

Article - Engineering, Technology and Techniques

Breast Cancer Prediction Using Dominance-based Feature Filtering Approach: A Comparative Investigation in Machine Learning Archetype

Kushangi Atrey¹

<https://orcid.org/0000-0001-6767-8447>

Yogesh Sharma¹

<https://orcid.org/0000-0002-7343-0373>

Narendra K. Bodhey²

<https://orcid.org/0000-0003-1736-3329>

Bikesh Kumar Singh^{1*}

<https://orcid.org/0000-0002-5052-9768>

¹National Institute of Technology Raipur; Biomedical Engineering Department, Raipur, CG, India. ²All India Institute of Medical Sciences Raipur, Radiodiagnosis Department, Raipur, CG, India.

Received: 2018.09.07; Accepted: 2019.07.08.

*Correspondence: bsingh.bme@nitrr.ac.in; Tel.: +91-9826469522(B.K.S.)

HIGHLIGHTS

- Dominance-based filtering approach is proposed for breast cancer detection.
- Extensive comparative study for breast cancer classification is conducted.
- High accuracy of 99.6% using only 5 dominant features is reported.
- Error between all nine and top 4 and 5 dominant features is less than 2%.

Abstract: Breast cancer is the most commonly witnessed cancer amongst women around the world. Computer aided diagnosis (CAD) have been playing a significant role in early detection of breast tumors hence to curb the overall mortality rate. This work presents an enhanced empirical study of impact of dominance-based filtering approach on performances of various state-of-the-art classifiers. The feature dominance level is proportional to the difference in means of benign and malignant tumors. The experiments were done on original Wisconsin Breast Cancer Dataset (WBCD) with total nine features. It is found that the classifiers' performances for top 4 and top 5 dominant-based features are almost equivalent to performances for all nine features. Artificial neural network (ANN) is come forth as the best performing classifier among all with accuracies of 98.9% and 99.6% for top 4 and top 5 dominant features respectively. The error rate of ANN between all nine

and top 4 & 5 dominant features is less than 2% for four performance evaluation parameters namely sensitivity, specificity, accuracy and AUC. Thus, it can be stated that the dominance-based filtering approach is appropriate for selecting a sound set of features from the feature pool, consequently, helps to reduce computation time with no deterioration in classifier's performance.

Keywords: Breast cancer; Computer aided diagnosis; Dominance-based filtering; Machine learning.

INTRODUCTION

Breast cancer is the most common type of cancer among females worldwide. The mortality-incidence observed to be more in rural and suburban areas. Physicians use common medical techniques like mammography, ultrasound imaging, MRI, biopsy, fine needle aspiration cytology etc. for diagnosis of breast cancer. The diagnosis efficiency is reduced by various man-made errors like inter-operation and intra-operation differences, fatigue, experience of the doctor. Hence, to correctly diagnose a fatal disease including breast cancer, a trustworthy second opinion is of utmost requirement. The evolution of computer aided diagnosis (CAD) with different machine learning units is a promising approach which can assist a radiologist to determine as well as predict the prognosis of cancerous cells in patient's body.

A lot of researches have already been carried out in the field of CAD with machine learning. Most of the classification techniques reported are supervised in which a predefined label is there to check the expert system's accuracy. Wisconsin breast cancer database (WBCD) from UCI machine learning repository is a standard dataset, used as a part of various investigations to approve the classifier's execution [1]. In [2], authors utilized supervised fuzzy clustering algorithm with 10-fold cross validation protocol and obtained an accuracy of 95.57%. In [3], authors proposed a genetic algorithm (GA) based approach to form different rules for diagnosing breast cancer. They secured better average classification accuracy (96.995%) compared with PolyAnalyst®, a commercially available tool for data mining. In [4], least square support vector machines (LS-SVM) was employed and an accuracy of 98.53% was achieved with 10-fold cross validation protocol. A new hybrid model of fuzzy artificial immune system with k-nearest neighbor presented by [5] for breast cancer diagnosis and for 10-fold cross validation they got classification accuracy of 99.14%.

