Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 3 (2021) 298–307

Journal of the Nigerian Society of Physical Sciences

Optimized Breast Cancer Classification using Feature Selection and Outliers Detection

A. B. Yusuf[a], R. M. Dima[b,*], S. K. Aina[c]

[a]*Department of Information and Communication Technology, Usmanu Danfodiyo University, Sokoto State*
[b]*Department of Computer Science, Federal University Dutsinma, Katsina State*
[c]*Department of Computer Science, Federal University Gashua, Yobe State*

Abstract

Breast cancer is the second most commonly diagnosed cancer in women throughout the world. It is on the rise, especially in developing countries, where majority of the cases are discovered late. Breast cancer develops when cancerous tumors form on the surface of the breast cells. The absence of accurate prognostic models to assist physicians recognize symptoms early makes it difficult to develop a treatment plan that would help patients live longer. However, machine learning techniques have recently been used to improve the accuracy and speed of breast cancer diagnosis. If the accuracy is flawless, the model will be more efficient, and the solution to breast cancer diagnosis will be better. Nevertheless, the primary difficulty for systems developed to detect breast cancer using machine-learning models is attaining the greatest classification accuracy and picking the most predictive feature useful for increasing accuracy. As a result, breast cancer prognosis remains a difficulty in today's society. This research seeks to address a flaw in an existing technique that is unable to enhance classification of continuous-valued data, particularly its accuracy and the selection of optimal features for breast cancer prediction. In order to address these issues, this study examines the impact of outliers and feature reduction on the Wisconsin Diagnostic Breast Cancer Dataset, which was tested using seven different machine learning algorithms. The results show that Logistic Regression, Random Forest, and Adaboost classifiers achieved the greatest accuracy of 99.12%, on removal of outliers from the dataset. Also, this filtered dataset with feature selection, on the other hand, has the greatest accuracy of 100% and 99.12% with Random Forest and Gradient boost classifiers, respectively. When compared to other state-of-the-art approaches, the two suggested strategies outperformed the unfiltered data in terms of accuracy. The suggested architecture might be a useful tool for radiologists to reduce the number of false negatives and positives. As a result, the efficiency of breast cancer diagnosis analysis will be increased.

DOI:10.46481/jnsps.2021.331

Keywords: Breast cancer, Machine learning, Accuracy, Feature selection and outliers

Article History :
Received: 05 August 2021
Received in revised form: 10 October 2021
Accepted for publication: 01 November 2021
Published: 29 November 2021

Communicated by: T. Latunde

1. Introduction

Cancer is a group of disorders characterized by the growth of abnormal cells that have the ability to infiltrate or spread throughout the body[1]. Breast cancer is the world's second most common disease and public health concern, especially among women, with high mortality and substantial morbidity[2]. Breast cancer is on the rise in emerging countries, and a five-year study indicated that it is the most common malignancy[3]. According to the American Cancer Society, 93,600 new instances of breast cancer are diagnosed each year in Africa, with

*Corresponding author tel. no:
Email address: rocinta976@gmail.com (R. M. Dima)

around 50,000 fatalities. Breast cancer was diagnosed in 2.3 million women, resulting in 685,000 deaths in 2020, according to the World Health Organization (WHO, 2021).

It is a disease characterized by aberrant cell proliferation in the breast[4], which is caused largely by DNA mutations. Breast cancer tumors are classified as either malignant or benign [5]. This classification is applied in the analysis of breast tumors, lumps, or any other abnormal development in the breast tissue. Cancer that is classified as benign is typically not life-threatening and has a greater chance of survival, whereas cancer that is classified as malignant is life-threatening [6]. A malignant tumor can develop fast, infiltrating the lymph system and encroaching on other healthy tissues in the surrounding area, causing disastrous effects; on the other hand, a benign tumor cannot grow beyond a specific size and remains confined inside its bulk. Early cancer identification guarantees successful therapy and enhances the likelihood of survival [7].

Scientists have attempted to pinpoint the specific cause of breast cancer since there are only a few risk factors that promote a woman's chances of developing the disease. Breast cancer risk factors include age, genetic risk, family history, obesity, gene variation, smoking and alcohol consumption. Due to the small size of the cancer cell as seen from the outside, it is nearly hard to identify breast cancer in its early stages. Mammography, ultrasound [8], dynamic MRI[9], and elastography are the only ways to detect cancer at an early stage[10]. In many cases, clinicians would be required to read a large amount of imaging data, which would compromise accuracy. This method is extremely time-consuming and, in some cases, it incorrectly diagnoses the cancer.

Medical professionals continue to make this sort of diagnosis in order to see which one has the most impact. In recent years, however, machine learning (ML) [11]–[14], deep learning [15], [16], and bio-inspired computing [17] approaches have been employed in a variety of medical diagnoses. Machine learning in the detection of breast cancer has been the subject of several studies [18]–[21]. Several studies use different datasets collected from the University of California-Irvine (UCI) repository for clinical prediction of this disease. Among these are Wisconsin Breast Cancer Dataset (WBCD), Wisconsin Diagnostic Breast Cancer (WDBC) and Wisconsin Prognostic Breast Cancer (WPBC) dataset to mention few. Regardless of the nature of the dataset, the focus of the research is always aimed at enhancing accuracy of the prediction in order to correctly diagnose the cancer. However, despite the popularity of ML algorithms modalities proven on different breast cancer dataset, it still cannot offer accurate and consistent outcome in diagnosis unless improved with some data mining techniques [22]. Among the popular dataset used is WDBC with continuous-valued data problem [23] hence finding the linearity among the features pose a difficulty [24], [25] when applied to ML algorithms thus leading to poor accuracy when applied on some algorithms.

