BioScientific Review (BSR)
Volume 4 Issue 1, 2022
ISSN$_{(P)}$: 2663-4198 ISSN$_{(E)}$: 2663-4201
Journal DOI: https://doi.org/10.32350/bsr
Issue DOI: https://doi.org/10.32350/bsr.0401
Homepage: https://journals.umt.edu.pk/index.php/bsr

Article: **Prediction of Breast Cancer Using Machine Learning Techniques**

Author(s): Tahir Iqbal[1], Asif Farooq[2], Nadeem Sarwar[1], Mohsin Ashraf [2], Asma Irshad[3]

Affiliation: [1]Department of Computer Sciences Bahria University, Lahore Campus, Pakistan

[2]Department of Computer Science, University of Central Punjab, Lahore, Pakistan

[3]School of Biochemistry and Biotechnology, University of the Punjab, Lahore, Pakistan

Article DOI: https://doi.org/10.32350/bsr.0401.04

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

Citation: Iqbal T, Farooq A, Sarwar N, Ashraf M, Irshad A. Prediction of breast cancer using machine learning techniques. *BioSci Rev.* 2022;4(1):59–75. https://doi.org/10.32350/bsr.0401.04

Article QR Code

Tahir Iqbal

Indexing

Crossref

EuroPub

INTERNET ARCHIVE

HJRS HEC Journal Recognition System

J-Gate INDEXED

LOCKSS

A publication of
The Department of Life Sciences, School of Science
University of Management and Technology, Lahore, Pakistan

Research Article

Prediction of Breast Cancer Using Machine Learning Techniques

Tahir Iqbal[1], Asif Farooq[2], Nadeem Sarwar[1*], Mohsin Ashraf [2], Asma Irshad[3]
[1]Department of Computer Sciences Bahria University, Lahore Campus, Pakistan
[2]Department of Computer Science, University of Central Punjab, Lahore Pakistan
[3]School of Biochemistry and Biotechnology, University of the Punjab, Lahore, Pakistan
Corresponding Author: nadeem_srwr@yahoo.com

Article Info

Received:10-11-2021
Revised: 23-03-2022
Accepted:25-03-2022

Keywords

breast cancer, machine learning, medical imaging, naïve Bayes, artificial neural network, support vector machine (SVM)

Abstract

Breast cancer affects a large number of women around the world who are more likely to die as a result of this condition. To seek out the main cause of breast cancer, samples were collected by employing a variety of cutting-edge procedures. The most modern techniques used in this regard are logistic regression, discriminant analysis and principal component analysis (PCA), all of which are useful in determining the causes of breast cancer. The Breast Cancer Wisconsin Diagnostic Dataset collects information about breast cancer via the machine learning repository approach. As a result of data correlation matrix, we were able to positively root our job. PCA, discriminant analysis, and logistic regression were utilized to extract the dataset features. Models such as decision tree, naive Bayes, logistic regression, support vector machine (SVM), and artificial neural networks were utilized and their performances were rigorously examined. The results suggested that the proposed strategy works effectively and reduces the training time. These new methods will help doctors to understand the origins of breast cancer and to distinguish between tumor kinds. Data mining techniques are used extensively, especially for feature selection. Finally, it was concluded that among all models, the hybrid discriminant-logistic (DA-LR) feature selection model outperforms SVM and naive Bayes.

