*I.J. Intelligent Systems and Applications*, **2021, 4, 24-37**
Published Online August 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2021.04.03

# A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer

**Avijit Kumar Chaudhuri**
Research Scholar, Department of Computer Application, SEACOM SKILLS UNIVERSITY, Kendradangal, Bolpur, Dist:Birbhum, PIN - 731 236, West Bengal
E-mail: c.avijit@gmail.com

**Dilip K. Banerjee**
Professor, Department of Computer Application, SEACOM SKILLS UNIVERSITY, Kendradangal, Bolpur, Dist:Birbhum, PIN - 731 236, West Bengal
E-mail: dkbanrg@gmail.com

**Anirban Das**
University of Engineering & Management, Kolkata
E-mail: anirban-das@live.com

Received: 14 May 2021; Accepted: 24 June 2021; Published: 08 August 2021

**Abstract:** World Health Organisation declared breast cancer (BC) as the most frequent suffering among women and accounted for 15 percent of all cancer deaths. Its accurate prediction is of utmost significance as it not only prevents deaths but also stops mistreatments. The conventional way of diagnosis includes the estimation of the tumor size as a sign of plausible cancer. Machine learning (ML) techniques have shown the effectiveness of predicting disease. However, the ML methods have been method centric rather than being dataset centric. In this paper, the authors introduce a dataset centric approach(DCA) deploying a genetic algorithm (GA) method to identify the features and a learning ensemble classifier algorithm to predict using the right features. Adaboost is such an approach that trains the model assigning weights to individual records rather than experimenting on the splitting of datasets alone and perform hyper-parameter optimization. The authors simulate the results by varying base classifiers i.e, using logistic regression (LR), decision tree (DT), support vector machine (SVM), naive bayes (NB), random forest (RF), and 10-fold cross-validations with a different split of the dataset as training and testing. The proposed DCA model with RF and 10-fold cross-validations demonstrated its potential with almost 100% performance in the classification results that no research could suggest so far. The DCA satisfies the underlying principles of data mining: the principle of parsimony, the principle of inclusion, the principle of discrimination, and the principle of optimality. This DCA is a democratic and unbiased ensemble approach as it allows all features and methods in the start to compete, but filters out the most reliable chain (of steps and combinations) that give the highest accuracy. With fewer characteristics and splits of 50-50, 66-34, and 10 fold cross-validations, the Stacked model achieves 97 % accuracy. These values and the reduction of features improve upon prior research works.

Further, the proposed classifier is compared with some state-of-the-art machine-learning classifiers, namely random forest, naive Bayes, support-vector machine with radial basis function kernel, and decision tree. For testing the classifiers, different performance metrics have been employed – accuracy, detection rate, sensitivity, specificity, receiver operating characteristic, area under the curve, and some statistical tests such as the Wilcoxon signed-rank test and kappa statistics – to check the strength of the proposed DCA classifier. Various splits of training and testing data – namely, 50–50%, 66–34%, 80–20% and 10-fold cross-validation – have been incorporated in this research to test the credibility of the classification models in handling the unbalanced data. Finally, the proposed DCA model demonstrated its potential with almost 100% performance in the classification results. The output results have also been compared with other research on the same dataset where the proposed classifiers were found to be best across all the performance dimensions.

**Index Terms:** Breast Cancer, Machine Learning, Feature Selection, Dataset Centric Approach, Ensemble Classifier.

## 1. Introduction

World Health Organisation declared breast cancer (BC) as the most frequent suffering among women and accounted for 15 percent of all cancer deaths. Its accurate prediction is of utmost significance as it not only prevents deaths but also stops mistreatments. The wrong diagnosis leads to exposure of radiation and drugs to benign patients. Hence, researchers need to find how to maximize prediction accuracy minimizing false-positive cases?

Researchers have widely used machine learning (ML) methods to predict diseases[1,2,3,4]. However, the attempts have been skewed towards time to compute, and accuracy of prediction. Studies show a high accuracy of prediction of breast cancer using ML methods. However, the researchers have not completed the analysis by validating the results. Table 1 gives an account of the incompleteness of the studies done so far. So the question remains – can the accuracy levels be further improved meeting the desired performance levels?

The database of diseases has many data items, starting from demographic to diagnostic test results to habits and similar. Researchers [5,6,7,8,9,10] substantiated that prediction improves with the choice of right features. As information and storage technology are making exponential progress, data sets are now omnipresent in pattern analysis, data processing, and machine learning (ML) systems, with a vast range of variables or features. Therefore there is a need for choosing a subset of many features that best suits the task. ML and training require a wide range of features that might take time to explore. Thus, FS marks the beginning of an ML analysis, followed by prediction using classifiers. There are several methods to select features, and these methods have shown varied outcomes when applied to different datasets. These results do not show the best method to use. Hence, researchers either apply one or some of them or all to carry out feature selections. These differences in results hint at the nature of the dataset. Thus, the question is how to determine a way to make dataset centric feature selection (FS) rather than the method based FS?

The classifiers evolved, and researchers eliminated the shortcomings of incomplete, incorrect, and non-standardization data, and binomial versus multi-domain attributes. DT, a supervised classifier, takes care of these issues as it predicts the dependent variable[11]. Outliers do not affect the results of the decision trees as the classification is based on the proportion of samples with split ranges and not absolute values. This approach needs no linearity in the relationship between the dependent attributes and the independent attribute. However, this method does not yield the desired accuracy for a large database. Bayes theorem, on comparatively small and simple datasets, is regarded as one of the most common classification techniques due to its simplicity, robustness, and prediction accuracy. The NB classifier's performance against large datasets and datasets with complex attribute dependencies is poor [12] . SVM model takes care of this issue [13,14] but does not show good results when large datasets contain noise. It needs combining with other ML techniques [15]. The concept of a combination of methods introduced the ensemble classifiers. RF gained importance as an ensemble classifier and demonstrated consistent results for different datasets [16, 17]. In the process of evaluating these classifiers, authors [18] observe that DT produced equivalent or better results than RF. This situation led to the use of all popular classifiers and compared the results. In the process, some ML techniques such as LR continued to be applied over broad types of variables. This method is easy to interpret as it gives the linear combination variables that predict the occurrence (or otherwise) of the disease. However, LR's problem is its tendency to generate over fitted models[19]. So the point to ponder is - which is the most preferred method?

In this paper, the authors introduce a dataset centric approach deploying a genetic algorithm (GA) method to identify the features and a learning ensemble classifier algorithm to predict using the right features. Adaboost is such an approach that trains the model assigning weights to individual records rather than experimenting on the splitting of datasets alone and perform hyper-parameter optimization. The authors compare the outcome of the Adaboost and hyper-parameter optimization-based GA approach to substantiate the same. The authors simulate the results by varying base classifier i.e, using LR, DT, SVM, NB and RF and 10-fold cross validations with different split of datasets as training and testing. The authors checked for noise and outliers to minimize any misfitting. This method is termed as a dataset-centric approach (DCA). There are some references [20,21] of GA used for FS and Adaboost for classification; however, authors did not find a comparison of the performance of a GA based FS with Adaboost classification with other approaches.

The DCA satisfies the underlying principles of data mining [22,23]. These principles are – i. the principle of parsimony – as it makes feature selection (FS) select the relevant features; ii. the principle of inclusion – as it allows most widely used ML methods; iii. the principle of discrimination – as it differentiates between weak and robust classifiers by assigning weights using the Adaboost algorithm and, iv. the principle of optimality – as it performs hyperparameter optimization to the trade-off between prediction accuracy and computation time. This DCA is a democratic and unbiased ensemble approach as it allows all features and methods in the start to compete, but filters out the most reliable chain (of steps and combinations) that give the highest accuracy.