Recently, in the research work carried out on WDBC dataset, study and analysis of ML algorithms were reported along with the approaches used in improving the performance of ML algorithms [25]–[27]. In their work, [25] proposed a method using clustering and noise removal on WDBC dataset before it was applied on some ML algorithms. In the research, Expectation Maximization (EM) was used for data clustering, Classification and Regression Trees (CART) automatically generated the fuzzy rules from the data hence causing removal of noise while the Principal Component Analysis (PCA) as a dimensionality reduction technique was used to overcome the multicollinearity issue in the data. The proposed technique was evaluated with WDBC and Mammographic mass datasets; then its effectiveness was demonstrated. When compared to PCA-Support Vector Machine (PCA-SVM,), PCA-K Nearest Neighbour (PCA-KNN) and Decision Tree (DT), EM-PCA-CART-Fuzzy Rule-Based had the greatest accuracy of 93.2%, whereas PCA-SVM, PCA-KNN and DT had an accuracy of 86.7%, 82.3% and 92.9%, respectively.

In order to improve classification accuracy of breast cancer disease, [27] applied preprocessing step on WDBC dataset. Here, features were selected using gain ratio and modeled with six algorithms using 10-fold cross-validation method. The accuracy of these algorithms was: SVM-Linear (98.07%), KNN at k=3 (97.36%), Naïve Bayes (95.08%), J48 (98.07%), Multilayer perception (98.41%) and Random Forest (98.77%). The results demonstrated a considerable improvement above the state-of-the-art since RF performed the best.

Similarly, [26] focused on integrating ML algorithms with different feature selection methods and compared their performances to identify the most suitable approach.. The selected features were Correlation based Feature Selection (CFS), Recursive Feature Elimination (RFE), Linear Discriminant Analysis (LDA) and PCA. The ML algorithms tested were SVM (using radial basis kernel), Neural Networks (NN) and Naïve Bayes (NB) carried out on WDBC Dataset. It was observed that SVM-LDA combination and NN-LDA combination obtained the best performance in terms of accuracy (98.82%).

In the comparative study carried out by [18] to determine how to improve classification algorithm on the WDBC dataset, the investigation was conducted on different level of cross validations and percentage of splitting the training dataset. The NB, J48, Random Forest (RF), SMO, Multilayer Perceptron algorithms when trained with 85.5% of the dataset at 10-fold cross validation, the evaluation result showed the NB as 97.28%, J48 as 94.27%, RF as 95.56%, SMO as 96.13% and Multilayer Perceptron as 96.13% accuracy. NB having the highest accuracy was further investigated with 5, 10 and 15 cross validation at 66.6% and 85.5% splitting. The result showed improvement when trained with 85.5% trained set resulting into 99% accuracy. It can be justified that such improvement exists as a result of overfitting of the training set.

In a similar investigation conducted on three distinct datasets: WBDC, WDBC and Coimbra by [23], the study proposed a fuzzy technique in improving the ML algorithms. In order to resolve the limitation of an existing method, where ID3 algorithm was unable to classify the continuous-valued data and increase the classification accuracy of the decision tree, FUZZY-DBD method an automatic fuzzy database was used to design the fuzzy database for fuzzification of data in the FID3 algorithm. It was used to generate a predefined fuzzy database

before the generation of the fuzzy rule base. The fuzzified dataset was applied to fuzzy-ID3 algorithm. The accuracy of fuzzy-ID3 applied to fuzzified dataset was 94.362% when compared with non-fuzzy WDBC dataset applied to ID3 (91.059%), SVM (86.1%), C4.5(92.97%), NB(91.81%), RF(91.66%) and KNN(92.57%).

The study conducted by [28] in order to improve accuracy of ML algorithms, investigated the best features suitable for WDBC dataset. Light Gradient Boosting Model (LGBM), Catboost, and Extreme Gradient Boosting (XGB) were applied as the feature selection approaches tested on Naive Bayes algorithm. The findings revealed best accuracy with LGBM (97%), followed by Catboost (96%) and XGB (96%).

Again, [29] experimented with feature selection techniques of Correlation based Feature Selection(CFS), univariate selection (selectKBest), and Recursive Feature Elimination (RFE). The RF was applied on these feature selection methods and evaluated on WDBC dataset. It was shown that when 5 features were picked, the RF model had the best accuracy with a CFS of 95.32%, selectKBest 94.15% and RFE 94.15%.

As reviewed from the literature on the state-of-the-art approaches employed, the problem of multi linearity in the WDBC dataset still exists because to the best of our knowledge, none of the existing work has investigated the presence of outliers on the WDBC dataset. In addition, the research investigates the extent to which the removal of outliers, when combined with feature selection method can improve the accuracy of other weak algorithms. Hence, this study analyzes improvement of ML algorithms for detecting disease based on outlier detection and feature selection method using Pearson Correlation based feature selection. The goal of this research was to create an optimal model that would fill a knowledge gap. In this study, the characteristics of the Wisconsin Diagnostic Breast Cancer (WDBC) dataset were examined in depth for the presence or absence of outliers. When outliers were discovered, some of the instances were dropped. The filtered dataset was further refined as the classification system's inputs using Correlation Feature Selection (CFS). The Pearson correlation technique was used to find the appropriate continuous features and their associated weight (importance) in order to discover variables that are relevant for prediction. This method assists in the resolution of overfitting and underfitting difficulties in ML. The accuracy was used to evaluate the performance of the unfiltered (conventional technique), filtered (outliers approach), and Outliers Correlation Feature Selection (OCFS) datasets. The findings were assessed and compared using seven classifiers: Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boost (GB), and Adaboost (AB).

