PERFORMANCE EVALUATION OF BREAST CANCER DIAGNOSIS USING RADIAL BASIS FUNCTION, C4.5 AND ADABOOST

Ameen A. O.[1], Olagunju M.[2], Awotunde J. B.[3], Adebakin T.O.[4], Alabi I.O.[5]
[1,4]Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[2]Department of Computer Science, Kwara State Polytechnic, Ilorin, Nigeria
[3]Department of Physical & Computer Science, McPherson University, Nigeria
[5]Department of Information & Media Technology, FUT Minna, Nigeria
[1]aminamed@unilorin.edu.ng, [2]awotunde.jb@unilorin.edu.ng

Keywords: Breast cancer diagnosis, Classification algorithm, Expert System, Radial basis function, Support vector machines, Data mining

Abstract: *This paper conducted a performance evaluation on the most commonly data mining algorithms: Support Vector Machines (Radial basis function), C4.5 decision tree algorithm and Adaboost, using the two previous algorithms as base classifiers (ensemble approach), on breast cancer diagnostic removing redundant or irrelevant features using Chi-square. Result shows that while C4.5 builds its classification model in a short time, The Adaboost with SVM as its base classifier when three features are removed proved to be the best algorithm in classifying breast cancer.*

1. INTRODUCTION

The healthcare system is adequately endowed with large quality of data, but little or no effort has been applied to the robust information contained in this data for solving some critical problems in medical diagnosis of different diseases (Adeyemo & Adeyeye, 2015). Among numerous techniques to achieve this task, data mining remains the most significant approach for prediction or diagnosis of several disease (Shouman, Turner & Stocker, 2012). Data mining has turned to be a crucial procedure in registering applications in the area of medicine (Zorluoglu & Agaoglu, 2015). Data mining approaches in medical domains is increasing rapidly due to the improvement effectiveness of these methods for classification and prediction (Lebbe, Saabith, Sundararajan & Bakar, 2014). Data mining is widely used in various application domains such as market analysis, credit assessment, stock market, fraud detection, fault diagnosis in production system, hazard forecasting, medical discovery, buying trends analysis, knowledge acquisition and science exploration (Lebbe et al., 2014)

It remains one of the most significant approaches for prediction or diagnosis of several diseases. Breast cancer is a serious ailment, which has been discovered to be second cause of death among women in the society. Several data mining classification approaches such as Neural Network, Support Vector Machine, Adaboost, Decision Tree, Naïve Bayes and KNN have been proposed by researchers to diagnose breast cancer disease. But there is a challenge to ascertain which of these data mining techniques perform effectively. It has been also identified that most time single data mining method may not provide desired result. In order to find a solution to this problem, the paper conducted a performance evaluation on the most commonly data mining algorithms:

In addition to its importance in finding ways to improve patient outcomes, it reduces the medical cost and enhances early disease discovery. In the recent years, the rising of breast cancer incidence in the developing countries is drastically on high increase, which eventually has been seen to be a great threat (Chaurasia & Pal,

2014). Breast cancer is a type of cancer which occurs due to change in normal cells at the breast region of the body as a result of uncontrolled growth of cells that give rise to tumour. Classification of it as a disease in medicine is based on the particular area where the cell or tissue cancer is formed. The most effective method to reduce ailment death is through earlier detection. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumours from malignant ones without going for surgical biopsy (Shukla, Gupta & Prasad, 2016).

Doctors and people suffering from this ailment need consistent and accurate knowledge of a person's possibility of having this disease. Developing an algorithm to estimate the risk of categorizing people with the chance of having the disease can be achieved through techniques in data mining. Techniques in data mining are used to analyze and deduce unknown relationship among features of clinical data to handle some problems such as prediction, diagnosis, control and treatment of diseases (Khaleel, Pradhan & Dash, 2013). Unlike the statistical methods, data mining approach searches for intriguing information without considering previous postulates on the type of patterns that can be found. The classification of the disease can be useful in the prediction or discovery of the genetic behaviour of tumour.

However, advanced and modern world may be today and are riding on the chariot of advancement, but the truth is that there are many things which are beyond our control even today and one of them is cancer. It is wrong to say that the cancer is incurable. But the death rate of patients of breast cancer is still very high. World Health Organization (WHO) has stated that the breast cancer is most frequently found cancer in the women and it is adversary affecting millions of women all over the world. But the positive trend is that the death rate is gradually declining after 1990 due to screening, early detection, awareness and continuous improvement in treatment (Breast Cancer Deadline 2020). Some of the key risk factors of breast cancers are age, gender, affluence, family history, breast conditions, alcohol consumption and obese (Breast Cancer Deadline 2020; Report to the nation-breast cancer).

Among the various classification algorithms the very famous algorithms SVM, Adaboost, ID3 and C4.5 play an essential role in breast cancer analysis (Shweta Kharya, 2012). A number of studies have been undertaken in order to understand the prediction and classification of breast cancer risks using data mining techniques. Hence, this study only applied three data mining techniques to predict breast cancer in patients using the C 4.5 Decision Tree, Kernel Basis Function Support Vector Machine (SVM) and Adaboost. The performance of these classification techniques was evaluated in order to determine the most efficient and effective model.

