International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 6, Issue 3, 2019, PP 8-21
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0603002
www.arcjournals.org

A Stack Ensemble Model for the Risk of Breast Cancer Recurrence

Omotayo Joseph Adeyemi[1*], Victor Oluwatimilehin Adebayo[2], Olawale Olaniyi[3], Olayinka Olufunmilayo Olusanya[4], Peter Adebayo Idowu[5]

[1,3,4]*Department of Computer Science, Tai Solarin University of Education, Nigeria*

[2,5]*Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.*

Corresponding Author:** ***Omotayo Joseph Adeyemi, *Department of Computer Science, Tai Solarin University of Education, Nigeria*

Abstract: *This study presents a stack-ensemble model for the classification of the recurrence of breast cancer patients' datasets that were collected from the UCI online Machine Learning Repository. Following the process of identification and collection of the data the data was pre-processed by converting all numeric attributes in the datasets into nominal values using appropriate bin intervals. Three supervised machine learning (ML) algorithms, namely: naïve Bayes (NB), C4.5 decision trees (DT) and support vector machines (SVM) were selected for the development of the stack ensemble models proposed for this study by adopting 2 ML algorithms as the base classifiers while the 3rd was adopted as the meta-classifier. The 10-fold cross validation technique was adopted during the simulation of the stack ensemble models using the WEKA simulation environment after which model validation was done based on a number of performance evaluation techniques and compared. It was concluded from the results that there was no difference in performance of the stack-ensemble models which adopted SVM and NB as the meta-classifiers however the stack ensemble model that was created using the NB and SVM as base classifiers and the DT as a meta-classifier had the overall best performance. Therefore, adopting the DT as a meta classifier showed a better capability to classify the recurrence of breast cancer compare to SVM and NB classifiers.*

Keywords: *Breast Cancer, Data Mining, Ensemble Model, Stacked Generalization, naïve Bayes, C4.5 Decision Trees, Support Vector Machines.*

1. Introduction

Breast cancer is the most common cancer in women, affecting about 10% of all women at some stages of her life. In recent years, the incidence rate keeps increasing and data show that the survival rate is 88% after 5 years from diagnosis and 80% after 10 years from diagnosis (Sarvestani *et al.*, 2010). Breast cancer recurrence refers to the reoccurrence of breast cancer in a patient whose previous cancer has gone into remission. Remission is usually the result of chemotherapy and regular treatment by oncologists (Siddhant and Mangesh, 2015). Recurrence of cancer is one of the issues that affect their quality of life. Early prediction of breast cancer is one of the most crucial works in the follow-up process of breast cancer for data mining techniques have been applied using various number of machine learning algorithms for the development of risk assessment, survival assessment and recurrence of breast cancer (Ahmad *et al.*, 2013).

As a result of this, data mining has significantly reduced the challenges of false positives and false negative decisions (Karabatak and Cevdet, 2009). Using machine learning algorithms for knowledge discovery within health databases, unseen patterns can be identified in order to exploit relationships among large number of variables, and for predicting the outcome of diseases using historical cases in datasets (Zhou and Jiang, 2003). An ensemble classifier is a method which uses or combines multiple classifiers to improve robustness as well as to achieve an improved classification performance from any of the constituent classifiers (Schapire, 2003). Furthermore, this technique is more resilient to noise compared to the use of a single classifier. This method uses a 'divide and conquer approach' where a complex problem is decomposed into multiple sub-problems that are easier to understand and solve. An ensemble classifier has better accuracy than single classification techniques. The success of the ensemble approach depends on the diversity in the individual classifiers with respect to misclassified instances (Lee and Cho, 2010).

According to Polikar (2006), there are four ways to achieve this diversity, the first is to use different training data to train single classifiers, the second is to use different training parameters, the third is to use different features to train the classifiers and the final one is to combine different types of classifier. Dietterich (1997), reported that there are three main reasons why an ensemble classifier is usually significantly better than a single classifier. Firstly, the training data does not always provide sufficient information for selecting a single accurate hypothesis. Secondly, the learning processes of the weak classifier might be imperfect, and thirdly, the hypothesis space being searched might not contain the true target function while an ensemble classifier can provide a good approximation.

Stacking or stacked generalization, is a different technique of combining multiple classifiers. Unlike bagging and boosting, stacking is usually used to combine various different classifiers, e.g. decision tree, neural network, rule induction, naïve Bayes, logistic regression, etc. Stacking consists of two levels which are base learner as level-0 and stacking model learner as level-1. Base learner (level-0) uses many different models to learn from a dataset. The outputs of each of the models are collected to create a new dataset. In the new dataset, each instance is related to the real value that it is supposed to predict. Then that dataset is used by stacking model learner (level-1) to provide the final output. Breast cancer recurrence can recur anytime among survivors of breast cancer but usually returns in the first 3 to 5 years after treatment. It includes regional/local recurrence and the distant metastasis. However, the recovery probability of breast cancer is very low, as soon as the recurrence happens. Therefore, it is important that a model is developed for the classification of the recurrence of breast cancer among cancer patients using a stack ensemble model of machine learning classifiers.

2. Related Works

Xiao et al. (2018), worked on the development of a deep learning-based multi-model ensemble method for cancer prediction. The study applied deep learning to an ensemble approach that incorporated multiple different machine learning models by supplying informative gene data selected by differential gene expression analysis to five different classification models. Then, a deep learning method was employed to ensemble the outputs of the five classifiers. The results showed that the proposed method in the study showed an average accuracy of 98%. The study was limited to the demonstration of the advantage of voting ensemble learning over traditional machine learning techniques.

Tseng et al. (2017) worked on the integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. The study collected ovarian cancer related data which were extracted from the Chung Shan Medical University Hospital Tumor Registry. The study adopted the use of ensemble learning and 5 data mining techniques, namely: support vector machine (SVM), C5.0, extreme learning machine (ELM), multivariate adaptive regression splines (MARS), and random forest (RF) to rank the importance of risk factors and diagnose the recurrence of ovarian cancer. The results showed that integrated C5.0 model had the best performance in predicting the recurrence of ovarian cancer. Moreover, the classification accuracies of C5.0, ELM, MARS, RF, and SVM indeed increased after using the selected important risk factors as predictors. The study was limited to research using data collected for the recurrence of ovarian cancer.

King (2015) applied ensemble learning methods to various structured and unstructured data using various machine learning algorithms. The study collected structured and unstructured datasets from the industry which were pre-processed, analyzed and followed by model evaluation. The study developed a method for classifying profitable campaigns and maximizing overall campaign portfolio profits. The study adopted the use of 4 traditional classifiers and 4 ensemble learning techniques to build models for identifying pay-per-click campaigns. The results if the study showed that using a Meta-Cost ensemble configuration, having the ability to integrate unequal classification cost, produced the highest campaign portfolio profit. The study was limited to the application of ensemble modeling on industrial data.

