Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776685

AN HYBRID MACHINE LEARNING TECHNIQUES FOR BREAST CANCER PREDICTION: A CONCEPTUAL APPROACH.

Okebule T., Adeyemo. O. A., Oguntimilehin A., Badeji-Ajisafe B, Obamiyi S.E.
Department of Mathematical and Physical Sciences, Afe Babalola University, Nigeria.
Emails: okebulet@abuad.edu.ng

Abstract
Breast cancer is the most prevalent cancer among women in most countries. It has been found through research that early and accurate detection of breast cancer can reduce the risk of death among women. It is therefore imperative to detect breast cancer at initial stages. Machine learning techniques is one of the most trending tools of the 21st century for solving problems and it also beneficial in most applications of use since it has the capability of making predictions and helps to make better decisions. There are several Machine Learning Techniques for identification of breast cancer through training and testing datasets. This paper presents a conceptual approach of a hybrid Machine Learning Techniques for breast cancer prediction. This method employed six different machine learning algorithms respectively, Support Vector Machine (SVM), Linear Regression (LR), K Nearest Neighbor (KNN), K-means, Naïve Bayes (NB) and Hierarchical clustering. The proposed concept combines both bagging and boosting with unsupervised Machine Learning Algorithms and clustering with unsupervised Machine learning algorithms. This is very essential to combine a hybrid Machine Learning Techniques to detect this disease at the early stage to increase the saving lives of women.
Keywords: Machine Learning, Model, Concept, Breast, cancer, Dataset.

1. Introduction

Breast cancer is one of the substantial worldwide health challenges in Nigeria women and other countries women. Comparing the majority of cancer related cases deaths according to researcher statistics shows that breast cancer is the most leading causes of death among women. Breast cancer is the principal cause of death among women globally and this has contributed 19.5% to the untimely death rate among women in Nigeria. Research showed that the early diagnosed with breast cancer are still alive 10 years after their diagnosis, which is due to early detection and prompt treatment and this can significantly reduce breast cancer mortality (Jemal et al., 2011).

As breasts are prominent to women's emotional life and are a symbol of womanhood and sexuality. Research showed that breast cancer determines important alterations in the body image and self-image of the women, which could affect their experience of sexuality and marital relationship. More so, breast cancer treatment causes important physical, social and psycho-emotional changes, with a subse¬quent decrease in the women's quality of life (Sharma et al, 2017).Therefore, this disease can cause negatively im¬pact on how woman performs her role of wife, mother and individual in the community, and so, affecting her socio-occupational functioning. (Van et al, 2015).

Indeed, breast cancer diagno¬sis and treatment can lead to a loss of identity that can disrupt women's career, professio¬nal and family life, (Fagbuagun et al, 2021). Research has confirmed a strict relationship between breast cancer diagnosis, treatment and women's social functioning. It has been highlighted that bre¬ast cancer treatment may cause adverse effects that sometimes cannot be avoided. According to Watters et al, (2003) role and social functioning can diminish especially when patients receive breast cancer treatments.

It is prominent to know that as debilitating as breast cancer disease is, majority of Nigerian women have little or no knowledge of the disease and even in situations where they are aware of The disease, their attitudes towards seeking health is negative causing their untimely or preventable death. The lack of awareness on the issue of vulnerability and susceptibility associated with breast cancer discourage many women from seeking intervention early or associate the symptoms they are experiencing with breast cancer (Ramathuba et al., 2015).

1.1 Machine Learning

Machine learning, a branch of artificial intelligence, relates the problem of learning from data samples to the general concept of reasoning (Witten et al., 2005). Machine learning (ML) is widely recognized as the best method in breast cancer pattern classification and prognostic modeling.

1.2 Every learning process consists of two phases:

(i). It estimates the unknown dependencies of the system from the given dataset and
(ii) uses the estimated dependencies to predict new outputs of the system.
Machine learning has also emerged as an interesting area of biomedical research, using various techniques and algorithms to search for a specific set of biological samples in an n-dimensional space and derive acceptable generalizations with several techniques and algorithms (Niknejad et al., 2013).

There are three main common types of ML methods known as:

Supervised Learning: This allows the machine to learn the tagged data. Supervised machine learning falls into two categories: Classification and regression. Supervised machine learning classification type is basically used to predict labels or classes. A classification task refers to the learning process of classifying data into a finite set of classes. For regression problems, the training function maps data to real-valued variables. Based on this process, predictor values can be estimated for each new sample. That is, regression is used to predict the continuous size. (Mohamed et al. 2020).

Unsupervised learning: In unsupervised learning, machines are trained based on unlabeled data without supervision. These types of machine learning are used to solve association and clustering problems. Association problems consisted of finding patterns in the data that found common occurrences. Clustering is a common unattended task that tries to find categories or clusters that describe data items. Based on this process, each new sample can be assigned to one of the clusters identified with similar characteristics. (Mohamed et al. 2020).

Reinforcement: In reinforcement learning, an agent interacts with its environment by producing actions and perceiving errors or rewards. It is a feedback-based machine learning technique in which an agent learns how it behaves in an environment by taking actions and seeing the results of those actions. (Mohamed A. et al. 2020).

2. Related works

In this section, several researches have been carried on detection of breast cancer. These studies have compared and used different machine learning methods to achieve best performance accuracy. Some of previous studies are given in the followings:

Seema et al., (2012) proposed an ARNN (Adaptive Resonance Neural Network) method for cancer detection using unsupervised learning. The dataset collected for this study contained a total of 699 cases, 600 of which were used to train the network. This database contains nine attributes, so the results are divided into two categories: benign or malignant. These adaptive neural resonance network techniques used in this study showed an accuracy of 75%. However, using a smaller data set also reduces.

