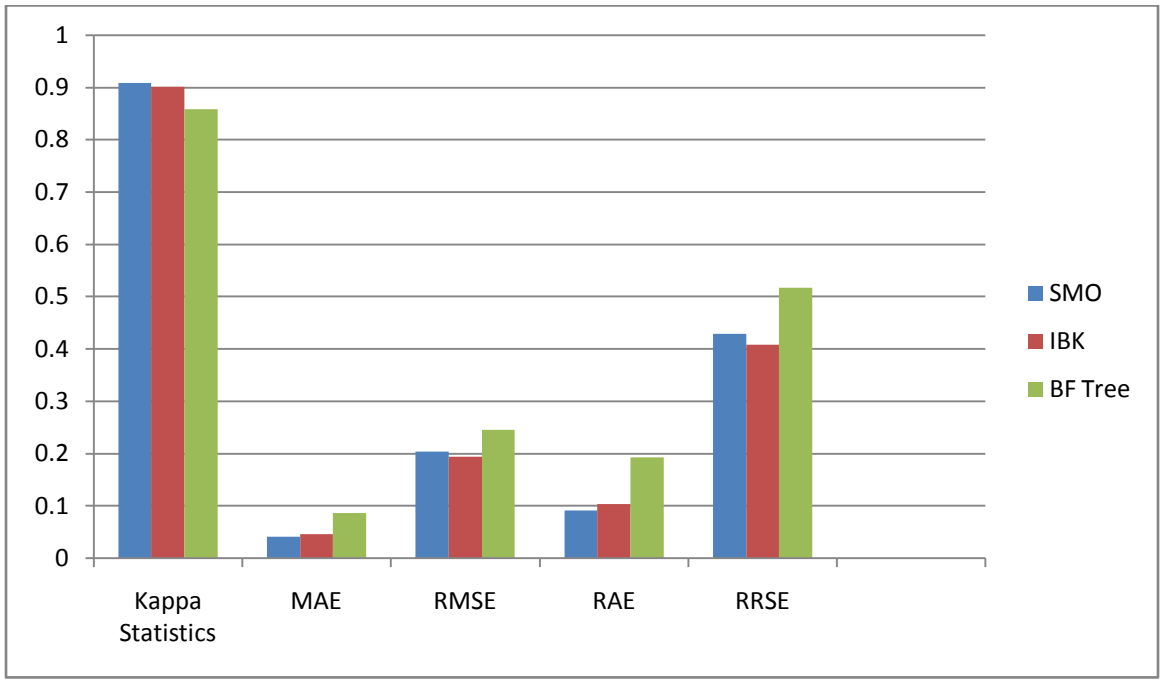

Fig. 6: Graph shows the comparison between parameters

$$Sensitivity(TPR) = \frac{TP}{TP + FN}$$

$$Specificity(TNR) = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where
TPR=True positive rate
TNR=True negative rate
True positive (TP) = number of positive samples correctly predicted.
False negative (FN) = number of positive samples wrongly predicted.
False positive (FP) = number of negative samples wrongly predicted as positive.
True negative (TN) = number of negative samples correctly predicted.

Table 6 shows the TP rate, FP rate, precision, recall, F-measures, ROC area and class value for SMO, IBK and BF Tree. Fig.7 shows the comparison between parameters.

Table 6: Comparison of accuracy measures

Technique	SMO	IBK	BF Tree
TP Rate	0.967	0.976	0.948
	0.942	0.917	0.913
FP Rate	0.058	0.083	0.087
	0.033	0.024	0.052
Precision	0.969	0.0957	0.954
	0.938	0.0953	0.902
Recall	0.967	0.976	0.948
	0.942	0.917	0.913
F-Measure	0.968	0.966	0.951
	0.94	0.934	0.907
ROC Area	0.955	0.984	0.923
	0.955	0.984	0.923
Class	2	2	2
	4	4	4