In [6], authors mined breast cancer patterns with a two phases approaches, regression analysis to eliminate insignificant features and discrete particle swarm optimization (PSO) to discriminate the sample between benign and malignant. They obtained an accuracy of 98.71%. Two different classification techniques were compared by [7] and they concluded that radial basis function neural network (RBFNN) classifier's performance is better than support vector machine (SVM) with polynomial order 2 for breast cancer classification. They found an average classification accuracy of 96.56% for RBFNN and 92.13% for SVM with 3-fold cross validation repeated for three times. In [8], a rough-set based feature selection approach was utilized with SVM as classifier and classification accuracies of 99.41% (50-50% training-test partition), 100% (70-30% and 80-20% training-test partition) were achieved. In [9], authors applied artificial meta-plasticity multilayer perceptron (AMMLP) algorithm in WBCD with 60-40% training-testing holdout protocol and got an accuracy of 99.26%. A new kernel approach for SVM known by combining Gaussian RBF with polynomial kernel was introduced and an accuracy of 99.88% was obtained [10].

Artificial neural networks (ANN) and SVM performances for micro-calcification clusters (MCC) in mammogram were studied by [11]. In [12], authors proposed a hybrid feature selection method comprises of association rules with principal component analysis (PCA) to detect breast cancer and obtained an accuracy of 99.29% for ANN classifier when 10-fold cross validation protocol was used. An intelligent hybrid clustering technique for feature selection using a comparison of coincident quantitative measurements was presented for the case when the true class labels of dataset are absent [13]. This approach gave an idea

of unsupervised learning and was proved to be better than many benchmark filter and wrapper based methods. In [14], authors have applied K-means algorithm for feature selection and later the abstracted features were used for breast cancer classification by SVM classifier. They achieved 97.38% classification efficiency for 10-fold cross validation protocol. Authors from [15], implemented a new, Genetically Optimized Neural Network (GONN) model for classification problems and with WBCD database the best average accuracy of 99.26% was obtained for 10- fold cross validation scheme. Recently in [16], Rotation Forest with GA as feature selection scheme was presented. For 14 significant features of Wisconsin diagnosis database they accomplished maximum efficiency of 99.48%. Also, in [17], authors modeled a novel optimized Gauss-Newton representation based algorithm (GNRBA). They computed optimal weighting coefficients of the significant training to reduce computer complexity. They got maximum accuracy of 100% for 70-30% training-test partition.

In summary, authors have utilized schemes like GA, PCA, regression analysis, rough-set technique, K-means algorithm and recently, rotation forest with GA for the optimal feature selection task. For classification, they applied fuzzy clustering algorithm, SVM with kernels, k-nearest neighbor, discrete PSO, GA, hybrid clustering technique and neural networks approaches. The highest obtained classification accuracy is 100%. However, the studies which report 100% accuracies have used holdout data division protocol to evaluate the classifier. The holdout data division protocol is not widely accepted by research community because it does not give a better approximation of generalization. Though simple to implement, statistically it is not much powerful. Though literature has provided enough information about the existing and state-of-the-art technologies on individual basis for breast cancer detection, we feel the need of applying a less complex feature selection mechanism for this work. Subsequently, in conjunction with the feature filtering approach, summing up and comparing the performances of some of well-known CAD tools on the basis of various performance measures would give a clear picture to the future researchers in this field.

In present work, detailed empirical study of major state-of-the-art classifiers namely naïve Bayes, ANN, linear SVM, quadratic SVM, cubic SVM, SVM with Gaussian radial basis function (RBF) (including coarse, medium and fine Gaussian schemes) were carried out for repeated holdout (RH), 5-fold cross validation and 10-fold cross validation data division protocols on WBCD (original) dataset. The dominance-based filtering technique was utilized to abstract the most imperative feature sets to lessen the computer complexity of CAD model [18]. It is hypothesized that by incorporating dominance-based filtering approach, the performances of state-of-art classifiers would improve for breast cancer prediction task.

