Original Article

Machine Learning-Based Decision Support System for Early Detection of Breast Cancer

Mochen Li[1,*], Gaurav Nanda[1], Santosh Chhajedss[2], Raji Sundararajan[1]

[1]School of Engineering Technology, Purdue University, Grant St. West Lafayette, USA.
[2]MET's Institute of Pharmacy, Bhujbal Knowledge City, Nashik, Maharashtra, INDIA.

ABSTRACT

Background: Breast cancer is one of the leading causes of death of women in the United States and also one of the most malignant cancer among women worldwide. Early, more accurate detection of breast cancer enables extended longevity at a reduced cost. Towards this, analyzing the available big data using tools, such as Machine learning-based decision support systems can improve the speed and accuracy of early detection of breast cancer. In this paper, we examined the prediction performance of various state-of-the-art machine learning models and a decision support system based on these models that provided the predicted category along with a prediction confidence measure. **Methods:** The various machine learning (ML) algorithms applied include Decision Tree, Naïve Bayes, k-Nearest Neighbors (kNN) and Support Vector Machine (SVM). We also analyzed the effect of multiple feature selection approaches on the prediction performance. We used the Breast Cancer Wisconsin Dataset from Wisconsin Prognostic Breast Cancer (WPBC) with 569 digitized images of a fine needle aspirate (FNA) of breast mass and 10 real-valued feature information. The performance of the ML model was evaluated using the ten-fold cross-validation approach and also on a prediction set comprising of 20% data with the models trained on remaining 80% data. Sensitivity and Specificity were used as the primary measures of performance. **Results:** Among all five machine learning methods, SVM had the best performance. Except for the kNN algorithm, the performance of the other three algorithms, Logistic Regressions, Naïve Bayes and Decision Trees, were also quite close to SVM. The prediction performance of the decision support system was better than any individual ML model where the prediction confidence was "High" or "Medium". **Conclusion:** We found that feature selection improved the performance and computation cost for all ML models. By building the ML-based decision support system with the optimal feature subset, the prediction performance for breast cancer can be improved to 96% which means it can provide powerful assistance to doctors and patinets. On the other hand, as the size of the data set increases, the processing of data with a lot of features can increase the computation cost as well as the possibility of classification errors.

Key words: Breast cancer, Data analysis, Machine learning, Feature selection, Decision support system.

INTRODUCTION

Breast cancer is one of the leading causes of death in the United States and one of the most malignant cancer among women worldwide. In 2015, about 41,000 women died of breast cancer in the U.S. and about 266,000 new cases were diagnosed in 2018.[1,2] Breast cancer is the most common cancer among women around the world and also one of the leading causes of cancer death in women. According to the Globocan 2018 data, the global female breast cancer incidence was about 2.09 million, ranking first between all cancer types among women.[1,2] Meanwhile, there are significant regional differences in the incidence of breast cancer around the world, the incidence rate in developed countries is significantly higher than that in developing

Submission Date: 19-04-2020;
Revision Date: 17-07-2020;
Accepted Date: 13-08-2020

DOI: 10.5530/ijper.54.3s.171

Correspondence:
Mr. Mochen Li
School of Engineering Technology, Purdue University, Grant St. West Lafayette, IN 47907, USA.
Phone: +1-765-4305406
E-mail: li1049@purdue.edu

www.ijper.org

countries. About 1 in 8 U.S. women have the risk of developing breast cancer during her lifetime. From 1999 to 2015, the data showed a continuous increase among the whole country and the diagnosed population was raised to 242,476.[2,3] However, the annual rate of new cases kept going into a relatively stable situation after decreasing for six consecutive years from 1999 to 2005.[3] Breast cancer can be diagnosed and treated before it causes obvious symptoms with regular breast cancer tests. Nowadays, artificial intelligence and machine learning techniques are widely applied for improving cancer detection. In the medical field, statistical machine learning methods have been found to be effective in classifying cancer data.[4-6] Using data volume and computation as the driving force, many machine learning methods, such as a convolutional neural network (CNN) or deep neural network (DNN), have surpassed traditional image recognition performance in case of medical imaging.[4] While the Artificial Neural Network (ANN) models have been found to yield good prediction performance for image data, other state-of-the-art ML models such as Logistic Regression, Support Vector Machine, Decision Trees and Naïve Bayes, have been found to yield good prediction performance for medical data consisting of numerical, ordinal and nominal attributes.[7,8] One of the previous studies using ANN in cancer diagnosis which made use of mammographic findings and demographic characteristics had shown that it can obtain a 0.965 value of area under the curve (AUC) in good accuracy even for a large dataset.[7] But some other research with ANN application in a small lung cancer dataset showed a lower accuracy.[8] Chen Y-C *et al.* presented an accuracy of 83.5% for lung cancer survival prediction with 440 patients' clinical and gene expression data. The key challenge in this research is the thousands of gene-expression data or the high-dimensional dataset. As they mentioned, even after data preprocessing, the obtained data was not good.[8] Many previous studies have used SVM for such applications including: Breast cancer,[4,9-11] Cervical cancer,[12] Oral cancer,[13] or Lung cancer.[14,15] Among those papers, the performances are quite different; in some research,[12] the accuracy was only about 68% with a sample of 168 patients; and in some research,[16] the number of patients is more than 200,000 but the accuracy was as large as 93%.