Bhola and Tiwari (2015), worked on the application of machine learning algorithms to a variety of cancer dataset. The study collected cancer-related data for Leukaemia, Prostate cancer, Breast cancer, Lung cancer and Lymphoma cancer from the Artificial Intelligence Lab, Ljubljan repository online. The study adopted the application of 7 machine learning algorithms which include: naïve Bayes' classifier, k-Nearest Neighbour, Support Vector Machine, Random Forest, Bagging and AdaBoost. The results showed that the models that adopted the use of naïve Bayes' and support vector machines had the best performance overall. The study was limited to a comparative analysis of machine learning algorithms on cancer data.

Al-Bahrani et al. (2013) worked on the development of a survival prediction model for colon cancer using ensemble data mining methods. The study collected data for the study from the Surveillance, Epidemiology, and End Results (SEER) program containing information collected about colon and rectum cancer between the years 1973 – 2009. The study adopted 20 classification schemes which include: 5 basic classifiers, and a combination of the 3 meta-classifiers with the 5 basic classifiers as underlying classifiers also an ensemble voting scheme using 3 classification schemes was also adopted following the application of information gain and consistency feature selection of relevant variables. The results of the study showed that prediction accuracies of 90.38%, 88.01%, and 85.13%and an AUC of 0.96, 0.95, and 0.92 were obtained for the 1-year, 2-year and 5-year colon cancer survival prediction using the ensemble voting classification scheme. The study was limited to the identification of the advantage of ensemble models over using isolated algorithms.

3. METHODS

3.1. Method of Data Collection and Pre-processing

For this study which required the development of a classification model for the recurrence of breast cancer, data was collected from an online resource which was accessed from the University of Chicago Illinois (UCI) machine learning repository located online. The required dataset which contained information about breast cancer patients with recurrence and those without recurrence was downloaded from the repository as a text file which was later preprocessed into an arff file format. Following the process of data collection, the data was pre-processed using feature selection techniques in order to identify the most relevant features among the initial input features in the dataset collected.

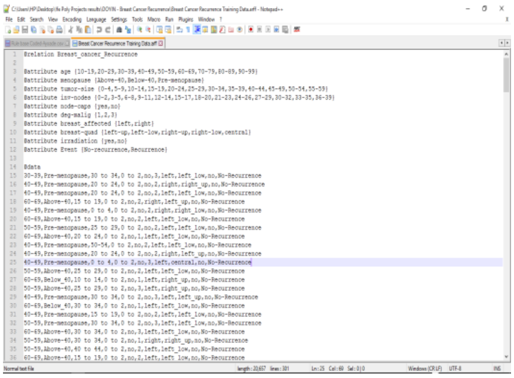
3.1.1. Method of Collection of Relevant Data

The dataset required for this study was downloaded from an online repository which was accessed from the University of Chicago Illinois (UCI) machine learning repository located online and retrieved from the location at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer. The dataset collected contained 286 breast cancer patient records consisting of 9 attributes which were either real or nominal valued. The dataset was downloaded from the repository as a comma separated variable (.csv) file format containing the attributes used to describe the data on the first row following which the data for each breast cancer patient was defined as either a recurrence or no recurrence.

The class label used to identify the recurrence of breast cancer was defined as event and was used to represent 201 records of no recurrence and 85 records of recurrence of breast cancer. The dataset was collected and was used to identify the features that were considered for the recurrence of breast cancer. In all there were 10 features among which 9 were used as the input variables while one was used as the target variable for the event of breast cancer recurrence. Table 3.1 presents the description of the variables that were identified for the classification of breast cancer recurrence.

3.1.2. Method of Pre-processing of Collected Data

Following the process of the identification and collection of the dataset required for the development of the ensemble model required for the classification of the recurrence of breast cancer. The description of the data presented in Table 3.1 was used to pre-process the dataset in order to identify valid and invalid values within the dataset collected in additional to the presence of missing data values also. Following the process of cleaning the data for missing and inconsistent values, the data was converted into a structured format which was required by the simulation environment. The data was converted into an attribute relation file format (.arff) which defined datasets using 3 different portions as shown in Figure 3.1.


C:\Users\HP\Desktop\fe Poly Projects results\DOYIN - Breast Cancer Recurrence\Breast Cancer Recurrence Training Data.arff - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
Rule base Coded-Ayoade.csv
Breast Cancer Recurrence Training Data.arff
Normal text file
length : 20,657 lines : 301

```
@relation Breast_cancer_Recurrence

@attribute age {10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99}
@attribute menopause {Above-40,Below-40,Pre-menopause}
@attribute tumor-size {0-4,5-9,10-14,15-19,20-24,25-29,30-34,35-39,40-44,45-49,50-54,55-59}
@attribute inv-nodes {0-2,3-5,6-8,9-11,12-14,15-17,18-20,21-23,24-26,27-29,30-32,33-35,36-39}
@attribute node-caps {yes,no}
@attribute deg-malig {1,2,3}
@attribute breast_affected {left,right}
@attribute breast-quad {left-up,left-low,right-up,right-low,central}
@attribute irradiation {yes,no}
@attribute Event {No-recurrence,Recurrence}

@data
30-39,Pre-menopause,30 to 34,0 to 2,no,3,left,left_low,no,No-Recurrence
40-49,Pre-menopause,20 to 24,0 to 2,no,2,right,right_up,no,No-Recurrence
40-49,Pre-menopause,20 to 24,0 to 2,no,2,left,left_low,no,No-Recurrence
60-69,Above-40,15 to 19,0 to 2,no,2,right,left_up,no,No-Recurrence
40-49,Pre-menopause,0 to 4,0 to 2,no,2,right,right_low,no,No-Recurrence
60-69,Above-40,15 to 19,0 to 2,no,2,left,left_low,no,No-Recurrence
50-59,Pre-menopause,25 to 29,0 to 2,no,2,left,left_low,no,No-Recurrence
60-69,Above-40,20 to 24,0 to 2,no,1,left,left_low,no,No-Recurrence
40-49,Pre-menopause,50-54,0 to 2,no,2,left,left_low,no,No-Recurrence
40-49,Pre-menopause,20 to 24,0 to 2,no,2,right,left_up,no,No-Recurrence
40-49,Pre-menopause,0 to 4,0 to 2,no,3,left,central,no,No-Recurrence
50-59,Above-40,25 to 29,0 to 2,no,2,left,left_low,no,No-Recurrence
60-69,Below_40,10 to 14,0 to 2,no,1,left,right_up,no,No-Recurrence
50-59,Above-40,25 to 29,0 to 2,no,3,left,right_up,no,No-Recurrence
40-49,Pre-menopause,30 to 34,0 to 2,no,3,left,left_up,no,No-Recurrence
60-69,Below_40,30 to 34,0 to 2,no,1,left,left_low,no,No-Recurrence
40-49,Pre-menopause,15 to 19,0 to 2,no,2,left,left_low,no,No-Recurrence
50-59,Pre-menopause,30 to 34,0 to 2,no,3,left,left_low,no,No-Recurrence
60-69,Above-40,30 to 34,0 to 2,no,3,left,left_low,no,No-Recurrence
50-59,Above-40,30 to 34,0 to 2,no,1,right,right_up,no,No-Recurrence
50-59,Above-40,40 to 44,0 to 2,no,2,left,left_low,no,No-Recurrence
60-69,Above-40,15 to 19,0 to 2,no,2,left,left_low,no,No-Recurrence
```