Boosted SVM dedicated for solving imbalanced results was suggested by Zięba et al., (2014). The result combined the advantages of ensemble classifiers with cost-sensitive support vectors for unbalanced data. More so, this method presented for extracting decisions from the boosted SVM. The solution compared the performance of the uneven data with different algorithms. In conclusion, an enhanced SVM was implemented for approximation after surgery life expectancy in patients with lung cancer.

Niranjana et al., (2015) proposed comparing the performance of artificial neural network (ANN), support vector machine (SVM), and K-nearest neighbor (ANN) models for cardiac ischemia classification. In this research approach, ANN, SVM, and ANN models were developed to classify cardiac ischemia based on morphological changes in ECG signals. The proposed ANN, SVM, and ANN models preserve the morphological features extracted from the preprocessed ECG beats. All model performances are compared and validated against the Physiobank database for accuracy, sensitivity, and specificity. The results of this study demonstrate that the proposed ANN-based model has greater potential than SVM and ANN classifiers in classifying cardiac ischemia. Experimental results confirm that the ANN model outperforms with a test classification accuracy of 96.62%. The accuracy achieved with this is significantly higher compared to SVM and ANN classifiers. However, this model has not been implemented in breast

Usha Rani (2010), suggested a parallel neural network method for improving the classification of breast cancer diagnoses. The researchers described various parallelization strategies measured in artificial neural networks, including block parallelism, neuron parallelism, and instance parallelism. Experiments were performed considering both single-layer and multilayer neural network models. We trained the network using a variable learning rate backpropagation algorithm and implemented a multi-layer perceptron to achieve an accuracy of 92. However, in this study only 11 attributes were used to train the mode

Tae-WooKim et al. (2010) have proposed a decision tree for occupational lung cancer parameters were to determine whether a condition was acceptable as lung cancer in relation to age, sex, years of smoking, histology, branch size, delay, working hours, and exposure to independent variables. . The Characterization and Recurrence Test (CART) universe is used

throughout the search for word-related indices of cellular deterioration in the lung. Presentation to a prominent pulmonary expert was the crowning achievement of the CART model. The decision tree method is easy to interpret. It can also be considered a minimum criterion for work relevance in lung cancer. However, this method was not conducted on breast cancer.

Ancy et al., (2018), performed classification of single-frame mammograms in pre-running datasets, feature extraction from grayscale co-occurrence matrices, region-of-interest segmentation, and support vector machine classification. Experimental results showed that the method used to classify tumor and non-tumor using SVM classifier, GCLM extracted features can provide accurate results. However, in this study, a method was proposed to evaluate her two datasets, tumor and non-tumor.

Anooj et al., (2012), adopted a weighted fuzzy rule for the detection of heart diseases using k-fold cross validation. The dataset used for this study gotten from UCI Respiratory and this dataset consists of 14 attributes of input and its output value is vary from 0 to 4 where 0 means no presence of diseases and from 1 to 4 it shows the existence of diseases. The datasets were used for the examination of heart disease and the result shoed the accuracy of 58.85%.

Azzaw et al., (2016), proposed a GEP model to predict microarray data on lung. To predict important lung cancer related genes with GEP model, in this proposed work two approaches were adopted for selecting genes and thus suggest specific GEP prediction models. The result of the cross-data collection was tested for consistency. The results gotten show that, precision, sensitivity, specialty, and region under the recipient functional property curve considered, the GEP model used fewer features that surpassed other models. The GEP model was a better approach to problems of diagnosis of lung cancer.

Selvi et al., (2006) proposed a frame work to detect breast cancer using KNN and SVM on the dataset collected from UCI repository to detect breast cancer with respect to the results of accuracy the efficiency of algorithm is also measured and compared. There method of machine learning algorithms applied on these datasets that shows different levels of accuracy range between 94.36% and 99.90%. However, the study was only used to determine the accuracy of the model.

Hafizah et al., (2013) employed data mining techniques to compare the performance of decision trees, SVMs, and ANNs, and analyzed data from the ICBC registry to develop a model to predict breast cancer recurrence. The dataset used was obtained from the Iranian center for breast cancer. Simulation results showed that SVM was the best classifier followed by ANN and decision tree. However, the study reported cases lost in the follow-up and there were records with missing values that were omitted in data collection.

David et al., (2019) conducted a comparative study of decision trees, NB, NN, and SVM using three different kernel functions as classifiers for classifying WPBC and Wisconsin breast cancer (WBC). Experimental results showed that NN (10X) has the highest accuracy of 98.09% on the WBC dataset and SVM-RBF (10X) has the highest accuracy of 98.32% on the WPBC dataset.

Rajendran et al., (2019) conducted a feasibility study of data mining techniques in the diagnosis of breast cancer. They reviewed a number of papers to provide an overview of the types of data mining techniques used for breast cancer prediction. Results show that commonly used data mining techniques include decision trees, naive Bayes, association rules, multilayer perceptrons (MLP), random forests, and support vector machines (SVM). The overall

performance of the technique varies from dataset to dataset. The Random Forest classifier performed better on the Wisconsin breast cancer data set, with an accuracy of 99.82%

3. Materials and Methods

There was a review of related materials on Machine Learning Techniques, Breast cancer, and risk factors of breast cancer. In this section, we describe the proposed conceptual approach by presenting an architectural overview of modules (M1, M2, M3, and M4) that correspond to the various processing stages of the system and a presentation of its main components and functions.

Figure 1, described the System Architecture for Hybrid Machine Learning Techniques for Breast Cancer Prediction: A Conceptual Approach. This system comprises of four modules, each module (M) runs independently and the last stage will provide the evaluation performances of the whole analysis.

3.1 Data Set: Datasets contain information about the data used to build models for evaluating different dataset choices. Abien Fred (2018). A total of 541 breast cancer records were collected locally at his Ekiti State University Teaching Hospital Breast Cancer Registry, each composed of 24 attributes and classes benign (no) or malignant (yes) were used for the comparative studies.