1. Introduction

After cervical cancer, breast cancer is the most common and one of the deadliest cancers among women. Around 12% of women in the US have a malignant tumor that can spread to their other organs [1]. Increasing the survival rate may be 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 stage, medical facilities must have a vast amount of data that can be processed. A mammogram is one of the most commonly used screening methods employed in the early detection of breast cancer. 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 features 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], since only 20-30% of biopsies are determined to be cancerous. The sensitivity of mammography is approximately 84%.The rest (16%) comprise false positive cases which are unduly referred to for further investigatory tests, such as a biopsy [6]. Although 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 the C4.5 decision tree method. Pena-Reyes and Sipper [11] proposed a fuzzy-genetic approach and obtained a success rate of 97.36%. Hamilton et al. [12] presented rule induction through approximate classification and obtained an accuracy of 96%. Abbass [13] applied an evolutionary multi-objective approach using an artificial neural network, achieving 98.1% accuracy with reduced computational cost as compared to the traditional backpropagation. Sahan 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 features 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 using five selected features including clump thickness, uniformity of cell shape, marginal adhesion, bare nucleoli, and mitosis. Jin et al. [17] achieved better results using two binary classifiers, that is, naïve Bayes and functional trees (FT) as compared to a multiclass classifier (one-step classifier) for predicting the diagnosis and prognosis of breast cancer. Kaya [18] proposed a hybrid RSELM model. RS was applied to reduce the attributes and ELM was utilized for classification. The proposed method obtained an accuracy of 100% with 80-20% training-test partition using four selected features including clump thickness, uniformity of cell shape, bare nucleoli, and 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 a 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 operates in three phases: instance selection, feature selection, and classification. The model obtained an accuracy of 99.715%. Hasan et al. [23] proposed a hybrid model of genetic algorithm and simulated annealing (GSA) and achieved 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, such as the top-down approach, follows its structure. It creates a model for predicting an 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. All nodes are compared until the leaf node is found with the value of N. To select the best attribute, we used a statistical property called Gain which helps to select a candidate attribute for each node as the tree grows [26]. Decision-making is the training phase of classification. The tree can be converted to if-then rules after training [26]. This algorithm gives a better understanding of the overall data structure, although it 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 [27]. Naïve Bayes (NB) algorithm is a machine learning classification technology based on the Bayes theory. It is a probabilistic (statistical) classification method used to determine the likelihood of the outcomes [28]. The properties are assumed to be independent and contribute to the resultant equivalent input, which reduces computational complexity to simple probability based 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. For most problems in the real world, an assumption of freedom is impractical because the characteristics are often dependent on one another. For example, in the healthcare field, the patient's health condition and characteristics are dependent on one another and may result in an improper classification of the independent assumption. Nevertheless, naive Bayes classification performs better in terms of classification accuracy.

Support vector machine (SVM) is a family of supervised learning algorithms based on the 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 different classes [30]. 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].

The concept of artificial neural network (ANN) originates with the biological network of neurons. ANN can be used to model and simulate the relationship between inputs and 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 outputs and the activation function. Each connection has a fixed weight which controls the signal between neurons. Learning is achieved by constantly updating the associated weight between different neurons. An Artificial Neural Network is a complex adaptive system and its architecture varies depending on the flow of information [32-34].

The output of logistic regression varies and there are 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 a 2 × 2 contingency table. In general, the logistic regression classification is well-suited to describe and test the hypotheses about the relationship between the outcome variable and one or more of the categorical or continuous predictor variables [35-38].

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. It was retrieved from Wohlberg University of Wisconsin Hospitals, Madison and contains 10 attributes and 699 examples. The main objective of the current study is feature reduction using logistic regression, discriminant analysis, and principal component analysis (PCA), together with the tools of machine learning, to access and classify breast cancer on the basis of its characteristics. Hybrid DA-LR feature reduction is proposed. Moreover, the models created with reduced features can be tested by classifying them using naive Bayes, SVM, logistic regression, decision tree, and artificial neural network.

2. Methodology

The authors of this study used the data from the UCI Machine Learning Repository. Dr. William H. Wolberg collected the data and donated it to the UCI Machine Learning Repository on the following date: 1992-07-15. The data is multivariate with 10 attributes and 699 instances which arrived periodically, as Dr. Wolberg reported his clinical cases in 8 different groups from 2016 to 2018. Group 1 included 367 instances as of January 2018 and the total number of instances in group 2 was 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, and mitosis. The output variable (diagnosis) has two levels, that is, malignant or benign. Figure 1 depicts the proposed methodology via a block diagram.

Data Preprocessing

Before applying statistical and data mining techniques, the selected dataset needed to be preprocessed because it contains missing values. The preparation of information missing data points necessitates a preprocessing procedure. However, removing the missing values is not an ideal option because the dataset isn't particularly large. Substituting the mean or the mode for any missing values remains an option. It is possible to obtain erroneous estimates of variance and covariance through these methods. Hence, instead of guessing the distribution of each variable in the dataset, it is preferable to estimate the distributions and then use the 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.