This paper attempts to answer the following research questions: Which is the best Data Mining Technique(DMT) for the prediction of diseases such as breast cancer? Which DMT framework can help meet the three criteria of consistency, sensitivity, and specificity? Author considers the most popular approaches and explores their ensemble to arrive at the highest levels of consistency, sensitivity, and specificity. Previous authors have emphasized only on the reduction of variables to improve prediction. However, this approach leads to a loss of information. Thus, in this paper, author establish a framework that proposes the application of data-mining approaches, the measurement of consistency using kappa and Wilcoxon rank-sum statistics, and suggested improvement of the specificity and sensitivity parameters using an ensemble learning approach.

In this research, author proposes a DCA approach to construct an ensemble of classifiers. This classifier utilizes n number of RF classifiers (instead of the DT as default classifier) with the adoptive boosting technique by changing its defined base classifier to RF and changing its weight iteratively. This proposed classifier will integrate the n individual decisions and generate a robust classifier with around 100% accuracy. Thus, the framework in this paper contributes to the well-being of humankind by enabling higher levels of prediction of diseases.

Table 1.Wisconsin (Diagnostic) Data Set for Breast Cancer Performance Comparison

| Year | Method | Classification Accuracy (%) | Sensitivity/ Specificity | Kappa | ROC/AUC | Wilcoxon |
|---|---|---|---|---|---|---|
| Sridevi &Murugan, 2014[24] | Multilayer perceptron(MLP) | 100 | × | × | × | × |
| Alickovic & Subasi, 2017[25] | Rotation Forest model classifies using GA | 99.48 | × | × | 0.993 | × |
| Hamsagayathri & Sampath, 2017[26] | Priority based decision tree classifier | 93.63% | 0.936/ 0.982 | 0.925 | 0.929 | × |
| Abdar et al., 2018[[27] | SV-Na¨ıve Bayes-3-MetaClassifiers | 98.07 | 0.981 | × | 0.976 | × |
| Zheng et al., 2014[28] | K-SVM | 97.38 | × | × | ✓ | × |
| Sewak et al., 2007[29] | Ensemble SVM | 99.29 | 1 / 0.981 | × | × | × |
| Jin et al., 2012[30] | Functional Trees (FT) | 97.33 | 0.946 | 98.85 | × | × |
| Obaid, et al., 2018[31] | Quadratic Kernel Based SVM | 98.1 | × | × | 0.984305 for benign tumor and 0.988352 for malignant tumor | × |
| Kumari & Arumugam, 2015[32] | Hybrid Krill Herd | 87.89 | 0.975/ 0.718 | × | × | × |

## 2. Relevant Literature

Data mining or machine-learning techniques help to generate predictive and diagnostic policies automatically. These methods can be beneficial in predicting risk at an early stage of breast cancer. Data-mining analysis has shown encouraging results, but the outcomes are not consistent with techniques and datasets. Researchers have tried and tested different classifiers for predicting the disease, namely: the SVM, DT, RF, LR, NB, GA, neural networks (NN) and K-nearest neighbor (KNN).

Christobel & Sivaprakasam(2011)[33] used the WDBC dataset for BC diagnosis. They use the classifiers DT, kNN, SVM & NB and compare the precision of their classification. The average accuracy of 96.99% of SVM was the highest accuracy among them. The suggested solution restricts the use of a single classifier and a single data set.

Lavanya and Rani (2011)[34] illustrated the DT classification results without using any feature selection techniques on the BC dataset. The study obtained a detection precision of 69.23 percent with no variety of features and 94.84 percent precision for Wisconsin breast cancer (WBC) and 92.97 percent for Wisconsin Diagnosis Breast Cancer(WDBC) dataset. The classification techniques boost the accuracy to 70.63%, 96.99%, and 92.09 %, respectively.

Keles et al. (2011)[35] developed a method and achieved 97% accuracy by using fuzzy laws as a useful diagnostic tool.

Chen et al. (2011)[36] suggested a classifier based on the rough-set vector to diagnose BC. The algorithm has also identified five features that can help doctors classify BC and achieve high detection accuracy.

Salama et al. (2012)[37] used NB, sequential-minimal optimization (SMO), DT(J48), Multi-layer Perception (MLP), and instance-based kNN classifiers to determine the precision of the classification of various BC datasets. In addition to the feature selection process, the introduction of the MLP and J48 classification systems has, among other factors, increased precision of accuracy using WDBC data sets. They suggested SMO, as most appropriate for the WDBC dataset.

In an ensemble classifier, Lavanya and Rani (2012)[38] presented data on BC. The hybrid approach proposed is based on the CART and bagging schemes. Pre-processing was used to improve efficiency and the collection of features and showed improved accuracy of the classification.

Kim et al. (2012)[39] used the SVM technique in their paper on a BC dataset consisting of 679 records that included clinical, pathological, and epidemiological data types. The precision achieved with the role of local tumor invasion was 99 %.

Katsis et al. (2013)[7] used a Correlation Feature Selection (CFS) technique for assessing the various derived characteristics in their suggested approach, and an Artificial Immune Recognition System (AIRS) classifier to assist with BC diagnosis. For the 4726 cases, data were obtained from 53 subjects to test the technique. The specific subjects, when evaluated on all modalities used, showed lesions that were not uniquely predictive of benignity or malignancy. The approach was validated using the findings of the biopsy for all 53 subjects. SVM methodology gave a 76.33 % accuracy considering the maximum number of variables and a 75.89% accuracy considering the subset of CFS selected variables.

Kumar et al. (2013)[40] used a dataset containing 699 patient studies in their research paper, and the training constitutes 499 records and 200 for testing. In that scenario, 241 or 34.5 % had BCs, while the remaining 458 or 65.5 %

were non-cancerous. The authors used a 10-fold crossover validation to verify the predictive results of the six conventional data mining techniques. The accuracy of 94.5% was achieved here by applying the NB and SVM algorithm.

Kharya et al. (2014)[41] established a probabilistic method for forecasting BC utilizing Naive Bayes Classifiers. This data set includes around 65.5 % of stable cases and 34.5 % of malignant cases. The approach showed a precision of 93 %.

Sivakami & Saraswathi(2015)[42] used the DT-SVM Hybrid Model for BC prediction on the WBCD dataset.The DT-SVM was 91% correct at an error rate of 2.58%. Many grouping algorithms have also been used, such as Instance-based Learning (IBL), Sequential Minimal Optimization (SMO), and NB. The precision of the IBL was 85.23%, with an error rate of 12.63%. The accuracy of the SMO is 72.56%, and the error rate is 5.96%. The NB was 89.48 % correct, with a 9.89 % error rate.

Ang et al. (2016)[43] tried to enhance Naive Bayes through the incorporation of links or connections to features like Tree Augmented Naive Bayes (TAN) into their research paper. As seen in the analysis, the precision of the General Bayesian Network (GBN), which did not impose any theoretical restrictions and better represented the data set, was applied to a hill-climbing system. The authors used seven minimal datasets to assess GBN's output against NB and TAN without missing values. These datasets were obtained from the UCI ML Repository and then classified with NB, GBN, and TAN as 10-fold cross-validation using the WEKA ML tool 286 instances of 10 attributes each. The NB model achieved a 71.68 % accuracy, followed by 69.58 % for TAN and 74.47 % for GBN.

Chaudhuri et al. (2018)[44] analyzed the dataset of BC to find a recurrence of the disease, using decision tree and discriminant analysis.