It is well-known that none of the machine learning models are 100% accurate. Therefore, it is difficult to build the trust of patients and doctors in a completely autonomous system, based on machine learning for a critical decision such as cancer detection. In this study, we examined the prediction performance of ML algorithms and propose a machine learning-based decision support system for medical practitioners for cancer detection, based on diagnostic data. This decision support system combines the prediction output of multiple machine learning algorithms to suggest the most likely category, with an associated confidence level of prediction. Such a system can help the diagnosis expert to spend more time and attention on cases where it cannot predict with high confidence and quickly examine the cases where it predicts with high confidence.

In the data pre-processing, there are actually two concepts related to features: feature selection and feature extraction. Although both can achieve data dimensionality reduction, the two are completely different. Feature selection refers to the process of removing irrelevant features and retaining related features. It can also be considered selecting the best feature subset from all features. Feature extraction refers to the process of converting raw data which cannot be recognized by one machine learning model into features that the algorithm can work. Feature extraction doesn't consider whether these features are useful or meaningful. Because of the importance of specificity in cancer diagnosis the meaning of each feature is important for the prediction model. Some of the previous studies examined only one feature selection method a certain feature selection method to one specific ML algorithm and have not considered determining the optimal subset by synthesizing the results of different feature selection methods.[17-20] From these articles, it can be found that even if it is with the same dataset, different feature selection methods may eliminate different features which means it is very hard to determine the optimal feature subset. According to this, the idea of ensemble learning is applied to the feature selection method in this study.

Several previous studies have compared the prediction performance of different ML models on cancer datasets.[7,18-22] And many ML methods such as SVM, Tree-based methods or Naïve Bayes have proven to be efficient and accurate for prediction;[18-21] but the performance of the prediction model can be improved with the idea of ensemble learning. Therefore, an ML-based decision support system is proposed in this research.

MATERIALS AND METHODS

Machine learning algorithms differ with application and nature, such as supervised learning, unsupervised learning and reinforcement learning. The four

algorithms discussed in this paper are all supervised learning methods that are trained on labeled data and generate a prediction output corresponding to a given value from the input space.[23,24] Typically, every instance from the input space is represented by a feature vector in which each dimension corresponds to a feature. Datasets for supervised learning are composed of pairs of input and output values and split into training data and test data.[23,25] The realization process of one supervised learning method is as follows:

1. Get a limited training dataset and determine the hypothesis space that includes all possible models.
2. Determine the criteria for model selection or assessment of the model.
3. Select the appropriate machine learning algorithm for solving the optimal model.
4. Generate the optimal model by learning datasets.
5. Assess and optimize the prediction model.

The overview of machine learning models used in this study include Logistic Regression, Decision Tree analysis, k-Nearest Neighbor, Naïve Bayes and Support Vector Machines, followed by Glimpse of Data, Feature Selection and ML-based Decision Support System. They are discussed briefly below.

Logistic Regression

The binomial logistic regression is a classical statistical approach to solve the binary classification. The fitted model is represented by the conditional probability distribution P(Y | X) with the logistic function:

$$P(Y=1 \mid x)=\frac{exp(wx)}{1+exp(wx)} \text{ and } P(Y=0 \mid x)=\frac{1}{1+exp(wx)}$$

With $log\frac{P(Y=1|x)}{1-P(Y=1|x)}=wx$, the log odds of can be represented by the linear function of input x. Training errors and test errors are usually used in learning method evaluation. The generated predictive models are often prone to over fitting due to high complexity. Therefore, one of the most common methods is to introduce the regularization or penalty terms into the model. In this study, L1-penalty (Lasso Regression) and L2-penalty (Ridge Regression) are applied.[25] On the other hand, these two methods can be used both as feature selection and predictive model training.

Decision Tree Analysis

The decision tree is a basic classification method and its tree structure represents the process of classifying instances based on features. In fact, this classification is very similar to a collection of an if-then statement. Decision tree learning usually involves three steps: feature selection, decision tree generation and decision tree pruning.[23] Commonly used algorithms for decision tree learning are ID3, C4.5 and CART. A classification decision tree model consists of nodes and directed edges. Nodes in the tree structure have two types: internal nodes that represent a feature or attribute and leaf nodes represent a class.