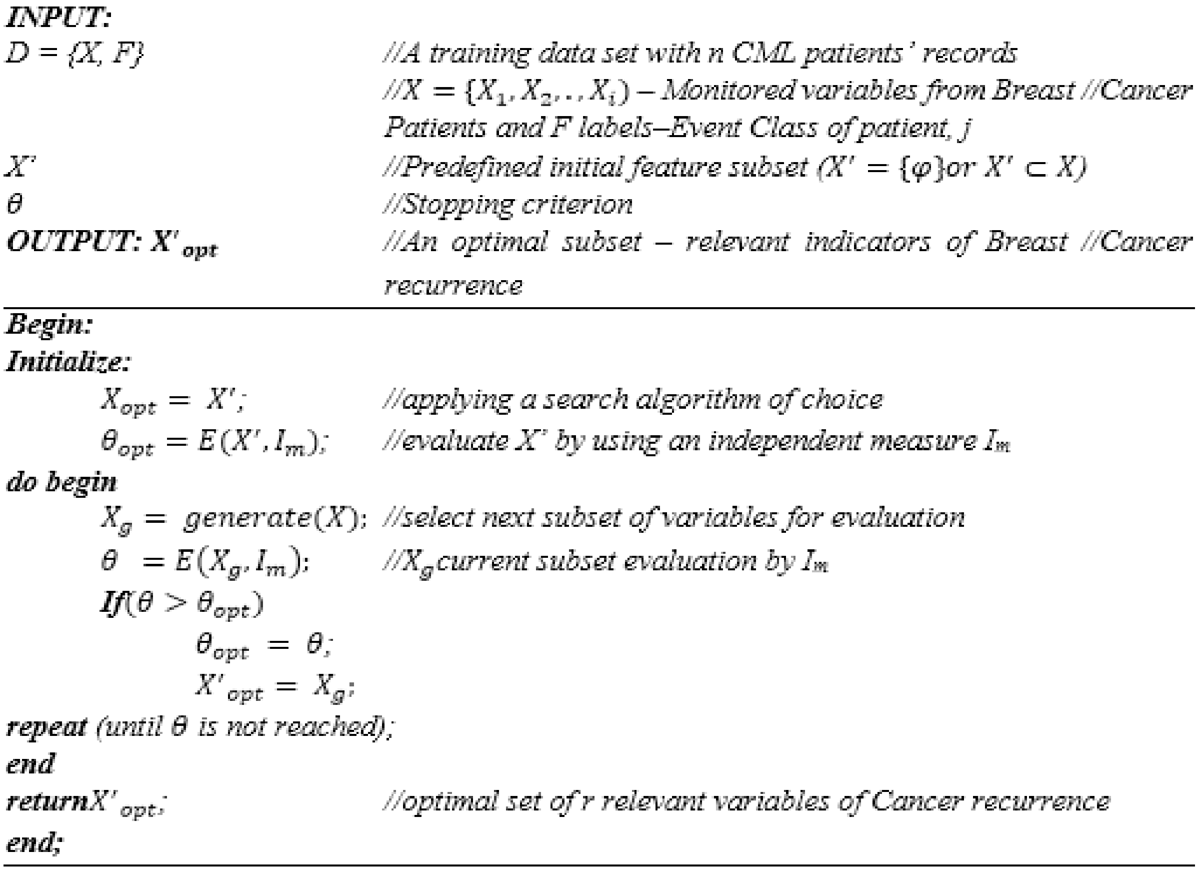
Figure3.1. *Screenshot of Attribute File Format (.arff) for the Dataset*

According to the figure, the name of the file is defined by the @relation tag of the .arff file which is followed by the @attributes tag which was used to define each attribute consisting of the inputs and the output at the last @attribute line. Following the description of the name of the attribute is the definition of the values that can be given to each attribute defined as shown in Figure 3.1 with the target class which describes the event of breast cancer recurrence on the last line of attributes defined. Following the process of the identification and collection of the data needed for developing the predictive model, it was necessary in order to determine which set of variables are deemed more predictive for breast cancer recurrence.

According to literature, identifying the most relevant variables improves the performance of the supervised machine learning algorithm's performance and also a reduction in model complexity. The Filter-based feature selection (FS) methods were used in this study to determine the relevant features among the features present in the data collected from the study location. This is motivated by the following reasons:

- Filter-based FS algorithms define relevance by identifying the attributes that are more correlated with the target class;
- Filter-based FS algorithms are less computationally expensive compared to wrapper-based FS algorithms which require an improvement of the supervised machine learning algorithm.

In this study, it is important to note that the purpose is to predict the event of recurrence or not. The performance of the predictive model will be determined by the identification of the most relevant variables from all the variables initially identified in the collected data. The motivation is the identification of the variables which improve the classification performance of the predictive model since they're more related to the target class than other variables. The feature selection algorithm that was adopted for the selection of the relevant variables in this study is shown as follows.



```
INPUT:
D = {X, F}                 //A training data set with n CML patients' records
                           //$X = \{X_1, X_2, .., X_i)$ – Monitored variables from Breast //Cancer
                           Patients and F labels–Event Class of patient, j
X'                         //Predefined initial feature subset ($X' = \{\varphi\}$or $X' \subset X$)
$\theta$                          //Stopping criterion
OUTPUT: $X'_{opt}$              //An optimal subset – relevant indicators of Breast //Cancer
                           recurrence
Begin:
Initialize:
        $X_{opt} = X'$;              //applying a search algorithm of choice
        $\theta_{opt} = E(X', I_m)$;      //evaluate X' by using an independent measure $I_m$
do begin
        $X_g = generate(X)$;    //select next subset of variables for evaluation
        $\theta = E(X_g, I_m)$;         //$X_g$current subset evaluation by $I_m$
        If($\theta > \theta_{opt}$)
                $\theta_{opt} = \theta$;
                $X'_{opt} = X_g$;
repeat (until $\theta$ is not reached);
end
return$X'_{opt}$;                //optimal set of r relevant variables of Cancer recurrence
end;
```

The feature selection algorithm required the dataset that contained the set of *i* initial features alongside the target event class for *j* breast cancer patient records. From the dataset, the set of array of *i* input features for *j* patients were tagged as *X* while the target class was tagged as *F*. Also, an initial subset of features *X'* was selected and defined as the initial relevant feature subset X'_{opt}. An initial stopping criterion θ_{opt} $(= E(X', I_m)$ was determined from the merit of the subset X'_{opt} using an independent measure I_m. Following this, another feature subset X_g was generated using a search algorithm (such as: greedy, hill climbing, genetic search etc.) and the independent measure I_m was used to estimate a new stopping criterion θ $(= E(X_g, I_m))$ which was compared with the initial θ_{opt}. If the new θ is greater than θ_{opt} then X_g becomes the new X'_{opt} otherwise another feature subset X_g is generated and tested using I_m**.** The process of generating new successive feature subsets X_g and evaluating the stopping criteria θ was repeated until the stopping criteria θ was not updated (since new $\theta < \theta_{opt}$) after a number of iterations following which the subset of features found in X'_{opt} was the set of relevant features.