3.2 Data Pre-Processing: Purification and modification of the dataset are required before applying ML algorithms to the dataset; it is a necessary step to pre-process the data. So that Performance and accuracy of the predictive model are not only affected by the algorithms used but also by the quality of the dataset and pre-processing. Dataset can be mathematically done as well as using the scikit-learn object Min-Max-Scaler (Kotsiantis et al, 2006). An imputation method was used for the missing value and calculated as:

$$\text{Expected cell value} = \frac{\text{Total No of corrected observation}}{\text{Total unit of observations}}$$

3.3 Feature Extraction Module

Feature extraction is the process of bringing out important or salient features from the preprocessed image. In this research facial features were extracted using three different methods namely: Principal Component Analysis (PCA) using the Singular Value Decomposition Techniques (SVD), Gabor Filter and Gray Level Co-occurrence Matrix (GLCM) method. In this, Principal Component Analysis (PCA) method was employed for enhancing classification effectiveness. PCA is a popular unsupervised linear technique which attempts to transform the original feature sets which include a large number of features to a new smaller feature space, so that the current data can be expressed with a few number of features variable.

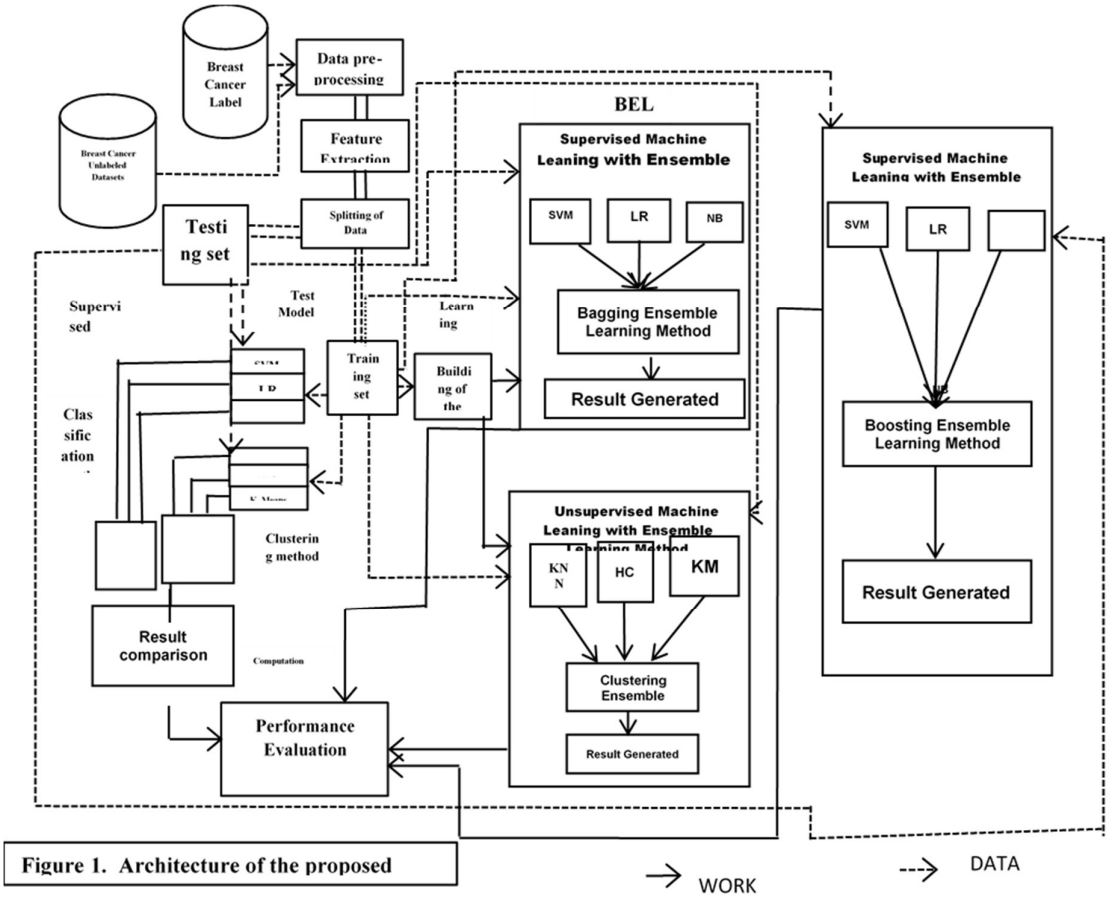

Figure 1. Architecture of the proposed

The details of PCA are mathematically explained as:

In a datasets of x1, $x_{2 \ldots} x_{541}$

Step 1: we compute the sample mean

$$\overline{\mathbf{x}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}_i \tag{1}$$

Step 2: subtract sample mean that is, when the center data is set at zero)

$$\Phi_i = \mathbf{x}_i - \overline{\mathbf{x}} \tag{2}$$

Step 3: compute the sample covariance matrix $\Sigma_{\mathbf{x}}$

$$\Sigma_{\mathbf{x}} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T = \frac{1}{M}\sum_{i=1}^{M}\Phi_i\Phi_i^T = \frac{1}{M}AA^T \tag{3}$$

Step 4: compute the eigenvalues or eigenvectors of $\Sigma_{\mathbf{x}}$

$$\Sigma_x u_i = \lambda_i u_i \tag{5}$$

Here, we will assume that:

$\lambda_1 > \lambda_2 > \ldots > \lambda_N$

(in N x M matrix) where:

N: reps features

M: reps data point

Step 5: dimensionality reduction step: approximate x using only the first K eigenvectors (K<<N) (that is, the corresponding to the K largest eigenvalues where K is a parameter):

$$\mathbf{x} - \overline{\mathbf{x}} = \sum_{i=1}^{N} y_i u_i = y_1 u_1 + y_2 u_2 + \ldots + y_N u_N \tag{6}$$

