Prediction of Breast Cancer Using Machine Learning Techniques

Tahir Iqbal¹, Asif Farooq², Nadeem Sarwar¹*, Mohsin Ashraf², Asma Irshad³

1. Department of Computer Sciences Bahria University, Lahore Campus, Pakistan
2. Department of Computer Science, University of Central Punjab, Lahore Pakistan
3. Centre of Excellence in Molecular Biology, University of Punjab, Lahore, Pakistan

Corresponding Author: Nadeem_srwr@yahoo.com

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: https://doi.org/10.32350/BSR.0401.04

Received: November 10, 2021; Revised: March 23, 2022; Accepted: March 25, 2022

Abstract

Breast cancer affects the majority of women around the world. Females are more likely to die as a result of this condition. By employing a variety of cutting-edge procedures, the samples are collected and the main cause of breast cancer is sought. The most modern techniques are logistic regression discriminant analysis and principal component analysis, both of which are useful in determining the causes of breast cancer. The Breast Cancer Wisconsin Diagnostic Dataset collects information via the Machine Learning Repository approach. As a result of the data correlation matrix processing, we are able to positively root our job. Principal component analysis, discriminant analysis, and logistic regression are utilized to extract the features. Models like Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machines, and Artificial Neural Networks are utilized and their performance is rigorously examined. The results suggest that the proposed strategy works effectively and reduces training time. These new methods help doctors understand the origins of breast cancer and distinguish between tumor kinds. Data mining techniques are used extensively, especially for feature selection. Conclusion: Among all models, the hybrid discriminant-logistic (DA-LR) feature selection model outperforms SVM and Naive Bayes.

Keywords: breast cancer, naïve Bayes, neural network, machine learning, medical imaging support, vector machine
1. Introduction