3.2. Description of Ensemble Model

The stack ensemble model formulated for this study adopted the use of the naïve Bayes' (NB), support vector machines (SVM) and the C4.5 decision trees (DT) algorithms as base learners following each algorithms were in turn used as meta-learners using the dataset collected for breast cancer recurrence. The study used a framework which involved the use of 2 ML algorithms as base learners while the 3rd ML algorithm was used as the meta-learner. The meta-learner took as input, the models which were developed using the 2 base-learners in order to develop the stack-ensemble model which was required for the classification of breast cancer recurrence. The meta-learner was used to identify how the predictions made by the base learners were combined to achieve the best classification accuracy. A description of the ML algorithms adopted for this study is presented in addition to how the algorithm if used as a meta-learner will handle the other 2 input base learners. The framework that was adopted for the development of the stack-ensemble model used in this study as shown in Figure 3.3.

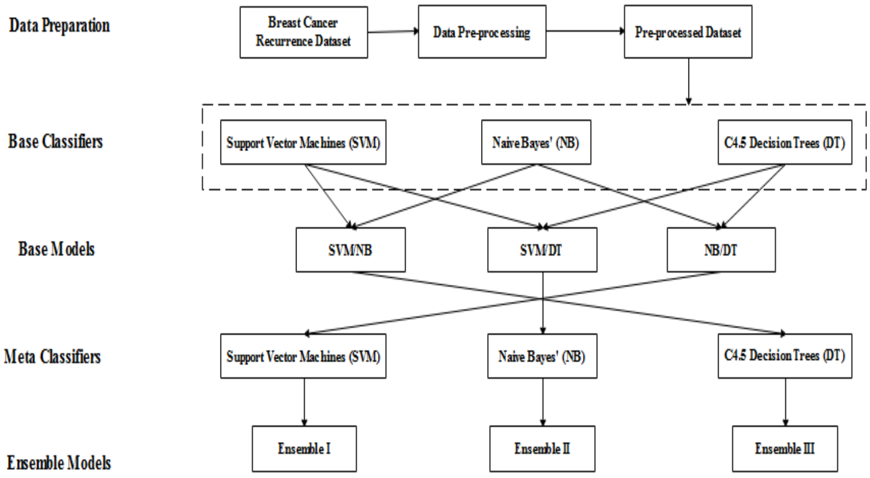


Figure3.3. *Framework of Stack-Ensemble Model using ML Algorithms*

3.2.1. C4.5 Decision Trees (DT) Algorithm

The decision tree (DT) composed of the following parts: a root node (which in this case is the predictions of the base classifier) which was used to identify the starting point of the DT and branches (defined by the values of variables used as nodes) used to connect to other nodes showing the flow from parent nodes to successive child nodes. The leaf or terminal node do not have child nodes and represent a possible value of the target variable given the variables represented by the node-to-node path from the root all the way to the leaf. Given a set X_{ij} of j number of cases consisting of i input features, the decision trees algorithm used to grow the DT using the divide-and-conquer algorithm is as follows:

- If all the cases in X_{ij} belong to the same class or X_{ij} is small, the tree is a leaf labeled with the most frequent class in X_{ij}.
- Otherwise, choose a test based on a single attribute X_i with two or more outcomes using the Information gain and gain ratio test. Make this test the root of the tree with one branch for each outcome of the test, partition X_{ij} into corresponding subsets according to the outcome for each case, and apply the same procedure recursively to each subset.

The two criteria used by the C4.5 decision trees are presented in equations (1) and (2) defined as the information gain and the split criteria respectively. Equation (1) shows the information gain which is used in determining which attribute is used to split the dataset at every iteration while equation (4) shows the split ratio that is used to determine which of the selected attribute split is most effective in splitting the dataset after attribute selection by equation (1). The gain ratio is determined by the dividing equation (1) by equation (2).

$$IG(X_i) = H(X_i) - \sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} \cdot H(X_i) \tag{1}$$

Where:

$$H(X_i) = -\sum_{t\epsilon T} \frac{|t, X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t, X_i|}{|X_{ij}|}$$

$$Split(T) = -\sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \tag{2}$$

T is the set of values for a given attribute X_i.

3.2.2. Support Vector Machines (SVM)

An SVM model is a representation of the datasets as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples were then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In formal terms, the SVM constructed a hyper-plane or set of hyper-planes in a high-dimensional space, which can be applied for classification, regression or any other task. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data points x called the support vectors since in general the larger the margin the lower the generalization error of the classifier.

Therefore, the SVM during model formulation attempts to minimize the cost by maximizing the distance between hyper-planes. A good separation is achieved by the hyperplane $< \boldsymbol{w}, \boldsymbol{x} > +b = 0$ that has the largest distance $\frac{2}{||w||}$ to the neighbouring data points of either classes at opposite ends, since in general the larger the margin the lower the generalization error of the SVM classifier. Figure 3.4 shows a clear description of the relationship between the parameters and the hyper-plane and separating margins from the support vectors x. Using the SVM as a meta-learner adopted the predictions made by the base-classifiers as the records required by the SVM for extracting the support vectors required for defining the hyperplane used for the binary classification of breast cancer recurrence.

The hyperplane created is defined as $< \boldsymbol{w}, \boldsymbol{x} > +b = 0$ where $\boldsymbol{w} \epsilon \mathbb{R}^p$ and $b \epsilon \mathbb{R}$ while $< \boldsymbol{w}, \boldsymbol{x} > +b = -1$ and $< \boldsymbol{w}, \boldsymbol{x} > +b = 1$ are the margins required for the separation w of support vectors x. Therefore, equation (3) was defined for a linearly separable function for which the decision function in equation (4) was used to propagate the output of equation (3) using a sigmoid function with interval {-1, 1}. The aim of the SVM was to maximize the separation of the hyper-planes in equation (5) subject to the decision function defined in equation (4).