Approximately, only the first K eigenvectors will be used, that is,

$$\hat{\mathbf{x}} - \overline{\mathbf{x}} = \sum_{i=1}^{K} y_i u_i = y_1 u_1 + y_2 u_2 + \ldots + y_K u_K$$

$$\mathbf{x}-\overline{\mathbf{x}}:\begin{bmatrix}x_1\\x_2\\.\\.\\.\\.\\.\\x_N\end{bmatrix}\rightarrow\begin{bmatrix}y_1\\y_2\\.\\.\\.\\.\\.\\y_N\end{bmatrix}\rightarrow\hat{\mathbf{x}}-\overline{\mathbf{x}}:\begin{bmatrix}y_1\\y_2\\.\\.\\y_K\end{bmatrix}$$

The features are extracted using PCA reduction method, that is

if K=N, then $\hat{\mathbf{x}}=\mathbf{x}$

3.4 Data splitting: The goal of dividing the data is to avoid overfitting the model during model testing on the training dataset. The dataset for this study will be split into two parts: training data (80%) and test data (20%). Training the Model will be based on the most six prevalent classifiers, three of which are Supervised Machine Learning (SVM, LR, NB) and three Unsupervised Machine Learning (KNN, KM, HC) that will eventually lead to the building of the hybrid; also in addition, infrequent techniques called Bagging Ensemble Learning (BEL) will be employed. Consequently, the application of these methods using hyper parameter can help the efficiency and accuracy performance of the proposal hybrid model. The proposed hybrid model will be trained to identify specific types of patterns using an algorithm in a machine learning system. The model's algorithm will use a collection of data for training and building of a model, and predicts the output whilst saving that procedure for future purposes.

3.5 Model Training and Testing: For model training and testing, certain percentage method will be adopted in this study to obtain the training and testing of datasets respectively. In order to describe the performance of the models, confusion metrics will be employed in this study; that is the accuracy, precision, recall and execution time for each model will be measured (Janghel et al., 2010).

3. 5.1 Performance Metrics of Machine Learning Classification Models

The performance metric of the models used for classification was based on the following metrics:

Accuracy: The ratio of sum of samples with no breast cancer and with the breast cancer that has been correctly predicted by ML model to the total of breast cancer predicted observations Hence, The accuracy will be calculated using the formula below:

$$\text{Accuracy} = \frac{TP + TN}{TN + TP + FP + TN + FN\ FP} \tag{7}$$

(ii) Precision: The ratio of sum of samples with no breast cancer that has been correctly predicted by ML model to the sum of samples with no breast cancer and with breast cancer that

has been correctly predicted by ML model. This measure is attractive, particularly within the clinical field because of what level of the observations will be accurately analyzed. Another name for precision is sensitivity.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

Recall: It is the ratio of samples with no breast cancer that has been correctly predicted by ML model to the sum of samples with no breast cancer and with breast cancer that has been correctly predicted by ML model. (that is, Breast cancer cases),

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

True Positive Rate (TP Rate): The true positive rate was computed as the ratio of the true positive (TP) to the sum of true positive and false negative (TP + FN) as presented in Equation (10).

$$TPR = \frac{TP}{TP + FN} \quad (10)$$

Where TP is the number of faces that were correctly recognized; FN is the number of faces that were incorrectly recognized. The true positive rate for the face class are the number of faces in the face class that are correctly classified as face, divided by the total number of faces in the face class.

False Positive Rate (FP Rate): The false positive rate was computed as the ratio of the number of faces that are incorrectly recognized as face under consideration (FP) to the sum of the number of faces that are correctly classified as not faces under consideration (TN), plus the FP. FPR is presented in Equation

$$FPR = \frac{FP}{TN + FP} \quad (11)$$

False positive rate is the total number of faces in the face class that are incorrectly classified divided by the total negatives in the dataset.

3.6 Support Vector Machine

Given a set of training data set x (a vectorz_i) of i subjects, and for each subjects i=1,2,…,541 in the training data set, and a class a label y_i belonging to classes y= ±1 (The pair of input feature vectors and the class label can be represented as tuple{x_i,y_i }.

In the input training data, there is of linear separable problem where exists a separating hyper-plane which defines the boundary between class 1 (labeled as y = 1) and class 2 (labeled as y = -1). The separating hyper-plane is thereby formulated as:

$$w.x + b = 0 \quad (12)$$

Which implies:

$$y_i(w.x + b) \geq 1, i = 1, 2, \dots, 541$$

The SVM method attempt to find a classifier using the following equation:

$$y(x) = sign\left[\sum_{i=1}^{N} \propto_i y_i \; k(x_i, x) + b\right] \quad (13)$$

Where xi is positive real constants and b is a real constant. Generally, k(xi, x) = (Ø(xi), Ø(x)), (',') represents the inner product operation, and Ø(x) is a nonlinear map from the original space to the high dimensional space.

Basically, there are numerous possible values of {w, b} that create separating hyperplane. In SVM only hyperplane that maximizes the margin between two sets is used. Margin is the distance between the closest data to the hyperplane. Figure 2 shows linearly separable binary classification problem with no possibility of miss-classification data.