Our findings indicate that the dominance-based filtering approach is a simple and successful practice for classification tasks and with fewer selected attributes it gives almost the identical performance to that of expert systems with all feature variables, hence, it supports our hypothesis. The rest of the paper is organized as follows: Section 2 presents materials and methods used in this work, Section 3 presents results and discussions and Section 4 presents conclusions and scope for future work.

MATERIAL AND METHODS

Data and Data Processing

The Wisconsin Breast Cancer (Original) cytological database (WBCD) has been acquired from UCI machine learning repository [1],[19]. To access the data and dataset information, one can visit the link given in [20]. The database consists of Fine Needle Aspirates of 699 subjects. The database has a total of 11 fields, the first column represents the sample code number i.e. patient's ID, column 2 to 9 are the different cytological attributes. Last column is for showing the true class of tumor (2- for benign, 4-for malignant) which is used as ground truth to predict the correct class of the sample. Each feature was measured in a scale of 1 to 10 where value 1 is closest to benign and 10 closest to malignant. The details of the 9 cytological attributes are shown in Table 1. As the cytological database

was acquired from UCI machine learning repository which is open access for research purposes, ethical approval is not needed for the present work.

Table 1. WBCD attributes.

Label	Attribute Name	Domain
x1	Clump Thickness	1-10
x2	Uniformity of Cell Size	1-10
x3	Uniformity of Cell Shape	1-10
x4	Marginal Adhesion	1-10
x5	Single Epithelial Cell Size	1-10
x6	Bare Nuclei	1-10
x7	Bland Chromatin	1-10
x8	Normal Nucleoli	1-10
x9	Mitoses	1-10

Normalization of the data is carried out to standardize the range of the features which contribute independently to classify the test sample. It is done to reduce the dominance effect of a particular feature in the classifier's performance [21]. Especially the use of neural network models for faster and stable response requires the input data in normalized form [22]. Here, we have used linear transformation technique to normalize all the features in between zero and one. Equation (1) shows the adapted method for normalization. Let x_i and \hat{x}_i are the actual and the scaled values of the particular feature, then

$$\hat{x}_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

Where $\min(x_i)$ and $\max(x_i)$ are lower and upper boundaries of feature x_i .

Statistical Significance Analysis

Statistical significance analysis is a method to deduce the most significant and critical attributes among the large pool of features used for determining the correct class of a tumor. It's well-known advantage is in machine learning paradigm where large feature sets are reduced to small and potentially significant space which could greatly advance the computation proficiency [23]. As the data has been normalized, therefore parametric t-test is performed here to check the significant level of feature variables. From the result of t-test, p-values are obtained for all features and if $p < .05$ (95% confidence interval), then the corresponding feature attribute is treated as the significant feature [24]. In present work, authors have performed independent t-test using SPSS to get the p-value, mean and standard deviation of all nine features for benign and malignant samples of WBCD dataset.

Feature Selection

The feature selection techniques, in general, do not alter the originality of the variables and only select the subset from their pool. The two very common approaches are filter and wrapper methods. Filter methods assign a score or rank to all the features in the feature space. The features here are considered invariantly by the methods and also it depends solely on the data. Filter methods are advantageous as these are relatively simple to execute and require less computational time but since it does not incorporate any inductive classification algorithms, it may be failed to choose the most useful features [25]. Filter methods include Chi square test, information gain, correlation coefficient scores etc. Wrapper methods prepare different combinations or subsets of the feature variables. A predictive model is inducted to compare the performance accuracy of different subsets and based on that scores are assigned. Since cross validation is used in wrapper methods so there is a high probability to select the most useful feature set among all. The disadvantage

is high computational cost, so if features are very large, it will consume a significant time to evaluate the best set of features [26]. Example of wrapper method is recursive feature elimination algorithm.