Feature selection, determining which feature will be used to divide the feature space, is the central choice in the ID3 algorithm and actually is the same detail as discussed later in the Feature Selection section. Here, the information gain is used to measure whether a feature is effective in classification. There is only one difference between two "feature selection": the previous one is data pre-processing for feature elimination and in decision tree all features will be applied in classification. For ease of explanation of information gain, entropy is defined firstly. Entropy is a measure of uncertainty or impurity of random variables. For a collection of random variables, X, the probability distribution is $P(X=x_i)=p_i$ and the entropy of X is $H(X)=-\sum_{i=1}^{n}p_i log p_i$. Given the entropy of a data collection X, the information gain of feature A for X is defined as: $G(X,A)=H(X)-H(X|A)=H(X)-\sum\frac{|X_A|}{|X|}H(X)$.[23,25]

Classification and regression tree model (CART) is a learning method for the conditional probability distribution of the output variable Y given the input variable X.[23,26] CART algorithm consists of the following two steps:

1. Generate a decision tree and make it as large as possible.
2. Pruning the generated tree with the test data to get the optimal subtree.

k-Nearest Neighbor Analysis

The k-Nearest Neighbor algorithm is another basic classification and regression analysis method which assumes that the class of every instance has been labeled given the training data. When classifying, a new instance is classified by the class of its k nearest neighbors training instances. Therefore, the kNN method doesn't have an explicit learning process. It actually divides the feature space according to the class of training data and uses this process as the classification model.

A kNN model is determined by three basic factors: the distance between test data and training data, selection of the value of *k* and classification statement. The feature space in kNN is an n-dimensional real vector space, so the distance between two instance points inside can use Euclidean distance or distance.[23] In this paper, Euclidean distance is applied in kNN analysis. Besides, the selection of *k* value will significantly affect the classification output of the model.

Naïve Bayes Analysis

Naïve Bayes algorithm is a classifier based on Bayes theorem and hypothesis of feature conditional independence. For a given training data set, the model will firstly get the joint probability distribution of input and output based on the hypothesis of conditional independence. Then for a given input instance x, it can classify the output y with the largest posterior probability based on the Bayesian theorem.[23,26] Here, the Naïve Bayes classifier makes an assumption of independence among features for the conditional probability distribution which is based on this assumption, maximizing the posterior probability is also precisely the minimization of the expected loss of model.

SVM Analysis

Support vector machines (SVM) is a binary classifier and the basic model of the SVM method is a linear support vector classifier that determines the biggest "margin" between two classes of training data. When the training data is linearly separable, this support vector classifier can be found through margin maximization. However, when the training data is data is linear non-separable, it's hard to find the hyperplane to satisfy the maximum margin for SVM. In terms of this situation, soft margin and the kernel trick are introduced to learn the nonlinear SVM model.[5] A slack variable ξ_i is introduced for every sample point, thus the objective function adds a new variable: $C\sum_i^N \xi_i$, where C controls margin distance and the number of misclassification points.[23] Kernel function can represent the inner product between the feature vectors which obtained by mapping the input from input space to the Hilbert space. In this paper, kernel trick will be discussed. From the linear separable case, it is known $\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$, then this can rewrite as $f(x) = (\sum_{i=1}^{N} \alpha_i y_i x_i)^T x + \beta_0 = \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + \beta_0$. When this function is mapped to high dimensional space, it can get $f(x) = \sum_{i=1}^{n} \alpha_i y_i \langle \varphi(x_i), \varphi(x) \rangle + \beta_0$. Therefore, kernel function can be defined like this: $K(x, z) = \langle \varphi(x), \varphi(z) \rangle = \varphi(x)^T \varphi(z) = (x^T z)^2$, so it just needs to calculate square of inner product of x and z.

Kernel functions can be different types: radial basis function, Gaussian, sigmoid, polynomial and so on. Table 1 displays some common kernel functions.[20,23,27]

While performing classifications in this research, 2 types of penalties with different ranges are considered: one, C, is from zero to infinity and the other, nu, is only between zero and one.

A Glimpse of Data

The breast cancer dataset used in this paper is created by Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasaria from the University of Wisconsin.[16] The original dataset contains 569 instances with the output value, diagnosis and 10 real-valued features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension.

In addition to the above-mentioned nuclear features, the mean, standard error and worst of these features were computed resulting in 30 features in total. For this paper, only the mean values were considered. A simple cross-validation was introduced: the dataset with 569 cases was divided into the Training set of 400 (70%) cases and remaining (30%) cases as prediction set. Among 400 training data instances, 226 instances were benign (B) and 173 instances are malignant (M). In Figure 1, two tumor types of data distribution within every feature data set is displayed.

From Figure 1, it can be observed that the two tumor types (B and M) can be clearly differentiated for the following 5 features: *radius_mean*, *perimeter_mean*, *area_mean*, *concavity_mean* and *concave points_mean*. This preliminary analysis indicated that through feature selection approaches, we can reduce the number of features present in the input space, as discussed next.

Feature selection

A key problem in machine learning is to decide the features to fit the predictive model. As the amount of

Table 1: Three common Kernel Functions.