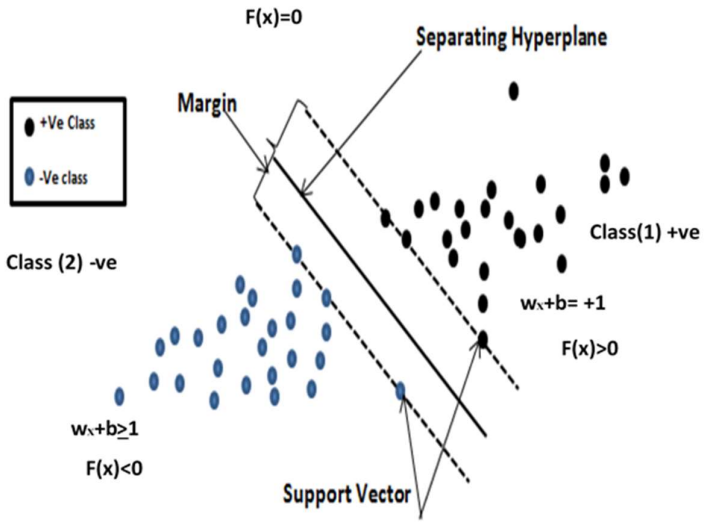

Figure 2: Support Vector Machine with linear separable data.

Point: setting of points for classification, where each point is represented by some feature vector (x), then it is been mapped to a more complex nonlinear space of phi (Ø) of X, then the feature is transformed.

$$x \in RD \tag{14}$$

$$Ø: RD \longrightarrow RM \; Ø(x) \in R$$

Decision boundary: Is the major separator that divides the points into their respective classes Decision boundary is calculated as:

$$h: w^{T} Ø(x) + b = 0 \tag{15}$$

Distance measured:
This provides a way of measuring the difference between two feature vectors

$$a(x) + b(y) + c \tag{16}$$

Optimizing: this ensures that data is seamlessly separable, that is when there exists at least a hyperplane separate the training data group of classes with 100% accuracy

$$w^{\mathbf{1}} = arg_{w,} max \quad [\min yn[w^{\mathbf{1}} \Phi(xn) + b]]$$

$$let \ \min yn[w^{\mathbf{T}} \Phi(xn) + b] = 1_1 \tag{17}$$

$$(cw)^{\mathbf{1}} \Phi(xn) + (cb) = c(w^{\mathbf{1}} \Phi(\mathrm{xn}+) + \mathrm{b} = 0$$

$$\mathrm{w} \leftarrow cw$$
$$\mathrm{b} \leftarrow \mathrm{cb}$$

$$w^{\mathbf{1}} = arg_{w,} max \frac{1}{\|w\|_2}$$

3.7 K-Nearest Neighbor (KNN) Model

K-Nearest Neighbor (KNN) is one of the simplest Machine Learning algorithms based on unsupervised Learning technique; it is a fundamental machine learning classification and regression technique. KNN required the dataset for design, the distance metric in computing distance between training and testing test, and the value of T, is also the number of nearest neighbors to retrieve. KNN classifier can be used to predict a class by means of Euclidean distance. The distance metrics is presented as:

$$d = \sqrt{\sum_i^n x_i - y_i} \tag{18}$$

Where (x, y) are points in the feature set.

To find the nearest k neighbors of y we assigned the point in the feature set to class of k nearest neighbors and majority decision rule is then used to classify the new sample. The vote is weighted based on its distance and it is presented as:

$$w = \frac{1}{d^2} \tag{19}$$

The value of k was calculated as:

$$k = \sqrt{T_s} \tag{20}$$

Where k is the nearest neighbors to get and Ts is the total training set.

3.8 K-means algorithm is used in extracting meaningful information from a large database using a cluster. Algorithm is an iterative technique that is used to partition an image into k clusters. It partitions n observations into k clusters distributing the observations among the clusters based on the nearest mean principle. The K-means clustering algorithm is a well-known data clustering technique. It is used in a variety of applications, including information retrieval and computer vision. K-means clustering divides n data points into k clusters, allowing for the grouping of comparable data points.

Step 1. Select k cluster centers, either randomly or based on some heuristic.

Step 2. Ascribe each pixel in the dataset to the cluster that minimizes the distance between a point and the cluster center.

Step 3. Re-compute the cluster centers.

Step 4. Repeat step 2 and 3

Given a set of n observations {x1, x2… xn}, K-means algorithm partitions the observations into k sets, that is {S1, S2… SK} where (k < n) in order to minimize the distance within the cluster

$$arg_s\, min \sum_{i=1}^{k} \sum_{x,\in S,} ||x_j - \mu_i||2,$$

Where μ_i is the mean of *ith* in S_i and is calculated in each iteration as presented:

$$\mu_i = \sum_{j=1}^{N} x_j /ni$$

Then, Gaussian distribution, and maximum likelihood method is used to calculate the parameters of S_i as:

$$\mu_i = \frac{\sum_{j=1}^{N} x_j}{n_i}$$

Therefore, Gaussian distribution can model only normally distributed data

The distance metrics is presented as:

$$d = \sqrt{\sum_i^n x_i - y_i} \quad (21)$$

Where (*x*, y) are points in the feature set.

3.9 Naïve Bayes

This algorithm is a supervised learning algorithm based on Bayes' theorem and is used to solve classification problems. It is primarily used in text classification with high-dimensional training datasets. Naive Bayes Classifier is one of the simplest and most effective classification algorithms that help you build fast machine learning models that can make fast predictions. It's a probabilistic classifier, meaning it makes predictions based on object probabilities. Bayes' Theorem:

Bayes' theorem, also known as Bayes' rule or Bayes' law, is used to determine the probability of a hypothesis based on prior knowledge. It depends on conditional probabilities. The formula for Bayes' theorem is given by:

$$P(h\backslash D) = \frac{P(D\backslash h)P(h)}{P(D)}$$

Where,

h is a class data specific

D is data with a class that is not yet known Hypothesis

P(h|D) is (Posterior (correct) probability): Probability of hypothesis h based on condition D.

P(D|h) is Likelihood (trial) probability: Probability of the evidence given that the probability of D hypothesis is true.

P(h) is Prior Probability: Probability of hypothesis before observing the evidence.