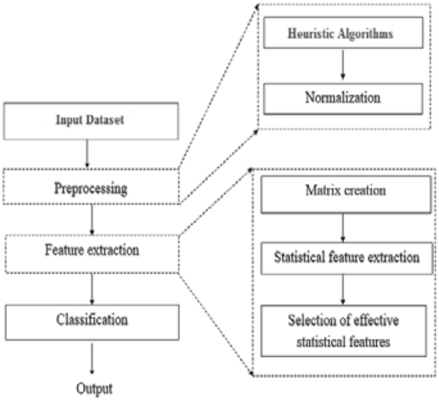

Figure 1: Block diagram of the proposed methodology

Correlation Matrix: In statistics, correlation is used to determine coordination 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 predictor's impact. When two or more variables are 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. Feature extraction is an effective technique to reduce the dimensionality of the data. 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) were three feature extraction techniques investigated in this research for their utilization in reducing dimensions and extracting informative features. Using PCA, it was possible to investigate and reduce the dimensionality of the information. The assumption of multivariate normality was used in discriminant analysis. It is not possible to sustain the multivariate normality assumption 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].

1. Results

The investigated data set is unbalanced as can be observed in Figure 2. It may lead to biased prediction because the prediction model tends to better predict the class with more observations. Due to the preponderance of observations, accuracy measures could not be fully trusted. Table 1 shows the correlation matrix of the dataset. It shows that apart from the uniformity of cell size and cell shape, other variables are not as highly correlated. Correlation values among some variables are still more than 0.5 and considered as moderately large. Therefore, feature selection and extraction are necessary for choosing the right inputs for the required classification.

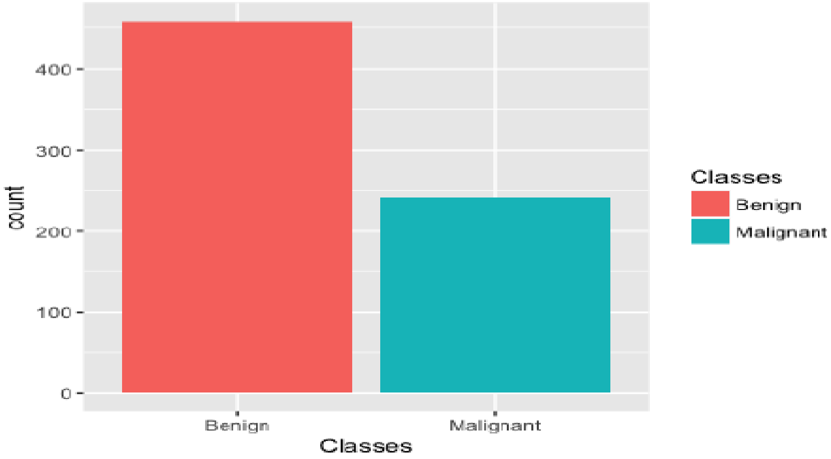

Figure 2: Distribution of data based on its classes

Table 1: Correlation Matrix

Variables	V1	V2	V3	V4	V5	V6	V7	V8	V9
Clump thickness (V1)	1	0.65	0.65	0.49	0.52	0.59	0.56	0.54	0.35
Uniformity of cell size (V2)	0.65	1	0.91	0.71	0.75	0.70	0.76	0.72	0.46
Uniformity of cell shape (V3)	0.65	0.91	1	0.68	0.72	0.71	0.73	0.72	0.44
Marginal adhesion (V4)	0.49	0.71	0.68	1	0.60	0.67	0.66	0.60	0.42
Single epithelial cell size (V5)	0.52	0.75	0.72	0.60	1	0.58	0.61	0.62	0.47
Bare nuclei (V6)	0.59	0.70	0.71	0.67	0.58	1	0.68	0.59	0.33
Bland chromatin (V7)	0.56	0.76	0.73	0.66	0.61	0.68	1	0.66	0.34
Normal nucleoli (V8)	0.54	0.72	0.72	0.60	0.62	0.59	0.66	1	0.42
Mitosis (V9)	0.35	0.46	0.44	0.42	0.47	0.33	0.34	0.42	1

The correlation matrix indicates multicollinearity. Hence, PCA was used to create new variables that comprise a linear combination of original variables. As shown in Figure 3 and Figure 4, the first two principal components represent 69% and 7% of the total variance, respectively. Figure 3 graphically shows how two new and independent variables cover the 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 in the data.

