Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 10, No. 2, May 2019), Pages: 31-52
www.jacr.iausari.ac.ir

ADABOOST Ensemble Algorithms for Breast Cancer Classification

Moshood. A. Hambali✉[1], Yakub. K. Saheed[2], Tinuke O. Oladele[3], Morufat. D. Gbolagade[2]
1) Computer Science Department, Federal University Wukari, P.M.B 1020, Katsina-Ala Road, Wukari, Taraba State, Nigeria
2) Department of Physical Sciences, Computer Science Programme, Al-Hikmah University, P.M.B 1601, Adewole Housing Estate, Ilorin, Kwara State, Nigeria
3) Department of Computer Science, University of Ilorin, P.M.B. 1515, Ilorin-Nigeria
hambali@fuwukari.edu.ng; yksaheed@alhikmah.edu.ng; tinuoladele@gmail.com; dammyconsult@gmail.com

Manuscript ID: JACR-1807-1628 **Received: 2018/07/04; Accepted: 2019/03/13**

Abstract

With advances in technologies, different tumor features have been collected for Breast Cancer (BC) diagnosis. The process of dealing with large data set suffers some challenges which include high storage capacity and time required for accessing and processing. The objective of this paper is to classify BC based on the extracted tumor features and to develop an ADABOOST ensemble Model to extract useful information and diagnose the tumor. In this research work, both homogeneous and heterogeneous ensemble classifiers (combining two different classifiers together) were implemented, and Synthetic Minority Over-Sampling Technique (SMOTE) data mining pre-processing is used to deal with the class imbalance problem and noise in the dataset. In this paper, the proposed method involve two steps. The first step employs SMOTE to reduce the effect of data imbalance in the dataset. The second step involves classifying using decision algorithms (ADTree, CART, REPTree and Random Forest), Naïve Bayes and their Ensembles. The experiment was implemented on WEKA Explore (Weka 3.6). Experimental results show that ADABOOST-Random forest classifies better than other classification algorithms with 82.52% accuracy, followed by Random Forest-CART with 72.73% accuracy while Naïve Bayes classification is the lowest with 35.70% accuracy.

Keywords: *Breast Cancer, ADABOOST, Synthetic Minority over Sampling Technique, Random Forest, Ensemble*

1. Introduction

Cancer disease is the main cause of death among humans in many developed countries. Cancer, sometimes referred to as malignant neoplasm, is a complex disease in which a set of cells exhibit certain traits of unrestrained growth and invasion which may possibly spread (metastasize) to other parts of the body. Cancer can develop in any part of human body which eventually give rise to various kinds of cancer like lung, prostate, breast, renal, brain, gastric, rectal, colon, and, head and neck cancers among others. Cancerous cells very often invade and destroy surrounding healthy tissues and organs. BC can occur due to an uncontrollable growth of cells in the breast tissue.

The potential to prevent, diagnose and treat any of these forms of cancers require in-depth knowledge of how changes occur in their genome. Over the past few decades, classification and diagnosis of cancer patients are done based on the examination of the organs where the tumor was developed. This often result in the exhaustive physical and histopathological assessments of the organs that harbour the tumor. Also, diagnoses are only achievable either through laboratory tests which might be too costly to bear or through surgical operations which might expose the patients to different kind of risks. In some instances, some of the test results, like autopsy can be available only after the passage of time, thus causing some delay before any diagnoses or cancer classification could be performed.

The traditional method of breast cancer diagnosis is to use mammography. Though, the radiologists demonstrate considerable inconsistency in the manner of interpreting a mammogram and analysis of its result [1]. Furthermore, Elmore, et al. [2] revealed that about 90% of radiologists can recognize cancer (less than 3% of cancer cases) and 10% detected cancer which account for just about 25% of the cases. Another alternative method adopted for breast cancer diagnosis is the fine needle aspiration cytology which has more reasonable prediction accuracy. However, the correct diagnosis rate is averagely 90% [3]. In general, the main drive of all these related research works is to be able to distinguish between patients with breast cancer disease (malignant group) and the one without breast cancer (benign group). The advent of DNA microarray technology in the recent past has introduced dramatic changes into cancer research. With this new technology, it is possible to simultaneously analyze the expressions of several thousands of genes at once and relate their expression patterns to clinical phenotypes [4].

In the past, statistically related approaches were mostly used with data mining techniques in building classification models. However, the breast cancer classification problem is greatly nonlinear in nature. It is very challenging to have a good model that will take into account all the independent variables using orthodox statistical modeling techniques. Also, traditional integration of statistical techniques and data management tools are no longer sufficient for analysing the enormous collection of data [5].

Data mining plays vital role in many research areas including the medical field to predict and detect various diseases [6]. Application of microarrays has made the study and diagnosis of cancer disease a lot easier to do. Having a hint on the signals that are symptoms for the disease phenotype and its progression needs the uses of robust techniques [7] that can improve the prediction accuracy of cancer. Breast cancer can be identified through the analysis of genetic data. The human genome contains more than 10 million single nucleotide polymorphisms which will be in charge of the difference that lies among human beings. Lots of researchers have tried to employ artificial intelligence and machine learning related approaches for predicting breast cancer diseases. Various machine learning techniques are reported for cancer classification, among them is support vector machine, k-nearest neighbour, neural network techniques and decision tree and so on.

Ensemble methods are one of machine learning models that merge multiple learning models together with the objective of improving predictive performance [8]. The main benefit of the ensemble model is that it is unlikely for all the models applied to commit the same error. Ensemble methods have been used extensively for medical diagnosis [9-10]. The ADABOOST technique is one of the widely used ensemble methods in

machine learning due to its minimum value of error rate. It performed excellently in the dataset with low noise [11-12].

Ensemble classification can be grouped into two: homogenous and heterogeneous techniques. Homogenous ensemble classification consists of only one classifier, while heterogeneous ensemble classification involves different classifiers. In this paper, both homogeneous and heterogeneous ensemble classifiers were implemented, and Synthetic Minority Over-Sampling Technique (SMOTE) data mining pre-processing was used to deal with the class imbalance problem and noise in the dataset.

The remaining part of this paper is organized as follows: Section 2 presents the related works on this study, Section 3 discusses preliminary knowledge to the study. In section 4, the methodology employed was described. Section 5 presents analysis of results and discussion. Finally, conclusion was presented in Section 6.

2. Related Works

A lot of researchers have developed various algorithms to aid healthcare experts in accurately diagnosing breast cancer. Wang et al. [13] studied a Support Vector Machine (SVM)-based ensemble learning algorithm for breast cancer diagnosis, in order to reduce the diagnosis variance and increase diagnosis accuracy. Twelve different SVMs, based on the proposed weighted Area under the Receiver Operating Characteristic Curve Ensemble (WAUCE) approach, were hybridized. The proposed WAUCE model reduces the variance by 97.89% and increased accuracy by 33.4% compared to the best single SVM model on the SEER dataset.

Nidhi, Mukesh and Saveta [14] used four different classification algorithms such as, J48, REPTree, Random Forest and Random Tree to build a classification model which was tested on the dataset taken from UCI for the purpose of diagnosing cancer, based on some diagnostic measurements integrated into the dataset. The maximum accuracy was 95.0791% for Random Forest and 93.4974%. Abed et al. [15] suggested a hybrid classification algorithm based on Genetic Algorithm (GA) and K-nearest neighbour (KNN). GA was used as an optimization technique for KNN by selecting the best features as well as the optimization of the k value, while KNN was used for classification purpose. The evaluation results of the algorithm achieved 99% accuracy.

The decision tree is one of the common classifiers used in the medical domain. For instance, in 2005, [16] predicted breast cancer survivability using classification and regression trees (CART) on SEER medical databases. Their results revealed that the decision tree algorithm was capable of extracting knowledge from the SEER dataset. Rajesh & Sheila [17] used the C4.5 classification algorithm to classify SEER breast cancer dataset into either “Carcinoma in situ” (beginning) or “Malignant potential” group. Syed et al. [18] employed decision tree algorithms such as the Random tree, ID3, CART, C4.5, and Naive Bayes to predict breast cancer. In this work, the experimental result showed that Random tree algorithm outperforms other algorithms used. Subasin et al. [19] also applied data mining techniques to diagnosis and prognosis of breast cancer disease using supervised learning algorithms such as C5.0, ID3, APRIORI, C4.5, and Naïve Bayes. The results of their experiment shows that the C4.5 algorithm performs better compared to other algorithms with the highest precision rate.

