ABUAD INTERNATIONAL JOURNAL OF NATURAL AND APPLIED SCIENCES
ISSN: 2955-1021
AIJNAS Volume 3, Issue 2, 2023 pp 76-86
https://doi.org/10.53982/aijnas.2023.0302.11-j

https://journals.abuad.edu.ng/index.php/aijnas

Machine Learning Techniques for Breast Cancer Prediction: A Concise Review

***Okebule Toyin, Adeyemo Oluwaseyi Adesina, Abiodun Oguntimilehin, Stephen Eyitayo Obamiyi, Bukola Badeji-Ajisafe**
Department of Mathematical and Physical Sciences, Afe Babalola University,
Ado-Ekiti, Ekiti State, Nigeria.
Corresponding Author: Okebule Toyin
Email: okebulet@abuad.edu.ng

Abstract
Breast cancer is the most trending type of cancer globally with close to two and half million cases recorded based on research by the World Health Organization in 2021 and it is also the most common cancer among women in all countries, posing a major cause for public health concern. In Nigeria and the world at large, over a hundred thousand new cases of cancer occur every year with high death among women. It has been researched that early and accurate detection of breast cancer can aid in the diagnosis of the disease for women and it may also reduce the risk of death rate among women. Literature shows that several machine learning techniques have been carried out on breast cancer diagnosis to help provide accurate technology solutions to early detection. The machine learning techniques used have different accuracy rate which varies for dissimilarity conditions. In this study, we compared different methods with many existing machine-learning techniques commonly used for breast cancer detection and diagnosis. Also, the aim of this review method will show an improvement in accuracy performances by implementing different methods and analysis in existing machine learning techniques to proficiently assist doctors in decision-making on an accurate detection and diagnosis of breast cancer and classifying tumors as benign or malignant thereby reducing the risk of death rate among women.
Keywords: *Breast Cancer, Machine Learning, techniques, Dataset, prediction, Wisconsin, algorithms, mammography, accuracy.*

INTRODUCTION

Breasts are prominent in women's emotional life and are a symbol of womanhood and sexuality (Karesen et *al.,* 1998). Breast cancer determines important alterations in the body image and self-image of the women, which could affect their experience of sexuality and marital relationships. More so, breast cancer treatment causes important physical, social and psycho-emotional changes, with a subsequent decrease in the women's quality of life (Cesnik*et al.,* 2013). Breast cancer is one of the substantial worldwide health challenges facing Nigerian women. The majority of cancer-related deaths according to researcher statistics show that breast cancer is the leading cause of death among women. Breast cancer is also 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 detection of breast cancer can prolong lives by 10 years,and prompt treatment may significantly reduce breast cancer mortality (Jemal*et al.,*2011).Therefore, this disease can negatively impact how a woman performs her role as a wife, mother, and individual in the community, which impacts her socio-occupational functioning. (Van *et al.,*2015).

Due to the complexity of breast cancer, there are four stages of Breast cancer. The stages of breast cancer refer to the size of the cancer and whether it has spread to other parts of the body or not.

Stage 1: This is the earliest stage and it means that the cancer is smaller than two centimeters and has not spread outside the breast.

Stage 2: This is also an early stage of breast cancer that refers to a tumor of more than two centimeters and cancer found in lymph nodes, that is, in the armpit near the breastbone.

Stage 3: This is locally advanced breast cancer, meaning that cancer has spread to the skin or chest wall and to ten or more axillary, breastbone, or collarbone lymph nodes.
Stage 4: This is often called metastatic cancer, meaning that cancer has spread to other parts of the body including the lung, brain, bones, or liver.

REVIEW OF RELATED WORKS USED IN BREAST CANCER PREDICTION

In this section,we present previous Machine Learning methodsin addressing breast cancer detection. These studies have compared and used different machine learning methods based on one factor, two factors, and three or more factors to achieve better performance and accuracy.

Machine Learning Techniques

Machine learning techniques are one of the most trending tools ofthe 21st century for solving problems, and also beneficial in most applications of use due to the capability to make predictions for better decisions (Hamsa*et al.,* 2021).

Types of Machine Learning

The main types of Machine Learning methods are shown in figure 1, known as (i) supervised learning and (ii) unsupervised learning (iii) Reinforcement

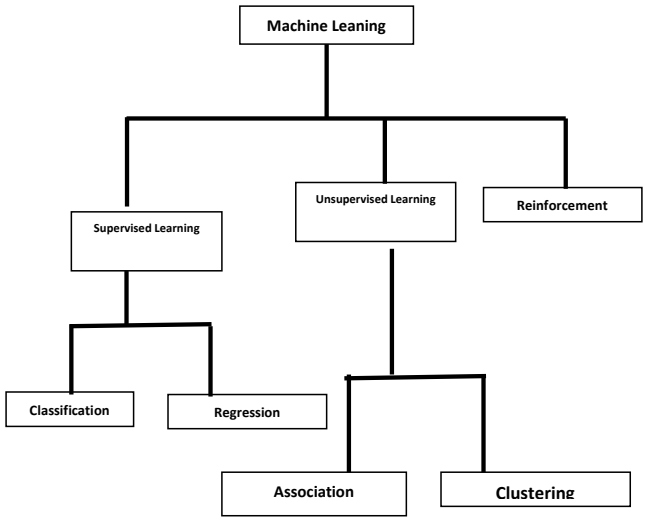

Figure 1. Types of Machine Learning Technique

Supervised Learning:this is a method oftraining the machine using labeled data. Supervised Machine Learning is of two categories, namely; classification and regression. Classification is used to predict a label or a class. The task of classification is to categorize the data into a set of finite classes. In the case of regression problems, a learning function maps the data into a real-value variable. Furthermore, for each new sample, the value of a predictive variable can be estimated and also asked to predict a continuous quantity (Mohamed *et al.,* 2020).