The dominance-based filtering method was incorporated in this work to recruit the appropriate features for classification. The advantage of this method is its simplicity and less time requirement. Dominance level of individual feature is proportional to the difference in mean value of each attribute for both the classes i.e. benign and malignant [18]. The features are arranged in sets according to their dominance rank in descending order.

Classification

Classification is generally the final stage of any CAD system which takes the decision whether the given sample is cancerous or not according to its training parameters. The classifier maps input feature vectors $\mathbf{x} \in \mathbf{X}$ to output class labels $\mathbf{y} \in \{1, \dots, m\}$, where \mathbf{X} is the feature space and m is total number of class. Two types of learning approaches are generally followed: supervised and unsupervised. When the training inputs are given to the classifier along with its target labels then this is known as supervised approach. The test data is mapped with its label based on the trained classifier parameters. In unsupervised case the target class is not known and in the training phase the classifier makes a group or cluster of different samples of identical characteristics which is treated as the class and the test samples are categorized with this previous knowledge [27].

In present study, supervised learning schemes are incorporated and results demonstrate the performances of naïve Bayes, support vector machines (SVMs), and back-propagation artificial neural network (BPANN) classifiers. These schemes are briefly discussed in following sections. All the experiments have been performed using MATLAB[®] software.

Naïve Bayes Classifier

Naïve Bayes classifier is a kind of statistical classifier which is based on Bayes' theorem. It predicts class membership probabilities that the particular sample belongs to a particular class by assuming that every attribute's value contributes independently to the given class in supervised scheme. The Bayes theorem of probability can be given as in equation (2)

$$P(A/B) = \frac{P(B/A) P(A)}{P(B)} \quad (2)$$

Where $P(A/B)$ is the posterior probability, $P(A)$ and $P(B)$ are probabilities of the occurrence of event A and B respectively and $P(B/A)$ is the likelihood.

The naïve Bayes classifier assigns the test sample $\mathbf{x} \in \mathbf{X}$ to the particular class \mathbf{y}_i as [27]:

$$y_i = \operatorname{argmax} \prod_{k=1}^l p(\mathbf{x}_k | y_j), j = 1, 2, \dots, m \quad (3)$$

Interestingly, because of some unknown reasons, the naive Bayes classifier can be very effective even when its independence assumption is violated[28].

Support Vector Machine

Support vector machines are the state-of-the-art techniques extensively used for classification and regression tasks with excellent outcomes. Authors even have incorporated SVM for the human identification by combining face, iris and finger features which could be used for various security applications [29]. It belongs to supervised learning model based on statistical theory of learning[30]. SVM separates two classes of a sample, for this, a hyperplane is formed which distinguish class members from non-members. The example is shown in Figure 1.

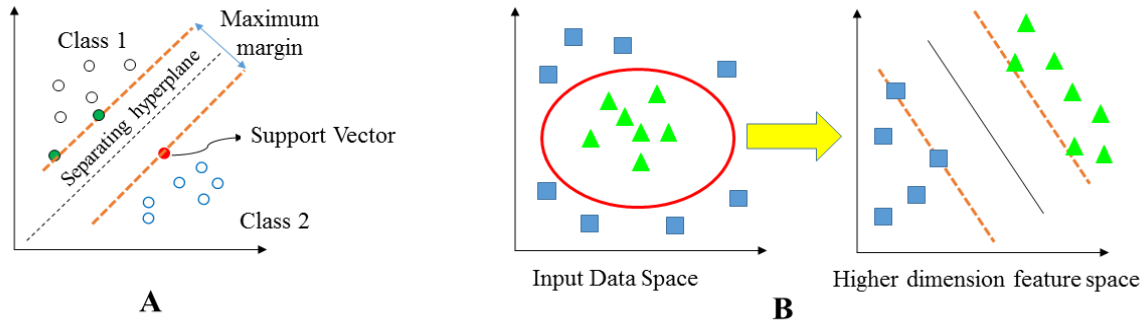


Figure 1. Support vector machine classification schemes **(A)** Linear **(B)** Non-linear.