The varied outcome from the different approaches led to different association rules. Thus, the existing studies lacked consistency. There is very little evidence of a framework that ensures the comparison of outcomes; identification of the right factors; measures to ensure consistency, specificity, and sensitivity; and, finally, prediction of the condition at close to 100% accuracy levels. In the present research, author proposed a framework to ensure all criteria are met, and introduce a new ensemble-based classifier – namely, the DCA – to produce an accurate classification. The paper draws on the new approach that shifts from the use of DT as the default classifier to the use of the RF classifier as a learner for training purposes.

Genetic Algorithm(GA) follows the iterative learning theory, which was first introduced by Holland. This methodology operates on a principle close to that of natural-system genetic simulations(Figure 1 represents architecture of the GA technique). This algorithm initially used to identify individuals with a permanent population using a space snapshot. The role of exercise is designed for individual evaluation. Any operations are carried out to develop new generations [45].

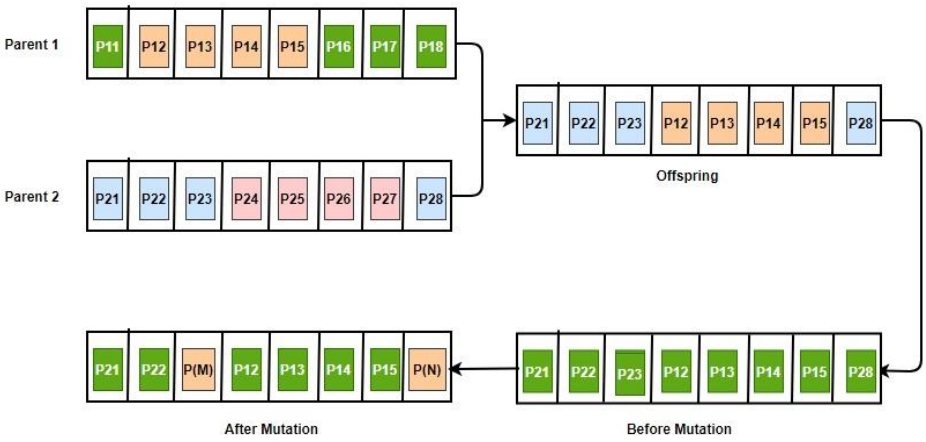


Fig.1. Architecture explaining the GA technique

Several studies have examined the use of GA in time-consuming tasks such as selecting features, which aim to select a specific small subset of features from the entire set of features[46]. AdaBoost[47] produces a variety of sequential base classifiers as one of the typical learning algorithms using ensemble technique by changing the weights over training instances. In AdaBoost, instances misclassified by the present classification techniques are given a greater weight and vice versa. This allows classifiers to concentrate on instances which are not adequately categorized in previous classifiers and to create new base classifiers. As long as the base classifier is slightly more reliable than random inference, the upper limit of training error in AdaBoost typically decreases monotonically. The distinguishing attribute of the final determination shall always be weighted by each base classifier when combined.

## 3. Proposed Classifier: Dataset-Centric-Approach (DCA)

The proposed approach (called DCA) refers to GA for selecting a collection of features to make Adaboost choose from rather than the list of all features available. Besides, Freund and Schapire (1996 )[48] proposed the original Adaboost algorithm as follows: AdaBoost uses small decision trees. Once a tree is created, the AdaBoost algorithm uses the tree's output for each training example. It compares the importance of one tree to the next tree. It assigns more weight to training data that are difficult to predict and less weight to instances that are easy to predict. The procedure is repeated until the set of selected features reaches the preset false alarm, and the rate of hit is set for the classification. The authors used the python programming language here to integrate the use of GA. Figure 1 displays a block diagram illustrating the operations of the proposed DCA technique.

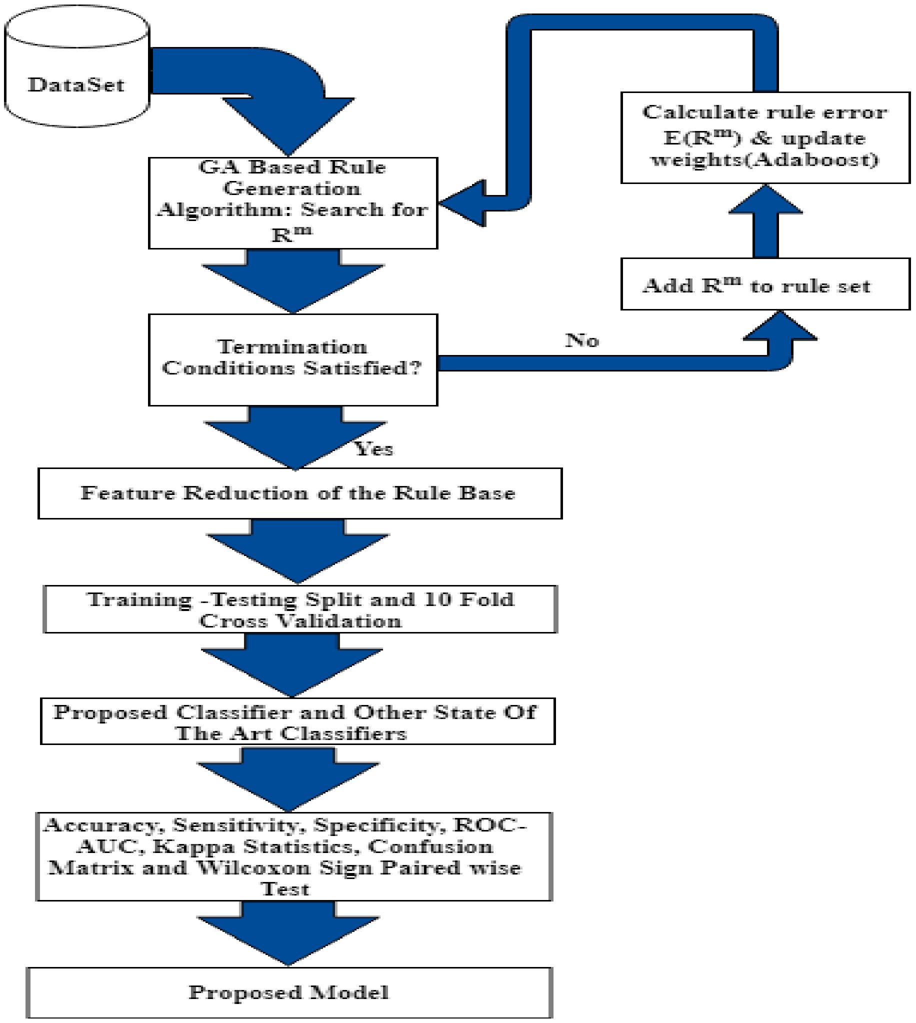

DataSet
GA Based Rule Generation Algorithm: Search for $R^m$
Calculate rule error $E(R^m)$ & update weights(Adaboost)
Add $R^m$ to rule set
Termination Conditions Satisfied?
No
Yes
Feature Reduction of the Rule Base
Training -Testing Split and 10 Fold Cross Validation
Proposed Classifier and Other State Of The Art Classifiers
Accuracy, Sensitivity, Specificity, ROC-AUC, Kappa Statistics, Confusion Matrix and Wilcoxon Sign Paired wise Test
Proposed Model

Fig.2. Diagram of the architecture explaining the proposed DCA technique

Figure 1 shows that after generating the best feature subset using the method above, the proposed model will create a new classifier with these features by inputting them into another Adaboost classifier. The base classifier for making the stacking ensemble of classifiers is random forest. For parameter tuning, we manually selected values for the number of iterations, i.e., 50 for all the splits and 10-fold cross-validation. From the training data, n numbers of samples are taken with stratified sampling with a replacement where a balance of different classes of the records is maintained for each subset. Each training sample is taken to train the individual RF classifier to determine the performance. Relative weight is assigned to each individual classifier based on performance, and a Meta RF classifier is created[49,50].

