Applied Medical Informatics
Original Research Vol. 41, No. 1 /2019, pp: 40-52

Breast cancer diagnosis using feature extraction techniques with supervised and unsupervised classification algorithms

Maryam SOLTANPOUR GHARIBDOUSTI[1,*], Syed M. HAIDER[2], Dieudonne OUEDRAOGO[3], and Susan LU[4]

Binghamton University, Department of System Science and Industrial Engineering, State University of New York 4400 Vestal Pkwy E, Binghamton, NY 13902, USA
E-mails: msoltan1@binghamton.edu; shaider1@binghamton.edu; douedra1@binghamton.edu; slu@binghamton.edu

* Author to whom correspondence should be addressed; Tel: +1 607-349-8051

Received: December 27, 2018 /Accepted: March 28, 2019/ Published online: March 31, 2019

Abstract

Background: Breast cancer is a serious disease that affects females around the globe. With the development of clinical technologies, different tumor features have been collected for breast cancer diagnosis. Filtering all the pertinent feature information to support the clinical disease diagnosis is a challenging and time-consuming task. The objective of this research was to diagnose breast cancer based on the extracted tumor features. The main contribution of our study is to use multivariate techniques such as principal component analysis, discriminant analysis and logistic regression for feature reduction combined with machine learning tools to classify and predict the tumor type. A hybrid DA-LR feature reduction is proposed, and models created with reduced features are tested by performing classification using Support Vector Machine, Naive Bayes, Decision Tree, Logistic Regression and Artificial Neural Network. ***Materials and Methods:*** Feature extraction and selection are critical to the quality of classifiers founded through data mining methods. To diagnose tumor through reduced features, a hybrid feature extraction is proposed. We tried to predict the disease based on relevant features in the data. The Breast Cancer Wisconsin Diagnostic Dataset obtained from the UCI Irvine Machine Learning Repository has been used in this study. After data pre-processing, the correlation matrix is generated that suggests the presence of multicollinearity. Feature reduction techniques including principal component analysis, discriminant analysis, and logistic regression are applied to extract features. Classification models namely Support vector machine, Naive Bayes, Decision Tree, Logistic Regression and Artificial Neural Network are created with extracted features, and their performance is compared. ***Result:*** The results not only illustrate the capability of the proposed approach on breast cancer diagnosis but also show time savings during the training phase. Physicians can also benefit from the mined abstract tumor features by better understanding the properties of different types of tumors. ***Conclusion:*** The Naive Bayes and Support Vector machine classification outperforms other classification methods and the model created with hybrid discriminant-logistic (DA-LR) feature selection performs best among all models.

Keywords: Breast cancer diagnosis; Feature extraction; Linear discriminant analysis; Logistic Regression analysis; Principal component analysis; Super vector machine; Artificial Neural Network; Supervised learning; Unsupervised learning

Introduction

Breast cancer is the second most common cancer among women after cervical cancer and one of the leading cancers with higher death rates. In the US, approximately 12% of women are diagnosed with a malignant tumor that can spread to other parts of the body [1]. Regular screening, coupled with appropriate diagnostic measures may have a significant impact on survival rates. The diagnosis involves a series of medical tests including an initial breast exam, mammograms, ultrasound, magnetic resonance imaging (MRI) scans, experimental breast imaging, and breast biopsy [2].

Data mining is extensively used in the diagnosis of breast cancer. The medical facilities have an enormous amount of data that can be utilized to help doctors in correctly diagnosing the disease at an early age. A mammogram is one of the most extensively used screening methods in detecting early breast tumor. If the mammogram detects the tumor, a further diagnosis which could include invasive technique is required to determine whether the tumor is malignant or benign. The breast cancer data consist of a large number of features. Some features are irrelevant or multicollinear these may cause the classification model to decrease its precision [3]. Feature selection is an essential preprocessing step in data mining and machine learning [4].

Statistically, only 20-30% of biopsies taken are found to be actual cancerous [5]. The mammography sensitivity is about 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. 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 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 Naïve 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 RS-ELM 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 Trees

Decision tree classifier is one of the most commonly used algorithms. Its construction follows a simple flowchart like top-down approach [26]. It creates the model to predict the output variable

based upon one or more input variables. The interior 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 testing and comparing the root attribute with the record attribute. The comparison is performed on all the nodes until a leaf node is found with the predicted value. To select the best attribute, a statistical property called information gain is used that helps in selecting the candidate attribute at each node while growing the tree [27]. Decision tree construction is the training step of classification. After the tree is trained, it can be converted to if-then rules [27]. This algorithm provides a better understanding of overall data structure but can becomes complex when the number of attributes increases. Tree pruning is one of the methods to overcome this problem. It also resolves the problem of overfitting [26].