Table 2 shows the eigenvectors of the first three principal components (PC). These three components explain 80% of the total variance in the data, although the eigenvector within the principal component is not distinguishable. Hence, PCA fails to provide sufficient motivation required for dimension reduction. Table 3 shows the variables used in discriminant analysis.

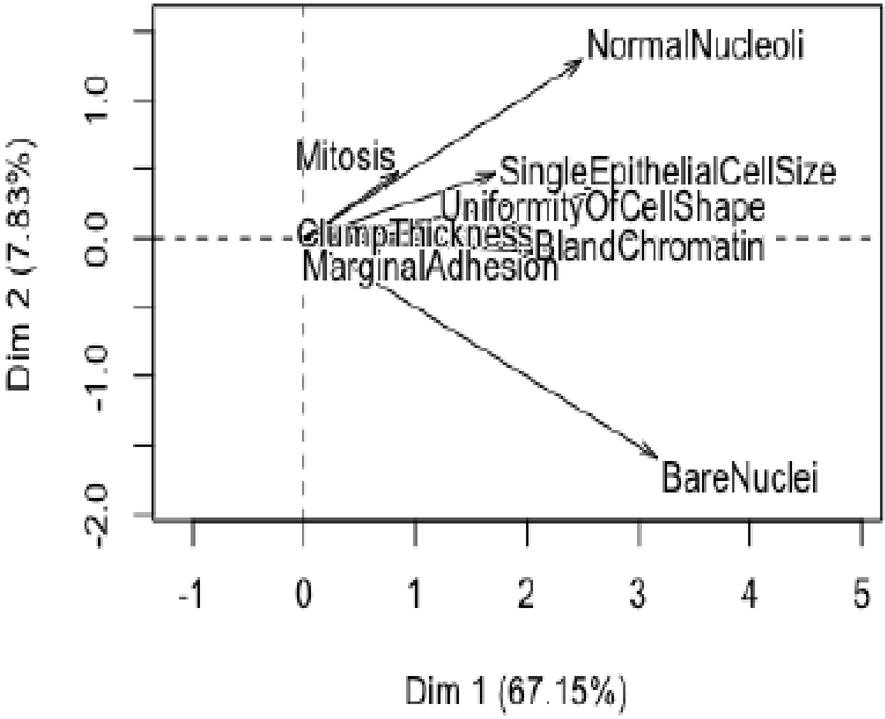

Figure 3: Coverage of original variables by PC1 and PC2

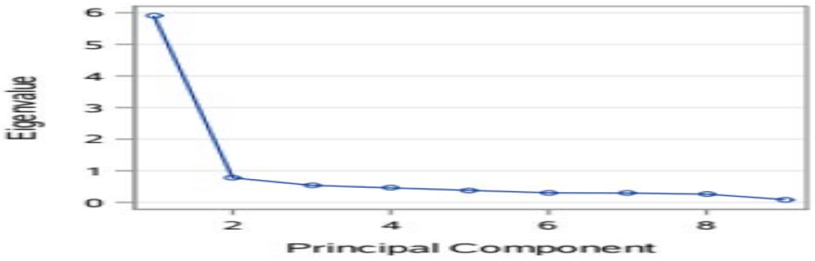

Figure 4: Eigenvector loadings and number of components in scree plot

Table 2. Eigenvectors of PC 1 and PC 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

Table 3. Stepwise Order of Entered Variables in the 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
V6	0.01	7.35	0.0069	0.159	<.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

Discriminant analysis suggested that the first five variables discriminated 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 stay if their p-value is less than the significance level of entry. Similarly, the variables entered in the model stay if the p-value of the overall model is less than the significance level of stay. Feature extraction with discriminant analysis is meaningful and follows the correlation matrix. 'Uniformity of cell shape', 'single epithelial cell size', and 'marginal adhesion' were not selected as they have a very high correlation of 0.9, 0.75 and 0.71, respectively with the variable 'uniformity of cell size'. The values of the identified logistic regression model expressed as Chi-square test for likelihood ratio, score, and Wald p-value should be within the acceptable significance value. These values are shown in Table .