Furthermore, efforts are also made based on ADABOOST to improve the performance and generalization of tree approaches by using an ensemble of decision tree approaches. The boosting [20], bagging [21] and ADABOOST algorithms are

popular approaches to constructing a random forest rather than a classifier based on one tree. The ADABOOST technique has gained more attention among other ensemble methods in machine learning, due to its minimal error rate and it performs excellently on the noisy dataset [11-12]. ADABOOST algorithm combines a set of weak classifiers in order to produce a model with better prediction outcomes [11]. Consequently, a lot of researches that were reported in the literature have successfully utilized ADABOOST algorithm to solve classification tasks which include face recognition, video sequences, and signal processing systems. For instance, Zhou & Wei [22] used ADABOOST algorithm to extract the 20 most significant features from the XM2VT face database. The results of their experiment indicated that the ADABOOST algorithm reduces computation time by 54.23 %. Moreover, Sun, Wang & Wong [23] utilized the ADABOOST algorithm on UCI Machine Learning database to extract high-order pattern. Their results revealed that ensemble classifiers have better classification accuracy compared to the High-order Pattern and Weight of evidence Rule based (HPWR) classifiers alone. Though, some research studies have exploited ADABOOST and random forests as data mining tools for predictions on medical databases. Ram´On, Genesrf & Varselrf [24], used random trees and bootstrap samples to perform gene selection and classification on 10 different cancer-related datasets. Random forests perform remarkably in microarray data analysis due to its robustness, even in situations where the predictor variables are characterized by noisy data. A major drawback of the random forest approach is that if the data features are correlated, it tends to be biased toward the smaller group [25]. This is the reason why it is imperative to use SMOTE algorithm to solve the problem of data imbalance and reduce the possibility of bias toward minority class in the dataset. Hambali & Gbolagade [7] applied a hybrid of Synthetic Minority Over-Sampling Technique (SMOTE) and Artificial Neural Network (ANN) to diagnose ovarian cancer from the public available ovarian dataset. Their study shows that the performance of Neural networks in the cancer classification can be enhanced by employing SMOTE pre-processing algorithm to lessen the influence of data imbalance in the dataset.

From several works of literature reviewed, it was observed that a lot of challenges are noted to be associated with dealing with large volumes of data. Many of these data sets consist of features that are irrelevant, redundant which upsurge the search time and subsequently result in difficulty to accurately classify medical datasets with class imbalance [7]. Furthermore, the presence of noise and unrelated features in large data sets usually result in complicated process analysis. For instance, microarray data holds thousands of genes with only a few feature samples that are relevant in the classification process [26]. There are lots of preprocessing and feature selection algorithms presented in the literature for dimensionality reduction of highly dimensional datasets. However, most of these approaches do not offer reliable results and some of the relevant features are missing too [27].

Most of the single classifier techniques have hitches of being computationally costly and high complexity on hefty datasets. Particularly, the classification methods do not produce reliable and consistent results for huge datasets which makes some single classifier systems inefficient and undependable [28]. For instance, the decision tree algorithm effectively handles the interaction between variables very well but have challenges in handling linear relations between variables [29]. Ensemble classifier has become a popular approach recently used in machine learning and pattern recognition. Basically, it is an approach that comprises integration of multiple classifier results. The

core purpose of the ensemble method is to augment classification accuracy by weighing numerous individual classifiers and then combine them as a single classifier that perform better than every individual classifier [28, 30-31]. Hence, to solve this problem a hybrid of SMOTE over sampling technique as data pre-processing was proposed and the ensemble ADABOOST approach was used for the classification.

3. Preliminary Knowledge

Cancer research is one of the prominent research areas in the medical research. It has gained a lot of attention. Earlier detection and precise predictions of various tumor types have a great impact in treatment and reduce the mortality rate of the cancer patients. In the earlier days, cancer detection had always been morphological and clinically based. These methods of cancer classification are characterized by several faults in their diagnostic capability. However, data mining techniques and machine learning has helped to correctly classify thousands of genes simultaneously without much stress. The following sections are brief descriptions of the algorithms employed in this research work to classify breast cancer.

3.1 Decision Tree

Decision trees (DT) algorithms are one of the famous classification techniques that are becoming progressively more popular in the data mining domain. Common decision tree algorithms include ID3, C4.5, C5 [33-34], and CART [35]. Generally, the DT technique recursively split data into branches to build a tree for the purpose of increasing the prediction accuracy. They performed this task by using mathematical algorithms (such as information gain, Gini index, and Chi-squared test) to find a variable and corresponding threshold for splitting the input data into two or more subgroups. This step was performed repeatedly at each leaf node until the complete tree is built. Decision trees are easy to build and comprehend due to their hierarchical structure. They are capable of model complex functions.

The decision tree classifier has two phases [36]:

i. Growth phase or Build phase.
ii. Pruning phase.

In the first phase, the tree is built by recursively splitting the training dataset based on the best criterion until all or most of the data belonging to each of the partitions have the same class label. Data overfitting may occur at this stage [37].

In the pruning phase, consecutive branches are minimized so that the tree is built to adequately generalize the model. Pruning generally involves bottom-up or top-down traversal of the decision tree while removing the noisy and outlier nodes to improve certain criteria in the tree. Popular pruning strategy that is commonly used includes cost-complexity pruning, reduced error pruning, minimum error pruning, minimum descriptive length pruning, minimum message length pruning and critical value pruning [38]. The pruning phase handles the problem of overfitting the data in the decision tree. Therefore, classification accuracy increases in the pruning stage. Pruning phase accesses the completely grown tree only. While multiple passes over the training data are required in the growth phase. The time complexity for pruning in the decision tree is less compared the one required to build the decision tree. Figure 1 shows the generic pseudo code of decision tree algorithm.

GenDecTree(Sample S, Features F)
Steps:

1. ***If****stopping_condition(S, F) = true* ***then***
   a. *Leaf = createNode()*
   b. *leafLabel = classify(s)*
   c. ***return*** *leaf*
2. *root = createNode()*
3. *root.test_condition = findBestSpilt(S,F)*
4. *V = {v | v a possible outcomecfroot.test_condition}*
5. ***For each*** *value v Є V:*
   a. S_v *= {s | root.test_condition(s) = v and s Є S };*
   b. *Child = TreeGrowth* (S_v, *F);*
   c. *Add child as descent of root and label the edge {root → child} as v*
6. ***return*** *root*

Figure 1: Pseudocode of Decision Tree Algorithm

3.2 Classification and Regression Tree (CART)

Breiman et al. [35] were the first to introduce the CART algorithm. The CART is based on Hunt's algorithm. It can process both categorical and continuous attributes to build a decision tree. Also, it takes care of missing values and builds the decision tree using the Gini Index as attribute selection measure. CART splits training datasets into binary, therefore, it generates binary trees. Gini Index measure is not involved in probabilistic assumptions used in ID3 and C4.5. However, CART uses cost-complexity pruning to eliminate the erratic branches from the decision tree in order to improve the classification accuracy.

$$\text{Gini Index} = \quad 1 - \sum j \;\; p^2 j \qquad (1)$$

The tree grows in CART algorithm by carrying out an exhaustive search of all variables for each decision node and all possible splitting values, then selecting the optimal split. It generates an estimate for the misclassification rate.

3.3 Reduced Error Pruning Tree (REPTree)

REPTree algorithm is built on the principle of computing the information gain with entropy and minimizing the error generated from variance with back-fitting [39]. This method has the benefit of reducing error pruning as the complexity of the decision tree model decreases and also, error generation from variance is minimized [40]. REPTree is one of the fast decision tree learners and can only process numeric attributes once. The missing data is dealt with by splitting the corresponding values into pieces.

In the growth phase, REPTree employs the regression tree logic and generates multiple trees in several iterations. Consequently, it chooses the optimal one for all

generated trees and considers it as the substance that represents the generated trees. In the pruning phase, the predictions made on the tree was measured by mean square error.

3.4 Alternating Decision Tree (ADTree)

ADTree is one of the machine learning classification considered to be another semantic representation of the decision tree. It is a generalization of data structure and decision tree [41]. Furthermore, each of the decision nodes in ADTree is replaced by two nodes (one for prediction node symbolized by an ellipse, and the other one for splitter node denoted by a rectangle). This tree predicts the nodes in the leaves and roots. In the decision tree, an instance is usually traversed along the path of the tree from the root to the leaves. ADTree is distinct from decision trees in the manner that classification is associated with the path, not with the label on the leaf. But, it is the symbol of summation of the prediction along the path.