	Equation
Polynomial kernel	$k(x_i, x_j) = (x_i \cdot x_j + 1)^d$
Radial basis function (RBF) kernel	$k(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2)$, $\gamma > 0$
Sigmoid kernel	$k(x, y) = \tanh(\alpha x^T y + c)$

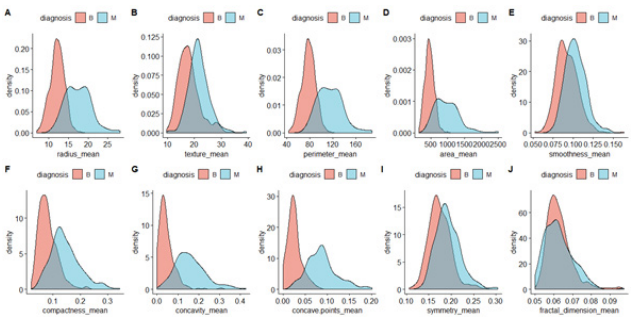


Figure 1: Data Distribution of Tumors.

data acquired grows larger, the contained information also can be easily disturbed by such as noisy data, high dimensional data or intrinsic relationships among the original features. The goal of feature selection is to find the optimal feature subsets and reduce the number of features to improve model accuracy and computation time.[28,29] In this paper, only 10 original features were considered so that the curse of dimensionality is not a concern and digitized images of FNA also are less affected with noisy data. In general, three major methods are applied in feature selection: filter approaches, wrapper approaches and embedded approaches.[30]

Filter approaches set thresholds or rankings to decide the optimal features according to the general characteristics of original features. Typically, it includes lots of statistical methods, such as chi-squared test, correlation coefficient or information gain. A preliminary analysis of features was conducted for our dataset based on the correlation coefficient and the following highly correlated (>0.75) features were removed: *concave points_mean, concavity_mean, perimeter_mean and radius_mean.*

Then, a further chi-squared test of feature independence and information gain were applied in this study, the results for which are displayed in Table 2.

Here, the chi-squared test and information gain will remove two different features separately. Through the above three statistical methods, there are great differences in feature selection.

Wrapper approaches will select a fitted model or algorithm to filter features step by step.[31] According to the predictive effect of the objective function, it selects or excludes some features. Usually, it includes stepwise regression, forward selection or recursive feature elimination (RFE) methods.[27]

Table 2: Chi-squared p-value and Information Gain of 10 Features.

Features	Chi-squared *p*-value	Information gain
Radius	0.0293	0.3058
Texture	0.2506	0.1436
Perimeter	0.1583	0.3539
Area	0.3071	0.3124
Smoothness	0.6158	0.0621
Compactness	0.4123	0.1637
Concavity	0.2444	0.2873
Concave Points	0.2521	0.3748
Symmetry	0.3953	0.0396
Fractal	0.4863	0

In fact, it can apply the predictive model with all sets of possible features combination to train data and select the best subset with the ordinary least squares. However, when the number of features is too large, the computation cost seems to be a burden that has to be considered. Accordingly, stepwise regression can be another choice. The stepwise selection methods with Akaike Information Criterion (AIC) are presented in the paper (Figure 2). The four features with AIC value larger than the predictive model threshold are eliminated: these include, *compactness_mean, perimeter_mean, concavity_mean and fractal_dimension_mean.*

RFE applies a base model to perform multiple rounds of training and eliminates some features after each round. Here two algorithms: linear regression and random forest, were chosen for each iteration to evaluate the model in this study. And the generated optimal feature subsets were actually quite different: for linear-regression-RFE, all 10 features were left; and for random-forest-RFE, 3 features, *compactness_mean, symmetry_mean* and *fractal_dimension_mean*, were eliminated.

Embedded approaches firstly use some machine learning algorithms to train data and obtain the weight coefficients of each feature.[31] Then they select feature subsets according to the coefficients from large too small. It is similar to the filter approaches, but only trained to determine the features through training data instead of statistical methods. Meanwhile, compare to RFE, embedded approaches train data with all original features not through eliminating features each round. The most common methods to select feature subsets are L1 regularization and L2 regularization or Lasso regression or Ridge regression.[23]

Generally, lasso regression and ridge regression are considered as shrinkage methods, but through those shrinkage methods, some features' coefficients can be as small as zero in which those features are also eliminated. Therefore, these two shrinkage methods can be applied in feature selection.[30] Through minimizing

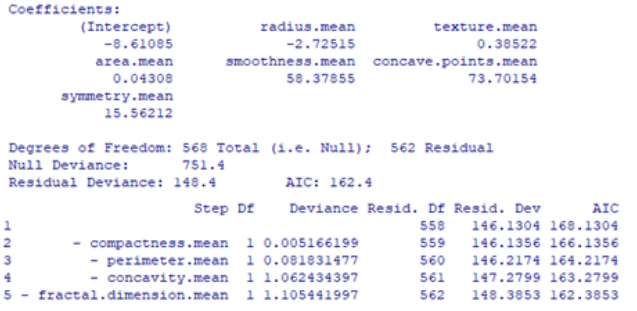


Figure 2: Stepwise selection with AIC.

the cross-validation error of both methods, the optimal feature subsets and the corresponding coefficients can be obtained. The results of those two methods are displayed in Table 3.

Because the Ridge regression removed too many features, it actually reduced the training and test error. In comparison, the Lasso regression provided better performance. As shown in the above results, feature selection results may vary with different approaches. In this study, features were selected according to how often each feature is selected in different feature selection methods. At last, three features were eliminated: *compactness_mean, symmetry_mean, fractal_dimension_mean*.