After cervical cancer, breast cancer is the most common female cancer and one of the most deadly. Around 12% of women in the US have a malignant tumor that can spread to other organs [1]. Increasing survival rates may be made possible by routine screening in conjunction with accurate diagnostics. Initial examination, mammograms, ultrasound, MRI scans, experimental breast imaging and breast biopsy are all part of the diagnostic process [2]. Breast cancer diagnosis relies heavily on data mining. To help doctors correctly diagnose the disease at an early age, medical facilities have a vast amount of data that can be accessed. In the early detection of breast cancer, a mammogram is one of the most commonly used screening methods. There must be further testing to determine whether or not the tumor is malignant or benign after the mammogram has detected it. There is a plethora of variables to consider when analyzing the breast cancer data. If some features are irrelevant or multi-collinear, the classification model may suffer from a loss of precision [3]. Feature selection is essential before data mining and machine learning. [4, 5] Only 20% to 30% of biopsies are found to be cancerous, statistically speaking. The sensitivity of mammography is approximately 84%. The rest of 16% false positive cases are incorrectly referred for further investigatory tests such as a biopsy [6]. Though very accurate, a biopsy is a painful, expensive and time-consuming surgical procedure. Artificial intelligence techniques have been successfully used in breast cancer diagnosis [7-9]. Quinlan [10] achieved 94.74% accuracy using 10-fold cross-validation with C4.5 decision tree method. Pena-Reyes and Sipper [12] proposed a fuzzy-genetic approach and obtained a success rate of 97.36% [11]. Hamilton et al. 1996 presents rule induction through approximate classification and obtained an accuracy of 96%. Abbass [13] applied an evolutionary multi-objective approach to artificial neural achieving 98.1% accuracy with reduced computational cost as compared to traditional backpropagation. Sahar et al. [14] proposed a hybrid K-NN algorithm and achieved an accuracy of 99.14% via 10-fold cross-validation. Akay [15] proposed SVM combined with feature selection using Bare nucleoli, Uniformity of cell shape, Uniformity of cell size, Clump thickness and Bland chromatin as selected feature and obtained an accuracy of 99.51% with 50-50% of training-test partition. Chen et al. [16] suggested rough set-based feature selection combined with support vector machine (RS_SVM) classifier. The classifier achieved an accuracy of 100% with 70%-30% training test partition with five selected features including Clump thickness, Uniformity of cell shape, Marginal adhesion, Bare nucleoli, and Mitosis. Jin et al. [17] concluded to have better results using two binary classifiers with Naive Bayes and Functional Trees (FT) as compared to multiclass classifier (one-step classifier) for predicting diagnosis and prognosis of breast cancer. Kaya [18] proposed a hybrid RSELM model. The RS was applied to reduce the attributes and ELM were utilized for classification. The proposed method obtained an accuracy of 100% with 80%-20% training-test partition with four selected features including Clump thickness, Uniformity of cell shape, Bare nucleoli, Normal nucleoli. Zheng [19] proposed a hybrid of K-means and SVM for feature reduction and classification with an accuracy of 97.38%. El-Baz [20] proposed a hybrid intelligent system that uses rough set-based feature selection and K-NN based classifier. Bhardwaj and Tiwari [21] proposed genetically optimized neural network and obtained an accuracy of 100% with 70-30 training-test partition. Onan [22] proposed a hybrid fuzzy-rough nearest neighbor classification model that consists of three phases: instance selection, feature selection, and classification. The model obtained an accuracy of 99.715%. Hasan et al. [23] proposed a hybrid genetic algorithms and simulated annealing (GSA and accomplished an accuracy of 98.84%. Aalaei et al. [24] applied genetic algorithm-based feature selection and obtained an accuracy of 96.9% with particle swarm classifier. Alickovic and Subasi [25] used genetic algorithm-based feature selection and achieved
an accuracy of 99.48% with rotation classifier. Decision Tree classification is the most commonly used algorithm for decision trees. A simple flowchart like the top-down approach follows its structure. This creates a model for predicting a variable output variable based on one or more input variables. The internal node represents the input variable and the leaf represents the output variable. The classification path is created from the root node to the leaf node by comparing the root attribute with the record attribute. Comparison of all nodes is done until the leaf node is found with the value of N. To select the best attribute, we use a statistical property called Gain, which helps to select a candidate attribute for each node as the tree grows [27]. Decision making is the training phase of classification. After training the tree, it can be converted to the if-then rules [27]. This algorithm gives a better understanding of the overall data structure, but becomes more complicated as the number of features increases. One way to overcome this problem is to use timber pruning. It also solves the problem of over-fitting [26]. New Bayes the New Bayes (NB) algorithm is a machine learning classification technology based on Bayes theory. It is a probabilistic statistical classification used to determine the likelihood of outcomes [28]. The properties are assumed to be independent and contribute to the resultant equivalent input, which reduces the computational complexity to the simple probability multiplication [29]. The training dataset is used to estimate the previous likelihood of a label, and the impact of each attribute meets this pre-probability to obtain a probability estimate. The posterior probability of each label is calculated using the naive Bayesian equation. The highest output labeled is the output of the reference. Little training data is needed to fit together. For most problems in the real world, freedom's assumption is not practical, because the characteristics are often dependent on one another. For example, in the healthcare field, the patient's health conditions and characteristics are dependent on one another and may result in an improper classification of independent assumption and appropriate bias. Nevertheless, the Nave Bayes classification performs better in terms of classification accuracy [30]. Support Vector Machine Support Vector Machine (SVM) is a family of supervised learning algorithms based on statistical learning theory. It is used for the classification and estimation of linear and nonlinear data. The algorithm works by creating a special hyperplane that serves as the boundary for the decision to separate the different classes. Optimal separation is tuned using hyperplane kernel, regularization, gamma and margin. The main advantage of SVM is its high classification accuracy and its ability to create complex linear boundaries, which are robust to over-fitting. The main drawback of this algorithm is that the training time for SVM is very slow [31]. Artificial neural network Artificial neural network (ANN) originates from the biological network of neurons. ANN can be used to model and simulate the relationship between inputs and P outputs. In the ANN model, a layer is a collection of nodes called neurons. The network consists of an input layer, an optional one or more hidden layer, and an output layer. There is a connection between the nodes that transmit the original number as an input signal. The input of each node is calculated based on the inputs and the activation function. Each connection has a fixed weight, which controls the signal between neurons. Known for its ability to study ANN, learning is achieved by constantly updating the associated weight between different neurons. The neural network is a complex adaptive system, and the architecture can vary depending on the flow of information [32 - 34]. Logistic Regression Logistic Output Logistic Regression The output of logistic regression differs with two possible outcomes. The mathematical concept that defines logistic regression is the natural logarithm of the logit-inequality ratio. A simple example of logit is taken from a 2 × 2 contingency table. In general, the logistic regression classification is well suited to describe and test hypotheses about the relationships between the outcome variable and one or more of the categorical or continuous predictor variables [35].
Feature selection is done to reduce the number of variables and to determine the important factors in the analysis phase. The dataset used in this research was provided by Dr. William H. Wright. Retrieved from Wohlberg University of Wisconsin Hospitals, Madison and contains 10 attributes and 699 examples. Main objective of this study is feature reduction by using Logistic Regression, Discriminant Analysis and Principal Component Analysis combined with the tools of machine learning to access and classify the breast cancer on the base of characteristics. Hybrid DA-LR feature reduction is proposed, and models created with reduced features can be tested by classifying those using Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree, and Artificial Neural Network.