3.5 Random Forests Algorithm

Random forest (RF) is a famous ensemble learning technique which is known to be a powerful technique in pattern recognition and machine learning for high dimensional dataset [42] and skewed classification problems [43]. RF is a family of classification methods, which rely on the combination of individual decision trees to build classifiers that employ CART algorithms [35, 44]. RF is also known as a generic principle of randomized ensembles of decision trees [43]. The base learner of RF (basic unit) is a binary tree built using recursive partitioning. An individual tree is built from training set by splitting the tree recursively into homogeneous or near homogeneous terminal (leaf) nodes partition. A good binary split ensures the improvement of homogeneity in the daughter nodes by traversing data from a parent tree-node to its two daughter nodes. RF comprises of hundreds to thousands of trees, where the individual tree is full-grown by applying a bootstrap sample on the original data. RF trees are different from the CART as the RF growth involve two stage non-deterministic randomization procedure. Apart from the randomization involved in the growing of the tree using a bootstrap sample of the original dataset, another stage of randomization is considered at the node level when growing the tree. RF selects a random subset of variables at each node of each tree, instead of splitting a tree node using all variables, and only those variables selected are used as candidates to determine the best split for the node. The aim of this two-level randomization is to de-correlate trees in order that the forest ensemble will yield a low variance model.

The basic element in these levels of randomization is the number of t tree and a random vector (Dt) using bootstrap sample are generated independently from the previous random vectors but with the same distribution, and a tree is grown using the training set and Dt.

3.6 Naïve Bayesian

Naive Bayesian classification is founded on the Bayesian theorem of posterior probability. It is a model that works excellently when the predictors contain independent classes. Although, sometimes it work well with predictors that have no distinct independent class. The Naive Bayesian method has two phases of classified data. The first stage involves the training (or learning) stage using the training input data to evaluate the parameters of a probability distribution, with the assumption that predictors are conditionally independent. The prediction stage is the second phase, where the

classifier predicts any unfamiliar data (test dataset) and estimates the posterior probability of each of the classes from the sample. Afterward, the test dataset is classified according to the largest posterior probability. The common functions used for tuning Naive Bayesian classification include the Kernel Density and Gaussian distribution Estimation functions. The function to be used is determined by the nature of the dataset. The Naïve Bayesian algorithm is presented in figure 2.

Input: *TS: training set, TS = u_i (i = 1, 2, …, n),*
Output: *Class label A and B*
Steps:

1. *Given training dataset TS which consists of genes belonging to different class say class A and B.*
2. *Compute the prior probability of class A = nob of features of class A / total nob of genes*
   Compute the prior probability of class A = nob of features of class B / total nob of genes
3. *Find n_i, the total nob of frequent features of each class.*
   n_a = the total nob of frequent features of class A
   n_b = the total of frequent features of class B
4. *Find conditional probability of occurrence of key gene given a class*
   $P_{(feature1/\ class\ A)} = geneCount / n_i(A)$
   $P_{(feature1/\ class\ B)} = geneCount / n_i(B)$
   $P_{(feature2/\ class\ A)} = geneCount / n_i(A)$
   $P_{(feature2/\ class\ B)} = geneCount / n_i(B)$
   " " "
   $P_{(featuren/\ class\ B)} = geneCount / n_i\ (B)$
5. *Avoid zero frequency problems by applying uniform distribution*
6. *Classify a new gene C based on the probability P(C/feature).*
   a) *Find* $P_{(A/feature)} = P_{(A)} * P_{(feature1/classA)} * P_{(feature2/classA)} * \ldots * P_{(featuren/classA)}$
   b) *Find* $P_{(B/feature)} = P_{(B)} * P_{(feature1/classB)} * P_{(feature2/classB)} * \ldots * P_{(featuren/classB)}$
7. *Assign gene to class that has higher probability*

Figure 2: Pseudocode of Naïve Bayesian Algorithm

3.7 ADABOOST

ADABOOST is one of the famous ensemble methods that show the ability to significantly augment the prediction accuracy of the weak learner algorithm. It is a successor of the boosting algorithm that combines a set of weak learning algorithms to build a model with better prediction outcomes. ADABOOST ensemble method has gained a lot of attention among the machine learning techniques due to its low error rate and performing excellently in noise data set [11-12].

The additional benefit of ADABOOST is that it requires fewer input parameters and little or no prior knowledge of the weak learner. With this, several researchers have successfully employed ADABOOST algorithm to proffer solution to classification problems such as object detection, which include face recognition, video sequences, and signal processing systems.

The ADABOOST algorithm is initiated by setting the weight of the training set. The training set (u_1, v_1), … (u_n, v_n) where each u_i belongs to instance space U, and each label v_i is in the label set V, which is equal to the set of {-1,+1}. It assigns the weight on the training example i on round t as $D_t(i)$. The same weight will be set at the starting point ($D_t(i)$=1/N, i=1,…,N). Then, the weight of the misclassified example from base learning

algorithm (called a weak hypothesis) is increased to concentrate the hard examples in the training set in each round. The ADABOOST algorithm is presented in figure 3.

***Input**: TS: training set, TS = u_i (i = 1, 2, … , n), labels $v_i \in V$*
t: Iteration number
Steps:
1. *Assign TS sample (u_1, v_1), ... , (u_n, v_n); $u_i \in U$, $v_i \in$ {-1,+1}*
2. *Initialise the weights of $D_1(i)$ = 1/N, i = 1,…,N*
3. *for t = 1, ..., T*
4. *Train weak learner using distribution D_t*
5. *Get weak hypothesis h_t: $U \rightarrow$ {-1, +1} with its error: $\varepsilon_t = \sum_{i=h_t(u)\neq v_i}^{n} D_t(i)$*
6. *Update distribution D_t: $D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_t\ T_t(u_t))}{C_t}$*
7. *Next t that, t +1*
8. *Output the final hypothesis: $H_{(u)}$ = sign $[\sum_{t=1}^{T} \propto_t h_t(u)]$*

Figure 3: Pseudocode of ADABOOST Algorithm [45]

Where $\propto_t = {}^1/_2 \ln\left[\frac{P_{+1} - P_{-1}}{P_{-1} + P_{-1}}\right]$, (2)

C_t is the normalization constant (select so that D_{t+1} will give a distribution). α_t is used to enable the generalization of the result and also offer a solution to the problem of overfitting and noise sensitive problems [46]. *P* denotes class probability estimate that builds the real value of $\alpha_t h(u)$.

Hence, the final hypothesis $H(u)$ produces a weighted majority vote of *t* weak hypotheses where it is the weight assigned to h_t. Furthermore, ADABOOST can also handle numerical class dataset apart from common binary class usually used in literature [47].

4. Methodology

In this paper, the proposed method has two steps. The first step employs SMOTE to reduce the effect of data imbalance in the dataset. The second step involves classification using decision tree algorithms (ADTree, CART, REPTree and Random Forest) and Naïve Bayes. Thereafter, ADABOOST Ensembles of those aforementioned algorithms were also implemented to compare their performance with single algorithm of decision trees and Naïve Bayes. The framework of the proposed method is shown in figure 4. WEKA Explore (Weka 3.6) was used to implement these algorithms.