Unsupervised Machine Learning: In unsupervised learning, the machine is trained on unlabeled data without any guidance. This type of machine learning can be used to solve association and clustering problems. The association problems involved discovering patterns in data finding co-occurrences. Clustering is a common unsupervised task in which one tries to find the categories or clusters in order to describe the data items. Based on this process each new sample can be assigned to one of the identified clusters concerning the similar characteristics that they share (Mohamed *et al.,* 2020).

Reinforcement: In reinforcement learning an agent interacts with its environment by producing actions and discovers errors or rewards. It is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions (Mohamed *et al.,* 2020).

Some previous studies are given as followsbased on the One-Factor Machine Learning Model

Seema *et al.* (2012) proposed an adaptive resonance neural network (ARNN) technique with unsupervised learning to detect cancer. The dataset collected for this study contains a total of 699 cases, of which 600 cases were used to train the network. This database contains 9 attributes, therefore categorizing the outputs into two categories benign or malignant. The adaptive resonance neural network techniques used in this study showed an accuracy of 75%. However, the reduction in the dataset used also decreases the accuracy of detection.

Zięba*et al.* (2014)suggested boosting SVM dedicated to solving imbalanced results. The result combined the advantages of ensemble classifiers with cost-sensitive support vectors for unbalanced data. More so, this method is 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.

Chaotan*et al.* (2000) discovered the probability of using decisionthumps as a deficient classification method and trail element analysis for predicting timely lung cancer with the combination of Adaboost. In the study, a cancer dataset was used to identify 9 trace elements in 122 urine samples. The Ada-boost projected results which were compared with the Fisher Biased Ana-lytic (FDA) results. In the test set, 100% of Adaboost's sensitivity was used for both cases that reaching 93.8% of accuracy, more so, 95.7% and 95.1% respectively for case A and case B 96.7%. The Adaboost appeared superior to the FDA and proved that combining Adaboost and urine analysis could be a valuable method through clinical practice for the diagnosis of early lung cancer.

Usha et*al. (*2010) proposed a method of parallel approach with a neural network technique to improvethe classification of diagnosis of breast cancer. The experiment was conducted by considering both single and multilayer neural network models. A back propagation algorithm with momentum and variable learning rate was used to train the networks and a multi-layer perceptron was implemented to yield a better accuracy of 92%. However, in this study, only 11 attributes were used to test the model.

Tae-WooKim*et al.* (2010) proposed a decision tree on occupational lung cancer). The parameter goal was to decide if the state was accepted as lung cancer linked to age, sex, smoking years, histology, industry size, delay, working time, and exposure to independent variables. The presentation was to known lung disease specialists the highest pointer of the CART model. Decision Tree techniques are simple to interpret. It can also be taken as the minimal decision standard of work-relatedness for lung cancer. However,they suffer from overfitting.

Ancy*et al.* (2018) performed classification on single view mammograms on the execution preprocessing of the data set, Gray Level Co-occurrences Matrix feature extraction, Region of Interest segmentation, and Support Vector Machine classification. The experiment results showed that the method used, GCLM extracted features for classifying tumor and non-tumor with SVM classifier could give accurate results. However, the study proposed a method to evaluate two datasets that is the tumor and non-tumor.

Anooj*etal.* (2012) adopted a weighted fuzzy rule for the detection of heart diseases using k-fold cross-validation. The dataset used for this study was obtained from UCI Respiratory this dataset consists of 14 attributes of input and its output value varies 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 showedan accuracy of 58.85%.

Hasan *et al.* (2019) proposed an ANN classifier using PCA preprocessed data as an optimal tool to improve differentiating between benign and malignant tumors on the WBC dataset. They employed the three rules of thumb of PCA namely scree test, cumulative variance, and Kaiser Guttman rule as feature selection. The proposed approach obtained an accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07%, and area under the receiver operating characteristic curve of 0.9994.The result obtained showed that the method can distinguish between benign and malignant cases. However, comparing Machine Learning algorithms was not used for breast cancer diagnosis in this study.

Arpit*et al.* (2015) proposed a genetically optimized neural network (GONN) for breast cancer classification (malignant and benign). They optimized the neural network architecture by introducing new crossover and mutation operators. To evaluate their work, they used WBCD and compared the classification accuracy, sensitivity, specificity, confusion matrix, ROC curves, and AUC under ROC curves of GONN with the classical model and classical Backpropagation model. This method presents a good accuracy classification. However, it can be improved by using a larger dataset than WBCD, and feature extraction to make GONN more efficient for real-time diagnosis of Breast Cancer.

Keles*et al.* (2019) conducted a comparative study on breast cancer prediction and detection using data mining classification. He runs and compares all the data mining classification algorithms in the Weka tool against an antenna dataset. His comparative result shows that the random forest algorithm became the most successful algorithm with 92.2%.

Karl *et al.* (2022)proposed Machine Learning Techniques for Breast Cancer Detection which explores a variety of machine learning techniques and compares their prediction accuracy and other metrics when using the Breast Cancer Wisconsin (Original) data set using 10-fold cross-validation methods. Support Vector Machine Model was employed as the radial basis function kernel which outperformed all others with an accuracy of 99%. However future work can be implemented with other methods to confirm the accuracy achieved in this study. Mehedi*et al.* (2021) proposed Pre-Trained Convolutional Neural Networks for Breast Cancer Detection Using Ultrasound Images. The study considered Grad-CAM and occlusion mapping techniques to examine how well the models extract key features from the ultrasound images to detect cancers using the Adam optimizer in classifying healthy and breast cancer patients. DenseNet201 and ResNet50 show 100% accuracy with Adam and RMSprop optimizers. However,the study was implemented on pre-trained Convolutional Neural Networks and not on Ensemble learning using dataset.