Naive Bayes

Naive Bayes (NB) algorithm is a supervised machine learning classification technique based upon Bayes' theorem. It is a probabilistic statistical classifier used to determine the probability of the outcomes [28]. It assumes that features contribute independently and of equal input to the outcome and thus reduces the computational complexity to simple probability multiplication [29]. The training dataset is used to calculate the prior probability of a label and the influence of each attribute is joined with this prior probability to get a likelihood estimate. The posterior probability is calculated for each label using a Naive Bayesian equation. The label with the highest posterior probability is the output of the prediction [30]. It requires a small set of training data to converge. The assumption of independence is not practical in most of the real-world problems as the features are often dependent on each other. For instance, in the healthcare sector, health conditions and symptoms of a patient are dependent on each other and Naive Bayes independence assumption could lead to invalid classification outcome. However, Naive Bayes classifier produces a good performance in terms of classification accuracy.

Support Vector Machine

Support vector machine (SVM) belongs to the supervised learning algorithm family and is based on statistical learning theory. It is used for both linear and nonlinear data classification and prediction. The Algorithm work by creating a separating hyperplane which acts as a decision boundary separating different class from one another. The optimal separating hyperplane is tuned using kernel, Regularization, Gamma and Margin. The major advantage of SVM is high classification accuracy and the ability to create complex nonlinear boundaries that are robust to overfitting. The main drawback of this algorithm is that the training time for SVM is extremely slow [31].

Artificial Neural Network

Artificial neural network (ANN) is derived from a biological network of neurons. ANN can be used to model and simulate relationships between inputs and outputs. In the ANN model, a collection of nodes termed as neurons constitute a layer. A network will have an input layer, optional one or more hidden layer and the output layer. A connection exists between nodes that transmit the real number as an input signal. The output of each node is calculated based upon by inputs and activation function. Each connection has a certain weight which regulates the signal between the neurons. ANN is known for its ability to learn, learning is achieved by continuously updating the weights associated with a connection between different neurons. Neural network is a complex adaptive system, and internal structure can change based on the information flow [32-34].

Logistic Regression

The output of the logistic regression is dichotomous with two possible outcomes. The mathematical concept that defines logistic regression is the logit-the natural logarithm of an odds ratio. The simplest example of a logit derives from a 2×2 contingency table. Generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables [35].

Feature selection is performed to reduce the number of variables and determine the significant factors in the diagnostic step. The dataset used in this research is obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg and contains 10 attributes and 699 instances.

The objective of this study was to use multivariate techniques such as principal component analysis, discriminant analysis and logistic regression for feature reduction combined with machine learning tools to classify and predict breast cancer based on the extracted tumor features. A hybrid DA-LR feature reduction is proposed, and models created with reduced features are tested by performing classification using Support Vector Machine, Naive Bayes, Decision Tree, Logistic Regression and Artificial Neural Network.

Material and Method

Data Description

The data set used in this research is from UC Irvine Machine Learning Repository. The data is collected in the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg and donated to 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 1989 to 1991. Group 1 included 367 instances as of January 1989, and it changed to 70 instances in October 1989 at group 2 to 86 instances in November 1991 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 dataset contains missing values and requires preprocessing before applying statistical and data mining technique.

Data Preprocessing

The dataset includes missing values and requires pre-processing. Dropping the missing value is doable but not preferable as the dataset is not large. Missing values can be imputed using mean or mode substitution. These methods can produce bias estimates of variances and covariances [36-38]. A better approach is to estimate the distribution of each variable in the dataset, and the missing values are filled based on those distributions. This method is considered as a heuristic algorithm which imputes missing values in a dataset without inserting much bias.

Correlation Matrix

Correlation is a statistical procedure to determine the degree of relationship between variables. Before building prediction models, analyzing the correlation matrix is very useful. The presence of multicollinearity interferes the precise effect of each predictor and makes the estimates very sensitive to minor changes in the model. Multicollinearity is a situation in which two or more variables are highly linearly related to each other.

Feature Extraction and Selection

The correlation matrix indicates the presence of multicollinearity. Feature extraction is a useful technique to reduce the dimensionality of the data. As the dimensions of the dataset increases, the amount of data to produce accurate result becomes large. In this section, feature extraction techniques namely *Principle component analysis* (PCA), *Discriminant analysis* (DA) and *Logistic regression* (LR) are explored to reduce the dimensions and to extract the informative features.

The principal component analysis was used to explore and reduce the dimensionality of the data.

The discriminant analysis has the multivariate normality assumption. If the data is a mixture of independent and dependent variables, the multivariate normality assumption will not hold. The objective of the discriminant analysis is to identify the variables that discriminate best between the two groups [39].