P(D) is Marginal Probability: Probability of Evidence

$$P(h\backslash D) = \frac{P(D\backslash h)P(h)}{P(D)}$$

The output of the model is to classify an individual to a class with the maximum posterior probability, through the use of Bayes' Rule as:

$$pr\big(c = c_1|x_1 = x_1, \dots, x_p\big) = \frac{pr(c{=}c_1)\Pr\,(x_1{=}x_1,\dots,x_p{=}x_p|c{=}c_1)}{\Pr\,(x_1{=}x_1,\dots,x_p{=}x_p}$$

(22)

Therefore, the ratio in this formula is expressed as:

$$pr\big(c = c_1|x_1 = x_1, \dots, x_p\big) = pr(c = c_1)\Pr\,(c = c_1)pr(c = c_1)\Pr\,(x_1 = x_1, \dots, x_p = x_p|c = c_i$$

The computation of prior class probabilities is formulated as follows:

$$Pr(c = c_1) = \frac{|\mathrm{T}^{ci}|}{T_N}$$

(23)

Where T^{ci} is the total number of breast cancer patients in class of C_i and T_N which is also the total number of breast cancer patients in the training set.

To formulate the independence assumption, the probability is then computed to be the relative frequency of breast cancer patients in class

$$\prod_{j=1}^{p}\Pr\big(x_{j-}x_i\big)\,|c = c_1)\prod_{j=1}^{p} = \frac{\mathrm{Pr}^{ci}x_{j-}{x_j}^{i}}{|\mathrm{Pr}^{ci}|}$$

(24)

Consequently, the conditional attribute value probability is calculated as:

$$\prod_{j=1}^{p}\Pr\big(x_{j-}x_i\big)\,|c = c_1)\prod_{j=1}^{p} = \frac{1}{\sqrt{2\pi\emptyset c_1}}e^{\frac{(x_j-\mu c_i)2}{2\sigma 2c_i}=0}$$

The output of a Naïve Bayes Classification model is the maximum posterior class probability, known as maximum a posterior which is presented as:

$$C_{MAP}(d_k) = \text{argmax}\,\mathrm{P}(C = C_i)\prod_{j=1}^{p} = \Pr(X_{j-}X_j \mid C_{-}C_i\,)$$

(25)

3.10 Linear Regression Algorithm

Linear regression is used to predict a quantitative response Y from the predictor variable X. Mathematically, we can write a linear regression equation as:

$$Y{=}a + B_x$$

(26)

Where a and b

$$b(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept)\frac{n\sum y - b(\sum x)}{n}$$

The true and estimated regression lines is shown in the figure 3below

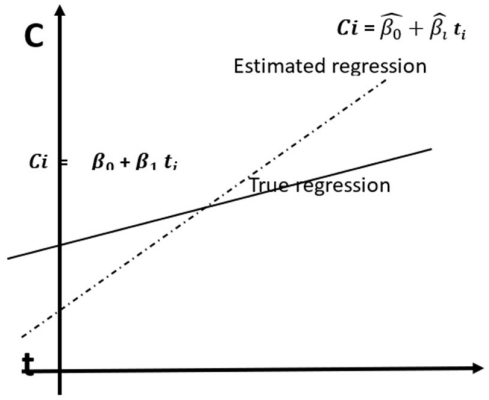

Figure 3: The true and estimated regression lines

Where:

= *dependent variable*

t_i = *independent variable*

β_0 = *C- intercept*

β_1 = *slope of the line*

The true regression line is represented as:

$$C_i = \beta_0 + \beta_1 t_i \quad (27)$$

The estimate relationship is represented as:

$$\widehat{C_\iota} = \widehat{\beta_0} + \widehat{\beta_\iota} + k_i e_i \quad (28)$$

where:

e_i = estimate of the random error μ

k_i= probability factor

μ = random error in C for observation i; i = 1, 2, 3…n., this depend on nature of the data (Ci) to so many factors k_i.

the estimate regression line is given as:

$$\widehat{C_\iota} = \widehat{\beta_0} + \widehat{\beta_\iota}\, t_i \quad (29)$$

Where, $\widehat{C_\iota}$ = estimated value of C_i given a specified value of t

A simple linear regression model relationship is represented as:

$$1 \ldots\ldots (y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 1,2,3 \ldots n \quad (30)$$

Here, y_i represents assumed variable and x_i is an independent variable which indicated the features of breast cancer dataset, this model is known as the general linear model.

$(\beta_0 + \beta_1 x)$ returns the amount of changes in y value whenever x value is changed with the amount a unit.

Where $\beta_0 + \beta_1$ are the parameters of the model and ε_i refers to the amount of error in explaining the independent variable y_i.

The estimated value of relationship is then given as:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_\iota}\, x \quad (31)$$

$\widehat{\beta_0} + \widehat{\beta_\iota}$ Represents the estimated values of the parameters of the model and mean can be obtained in each of the formulas as:

$$\widehat{\beta_\iota} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_{i^2} - \sum_{i=1}^{n} x_i)2} \text{ and } \widehat{\beta_0} = \hat{y} - \widehat{\beta_\iota}\, \hat{x} \quad (32)$$

$\hat{y}$ and $\hat{x}$ are the mean for each x and y respectively:

The estimated value of the error is therefore presented as follows:

$$\widehat{\varepsilon_\iota} = Y_i = \widehat{Y_\iota} \quad (33)$$

3.11 Hierarchical clustering is a method of grouping identical elements into clusters. An endpoint is a set of clusters, each cluster is isolated from all other clusters, and the objects within each cluster are roughly similar to each other. Clustering is performed by using a distance function to determine an approximate matrix containing the distances between each point. For each pair of clusters, the model computes and joins them all to minimize the maximum distance between clusters.

$$C_{ij}\text{Lij} = \max\{\text{d}(x_a, x_b)\forall(x_a \epsilon\, C_i \; and \; x_b \epsilon\, C_j\}$$