1.1 Data Mining

Data mining is an essential step in the process of knowledge discovery in database that is used for the extraction of patterns from data (Karim & Zand, 2015). The patterns that can be discovered depend on the data mining tasks applied. There are many approaches used in the breast cancer prediction or diagnosis using data mining methods. These methods include Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Naïve Bayes (NB) (Zorluoglu & Agaoglu, 2015).

1.2 Classification Algorithms in Data mining

Classification algorithms are used to categorize or find out group each data instance is belong within a given dataset. The algorithm is used to categorized or classified data into different classes by a given constrains. The techniques is capable of processing large amount of data, to predict class labels and classifies data based on training set . The class labels and training dataset can be used for classifying newly available data. Furthermore, it covers any context to make decision or forecast base on present available information. There are many types of classification algorithms among them are: C4.5, ID3, k-nearest neighbor classifier, Naive Bayes, SVM, and ANN etc. the methods normally follows three approaches Statistical, Machine Learning and Neural Network methods.

1.2.1 Support Vector Machine (SVM)

The support vector machine algorithms apply 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 (Karatzoglou et al., 2005).

SVM builds its method on the principle of VC dimension from statistical learning and Structural Risk Minimization. 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.

1.2.2 Artificial Neural Network (ANN)

Artificial neural networks are models which draws its inspiration from biological nervous systems which comprises of neural network. 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.

1.2.3 Naïve Bayes (NB)

Naïve Bayes is based on Bayesian theorem rule and it assumes independence naively. This classification technique analyses 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 eliminate redundant and unimportant features.

1.2.4 Decision Tree

Decision tree is one of the classification techniques in data mining method that is employed for decision support system and machine learning process (Seema et al., 2012). This technique plays a significant role in process of data mining and data analysis (Singh, Naveen, & Samota, 2013). Generally, the structure of decision tree allows the applicability to understand the structure of trained knowledge models.

1.2.5 ADABOOST

It is a technique in data mining classification method for constructing a strong classifier as linear combination of a weak classifier (E. R. Kaur & Chopra, 2015). 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. It works by successively introduced it to any classification algorithm and produced a new classified versions of the training data and then taking a weighted reliable of the order of classifiers thus produced. This system of weighting of classification algorithms have results to bring above dramatic improvements in those algorithms performance.

1.3 Techniques in Data Mining

Techniques in data mining are methods employed for sorting data in order to identify patterns, these techniques include association, classification, prediction and clustering (Khaleel et al., 2013).

Association: looks for patterns based on connection of a particular event to other events. It is commonly used approach for prediction of heart disease as it gives the relationship of unlike features for analysis and grouping out patient with all risk factor needed for prediction (Bindushree, 2016).

Clustering: - This is unsupervised machine learning in which no class labels are given. It locates and visually documents collections of facts that not previously known.

Classification: - The technique is supervised learning, where the class labels of some training samples are supplied. Mathematical methods such as linear programming, Naïve Bayes, Decision trees and artificial neural network are employed for classification method (Bindushree, 2016).

Prediction (Forecasting):- determines patterns in data that can result into a reasonable prediction about the future. The prediction technique fits in the prognostic model of data mining (Karthikeyan, Ragavan, & Kanimozhi, 2016).

2. RELATED WORK

The most recent related work was reviews in other to have a direction to this research work the paper reviewed fall within 2011-2016. Several studies have been conducted on the performance analysis of data mining techniques in medical diagnosis of breast cancer. These include:

(Shukla, Gupta & Prasad 2016) performed a survey on the importance and usefulness of data

mining techniques using 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.

(El-hasnony, El-bakry, & Saleh 2016) proposed a system which combines K-means clustering algorithm, fuzzy rough feature set and discernibility nearest neighbor classifier. The proposed model was compared with previous studies and shown to perform better than others with an accuracy of 98.9%

(Zorluoglu & Agaoglu, 2015) 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 evaluation metric which is the accuracy. 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.

(Arutchelvan & Periyasamy 2015) presented a novel multi layered method combining clustering and decision tree technique 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.

(Porkodi & Suganya 2015) presented a cancer prediction using data mining techniques, the study classified colon cancer microarray dataset in bioinformatics using five different classification algorithms namely Naïve Bayesian, K-Nearest Neighbors, Support Vector Machine, Random Forest and Neural Network. The performance of these classification algorithms were calculated based on the Performance measures namely Classification Accuracy (CA), Precision and Recall. The experimental result showed that the highest accuracy was found in both KNN and Neural Network classifier among all other classification algorithms.

(Karim & Zand 2015) conducted a comparative survey on data mining techniques for breast cancer diagnosis and prediction using Naïve Bayes, back propagated neural network and C4.5 Decision tree algorithms which was implemented in Weka toolkit. Experimental result indicated that C4.5 Decision tree outperformed other methods.