The stepwise discriminant analysis is applied in SAS version 9.2 software. The significance level of entry and stay is set to 0.01. The analysis reported that 6 variables meet the 0.01 significance level of entry. Wilks' lambda is calculated at each step to test for significant difference between groups. Smaller values indicate a greater discriminatory ability of the function.

Logistic regression (LR) is recommended when the independent variables do not satisfy the multivariate normality test [40]. The stepwise logistic regression is applied with a 0.05 significance level for entry in SAS software. Akaike Information and Schwarz Criterion are reported as model fit statistics.

Classification Methods

The classification methods used in this study were Support Vector Machine (SVM), logistic regression, Neural Network (NN), Decision Trees (DT) and Naive Bayes (NV). Data was split into training and testing set in 70-30 % ratio. First, classification models are applied to the original dataset including all features. Performance measurement criteria such as accuracy, sensitivity, specificity and Area Under Curve (AUC) for Receiver Operating Characteristic (ROC) Curve are used to evaluate various techniques. Second, the experiment is repeated for the features extracted from Logistic Regression and Discriminant Analysis. Last, a hybrid "DA-LR" model is proposed, and the above process is repeated.

Keeping the split percentage constant, training and testing data are selected five times randomly from the dataset and the average performance measures are reported.

A hybrid DA-LR feature selection is proposed to include features selected from Discriminant and

In the proposed hybrid method, aâ logistic regression analysis correlation-based feature selection methodology is employed. The motivation is to remove features that are irrelevant and redundant with respect to classification purpose [41-43]. A rule is created to use a correlation threshold of ± 0.75. The idea is to remove highly correlated features from the hybrid model. A feature will be included in the hybrid model if its correlation with all other features is less than the threshold.

Results

The investigated data set was unbalanced as can be observed in Figure 1, 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.

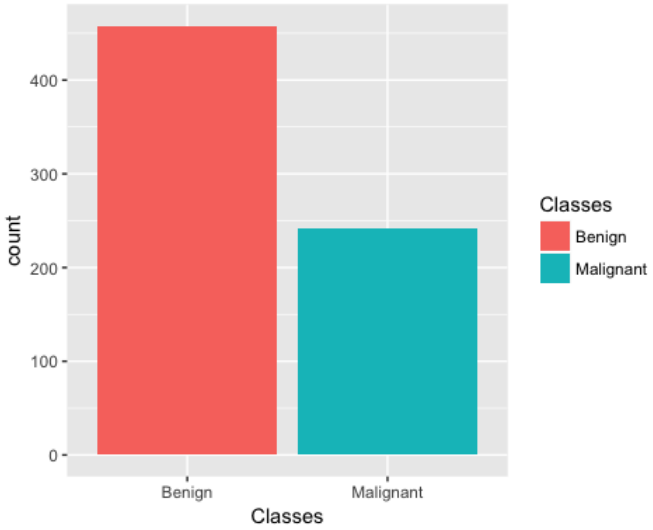

Figure 1. The distribution of data based on their classes

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	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 is used to create new variables that are a linear combination of original variables. As it is shown in Figure 2-3, the first two principal components represent %69 and %7 of total variance respectively. Figure 2 shows graphically how two new independent variables cover original variables. Figure 3 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.

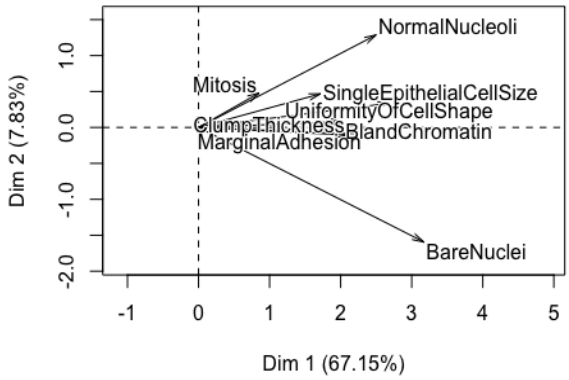

Figure 2. Coverage of original variables by PC1 and PC2

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.

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

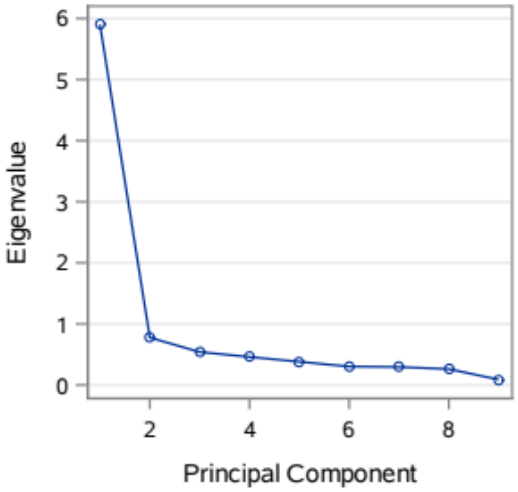

Figure 3. Eigenvector loadings and number of components in scree plot

The variables associated to the discriminant analysis model are entered in the order presneted in Table 3.