2. Methodology

Descriptors of this study used data from the UC Irvine Machine Learning Repository. Dr. William H. Wolberg collected the data and donated it to the UCI Machine Learning Repository on 1992-07-15. The data is multivariate with 10 attributes and 699 instances. Instances arrive periodically as Dr. Wolberg reports his clinical cases in 8 different groups from 2016 to 2018. Group 1 included 367 instances as of January 2018, and it changed to 70 instances in October 2018 at group 2 to 86 instances in November 20 which in total number of instances added up to 699. The database, therefore, reflects this chronological grouping of the data. The attributes include Sample ID, Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single epithelial cell size, Number of bare nuclei, Bland chromatin, Number of normal nuclei, Mitosis. The output variable Classes(diagnosis) has two levels, Malignant or Benign. The block diagram of the proposed methodology is demonstrated in Figure 1.

![Figure 1: Block diagram of proposed methodology](image-url)
Data Preprocessing: Before applying statistical and data mining techniques, the dataset needs to be preprocessed because it contains missing values. Preparation of Information Missing data points necessitate a pre-processing procedure. However, removing the missing value is an option, but it isn't ideal because the dataset isn't particularly large. Substituting the mean or the mode for any missing values is an option. As a result of these methods, it is possible to obtain erroneous estimates of variance and covariance. Instead of guessing at the distribution of each variable in the dataset, it is preferable to estimate those distributions and then use those estimates to fill in the blanks. Heuristic algorithms, such as this one, are used to fill in the blanks in a dataset without introducing significant bias.

Correlation Matrix: In statistics, correlation is used to determine the correlation between two variables. Analyzing the correlation matrix is extremely beneficial prior to developing prediction models. Having multiple predictors in the model leads to more uncertain estimates because of the way multicollinearity affects the precision of each one's impact. When two or more variables are highly linearly related to each other, it is known as multicollinearity.

Feature extraction and Selection: The correlation matrix indicates whether or not there is multicollinearity in the data. In order to reduce the dimensionality of the data, feature extraction is an effective technique. The amount of data required to produce an accurate result grows in proportion to the size of the dataset. Principal component analysis (PCA), discriminant analysis (DA), and logistic regression (LR) are three feature extraction techniques that are investigated in this section for the purpose of reducing dimensions and extracting informative features. Using principal component analysis, it was possible to investigate and reduce the dimensionality of the information. The assumption of multivariate normality is used in the discriminant analysis process. It is not possible to make the multivariate normality assumption hold if the data contains a mixture of independent and dependent variables. The goal of discriminant analysis is to identify the variables that are most effective at distinguishing between the two groups [39].

3. Results

The investigated data set was unbalanced as can be observed in Figure 2, which could lead to bias prediction because the prediction model will tend to predict the class with more observations and accuracy measure, in that case, could not be fully trusted.