(Demigha, 2015) presented concepts and techniques used in developing a data mining system particularly in the medical field and imaging showing that the different tasks of Data mining are beneficial to the medical field for diagnosis, decision making, screening, monitoring, therapy support, patient management amongst others thus improving quality and decreasing cost.

(Kaur & Bawa, 2015) summarizes various technical articles on medical diagnosis and prognosis. It has also been focused on current research being carried out using the data mining techniques to enhance the disease(s) forecasting process. Future trends of current techniques of KDD were discussed in using data mining tools for healthcare, significant issues and challenges associated with data mining and healthcare in general. The research found a growing number of data mining applications, including analysis of health care centers for better health policy-making, detection of disease outbreaks and preventable hospital deaths.

(Chaurasia & Pal, 2014) 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 analyzed 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.

(Lebbe et al 2014) 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).

(Krishna, Nookala, Orsu, Pottumuthu & Mudunuri, 2013) conducted a comprehensive comparative analysis of 14 different classification algorithms and their performance was evaluated by using 3 different cancer data sets. The results indicate that none of the classifiers outperformed all others in terms of the accuracy when applied on all the 3 data sets. The study mentioned that most of the algorithms performed better as the size of the data set was increased. Finally, it was recommended to the users not to stick to a particular classification method and should evaluate different classification algorithms in order to select the better algorithm.

(Omari, 2013) introduced an approach to improve and support decision-making process for breast cancer management in the Kingdom of Saudi Arabia. This can be accomplished by applying different association rule mining algorithms on the cancer information system in Saudi Arabia. It also provides valuable information about predicted distribution and segmentation of cancer in Saudi Arabia, which may be linked to possible risk factors. From the extracted patterns, the information need to be considered in the decision- making process can be identified and recognized as well, which yields to knowledge based decisions. Consequently, identifying health risk behaviours among target group of patients and adopting interventional and preventive measures can be initiated in order to decrease breast cancer incidence and prevalence and ultimately.

(Shajahaan, Shanthi, & Manochitra, 2013) 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.

(Kharya, 2012) discussed the use of data mining techniques for diagnosis and prognosis of cancer diseases in their research work. The study mentioned that the prediction of the outcome of a disease is the most interesting and challenging tasks to develop data mining applications. An overview of the current research was carried out on various types of breast cancer datasets using the data mining techniques to enhance the diagnosis and prognosis. The accuracy of three data mining techniques was compared and experimental results of their approach and the preliminary results capable for the application of the data mining methods into the survivability prediction problem in medical databases. The performance of C4.5 algorithm has a much better than the other two techniques.

(Einipour, 2011) 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.

(Devi, 2011) 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 net classifier outperforms other algorithm used in the research, the other algorithms are J48, Classification via Regression, and Logistic.

Hence, there is need in the diagnostic algorithm of breast cancer using data mining techniques to consider properly the effectiveness of machine learning algorithms that will function optimally during diagnosis. 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 machine learning algorithms must be carried out.

The aim of this study is to conduct a performance analysis on breast cancer diagnostic system using data mining classification methods and the objectives are to:

- Use Chi-square to select relevant attributes from breast cancer dataset.

- Implement the data mining classification methods; Radial Basis Function (SVM), C4.5 Decision Tree and Adaboost in Weka data mining tool environment.

- Conduct comparative performance evaluation on classification algorithms.

3. RESEARCH METHODOLOGY

3.1 The Approach

The development of this experimental framework incorporates the Weka 3.6 platform with interaction with data mining techniques. Data was converted into arff-format and then later loaded into the system. The selection of relevant features using the chi-square filter selector to rank the features according to their relevance. The proposed performance analysis of breast cancer predictive system using Adaboost, C4.5 Decision Tree and Support Vector Machine was achieved by considering several chronological procedures. This study employed three data mining techniques: Adaboost, C4.5 Decision Tree and Support Vector Machine algorithms together with ensemble method to predict the probability of a woman having breast cancer. The selection of three classification methods to conduct performance analysis on each in order to find the most suitable one for the prediction of breast cancer disease. Weka toolkit was used to experiment three data mining algorithms. Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization.

Pre-processing of the input data set for a knowledge discovery goal using a data mining approach usually consumes the biggest portion of the effort in the proposed study. A chi-square filter algorithm will be applied to extract and clean the raw data from the dataset. After the input data normalization, classification will be performed using the three machine learning techniques.

3.2 WEKA Data Mining Tool

This is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, classification, regression, clustering and association rules; it also includes visualization tools. The new machine learning schemes can also be developed with this package. WEKA is an open source software issued under General Public License. The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file (foremost: attribute names, attribute types, attribute values and the data). The main interface in Weka is the Explorer. It has a set of panels, each of which can be used to perform a certain task. Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis.