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

	Partial R^2	**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

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

V1 = bare nuclei, V2 = uniformity of cell shape,
V3 = clump thickness V4 = bland chromatin, V5 =marginal adhesion, V6 = normal nucleoli

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
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 the round brackets are the lower and upper 95% confidence interval bounds

The logistic regression 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 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 (0.62to 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

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 the round brackets are the lower and upper 95% confidence interval bounds

The discriminant analysis selected the following variables:

1. Bare Nuclei
2. Uniformity of Cell Size
3. Clump Thickness
4. Normal Nuclei

Table 7 shows the results of classification for features selected by DA. This table shows the classification result with DA selected features in the model. NB and SVM perform better than other classifiers.

Table 7. Performance of classifiers for DA features

Method	Accuracy	Sensitivity	Specificity	AUC	P-value of AUC
NB	0.961 (0.64 to 0.98)	0.985 (0.64 to 0.99)	0.951 (0.34 to 0.95)	0.968 (0.64 to 0.98)	0.0116
DT	0.928 (0.51 to 0.95)	0.850 (0.47 to 0.89)	0.965 (0.62 to 0.88)	0.900 (0.29 to 0.94)	0.0349
LR	0.957 (0.32 to 0.97)	0.895 (0.42 to 0.93)	0.986 (0.72 to 0.99)	0.940 (0.52 to 0.97)	0.0478
SVM	0.973 (0.68 to 0.99)	1.0 (0.99 to 1.00)	0.958 (0.69 to 0.97)	0.979 (0.74 to 0.99)	0.0239
ANN	0.323 (0.21 to 0.62)	1.0 (0.93 to 1)	0.006 (0.0 to 0.42)	0.503 (0.35 to 0.72)	0.0245

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 the round brackets are the lower and upper 95% confidence interval bounds

The best performing variables identified by the Hybrid Selected Features were:

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

Table 8 shows the results of classification for features selected by a hybrid algorithm based on DA and LR.

Table 8. Performance of classifiers for the hybrid method

Method	Accuracy	Sensitivity	Specificity	AUC	P-value of AUC
NB	0.966 (0.45 to 0.98)	0.985 (0.41 to 0.99)	0.985 (0.46 to 1.00)	0.971 (0.35 to 0.99)	0.0144
DT	0.923 (0.61 to 0.96)	0.805 (0.49 to 0.88)	0.958 (0.46 to 0.98)	0.911 (0.71 to 0.87)	0.0474
LR	0.961 (0.43 to 0.98)	0.910 (0.70 to 0.96)	0.986 (0.44 to 1.00)	0.948 (0.61 to 0.96)	0.0237
SVM	0.979 (0.68 to 0.99)	1.0 (0.97 to 1.00)	0.986 (0.69 to 0.99)	0.985 (0.74 to 0.99)	0.0446
ANN	0.680 (0.47 to 0.81)	0.0 (0.0 to 0.12)	1.0 (0.83 to 1)	0.502 (0.23 to 0.62)	0.0220

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 the round brackets are the lower and upper 95% confidence interval bounds

Discussion

The paper presents an extensive comparative data mining and machine learning analysis is 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 [44]. 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 LR-DA. 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 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.

List of abbreviations

MRI= Magnetic resonance imaging
NB = Naive Bayes
DT = Decision Tree
LR = Logistic Regress
SVM = Support Vector Machin
ANN = Artificial Neural Network
MRI= Magnetic Resonance Imaging

PCA= Principal Component Analysis
AUC= Area Under Curve
DA= Discriminant Analysis

Conflict of Interest

The authors declare that they have no conflicts of interest.

Acknowledgements

This research was supported by the Watson School's System Science and Industrial Engineering Department at the State University of New York at Binghamton. The breast cancer databases were obtained from the University of Wisconsin Hospitals, Madison from dr. William H. Wolberg.

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/breast-biopsy.
6. Breast Cancer Surveillance Consortium. [Online] [23 September 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. Radiation Medicine 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 Letters 2001;171(2):183-191.
9. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr FE, Marks JR, Winchester DP, Bostwick DG. 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 Mİ, Örkçü M. A Hybrid Applied Optimization Algorithm for Training Multi-Layer Neural Networks in the Data Classification. Gazi University Journal of Science 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. 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, NewZealand. 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.