![Figure 2: The distribution of data based on their classes](image-url)
Table 1 shows the correlation matrix of the dataset. Apart from the uniformity of cell size and uniformity of cell shape, other variables are not as highly correlated. The correlations among some variables are still more than 0.5 and considered moderately large. Therefore, feature selection and extraction are necessary for choosing inputs for the classification.

**Table 1: Correlation Matrix**

| Variables                  | V1  | V2  | V3  | V4  | V5  | V6  | V7  | V8  | V9  |
|----------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| clump thickness (V1)       | 1.00| 0.65| 0.65| 0.49| 0.52| 0.59| 0.56| 0.54| 0.35|
| uniformity of cell size (V2)| 0.65| 1.00| 0.91| 0.71| 0.75| 0.70| 0.76| 0.72| 0.46|
| uniformity of cell shape (V3)| 0.65| 0.91| 1.00| 0.68| 0.72| 0.71| 0.73| 0.72| 0.44|
| marginal adhesion (V4)     | 0.49| 0.71| 0.68| 1.00| 0.60| 0.67| 0.66| 0.60| 0.42|
| single epithelial cell size (V5)| 0.52| 0.75| 0.72| 0.60| 1.00| 0.58| 0.61| 0.62| 0.47|
| bare nuclei (V6)           | 0.59| 0.70| 0.71| 0.67| 0.58| 1.00| 0.68| 0.59| 0.33|
| bland chromatin (V7)       | 0.56| 0.76| 0.73| 0.66| 0.61| 0.68| 1.00| 0.66| 0.34|
| normal nucleoli (V8)       | 0.54| 0.72| 0.72| 0.60| 0.62| 0.59| 0.66| 1.00| 0.42|
| mitosis (V9)               | 0.35| 0.46| 0.44| 0.42| 0.47| 0.33| 0.34| 0.42| 1.00|

The correlation matrix indicates multicollinearity, hence PCA is used to create new variables that are a linear combination of original variables. As it is shown in Figure 3-4, the first two principal components represent 69% and 7% of total variance respectively. Figure 3 shows graphically how two new independent variables cover original variables. Figure 4 shows the scree plot that has a steep curve followed by a bend and horizontal line. The steep curve has two principal components that are retained to explain most of the variability of the data.

![Figure 3: Coverage of original variables by PC1 and PC2](image)

Table 2 shows the Eigen Vectors of first three principal component. The first three components explain 80% of the total variance in the data, but the Eigen Vectors within the principal component
is not distinguishable. Hence PCA did not provide enough motivation to make dimension reduction. In Table 3 the variables used in discriminant analysis model are shown.

Table 2. Eigen Vectors of Principle Component 1 and 2

| Variable                  | Prin1  | Prin2  | Prin3  |
|---------------------------|--------|--------|--------|
| clump thickness           | 0.302  | -0.140 | 0.866  |
| uniformity of cell size   | 0.381  | -0.046 | -0.019 |
| uniformity of cell shape  | 0.376  | -0.082 | 0.033  |
| marginal adhesion         | 0.333  | -0.052 | -0.412 |
| single epithelial cell size| 0.336  | 0.164  | -0.087 |
| bare nuclei               | 0.335  | -0.261 | 0.0006 |
| bland chromatin           | 0.345  | -0.2281| -0.2130|
| normal nucleoli           | 0.335  | -0.033 | -0.1342|
| mitosis                   | 0.230  | 0.905  | 0.0804 |

Figure 4: Eigenvector loadings and number of components in scree plot

Table 3. Stepwise order of entered variables in Discriminant analysis model

|       | Partial R2 | F Value | Pr > F | Wilks Lambda | Pr < Lambda | Average Squared Canonical Correlation | Pr > ASCC |
|-------|------------|---------|--------|--------------|-------------|---------------------------------------|----------|
| V1    | 0.68       | 1426.2  | <0.0001| 0.32         | <0.0001     | 0.68                                 | <0.0001  |
| V2    | 0.38       | 409.7   | <0.0001| 0.20         | <0.0001     | 0.80                                 | <0.0001  |
| V3    | 0.12       | 92.5    | <0.0001| 0.18         | <0.0001     | 0.82                                 | <0.0001  |
| V4    | 0.07       | 51.01   | <0.0001| 0.17         | <0.0001     | 0.83                                 | <0.0001  |
| V5    | 0.03       | 19.23   | <0.0001| 0.16         | <0.0001     | 0.84                                 | <0.0001  |
V1 = bare nuclei, V2 = uniformity of cell size, V3 = clump thickness V4 = normal nucleoli, V5 = bland chromatin, V6 = uniformity of cell shape