(34)

The average distance between the pairs of clusters is presented as:

.

$$C_i\text{CjLj} = \frac{1}{|C_i||C_j|}\sum_{x_a \epsilon\, C_i}\sum_{x_b \epsilon\, C_j}\; \text{d}(x_a, x_b)$$

(35)

All formed clusters should be considered and an algorithm should be used to compute the sum of the squared distances within the clusters and merge them to reduce the variance of each cluster in the result.

$$\forall C_i\text{CjLj} = \sum_{x_a \epsilon\, C_i}\sum_{x_b \epsilon\, C_j}\; \|x_a - x_b\|^2$$

(36)

Euclidean Distance: The ordinary straight-line distance between two points in Euclidean space is the metric space and is calculated by Equation 37.

$$^{d}\text{Euclidean}\left(\bar{\text{x}}_1, \bar{\text{x}}_2\right) = \|\bar{\text{x}}_i - \bar{\text{x}}_2\|_2 = \sqrt{\sum_t \quad (x_1^i - x_2^1)^2}$$

(37)

3.12 Ensembles: this is a method of combining multiple models to produce a single output. The idea is that combining multiple models yields better results than a single model. Various assembly methods such as bagging, boosting and stacking are possible.

3.13 Bagging or Bootstrapping Ensemble technique: It is a method of creating multiple models on the same set of observations to produce an approximately the same results. Therefore, the whole set is broken down to smaller sets that is with replacement. Various algorithms which use bagging technique are bagging estimator, Random Forest, Extra trees. Mathematically, bagging is calculated as:

$F_{bag} = f_1(x) + f_2(x) + \ldots + f_m(x)$

(38)

3.13 Boosting Ensemble technique. It is a sequential process in which each subsequent model attempts to correct the previous model's errors. This is done by giving higher weights to the observations which were incorrectly predicted. Final model (strong learner) is the weighted mean of all the models (weak learners). AdaBoost GBM, XGBoost, are some of the algorithms which use boosting technique.

Mathematically, Boosting is calculated as:

Formula: (a^t = 1/2 ln 1-e(t)/e(t))

(39)

where,

a ^t – alpha to the power of t , *e(t) – error* with respect to 't'

Weight after time. 't' is given as :

W (i) ^ t+1 / Z. e ^ -at.h1(x).y(x)
(40)

Where,

Z – Normalizing Factor, *W(i) ^t+1* – weight to the power of t + 1 ,

h1(x).y(x) – is sign of current. Output. e^-at.h1(x).y(x) – error to the power of '-' (minus) alpha multiply sign of current output.

3.14 Building a model

Building a model: Building a machine learning models in this study will assist us to understand the problem (and its surrounding system). This stage contains a file that will be trained to recognize certain types of patterns. In this study, models will be trained over a set of data gotten from source, then providing it an algorithm that it will use to reason over and learn from those data. The purpose of this is to discover the detective relationship by using such model. Once this is done, the test dataset will be used to get the accuracy of the hypothesis.

4. Result Discussion

The evaluation of a ML algorithm performance involves testing the proposed model(s) built. In this proposed study, the evaluation will be done by comparing the Machine Leaning Models (Support Vector Machine (SVM), K Nearest Neighbour (KNN), K Means, decision Tree and Naïve Bayes) results with the real data value. In the prediction phases, the test dataset will be used to assess the performance of the models in classifying the classes of breast cancer. Then, the different performance will be used to evaluate the result of the ML model in each module based on their accuracy, precision and Time measure and otherwise as the research continues.

4.1 Application of the conceptual Approach

4.1.1 Virtual Personal Assistants

This concept can be used in the area of virtual personal assistants to aids in discovery of valuable facts, especially when it is been asked through voice or text. A good example of this is speech Recognition, speech to Text Conversion, Natural Language Processing and text to Speech Conversion.

4 .1.2 Fraud Detection

Experts predict online credit card fraud to soar to a whopping $32 billion in 2020. That's more than the profit made some organization in Nigeria and this is more worrisome on the part of the business owners. Fraud Detection is one of the most necessary applications of Machine Learning due to the fact that the number of transactions has through the payment channels such as credit or debit cards, numerous wallets, and smartphones, among others. At the same time,

the rate at which the criminals have become practiced at finding escapes has also increased. By applying this concept will drastically reduce the rate of fraud most especially in Nigeria.

4. .1.3 Medical Diagnosis:

With this advanced of machine learning techniques, medical technology will grow very fast and able to build system models that can predict the exact situation of the aliments thereby creating new approaches of predictions, diagnosis and classifying patients into novel phenotypic groups, and improving prediction capabilities. There is a need for capacity development in this area by providing a conceptual analysis of machine learning alongside with a practical guide to developing and evaluating predictive.

4. 1.4 Decision support

Decision support is another application area where machine learning. Here, algorithms trained on historical data and any other relevant data sets can analyze information and run through multiple possible scenarios at a scale and speed impossible for humans to make recommendations on the best course of action to take. For instance, in agriculture, machine learning-enabled decision support tools incorporate data on climate, energy, water, resources and other factors to help farmers make decisions on crop management. In businesses, decision support systems help management anticipate trends, identify problems and speed up decisions. Information is presented via executive dashboards in the form of charts and other graphics.

4.1.5 Predictions of Traffic:

Global Positioning System navigation services have been using globally, the current locations and velocities are normally kept and saved at a central server for managing traffic. This data is then used to build a map of current locations and velocity traffic. This will prevented the traffic and also will also perform congestion analysis; there are fundamental problem due to less number of cars that are provided with Global Positioning System (GPS). Machine learning in such developments will aids to evaluate the locations where congestion can be found on day-to-day practices.