Table 4. Stepwise Order of Entered Variables in the 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 suggested that six variables are essential to distinguish effectively between malignant and benign tumors. With 0.01 significance level for entry, the first four variables were selected in the model. Classification and prediction of breast cancer type was performed using all features in the dataset with methods named in the prior section. A comparison of their performance is given in Table 5, which shows the classification result using all features in the model. As can be seen in Table 5, SVM performs the best with the highest accuracy and AUC. In tables 5-8, when the upper bound of CI is 1, it is the round number of 0.9 with 5 digits that makes 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
LR	0.966 (0.73 to 0.97)	0.910 (0.67 to 0.93)	0.993 (0.82 to 0.99)	0.951 (0.72 to 0.96)	0.033
SVM	0.971 (0.91 to 0.99)	0.955 (0.77 to 0.98)	0.979 (0.81to 0.99)	0.967 (0.73 to 0.98)	0.014
ANN	0.680 (0.52 to 0.75)	0.013 (0.19 to 0.34)	1.00 (0.88 to 1.00)	0.50 (0.38 to 0.69)	0.011

NB = Naive Bayes, DT = Decision Tree, LR = Logistic Regression, SVM = Support Vector Machine, ANN = Artificial Neural Network, AUC= Area Under Curve

Values represent the point estimators and the values in round brackets are the lower and upper 95% confidence interval bounds

The logistic regression (LR) model selected the following variables:
1. Bare nuclei
2. Uniformity of cell shape
3. Clump thickness
4. Bland chromatin
5. Marginal adhension

Table 6 shows the performance of classifiers with LR feature selection. As shown in Table 6, naïve Bayes and SVM perform better than all other classifiers. Comparing the LR feature selection classification with all feature classification, it was determined that performance improved with LR selected features. The significance level of alpha equal to 0.05 was considered when the null hypothesis was tested.

Table 6. Performance of Classifiers for LR Features

Method	Accuracy	Sensitivity	Specificity	AUC	P-value of AUC
NB	0.961 (0.62to 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 (0.41to 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

2. Discussion

The current paper presents an extensive and comparative data mining and machine learning analysis performed on breast cancer dataset. The correlation matrix of features indicate the presence of multicollinearity. Therefore, feature reduction was investigated using PCA, logistic regression, and discriminant analysis to reduce the dimensions and increase the classification power. Comparing the results of the classification performance metrics of artificial neural network, decision tree, logistic regression, SVM, and naïve Bayes based on four different sets of features showed that both naïve Bayes and SVM manifested superior performance when fed with DA selected features. These four sets of features included a set of all features selected, features selected by logistic regression, features selected by discriminant analysis, 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 that a positive diagnosis without the patient having cancer comes with a huge load of mental and emotional stress and discomfort [40]. During the last decades, researchers have invested their efforts in breast cancer diagnosis using data analytics and machine learning. To this aim, the data of patients that might have breast cancer has been analyzed using different techniques. Data sets used in the literature might vary in terms of their size and types of variables. Collecting the related data is a time-consuming activity and a higher number of features would not necessarily lead to a higher accuracy in diagnosis [41, 42]. For this reason, in many of the breast cancer diagnosis studies or similar applied health-related studies, feature selection is an important part of the methodology.

In this study, we attempted to combine different feature selection methods with different classification models to find out which one of these combinations leads to higher accuracy. Feature selection and classification models were 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 in our case study. Hence, we did not use its outputs accordingly [43-45]. From physical examination to biopsy to imaging tests such as mammogram and MRI, diagnostic methods have evolved over the years. The chances of the survival of breast cancer patients as well as other types of cancer patients skyrocket when their cancer is detected early [46].

We applied all five classifiers, that is, naïve Bayes, decision tree, logistic regression, SVM, and artificial neural network to obtain a fair assessment of the impact of features selection on classification results when all 10 features were included within the dataset. As shown in Table 5, SVM outperformed other classifiers with a significantly better accuracy and AUC. Then, we selected 5 features using logistic regression. As shown in Table 6, the overall performance of all classifiers significantly improved, especially that of artificial neural network. However, SVM still held the best rank among other classifiers, while naïve Bayes also achieved a high accuracy comparable with SVM. Furthermore, the analysis was conducted by feeding the classifiers with the results of logistic regression, choosing 4 features. While SVM still gave the best performance keeping in view the performance evaluation metrics, the performance of artificial neural network significantly declined. It shows the sensitivity of this classifier as compared to other classification models.