$$QOS_i = f(\boldsymbol{x}_i) = (< \boldsymbol{w}, \boldsymbol{x}_\mathbf{i} > +b) > 0, \qquad \forall i \ \epsilon \ [1, n] \tag{3}$$

$$f_d(\boldsymbol{x}_i) = sign(QOS_i) = (< \boldsymbol{w}, \boldsymbol{x}_\mathbf{i} > +b) > 0, \quad \forall i \ \epsilon \ [1, n] \tag{4}$$

$$maximize \ \frac{1}{2} ||\boldsymbol{w}||^2 \tag{5}$$

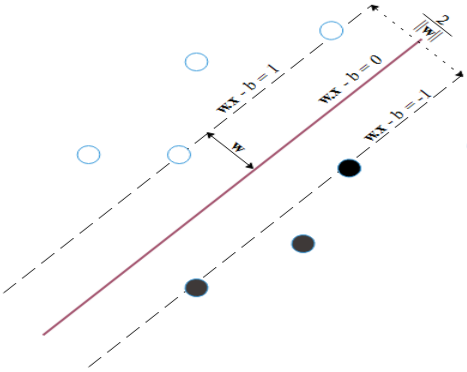


Figure3.4. *Separation of Support Vectors Using Hyper-planes*

3.2.3. Naïve Bayes (NB) Classifier

Naive Bayes' Classifier is a probabilistic model based on Bayes' theorem. It is defined as a statistical classifier which provides practical learning algorithms and prior knowledge on observed data. Let X_{ij} be a dataset sample containing predictions of the base classifiers i alongside their respective event class, C (target class) collected for j number of records/patients and $H_k = \{H_1 = Recurrence, H_2 = No\ Recurrence\}$ be a hypothesis that X_{ij} belongs to class C. For the classification of the recurrence of breast cancer given the values of the input variables of the jth record, Naïve Bayes' classification required the determination of the following:

- $P(H_k|X_{ij})$ – Posteriori probability: is the probability that the hypothesis, H_k holds given the observed data sample X_{ij} for $1 \leq k \leq 2$.
- $P(H_k)$ - Prior probability: is the initial probability of the target class $1 \leq k \leq 2$;
- $P(X_{ij})$ is the probability that the sample data is observed for each attribute (which in this case are the predictions of the base-classifiers), i; and
- $P(|X_{ij}|H_k)$ is the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X_{ij}.

Therefore, the posteriori probability of an hypothesis H_k is defined according to Bayes' theorem as shown in equation (6) while the classification of breast cancer recurrence for a record was determined with the class label with maximum probability according to equation (7).

$$P(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(X_i)}{P(H_k)} \quad for\ k = 1,2 \tag{6}$$

$$Risk = MAX[P(Recurrence|X_k), P(No\ Recurrence|X_k)] \tag{7}$$

3.3. Model Simulation Environment

Following the identification of the ensemble model of supervised machine learning algorithms that was needed to formulate the classification model for breast cancer recurrence, the simulation of the predictive model was performed using the data collected online from the UCI repository. The Waikato Environment for Knowledge Analysis (WEKA) software – a suite of machine learning algorithms was used as the simulation environment for the development of the predictive model.

WEKA is open source software under the GNU General Public License. The system was developed at the University of Waikato in New Zealand. WEKA stands for the Waikato Environment for Knowledge Analysis. The software is freely available at http://www.cs.waikato.ac.nz/ml/weka. The system was written using object-oriented language, Java. WEKA provides implantations of state-of-the-art data mining and machine learning algorithms which includes: 49 data preprocessing tools, 76 classification/regression algorithms, 8 clustering algorithms, 15 attribute/subset evaluators + 10 search algorithms for feature selection, 3 algorithms for finding association rules and 3 graphical user interfaces from which the explorer was used for this study using a 10-fold cross validation process.

3.3.1. Extraction of Evaluation Results using Confusion Matrix

In order to evaluate the performance of the supervised machine learning algorithms used for the classification of the diagnosis of hypertension, there was the need to plot the results of the classification on a confusion matrix as shown in Figure 3.5. A confusion matrix is a square which shows the actual classification along the vertical and the predicted along the vertical. All correct classifications lie along the diagonal from the north-west corner to the south-east corner also called True Positives (TP) and True Negatives (TN) while other cells are called the False Positives (FP) and False Negatives (FN). In this study, recurrence cases are considered as the positive case while the no recurrence cases are the negative cases; the definitions are presented as follows:

- True Positive (TP): Recurrence cases correctly classified as recurrence cases;
- True Negative (TN): No Recurrence cases correctly classified as no recurrence cases;
- False Positive (FP): No recurrence cases incorrectly classified as recurrence cases; and
- False Negative (FN): Recurrence cases incorrectly classified as m o recurrence cases.

3.3.2. Description of Performance Evaluation Metrics used for Model Validation

The true positive/negative and false positive/negative values recorded from the confusion matrix can then be used to evaluate the performance of the prediction model.

Recurrence	No recurrence	← Predicted as
TP	FN	Recurrence
FP	TN	No recurrence

Figure3.5. *Confusion Matrix for Model Performance Evaluation*

A description of the definition and expressions of the metrics are presented as follows:

- True Positive (TP) rates (sensitivity/recall) – is the proportion of the actual recurrence (or no recurrence) cases correctly classified.

$$TP\ (recurrence) = \frac{TP}{TP + FN} \tag{8a}$$

$$TP\ (no\ recurrence) = \frac{TN}{TN + FP} \tag{8b}$$

- False Positive (FP) rates (1-specificity/false alarms) – proportion of actual no recurrence (or recurrence) cases misclassified.

$$FP\ (recurrence) = \frac{FP}{FP + TN} \tag{9a}$$

$$FP\ (no\ recurrence) = \frac{FN}{FN + TP} \tag{9b}$$

- Precision – proportion of predicted recurrence (or no recurrence) cases that were correct classified.

$$Precision\ (recurrence) = \frac{TP}{TP + FP} \tag{10a}$$

$$Precision\ (no\ recurrence) = \frac{TN}{FN + TN} \tag{10b}$$

- Accuracy – proportion of the total predictions that were correct.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

4. Results and Discussion

4.1. Results of Data Collection and Pre-processing

For this study, data was collected from an online repository created by the UCI Machine Learning Repository. The data collected for this study consisted of 286 records which consisted of 10 attributes including the target class which defined the event of breast cancer recurrence or not. The data collected contained records for 201 non recurrent breast cancer cases and 85 recurrent breast cancer cases as shown in Table 4.1 which showed a ratio of about 3 to 1 for no recurrence to recurrence data. Hence there were more data records for the non-recurrent cases than there were data for the recurrence cases. Among the 9 attributes identified in the collected dataset, there were nominal and numeric values used for labelling. Among the numeric labels identified there were the present age of the patient, breast cancer tumor size, the number of inverse node caps and the age at menopause while the nominal attributes consisted of the presence of node caps, degree of malignancy, breast affected, breast quadrant of breast affected and irradiation.

All numeric attributes were discretized using a number of intervals as follows: the age of the patient was discretized using 10 bins such that the lowest age was 10 and oldest age 99, for example using the intervals 10 – 19, 20 – 29 etc., the age of menopause was also discretized by using a threshold age of 40 years such that 3 nominal values were created namely: below 40 years, above 40 years and pre-menopause of the patient has not experienced menopause; the tumor size was also discretized using 5 bins such that the lowest size was 0 and largest size was 59 with 12 intervals created in all, for example, using the intervals 0 – 4, 5 – 9, 10 – 14 etc.; the number of inverse node caps was discretized using 3 bins such that the lowest was 0 and highest was 39 with 13 intervals created in all, for example using the intervals 0 – 2, 3 – 5, 6 – 8 etc. All nominal values were presented as discovered in the dataset following which a frequency distribution table was used to present a description of the dataset based on the identified attributes.