Liu Lei (2018) proposed a model that uses machine learning for cancer detection. In this research, the Logistic Regression algorithm of the Sklearn machine learning library has been used to classify the data sets of breast cancer. Two features of maximum texture and minimum perimeter were selected and the classification accuracy stood at 96.5%.

Yogendra*et al.* (2021) proposed A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches. Data Pre-processing, data imbalance handling feature Selection, Machine Learning Classifiers, and classifier's performance evaluation method was employed in this study. Multilayer Perceptron ANN classifier with Genetic Search algorithm for feature selection achieves an accuracy of 98.59%. However, this method was not implemented on the prognosis of breast cancer using thermal images and IoT-based sensors.

Zemouri*et al.* (2018)proposed a model that uses a Breast Cancer Computer-Aided Diagnosis (BC-CAD) based on joint variable selection and a Constructive Deep Neural Network. Wisconsin Breast Cancer Dataset (WBCD) and real data from the north hospital of Belfort (France) were used to predict the recurrence score of the Oncotype DX. They applied a method to lower the number of inputs for training a deep-learning neural network. Accordingly, the performance of the use of the Deep Learning architecture alone was exceeded by the use of a joint variable algorithm with ConstDeepNet.

Eltonsy*et al*. (2009)proposed a technique for the automated detection of malignant masses in screening mammography. The technique is based on the presence of concentric layers surrounding a focal area with suspicious morphological characteristics and low relative incidence in the breast region. Malignant masses were detected with 92%, 88%, and 81% sensitivity.

Osmanović*et al.* (2019) implemented Machine Learning Techniques for the classification of breast cancer by deploying an artificial neural network with a high and acceptable level of accuracy designed by testing different numbers of hidden layers, and the number of neurons in the hidden layer. the result of the study demonstrated that a feed-forward backpropagation single hidden layer neural network through 20 neurons and TANSIG transfer function which has the highest classification accuracyof 98.9 and 99% accuracy in training and test sets separately. the ANN within this study was configured with nine input neurons (number of attributes) and one output neuron (benign or malignant).however, a graphical user interface (GUI) was not developed for this study.

Hamid *et al.* (2020) proposed Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. They described a machine learning method for identifying a combination of interacting genetic variants (SNPs) and demographic risk factors for BC using the Kuopio Breast Cancer Project (KBCP) dataset. The proposed ML approach quickly evaluates the importance of features (SNPs) to the BC risk prediction accuracy using an XGBoost model and the results also show that demographic risk factors are individually more important than genetic variants in predicting BC risk.

Nawel*et al*. (2016) proposed a conception and implementation of Computer Assisted Detection (CAD) for mammogram image classification. GA-based features selection algorithm on Images from digital X-ray images. The proposed method improved the classification accuracy but was computationally expensive.

Shahnorbanun et*al.* (2018)present Machine Learning Methods for Breast Cancer diagnosis on computer-aided detection (CADe) and computer-aided diagnosis (CADx) techniques using X-ray images. The proposed SVM algorithm identified the different tissue components and modeled the pattern of the relationship between these components spatially and statistically. The technique helps the radiologist and pathologist reduce their workload by automating the automation for decision-making, especially for common and mundane cases but combining these tissue components' features resulted in dense feature vectors, which suffer from overfitting.

Tahmooresi*et al.* (2018) proposed the Early Detection of Breast Cancer Using Machine Learning Techniques with a hybrid model combined several Machine Learning for testing data set using the BCD dataset. The findings in this research showed that SVM is the most popular method used for cancer detection applications. The authors stated that SVM can either be used alone or pooled with another technique to advance the performance. The maximum achieved accuracy of SVM was 99.8%.This method was applied and tested on other data sets to check the performance of different data types but not on datasets like mammograms and ultrasound.

Khourdif et*al.* (2018) proposed the best machine learning for breast cancer prediction. The Dataset was divided into two: K-fold validation technique and was applied before the features selection and extraction method. SVM provides more accuracy for prediction. The predictive model was designed; SVM provided 99.7% accuracy for the benign class and 94.6% for the malignant class. The turnaround time and error rate of SVM are lesser than other algorithms and SVM provided 99.7% accuracy for the benign class and 94.6% for the malignant class. The good and appropriate selection of method was important for the evaluation of the machine learning algorithm Confusion matrix and was designed for expected class result, the matrix correctively predicted the instances but with prediction time was maximum.

Said *et al*. (2018) experiments hyper parameter Optimization for Breast Cancer Prediction. The HPO technique through the clustering method was used to get the best suitable prediction algorithm for Breast cancer with the WBCD dataset. Hyperparameters through the clustering method provided the highest accuracy. Hyperparameters handled both categorical and continuoustypes of data more effectively but selected futures also provided some redundant data. However, the BCOAP model consists of too many phases each phase took a lot of time for the evaluation of breast cancer data.

Asoke*et al.* (2008) presented a Classification of Breast Masses Using Selected Shape, Edge-sharpness, and Texture Features with Linear and Kernel-based Classifiers. Principal component analysis (PCA) was adopted as feature selection performed by a genetic algorithm based on several criteria on the Breast cancer dataset. Principal Component Analysis (PCA) was used asa feature dimensionality reduction tool, which concentrates on significant projections of features. However, PCA does not always improve the classification performance of most classifiers and, PCA does not identify specific features.