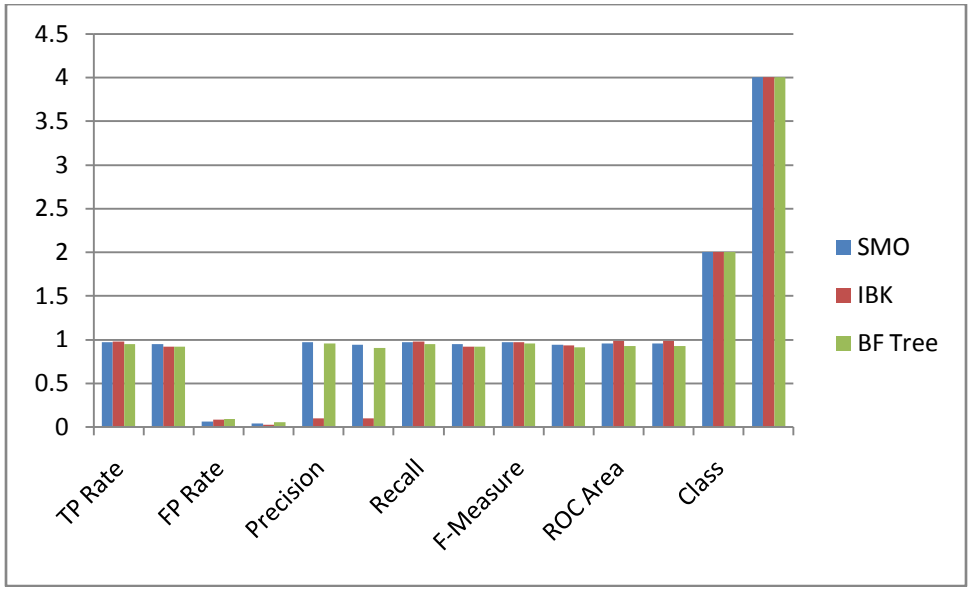

Fig. 7: Graph shows the comparison between parameters

Confusion matrix is a visualization tool and is used to displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model. The columns represent the actual class and the rows represent the predicted values. The usual methodology is performed to evaluate the robustness of classifier as shown in the table 7. Figure 8 shows the graphical representation of different classifier of confusion matrix.

Table 7: Confusion matrix

Technique	SMO		IBK		BF Tree	
Benign	443	15	447	11	434	24
Malignant	14	227	20	221	21	220
Class	2	4	2	4	2	4