Later, we attempted to feed the classification models with the features chosen by the hybrid method of LR-DA, which led to 6 features being selected. ‘Bare nuclei’ and ‘clump thickness’ were the two features selected with logistic regression, discriminant analysis, and the hybrid LR-DA. While ‘bland chromatin’, ‘marginal adhesion’, and ‘uniformity of cell shape’ were selected by logistic regression but not discriminant analysis, although they were selected in the hybrid LR-DA and the ‘uniformity of cell size’ was selected by DA but removed from the hybrid selection. It is worth mentioning that although ‘normal nuclei’ was selected by both logistic regression and discriminant analysis, it was not selected by the hybrid LR-DA model. This selection is worthy because after running all the classification models for 2000 times, there was no significant variation in confidence intervals and p-values, which shows the significance of the results.

As shown in Table 8, naïve Bayes and SVM outperformed other classifiers by achieving improved accuracy and AUC through hybrid feature selection, as compared to solely logistic regression or discriminant analysis based feature selection. The proposed DA-LR feature selection performed best among all techniques using SVM classifier. Therefore, based on the results, SVM is the most suitable method for the classification of breast cancer data, while the proposed hybrid DA-LR is the best technique for feature reduction. As shown and discussed in this study, the power of SVM in analyzing breast cancer data with high accuracy is aligned with the findings of the reviewed literature [15, 16, 19]. Moreover, when the right features are selected, SVM can achieve high accuracy in predicting a patient’s malignancy in a short amount of time. In the future, we intend to use a data set with a high number of observations and to try different multivariate-classification methods. Furthermore, running sensitivity analysis on the parameters of each classification model can help to validate the robustness of each model.

Conclusion

The most advanced techniques available for accurate prediction are logistic regression, discriminant analysis, and principal component analysis (PCA), all of 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 the reported techniques. Data mining is the most applicable methodology to extract and select such features. Many techniques have been developed and analyzed for the diagnosis of tumors. Accurate diagnosis of breast cancer depends on the extraction and selection of relevant features from the already existing data. Machine learning repository method

References

1. DeSantis C, Ma J, Bryan L, Jemal A. Breast cancer statistics. *CA Cancer J Clin*. 2013;64(1):52-62. https://doi.org/10.3322/caac.21203
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;668-676. https://doi.org/10.1007/3-540-44522-6_69
4. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Lear Res.* 2003;3:1157-1182.
5. C Society. Breast Biopsy. [Online]. August 18, 2016. Available from: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breastbiopsy.
6. Breast Cancer Surveillance Consortium. [Online]. September 23, 2013. Available from: http://www.bcsc-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. *Radiat Med.* 1997;15(5):283-294.
8. Abdolmaleki P, Buadu LD, Naderimansh H. Feature extraction and classification of breast cancer on dynamic magnetic resonance imaging using artificial neural network. *Cancer Lett.* 2001;171(2):183-191. https://doi.org/10.1016/S0304-3835(01)00508-0
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. https://doi.org/10.1002/(SICI)1097-0142(19970215)79:4<857::AID-CNCR24>3.0.CO;2-Y
10. Quinlan JR. Improved use of continuous attributes in C4.5. *J Artif Intell Res.* 1996;4:77-90. https://doi.org/10.1613/jair.279
11. Pena-Reyes CA, Sipper M. A fuzzy-genetic approach to breast cancer diagnosis. *Artif Intell Med.* 1999;17(2):131-155. https://doi.org/10.1016/S0933-3657(99)00019-6
12. Hamilton HJ, Cercone N, Shan N. *RIAC: a rule induction algorithm based on approximate classification.* University of Regina. 1996.
13. Abbass HA. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artif Intell Med.* 2002;25(3):265-281. https://doi.org/10.1016/S0933-3657(02)00028-3
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. *Comput Biol Med.* 2007;37(3):415-423. https://doi.org/10.1016/j.compbiomed.2006.05.003
15. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst Appl.* 2009;36(2):3240-3247. https://doi.org/10.1016/j.eswa.2008.01.009
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. https://doi.org/10.1016/j.eswa.2011.01.120
17. Jin SY, Won JK, Lee H, Choi HJ. Construction of an automated

screening system to predict breast cancer diagnosis and prognosis. *Basic Appl Pathol.* 2012;5(1):15-18. https://doi.org/10.1111/j.1755-9294.2012.01124.x