2. Materials and Method

The research architecture for predicting the presence of breast cancer disease is shown in Figure 1. The approach includes acquiring a breast cancer disease dataset and preprocessing it to remove missing values and outliers. Also, using the already processed dataset, an algorithm to discover strongly correlated features was applied, and the results were engaged in ML techniques to predict whether a patient had Benign or Malignant tumors. Finally, the outcome was compared using a performance score based on the confusion matrix. Figure 1 depicts the process of recommended techniques for implementing ML Algorithms.

2.1. Data Description

The data for this study is acquired from the UCI repository. This dataset, identified as the WDBC dataset, has 569 cases that are either Benign or Malignant. In these situations, 357 cases (62.74%) are Benign and 212 cases (37.26%) are Malignant. The distribution of the number of Benign and Malignant classes in the dataset is displayed in Figure 2.

The dataset contains 33 attributes: class attribute labels (diagnosis: B= Benign, M= Malignant), id, and 31 real value attributes. These attributes are derived from a digitized image of a biopsy procedure for a breast mass and are used to describe the characteristics of the cell nuclei in the image. The WDBC dataset is a computation of ten real-valued features of cell nucleus: radius, texture, perimeter area smoothness, compactness, concavity, concave points, and symmetry fractal dimension. Each of these qualities was estimated of their respective mean, standard error, and worst values, resulting in a total of 30 attributes. The attributes of the WDBC dataset, as well as their datatypes, are listed in Table 1.
The unique id numbers of the instances and the accompanying class label (diagnosis: M=Malignant, B=Benign) are stored in the first two columns of the dataset, respectively. Columns 3-32 contain 30 real-value features derived from digitized images of cell nuclei which can be used to create a model to predict whether a tumor is benign (i.e., cancer-free) or malignant (i.e., cancerous).

2.2. 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. 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. The phases of pre-processing used in this investigation are as follows:

2.2.1. Missing Values Checking

The dataset contains 569 instances of 33 variables. However, it was discovered that the variable id had no effect on the dataset description or on disease prediction because it merely keeps a serial record of the instances. As a result, the dataset's id feature was removed. Additionally, while conducting additional preprocessing operations on the dataset, it was discovered that the last feature, unnamed:32, had the value null for all occurrences. This might be a mistake in the data collection process, because of this the feature was also removed from the dataset.

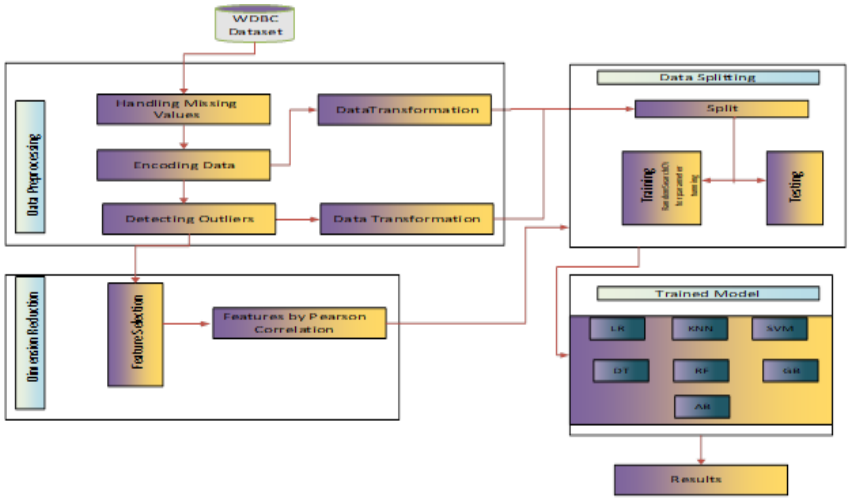

Figure 1. Proposed Architecture

Table 1. Attributes and their Description on WDBC Dataset

S/n	Attributes	Datatypes	S/n	Attributes	Datatypes
1	Id	numeric	18	compactness_se	numeric
2	Diagnosis	nominal	19	concavity_se	numeric
3	radius_mean	numeric	20	concave points_se	numeric
4	texture_mean	numeric	21	symmetry_se	numeric
5	perimeter_mean	numeric	22	fractal_dimension_se	numeric
6	area_mean	numeric	23	radius_worst	numeric
7	smoothness_mean	numeric	24	texture_worst	numeric
8	compactness_mean	numeric	25	perimeter_worst	numeric
9	concavity_mean	numeric	26	area_worst	numeric
10	concave points_mean	numeric	27	smoothness_worst	numeric
11	symmetry_mean	numeric	28	compactness_worst	numeric
12	fractal_dimension_mean	numeric	29	concavity_worst	numeric
13	radius_se	numeric	30	concave points_worst	numeric
14	texture_se	numeric	31	symmetry_worst	numeric
15	perimeter_se	numeric	32	fractal_dimension_worst	numeric
16	area_se	numeric	33	unnamed:32	numeric
17	smoothness_se	numeric			

1

2.2.2. *Encoding data*

The performance of machine models depends on various aspects. One element that influences performance of the models are the methods used to analyze data and feed it to the model. As such, vital step in encoding data is turning data into categorical variables understood by ML models. Encoding data elevates model quality and helps in feature engineering. The class label "diagnosis" was expressed as strings of (B= Benign, M= Malignant). This category characteristic must be converted to restricted numbers. This is done to transform data into a format that ML algorithms can understand. Label encoding was used to encode the diagnostic occurrences in this study, and the result was (M=1, B =0).