Decision Support System

The proposed decision support system combines the prediction output from the following five machine learning algorithms used with feature selection: Support Vector Machine with Radial Basis Function Kernel, Lasso Regression, Logistic Regression, Decision Tree (C4.5) and Naïve Bayes. A schematic diagram of the proposed decision support system is provided in Figure 3.

As shown in Figure 3, the decision support system combines the prediction output from multiple ML models to provide a more robust and reliable prediction of tumor category. The prediction output and confidence score of the decision support system are calculated based on following condition:

- If prediction from all 5 models agree:
  - Prediction output = prediction of the 5 models, Prediction confidence = "High"
- If prediction of 4/5 models agree:
  - Prediction output = prediction of 4 agreeing models, Prediction confidence = "Moderate"
- If prediction of less than 4 models agree:
  - Prediction output = "Not sure, need manual review", Prediction confidence = "Low"

The results from the machine learning models and the decision support system are discussed in the next section. For their prediction performance evaluation, following widely-used measures were used:

Sensitivity = True Positive/ Condition Positive

Specificity = True Negative/ Condition Negative

Accuracy = (True Positive + True Negative) / (Condition Positive + Condition Negative)

Table 3: Coefficients of Lasso and Ridge Regression.

Features	Lasso	Ridge
Radius	0	0.25
Texture	0.32	0.108
Perimeter	0	0.002
Area	0.009	0
Smoothness	55.418	0
Compactness	0	0
Concavity	8.044	0
Concave Points	52.247	49.311
Symmetry	12.735	0
Fractal	44.33	0

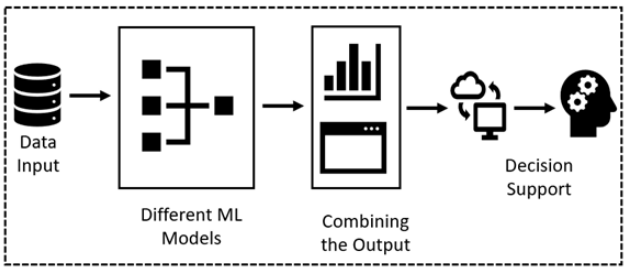


Figure 3: Schematic diagram of machine learning based decision support system for cancer detection.

RESULTS AND DISCUSSION

In order to present a summary of prediction results for all machine learning models, methods' confusion matrices and three key statistical measures of the performance are combined in Table 4. The two values of categorical variable *diagnosis*, *Benign* and *Malignant* are represented as 0 and 1 respectively. To indicate the effect of feature selection, the predictive models with feature elimination are labeled as "*modified*".

The prediction performance of individual models are:

Linear Regression Models

Among all three models in linear regression analysis, compared with the full-feature model, two feature-selected models, Lasso and modified logistic regression, had relatively better performance. However, the two model performances were slightly different: the modified logistic regression model had a higher sensitivity of 92.63% and the Lasso regression model had a higher specificity of 95.56%. According to the purpose of this study is to diagnose breast cancer, the Lasso regression model may be a better choice.

Decision Tree Models

Since the decision tree is learned by recursively selecting the optimal features, the predictive model is built until either all data set can be correctly classified or there are no more features available. As a result, the full-feature decision tree was good as 92.3% in accuracy, 94.57% in sensitivity and 88.24% in specificity. Compared

Table 4: Confusion Matrix and Prediction Performance of Machine Learning Models.

Confusion Matrix				Performance		
Predictive Model	Target Class	Test		Accuracy	Sensitivity	Specificity
		0	1			
Logistic	0	87	4	91.61%	91.58%	91.67%
	1	8	44			
Logistic (Modified)	0	88	3	93.01%	92.63%	93.75%
	1	7	45			
Lasso	0	90	2	93.01%	91.84%	95.56%
	1	8	43			
k-NN	0	86	13	86.71%	93.48%	74.51%
	1	6	38			
k-NN (Modified)	0	85	12	86.71%	92.39%	76.47%
	1	7	39			
Naïve Bayes	0	88	11	89.51%	95.65%	78.43%
	1	4	40			
Naïve Bayes (Modified)	0	88	9	90.91%	95.65%	82.35%
	1	4	42			
Nu-SVM-Polynomial	0	92	28	80.42%	100%	45.10%
	1	0	23			
Nu-SVM-Polynomial (Modified)	0	92	28	80.42%	100%	45.10%
	1	0	23			
Nu-SVM-RBF	0	90	10	91.61%	97.83%	80.39%
	1	2	41			
Nu-SVM-RBF (Modified)	0	91	9	93.01%	98.91%	82.35%
	1	1	42			
C-SVM-Polynomial	0	90	17	86.71%	97.83%	66.67%
	1	2	34			
C-SVM-Polynomial (Modified)	0	92	17	88.11%	100%	66.67%
	1	0	34			
C-SVM-RBF	0	88	7	92.31%	95.65%	86.27%
	1	4	44			
C-SVM-RBF (Modified)	0	90	6	94.41%	97.83%	88.24%
	1	2	45			
Decision Tree (ID3)	0	87	6	92.3%	94.57%	88.24%
	1	5	45			
Decision Tree (ID3 Modified)	0	86	6	91.61%	93.48%	88.24%
	1	6	45			

with the full-feature model the modified model with feature selection, feature selection did not improve the performance but reduced it a little. The decision tree model for CART and C4.5 are presented in Figures 4 and 5 respectively.

k-Nearest Neighbor Analysis.