3.3 System Architecture

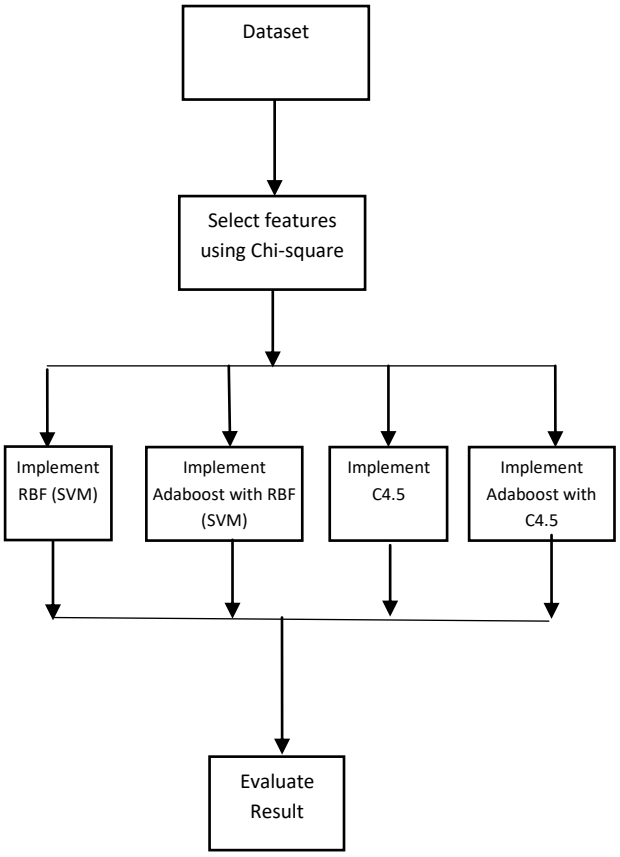

Fig 1: Block Diagram for System Architecture

3.4 Dataset Description

This study used the Wisconsin breast cancer dataset to carry out the performance

evaluation on the data mining classification techniques. This is further explained explicitly in subsection 3.4.1

3.4.1 Wisconsin Breast Cancer Dataset

The details of the attributes found in Wisconsin Breast Cancer Dataset (WDBC) dataset: ID number, Diagnosis (M = malignant, B = benign) and ten real valued features are computed for each cell nucleus: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension. These features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. When the radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points. The total distance between consecutive snake points constitutes the nuclear perimeter.

The area is measured by counting the number of pixels on the interior of the snake and adding one-half of the pixels on the perimeter. The perimeter and area are combined to give a measure of the compactness of the cell nuclei using the formula. Smoothness is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it. This is similar to the curvature energy computation in the snakes. Concavity captured by measuring the size of the indentation (concavities) in the boundary of the cell nucleus. Chords between nonadjacent snake points are drawn and measure the extent to which the actual boundary of the nucleus lies on the inside of each chord. Concave Points: This feature is Similar to concavity but counted only the number of boundary point lying on the concave regions of the boundary. In order to measure symmetry, the major axis, or longest chord through the center, is found. Then the length difference between lines perpendicular to the major axis to the nuclear boundary in both directions is measured. The fractal dimension of a nuclear boundary is approximated using the "coastline approximation" described by Mandelbrot. The perimeter of the nucleus is measured using increasingly larger "rulers". As the ruler size increases, decreasing the precision of the measurement, the observed perimeter decreases. Plotting log of observed perimeter against log of ruler size and measuring the downward slope gives (the negative of) an approximation to the fractal dimension. With all the shape features, a higher value corresponds to a less regular contour and thus to a higher probability of malignancy. The texture of the cell nucleus is measured by finding the variance of the grey scale intensities in the component pixels.

3.4.2 Samples of Wisconsin Breast Cancer Dataset

@attribute Cell_Shape_Uniformity integer [1,10]
@attribute Marginal_Adhesion integer [1,10]
@attribute Single_Epi_Cell_Size integer [1,10]
@attribute Bare_Nuclei integer [1,10]
@attribute Bland_Chromatin integer [1,10]
@attribute Normal_Nucleoli integer [1,10]
@attribute Mitoses integer [1,10]
@attribute Clump thickness integer [1,10]
@attribute Uniformity of cell size integer [1,10]
@attribute Type of cell integer [1,10]
@attribute Class {benign, malignant}
@data