2.2.3. *Outliers Checking*

An outlier is a statistic or observation that deviates from a distribution's overall pattern. If few data are significantly different or not in range of main trend then those are termed outliers. There skewness results, affecting the mean and standard deviation of the distribution. As shown in Figure 3, this study detects the existence of outliers in the dataset. As a result, outliers were identified and eliminated from their respective features.

2.2.4. *Data Transformation*

Data must be normalized or standardized before ML algorithms can be applied. The data is standardized to have a mean of 0 (μ) and a standard deviation ($\sum$) of 1. Equation 1 gives the

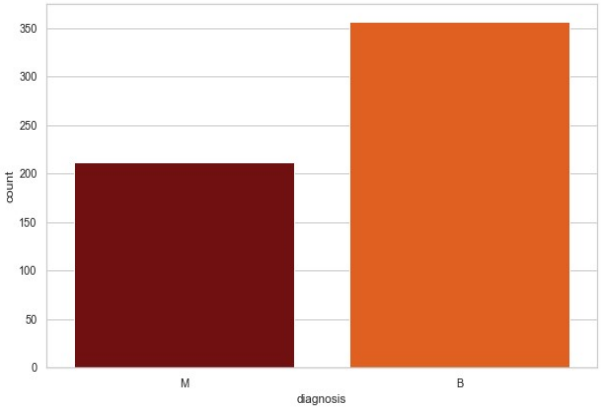

Figure 2. Dataset's Class Level Showing Malignant and Benign in WDBC Dataset

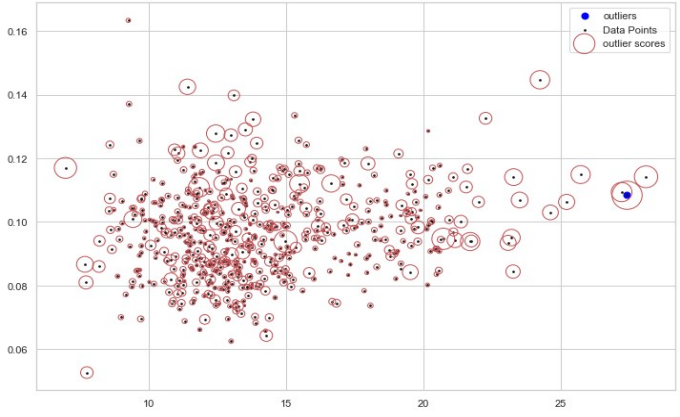

Figure 3. Plots of data points to show presence of outliers

conversion formula:

$$X = \frac{X - \mu}{\sigma} \tag{1}$$

2.3. Dimension Reduction

Dimension reduction is delineated as the mapping of data to a lower dimensional space by removing irrelevant variance in data, resulting in the detection of a subspace in which the data exist [30]. Feature extraction and feature selection are two types of dimensional reduction [30]. The process of identifying and discarding irrelevant, less relevant or duplicated features of dimensions in a dataset is known as feature extraction. Creating strong learning models, feature selection may be used to detect and remove as much unnecessary and redundant information as feasible. As a result, feature selection not only decreases computational and processing costs, but also improves the model created from the chosen data [31], [32].

On healthcare data, the feature selection method has been used in a number of previous studies [27]–[29], [33]. Previous research that are partly relevant to this study and deal with the datasets utilized here; nevertheless, in most cases, the performance of such systems was not as predicted. One of the reasons for some systems' poor performance is their inability to recognize the most important and highly correlated features.

The goal of this research is to devise a method for identifying the best set of features and then investigate which algorithms work best with those features.

Filters, wrappers, and embedding techniques are the three types of algorithms that may be used to select features[34]. The Correlation Feature Selection method is used in this study to identify the best predictive features. Correlation feature selection is a technique that uses the filter approach. The relationship between the independent and dependent variables is determined using a mathematical function. The features are chosen based on the values of their correlation coefficients. The most predictive feature with the class variable is considered to be highly associated, and it is included in the final feature set. Pearson's Correlation: Consider a dataset D having feature set F

$$F = \{x_1, x_2, x_3 \ldots, x_n\} \tag{2}$$

and classes C with values c, where X, C are treated as random variables, Pearson's linear correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^{n} (x_i - x)^2 (c_i - c)}{\sqrt{\left[\sum_{i=1}^{n} (x_i - x)^2\right] \left[\sum_{i=1}^{n} (c_i - c)^2\right]}} \tag{3}$$

Where x_i and c_i the i^{th} value of X and C respectively. Percentage of $(r) = \pm 1$ if X and Y are linearly dependent and zero if they are completely uncorrelated. The values of the correlation coefficients between independent features and the dependent class variable in the WDBC data are shown in Figure 4.
A filter approach was used to identify the most predictive characteristics. The correlation between all features is calculated and displayed in this article. The correlation criterion utilized is 0.6, and features having a correlation of less than 0.6 are removed from the training dataset. While the other qualities that have a higher threshold are chosen. The following Figure 5 depicted based on the highly correlated 10 features with predicted attribute (diagnosis).

2.4. Data Splitting

The goal of dividing the data is to avoid overfitting the model during model testing on the testing dataset. The dataset for this study was split into two parts: training data (80%) and test data (20%).