The discriminant analysis suggests that the first five variables discriminate best between the malignant and benign cases with a 0.05 significance level of entry. The F-statistic score determines the order of the variables. The variables entered in the stepwise discriminant analysis will stay if their p-value is less than the significance level of entry. Similarly, the variables entered in the model will stay if their p-value of the overall model is less than the significance level of stay. Feature extraction with Discriminant Analysis is meaningful and is following the correlation matrix. “Uniformity of Cell Shape”, “Single epithelial cell size” and “Marginal adhesion” are not selected as they have a very high correlation of 0.9, 0.75 and 0.71 respectively with “Uniformity of Cell Size”. The values of the identified logistic regression models expressed as Chi-square test for likelihood ratio, Score and Wald p-value should be within the acceptable significance value and are shown in Table 4.

Table 4. Stepwise order of entered variables in Logistic regression model

|    | DF | Order | Chi-Square Score | Pr > Chi Sq |
|----|----|-------|------------------|-------------|
| V1 | 1  | 1     | 462.2739         | <0.0001     |
| V2 | 1  | 2     | 180.4102         | <0.0001     |
| V3 | 1  | 3     | 30.0228          | <0.0001     |
| V4 | 1  | 4     | 16.6428          | <0.0001     |
| V5 | 1  | 5     | 10.2061          | 0.0014      |
| V6 | 1  | 6     | 5.2962           | 0.0214      |

The logistic regression analysis suggests that six variables are essential to classify effectively between malignant and benign tumors. With 0.01 significance level for entry, the first four variables are selected in the model. Classification and prediction of breast cancer type is performed using all features in the dataset with methods named in the prior section, and their performance is compared in Table 5. Table 5 shows the classification result using all features in the model. As can be seen in table 5, SVM performs the best with highest accuracy and AUC. In Tables 5-8 when the upper bound of CI is 1, it is a round number of 0.9 with 5 digits that make it close enough to 1.

Table 5. Performance of classifiers for all features

| Method | Accuracy | Sensitivity | Specificity | AUC | P-value of AUC |
|--------|----------|-------------|-------------|-----|----------------|
| NB     | 0.961 (0.81 to 0.97) | 0.970 (0.69 to 0.97) | 0.958 (0.81 to 0.96) | 0.964 (0.77 to 0.96) | 0.012 |
| DT     | 0.952 (0.90 to 0.96) | 0.865 (0.71 to 0.89) | 0.993 (0.75 to 0.99) | 0.933 (0.62 to 0.95) | 0.032 |
The logistic regression selected the following variables:

1. Bare Nuclei
2. Uniformity of Cell Shape
3. Clump Thickness
4. Bland Chromatin
5. Marginal Adhesion

Table 6 shows the performance of classifiers with LR feature selection which shows the results with LR selected features in the classification model. As it is seen in table 6, NB and SVM perform better than all other classifiers. Comparing the LR feature selection classification with all feature classification, performance is improved with LR selected features. The significance level of alpha equal to 0.05 is considered when the null hypothesis is tested.

### Table 6. Performance of classifiers for LR features

| Method | Accuracy      | Sensitivity   | Specificity   | AUC          | P-value of AUC |
|--------|---------------|---------------|---------------|--------------|----------------|
| NB     | 0.961 (0.73 to 0.97) | 0.940 (0.54 to 0.95) | 0.972 (0.61 to 0.97) | 0.960 (0.71 to 0.97) | 0.0277         |
| DT     | 0.923 (0.57 to 0.96) | 0.805 (0.59 to 0.88) | 0.979 (0.53 to 0.98) | 0.903 (0.53 to 0.99) | 0.0448         |
| LR     | 0.961 (0.67 to 0.99) | 0.910 (0.62 to 0.96) | 0.986 (0.66 to 0.99) | 0.933 (0.47 to 0.99) | 0.0395         |
| SVM    | 0.971 (0.81 to 0.98) | 0.985 (0.79 to 0.99) | 0.972 (0.84 to 0.98) | 0.967 (0.76 to 0.98) | 0.0015         |
| ANN    | 0.938 (0.41 to 0.95) | 0.865 (0.43 to 0.92) | 0.720 (0.43 to 0.89) | 0.907 (0.51 to 0.84) | 0.0458         |