There may be so many separating hyperplanes but the one with maximum separation from data points of both the classes are chosen [31]. The data points from both the classes which fall on margin hyperplanes (orange dashed lines) formed in parallel with separating hyperplane are called Support Vectors (SVs). For the maximum margin between the SVs of two classes the parallel hyperplanes which are shown in Figure 1(A) can be described by following equations (vectors are shown in bold letters)[32]:

$$\mathbf{w}\mathbf{x} + b = 1 \quad (4)$$

$$\mathbf{w}\mathbf{x} + b = -1 \quad (5)$$

Where, \mathbf{w} is the coefficient vector of the hyperplane, $\mathbf{x} \in \mathbb{R}^n$ is the data vector and b is the offset.

The maximum margin can be achieved by maximizing the distance between two hyperplanes i.e. $2/(\|\mathbf{w}\|)$ or by minimizing $\frac{1}{2}\|\mathbf{w}\|^2$. Hence the perfect classification is done with this if no points come in between the two parallel hyperplanes. The required constraint may be given as [7]:

$$\mathbf{w}\mathbf{x} + b \geq 1 \quad (6)$$

$$\mathbf{w}\mathbf{x} + b \leq -1 \quad (7)$$

By combining equation (7) and (8),

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1, \forall i \quad (8)$$

Where, $y_i \in \{-1, 1\}$.

Not all the classification problems are linearly separable, so to construct non-linear classifiers for categorizing the data points spread in a haphazard manner, SVM with kernel tricks were introduced [33]. The input data space is mapped into higher dimensional feature space denoted by: $\phi: \mathbf{X} \rightarrow \phi(\mathbf{X})$. The kernel function is nothing but the inner product of the data variables in feature space such that $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$, this is shown in Figure 1 (B). The introduction of non-linear classifiers using kernel functions made data points linearly separable at higher dimension feature space which are inseparable at input space by a linear classifier [32]. Some popular kernel functions are linear, polynomial, Gaussian radial basis function (RBF) and Sigmoid function.

Artificial Neural Network

An Artificial Neural Network (ANN) is a biologically motivated computational mechanism which is employed for a variety of applications such as pattern reading, prediction of certain outputs (stock market prediction), clustering and optimization problems. One of the very popular models of neural network is Back-Propagation artificial neural network (BPANN) which has a number of advantages [34]. It has also been employed for malignant lung nodule detection system [35], hence found application in clinical research. The structure consists of an input layer, one hidden layer and an output layer as shown in Figure 2.

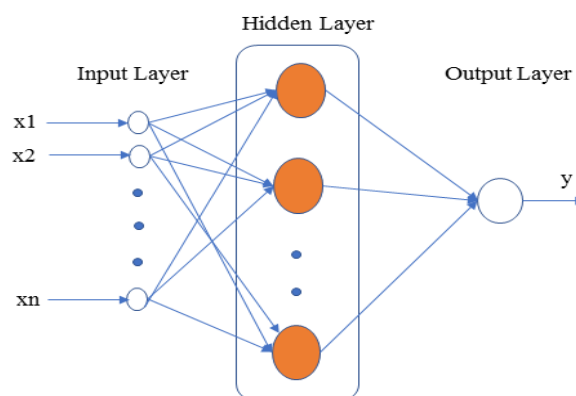


Figure 2. BP-ANN architecture.

The output of each neuron of neural network depends upon the input value, bias value, link weights and the activation function or transfer function of each neuron. Here, we have used hyperbolic tangent segment transfer function at the output of hidden layer as well as the output layer respectively. The default performance goal and maximum epochs were decided arbitrarily as 0 and 100 respectively. The initial learning rate is set to 0.01.

The overall process done in this work is illustrated in Figure 3. At first, the mean and standard deviation of individual feature of WBCD dataset for benign and malignant class is calculated through SPSS software. Feature combinations (feature sets) are then prepared according to the rank of features by applying dominance-based filtering approach. Each feature set is divided into training-test combination via different data division protocols (repeated holdout (RH), 5-fold, 10-fold). Finally, the performances of classifiers would be judged for the different subsets and effectiveness of dominance-based filtering is monitored.