Singhal*et al.* (2018) proposed Artificial Neural Network for breast cancer with backpropagation algorithm that used prediction through the ANN algorithm, every hidden layer provided a different accuracy during evaluation. Multi-layered Neural network created weight arbitrary that provided the Mean Square Error whose rate is too low and the feed-forward algorithm helped to reduce the error through weight modification. However, it requires high processing and time for a large number of data that affect the overall accuracy of data and to achieve good accuracy, precision, and sensitivity of data, a large number of samples are needed for computations.

Ahmed *et al*. (2019)proposed Machine Learning Techniques for the Classification of Breast Cancer. The use of Artificial Neural Network was implemented with a high and acceptable level of accuracy and was designed by testing different numbers of hidden layers, the number of neurons in the hidden layer using the WBCD dataset. The result of the study demonstrated the highest classification accuracy of 98.9 and 99% accuracy in training and test sets separately and 98.9 and 99% accuracy was achieved. However, Graphical User Interface (GUI) was not developed for this study.

Two-Factor Machine Learning Model

Rashmi *et al.* (2020)aimed at optimizing the testing algorithm on SVM and Ga-clustering-based feature selection approach for breast cancer detection. In the experiment, four Weka clustering strategies with genetic clustering were equated then a comparison of results revealed that sequential minimal optimization (S.M.O.) is better than I.B.K. and B.F. The research focused on genetic programming and machine learning algorithms that identified benign and malignant breast cancer. A comparison of results reveals that sequential minimal optimization (S.M.O.) is better than I.B.K. and B.F. Tree processes, i.e. 97.71%. However, the approach was based only on a selection approach based on genetic algorithms combined with 5k fold-cross-validation SVM classification.

Tan *et al.* (2009) experiments with early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm. The method employed in this study was Adaboost classifiers and Fisher Discriminant Analysis (FDA). Lung using cancer dataset. The Adaboost method employed in this study achieved high sensitivity and best performance and was very simple to implement 95.1%, 96.7%, for case A and case B though It is very sensitive to noisy data.

Mariam *et al.* (2018) proposed Naive Bayes and K Nearest Neighbors classifiers for breast cancer by comparing the accuracy using cross-validation in which KNN outperformed Naive ayes by 97.51%accuracy while Naive Bayes Classifier achieved 96.19% accuracy.

Yasmeen *et al.*(2012) presented an intelligent diagnosis system for breast cancer classification. The authors applied different machine learning algorithms on six data models and predicted that PNN and SVM were used to produce an efficiency of 99.7% in terms of sensitivity and accuracy. The authors advised that the methodology employed in this study can be implemented in different medical datasets in future related works.

Selvi*et al.* (2006) proposeda framework to detect breast cancer using KNN and SVM on the dataset collected from the UCI repository to detect breast cancer with respect to the results of accuracy the efficiency of the algorithm is also measured and compared. The method of machine learning algorithms applied to these datasets show different levels of accuracy ranging between 94.36 and 99.90%. However, the study was only used to determine the accuracy of the model.

Ojha*et al.* (2017)use different ML algorithms to predict recurrent cases of breast cancer using the Wisconsin Prognostic Breast Cancer (WPBC) data set. The evaluation result produced SVM and decision tree (C 5.0) as the best predictors with 81% accuracy, while fuzzy c-means were found to have the lowest accuracy of 37%.

RafeeK*et al.* (2019) proposed A Breast Cancer Using Machine Learning Techniques. The study focused on the advancement of predictive models by using a supervised machine learning method to achieve better accuracy. The classifiers; J48 and LMT were compared with or without a random projection filter. LMT outperformed others with an accuracy of 97.368%. The accuracy of using filter is secondary. The result concluded that the classifiers without filter are efficient for the breast cancer detection.

Ahmad *et al.* (2015) compared the performance of decision tree (C4.5) SVM, and ANN using data mining techniques, the authors developed models to predict the recurrence of breast cancer by analyzing data collected from the ICBC registry. 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.

Susmitha*et al.* (2019) proposed the Analysis of Machine Learning Algorithms for Breast Cancer Data which developed and compared two machine learning techniques to classify the dataset WBCD (Wisconsin Breast Cancer Diagnosis) as either malignant or benign breast lumps. Confusion matrix was used to plot both algorithms to determine the best accuracy which was achieved by the logistic regression algorithm with 94.73% and by the decision tree classification algorithm with 92.39%.

Nikita *et al.* (2022) proposed Breast Cancer Detection Using Machine Learning. The study was tailored to two machine learning techniques for breast cancer classification, namely KNN (k-Nearest-Neighbor) and Naive Bayes using the Wisconsin Breast cancer dataset. The goal of this approach was based on cancer classification using two classifiers in a data set. Each classification uses two classifiers in the data set and each classifier's performance was evaluated in terms of accuracy, training process, and testing process.

Akshay*et al.* (2020) surveyed Breast cancer classification using Machine learning with GLCM feature extraction and wavelet transform to detect both mass lesions and microcalcifications. The result shows a fair improvement in the recall rate. The results obtained are, however, the result were highly dependent on the dataset and was not implemented with Ensemble learning.

Majid *et al.* (2007) experimented Breast Cancer Detection from FNA Using SVM and RBF Classifier. The benefits application of support vector machines (SVMs) and radial basis function (RBF) for breast cancer detection were considered with the Breast cancer dataset. The results showed that SVM classifiers and the RBF (Radial Basis Function) classifier can be used as efficient tools for breast cancer detection with a detection accuracy of up to 98%. Without the value of the regularization parameter to control the tradeoff between the complexity of an SVM and the number of non-separable points, the method of the SVM classifier cannot be successful.