The discriminant analysis selected the following variables: 1. Bare Nuclei 2. Uniformity of Cell Size 3. Clump Thickness 4. Normal Nuclei

4. **Discussion**

The paper presents an extensive comparative data mining and machine learning analysis performed on breast cancer dataset. The correlation matrix of features indicates the presence of multicollinearity. Therefore, feature reduction is investigated using PCA, LR, and DA to reduce
the dimension and to increase classification power. Comparing the results of classification performance metrics of ANN, DT, LR, SVM and NB on four different sets of features, showed both NB and SVM have superior performance when they are fed with DA selected features. These four sets of features include a set of all features selected, features that are selected by LR, features that are selected by DA and hybrid DA-LR feature selection. The diagnosis of breast cancer can be very expensive and risky through mammography and biopsy [1,2]. The risk of biopsy is when the diagnosis is positive, but the patient does not have cancer this comes with a huge load of mental and emotional stress and discomfort [40]. During the last decades, many kinds of research have invested in breast cancer diagnosis using data analytics and later on machine learning. To this aim, data of patients that might have breast cancer is analyzed using different techniques. Data sets that are used in the literature might vary in terms of size and variety of variables. Collecting related data is time-consuming and a higher number of features would not necessarily lead to higher accuracy in diagnosis [3,4]. For this reason, in many of breast cancer diagnosis studies, or similar applied health-related studies feature selection is an important part of the methodology. In this study, we tried combining different feature selection methods with different classification models to find out using which of these combinations leads to higher accuracy. Feature selection and classification models are chosen based on their frequency of use in highly cited journal papers [7-25]. Although PCA has been proved to be a strong dimension reduction technique, we did not find it very insightful with our case study and hence did not use its outputs in further steps [45]. From physical examination to biopsy to imaging tests such as mammogram and MRI, diagnostic methods have evolved over the years (MRI). Breast cancer and other types of cancer patients' chances of survival skyrocket when detected early [46]. To fairly get an assessment of the impact of features selection on classification results, first, we applied all five classifiers, NB, DT, LR, SVM and ANN when all 10 features are included within the dataset. As shown in Table 5, SVM outperforms other classifiers with a significantly better accuracy and AUC. Next, after selecting 5 features using LR. As shown in Table 6, the overall performance of all classifiers has significantly improved especially ANN while SVM still has the best rank among other classifiers, NB has also reached a high accuracy as SVM. Furthermore, the analysis was conducted by feeding the classifiers with the results of LR choosing 4 features. While SVM still has the best performance regarding the performance evaluation metrics, the performance of ANN has significantly dropped down which shows the sensitivity of ANN on a number of times the models is being run comparing to other classification models. Later we tried feeding the classification models with the features that are chosen by a hybrid method of LR-DA which lead to 6 features selected. Bare Nuclei and Clump Thickness are the two features that are selected with LR, DA and hybrid LRDA. While Bland Chromatin, Marginal Adhesion and Uniformity of Cell Shape are selected by LR but not DA, it is selected in hybrid LA-DR and Uniformity of Cell Size that was selected by DA is removed from the hybrid selection. It is worth mentioning that Normal Nuclei that was selected by both LR and DA, is not selected by the hybrid model. This selection is worthy because after running all the classification models for 2000 times, there is not much variation in the confidence intervals and p-values show the significance of the results. As shown in Table 8, with hybrid feature selection, NB and SVM outperforms other classifiers and have improved accuracy and AUC as compared to LR and DA feature selection. The proposed DA-LR feature selection performs best out of all techniques using SVM classifier. Therefore, based on the results, SVM is the most suitable method for classification of breast cancer data while proposed hybrid DA-LR is the best technique for feature reduction. As shown and discussed in this study, the power of SVM in diagnosing breast cancer data with high accuracy is aligned with the results of the reviewed
literature [15,16,19] and that when right features are selected, SVM can achieve high accuracy in predicting patient’s malignancy in a short amount of time. As a future direction to extend this study, we intend to use a data set with a high number of observations and to try different multivariate-classification methods. In addition, running sensitivity analysis on parameters of each classification model can help validate the robustness of each model.