Among all the machine learning models, the performance of k-NN models was the worst. From the accuracy curve of the k-NN model presented in Figure 6, the effect of different values of k is displayed and the best result was generated with k=5.

As shown in Table 4, feature selection did improve the performance of the k-NN model. However, the overall prediction performance of kNN model was still not satisfactory for the binary classification.

Naïve Bayes

From the analysis of prediction results on test data, the accuracy of the Naïve Bayes model with feature selection

was 90.91%, the sensitivity was 95.65% and specificity was 82.35%. Compared with the model with full features, i.e., without feature selection, the performance was slightly better with Accuracy=89.51%, Sensitivity = 95.65% and Specificity= 78.43%.

SVM Models

The performance of C-SVM and nu-SVM with two kernel functions on test data indicate the following:

1. When the penalty C had a large value range, the SVM model performed better
2. With the kernel function being radial basis function (RBF), the SVM model had a much better performance
3. Feature selection improved the performance of SVM

Based on the result of SVM models, C-SVM with RBF and feature selection had the best accuracy among all predictive models that is 94.41%, but the specificity was not so good, being only 88.24%.

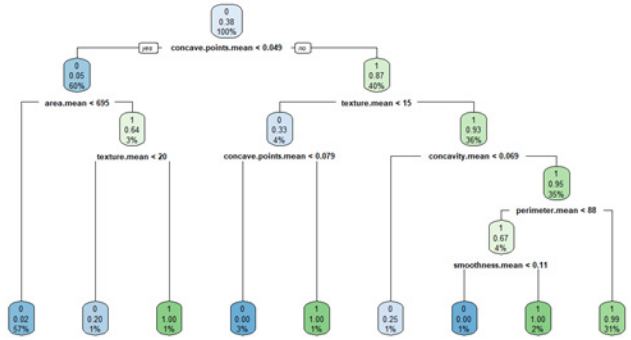


Figure 4: CART Model.

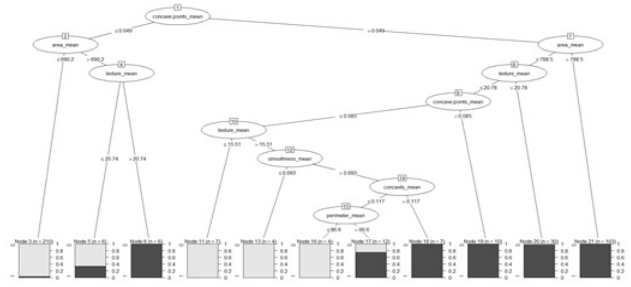


Figure 5: C4.5 Model.

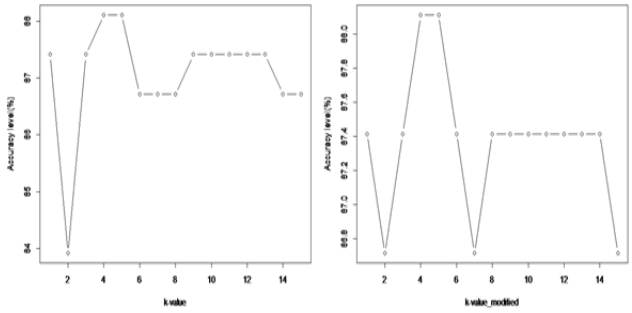


Figure 6: Accuracy Curve with k-values.

Table 5: Results of Decision Support System.

Confusion Matrix				Performance		
Predictive Model	Target Class	Test		Accuracy	Sensitivity	Specificity
		0	1			
5/5 Agree	0	78	2	96.18%	96.30%	96.00%
	1	3	48			
4/5 Agree	0	82	3	94.20%	94.25%	94.12%
	1	5	48			

Decision Support System

For the decision support system, the outputs from the following five machine learning models were combined as an ensemble: C-SVM-RBF (Modified), Lasso Regression, Logistic (Modified), Decision Tree (C4.5) and Naïve Bayes (Modified). The decision support system was designed to output the cases with 'high confidence' where the results from all five models agreed, with 'moderate confidence' when the output from four out of the five models agreed and with 'low confidence' when less than four models agreed. The results for the decision support system are presented in Table 5.

As shown in Table 5, out of the 143 cases in the prediction set, all the five models agreed for 131 cases, i.e. 92% cases. This means that the remaining 12 cases with 'low confidence' of predictions will be filtered for expert review and prediction from the machine learning side will not be provided. Among the 131 cases with 'high confidence' predictions, the Accuracy was 96.2%, Sensitivity was 96.3% and Specificity was 96%, which is higher than any machine learning model and also both measures of performance Sensitivity and Specificity are relatively higher than any individual model.