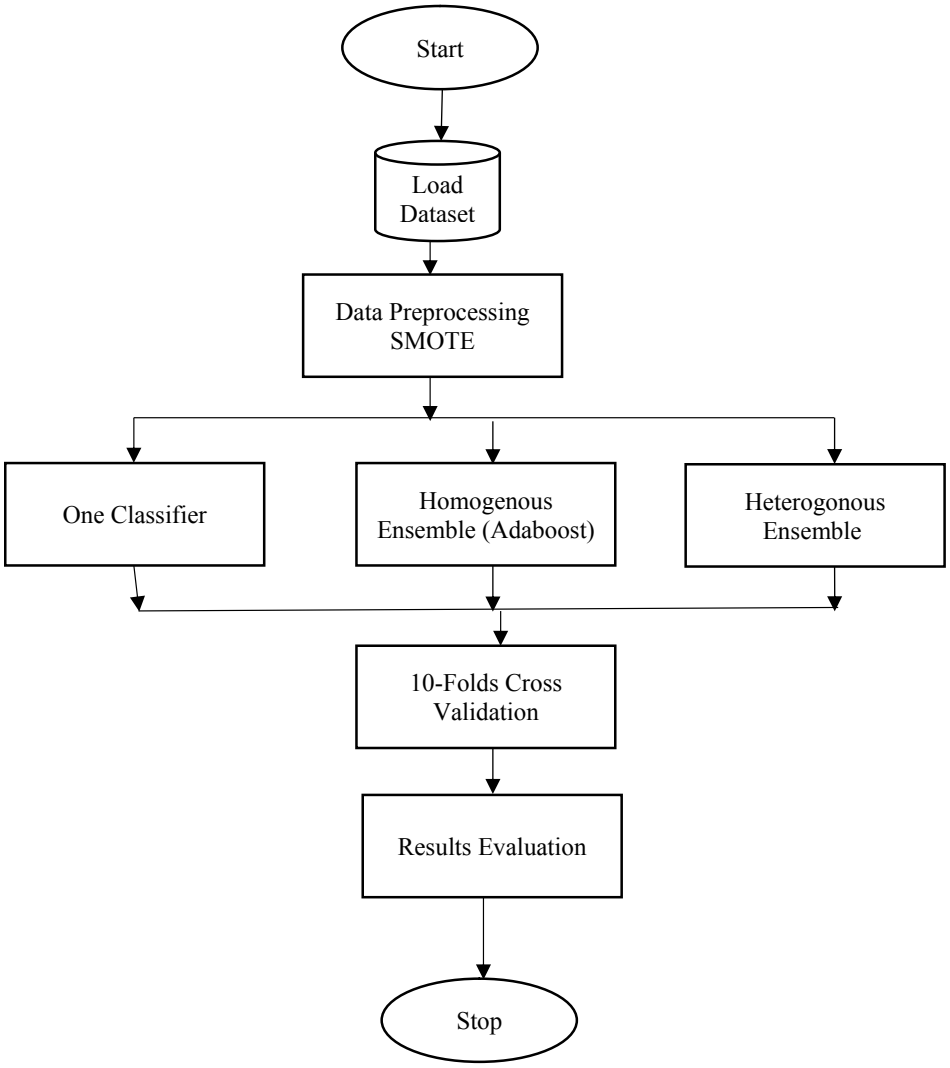

Figure 4: Proposed Framework

5. Results Analysis

In this research work, five different classification algorithms and their ensembles were proposed for the purpose of diagnosing publicly available breast cancer dataset. The dataset was first pre-processed with SMOTE algorithm before classification algorithms were employed. Performance evaluation of the models were carried out by using 10-fold cross-validation test option based classification accuracy, error reports, F-measure, ROC area and execution time. Tables 1 - 3 summaries the results of the experiments carried out.

5.1 Dataset

To evaluate the proposed approach, the experiments were carried out using gene expression profile dataset obtained from [32]. The breast cancer dataset used consist of 24, 481 genes and 97 instances with two class labeled (relapse and non-relapse). There is an enormous difference between the genes' number and the samples' number in the

selected dataset, which means that the experiment reveals the challenge of effectively dealing with such varying dimensionalities.

5.2 Performance Evaluation Metrics

The performance of classifier algorithms are measured based on the following metrics:

The confusion matrix is used to evaluate a classifier as illustrated in table 1. The columns specify the predicted class and the rows show the actual class. In the confusion matrix, True Negative (TN) is the number of negative samples correctly classified, False Positive (FP) is the number of negative samples incorrectly classified as positive, False Negative (FN) is the number of positive samples incorrectly classified as negative and True Positive (TP) is the number of positive samples correctly classified.

Table 1: Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	**TN**	**FP**
Actual Positive	**FN**	**TP**

From the confusion matrix in table 1, the expressions for Accuracy, FP rate, Recall, and Precision are derived and are presented in equations 3, 4, 5 and 6.

$$Accuracy = \frac{(TP + TN)}{(TP+FP+TN+FN)} \quad (3)$$

Precision: Precision or Positive Predictive Value (PPV) is calculated as in equation 4.

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (4)$$

$$\text{TP Rate} = \text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

F-Measure: The F-measure or F-score is the harmonic mean between precision and recall.

$$\text{F-Measure} = \frac{2(Precision-Recall)}{(Precision+Recall)} \quad (6)$$

Time Taken to Build the Model (TTBM): This is referred to as the time required to complete training or modeling of a dataset. It is represented in seconds.

Receiver Operating Curve (ROC): these are the curves used to compare the usefulness of tests.

To determine error rates (equations 7 - 11)in predicted value, let P^N represent a set of test data of the form (t_1, r_1), $(t_2, r_2)\ldots (t_p, r_p)$, such that t_i is n-dimensional test tuples with corresponding values of r_i, for a response value, r, and p is the number of tuples in P^N.

Mean Absolute Error (MAE): Mean absolute error is the average of the difference between the predicted and the actual value in all test cases. It is the average prediction error.

$$\text{Mean Absolute Error (MAE)} = \sum_{i=1}^{p} |r_i - r_i^{\mathrm{T}}| \quad (7)$$

Root Mean Squared Error (RMSE): Mean-squared error is one of the famous methods for measures of success for numeric prediction. This value is computed by

taking the average of the squared differences between each computed value and its corresponding correct value. The mean squared error is simply the square root of the mean squared error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\frac{\sum_{r=1}^{p}(r_i - r_i^{\mathrm{T}})^2}{p}} \quad (8)$$

Relative Absolute Error (RAE): Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.

$$\text{Relative Absolute Error (RAE)} = \frac{\sum_{i=1}^{p}|r_i - r_i^{\mathrm{T}}|}{\sum_{i=1}^{p}|r_i - \bar{\bar{r}}|^2} \quad (9)$$

Root Relative Squared Error (RRSE): Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root-mean-squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted value.

$$\text{Root relative squared error (RRSE)} = \sqrt{\frac{\sum_{i=1}^{p}(r_i - r_i^{\mathrm{T}})^2}{\sum_{i=1}^{p}(r_i - \bar{\bar{r}})^2}} \quad (10)$$

where r_i^{T} is the predicted value, $\bar{\bar{r}}$ is the mean value for r_i's of the training data, that is

$$\bar{\bar{r}} = \frac{\sum_{i=1}^{p} r_i}{p} \quad (11)$$

Table 2: Comparison of Evaluation Measure for Single Classifiers

Performance Metrics	Naïve Bayes	ADTree	Random Forest	REPTree	CART
TTBM (Sec)	4.66	68.61	2.27	11.27	52.94
Accuracy (%)	35.70	**73.43**	72.73	69.93	69.23
Mean absolute error (MAE)	0.64	0.31	0.36	0.37	0.32
Root mean squared error (RMSE)	0.80	0.47	0.42	0.46	0.52
Relative absolute error (%)	140	67.13	77.45	81.33	70.08
Root relative squared error (%)	167.42	97.77	88.13	96.69	107.9
F-Measure	0.20	0.72	0.70	0.68	0.69
ROC Area	0.50	0.74	0.78	0.70	0.67

Table 2 reveals that Random Forest has the best time taken to build the model 2.27 Seconds while the ADTree has the worst time taken to build the model 68.61 Seconds.

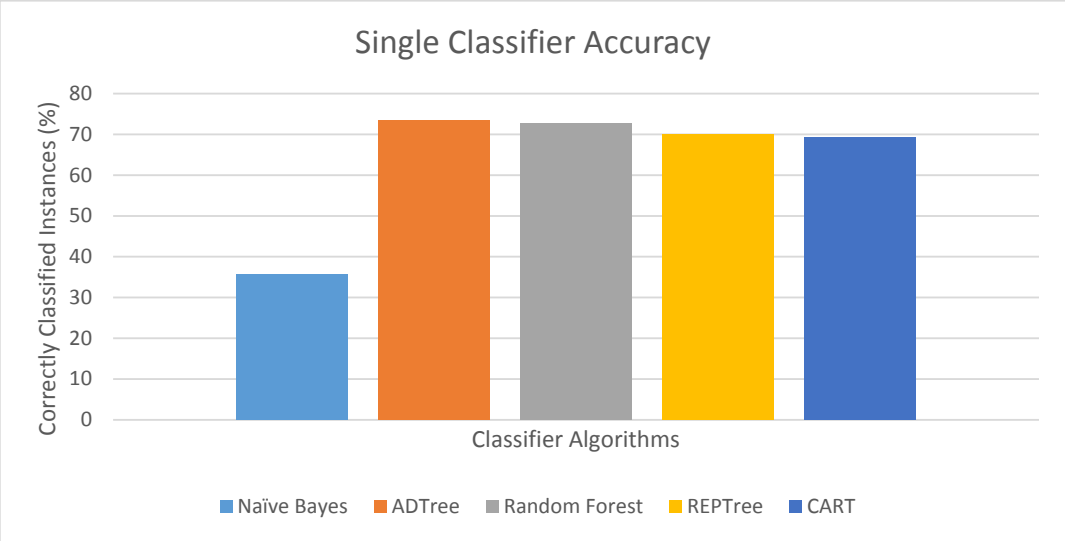

Figure 5: Prediction Accuracy for single classifiers

Figure 5 shows the prediction accuracy for single classifiers. Out of the five single classifiers used in this research work, ADTree predicts better than other classification algorithms with 73.43% accuracy, followed by Random Forest with 72.73%. While Naïve Bayes prediction is the lowest with 35.70%.