Shereen*et al.* (2015) developed a Prototype for Breast Cancer Detection and Development Probability Expert System – Towards a Supportive Tool. The proposed prototype for breast cancer detection was to identify the stage of breast cancer using the Breast cancer dataset. The research proposed a prototype that was able to detect the existence of breast cancer in the patient using periodic mammographic examinations with identifying the stage of the disease based on the size of the cancerous tissues. A comprehensive validated expert system was not the focus in this paper to provide more useful information to users for the required lifestyle to avoid disease development.

Shahariar*et al.* (2019)conducted early Detection of Breast Cancer Using Artificial Intelligence. The research proposed Artificial Neural Network and Convolutional Neural Network algorithms with and without PCA on a dataset. The research was proposed for prognosis, and diagnosis and to assist doctors in making the final decision more accurately in a shorter time span with less human and monetary resources. However, the inadequate instances of the data and the alteration of data in order to use it for CNN proved to be a challenging part.

Three or more-factor Machine Learning Model

Shanjida*et al.* (2019)experimented prediction of breast cancer using Naive Bayes, KNN and J48. The dataset was divided into two parts, one is training data and another one is testing data.Ten-fold cross-validation was applied for the evaluation algorithms with the WBCD dataset. The most suitable technique for the prediction of cancer dataset is to classify the data according to the similarity of each instanceproviding good accuracy for both training data and testing data. However, the testing phase was slow and also took much time, it was difficult to choose the required K value and to predict the new data K-nearest only find the nearest neighbor from training data.

Niranjana *et al.* (2015) proposed a comparison of the performance of Artificial Neural Network (ANN) Support vector machine (SVM) and K-Nearest-Neighbour (KNN) models for cardiac ischemia classification. The proposed ANN, SVM, and KNN models receive the

morphological features extracted from preprocessed ECG beats. The performances of all models are compared and validated on the physiobank database in terms of accuracy, sensitivity and specificity. The experimental results confirmed that the ANN model outperformed with testing classification accuracy of 96.62%. This accuracy obtained is considerably high in comparison with SVM and KNN classifiers. However, the model was not performed on breast cancer.

Deepa *et al* .(2021)proposed research work on supervised learning algorithm for four different classifiers, Artificial Neural Network, Support Vector Machine, K- Nearest Neighbor (KNN) Support Vector Machine, Linear Discriminant Analysis, and Weighted K- Nearest Neighbor for breast cancer classification. The study proposes the difference between the abovementioned classifiers also define their accuracy. The performance of the different classifiers was determined by their specificity, accuracy, precision, sensitivity, and recall. The result showed that ANN outperformed other classifications with the highest accuracy of 97.60 %.

Sachin*et al*. (2017) proposedthe Detection and Classification of Blood Cancer from Microscopic Cell Images Using a Support Vector Machine, K- Nearest Neighbour and Neural Network Classifier. Automatic detection and classification of AML approach in blood smear is presented. The proposed method performed the segmentation and classification of WBCs and RBCs well when results were compared with the ground truth, KNN with 61.11% and SVM with 83.33% accuracy. However, accuracy measurement was taken with Neural Network.

Nematzadeh*et al.* (2020) conducted a comparative study on decision trees, NB, NN, and SVM with three different kernel functions as classifiers to classify WPBC and Wisconsin Breast Cancer (WBC). The experimental result showed that NN (10-fold) had the highest accuracy of 98.09% in the WBC dataset, while SVM-RBF (10-fold) had the highest accuracy of 98.32% in the WPBC dataset.

Gopi*et al.* (2020)experimentedwith the three enhanced single methods of machine learning algorithms that classify highly correlated features closely related to malignant identification. The proposed methods assist health care and medical researchers in breast cancer identification. In their study, three different supervised machine learning models were adopted namely support vector machine, Linear Regression, and K-Nearest Neighbour. Two different experiments (total features and limited features) were conducted. The result was documented as 98.44% for SVM, 100% for LR, and 93.75% for KNN by selective features. However, the linear regression model with limited features could be the best solution to improve the diagnostic accuracy of breast FNA and studies are needed for confirming the study outcomes using large data of biomarkers, or multi-centric databases.

Westerdijk*et al.* (2017) studied several machine-learning techniques for the prediction of breast cancer cells. The performances of the models were tested by looking at their accuracies, sensitivities, and specificities. Accuracy scores of LR, Random Forest, SVM, neural network, and ensemble models were compared. The prediction of breast cancer should be improved with the accuracy score and other methods in future works.

Bazazeh*et al.* (2016) investigated SVM, random forest (RF) and Bayesian networks (BN) for breast cancer diagnosis and performed a comparative analysis of them. The WBC dataset was used as a training set to evaluate the performance of the machine learning classifiers. The experimental results showed that SVM had the best performance in terms of accuracy, specificity, and precision, while RF had the highest probability of correctly classifying tumors.

Rajendran*et al. (*2019)conducted a feasibility study on data mining techniques in the diagnosis of breast cancer. They have reviewed a lot of papers to provide a holistic view of the types of data mining techniques used in the prediction of breast cancer. The result shows that the data mining techniques that are commonly used include Decision Tree, Naïve Bayes, Association rule, Multilayer Perceptron (MLP) Random Forest, and Support Vector Machines (SVM). The overall performance of the techniques differs for every dataset. On the Wisconsin Breast Cancer Dataset, the random forest classifier produced a better performance with an accuracy of 99.82%.