The DCA strategy used here is to use the random forest as a base classifier to involve reducing error-rate in prediction models. The steps of the DCA (i.e., GA-AdaBoost-Random Forests algorithm) are shown in Figure 3 below.

```
Input: T: training set, T=m_i(i=1,2,…,p), labels n_i Î N
X: Iterations number
R:Learn (Random Forests algorithm as base learner)
f: number of input instance to be used at each of the tree
G: number of generated trees in random forest
Step I: Assign P sample $(m_1,n_1),..,(m_p,n_p)$;  $m_i$ÎM, $n_i$Î {-1,+1}
Step II: Initialize the weights of $DTN(i)=1/p$, i=1,…,p)
Step III: for x=1,…,X
    Step IV: Empty E with the distribution $DTN_x$
    Step V: for g=1 to G
        Step VI: Tg = booststrapSample($T$)
        Step VII: $C_g$ = BuildRandomTreeClassifiers($T_g$,f)
        Step VIII: E=EÈ{$C_g$}
    Step IX: next g
    Step X: Get weak hypothesis $h_x$:M®{-1,+1} with its error:$\varepsilon_x = \sum_{i=h_x(m_i)¹n_i} DTN_x(i)$
    StepXI:Update distribution $DTN_x$:$DTN_{x+1}(i) = \frac{DTN_x(i)\exp(-\alpha_x n_x x_x(m_x))}{Z_x}$
StepXII:next x
Step XIII: Output : $H_{(m)}=\text{sign}\left(\sum_{x=1}^{X} \alpha_x h_x(m)\right)$
```

Fig.3. DCA (i.e., GA-AdaBoost-Random Forests) algorithm

## 4. Assessment of Performance of ML Algorithms

The authors, in this paper, applied statistical metrics to evaluate the classification performance of ML algorithms. The metrics include –

4.1. Accuracy
4.2. Kappa Statistic for each model
4.3. ROC Curve and AUC values
4.4. Wilcoxon ranked sum(WLS) test

### *4.1. Accuracy*

For the binary classification, evaluation metrics include Accuracy, Sensitivity, and Specificity.
The metrics are defined as follows:

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

$$\text{Sensitivity} = \frac{T_P}{T_P + F_N}$$

$$\text{Specificity} = \frac{T_N}{T_N + F_P}$$

where:

$T_P$, $T_N$, $F_P$ and $F_N$: indicates True Positive, True Negative, False Positive and False Negative respectively
+VE: patient with malignancy correctly classified; $F_P$-False
+VE: patient with benignity incorrectly classified; $T_N$-True
-VE: patient with benignity correctly classified; $F_N$-False
-VE: patient with malignancyincorrectly classified; $T_P$-False

In other words: the term 'accuracy' measures the rate of rightly classified instances, 'sensitivity' is the rate of wrongly classified instances with BC and 'specificity' is the rightly classified instances without BC

### *4.2. Kappa Statistic for each model*

Cohens Kappa statistic enables accuracy checks of classification. The Kappa score gives the measure of the accuracy of classification in the range [-1,1]. This statistic compares the observed (between the coders) and predicted agreement (the probability that the coders agree by chance)[51]. A value of -1 indicates a sharp disparity between observed and predicted outcomes, while a score of 1 indicates the vice versa. The zero value indicates that observed and predicted outcomes are equal.

### *4.3. ROC Curve and AUC values*

In a binary classifier, the area under the curve (AUC) is a performance metric. It measures the degree to which the curve is up in the northwest corner by contrasting the ROC curves with the area below the curve. A score of 0.5 is referred to as better than a random guess. For the value close to 0.9, it would be a good model, but for a score equal to 1, it would be for an excellent model.

### *4.4. Wilcoxon ranked sum (WLS) test*

The WLS test is a non-parametric statistical hypothesis check comparing two identical and matched samples or perennial measurements on one sample to see if their mean population ranks vary. This test applies to a small sample and non-normal distribution. In this study, the authors use this implication test to compare different models of ML. K-fold cross-validation can be used to generate accuracy scores for each model. The effect will be two samples, one for each pattern. After that, the WLS test method is used to determine whether the two samples vary considerably from each other. If they do, one is more accurate than the other.

## 5. Description of the Dataset

Features are derived from a digitized representation of the fine needle aspirate (FNA) of the breast mass dataset available in http://archive.ics.uci.edu/ml/datasets/Breast?Cancer?Wisconsin?(Diagnostic). We identify the characteristics of the nucleus present in the image.

*Attribute Information:*

1) ID number
2) Diagnosis (M = malignant type, B = benign type)
3-32)

Ten real-value characteristics are calculated for each cell nucleus:

a) radius (Average lengths from center to perimeter points)
b) texture (Standard deviation of the values of the gray scale)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (The intensity of the concave portion of the contour)
h) concave points (The number of concave parts of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The authors use the Wisconsin (Diagnostic) Data Set for Breast Cancer, Created at Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792 wolberg@eagle.surgery.wisc.edu available at the UCI ML Repository website
(https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). There are no missing values present in that dataset. Features are measured from a digitized image of a fine needle aspirate (FNA) of a breast mass. We identify the features of the cell nuclei present in the image. In this BC study, the authors consider a sample comprising 569 patients with 212 suffering from malignancy. Table 1 displays the biometric data obtained during the physical examination of the patients. The dataset consists of 32 features; the first two are ID, the diagnosis results, i.e., benign or malignant, and rest 30 are real-valued input features.

Fine needle biopsy remains a commonly used diagnostic tool involving minor surgical procedures and providing reliable results with the 'triple test' (combined cytological, clinical, and radiological findings)[32]. The DT, based on the classification law, will further improve BC detection by making statistical decisions. The BC dataset used in our analysis consists of functions that define the characteristics of the cell nuclei. The dataset contains not only the mean of the assessed features from each image, but also the standard error and the highest value to ensure that the averaging process does not ignore even rare anomalies. For example, if we relate two images of the elongated and circular nucleus by their characteristic values, then the mean nucleus radius alone cannot distinguish them. However, they will be conveniently distinguished from the mean radius of the features compared with the maximum radius value, as in circular situations, the difference between the mean and the maximum radius will be slight, and the mean and highest radius value will be dramatically more significant. Therefore, the standard error and the higher (worst) value are important properties for the analysis.

## 6. Training-Testing Partition

In ML, obtaining separate (independent) samples from an original set is a common practice for predictive (supervised) models to be constructed and tuned[52]. The "train and test" approach is the most conventional(Refer to Table 2). The essential thought is to utilize an independent test specimen to build a prescient model with a preparation test and afterward approve the model. This method will decrease the probability of the model's over-fitting, provide a reasonable model accuracy estimate, and improve generalization when using the model on new data. The authors apply this idea by dividing the BC dataset into two sub-sets which are disjointed as depicted in Table 3 below.