PERFORMANCE EVALUATION PARAMETERS

Machine learning algorithms may have a tendency to attain different outcomes for different trials, resulting reliability concern. Therefore, to avoid any ambiguous inference, each experiment is repeated for 10 times to compute the mean results. A lot of benchmarking parameters are used to prove the worthiness of the classifier. Some of them which have been used in present work are listed below:

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{Sensitivity (Se)} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{Specificity (Sp)} = \frac{TN}{TN + FP} \quad (11)$$

$$\text{Area Under ROC Curve (AUC)} = \frac{Se + Sp}{2} \quad (12)$$

Where TP = True positives (Correctly classified as malignant), TN = True negatives (Correctly classified as benign), FP = False positives (Sample is benign but misclassified as malignant), FN = False negatives (Sample is malignant but misclassified as benign).

The performance parameters are calculated from "Confusion Matrix" which contains the information about actual and classifier's predicted class. The format is shown in Table 2.

Table 2. Confusion Matrix.

Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Youden's Index

Youden's Index (γ) [7], [36] denotes how good the algorithm is in avoiding false results.

$$\gamma = \text{Sensitivity} + \text{specificity} - 1 \quad (13)$$

The closeness of Youden's Index value with unity is directly proportional to the algorithm's ability to avoid failure.

Discriminatory Power

The distinguishability of an algorithm for positive and negative samples are calculated through discriminatory power value (DP) [37].

$$D. P. = \frac{\sqrt{3}}{\pi} (I_n A + I_n B) \quad (14)$$

Where,

$$A = Se/(1 - Se), B = Sp/(1 - Sp) \quad (15)$$

RESULTS AND DISCUSSIONS

The experiments were carried out to demonstrate the classification capabilities of naïve Bayes, ANN and SVM classifiers on WBCD (original) dataset consists of 699 subjects. There are 16 subjects with missing values of one of the attributes which were replaced by mean value of that attribute column. The independent t-test analysis result is shown in Table 3. It is observed that feature #6 (Bare nuclei) is having the highest variance between mean values of benign and malignant classes. After applying the dominance-based filtering approach the feature vectors were arranged in sets in descending order according to their dominance rank. The arrangement is displayed in Table 4.

The performances of naïve Bayes, ANN and SVM classifiers for three validation protocols were calculated along with their ROC plots as shown in Figure 4. The feature set #1 (most dominant feature "Bare nuclei") attained the highest classification accuracy of 91.4% for naïve Bayes and ANN classifiers with RH protocol. The AUC is highest (0.901) for ANN and for naïve Bayes it is 0.896. Their ROC plot is shown in Figure 4 **(A)** and **(B)**.

Evaluating feature sets #4 and #5 (top 4 & 5 dominant feature sets), the ANN clearly outperformed all other classifiers with highest classification accuracies 98.9% and 99.6% respectively for 10-fold cross validation protocol. The AUC values for the above cases are 0.986 and 0.995 respectively. However, for feature set #5 the quadratic SVM classifier obtained the highest classification accuracy and AUC namely 97% and 0.967 if subjected to RH validation protocol. Tables 5 and Table 6 display the results in detail along with the values of sensitivities and specificities.