Saniya *et al.* (2020)examined the Prognosis of Breast Cancer from Mammograms with predictive models; Artificial Neural Networks (ANNs) Bayesian Networks (BNs) Support Vector Machines (SVMs) and Decision Trees (DTs) on different input features. The result shows that the Support Vector Machine got the highest accuracy of 92.78%. The study focused on the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. However, the Ensemble learning method was not implemented in this study to comprehend the accuracy and precision.

Krant*iet al.* (2020) worked on the Prediction of Breast Cancer Using Ensemble Learning. This was done on a

relative study of the implementation of models using a Support Vector Machine (SVM) k-Nearest Neighbor (k-NN) and Decision tree on the dataset. The results of accuracy, precision, sensitivity, specificity, and False Positive Rate and the efficiency of each algorithm were measured and compared, and the models were integrated with the help of ensemble learning. Of all the three applied algorithms, SVM gives the highest accuracy of 99% when compared to other two algorithms. However, the Ensemble learning was only integrated on Supervised Machine Learning.

RamikRawal (2020) worked on Breast Cancer Prediction Using Machine Learning using SVM, Logistic Regression, Random Forest and KNN for predicting breast cancer and the outcome were compared in the study using different datasets. The accuracy obtained by SVM with 97.13% is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN. However, the outcome was only measured on accuracy and other parameters were not used.

Toukir*et al.* (2020) discussed an analysis of the Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction with five classification algorithms that is Naïve Bayes, Support Vector Machine (SVM) Multilayer Perceptron (MLP) J48 and Random Forest. The result shows that Naïve Bayes is superior to others compared with standard parameters of 95.99% accuracy.

Reza *et al.* (2022) worked on the Prediction of Breast Cancer using Machine Learning Approaches with The random forest (RF) neural network (MLP) gradient boosting trees (GBT) and genetic algorithms (GA). The result showed that RF presented higher performance compared to other techniques with an accuracy of 80%. For future work, applying different machine learning approaches, accessibility to bigger datasets from different institutions (multi-center study) and considering key features from a variety of relevant data sources could improve the performance of modeling.

Nanchen*et al.* (2021) proposed Breast Cancer Diagnosis using a Machine Learning Approach with Logistic Regression, Random Forest Classifier, and K-Nearest Neighbors. Random Forest Classifier outperformed with an accuracy of 96%. However, the dataset could not capture the demographic effects of breast cancer on the diagnosis.

Sweta*et al.* (2021) employed CNN as a classifier model and Recursive Feature Elimination (RFE) for feature selection with five algorithms namely, SVM, Random Forest, KNN, Logistic Regression, Naïve Bayes classifier were compared in the study. The experiment proved that CNN outperforms the existing methods when in accuracy and precision. In future work, this method could be implemented on different breast cancer dataset.

Deneshkumar*et al.* (2018) predicted breast cancer using five prediction algorithms. That is, Naive Bayes, Logistic regression, Decision tree, Random forest, and Support vector Machine. The prediction was done on the Wisconsin breast cancer dataset. The result shows that, without any feature selection, the support vector machine is the best algorithm with an accuracy of about 95.6%. While logistic regression showed a better performance compared to others with feature selection, which was nearly 97%.

Dharambir*et al.* (2022) discussed various risk factors and advanced technology available for breast cancer diagnosis to combat the worst breast cancer status and areas that need to be focused on for the better management of breast cancer by proposed Global Increase in Breast Cancer. The effectiveness of preventive and screening programs also depends on the economic condition of the country. Therefore, good validation of the biomarkers is required to decide the region-specific cut-off values.

Kemal *et al.* (2018) proposed a hybrid approach based on mad normalization, KMC-based feature weighting and AdaBoostM1 classifier. The detection of the presence of breast cancer is done in three steps: In the first step, the dataset was first normalized by the MAD normalization method. In the second step, k-means clustering (KMC) based feature weighting has been used for weighting the normalized data. Finally, the AdaBoostM1 classifier has been used to classify the weighted data set. The Breast Cancer Coimbra dataset (BCC) taken from the UCI machine learning database was used. This method shows good results in terms of accuracy. However, it is a computationally expensive method.

Osareh*et al.* (2010) investigated the issues of breast cancer diagnosis and prognostic risk evaluation of recrudescence and metastasis using SVM, K-nearest neighbor (KNN) and probabilistic neural network (PNN). These classifiers were combined with signal-to-noise ratio (SNR) feature ranking method, sequential forward selection-based (SFS) feature selection and PCA feature transformation. The SVM-RBF was found to obtain the best overall accuracies of 98.80%.

Manav*et al.* (2022) proposed Data Visualization and comparative analysis between Support Vector Machine (SVM) Decision Tree, Naive Bayes (NB) K Nearest Neighbours (k-NN) Ensemble learning method and Random Forest conducted on Wisconsin breast cancer

Dataset. Experimental results show that the ensemble method offers the highest accuracy (98.24%) with the lowest error rate. However, the study was not combined with unsupervised machine learning techniques.

Bellaachia*et al.* (2006) looked into the use of Naïve Bayes, the back-propagated neural network and the C4.5 decision tree algorithms on SEER dataset which contained 16 attributes and 482,052 records. The dataset is considered to be ideal due to the large amount of patients and a moderate number of attributes. From their experiment, the C4.5 algorithm outperformed the rest with an accuracy of 86.7%.

Endo *et al.* (2007) proposed predicting Breast Cancer Survivability: Comparison of Five Data Mining Techniques. Statistical methods (Logistic Regression) were used to evaluate the prediction models using the Breast cancer dataset. The accuracy was 5.8±0.2%, 84.3±1.4%, 83.9±0.2%, 82.3±0.2%, 75.1±0.2% for the Logistic Regression, Artificial Neural. Naive Bayes, Decision Trees (ID3) Decision Trees (J48) respectively, 5.8±0.2%, 84.3±1.4%, 83.9±0.2%, 82.3±0.2%, 75.1±0.2%. The accuracy of Decision Trees was the worst among the prediction models used.