5,1,1,1,2,1,3,1,1,benign
5,4,4,5,7,10,3,2,1,benign
3,1,1,1,2,2,3,1,1,benign
6,8,8,1,3,4,3,7,1,benign
4,1,1,3,2,1,3,1,1,benign
8,10,10,8,7,10,9,7,1,malignant
1,1,1,1,2,10,3,1,1,benign
2,1,2,1,2,1,3,1,1,benign
2,1,1,1,2,1,1,1,5,benign
4,2,1,1,2,1,2,1,1,benign
1,1,1,1,1,1,3,1,1,benign
2,1,1,1,2,1,2,1,1,benign
5,3,3,3,2,3,4,4,1,malignant
1,1,1,1,2,3,3,1,1,benign
8,7,5,10,7,9,5,5,4,malignant
7,4,6,4,6,1,4,3,1,malignant
4,1,1,1,2,1,2,1,1,benign
4,1,1,1,2,1,3,1,1,benign
10,7,7,6,4,10,4,1,2,malignant
6,1,1,1,2,1,3,1,1,benign
7,3,2,10,5,10,5,4,4,malignant
10, 5,5,3,6,7,7,10,1,malignant
3,1,1,1,2,1,2,1,1,benign
8,4,5,1,2,?,7,3,1,malignant
1,1,1,1,2,1,3,1,1,benign
5,2,3,4,2,7,3,6,1,malignant
3,2,1,1,1,1,2,1,1,benign
5,1,1,1,2,1,2,1,1,benign
2,1,1,1,2,1,2,1,1,benign
1,1,3,1,2,1,1,1,1,benign
3,1,1,1,1,1,2,1,1,benign
2,1,1,1,2,1,3,1,1,benign

10,7,7,3,8,5,7,4,3,malignant
2,1,1,2,2,1,3,1,1,benign
3,1,2,1,2,1,2,1,1,benign
2,1,1,1,2,1,2,1,1,benign

4. RESULTS AND DISCUSSION

4.1 Performance Measurement Terms

Accuracy: this is the percentage of the instances which are classified correctly by the classifier.

Time taken to build model: This is the time taken by the classifier to build the model to be used for classification.

4.2 10-Fold Cross Validations

10-fold cross validation is used in the field of machine learning to determine how accurately a learning algorithm will be able to predict data that it is not trained on. The training dataset is randomly partitioned into 10 groups, the first 9 groups are used for training the classifier and the other group is used as the dataset to test on. The process is repeated until all the groups have been used as testing dataset, then the classifier's performance is measured as an aggregate of all the 10 folds.

4.3 Performance Evaluation

Table 4.1: Classification Accuracy and time taken to build classification model when First feature is removed (Mitosis)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	95.9943	0.73
Adaboost (SVM)	96.1373	3.76
C4.5	94.9928	0.49
Adaboost (C4.5)	95.7082	0.24

Table 4.2: Classification Accuracy and time taken to build classification model when second feature is removed (Clump thickness)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	95.9943	0.73
Adaboost (SVM)	95.9943	3.76
C4.5	94.5637	0.49
Adaboost (C4.5)	95.1359	0.24

Table 4.3: Classification Accuracy and time taken to build classification model when third feature is removed (Marginal adhesion)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	96.1373	0.25
Adaboost (SVM)	96.5665	3.56
C4.5	94.7067	0.01
Adaboost (C4.5)	95.5651	0.19

Table 4.4: Classification Accuracy and time taken to build classification model when fourth feature is removed (Normal Nucleoli)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	95.9943	0.73
Adaboost (SVM)	95.9943	3.76
C4.5	94.5637	0.49
Adaboost (C4.5)	95.1359	0.24

Table 4.5: Classification Accuracy and time taken to build classification model when fifth feature is removed (Single EPI cell size)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	95.5651	0.22
Adaboost (SVM)	93.8484	4.66
C4.5	95.1359	0.01
Adaboost (C4.5)	94.5637	0.13

Table 4.6: Classification Accuracy and time taken to build classification model when sixth feature is removed (Bland chromatin)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	95.7082	0.08
Adaboost (SVM)	93.9914	3.42
C4.5	95.4222	0.01
Adaboost (C4.5)	94.4206	0.10

Table 4.7: Classification Accuracy and time taken to build classification model when seventh feature is removed (Bare Nuclei)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	93.5622	0.12
Adaboost (SVM)	93.4192	3.46
C4.5	93.5622	0
Adaboost (C4.5)	93.2761	0.05

Table 4.8: Classification Accuracy and time taken to build classification model when eighth feature is removed (Cell shape uniformity)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	95.9943	0.46
Adaboost (SVM)	95.9943	3.09
C4.5	94.5637	0.01
Adaboost (C4.5)	95.1359	0.22

Table 4.9: Classification Accuracy and time taken to build classification model when eighth feature is removed (uniformity of cell shape)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	94.4123	0.68
Adaboost (SVM)	94.4123	3.30
C4.5	93.1278	0.23
Adaboost (C4.5)	94.5462	0.44

Table 4.10: Classification Accuracy and time taken to build classification model when eighth feature is removed (Cell size)

Classification Techniques	Accuracy (%)	Time Taken to build model (secs)
SVM	95.9943	0.46
Adaboost (SVM)	95.9943	3.09
C4.5	94.5637	0.01
Adaboost (C4.5)	95.1359	0.22