Table 2. Description of the dataset

| Sl. No | Attributes | Description | Range of Values | Mean | Standard Deviation |
|---|---|---|---|---|---|
| 1 | id | ID number | | | |
| 2 | diagnosis | The diagnosis of breast tissues | M = malignant, B = benign<br>B - 63%<br>M - 37% | | |
| 3 | radius_mean | Average lengths from center to perimeter points | | 14.13 | 3.52 |
| 4 | texture_mean | standard deviation of the magnitude of the gray scale | | 19.29 | 4.3 |
| 5 | perimeter_mean | Average size of the central tumor | | 91.97 | 24.3 |
| 6 | area_mean | | | 654.89 | 351.91 |
| 7 | smoothness_mean | Average local difference in radius length | | 0.10 | 0.01 |
| 8 | compactness_mean | Average of perimeter^2 / area - 1.0 | | 0.10 | 0.05 |
| 9 | concavity_mean | Average extent of severity of concave sections of contour | | 0.09 | 0.08 |
| 10 | concave points_mean | Average number of concave parts of the contour | | 0.05 | 0.04 |
| 11 | symmetry_mean | | | 0.18 | 0.03 |
| 12 | fractal_dimension_mean | Average of 'coastline approximation' - 1 | | 0.06 | 0.01 |
| 13 | radius_se | Normal error for the average distance between the middle and perimeter | | 0.41 | 0.28 |
| 14 | texture_se | Standard fault for grayscale standard deviation | | 1.22 | 0.55 |
| 15 | perimeter_se | | | 2.87 | 2.02 |
| 16 | area_se | | | 40.34 | 45.49 |
| 17 | smoothness_se | Standard error for local variance in radius length | | 0.01 | 0 |
| 18 | compactness_se | standard error for perimeter^2 / area - 1.0 | | 0.03 | 0.02 |
| 19 | concavity_se | standard error for severity of concave portions of the contour | | 0.03 | 0.03 |
| 20 | concave points_se | Ordinary fault in the contour of the number of concave sections | | 0.01 | 0.01 |
| 21 | symmetry_se | Fractal-dimension-se is a normal error for "coastline approximation"-1 | | 0.02 | 0.01 |
| 22 | fractal_dimension_se | standard error for "coastline approximation" - 1 | | 0.00 | 0.00 |
| 23 | radius_worst | "worst" or largest mean value for mean of distances from center to points on the perimeter | | 16.27 | 4.83 |
| 24 | texture_worst | "worst" or largest mean value for standard deviation of gray-scale values | | 25.68 | 6.15 |
| 25 | perimeter_worst | | | 107.26 | 33.6 |
| 26 | area_worst | | | 880.58 | 569.36 |
| 27 | smoothness_worst | "worst" or greater mean value for local radius length variations | | 0.13 | 0.02 |
| 28 | compactness_worst | "worst" or largest mean value for perimeter^2 / area - 1.0 | | 0.25 | 0.16 |
| 29 | concavity_worst | "worst" or largest mean value for severity of concave portions of the contour | | 0.27 | 0.21 |
| 30 | concave points_worst | "worst" or greater average value for the number of concave contour portions | | 0.11 | 0.07 |
| 31 | symmetry_worst | | | 0.29 | 0.06 |
| 32 | fractal_dimension_worst | "worst" or largest mean value for "coastline approximation" - 1 | | 0.08 | 0.02 |

Table 3. Training and Testing Set Partition

| **Training-Testing Partition** | **Total Training Records** | **Positive Records in Training Set** | **Negative Records in Training Set** |
|---|---|---|---|
| 50-50 | 284 | 111(39%) | 173(61%) |
| 66-34 | 364 | 135(37%) | 229(63%) |
| 80-20 | 455 | 169(37%) | 286(63%) |
| 10 fold cross validation | 569 | 212(37%) | 357(63%) |

## 7. Results and Discussions

The authors developed and simulated a proposed model by using the Python programming language. In this model, the authors perform a comparative study between six state-of-the-art ML algorithms, namely GA, LR, RF, NB, SVM, DT, and the proposed model DCA. Among these six popular ML techniques, some techniques show better accuracy, whereas performances of some other techniques are inferior. To boost up the accuracy and performance of the weak classifier, the authors used advanced ensemble ML.

The ensemble of ML techniques applied to the Wisconsin (Diagnostic) dataset available at the UCI machine learning repository showed 97% accuracy. As shown in Table 4, different approaches yielded different levels of accuracies as NB revealed 91%, while the proposed DCA exhibited 97% accuracy. Table 5 presents the comparison of the standard deviation of accuracies among these classifiers.

Table 4. Comparison of Accuracies

| Training-Testing Partition | Accuracy | | | | | |
|---|---|---|---|---|---|---|
| | LR | NB | SVM | DT | RF | DCA |
| 50-50 | 0.93 | 0.93 | 0.93 | 0.93 | 0.95 | 0.97 |
| 66-34 | 0.94 | 0.91 | 0.93 | 0.93 | 0.94 | 0.97 |
| 80-20 | 0.95 | 0.92 | 0.94 | 0.93 | 0.93 | 0.96 |
| 10 fold cross validation | 0.94 | 0.94 | 0.92 | 0.92 | 0.96 | 0.97 |

Table 5. Comparison of Standard Deviation

| Training-Testing Partition | Accuracy | | | | | |
|---|---|---|---|---|---|---|
| | LR | NB | SVM | DT | RF | DCA |
| Standard Deviation(10 fold cross validation) | 0.044 | 0.048 | 0.103 | 0.026 | 0.043 | 0.021 |

A confusion matrix represents the statistics of real and projected classifications achieved from the analysis of different classification systems. The performance of all such systems is generally assessed by using the data generated in this matrix. Table 6 shows the results generated from confusion matrices by using different ML algorithms. The performance of our proposed model, along with other methods was evaluated based on sensitivity, specificity, and accuracy tests, which use the true negative (TN), true positive (FP), false negative (FN), and false-positive (FP) terms.

Table 6. Comparison of Sensitivity and Specificity

| Training-Testing Partition | LR | | NB | | SVM | | DT | | RF | | DCA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity | Specificity |
| | | | | | | | | | | | | |
| 50-50 | 0.94 | 0.93 | 0.88 | 0.92 | 0.93 | 0.93 | 0.90 | 0.95 | 0.92 | 0.97 | 0.97 | 0.96 |
| 64-36 | 0.95 | 0.94 | 0.86 | 0.95 | 0.90 | 0.95 | 0.94 | 0.92 | 0.87 | 0.98 | 0.99 | 0.94 |
| 80-20 | 0.95 | 0.94 | 0.84 | 0.97 | 0.91 | 0.96 | 0.91 | 0.94 | 0.84 | 0.99 | 0.99 | 0.91 |
| 10 fold cross validation | 0.96 | 0.90 | 0.97 | 0.90 | 0.98 | 0.81 | 0.93 | 0.91 | 0.98 | 0.94 | 0.99 | 0.96 |

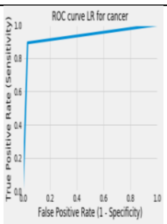
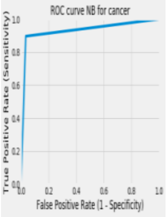
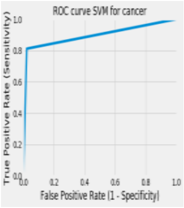
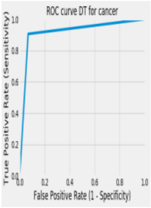
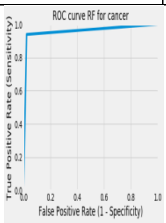
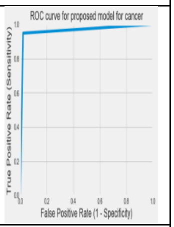
The results demonstrate the potential of our proposed model in the classification of two classes. It is clear from the comparative results that our proposed classification technique has the highest accuracy, sensitivity, and specificity values (accuracy 97%, sensitivity=0.99, specificity=0.96) for breast cancer dataset.