For the 'moderate confidence' predictions by the decision support system, i.e., the 138 cases out of total 143 cases in prediction set where 4/5 models agreed and the Accuracy was 94.2%, the Sensitivity was 94.3% and the Specificity was 94.1%. Thse values are also higher than any individual machine learning model and are better on both measures of Sensitivity and Specificity. This indicates that the decision support system performs better than an individual model and provides more robust and reliable predictions. It also successfully filters out cases where the predictions from a machine learning model are relatively weaker.

CONCLUSION

This paper has explored 579 cases with 10 features for prediction. Five popular machine learning algorithms: logistic regression, Decision Tree, kNN, SVM and Naïve Bayes classifier, were applied to learn the classification model. Overall, all models performed better on test data than training data which indicates those learned models were not overfitting and are suitable for classification of tumor types. Among various models, SVM with feature selection showed the best accuracy. Feature selection did improve the performance of all machine learning models. This indicates that even for small amounts of data, the data preprocessing is very necessary.

However, each machine learning model had its own limitations. Because of the huge cost of computing in SVM, it will be very hard to apply SVM for large data sets. For our dataset with a small number of cases, SVM took a much longer time than the other three models in learning data. Sometimes the selection of kernel function in SVM also affects the performance of prediction. While kNN is easy to understand, but it is sensitive to outliers of the dataset. For the reason that kNN presents the worst performance in this research. By comparison of prediction performance, the Naïve Bayes method was slightly worse than Decision Tree and SVM, but it is highly efficient and easy to implement.

The decision support system that combines the prediction from multiple machine learning model performed better than any individual machine learning model and can effectively filter out cases where the machine learning models may not be accurate, indicating that medical professionals may need to examine those cases more carefully as they may not be so straightforward. The cases where multiple machine learning models make a consistent prediction may be more clearly belonging to one category and can be examined quickly by medical professionals.

In this study, the size of a dataset is a limitation for the machine learning algorithm. In fact, when the amount of cases increases, the machine learning model may generate significant differences. In future work, as the size of the dataset is enlarged, many hiding factors such as computational cost, processing time and data preprocessing can also be fully considered. The behavior of the decision support system can also be tested with more data, whether it still performs considerably better than individual machine learning models or with large training dataset, all the machine learning models perform similarly.

CONFLICT OF INTEREST

Authors have no conflict of interest with the content and publication of this article.

ABBREVIATIONS

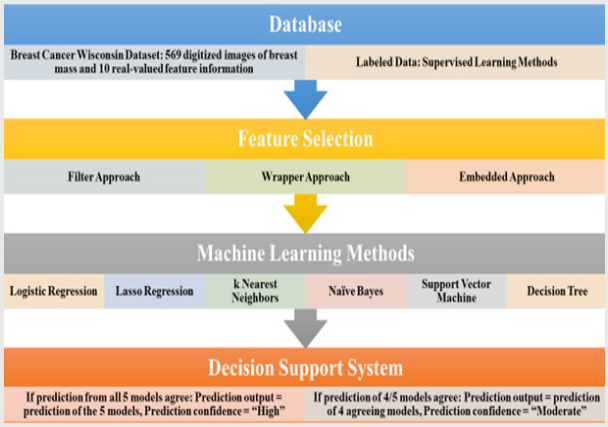
ML: Machine Learning; **kNN:** k-Nearest Neighbors; **SVM:** Support Vector Machine; **ANN:** Artificial Neural Network; **CART:** Classification And Regression Tree; **RFE:** Recursive Feature Elimination; **AIC:** Akaike Information Criterion; **RBF:** Radial Basis Function.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics. Cancer J Clin. 2018;68(1):7-30.
2. Siegel R, Miller K, Jemal A. Cancer statistics. Cancer J Clin. 2015;65(1):29.
3. Breast Cancer Statistics. Centers for Disease Control and Prevention. 2019 Available: https://www.cdc.gov/cancer/breast/statistics/index.htm.
4. Lg A, At E. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. J Heal Med Informatics. 2013;4(2):2-4.
5. Asri H, Mousannif H, AlMoatassime H, Noel T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Comput Sci. 2016;83:1064-9.
6. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif Intell Med. 2005;34(2):113-27.
7. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration. Cancer. 2010;116(14):3310-21.
8. Chen YC, Ke WC, Chiu HW. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med. 2014;48(1):1-7.
9. Listgarten J, *et al.* Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. Clin Cancer Res. 2004;10(8):2725-37.
10. Kim W, *et al.* Development of Novel Breast Cancer Recurrence Prediction Model using Support Vector Machine. J Breast Cancer. 2012;15(2):230-8.
11. Xu X, Zhang Y, Zou L, Wang M, Li A. A gene signature for breast cancer prognosis using support vector machine. 2012 5th. Int Conf Biomed Eng Informatics. BMEI. 2012; 928-31.
12. Tseng CJ, Lu CJ, Chang CC, Den CG. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Comput Appl. 2014;24(6):1311-6.
13. Chang SW, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinformatics. 2013;14(1):170.
14. Sun T, Zhang R, Wang J, Li X, Guo X. Computer-Aided Diagnosis for Early-Stage Lung Cancer Based on Longitudinal and Balanced Data. PLoS One. 2013;8(5):e63559.
15. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci Rep. 2015;5:1-11.
16. UCI. Breast Cancer Wisconsin (Diagnostic) Data Set. 2019 Available: http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29.
17. Karthik S, Srinivasa PR, Chandra M. Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods. Knowl Comput its Appl Knowl Manip Process Tech. 2018;2(3):227-41.
18. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. J Algorithms Comput Technol. 2018;12(2):119-26.