Conclusion
The most advanced techniques involved are logistic regression discriminant analysis and principal component analysis, which are handy to find out the reasons of breast cancer. Different types of cancer can be diagnosed by studying different features with the help of reported techniques. To extract and select the features the most applicable methodology is to use Data mining. Many techniques were developed and analyzed for the diagnosis of tumors. The main purpose to diagnose the breast cancer depends on the relevant features extraction and selection from already fed data. Machine Learning Repository method is used by the Breast Cancer Wisconsin Diagnostic Dataset for the collection of information. The Naive Bayes and Support Vector machine classification outperforms other classification methods, and the hybrid discriminant-logistic (DA-LR) model feature selection performs best among all models.

Conflict Of Interest
The Authors declare no conflict of intrest.

References
[1]. DeSantis C, Ma J, Bryan L, Jemal A. Breast cancer statistics, 2013. CA: A Cancer Journal for Clinicians. 2014;64(1):52-62.
[2]. T.A.C. Society. Breast Cancer Early Detection and Diagnosis. [Online]. Available from: https://www.cancer.org/cancer/breast-cancer.
[3]. Abe N, Kudo M, Toyama J, Shimbo M. A divergence criterion for classifier-independent feature selection. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer. Berlin, Heidelberg, 2000; pp. 668-676.
[4]. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research 2003;3:1157-1182.
[5]. C Society, "Breast Biopsy" [Online] [18 August 2016]. Available from: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breastbiopsy.
[6]. Breast Cancer Surveillance Consortium. [Online] [23 September 2013]. Available from: http://www.bsc-research.org/statistics/performance/screening/2009/rate_age.
[7]. Abdolmaleki P, Buadu LD, Murayama S, Murakami J, Hashiguchi N, Yabuuchi H, Masuda K. Neural network analysis of breast cancer from MRI findings. Radiation Medicine 1997;15(5):283294.
[8]. Abdolmaleki P, Buadu LD, Naderimansh H. Feature extraction and classification of breast cancer on dynamic magnetic resonance imaging using artificial neural network. Cancer Letters. 2001;171(2):183-191.
[9]. Burke HB, Goodman PH, Rosen DB, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer. 1997;79(4):857-862.
[10]. Quinlan JR. Improved use of continuous attributes in C4. 5. Journal of Artificial Intelligence Research. 1996;4:77-90.
[11]. Pena-Reyes CA, Sipper M. A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine*. 1999;17(2):131-155.

[12]. Hamilton HJ, Cercone N, Shan N. RIAC: a rule induction algorithm based on approximate classification. *Computer Science Department, University of Regina*. 1996.

[13]. Abbass HA. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence in Medicine*. 2002;25(3):265-281.

[14]. Şahan S, Polat K, Kodaz H, Güneş S. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Computers in Biology and Medicine*. 2007;37(3):415-423.

[15]. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*. 2009;36(2):3240-3247.

[16]. Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*. 2011;38(7):9014-9022.

[17]. Jin SY, Won JK, Lee H, Choi HJ. Construction of an automated screening system to predict breast cancer diagnosis and prognosis. *Basic and Applied Pathology*. 2012;5(1):15-18.

[18]. Kaya Y. A new intelligent classifier for breast cancer diagnosis based on a rough set and extreme learning machine: RS+ ELM. *Turkish Journal of Electrical Engineering & Computer Sciences*. 2013;21(Sup. 1):2079-2091.

[19]. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*. 2014;41(4):1476-1482.

[20]. El-Baz AH. Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis. *Neural Computing and Applications*. 2015;26(2):437-446.