We found from different analyses in our study that the maximum classification accuracy of the LR was 95% with 96% sensitivity and 94% specificity. The RF reached an accuracy of classification of 96% with a sensitivity of 98% and 99% specificity. The SVM achieved an accuracy of classification of 94% with 98% sensitivity and 96% specificity. The

NB reached an accuracy of classification of 94% with a sensitivity of 97% and 97% specificity. The DT achieved an accuracy of classification of 93% with 94% sensitivity and 95% specificity. However, the best of the six classifiers evaluated was performed by our proposed classifier. It reached 97% accuracy in classification with 99% sensitivity and 96% specificity. Table 4, Table 5 and Table 6 show the complete set of results.

Our proposed classifier's analytical findings are also positive. A significant result is to improve the specificity and sensitivity to predict BC using our proposed classifier. Without the disorder, fewer patients would need to be tested for BC due to higher specificity. Simultaneously, higher sensitivity value would also save money and shorten the waiting times of the genuinely ill patients that are critical to saving lives.

Table 7. Comparison of ROC Curve and AUC values

| Training-Testing Partition | LR | NB | SVM | DT | RF | DCA |
|---|---|---|---|---|---|---|
| 10 fold cross validation |  |  |  |  |  |  |
| AUC | 0.99 | 0.99 | 0.97 | 0.92 | 0.99 | 1 |

The ROC charts for these experiments with individual ML techniques are depicted in Table 7. In this table, six ROC charts drawn in different parts for 10-fold cross-validation in blue color. Experimental results show that our proposed classifier outperformed all other state of the art classifiers discussed in the literature study, in terms of cross-validation accuracy. With the proposed model, the generated AUC value reaches 1.

Comparing performances of different ML classifiers might generate an ambiguous result if it has been produced based only on accuracy-based metrics. The Cohen's Kappa Statistics(CKS) value is used to help to produce error-free comparative efficiency of different classifiers. The cost of error must be considered in such evaluations. CKS is an excellent measure in this respect for inspecting classifications that may be due to chance. Usually, CKS takes a value between -1 to +1. As the classifiers calculated Kappa value approaches' 1,' the classifier's performance is assumed to be more realistic than 'by-chance'. Therefore, CKS value is a suggested metric for measurement purposes in the performance analysis of classifiers[51]. This Kappa value is calculated by using Equation 1.

$$CKS = \frac{(pa - pbc())}{(1 - pbc())} \tag{1}$$

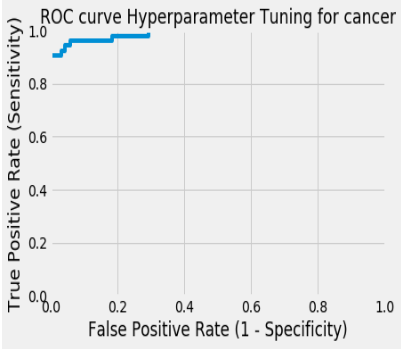
Where pa represents total agreement probability and pbc represents probability 'by-chance'.

The results of CKS analysis of the five popular ML techniques and our proposed model are shown in Table 8. According to this study, it can be proved easily that the proposed model performed much better than other classifiers (value=0.92).

Table 8. Kappa Statistic for each model

| Training-Testing Partition | LR | NB | SVM | DT | RF | DCA |
|---|---|---|---|---|---|---|
| Kappa Statistic | | | | | | |
| 50-50 | 0.86 | 0.85 | 0.86 | 0.85 | 0.89 | 0.92 |
| 66-34 | 0.88 | 0.81 | 0.84 | 0.85 | 0.87 | 0.91 |
| 80-20 | 0.89 | 0.83 | 0.87 | 0.85 | 0.85 | 0.91 |
| 10 fold cross validation | 0.87 | 0.87 | 0.81 | 0.83 | 0.92 | 0.92 |

Table 9. Results with Hyperparameter Tuning

| Parameter | Hyperparameter optimization |
|---|---|
| Accuracy | 0.971 |
| Sensitivity | 1 |
| Specificity | 0.90 |
| roc_auc_score<br>0.99 |  |
| Time required for execution | 0:21:44.020735 |

In this paper, the authors compare the results with genetic algorithm-based hyperparameter optimization is used to derive an optimal solution for predicting breast cancer. This approach has been widely accepted (Han et al. 2020) as it relies on simple iterations and probabilities, leading to an appropriate level of accuracy. The authors considered a population size of 50 with ten generations, thus leading to 500 pipeline or hyperparameter configurations. That is 500 model configurations to evaluate with 5-fold cross-validations. This implies that 2500 models are fit and evaluated on training data in one grid search process. Under this scenario, optimization at different levels, namely, 33%, 50%, 67%, and 84%, yielded an accuracy level of 97.1% in prediction of the disease using logistic regression but throughput is very low i.e., it is time consuming. The authors used the TPOT classifier to estimate accuracy. The test set was set to 30%.

The authors have also used a statistical test proposed by Wilcoxon, which is a non-parametric test to compare the performance of different ML algorithms used in this study[53]. This hypothesis test is used to evaluate statistical differences among two populations by comparing the median of a single column of numerical values against a hypothetical median. In this research, the WLS test is used to solve a six-class problem, while WLS is a feature selection technique for binary values. To handle this problem, authors utilized a one against one approach, which will break n number of classes into n(n-1)/2 binary classes containing a set of all possible pairs of n classes. In this study, 6-class classification is broken down into 15 binary sub-problems. Using the WLS test, the significant features were evaluated separately for each binary issue. The selected features for each binary issue, as explained in the following subsection, were used as an input to one LR, RF, DCA, NB, SVM, and DT classifier. At p values of 1, 0.007, 0.074, 0.720, 0.020, 0.009, 0.858, 0.012 And 0.013, the 6-class problems were solved (Table 10).

Table 10. Wilcoxon signed-rank test

| LR | RF | DCA | NB | SVM | DT |
|---|---|---|---|---|---|
| LR | 1 | 0.013 | 0.074 | 0.720 | 0.020 |
| RF | | 0.007 | 0.009 | 0.858 | 0.012 |
| DCA | | | 0.007 | 0.013 | 0.007 |
| NB | | | | 0.020 | 1 |
| SVM | | | | | 0.013 |

Table 10 lists the p-values of the WLS test for the pairs of accuracies originating from the analysis of different ML algorithms performed in this research work by splitting the BC dataset into various training and testing partitions. According to the results, there is no significant difference between LR and RF, LR and NB, LR and SVM, RF and SVM, and between NB and DT obtained as the p-values are higher than 0.05. WLS test indicates that those algorithm pairs are mutually convergent. The authors cannot reject the null hypothesis for the above cases. However, our proposed model produces different results with all five other classifiers, and in all such cases, the p-value is less than 0.05, which means the medians of these two distributions differ. Thus, the null hypothesis H0 for all these pairs can be rejected, which indicates our proposed model outperforms the compared classification models considerably.

Table 11. Wisconsin (Diagnostic) Data Set for Breast Cancer Performance using proposed model(DCA)

| Year | Method | Classification Accuracy (%) | Sensitivity/ Specificity | Kappa | ROC/AUC | Wilcoxon |
|---|---|---|---|---|---|---|
| Our proposed model(DCA) | DCA | 97 | 0.99/0.96 | 0.92 | 1 | ✓ |

## 8. Conclusion

This paper lays forth a new approach to the collection of features using GA. A blend of several Adaboost classifiers serves as a fitness feature. An increasing data quality, decreasing attributes and a dataset based trained classifier lead to achieving higher performance and prediction of breast cancer shown in Table 11. The DCA system can be used by oncologists to carry out accurate diagnostics in lesser periods using this artificial intelligence technology as a scientific decision-making method. Results indicate that the method used to classify breast cancer into benign vs. malignant tumours' is effective. If the dataset size and features are too large, it is difficult for the oncologists to identify the important and relevant features. However, using the DCA classifier, this can be done with greater accuracy and less time. We used various train-test partitions, 10-fold cross-validation, accuracy, sensitivity, and specificity to demonstrate the proposed method's capacity and effectiveness. The results showed that the proposed method increases the accuracy by up to 97 %, which is considerably higher than that of several single and combined classifiers. The predictive accuracy changes obtained from the ROC and AUC experiments were significant, and with our proposed model, numerically, it reached 100 percent.