2.5. Trained Model

Based on the seven classifiers, two intelligent systems were built. Different forms of ML algorithms have previously been the subject of numerous studies. Five of the most prevalent approaches (LR, KNN, SVM, DT, and RF) were chosen, as well as two infrequent techniques (AB and GB). When combined with feature selection approaches, several prior research have demonstrated that the projected accuracy of LR, KNN, SVM, DT[25], [26] and RF[27] algorithms was fairly high. Furthermore, to the best of our knowledge, no research in this field have shown that AB and GB can perform very well with a high degree of accuracy. These methods were investigated in this study using hyperparameter tuning to improve the proposed model's efficiency. The following are the algorithms that will be discussed:

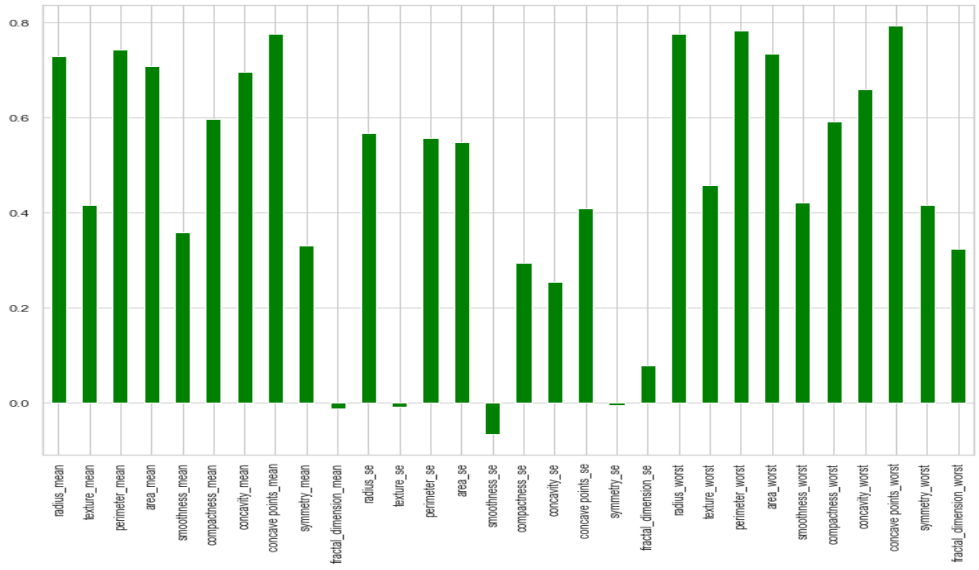

Figure 4. Correlation Coefficient Values between Independent Features and Dependent Class Variable

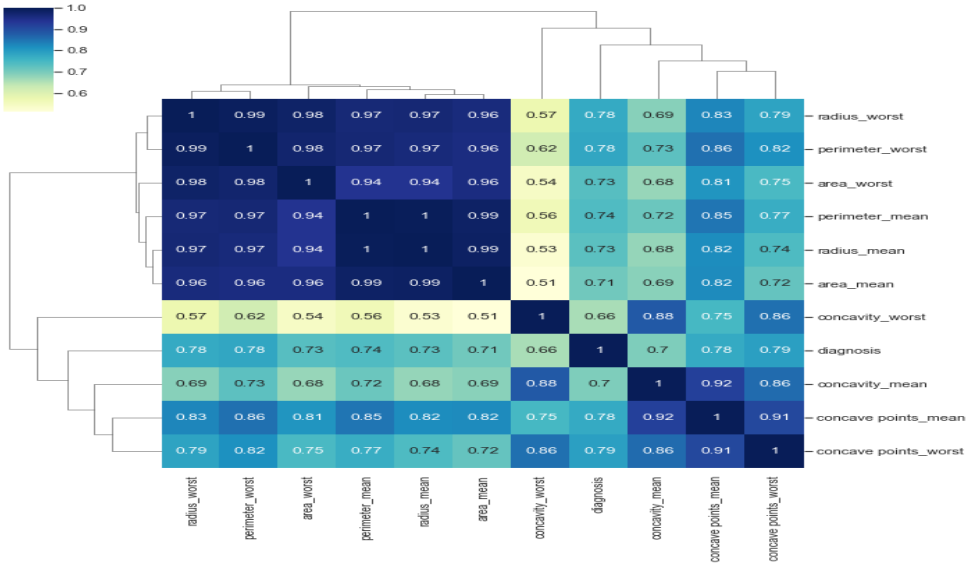

Figure 5. Plot of Highly Correlated Features

2.5.1. *Linear regression*

LR is a simple supervised learning method for projecting the relationship between explanatory variables and dependent variables by fitting a linear equation to experience data. LR can be mathematically modelled as shown in equation 4 [35]:

$$y = \beta_0 + \beta_1 x_1 + e \tag{4}$$

Where y is the response variable, β_0 and β_1 are the model coefficients, and e is the model coefficients error. The intercept and slope are represented by these unknown constant values, which are learnt during the training phase.
Equation 5 is used to predict after the model has been fitted in the training phase:

$$y = \beta_0 + \beta_1 x \tag{5}$$

Where y is the predicted value based on x, and the error is calculated as shown in equation 6 as follows:

$$e_i = y_i + \hat{y}_1 \tag{6}$$

In this study, the solver and the maximum number of iterations were the two LR parameters used. The solver is the algorithm that is utilized to solve the optimization issue. The algorithm options include "newton-cg", "sag", "saga", "liblinear" and "lbfgs". The number of iterations required for the solvers to converge is specified between 100 and 10000.