Figure 6 shows the error rates reported for the single classifiers, ADTree has Mean Absolute Error (MAE) of 0.31 and Root Mean Squared Error (RMSE) of 0.47 respectively. This shows minimal error reported during the prediction processes while Naïve Bayes has the high error rate of 0.64 and 0.80 MAE and RMSE respectively.

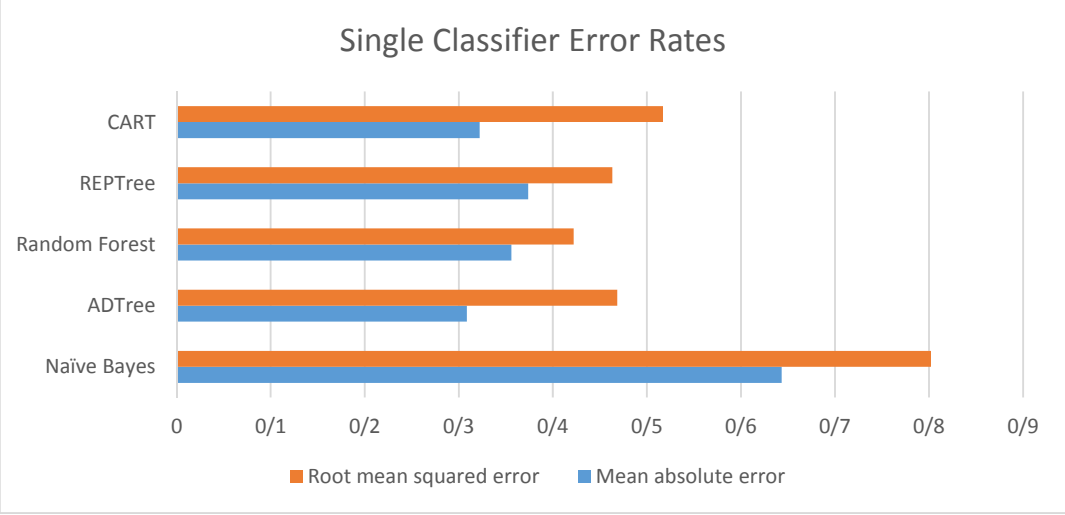

Figure 6: Error Rate for Single Classifier

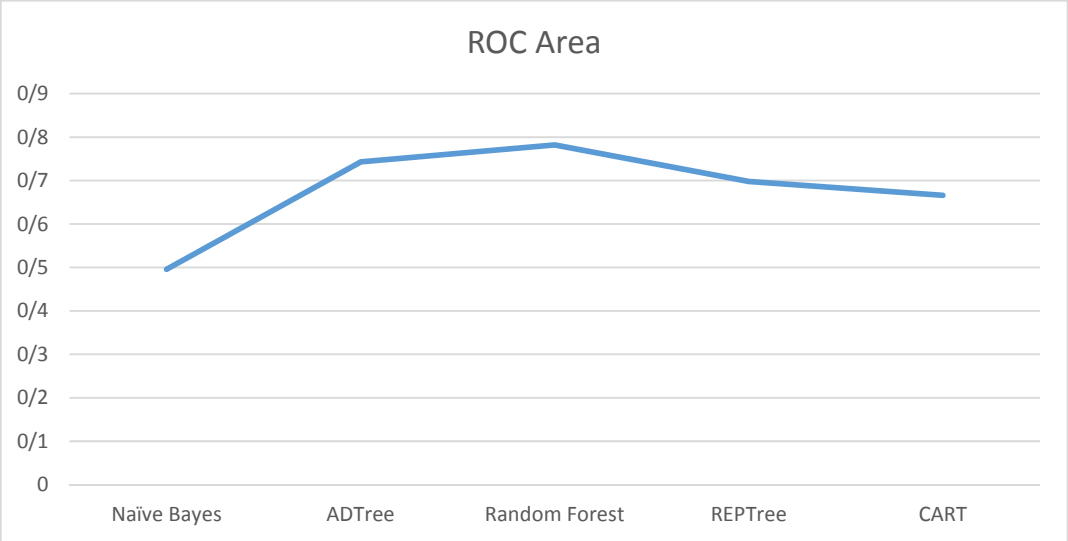

Figure 7: ROC Area for Single Classifiers

Figure 7 shows the ROC Area curve for single classifiers. In figure 7, Random Forest recorded a high ROC area with 0.78 followed by ADTree with 0.74 while Naïve Bayes has the lowest ROC area of 0.50. Therefore, the ADTree performs better than other algorithms when compared with Classification accuracy and error rates, while Random Forest performs better when using ROC, F-Measure and time taken to build the model metrics.

Table 3: ADABOOST Ensemble Classifiers

Performance Metrics	**AB-Naïve Bayes**	**AB-ADTree**	**AB-Random Forest**	**AB-REPTree**	**AB-CART**
TTBM (Sec)	20.37	33.27	12.57	67.09	300.59
Accuracy (%)	35.66	73.43	**82.52**	77.62	77.62
Mean absolute error (MAE)	0.64	0.31	**0.21**	0.25	0.24
Root mean squared error (RMSE)	0.80	0.47	0.42	0.45	0.45
Relative absolute error (%)	139.98	67.13	45.90	55.17	51.16
Root relative squared error (%)	167.42	97.77	87.13	93.07	94.28
F-Measure	0.20	0.72	0.81	0.76	0.76
ROC Area	0.50	0.74	0.86	0.78	0.83

Table 3 shows that ADABOOST-Random Forest has the best time taken to build the model with 12.57 Seconds while the ADABOOST-CART has the worst time taken to build the model with 300.59 Seconds. ADABOOST-Random Forest has a high F-measure score of 0.81 while ADABOOST-Naïve Bayes has the worst F-measure score of 0.2.