5.0 CONCLUSION

In this paper, we presented the conceptual architectural design and the sequence for an hybrid Machine Learning algorithms, six classification parameters were examine namely; Support Vector Machine (SVM), Linear Regression (LR), Naive Bayes (NB), k Nearest Neighbor (kNN), K_means and Hierarchical clustering with Bagging and Boosting ensemble Machine Learning Techniques for breast cancer prediction. The target of this concept research is to implement this concept so as to compare different machine learning models performance and determine the foremost accurate machine learning algorithm for the diagnosis of breast cancer. We believed that if this concept is implemented, it will compare, evaluate and establish a hybrid method among many Machine Learning techniques commonly used for breast cancer detection therefore, false positives will be minimized. As future work, this approach needs a more specific way of systematization, perhaps through the aforementioned sophisticated machine learning techniques. Additionally, the used assumptions in this contribution must face well-grounded scientific evaluations in order to ensure their stability since the described approach is conceptual and has not been evaluated yet.

Acknowledgement

We would like to appreciate the Founder of Afe Babalola University, Nigeria - Aare Afe Babalola CON, OFR, SAN for providing financial support for this research. Without this, it would have been impossible for us to complete this work.

References

- Abien Fred (2018).”On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset" International Conference on Advanced Machine Learning and Soft Computing
- Anooj P. et at., (2012): “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules”, Journal of King Saud University - Computer and Information Sciences. Vo 24 (1), Pp 27-40,.
- Azzaw H., Hou J., Xiang Y., Alanni R. (2016): “Lung cancer prediction from microarray data by gene expression programming” IET Syst. Biol., 10 (5) pp. 168-178,
- Chao Tan, Hui Chen and Chengyun Xia (2009): “Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm”. Journal of Pharmaceutical and Biomedical Analysis, 89(3), Pp. 746-752,
- David A.O., Shanmugam V. and Amandeep S. S. (2019): Machine Learning Classification Techniques for Breast Cancer Diagnosis. Materials Science and Engineering Vol 495, No 1.
- Delen, D., Walker, G. and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, 34(2), pp.113-127.
- Dominguez A. and Nandi A.(2009.): Toward breast cancer diagnosis based on automated segmentation of masses in mammograms, Pattern Recognition, vol.42, no.6, pp.1138-1148,
- Endo, A., Takeo, S. and Tanaka, H. (2007). Predicting Breast Cancer Survivability: Comparison of Five Data Mining Techniques. Journal of Korean Society of Medical Informatics, 13(2), p.177.
- Fagbuagun O.A, Folorunsho O.,Adewole L.B and Akin-Olayemi T.H, (2022): Breast Cancer Diagnosis in Women Using Neural Networks and Deep Learning, IRCS-ITB. Vol. 16, No. 2, 152-166.
- Hafizah S., Ahmad S., Sallehuddin R., and Azizah N. (2013) “Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study,” J. Teknol, vol. 65, pp. 73–81.
- Ketan S., Ainhoa C., Lawrence N., and John G. (2012). “A Systematic Review of Barriers to Breast Cancer Care in Developing Countries Resulting in Delayed Patient Presentation” Content syndication partnership advances discovery of research, pp(8),
- Kotsiantis S. B., Kanellopoulos D., and Pintelas P. E. (2006): Data Preprocessing for Supervised Leaning International Journal of Computer Science, Vol 1, No 1, pp 111- 117.
- M.A. Richards, A.M. West combe, S.B. Love, P. Little Johns, and A.J. Ramirez.(2016)., “Influence of delay on survival in patients with breast cancer: a systematic review,” The Lancet, vol. 353, no. 9159, pp. 1119- 1126,
- Nawel Z., Nabiha A., Nilanjan D., and Mokhtar S.: Adaptive Semi Supervised Support Vector Machine Semi Supervised Learning with Features Cooperation for Breast Cancer

Classification. Journal of Medical Imaging and Health Informatics, American scientific publisher, pp. 53-62,

- Niranjana H. and Meenakshi M. (2015).: "ANN, SVM and KNN Classifiers for Prognosis of Cardiac Ischemia- A Comparison": Bonfring International Journal of Research in Communication Engineering, Vol. 5(2), PP 7-11,
- Rani K. U (2010): Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique International Journal of Computer Applications , volume 10, No.3, pp 1-5.
- Seema S,, Sunita S. and Mandeep S (2006): "Cancer detection using adaptive neural network", International journal of advancements in research and technology, Volume 1, (2012).
- Sharma N., and Purkayastha A (2017): Factors affecting quality of life in breast cancer patients: A descriptive and cross-sectional study with review of literature. J. Midlife Health; 8:75–83
- Selvi, S.T., Malmathanraj, R.: Segmentation and SVM classification of mammograms. In: proceedings of the IEEE International Conference on Industrial Technology, pp. 905–910,
- Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. & Bray, F., Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA Cancer J. Clin, 71, pp. 209-249, 2020.
- Tae-WooKim D.H. , Chung-Yill Park: "Decision tree of occupational lung cancer using classification and regression analysis" Safety Health Work, 1 (2) pp. 140-148, (2010).
- Rajendran K, Jayabalan M., Thiruchelvam V, and Sivakumar V., (2019): Feasibility Study on Data Mining Techniques in Diagnosis of Breast Cancer. International Journal of Machine Learning and Computing, vol. 9, No. 3, PP 328-333.
- Venketkumar H., Wan K., and Vikneswaran Z. (2018):"Fuzzy Multi-Layer SVM Classification of Breast Cancer Mammogram Images" International Journal of Mechanical Engineering and Technology (IJMET) Vol 9(8) pp. 1281–1299
- WHO, WHO Position Paper on Mammography Screening, World Health Organization, 2014
- Zięba M., Tomczak J, Marek Lubicz, and Jerzy Świątek (2014).: "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients" Appl. Soft Compute., 14, pp. 99-108.