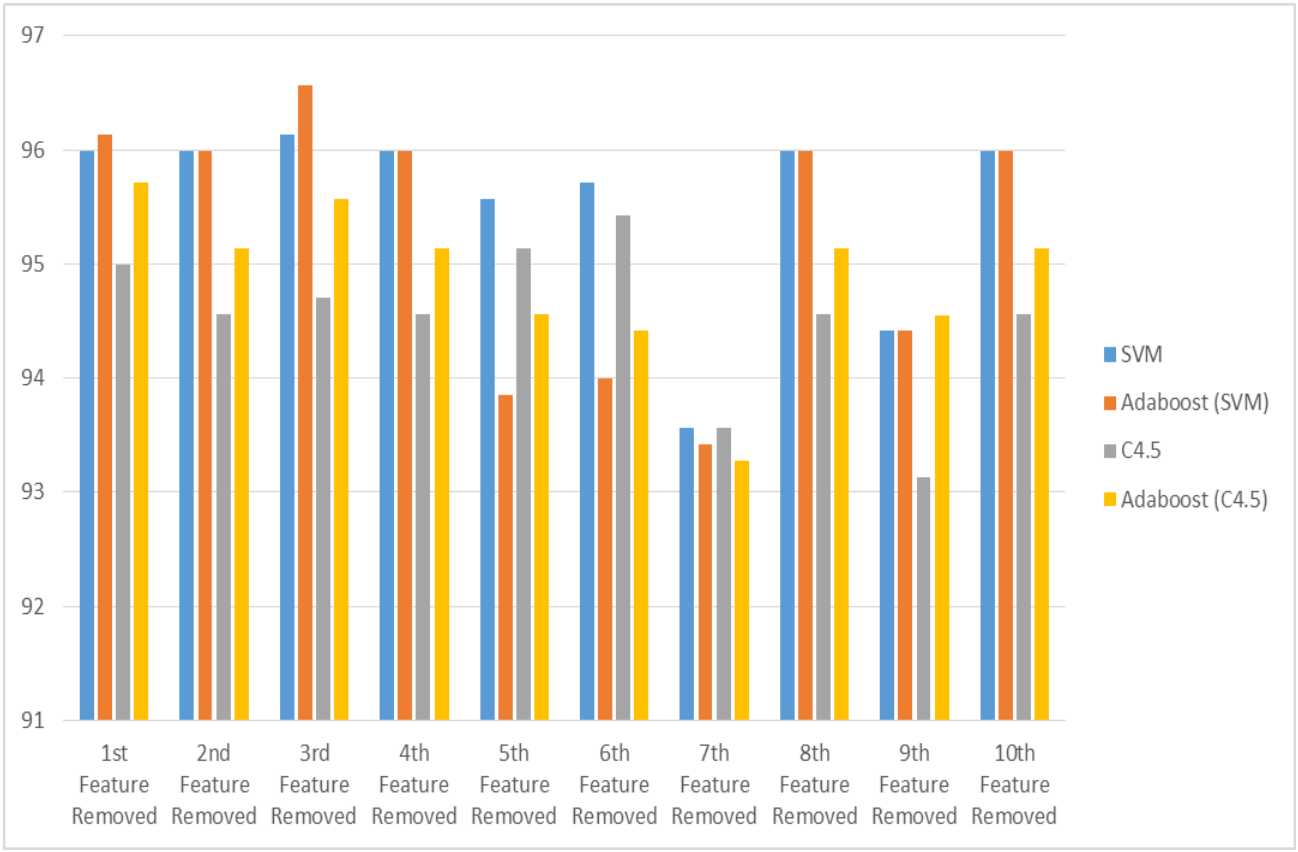

Fig. 2: Classification Accuracy

After removal of the ten (10) features, it was observed that none of the classifiers had less than 92% classification accuracy; this shows that the "cell size uniformity" attribute is the most important feature in determining whether a cancerous cell in the breast is malignant or benign. But being a medical condition where increased accuracy in classification can mean the difference between life and death, other features need to be considered.

Boosting SVM performed almost equally when one feature and two features was removed, and performed better when three features were removed but subsequently, removal of features reduced its accuracy.