2.5.2. *K-Nearest Neighbor*

KNN is a supervised classifier that learns from data samples that have been labeled. It is a lazy method since it makes

no generalizations about the sample data points and all calculations are put on hold until the classification is completed. KNN works by determining the K closest neighbors given n training vectors. KNN converts the training data set into a multi-dimensional feature space and divides it into various areas based on the training dataset's classifications. The concept of number of neighbors is fundamental to this method. The number of neighbors specifies how many neighbors should be checked when an item is classified. The parameter range in this study was randomly searched between 1 and 30, with the case being assigned to the most frequent class among its K nearest neighbors, as determined by a distance role.

2.5.3. Support Vector Machine

SVM are a classification approach that involves projecting input data points into n-dimensional vector space and determining the optimal hyper-plane that maximizes the difference between the two classes. The choice of parameters such as kernel, C, and gamma has a significant impact on SVM performance. Kernels are a function that converts a low-dimensional space into a high-dimensional one, making categorization simple. The non-linearity is controlled by the kernel. "rbf", "poly" or "sigmoid" are all possible kernel coefficients. The kernel coefficients were adjusted in this investigation.

2.5.4. Decision Tree Algorithm

DT is a strong predictive learning tool used to solve classification and regression problems. It uses a tree-based top-down progression method. It employs a tiered splitting method to divide data into two or more groups at each layer, ensuring that data in each group is comparable. Every inner node in a DT's tie ups to a test attribute, every branch to a test result, and each leaf node to a different class. Before applying 'splitting', the tree develops from the root node by selecting a 'best feature' or 'best attribute' from the set of accessible attributes using entropy and information gain measures. The most useful information is provided by the 'best attribute'. Information Gain is the pace at which the entropy of attributes increases or decreases, and entropy reflects how homogenous the dataset is. The key parameters that are optimized are the maximum depth of the tree, the number of features to check when looking for the optimum split, the lowest number of samples required to divide an internal node, and the criterion used for splitting.

2.5.5. Random Forest

RF is an ensemble of several separate randomized decision trees that work together. Bootstrap sampling of the data is used to create the trees. Based on the set of predictor values entered, each individual tree in the random forest casts a unit vote, and the class with the most votes becomes the model's prediction for categorizing an input vector[36]. The recurrent division of a binary tree into comparable nodes is used to create RF. By inheritance, the parent node impacts the similarity of the child node.

2.5.6. Gradient Boosting

GB is a strategy for enhancing ideas that have poor learning or predictability. The goal of GB is to combine numerous concepts with a weak predictive component and a clever algorithm to create a decision tree with a considerably greater connectivity. If there aren't many ideas in common across the data components, this notion is especially effective with large datasets. Search engine rankings are one of the most popular uses of this technology. Search engine rankings must filter a large number of possible queries, some of which may or may not be related, into a limited number of rankable words.

2.5.7. AdaBoost

AB is a boosting method that uses weight modification to solve classification problems without requiring any prior information of the learner's learning. The goal of AdaBoost is to enhance classification performance by combining different weak learners or classifiers. A basic collection of training examples is used to train each weak learner. Each sample has a weight, which is adjusted iteratively across all samples. The robustness of the weak learner is represented by this weight. The AdaBoost algorithm consists of the following main steps: (i) sampling, which involves selecting some samples from the training set while iterating. (ii) The sample data is used to train different classifiers, and the error rates for each classifier are computed. (iii) The last stage is the combination of all trained models.

3. Performance evaluation metric

The metric accuracy was calculated using the 2 *X* 2 confusion matrix to test the validity of the prediction models, as shown in Table 2. The accuracy determines the proportion or possibility of a total number of correct predictions [37]. As seen in equations 7, the following formula is used to quantitatively represent this measurement. Where TP, TN, FP, and FN stand for True Positive (number of positive data correctly labeled by the classifier), True Negative (number of negative data correctly labeled by the classifier), False Positive (number of negative data incorrectly labeled as positive), and False Negative (number of positive data incorrectly labeled as negative) respectively.

Table 2. Illustration of Confusion Matrix Table

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

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

Table 3. Comparison Table between the Accuracy of the Proposed Models and Existing Techniques

Authors	Dataset	Techniques	Accuracy (%)
[25]	WDBC	DT	92.9
		PCA-SVM	86.7
		PCA-KNN	82.3
		EM-PCA-CART-Fuzzy Rule-Based	93.2
[27]	WDBC	RF	98.7
[26]	WDBC	SVM	96.47
		CFS-SVM	96.47
		RFE-SVM	96.47
		LDA-SVM	98.82
[23]	WDBC	SVM	61.96
		RF	89.37
		KNN	92.77
		Fuzzy-ID3	94.53
[29]	WDBC	CFS-RF	95.32
		UFS-RF	94.15
		RFE-RF	94.15
Proposed work with Outliers	WDBC	LR	99.1
		KNN	96.5
		SVM	95.6
		DT	96.5
		RF	99.1
		GB	98.3
		AB	99.1
Proposed work with Pearson Correlation Feature Selection	WDBC	OCFS-LR	96.5
		OCFS-KNN	96.5
		OCFS-SVM	95.6
		OCFS-DT	94.7
		OCFS-RF	100
		OCFS-GB	99.1
		OCFS-AB	98.3

4. Result and Discussion

The two proposed approaches were subjected to various ML techniques. In order to create the performance statistic, a 2 x 2 confusion matrix was created, which allowed all of the algorithms to be compared. The suggested models were evaluated using the performance indicator "accuracy".