Table4.1. *Distribution of Breast Cancer recurrence Data Collected*

Class Label	Frequency	Percentage (%)
Recurrence	85	29.7
No Recurrence	201	70.3
Total	**286**	**100.0**

4.2. Discussion of Data Identification and Collection

Following the identification and collection of the dataset required for this study, a frequency distribution table was used to present the description of the attributes of the dataset collected for this study. Table 4.2 shows the frequency distribution table that was used to describe the dataset based on the attributes identified. The results on that table showed that based on the discretization of the age of patients using the interval of 10 bins, there were 6 unique intervals described out of the identified intervals for the age of patient which showed that majority of the patients were between the ages 30 and 69 years of age with the highest number of patients occurring within the interval of 50 – 59 followed by 40 – 49 and 60 – 69. The results of the discretization of the age of menopause showed that majority of the patients had not experienced menopause followed by the patients which had menopause above 40 years with the least been patients whom had experienced menopause.

The results of the description of the size of tumor showed that majority of the patients has tumor sizes within the interval of 30 – 34 followed by those with sizes within the interval of 25 – 29 and 20 – 24 which in general shows an interval of about 20 – 34 patients among whom are majority of the patients who experienced breast cancer recurrence. The results of the number of inverse nodes showed that majority of the patients in the datasets lied within the interval of 0 – 2 and 3 – 5 among whom were majority of the patients who had breast cancer recurrence. The results of the presence or absence of node caps revealed that majority of the patients had no node caps however; majority of patients without node caps had breast cancer recurrence. The results of the distribution of the degree of malignancy revealed that majority of the patients had 2nd degree followed by those with 3rd degree however the majority of patients who experienced breast cancer recurrence were those with 3rd degree malignancy.

Table4.2. *Distribution of Breast Cancer Recurrence Dataset*

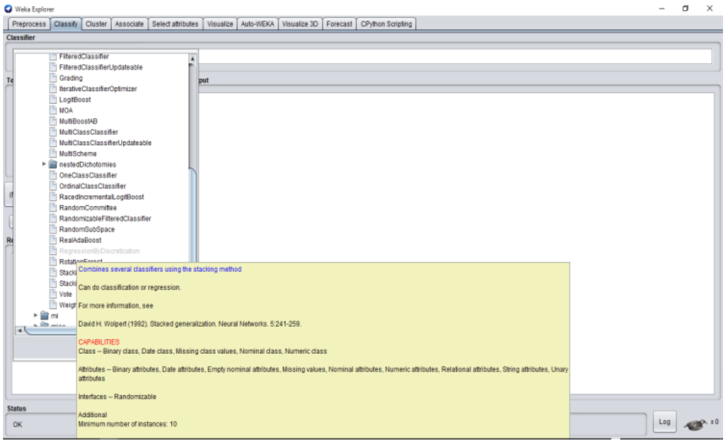
Variable Name	Values	Frequency	Percentage (%)
Age	20 – 29 30 – 39 40 – 49 50 – 59 60 – 69 70 – 79	1 36 90 96 57 6	0.35 12.59 31.47 33.57 19.93 2.10
Age of Menopause	Above 40 years Below 40 years Pre-menopause	129 7 150	45.10 2.45 52.45
Size of Tumor	0 – 4 5 – 9 10 – 14 15 – 19	8 4 28 30	2.80 1.40 9.79

	20 – 24 25 – 29 30 – 34 35 – 39 40 – 44 45 – 49 50 – 54	50 54 60 19 22 3 8	10.49 17.48 18.88 20.98 6.64 7.69 1.05 2.80
Number of Inverse Nodes	0 – 2 3 – 5 6 – 8 9 – 11 12 – 14 15 – 17 18 – 20 21 – 23 24 – 26	213 36 17 10 3 6 0 0 1	74.48 12.59 5.94 3.50 1.05 2.10 0.00 0.00 0.35
Presence of Node caps	Yes No Missing	56 222 8	19.58 77.62 2.80
Degree of malignancy	First Second Third	71 130 85	24.83 45.45 29.72
Breast Affected	Left Right	152 134	53.15 46.85
Breast Quadrant	Left-Up Left-Low Right-Up Right-Low Central Missing	97 110 33 24 21 1	33.92 38.46 11.54 8.39 7.34 0.35
Irradiation	Yes No	68 218	23.78 76.22

The results of breast affected showed that majority of the patients had their left part affected which was almost the same as patients with the right part affected however it was observed that the recurrence of cancer was more common among those with the left breast affected. The results of the breast quadrant affected revealed that majority had the left-lower part affected followed by those with the left-upper part affected however the majority of patients with breast cancer recurrence were among those who had their left-upper and left-lower part of breast affected. The results of the exposure to irradiation showed that majority of the patients were nor exposed to irradiation however majority of patients with breast cancer recurrence were those that had no exposure to radiation.

4.3. Results of Formulation and Simulation of Ensemble Model

Following the process of data identification and description, the stack ensemble model required for the classification of breast cancer recurrence as presented in the framework is presented in this section. The WEKA simulation environment was used for the development of the stack ensemble model using the *meta* Class from which the stack ensemble set-up was carried out as shown in Figure 4.1(a). Following the process of loading the stack ensemble set-up, the base classifiers alongside the meta classifiers were also selected by choosing 2 out of the identified ML algorithms while using the 3rd as the *meta* classifiers as shown in Figure 4.1(b). The process was repeated 3 times by combining any 2 ML algorithm as base classifiers and the 3rd as a *meta* classifier.



Combines several classifiers using the stacking method

Can do classification or regression.

For more information, see

David H. Wolpert (1992). Stacked generalization. Neural Networks. 5:241-259.