Hiba *et al.* (2020) proposed Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis using Support Vector Machine (SVM) Decision Tree, Naive Bayes (NB) and k-nearest neighbors (k-NN). The accuracy obtained by SVM (97.13%) is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 %, SVM with (97.13%) Naïve Bayes with 95.12 % and k-NN with 95.28 %. However, Implementation was done only on supervised machine learning Algorithm.

Delen*et al.* (2005) proposed predicting breast cancer survivability: a comparison of three data mining methods, Data mining algorithms (artificial neural networks and decision trees) along with a most statistical method (logistic regression) were used to develop the prediction models with Breast cancer dataset. The results indicated that the decision tree is the best predictor with 93.6% accuracy on the holdout sample artificial neural networks came out to be the second with 91.2% accuracy.

Padmapriya*et al.*(2016) conducted an analysis of Breast cancer through Classification Algorithm performance classification algorithm that was analyzed in terms of their accuracy, sensitivity and precision with the Breast cancer data set. Comparison of each classification algorithm was done through the evaluation of weighted average values. CART algorithm provided better accuracy for prognoses of breast cancer with minimum time. Tree algorithm J48, CART, ADtree. However, the evaluation phase took too much time and the model was designed for comparative analysisof the data mining decision.

Bharat *et al.* (2018)experimented Breast Cancer risk prediction and Diagnosis using Performancemetrics evaluation on C4.5, SVM, NB and KNN in terms of their accuracy, precision and sensitivity, SVM provided the highest accuracy result than others on BCD dataset ROC curve provided the good evaluation of each algorithm and prediction of correctively classified instances rate higher through SVM algorithm. Also this algorithm provided lower error rate value. Processing time of SVM was 0.007 while KNN was 0.01sProcessing time of SVM was 0.007 while KNN was 0.01s.Model was designed to train data for the evaluation of correctly and in correctively classify the instance that was difficult and complex task.

Table 1: Summary of the Machine Leaning Factors
The chart below shows the summary of the Machine Learning reviewed based on one factor, two factors and three or more factors with various accuracy achieved on tested data.

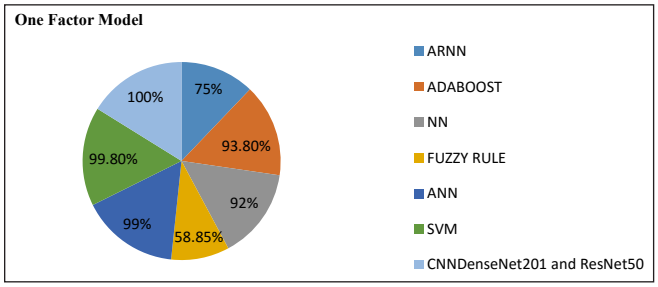

Figure 2. One Factor Model

In one factor machine learning, we reviewed the performance of each single model where the CNN with DenceNet201 and ResNet50 outperformed with 100% accuracy for detection of breast cancer than others. However, Fuzzy rule acquired the least accuracy with 58.85%. The different factors here were tested on different dataset of breast cancer. To establish these performances, there is a need for these factors to be tested on a single datasets.

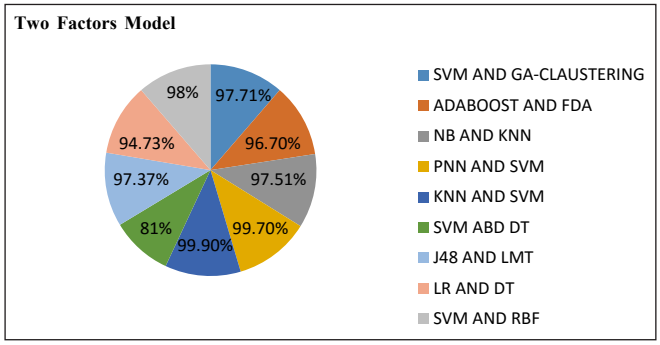

Figure 3. Two Factors Model

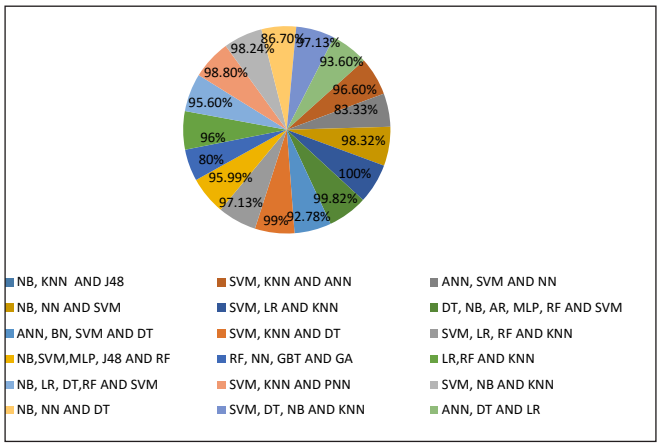

Figure 4: Three or more Factors model

CONCLUSION

We present different Machine Learning Techniques based on one, two three or more factors with different datasets used, the accuracies and limitations in the most recent studies related to breast cancer diagnosis. SVM shows the best accuracy as a single factor, PNN and SVM indicated the best accuracy for two-factors, while combination of SVM, k-NN, LR and as three or more factors indicated the best accuracy. This work recommends large consistent datasets with hybrid method in order to develop a robust model for early breast cancer detection and prediction.