The proposed method is, therefore, capable of extracting the most important features to classify malignancy in breast cancer patients, but these findings would be useless without the medical practitioner's cooperation and input. Also, this approach does not intend to replace medical practitioners and physicists, but rather to complement their invaluable efforts to save more human lives. The medical practitioners will, therefore, examine the patterns found through this ensemble approach. As for further research, we plan to examine the variety of the number of classifiers within the ensemble and to compare them in this dimension with other ensemble methods. One way of researching is by using the DCA algorithm in larger data sets.

So, the main contribution of this research paper is not only the development of an ensemble learning model but also the restructuring of the default structure of the boosting algorithm by changing the base estimator. This proposed classifier also has important features, such as its ability to determine the best possible ratio on the training dataset versus the testing dataset to simultaneously find the optimum combination of both sets based on the defined ratio, as well as its experimentation to find an accurate rule using the DCA. Test results show that the proposed ensemble learning classification model is effective in improving performance metrics and classification accuracy compared to its foundation learner and other independent learners specified in the literature. The following are some of the potential implications of the present research work and the proposed classifier.

1. Breast cancer detection is frequently accompanied by several regular medical tests in laboratories in the presence of experts or doctors or during admission at hospital. However, this is usually a very costly and tedious process. This proposed model uses some features extracted from regular lifestyles and a few medical test reports from laboratories in text or number format to predict the disease with greater accuracy.
2. For medical service providers or doctors, this proposed model is intended to provide accurate classification of breast cancer based on fewer precise and explanatory test data from patients. As a result, with the aid of this model, the primary consumer (i.e., medical practitioners or doctors) can predict breast cancer more quickly and accurately (including in cases of clinical assumption) and they can indicate the risk level of the disease efficiently. This proposed model can be used as an electronic doctor; that is, the disease can be diagnosed in the absence of medical practitioners. So, it can help to save lives and significantly minimize medical costs.
3. For research scholars/academics, this proposed model for classification provides a wide scope for future research on improving the prediction accuracy not only of breast cancer but of other medical conditions too.

In summary, in relation to other healthcare datasets, the proposed model is capable of successfully performing medical decision support tasks for various diseases; however, it may produce better classification accuracy with real-world datasets. Researchers can restructure other boosting algorithms by using different weak classifiers as the base classifier to improve the classification accuracy for medical and non-medical datasets.

## 9. Future Scope

The machine-learning methods used in this proposed classifier to predict the early diagnosis of breast cancer are shown to be highly effective. The proposed model in this study was tested with small datasets with a limited number of features. However, the same model needs testing on more divergent real-life cancer datasets, on data from various other disease datasets, or on data that can be obtained from specific electronic health record repositories for non-medical data.

Data collected from patient surveys, observational cohort studies and clinical trials can also help understand the success, safety, and cost of the model. Consequently, I am looking for real-world data in the future research work.

## References

[1] M. U. Sarwar, M. K. Hanif, R. Talib, A. Mobeen, and M. Aslam, “A survey of big data analytics in healthcare,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, pp. 355-359, 2017.

[2] E. W. Steyerberg, *Clinical prediction models*. Cham: Springer International Publishing, 2019, pp. 297-308.

[3] R. Ramani,N.Suthanthira Vanitha,S. Valarmathy,"The Pre-Processing Techniques for Breast Cancer Detection in Mammography Images", IJIGSP, vol.5, no.5, pp.47-54, 2013.DOI: 10.5815/ijigsp.2013.05.06.

[4] Prabhjot Kaur, Yashita Pruthi, Vidushi Bhatia, Janmjay Singh,"Empirical Analysis of Cervical and Breast Cancer Prediction Systems using Classification", International Journal of Education and Management Engineering(IJEME), Vol.9, No.3, pp.1-15, 2019.DOI: 10.5815/ijeme.2019.03.01

[5] A. K. Chaudhuri, D. Sinha, K. Bhattacharya, and A. Das, “An Integrated Strategy for Data Mining Based on Identifying Important and Contradicting Variables for Breast Cancer Recurrence Research,” *Int. J. Recent Tech. Eng.*, vol. 8, March 2020.

[6] Bhagwati Charan Patel,G. R. Sinha,"Energy and Region based Detection and Segmentation of Breast Cancer Mammographic Images", IJIGSP, vol.4, no.6, pp.44-51, 2012.

[7] C.D. Katsis, I. Gkogkou, C.A. Papadopoulos, Y. Goletsis, P.V. Boufounou, G. Stylios, "Using Artificial Immune Recognition Systems in Order to Detect Early Breast Cancer", International Journal of Intelligent Systems and Applications(IJISA), vol.5, no.2, pp.34-40, 2013.DOI: 10.5815/ijisa.2013.02.04

[8] D. Tripathi, I. Manoj, G. R. Prasanth, K. Neeraja, M. K. Varma, and B. R. Reddy, “Survey on classification and feature selection approaches for disease diagnosis,” in *Emerging Research in Data Engineering Systems and Computer Communications*, Singapore: Springer, 2020, pp. 567-576.

[9] M. Tubishat, N. Idris, L. Shuib, M. A. Abushariah, and S. Mirjalili, “Improved Salp Swarm Algorithm based on opposition based learning and novel local search algorithm for feature selection,” *Expert Syst. Appl.*, vol. 145, pp. 113122, 2020.

[10] S. Maldonado, J. López, A. Jimenez-Molina, and H. Lira, “Simultaneous feature selection and heterogeneity control for SVM classification: An application to mental workload assessment,” *Expert Syst. Appl.*, vol. 143, pp. 112988, 2020.

[11] M. Shouman, T. Turner, and R. Stocker, “Applying k-nearest neighbour in diagnosing heart disease patients,” *Int. J. Inf. Educ. Technol.*, vol. 2, pp. 220-223, 2012.

[12] Y. Ji, S. Yu, and Y. Zhang, “A novel naive bayes model: Packaged hidden naive bayes,” in *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, vol. 2, IEEE, August 2011, pp. 484-487.

[13] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, “New support vector algorithms,” *Neural Comput.*, vol. 12, pp. 1207-1245, 2000.

[14] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, and A. Plaza, “On understanding big data impacts in remotely sensed image classification using support vector machine methods,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 8, pp. 4634-4646, 2015.

[15] Y. Tang and J. Zhou, “The performance of PSO-SVM in inflation forecasting,” in *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*, IEEE, June 2015, pp. 1-4.

[16] L. Breiman, “Random forests,” *Mach. Learn*, vol. 45, pp. 5-32, 2001.

[17] X. Chen, and H. Ishwaran, “Random forests for genomic data analysis,” *Genomics*, vol. 99, pp. 323-329, 2012.

[18] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, “Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation,” *Biomed. Signal Process. Contr.*, vol. 52, pp. 456-462, 2019.

[19] M. A. Babyak, “What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models,” *Psychosom. Med.*, vol. 66, pp. 411-421, 2004.

[20] H. Chouaib, O. R. Terrades, S. Tabbone, F. Cloppet, and N. Vincent, “Feature selection combining genetic algorithm and adaboost classifiers,” in *2008 19th International Conference on Pattern Recognition*, IEEE, December 2008, pp. 1-4.