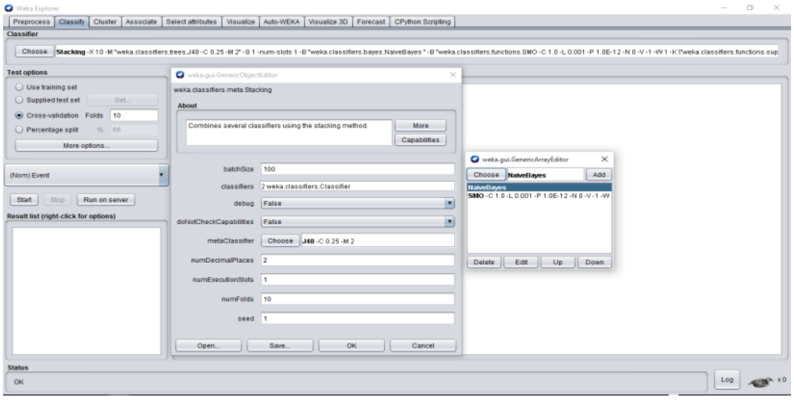
CAPABILITIES
Class -- Binary class, Date class, Missing class values, Nominal class, Numeric class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

Interfaces -- Randomizable

Additional
Minimum number of instances: 10

Figure4(a). *Setting up the stack-ensemble model*



Stacking -X 10 -M "weka.classifiers.trees.J48 -C 0.25 -M 2" -S 1 -num-slots 1 -B "weka.classifiers.bayes.NaiveBayes " -B "weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"weka.classifiers.functions.sup

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Event

weka.classifiers.meta.Stacking

Combines several classifiers using the stacking method.

batchSize 100
classifiers 2 weka.classifiers.Classifier
debug False
doNotCheckCapabilities False
metaClassifier Choose J48 -C 0.25 -M 2
numDecimalPlaces 2
numExecutionSlots 1
numFolds 10
seed 1

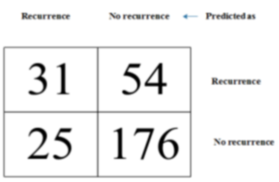
Choose NaiveBayes
NaiveBayes
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W

OK

Figure4(b). *Setting up the base classifiers and meta classifier*

4.3.1. Results of Simulation of the Stack Ensemble Models

Using the stack ensemble model provided by the WEKA simulation environment, the first stack ensemble model was simulated by adopting the NB and SVM as the base classifiers while the DT was adopted as the meta classifier using the 10-fold cross validation training process. Therefore, the predictions that were provided by the base learners was used as the input provided to the DT in creating the final stack ensemble model for breast cancer recurrence. The simulation run for the stack ensemble model using NB and SVM as base learners and DT as the meta learner is shown in Appendix I. The results of the study showed that the model ran for a period of 3.12 seconds with a root mean square error of 0.4358 owing for an accuracy of 72.3% for 207 correct classifications out of 286 patients' records. Also, the results of the simulation run showed that the stack ensemble model was able to predict 176 cases correctly out of the actual 191 no recurrence cases and 31 cases correctly out of the actual 85 recurrence cases as shown in the confusion matrix in Figure 4.2.

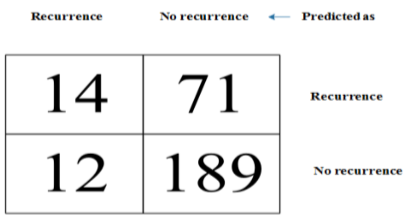


Recurrence	No recurrence	← Predicted as
31	54	Recurrence
25	176	No recurrence

Figure4.2. *Result of Stack Ensemble Model using NB and SVM as Base Learners and DT as a Meta Learner*

The second stack ensemble model was simulated by adopting the NB and DT as the base classifiers while the SVM was adopted as the meta classifier using the 10-fold cross validation training process.

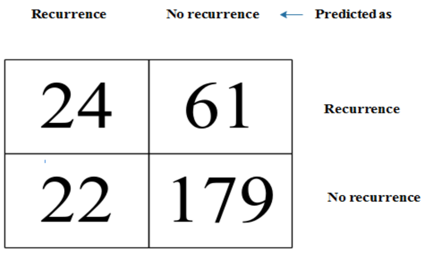
Therefore, the predictions that were provided by the base learners was used as the input provided to the SVM in creating the final stack ensemble model for breast cancer recurrence. The results of the study showed that the model ran for a period of 0.53 seconds with a root mean square error of 0.5387 owing for an accuracy of 71.0% for 203 correct classifications out of 286 patients' records. Also, the results of the simulation run showed that the stack ensemble model was able to predict 189 cases correctly out of the actual 191 no recurrence cases and 14 cases correctly out of the actual 85 recurrence cases as shown in the confusion matrix in Figure 4.3.



Recurrence	No recurrence	← Predicted as
14	71	Recurrence
12	189	No recurrence

Figure4.3. *Result of Stack Ensemble Model using NB and DT as Base Learners and SVM as a Meta Learner*

The third stack ensemble model was simulated by adopting the SVM and DT as the base classifiers while the NB was adopted as the meta classifier using the 10-fold cross validation training process. Therefore, the predictions that were provided by the base learners was used as the input provided to the NB in creating the final stack ensemble model for breast cancer recurrence. The simulation run for the stack ensemble model using SVM and DT as base learners and NB as the meta learner is shown in Appendix III. The results of the study showed that the model ran for a period of 1.5 seconds with a root mean square error of 0.4775 owing for an accuracy of 71.0% for 203 correct classifications out of 286 patients' records. Also, the results of the simulation run showed that the stack ensemble model was able to predict 179 cases correctly out of the actual 191 no recurrence cases and 24 cases correctly out of the actual 85 recurrence cases as shown in the confusion matrix in Figure 4.4.



Recurrence	No recurrence	← Predicted as
24	61	Recurrence
22	179	No recurrence

Figure4.4. *Result of Stack Ensemble Model using DT and SVM as Base Learners and NB as a Meta Learner*

4.3.2. Results of Model Validation using Performance Evaluation Metrics

The results of the validation of the 3 stack ensemble models was evaluated using a number of performance evaluation metrics that were estimated from the confusion matrices of the simulation results. The results of the model validation was assessed based on the number of correct classifications, accuracy (recorded in percentage), true positive (TP) rate, false positive (FP) rate and the precision. The best model is the ensemble model with the highest number of correct classifications, accuracy, TP rate and precision alongside with lowest FP rates. The summary of the results of the model validation of the 3 stack ensemble models developed based on the performance evaluation metrics is presented in Table 4.2.

The results of the first stack ensemble model using the NB and SVM as base learners and the DT as the meta learner had 207 correct but 79 incorrect classifications out of the 286 dataset records used in the study. The correct classifications were composed of 176 no recurrence and 31 recurrent cases while the incorrect classifications composed of 25 no recurrent cases as recurrence and 31 recurrent cases as no recurrence which owed for an accuracy of 72.4% and misclassification rate of 27.6%. the TP rate which showed how much of actual cases were correctly classified showed that 35.5% and 87.6% of the actual recurrent and non-recurrent cases respectively were correctly classified owing for an average of 72.4% of actual cases correctly classified. The FP rate which showed how much cases were misclassified showed that 12.4% of no recurrent cases were misclassified as recurrent while 63.5% of recurrent cases were misclassified as no recurrent owing for an average of 48.3% of actual cases misclassified. The precision which showed how much of the predictions made by the stack-ensemble model is correct showed that 55.4% and 76.5% of the predicted recurrent and non-recurrent cases were correct owing for an average precision of 70.2% of prediction correct.