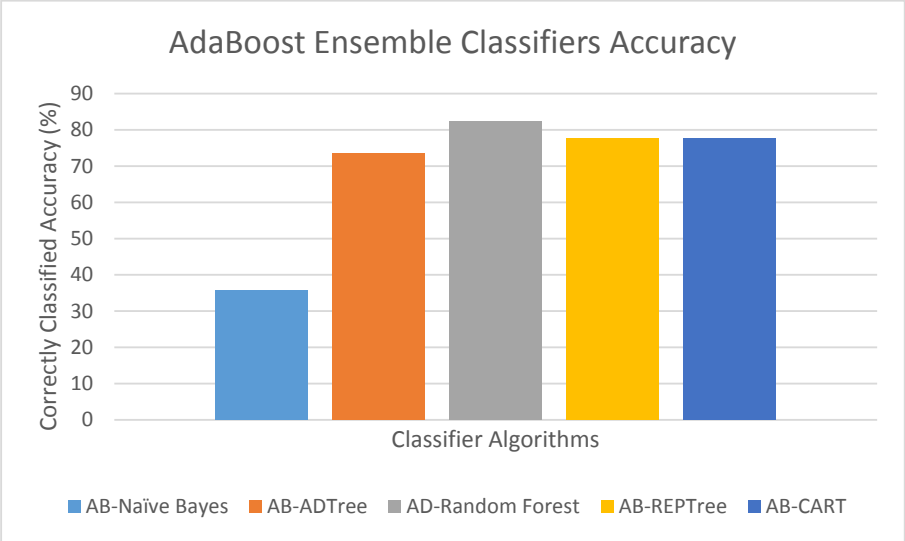

Figure 8: Accuracy for ADABOOST Ensemble Classifiers

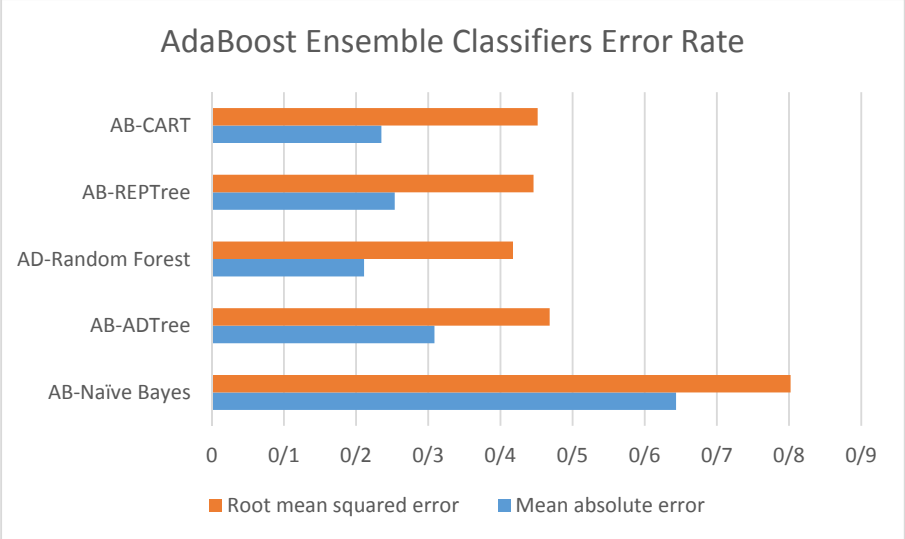

Figure 9: Error Rate for ADABOOST Ensemble Classifiers

From Figure 8, ADABOOST-Random Forest predicts better than other classification algorithms with 82.52% prediction accuracy, followed by ADABOOST-REPTree with 77.62% prediction accuracy. While ADABOOST-Naïve Bayes prediction is the lowest at 35.66%.

Figure 9 shows the error rates reported for the ADABOOST Ensemble classifiers, with the lowest error rate of 0.21 and 0.42 MAE and RMSE respectively reported for ADABOOST-Random Forest classifier. This shows minimal error reported during the prediction processes, while ADABOOST-Naïve Bayes has a high error rate of 0.64 and 0.80 MAE and RMSE respectively which is the same as the Naïve Bayes single classifier.

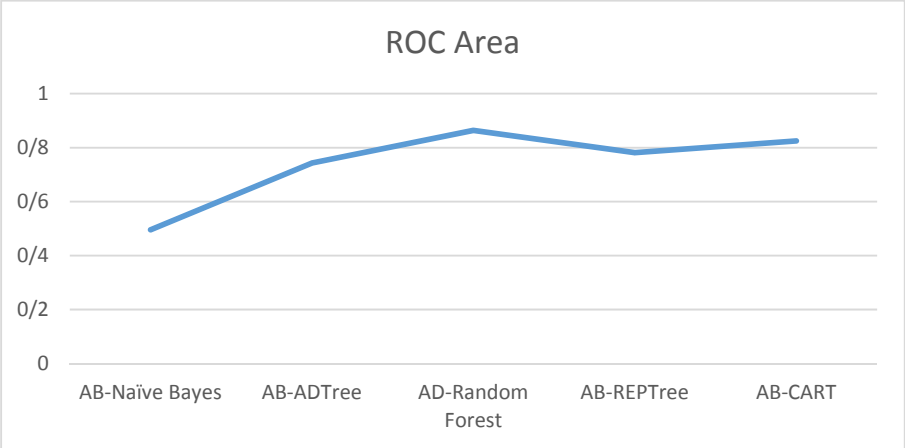

Figure 10: ROC Area for ADABOOST Ensemble Classifiers

Figure 10 shows ROC Area curve for ADABOOST Ensemble classifiers. ADABOOST-Random Forest has the highest ROC area with 0.86 followed by ADABOOST-CART with 0.83 while ADABOOST-Naïve Bayes has the lowest ROC area of 0.50 which is also the same with Naïve Bayes single classifier. Therefore, the ADABOOST-Random Forest performs better than other algorithms in terms of all metrics considered.

Table 4: Heterogeneous Ensemble Classifiers

Performance Metrics	**NB + ADTree**	**NB + RF**	**ADTree + RF**	**RF + REPTree**	**RF +CART**	**ADTree + REPTree**	**ADTree + CART**	**NB + REPTree**	**NB + CART**	**CART + REPTree**
TTBM (Sec)	33.06	35.47	417.63	8.59	8.34	371.86	610.11	42.53	55.16	62.17
Accuracy (%)	64.34	64.34	59.44	72.03	**72.73**	69.93	69.93	64.34	64.34	65.73
Mean absolute error (MAE)	0.48	0.46	0.40	0.37	0.36	0.40	0.45	0.46	0.46	0.45
Root mean squared error (RMSE)	0.49	0.48	0.55	0.45	0.45	0.46	0.47	0.48	0.48	0.48
Relative absolute error (%)	104.37	99.49	87.46	79.70	78.02	86.16	98.68	99.86	99.86	97.75
Root relative squared error (%)	101.72	100.7	114.69	94.62	93.80	96.88	98.70	100	100	100.42
F-Measure	0.50	0.50	0.59	0.71	0.72	0.69	0.68	0.50	0.50	0.58
ROC Area	0.48	0.49	0.61	0.70	0.69	0.65	0.63	0.48	0.48	0.54

Table 3 shows the results of Heterogeneous Ensemble classifiers with the best time taken to build the model of 8.34 and 8.59 Seconds for Random Forest-CART and Random Forest-REPTree Ensemble respectively. While the ADTree-CART has the worst time taken to build the model with 610.11 Seconds. Random Forest-CART has a high F-measure score of 0.72 followed by Random Forest-REPTree with 0.71 scores. While Naïve Bayes-ADTree, Naïve Bayes-Random Forest, Naïve Bayes-REPTree and Naïve Bayes-CART have the worst of F-measure of 0.50.

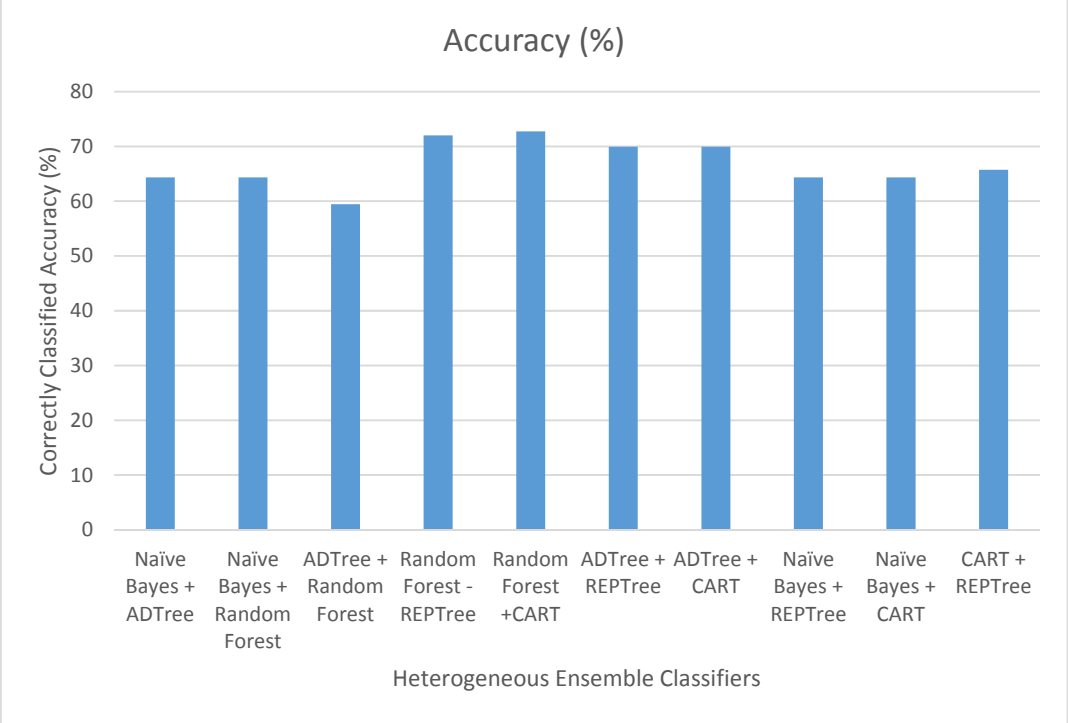

Figure 11: Accuracy for Heterogeneous Ensemble Classifiers

From Figure 11, Random Forest-CART predicts better than other classification algorithms with 72.73% prediction accuracy, followed by Random Forest-REPTree with 72.03% prediction accuracy. While Naïve Bayes-ADTree, Naïve Bayes-REPTree, and Naïve Bayes-CART have the same prediction accuracy of 64.34% which is the lowest.