18. Kaya Y. A new intelligent classifier for breast cancer diagnosis based on a rough set and extreme learning machine: RS+ ELM. *Turk J Elec Eng & Comput Sci.* 2013;21:2079-2091 https://doi:10.3906/elk-1203-119
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 Syst Appl.* 2014;41(4):1476-1482. https://doi.org/10.1016/j.eswa.2013.08.044
20. El-Baz AH. Hybrid intelligent system-based rough set and ensemble classifier for breast cancer diagnosis. *Neural Comput Appl.* 2015;26(2):437-446. https://doi.org/10.1007/s00521-014-1731-9
21. Bhardwaj A, Tiwari A. Breast cancer diagnosis using genetically optimized neural network model. *Expert Syst Appl.* 2015;42(10):4611-4620. https://doi.org/10.1016/j.eswa.2015.01.065
22. Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Exp Syst Appl.* 2015;42(20):6844-6852. https://doi.org/10.1016/j.eswa.2015.05.006
23. Örkçü HH, Doğan Mİ, Örkçü M. A Hybrid Applied Optimization Algorithm for Training MultiLayer Neural Networks in the Data Classification. *Gazi Uni J Sci.* 2015;28(1):115-132.
24. Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. *Iran J Basic Med Sci.* 2016;19(5):476-482.
25. Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Comput Appl.* 2017;28(4):753-763. https://doi.org/10.1007/s00521-015-2103-9
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. *J Med Syst.* 2012;36(4):2431-2448 https://doi.org/10.1007/s10916-011-9710-5 .
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. *Int J Comput Appl.* 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. https://doi.org/10.1007/s10916-018-1003-9
30. Han J, Pei J, Kamber M. *Data Mining: Concepts and Techniques.* Elsevier. 2011.

31. Shiffman D, Fry S, Marsh Z. Cellular Automata. *The Nature of Code*. 2012:323-330.
32. Sharma S, Sharma V, Sharma A. Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. *Int J Mod Comput Sci*. 2016;4(3):11-16 https://doi.org/10.48550/arXiv.1606.09581
33. Peng CY, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. *J Educ Res*. 2002;96(1):3-14. https://doi.org/10.1080/00220670209598786
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. *Int J Heat Mass Transf*. 2018;125:920-932. https://doi.org/10.1016/j.ijheatmasstransfer.2018.04.142
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. *J R Stat Soc Series B Stat Methodol*. 1968;30(1):67-82. https://doi.org/10.1111/j.2517-6161.1968.tb01507.x
38. Hansen J. Using SPSS for Windows and Macintosh: Analyzing and Understanding Data. *Am Stat*.1999;59(1):113-113. https://doi.org/10.1198/tas.2005.s139
39. Liong CY, Foo SF. Comparison of linear discriminant analysis and logistic regression for data classification. *InAIP Conference Proceedings*. 2013;1522(1):1159-1165. https://doi.org/10.1063/1.4801262
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. *Appl Soft Comput*. 2018;72:108-120. https://doi.org/10.1016/j.asoc.2018.07.060
41. Hall MA. Correlation-Based Feature Selection for Machine Learning. [PhD thesis]. Hamilton, New Zealand: The University of Waikato; 1999. 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;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. *Int J Heat Mass Transf*. 2018;125:920-932. https://doi.org/10.1016/j.ijheatmasstransfer.2018.04.142
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. *Nutr Neurosci*. 2018;23(4):295-308. https://doi.org/10.1080/1028415X.2018.1500198

45. Mangal, Anuj, and Vinod Jain. Prediction of Breast Cancer using Machine Learning Algorithms. *Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).* 2021;464-466. https://doi/10.1109/I-SMAC52330.2021.9640813

46. Mridha, Krishna. "Early Prediction of Breast Cancer by using Artificial Neural Network and Machine Learning Techniques. *10th IEEE International Conference on Communication Systems and Network Technologies (CSNT).* 2021;582-587.