Table4.2. *Results of Validation of Ensemble Models*

Stack Ensemble	Classifiers		Correct Classification	Accuracy (%)	Target Class	Performance Evaluation Metrics		
	Base	**Meta**				**TP rate**	**FP rate**	**Precision**
Ensemble I	**NB**	**DT**	207	72.38	**Recurrence**	***0.365***	***0.124***	***0.554***
	SVM				**No recurrence**	***0.876***	***0.653***	***0.765***
					Class Average	0.724	0.483	0.702
Ensemble II	**NB**	**SVM**	203	70.98	**Recurrence**	0.165	0.060	0.538
	DT				**No recurrence**	0.940	0.835	0.727
					Class Average	0.710	0.605	0.671
Ensemble III	**SVM**	**NB**	203	83	**Recurrence**	0.282	0.109	0.522
	DT				**No recurrence**	0.891	0.718	0.746
					Class Average	0.710	0.537	0.679

The results of the second stack ensemble model using the NB and DT as base learners and the SVM as the meta learner had 203 correct but 83 incorrect classifications out of the 286 dataset records used in the study. The correct classifications were composed of 189 no recurrence and 14 recurrent cases while the incorrect classifications composed of 12 no recurrent cases as recurrence and 71 recurrent cases as no recurrence which owed for an accuracy of 71.0% and misclassification rate of 29.0%. the TP rate which showed how much of actual cases were correctly classified showed that 16.5% and 94.0% of the actual recurrent and non-recurrent cases respectively were correctly classified owing for an average of 71.0% of actual cases correctly classified. The FP rate which showed how much cases were misclassified showed that 6% of no recurrent cases were misclassified as recurrent while 83.5% of recurrent cases were misclassified as no recurrent owing for an average of 60.5% of actual cases misclassified. The precision which showed how much of the predictions made by the stack-ensemble model is correct showed that 53.8% and 72.7% of the predicted recurrent and non-recurrent cases were correct owing for an average precision of 60.5% of prediction correct.

The results of the third stack ensemble model using the SVM and DT as base learners and the NB as the meta learner had 203 correct but 83 incorrect classifications out of the 286 dataset records used in the study. The correct classifications were composed of 179 no recurrence and 24 recurrent cases while the incorrect classifications composed of 22 no recurrent cases as recurrence and 61 recurrent cases as no recurrence which owed for an accuracy of 71.0% and misclassification rate of 29.0%. the TP rate which showed how much of actual cases were correctly classified showed that 28.2% and 89.1% of the actual recurrent and non-recurrent cases respectively were correctly classified owing for an average of 71.0% of actual cases correctly classified. The FP rate which showed how much cases were misclassified showed that 10.9% of no recurrent cases were misclassified as recurrent while 71.8% of recurrent cases were misclassified as no recurrent owing for an average of 53.7% of actual cases misclassified. The precision which showed how much of the predictions made by the stack-ensemble model is correct showed that 52.2% and 74.6% of the predicted recurrent and non-recurrent cases were correct owing for an average precision of 67.9% of prediction correct.

5. Conclusion

This study presented a stack-ensemble model by combining 3 ML algorithms such that 2 were adopted as base classifiers and the 3rd as the meta classifiers using the WEKA simulation environment. It was concluded from the study that based on the description of the dataset, it was observed that majority of patients with breast cancer recurrence were among patients within the age group of 40 to 69 years of age, patients whom were pre-menopause, patients with tumor sizes between the intervals 20 to 34, patients with number of inverse nodes within the interval of 0 – 2, patients with no node caps, patients with 3rd degree malignancy, patients who have their left breast affected alongside the left-side quadrant and patients who did not get exposed to radiation. It was also concluded form the results that there was no difference in performance of the stack-ensemble models which adopted SVM and NB as the meta-classifiers however the stack ensemble model that was created using the NB and SVM as base classifiers and the DT as a meta-classifier had the overall best performance. Therefore, adopting the DT as a meta classifier showed a better capability to classify the recurrence of breast cancer compare to SVM and NB classifiers. This could be partly due to the fact that the predictions made using SVM and NB were more reliable for the DT meta classifier adopted in this study.

References

[1] Ahmad, L.G., Poorebrahimi, A.T., Ebrahimi, A. and Razavi, A.R. (2013). Using Three Machine Learning Techniques for predicting Breast Cancer Recurrence. *Journal of Health and Medical Informatics 4*(2): 1 – 3.

[2] Al-Bahrani, R., Agrawal, A. & Choudhary, A. (2013). Colon Cancer Survival Prediction using Ensemble Mining on SEER Data. In Proceeding of *IEEE International Conference on Big Data*: 9 – 16.

[3] Bhola & Tiwari (2015). Machine Learning Based Approaches for Cancer Classification using Gene Expression Data. *Machine Learning and Applications: An International Journal (MLAIJ) 2*(3/4): 1 – 12.

[4] Dietterich, T.G. (1997). Machine learning research: Four current directions. AI Magazine 18(4): 97–136.

[5] Karabatak, M. and Cevdet, M. (2009). An Expert System for Detection of Breast Cancer based on Association Rules and Neural Network. *Expert Systems with Applications 36:* 3465 – 3469.

[6] King, M.A. (2015). Ensemble Learning Techniques for Structured and Unstructured Data. PhD Thesis for Business Information Technology, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, United States.

[7] Lee, K.C. and Cho, H. (2010). Performance of Ensemble Classifier for Location Prediction Task: Emphasis on Markov Blanket Perspective. *International Journal of u- and e- Service, Science and Technology 3*(3): 1 -12.

[8] Polikar R. (2006). Ensemble Based Systems in Decision Making. IEEE Circuits and Systems Magazine, 6(3).

[9] Sarvestani, A.S., Safavi, A.A., Parandeh, N.M. and Salehi, M. (2010). Predicting Breast Cancer Survivability using Data Mining Techniques. *2nd International Conference on Software Technology and Engineering 2*: 227 – 231.

[10] Schapire, R.A. (2003). *The Boosting Approach to Machine Learning an Overview*. Nonlinear Estimation and Classification. Springer.

[11] Siddhant, K. and Mangesh, B. (2015). Predicting Breast Cancer Recurrence using Data Mining Techniques. *International Journal of Computer Applications 122*(23): 26 – 31.

[12] Tseng, C.J., Chang, J., Chen, G.D. & Cheewakriangkrai, C. (2017). Integration of Data Mining Classification and Ensemble Learning to Classify Risk Factors and Diagnose Ovarian Cancer Recurrence. *Journal of Artificial Intelligence in Medicine 78*: 47 – 54.

[13] Xiao, Y., Wu, J., Lin, Z. & Zhao, X. (2018). A Deep Learning-Based Multi-Model Ensemble Method for Cancer Prediction. *Journal of Computer Methods and programs in Biomedicine 153*: 1 – 9.

[14] Zhou, Z.H. and Jiang, Y. (2003). Medical Diagnosis with C4.5 Rule preceded by Artificial Neural Network Ensemble. *IEEE Trans Information Technology in Biomedical Sciences 7*: 37 – 42.

Citation: *Omotayo Joseph Adeyemi, et.al., (2019) A Stack Ensemble Model for the Risk of Breast Cancer Recurrence. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 6(3), pp.8-21. http://dx.doi.org/10.20431/2349-4859.0603002*