Figure 12 shows error rates reported by the Heterogeneous Ensemble classifiers. Random Forest-CART has the (MAE) of 0.36 which is the lowest error rate and also have a (RMSE) of 0.45. This is followed by Random Forest-REPTree with MAE of 0.37 and RMSE of 0.45. While Naïve Bayes-ADTree has the highest MAE of 0.48 and RMSE of 0.49.

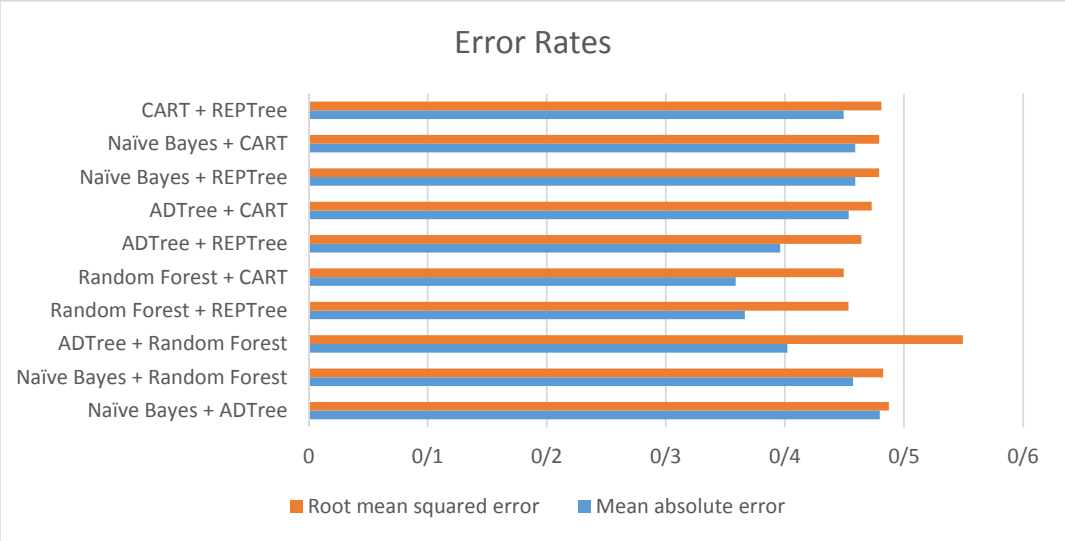

Figure 12: Error Rates for Heterogeneous Ensemble Classifiers

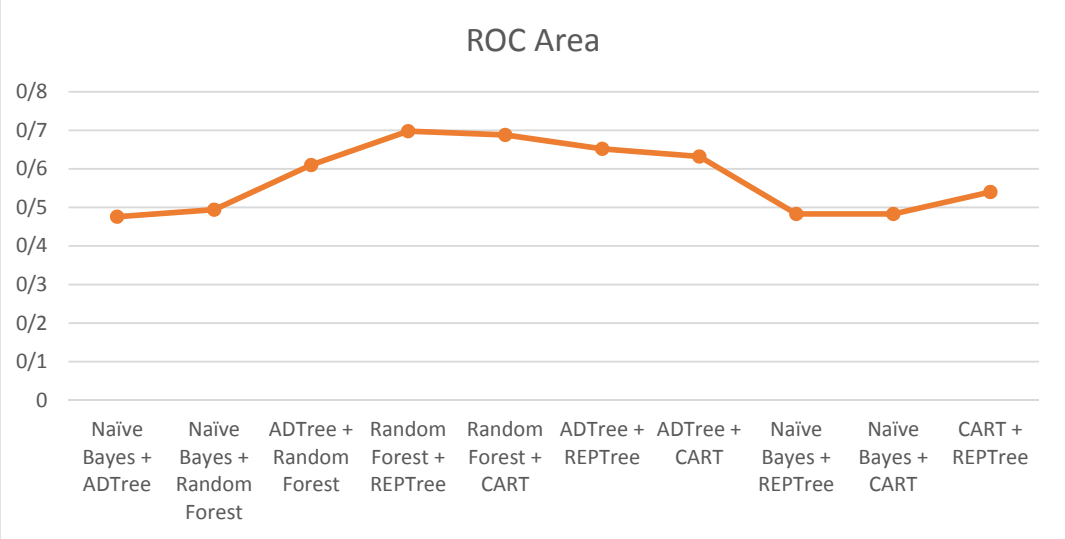

Figure 13: ROC Area for Heterogeneous Ensemble Classifiers

Figure 13 shows ROC Area curve for Heterogeneous Ensemble classifiers, Random Forest-REPTree classification has the high ROC area with 0.70 followed by Random Forest-CART with 0.69 while Naïve Bayes-ADTree, Naïve Bayes-REPTree, and Naïve Bayes-CART have the lowest ROC area of 0.48 which is lower than the Naïve Bayes single classifier. Therefore, the Random Forest-REPTree ensemble performs better than other algorithms in the Heterogeneous ensemble classifier.

In summary, when the performance of the classifiers were compared in the overall experiments, the results showed that ADABOOST-Random Forest outperforms other Classifiers.

6. Conclusion

In this paper, the ADABOOST Ensemble model based on the recognized feature patterns has been proposed for breast cancer classification. It can be compared with

traditional data mining methods in cancer diagnosis. For the phase of feature extraction, the conventional methods of extracting useful information are replaced by Ensemble classification techniques which was used to extract the symbolic tumor objects to represent tumor classifiers. In this research work, both homogeneous and heterogeneous ensemble classifiers (combination of two different classifiers together) were implemented, and Synthetic Minority Over-Sampling Technique (SMOTE) data mining pre-processing was used to deal with the class imbalance problem and noise in the dataset. The evaluation criteria of the model were done using 10-fold cross-validation test option, classification accuracy, error reports, F-measure, ROC area and execution time. The experimental results show that Heterogeneous Ensemble classifiers have the best time taken to build the model of 8.34 seconds and 8.59 seconds for Random Forest-CART and Random Forest-REPTree Ensemble respectively. The ADTree-CART has the worst time taken to build the model with 610.11 seconds. Random Forest-CART has a high F-measure score of 0.72 followed by Random Forest-REPTree with 0.71 scores. While Naïve Bayes-ADTree, Naïve Bayes-Random Forest, Naïve Bayes-REP Tree, and Naïve Bayes-CART have the worst of F-measure of 0.50. ADABOOST-Random Forest predicts better than other classification algorithms with 82.52% prediction accuracy, followed by ADABOOST-REPTree with 77.62% prediction accuracy. While ADABOOST-Naïve Bayes prediction is the lowest at 35.66%.

The result also indicates that error rates reported for the ADABOOST Ensemble classifiers, with the lowest MAE of 0.21 and RMSE of 0.42 was reported for ADABOOST-Random Forest classifier. This shows minimal error reported during the prediction processes, while ADABOOST-Naïve Bayes has the highest MAE of 0.64 and RMSE of 0.80 which is the same as the Naïve Bayes single classifier. Therefore, the Random Forest-REPTree ensemble performs better than other algorithms of the Heterogeneous ensemble classifier. The performance of classifiers were compared in the overall experiments and the results showed that ADABOOST-Random Forest outperforms other classifiers.

Acknowledgment

The authours would like to appreciate the editor and reviewers for their valuable comments and suggestions, which have led to the improvement of the paper. This work was supported by the Tertiary Education Trust Fund (TETFUND) with Reference FUW/REG/T.5/VOL.1/T11

References

[1] Pendharkar, P. C., Rodger, J. A., Yaverbaum, X. X., Herman, N., & Benner, M. (1999). Association, Statistical Mathematical and Neural Approaches, for Mining Breast Cancer Patterns. Expert Systems with Applications, 17, 223–232.

[2] Elmore, J., Wells, M., Carol, M., Lee, H., Howard, D., & Feinstein, A. (1994). Variability in Radiologists' Interpretation of Memograms. New England Journal of Medicine, 331(22), 1493 - 1499.

[3] Fetiman, I. S. (1998). Detection and Treatment of Breast Cancer (2nd Ed.). London: Martin Duntiz.

[4] Lonning P. E., Sorlie T. & Borresen-Dale A-L. (2005). Genomics in Breast Cancer Therapeutic Implications. Nature Clinical Practice Oncology, **2(**1): 26-33.

[5] Abdulsalam S. O., Babatunde A. N., Hambali M. A. & Babatunde R. S. (2015). Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students' Performance in Computer Programming. Journal of Advances in Scientific Research & Its Application (JASRA), 2, Pg. 79 – 92.