4.1. Comparison between Different Machine Learning Algorithms Based on Accuracy

The most essential metrics for evaluating ML algorithms is accuracy. Seven classifiers were applied to the WDBC features and processed by resolving the problem of missing values and scaling their instances, as previously indicated, this is called Conventional approach. Second, the outliers instances were detected in five input features and were dropped from their respective features this is termed Outliers approach. Finally, the remaining instances of the outliers technique were subjected to Pearson Correlation Feature Selection, which resulted in the selection of 10 features, which is known as the OCFS approach. The accuracy of several types of classifiers was carried out on each of these three phases and the result is plotted in Figure 6.

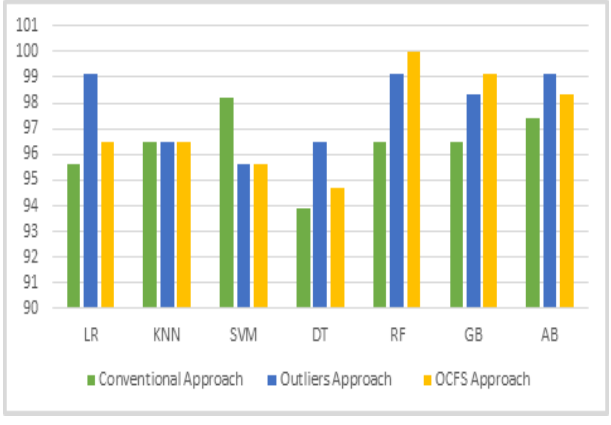

Figure 6. Comparison between Different Machine Learning Algorithms Based on Accuracy

As demonstrated in Figure 6, the conventional approach performs the poorest of all the strategies owing to the presence of outliers, which increases data variability and reduces statistical power. KNN, RF, and GB all have similar accuracy

(96.5%); however, they perform poorly when compared in other approaches. Furthermore, the conventional approach has an exceptional high accuracy of 98.2% in the SVM classifier and the least accuracy of 93.9% in the DT. The outlier approach produces considerably higher results for all its classifiers, with three classifiers having maximum accuracy of 99.1% for LR, RF, and AB, respectively. Lastly, when the approach of outlier is used with the outcomes of Pearson Correlation technique (10 features), the accuracies for all predictive classifiers improve the most compared to their foils in other approaches. This is feasible because feature selection allows for the removal of noise from data while also picking the most valuable features; this strategy, when paired with the outlier approach, yields the greatest results. Figure 6 shows that OCFS approach RF has the best accuracy (100%); this is because RF added additional randomness to the model when developing the trees. Instead of working with the suggested features, it also performs additional searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.

4.2. Comparison Table between the Accuracy of the Proposed Models and Existing Techniques

For this research, WDBC datasets were used. The architecture of our suggested system is depicted in Figure 1. Table 3 shows the outliers and OCFS results, as well as previous work on the WDBC dataset. As a result, distinct results based on the WDBC dataset have been published in the literature. It was reported that after changing the number of selected features by implementing selection algorithms like Principal Component Analysis(PCA), Correlation Feature Selection(CFS), Recursive Feature Selection(RFE), Linear Discriminant Analysis(LDA), Univariate Feature Selection(UFS), Fuzzy Rule Based and Expectation Maximization (EM)-PCA-CART-Fuzzy Rule-Based, there were significant improvements in the accuracy. So, this work compared the accuracy of our two approaches.

When it came to removing outliers, the LR model had the greatest accuracy (99.1%), while the SVM model had the lowest accuracy score (95.6%). When the OCFS is used, it results in some significant modifications. The RF model achieved the best accuracy (100%), whereas the SVM model fared the poorest. Table 3 shows a comparison of our findings to current models and datasets. The table's "Techniques" column contains information on the methodologies that were utilized in the previous study, as well as our own methodology and the findings that were published. The table depicts the overall performance of the algorithms in our study in comparison to other comparable works. The greatest outcome of previous RF findings is 94.15% [29], and our OCFS performance has improved to 100%.

5. Conclusion

The research primarily focuses on improving ML models in order to improve accuracy in forecasting breast cancer disease outcomes. The results show that outlier detection and OCFS methods, in combination with various classification algorithms, might provide useful tools for inference in this area. More study in this area is needed to improve the classification systems' performance on diverse feature selection approaches so that they can predict on more variables.

Acknowledgments

The authors will like to appreciate the handling editor and the anonymous referees for their contributions to the success of this research.

References

[1] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas & F. Mokarian, "A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning," Computational and Structural Biotechnology Journal **15** (2017) 75.

[2] S. Amin, H. S. Ewunonu, E. Oguntebi & I. Liman, "Breast cancer mortality in a resource-poor country: a 10-year experience in a tertiary institution," Sahel Medical Journal **20** (2017) 9.

[3] M.W. Huang, C.W. Chen, W.C. Lin, S.W. Ke & C.F. Tsai, "SVM and SVM ensembles in breast cancer prediction," PLoS ONE **12** (2017) 161501.

[4] CDC, "What is breast cancer?" (2021).

[5] R. J. Oskouei, N. M. Kor & S. A. Maleki, "Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges," American Journal of Cancer Research **7** (2017) 610.

[6] L. A. Aaltonen, R. Salovaara, P. Kristo, F. Canzian, A. Hemminki, P. Peltomäki, R. B. Chadwick, H. Kääriäinen, M. Eskelinen, H. Järvinen, J. P. Mecklin, & A. De la Chapelle, "Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease," New England Journal of Medicine **338** (1998) 1481.

[7] A. Khamparia, S. Bharati, P. Podder, D. Gupta, A. Khanna, T. K. Phung & D. N. H. Thanh, "Diagnosis of breast cancer based on modern mammography using hybrid transfer learning," Multidimensional Systems and Signal Processing **32** (2021) 747.