19. Salama G, Abdelhalim MB, Zeid MA. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifications. Int J Comput Inf Technol. 2012;1(1):36-43.
20. Mu T, Nandi AK. Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier. J Franklin Inst. 2007;344(3-4):285-311.
21. Azmi MSBM, Cob ZC. Breast cancer prediction based on backpropagation algorithm," Proceeding, 2010 IEEE Student Conference on Research and Development. Eng Innov Beyond: SCOReD. 2010;164-8.
22. Aruna S, Rajagopalan SP, Nandakishore LV. In, Knowledge Based Analysis of Various Statistical Tools in Detecting Breast Cancer. Computer Science & Information Technology. 2011;2(2011):37-45.
23. Hastie T. The Elements of Statistical Learning Data Mining, Inference and Prediction, Second Edition. 2nd ed. Springer New York: Imprint: Springer. 2009.
24. Mitchell TM. Machine Learning, 1st ed. New York, NY, USA: McGraw-Hill, Inc. 1997.
25. Casella G, Fienberg S, Olkin I. An Introduction to Statistical Learning. 2013.
26. Al-Khasawneh A. Diagnosis of breast cancer using intelligent information systems techniques. Nature-Inspired Comput. Concepts, Methodol. Tools, Appl. 2016;1-3(3):203-14.
27. Liu Q, Chen C, Zhang Y, Hu Z Feature selection for support vector machines with RBF kernel. Artif Intell Rev. 2011;36(2):99-115.
28. Mu T, Nandi AK. Breast cancer diagnosis from fine-needle aspiration using supervised compact hyperspheres and establishment of confidence of malignancy. Eur Signal Process Conf Eusipco. 2008;1-5.
29. Wolberg WH, Street WN, Mangasarian OL. Machine learning techniques to diagnose breast cancer from fine needle aspirates. Cancer Lett. 1994;77(2-3):163-71.
30. Wang S, Wang KY, Zheng L. Feature selection via analysis of relevance and redundancy. J Beijing Inst Technol. 2008;17(3):300-4.
31. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507-17.

PICTORIAL ABSTRACT



SUMMARY

- Among various models, SVM with feature selection showed the best accuracy (94.41%)
- Feature selection did improve the performance of all machine learning models
- Limitation for the ML algorithms:
- Because of the huge cost of computing in SVM, it will be very hard to apply SVM for large data sets
- kNN is easy to understand, but it is sensitive to outliers of dataset
- Naïve Bayes method was slightly worse than Decision Tree and SVM, but it is highly efficient and easy to implement
- The size of data set is a limitation for machine learning algorithm
- The decision support system that combines the prediction from multiple machine learning model performed better than any individual machine learning model.

About Authors

Mochen Li, a Ph.D. candidate at the School of Engineering Technology, Purdue University. He received an M.S. from Purdue University and a B.E. from China University of Mining & Technology-Beijing. His research area focuses on applications of machine learning in cancer diagnosis and prevention.

Dr. Gaurav Nanda is an Assistant Professor of Practice in the School of Engineering Technology at Purdue University. His research interests include applications of machine learning, text mining, and intelligent decision support systems in the areas of healthcare, safety, and education. He obtained his Ph.D. in Industrial Engineering from Purdue University and his Bachelors (B.Tech.) and Masters (M.Tech). from Indian Institute of Technology (IIT) Kharagpur. He has worked as a postdoctoral researcher at Purdue University for two years and in the software industry for five years.

Dr. Santosh S. Chhajed is working as Associate Professor at Mumbai Education Trust's Institute of Pharmacy, Nashik. Dr. Santosh Chhajed is having teaching experience of 13 years of various Pharmaceutical Chemistry subjects at undergraduate as well postgraduate level. Dr. Chhajed has to his credit 08 books on pharmacy and has published 40 papers in national and international journal of repute. Research area is Drug Discovery. Design and synthesis of small molecule heterocycles. Analytical method development.

Raji Sundararajan, Ph.D is a Professor at the School of Engineering Technology Department, Purdue University. Her research interests include cancer treatment using Electrochemotherapy and studying machine learning-based breast cancer analyses. She is a reviewer of NIH, NSF, US International Science & Technology Center, and US National Research Council proposals, and various scholarly journals including International Journal of Cancer, Molecular Biotechnology, Journal of Biomedical Microdevices, Journal of Anticancer drugs, and several IEEE.

Cite this article: Li M, Nanda G, Chhajedss S, Sundararajan R. Machine Learning-Based Decision Support System for Early Detection of Breast Cancer. Indian J of Pharmaceutical Education and Research. 2020;54(3s):s705-s715.