[6] Ba-Alwi F. M. & Hintaya M. (2013). Comparative Study for Analysis: The Prognostic In Hepatitis Data: Data Mining Approach. International Journal of Scientific &Engineering Research, 4(8).

[7] Hambali Moshood A., & Gbolagade Morufat D. (2016). Ovarian Cancer Classification Using Hybrid Synthetic Minority Over-Sampling Technique and Neural Network. Journal of Advances in Computer Research (JACR), 7(4), 109 – 124

[8] Daumé III, H. (2012). A Course in Machine Learning. Department of Computer Science, University of Maryland.

[9] Das, R., & Sengur, A. (2010). Evaluation of Ensemble Methods for Diagnosing of Valvular Heart Disease. Expert Systems with Applications, 37(7), 5110–5115.

[10] West, D., Mangiameli, P., Rampal, R., & West, V. (2005). Ensemble Strategies for a Medical Diagnostic Decision Support System: A Breast Cancer Diagnosis Application. European Journal of Operational Research, 162(2), 532–551.

[11] Ma Y. & Ding X. (2003). Robust Real-Time Face Detection Based On Cost Sensitive ADABOOST Method. In Proc. The International Conference on Multimedia and Expo, 465 - 473.

[12] Vezhnevets A. & Vezhnevets V. (2005). Modest ADABOOST' – Teaching ADABOOST to Generalize Better. Novosibirsk Akademgorodok, Russia.

[13] Wang H., Zheng B. Yoon S. W. And Ko H. S. (2017). A Support Vector Machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research. 687-699. www.elsevier.com/locate/ejor. https://doi.org/10.1016/j.ejor.

[14] Nidhi W., Mukesh K. And Shaveta M. (2018). Classification of Breast Cnacer Tissues using Decision Tree Algorithms. International Journal of Research in Engineering Applications and Management. (IJREAM). ISSN: 2454-9150. Vol. Issue 05. 342-346.

[15] Abed B. M., Shaker K., Jalas H. A., Shaker H., Mansoor A. M., Alwan A., and Al-Gburi I. S. (2016). A Hybrid Classification Algorithm Approach for Breast Cancer Diagnosis. IEEE. 978-1-5090-0925-. 264-268.

[16] Delen D., Walker G. & Kadam A. (2005). Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2), 113 - 127.

[17] Rajesh K., & Sheila Anand, (2012). Analysis of SEER Dataset for Breast Cancer Diagnosis Using C4.5 Classification Algorithm. Int. Journal of Advanced Research in Computer and Communication Engineering, 1(2), 72 - 77.

[18] Syed Shajahaan. S, S. Shanthi, &V. Manochitra, (2013).Application of Data Mining Techniques to Model Breast Cancer Data. International Journal of Emerging Technology and Advanced Engineering, 3(11), 362-369.

[19] Subasini. A, Nirase Fathima, & Abubacker Rekha, (2014). Analysis of Classifier to Improve Medical Diagnosis for Breast Cancer Detection Using Data Mining Techniques. Int. Journal Advanced Networking and Applications, 5 (6), 2117 - 2122.

[20] Salama G. I., Abdelhalim M. B., &Zeid M. A. E, (2012). Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. Int. Journal of Computer and Information Technology, 1(1), 36-43.

[21] Isotonic Separation Technique. Journal European Operational Research, 181, 842 - 854.

[22] Leo Breiman, (1996). Bagging Predictors. Machine Learning 24, No. 2, 123–140.

[23] Zhou M. & Wei M. (2006). Face Verification Using Gabor Wavelets and ADABOOST. In The Eighteenth International Conference on Pattern Recognition, Hong Kong, 404 - 407.

[24] Sun Y., Wang Y. & Wong A. K. C. (2006). Boosting an Associative Classifier. IEEE Trans. Knowledge and Data Engineering, 18, 988 - 992.

[25] Ram´On Diaz-Uriarte, Genesrf & Varselrf (2007). A Web-Based Tool and R Package for Gene Selection and Classification Using Random Forest, BMC Bioinformatics 8 (1), 328.

[26] Laura Tolo¸ Si & Thomas Lengauer, (2011). Classification with Correlated Features: Un-Reliability of Feature Ranking and Solutions, Bioinformatics 27 (14), 1986 – 1994.

[27] Somorjai, R .L., Dolenko B.& Baumgartner R., (2003). Class Prediction and Discovery Using Gene Microarray and Proteomics Mass Spectroscopy Data: Curses, Caveats and Cautions. Bioinformatics, 19(12), 1484-1491.

[28] Stefano, C. D., Fontanella F. & Marrocco C., (2008). A GA-Based Feature Selection Algorithm for Remote Sensing Images. In: Giacobini, M. Et Al. (Ed.): Evo Workshops. LNCS 4974, Springer Verlag, Berlin, Heidelberg, 285-294.

[29] Vandar KuzhaliJ. & Vengataasalam S. (2014). A Novel Ensemble Classifier based Classification on Large Datasets with Hybrid Feature Selection Approach. Research Journal of Applied Sciences, Engineering and Technology 7(17), 3633-3642.

[30] Lavanya D. and Dr. Usha Rani K. (2012). Ensemble Decision Tree Classifier for Breast Cancer Data. International Journal of Information Technology Convergence and Services (IJITCS), 2 (1), 17 -24.

[31] Shipp, C. A. & Kuncheva L. I. (2002). Relationships between Combination Methods and Measures of Diversity in Combining Classifiers. Inform. Fusion, 3, 135-148.

[32] Lior, R., 2010. Ensemble-Based Classifiers. Artif. Intell. Rev., 33, 1-39.

[33] Zexuan Zhu, Y. S. Ong and M. Dash (2007). Markov Blanket-Embedded Genetic Algorithm for Gene Selection. *Pattern Recognition*, 49 (11), 3236-3248.

[34] Quinlan J. (1986). Induction of Decision Trees. Machine Learning, 1, 81—106.

[35] Quinlan J. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

[36] Breiman L., Friedman J. H., Olshen R. A., & Stone C. J. (1984). Classification and Regression Trees. Monterey, C. A. Wadsworth & Brooks/Cole Advanced Books & Software.

[37] Abdulsalam S. O., Babatunde A. N., Hambali M. A. & Babatunde R. S. (2015). Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students' Performance in Computer Programming. Journal of Advances in Scientific Research & Its Application (JASRA), 2, 79 – 92.

[38] Han J. & Kamber M., (2006). Data Mining: Concepts and Techniques. 2nd. Ed. San Francisco: Morgan Kaufmann, Elsevier Science.

[39] Lior R. & Oded M. (2005). Decision Trees, Data Mining and Knowledge Discovery Handbook, Springer, Pp. 165–192.

[40] Witten I. H. & Frank E. (2005). Data Mining Practical Machine Learning Tools and Techniques– 2nd Ed. The United States of America, Morgan Kaufmann Series in Data Management Systems.

[41] Bouckaert R. R., Frank E., Hall M., Kirkby R., Reutemann P., Seewald A. & Seuse D. (2008). WEKA Manual for 3.6.0, Http://Prdownloads.Sourceforge.Net/Weka/Wekamanual3.6.0.Pdf?Download [Access: 24 June 2016].

[42] Freund, Y. & Mason, L. (1999). The Alternating Decision Tree Learning Algorithms [On Line]. S. N. Available From: www1.Cs.Columbia.Edu/Compbio/Medusa/Non_Html_Files/Freund_Atrees.pdf

[43] Meinshausen, N. (2006). Quantile Regression Forests. Journal Machine Learning Research, 7, 983–999.

[44] Breiman, L. (2001) Random Forests. Machine Learning Journal Paper, **45**, 5-32.

[45] Wu, X.D. & Kumar, V. (2009). The Top Ten Algorithm in Data Mining. Chapman & Hall/CRC, London.

[46] Zhou, M., Wei, H., & Maybank, S. (2006). Gabor wavelets and ADABOOST in Feature Selection for Face Verification. In *Proceedings of Applications of Computer Vision 2006 workshop in conjuction with ECCV 2006 May, Graz, Austria* (pp. 101-109).

[47] Friedman J., Hastie T. & Tibshirani, R. (2000). Additive Logistic Regression: A Statistical View of Boosting. Journal the Annals of Statistics, 38, 337-374.

[48] Schapire R. E. & Singer, Y. (1999). Improved Boosting Algorithms Using Confidence-Rated Predictions. J. Machine Learning, Vol. 37(3), 297-336.