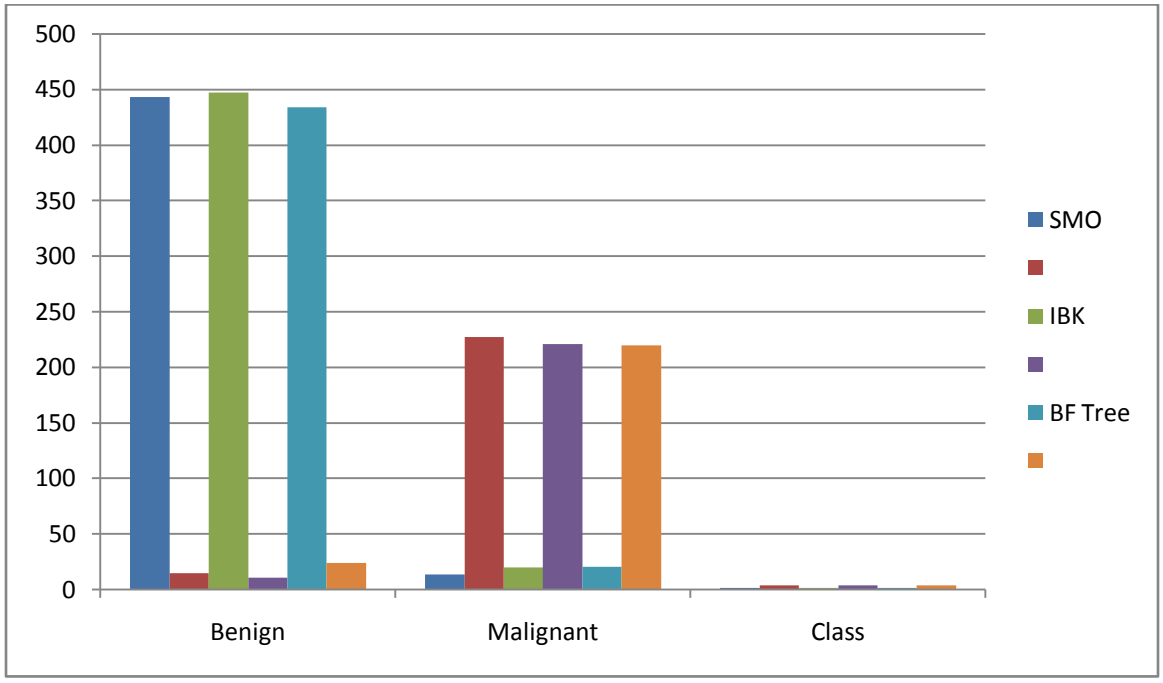

Fig. 8: Graph of different classifiers showing confusion matrix

4. CONCLUSION

In this paper it has been concluded that SMO has a high level performance than other two classifiers IBK and BF Tree and SMO shows the concrete results with Breast Cancer disease of patient records. Therefore SMO classifier is suggested for diagnosis of Breast Cancer disease based classification to get better results with accuracy and low error rate.

REFERENCES

[1] Vikas Chaurasia, Saurabh Pal "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", International Journal of Computer Science and Mobile Computing, (2014) Vol. 3, Issue 1, pp. 10 – 22.

[2] Vikas Chaurasia, Saurabh Pal, "Novel approach for breast cancer detection using data mining techniques", International Journal of Innovative Research in Computer and Communication Engineering, (2014) Vol. 2, Issue 1, pp. 2456-2465.

[3] Sushil Kumar.R. Kalmegh, "Analysis of Weka data mining algorithm REPTree,Simple Cart and Random Tree for Classification of Indian News", International Journal of Innovative Science, Engineering & Technology, (2015) Vol. 2, Issue 2.

[4] Gouda I. Salama, M.B. Abdelhalim, Magdy Abd-elghany Zeid "Breast cancer Diagnosis on three different datasets using Multi-Classifiers", International Journal of Computer and information Technology, (2012) Vol. 1, Issue 1,pp.36-43.

[5] Sushil Kumar. R. Kalmegh, "Successful Assessment of Categorization of Indian News Using JRip and Nnge Algorithm", International Journal of Emerging Technology and Advanced engineering, (2014) Vol. 4, Issue 12, pp. 395-402.

[6] Shelly Gupta, Dharminder Kumar and Anand Sharma "Performance analysis of various data mining classification techniques on health care data", International Journal of Computer Science & Information Technology, (2011) Vol. 3, Issue 4.

[7] V.Vaithiyanathan, K.Rajeswari, Rashmi Phalnikar and Swati Tonge **"**Improved Apriori algorithm based on Selection Criterion", IEEE International Conference on Computational Intelligence and Computing Research (2012).

[8] M. Halkidi, "Quality assessment and uncertainty handling in data mining process," in Proc, EDBT Conference, Konstanz, Germany (2000).

[9] Zhou, Z.H., "Three perspectives of data mining", Artificial Intelligence, (2003) Vol. 143, Issue 1, pp.139-146.

[10] Tan AC, Gilbert D. "Ensemble machine learning on gene expression data for cancer classification", Appl Bioinformatics (2003).

[11] Rajni Bedi and Ajay Shiv Sharma "Classification Algorithms for Prediction of Lumbar Spine Pathologies", IEEE International Conference on advanced informatics for computing research, (2017) pp. 42–50.

[12] D. Wolpert and W. Macready, No Free Lunch Theorems for Search, Santa Fe Institute, (1995) Technical report No., No. SFI-TR-95-02-010.

[13] Haijian Shi. "Best-first decision tree learning", Master's thesis, University of Waikato, Hamilton,NZ, (2007) COMP594.New Zealand. Retrieved from http://hdl.handle.net/10289/2317

[14] UCI Machine Learning Repository [online]. Available http://archive.ics.uci.edu/ml/datasets.html