Declaration

The corresponding author declares that there is no conflict of interest on behalf of the other authors.

Acknowledgement

We would like to express our gratitude to Aare Afe Babalola CON, OFR, SAN, the founder of Afe Babalola University-Nigeria, for providing an appreciative environment for research work.We also want to thank Dr. Azeez T.M. from the College of Engineering at Afe Babalola University, Nigeria, for his valuable suggestions on this study.

REFERENCES

Aastha, G,Himanshu, S. & Anas, A. (2021). A Comparative Analysis of K-means and Hierarchical Clustering. EPRA International Journal of Multidisciplinary Research (IJMR).Jagan Institute of Management Studies, sec-5, Rohini, 7(8):412-418.

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.

Akram, M., Iqbal. M.,& Daniyal. M., & Kha, A.U(2017). Awareness and current knowledge of breast cancer. Biological Research 50(33).

Ali, E., & Feng, W. (2013). Breast Cancer classification using Support Vector Machine and Neural Network: International Journal of Science and Research,23:19-7064.

Aminikhanghahi S., Shin S., Wang W., Jeon S., Son S., and Pack C., Study of wireless mammography image transmission impactson robust cyber-aided diagnosis systems: Proc. 30th Annu. ACM Symp. Appl. Comput. - SAC '15: 2252–2256, 2015.

Andre, R., &Rangayyan, M.(2006). Classification of breast masses in mammograms using neural networks with shape, edge sharpness, and texture features, J. Electron. Imaging 15(1) 13019.

Avramov, T.,&Si, D. (2017). Comparison of Feature Reduction Methods and Machine Learning Models for Breast Cancer Diagnosis. Proc. Int. Conf. Comput. Data Anal. – ICCDAL.17, 69–74.

Ayeldeen, H., Elfattah, M., Shaker, O., Hassanien, A., &Kim, T. (2015). Case-Based Retrieval Approach of Clinical Breast Cancer Patients. 2015 3rd International Conferenceof Computer. Information and. Application, 38–41.

Azar, A., &El-Said, S. (2014). Performance analysis of support vector machines classifiers in breast cancer mammography recognition, Neural Comput. Appl.24 (5):1163–1177.

Bevilacqua, V., Brunetti, A., Triggiani, M., Magaletti, D., Telegrafo,M., &Moschetta, M. (2016). An Optimized Feed-forward Artificial Neural Network Topology to Support Radiologists in Breast Lesions Classification. In Proceeding 2016 Genetic Evolution of Computer Conference Companion - GECCO '16 Companion, 1385–1392.

Caplan, L. (2014). Delay in breast cancer: implications for the stage at diagnosis and survival. Frontiers in Public Health, 2(87)1–6.

Bojana, R.,&Andjelkolvic, C.(2020). Machine Learning Approach for Breast Cancer Prognosis Prediction. Computational Modeling in Bioengineering and Bioinformatics, 41-68.

Cesnik, V.,Vieira, E., Giami, A., Almeida, A., Santos, D., &Santos, M. (2013). The sexual life of women with breast cancer: meanings attributed to the diagnosis and its impact on sexuality. EstudPsicol, 30(2)187–197.

Chowdhary, C., &Acharjya, D. (2016). Breast Cancer Detection using Intuitionistic Fuzzy Histogram Hyperbolization and Possibility Fuzzy c-mean Clustering algorithms with texture feature-based Classification on Mammography Images. In Proceedings of the International Conference on Advances in Information Communication Technology and Computing, Bikaner, India,1–6.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, 34(2) 113-127.

Duda, R., Hart, P.,&Stork, D. (2000). Text of Dimensionality Reduction, Pattern classification, 2nd edition, Wiley-Inter science. ISBN 0-471-05669-3, SECTION (8)79/679.

Hamsa, B.,&Wichinpong, P.(2021). Improving Human Decision-Making with Machine Learning University of Pennsylvanial. Berkeley, 4(2)1-2.

Harry Zhang. (2005). Machine Learning and Neural Network Approaches to Feature Selection and Extraction for Classification. Exploring Conditions for the Optimality of Naïve Bay. International Journal of Pattern Recognition and Artificial Intelligence, 19(2) 183-198.

Mohamed, A., Jumeily, O.,Jumeily, O., Ahmed, J., &Aljaaf A., J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science Computer Science Department. Conway, Arkansas, US, 1-30.

Mounita, G., Mohsin, S., Laboni, A.&Alshamrani, M.(2021). A Comparative Analysis of Machine Learning Algorithms to Predict Liver Disease. Department of Computer Science, College of Computers and Information Technology Intelligent Automation & Soft Computing, 30(3) 918-928.

Murugan, S., Muthu, B.,&Amudha, S. (2017). Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest. International Conference on Current Trends in Computer, Electrical. Electronics and Communication (CTCEEC) 1-25.

Kaminskal. M., Ciszewski, T., Łopacka-Szatan, K., Miotłal. P.,&Starosławskal. E. (2015). Breast cancer risk factors.PrzegladMenopauzalny-Menopause Review, 14(3)196,

Witten, I., &Frank, E. (2006). Data mining: practical machine learning tools and techniques.BioMedical Engineering on Line. Computer Science Research Institute, University of Ulster, Jordanstown, Co. Antrim, BT37 0QB, Northern Ireland, UK, 1-2.

Seo, L., Paulina, M., Ryan, P., and James, J. (2020). Towards Standardization of Data Normalization Strategies to Improve Urinary Metabolomics Studies by GC×GC-TOFMS. Development and Application of Statistical methods for Analyzing Metablolomicsdata. Department of Chemistry, University of Albertal. Edmonton, AB T6G 2G2, Canada. 10(9) 376.