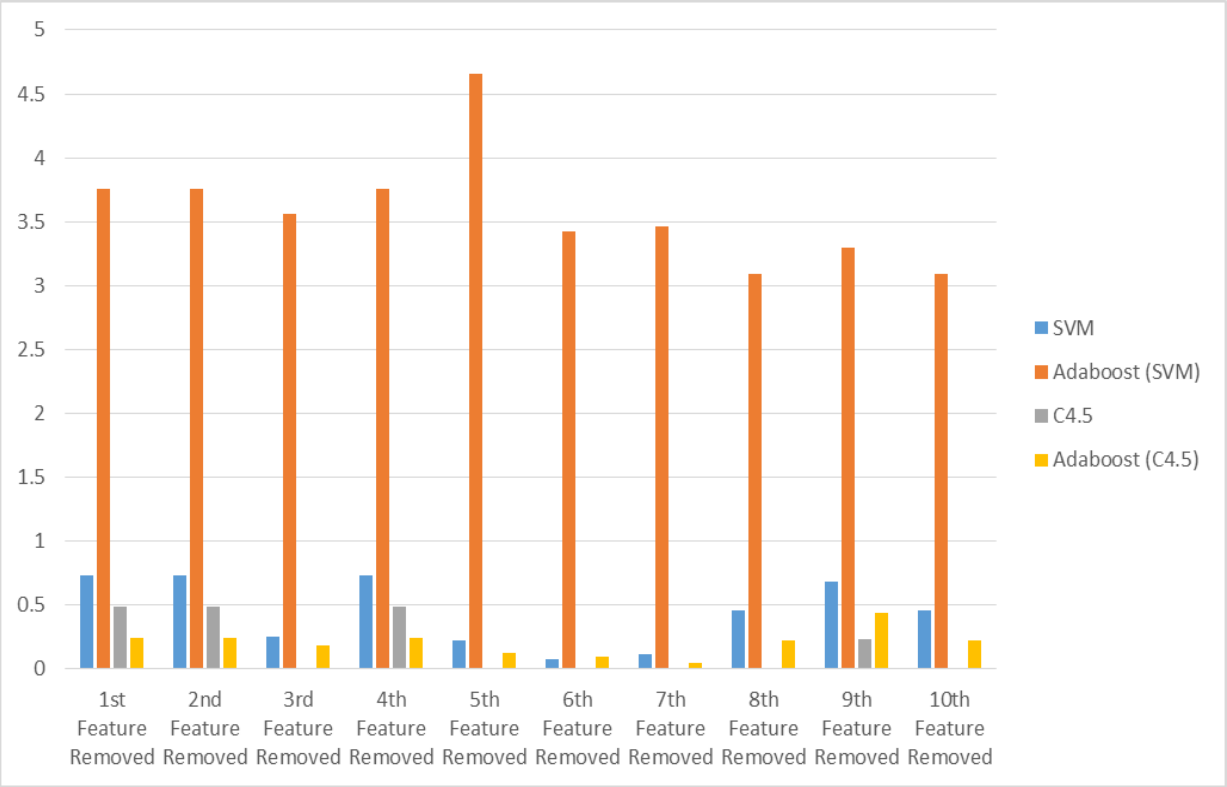

Fig. 3: Time taken to build classification model

C4.5 was extremely fast in building its classification model, thus the fastness of its boosting, while SVM took a slower time in building its model but its boosting took a much longer time. In conditions in which time is very critical such that further analysis can be done later, C4.5 can be used to determine the class of the cancerous cell. As expected, removal of more features generally reduced the time it takes for algorithms to build their models.

5. CONCLUSION

Classification of a disease like breast cancer is very crucial and thus this research work analyzed the performance of two data mining algorithms and their boosted version in the determination of breast cancer based on malignant and benign cases. Result shows that C4.5 is very fast in building its classification model, it is not as effective in classifying breast cancer as much as the boosted version of SVM. Because of the possible causes of breast cancer to human health, getting a high accuracy is of crucial importance and from all the experiments removing three features (mitosis, clump thickness and Marginal adhesion) as determined by the feature selector yields the highest classification accuracy overall from the boosted version of SVM, thus Boosting SVM with the three features removed should be the algorithm used for breast cancer classification. From the study the performance of the boosted version of SVM gave the highest accuracy, other methods that can be applied to improve the classification accuracy of SVM for a better

performance should further be investigated. Other classification algorithms such as Neural network, KNN and Naïve Bayes can also be considered to see if they can yield better high accuracy than classification approaches used in this study.

6. REFERENCES

[1]. Adeyemo, O. O., Adeyeye, T. O., "*Comparative Study of ID3 / C4.5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever"*. African Journal of Computing & ICT, 8(1), 103–112, 2015.

[2]. Arutchelvan, K., Periyasamy, R., "*Cancer Prediction System Using Datamining Techniques",* International Research Journal of Engineering and Technology, 2(8), 1179–1183, 2015.

[3]. Bindushree, D. C., "Prediction of Cardiovascular Risk Analysis and Performance Evaluation Using Various Data Mining Technioques: A Review" International Journal of Engineering Research, 5013(5), 796–800, 2016.

[4]. Chandna, D., "*Diagnosis of Heart Disease Using Data Mining Algorithm'* Inetrnational of Computer Science and Information Technologies, 5(2), 1678–1680, 2014.

[5]. Chaurasia, V., Pal, S., "*A Novel Approach for Breast Cancer Detection using Data Mining Techniques",* InternatioInterain Computer and Communication Engineering, 2(1), 2456–2465, 2014a.

[6]. Chaurasia, V., Pal, S., "*Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability"*. International Journal of Computer Science and Mobile Computing, 3(1), 10–22, 2014b.

[7]. Demigha, S., "*Data Mining for Breast Cancer Screening",* In The 10th International Conference on Computer Science & Education 65–69, 2015.

[8]. Devi, G., "*Breast Cancer Prediction System using Feature Selection and Data Mining Methods",* International Journal of Advanced Research in Computer Science, 2(976), 81–87, 2011.

[9]. Einipour, A., "*A Fuzzy-ACO Method for Detect Breast Cancer",* Global Journal of Health Science, 3(2), 195–199, 2011.