[21] M. Tolba and M. Moustafa, “GAdaBoost: accelerating adaboost feature selection with genetic algorithms,” arXiv preprint arXiv:1609.06260, 2016.

[22] M. Bramer, *Principles of data mining*, vol. 180. London: Springer, 2007.

[23] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining: Adaptive Computation and Machine Learning*. ISBN: 026208290X, 2001.

[24] T. Sridevi and A. Murugan, “A novel feature selection method for effective breast cancer diagnosis and prognosis,” *Int. J. Comput. Appl.*, vol. 88, 2014.

[25] E. Aličković and A. Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest,” *Neural. Comput. Appl.*, vol. 28, pp. 753-763, 2017.

[26] P. Hamsagayathri and P. Sampath, “Performance analysis of breast cancer classification using decision tree classifiers,” *Int. J. Curr. Pharm. Res.*, vol. 9, pp. 19-25, 2017.

[27] M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, and R. Gururajan, “A new nested ensemble technique for automated diagnosis of breast cancer,” *Pattern Recognit. Lett.*, vol. 132, pp. 123-131, 2020.

[28] B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,” *Expert Syst. Appl.*, vol. 41, pp. 1476-1482, 2014.

[29] M. Sewak, P. Vaidya, C. C. Chan, and Z. H. Duan, “SVM approach to breast cancer classification,” in *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, IEEE, August 2007, pp. 32-37.

[30] S. Y. Jin, J. K. Won, H. Lee, and H. J. Choi, “Construction of an automated screening system to predict breast cancer diagnosis and prognosis,” *Basic Appl. Pathol.*, vol. 5, pp. 15-18, 2012.

[31] O. I. Obaid, M. A. Mohammed, M. K. A. Ghani, A. Mostafa, and F. Taha, “Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer,” *Int. J. Eng. Tech.*, vol. 7, pp. 160-166, 2018.

[32] S. Kumari and M. Arumugam, “Application of bio-inspired krill herd algorithm for breast cancer classification and diagnosis,” *Indian J. Sci. Technol.*, vol. 8, pp. 30, 2015.

[33] A. Christobel and Y. Sivaprakasam, “An empirical comparison of data mining classification methods,” *Int. J. Comput. Inf. Syst.*, vol. 3, pp. 24-28, 2011.

[34] D. Lavanya and D. K. U. Rani, “Analysis of feature selection with classification: Breast cancer datasets,” *Indian J. Comput. Sci. Eng.*, vol. 2, pp. 756-763, 2011.

[35] A. Keleş, A. Keleş, and U. Yavuz, “Expert system based on neuro-fuzzy rules for diagnosis breast cancer,” *Expert Syst. Appl.*, vol. 38, pp. 5719-5726, 2011.

[36] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu, “A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis,” *Expert Syst. Appl.*, vol. 38, pp. 9014-9022, 2011.

[37] G. I. Salama, M. Abdelhalim, and M. A. E. Zeid, “Breast cancer diagnosis on three different datasets using multi-classifiers,” *Int. J. Comput. Inform. Tech.*, vol. 1, pp. 36-43, September 2012.

[38] D. Lavanya and K. U. Rani, “Ensemble decision tree classifier for breast cancer data,” *Int. J. Inf. Technol. Converg. Serv.*, vol. 2, pp. 17-24, 2012.

[39] W. Kim, K. S. Kim, J. E. Lee, D. Y. Noh, S. W. Kim, Y. S. Jung, and R. W. Park, “Development of novel breast cancer recurrence prediction model using support vector machine,” *J. Breast Canc.*, vol. 15, pp. 230-238, 2012.

[40] G. R. Kumar, G. A. Ramachandra, and K. Nagamani, “An efficient prediction of breast cancer data using data mining techniques,” *Int. J. Innov. Eng. Technol.*, vol. 2, pp. 139, 2013.

[41] S. Kharya, S. Agrawal, and S. Soni, “Naive Bayes classifiers: a probabilistic detection model for breast cancer,” *Int. J. Comput. Appl.*, vol. 92, pp. 0975-8887, 2014.

[42] K. Sivakami and N. Saraswathi, “Mining big data: breast cancer prediction using DT-SVM hybrid model,” *Int. J. Sci. Eng. Appl. Sci.*, vol. 1, pp. 418-429, 2015.

[43] S. L. Ang, H. C. Ong, and H. C. Low, “Classification Using the General Bayesian Network,” *Pertanika J. Sci. Technol.*, vol. 24, 2016.

[44] A. K. Chaudhuri, D. Sinha, and K. S. Thyagaraj, “Identification of the recurrence of breast cancer by discriminant analysis,” in *Emerging technologies in data mining and information security*, Singapore: Springer, 2019, pp. 519-532.

[45] S. K. Trivedi and S. Dey, “A study of ensemble based evolutionary classifiers for detecting unsolicited emails,” in *Proceedings of the 2014 conference on research in adaptive and convergent systems*, October 2014, pp. 46-51.

[46] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A survey on evolutionary computation approaches to feature selection,” *IEEE Trans. Evol. Comput.*, vol. 20, pp. 606-626, 2015.

[47] B. Yuan and X. Ma, “Sampling+ reweighting: Boosting the performance of AdaBoost on imbalanced datasets,” in *The 2012 international joint conference on neural networks (IJCNN)*, IEEE, June 2012, pp. 1-6.

[48] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in *ICML*, vol. 96, July 1996, pp. 148-156.

[49] R. Sikora, “A modified stacking ensemble machine learning algorithm using genetic algorithms,” in *Handbook of research on organizational transformations through big data analytics*, IGI Global, 2015, pp. 43-53.

[50] B. Bhasuran, G. Murugesan, S. Abdulkadhar, and J. Natarajan, “Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases,” *J Biomed. Informat.*, vol. 64, pp. 1-9, 2016.

[51] A. Ben-David, “Comparison of classification accuracy using Cohen’s Weighted Kappa,” *Expert Syst. Appl.*, vol. 34, pp. 825-832, 2008.

[52] S. K. Trivedi and S. Dey, “Effect of feature selection methods on machine learning classifiers for detecting email spams,” in *Proceedings of the 2013 Research in Adaptive and Convergent Systems*, 2013, pp. 35-40.

[53] F. Wilcoxon, “Individual comparisons by ranking methods,” in *Breakthroughs in statistics*, New York: Springer, 1992, pp. 196-202.

## Authors’ Profiles

**Avijit Kumar Chaudhuri** received his B.SC degree from Calcutta University, A.M.I.E degree in computer engineering from The Institution of Engineers(India). He has completed M.Tech. and MBA degrees in computer science and E-Business from the Sam Higginbottom University of Agriculture, Technology & Sciences – SHUATS , formerly known as AAIDU, India and Annamalai University, India in 2007 and 2014, respectively. Since July 2010, he has been with the Department of Computer Science and Engineering, Techno Engineering College Banipur, where he is an Assistant Professor.

Prior to that, he was the Academic In-Charge, Sikkim Manipal University Learning Centre, Kolkata. Mr. Chaudhuri started his career as a Centre Academic Head in Aptech Computer Education and worked as Visiting Faculty in University Learning Centres of Vidyasagar University and IGNOU. His current research interests include Artificial Intelligence (AI), Data Mining, Machine Learning and Hybrid Learning.

**How to cite this paper:** Avijit Kumar Chaudhuri, Dilip K. Banerjee, Anirban Das, "A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer", International Journal of Intelligent Systems and Applications(IJISA), Vol.13, No.4, pp.24-37, 2021. DOI: 10.5815/ijisa.2021.04.03