[21]. Bhardwaj A, Tiwari A. Breast cancer diagnosis using genetically optimized neural network model. *Expert Systems with Applications*. 2015;42(10):4611-4620.

[22]. Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications*. 2015;42(20):6844-6852.

[23]. Örkçü HH, Doğan MI, Örkçü M. A Hybrid Applied Optimization Algorithm for Training MultiLayer Neural Networks in the Data Classification. *Gazi University Journal of Science*. 2015;28(1):115-132.

[24]. Aalaei S, Shahraki H, Rowhani-manesh A, Eslami S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iranian Journal of Basic Medical Sciences*. 2016;19(5):476-482.

[25]. Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Computing and Applications*. 2017;28(4):753-763.

[26]. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of Medical Systems*. 2012;36(4):2431-2448.

[27]. Mitchell TM, Learning M. McGraw-Hill Science. *Engineering/Math*. 1997;1:27.

[28]. Dey A, Singh J, Singh N. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. *International Journal of Computer Applications*. 2016;140(2):27-31.

[29]. Lan K, Wang DT, Fong S, Liu LS, Wong KKL, Dey N. A Survey of Data Mining and Deep Learning in Bioinformatics. *J Med Syst*. 2018;42(8):139.
[30]. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier. 2011.
[31]. Shiffman D, Fry S, Marsh Z. The nature of code. Chapter 7 Cellular Automata. D. Shiffman. 2012:323-330.
[32]. Sharma S, Sharma V, Sharma A. Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. arXiv preprint arXiv:1606.09581. 2016 Jun 28.
[33]. Peng CY, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. The Journal of Educational Research. 2002;96(1):3-14.
[34]. Alrashed AA, Gharibdousti MS, Goodarzi M, de Oliveira LR, Safaei MR, Bandarra Filho EP. Effects on thermophysical properties of carbon based nanofluids: Experimental data, modelling using regression, ANFIS and ANN. International Journal of Heat and Mass Transfer. 2018;125:920-932.
[35]. Enders CK. Applied Missing Data Analysis. Methodology in the Social Sciences Series. Guilford Press. 2010.
[36]. Allison PD. Missing data. Sage Publications. 2001.
[37]. Haitovsky Y. Missing data in regression analysis. Journal of the Royal Statistical Society: Series B (Methodological). 1968;30(1):67-82.
[38]. Hansen J. Using SPSS for Windows and Macintosh: Analyzing and Understanding Data. Pearson College Div. 1999.
[39]. Liong CY, Foo SF. Comparison of linear discriminant analysis and logistic regression for data classification. InAIP Conference Proceedings. 2013;1522(1):1159-1165.
[40]. Jafari-Marandi R, Davarzani S, Gharibdousti MS, Smith BK. An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals. Applied Soft Computing. 2018;72:108-120.
[41]. Hall MA. Correlation-based feature selection for machine learning. PhD thesis. The University of Waikato, Department of Computer Science, Hamilton, New Zealand. Available from: https://www.cs.waikato.ac.nz/~mhall/thesis.pdf
[42]. Gharibdousti MS, Azimi K, Hathikal S, Won DH. Prediction of chronic kidney disease using data mining techniques. Proceedings of the 2017 Industrial and Systems Engineering Conference. 2017, pp. 2135-2140.
[43]. Alrashed AA, Gharibdousti MS, Goodarzi M, de Oliveira LR, Safaei MR, Bandarra Filho EP. Effects on thermophysical properties of carbon based nanofluids: Experimental data, modelling using regression, ANFIS and ANN. International Journal of Heat and Mass Transfer. 2018;125:920-932.
[44]. Begdache L, Kianmehr H, Sabounchi N, Chaar M, Marhaba J. Principal component analysis identifies differential gender-specific dietary patterns that may be linked to mental distress in human adults. Nutritional Neuroscience. 2018:1-4
[45]. Mangal, Anuj, and Vinod Jain. Prediction of Breast Cancer using Machine Learning Algorithms. 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). IEEE, 2021.
[46]. Mridha, Krishna. "Early Prediction of Breast Cancer by using Artificial Neural Network and Machine Learning Techniques. 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). IEEE. 2021.