[10]. El-hasnony, I. M., El-bakry, H. M., Saleh, A. A,. "*Classification of Breast Cancer Using Soft computing Techniques"* International Journal of Electronics and Information Engineering, 4(1), 45–54, 2016.

[11]. Imandoust, S. B., Bolandraftar, M., "*Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background"*. Int. Journal of Engineering Research and Applications, 3(5), 605–610, 2013.

[12]. Karatzoglou, A., Meyer, D., Hornik, K., "*Support Vector Machines",* in R. Journal of Statistical Software, 21(9), 1–26, 2005.

[13]. Karim, H., Zand, K., "*A Comparative Survey on Data Mining Techniques for Breast Cancer Diagnosis and Prediction",* Indian Journal of Fundamental and Applied Life Sciences, 5(1), 4330–4339, 2015.

[14]. Karthikeyan, T., Ragavan, B., Kanimozhi, V. A., "*A Study on Data mining Classification Algorithms in Heart Disease Prediction",* International Journal of Advanced Research in Computer Engineering & Technology, 5(4), 1076–1081, 2016.

[15]. Kaur, E. R., Chopra, V., "*Implementing Adaboost and Enhanced Adaboost Algorithm in Web Mining"* International Journal of Adanced Research in Computer and Communication Engineering, 4(7), 306–311, 2015.

[16]. Kehinde, W., Peter, A. I., Jeremiah, A. B. Adeniran, I. O., "*Breast Cancer Risk Prediction Using Data Mining"* Transactions on Networks and Communications, 3(2), 1–12, 2015.

[17]. Khaleel, M. A., Pradhan, S. K., Dash, G. N., "*Finding Locally Frequent Diseases Using Modified Apriori Algorithm"* International Journal of Advanced Research in Computer and Coomunication Engineering, 2(10), 3792–3797, 2013.

[18]. Khamis, H. S., Cheruiyot, K. W., Kimani, S., "Application of k-Nearest Neighbour Classification in Medical Data Mining Application of k- Nearest Neighbour Classification in Medical Data Mining", International Journal of Infromation and Communication Technology Research, 4(4), 121–128 ,2014.

[19]. Kharya, S., "*Using Data mining Techniques for Diagnosis and Prognosis of Cancer Disease"* International Journal of Computer,

Engineering and Information Technology, 2(2), 2012.

[20]. Krishna, G., Nookala, M., Orsu, N., Pottumuthu, B. K., Mudunuri, S. B., "*Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification"* IJARAI) International Journal of Advanced Research in Artificial Intelligence, 2(5), 49–55, 2013.

[21]. Lebbe, A., Saabith, S., Sundararajan, E., Bakar, A. A., "*Comparative Study on Different Classification Techniques for Breast Cancer Dataset"*, International Journal of Computer Science and Mobile Computing, 3(10), 185–191, 2014.

[22]. Li, X., Wang, L., Sung, E., "*AdaBoost with SVM-based component classifiers"* Engineering Applications of Artificial Intelligence, 21, 785–795, 2008.

[23]. Majali, J., Niranjan, R., Phatak, V., Tadakhe, O., "*Data Mining Techniques For Diagnosis And Prognosis Of Cancer"* International Journal of Advanced Research in Computer and Communication Engineering, 4(3), 613–616, 2015.

[24]. Omari, A., "*A Knowledge Discovery Approach for Breast Cancer Management"*, Health Informatics International Journal, 2(3), 1–7, 2013.

[25]. Padmapriya, B., "*A survey on breast cancer analysis using data mining techniques"*, IEEE International Conference on Computational Intelligence and Computing Research, 4–5, 2014.

[26]. Porkodi, R Suganya, G.,"*A Comparative Study of Different Deployment Models in a Cloud"* International Journal of Advanced Research in Computer Science and Software Engineering, 3(5), 512–515, 2015.

[27]. Seema, Rathi, M., Mamta., "*Decision Tree: Data Mining Technique*,. International Journal of Latest Trends in Engineering and Technology (IJLTET), 1(3), 150–155, 2012.

[28]. Shajahaan, S. S., Shanthi, S., Manochitra, V., "*Application of Data Mining Techniques to Model Breast Cancer Data"* International Journal of Emerging Technology and Advanced Engineering, 3(11), 1–8, 2013.

[29]. Shouman, M., Turner, T., Stocker, R., "*Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients"*, International Journal of Information and Education Technology, 2(3), 220–223, 2012.

[30]. Shukla, S., Gupta, D. L., & Prasad, B. R. (2016). C*omparative Study of Recent Trends on Cancer Disease Prediction using Data Mining Techniques*. International Journal of Database Theory and Application, 9(9), 107–118.

[31]. Singh, D., Naveen, H., Samota, J., "*Analysis of Data Mining Classification with Decision Tree Technique"*, Global Journal of Computer Science and Technology, 13(13), 1–6, 2012.

[32]. Zorluoglu, G., & Agaoglu, M., "*Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods"* International Journal of Bioinformatics and Biomedical Engineering, 1(3), 318–322' 2015.