[8] H. Kurihara, C. Shimizu, Y. Miyakita, M. Yoshida, A. Hamada, Y. Kanayama, K. Yonemori, J. Hashimoto, H. Tani, M. Kodaira, M. Yunokawa, H. Yamamoto, Y. Watanabe, Y. Fujiwara & K. Tamura, "Molecular imaging using PET for breast cancer," The Japanese Breast Cancer Society **23** (2016) 24.

[9] T. Nagashima, M. Suzuki, H. Yagata, H. Hashimoto, T. Shishikura, N. Imanaka, T. Ueda & M. Miyazaki, "Dynamic-enhanced MRI predicts metastatic potential of invasive ductal breast cancer," Breast Cancer **9** (2002) 226.

[10] C. S. Park, S. H. Kim, N. Y. Jung, J. J. Choi, B. J. Kang & H. S. Jung, "Interobserver variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions," Breast Cancer **22** (2015) 153.

[11] S. I. Ayon, M. Islam & M. R. Hossain, "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques," IETE Journal of Research (2020) 1.

[12] M. M. Islam, H. Iqbal, R. Haque & K. Hasan, "Prediction of breast cancer using support vector machine and k-nearest neighbors," IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (2017) 226.

[13] L. J. Muhammad, M. M. Islam, S. S. Usman & S. I. Ayon, "Predictive data mining models for novel coronavirus (covid-19) infected patients' recovery," SN Computer Science **1** (2020) 206.

[14] A. Yusuf & O. Akande, "Hyper-parameter optimization and evaluation on selected machine learning algorithm using hepatitis dataset," FUDMA Journal of Sciences **5** (2021) 447.

[15] S. I. Ayon & M. Islam, "Diabetes prediction: a deep learning approach," International Journal of Information Engineering and Electronic Business **11** (2019) 2.

[16] Z. Islam, M. Islam & A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (covid-19) using x-ray images," Informatics in Medicine Unlocked **20** (2020) 100412.

[17] K. Hasan, M. Islam & M. M. A. Hashem, "Mathematical model development to detect breast cancer using multigene genetic programming," International Conference on Informatics, Electronics and Vision (2016) 574.

[18] M. T. Ahmed, M. N. Imtiaz & A. Karmakar, "Analysis of wisconsin breast cancer original dataset using data mining and machine learning algorithms for breast cancer prediction," Journal of Science Technology and Environment Informatics **9** (2020) 665.

[19] M. M. Islam, Md. R. Haque, H. Iqbal, Md. M. Hasan, M. Hasan & M. N. Kabir, "Breast cancer prediction: A comparative study using machine learning techniques," SN Computer Science **1** (2020) 290.

[20] N. Khuriwal & N. Mishra, "Breast cancer diagnosis using deep learning algorithm," International Conference on Advances in Computing, Communication Control and Networking (2018) 98.

[21] C. Shah & A. G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction," Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (2013) 1.

[22] F. A. Muhammet, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," Healthcare **8** (2020) 111.

[23] N. F. Idris & M. A. Ismail, "Breast cancer disease classification using Fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition," PeerJ Computer Science **7** (2021) 427.

[24] R. Harikumar & C. Sannasi, "Effective classification framework for breast tumors using optimized multi-kernel SVM with controlled skewness," International Journal of Aquatic Science **12** (2021) 1604.

[25] M. Nilashi, O. Ibrahim, H. Ahmadi & L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," Telematics and Informatics **34** (2017) 133.

[26] D. A. Omondiagbe, S. Veeramani & A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," IOP Conference Series: Materials Science and Engineering **495** (2019) 012033.

[27] A. Saygılı, "Classification and diagnostic prediction of breast cancers via different classifiers," International Scientific and Vocational Studies Journal **2** (2018) 56.

[28] A. Derangula, S. Edara & P. K. Karri, "Feature selection of breast cancer data using gradient boosting techniques of machine learning," Clinical Medicine **7** (2020) 17.

[29] S. Raj, S. Singh, A. Kumar, S. Sarkar & C. Pradhan, "Feature selection and random forest classification for breast cancer disease," Data Analytics in Bioinformatics (2021) 191.

[30] T. H. Cheng, C. P. Wei & V. S. Tseng, "Feature selection for medical data mining: comparisons of expert judgment and automatic approaches," 19th IEEE Symposium on Computer-Based Medical Systems (2006) 165.

[31] S. N. Ghazavi & T. W. Liao, "Medical data mining by fuzzy modeling with selected features," Artificial Intelligence in Medicine **43** (2008) 195.

[32] S. M. Vieira, J. M. C. Sousa & U. Kaymak, "Fuzzy criteria for feature selection," Fuzzy Sets and Systems 189**189** (2012)1.

[33] S. B. Sakri, N. B. Abdul Rashid & Z. Muhammad Zain, "Particle swarm optimization feature selection for breast cancer recurrence prediction," IEEE Access **6** (2018) 29637.

[34] E. E. Bron, M. Smits, W. J. Niessen & S. Klein, "Feature selection based on the SVM weight vector for classification of dementia," IEEE Journal of Biomedical and Health Informatics **19** (2015) 1617.

[35] M. Kumari, V. Singh & P. Ahlawat, "Automated decision support system for breast cancer prediction," International Journal on Emerging Technologies **11** (2020) 193.

[36] L. Breiman, "Random forests: random features," Technical Report 567, Statistics Department, University of California, Berkeley (1999) 29.

[37] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," Remote Sensing of Environment **62** (1997) 77.