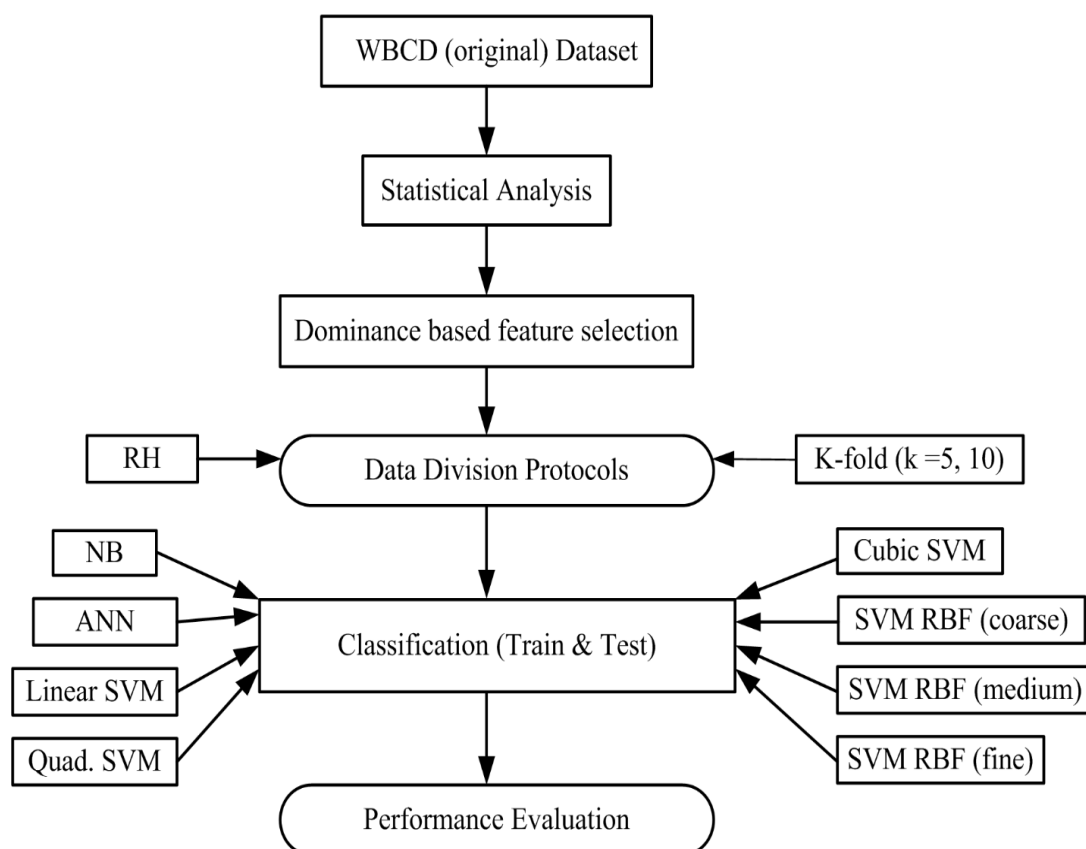


Figure 3. Flow diagram of overall procedure.

Table 7 presents the effectiveness of the classifiers for feature set #9 which consist of all nine features. Though all the classifiers performed well in this case, the highest value of accuracy and AUC were achieved by ANN equal to 99.7% and 0.996 for 10-fold cross validation protocol.

The ROC curves of feature subsets #4, #5 and #9 for best performing classifier (ANN) are given in Figure 4 (C), (D), (E). Figure 4 (F), displays the ROC for feature set #5 when subjected to quadratic SVM for RH validation protocol. As it is observed from different tables (Table 5 to Table 7), the ANN classifier with 10-fold cross validation protocol possesses the highest values of accuracy and AUC (Both > 0.95).

Therefore, it may be affirmed that ANN gives the best classifier model for WBCD database when top 4 and top 5 features based on their dominance rank and otherwise all 9 features are considered. Also, the average value of accuracies of all feature sets (feature set #1 to #9) followed incremental order (95.9%, 96.8% and 97.7%) for ANN when the validation protocol shifted from RH to 5-fold to 10-fold whereas for all other classifiers it was almost constant (around 95.5%). Accuracies of ANN classifier for three data division protocols (RH, 5-fold, 10-fold) is depicted in Figure 5.

Though, ANN's performance increases with increase in number of feature vectors in scope, the results also show that there is very minute error difference (< 2%) between the performance of classifier for overall nine features and for top 4 or top 5 features obtained from dominance-based filtering approach. Table 8 shows the comparative statistics along with error values for feature sets #4, #5 and #9 for ANN classifier.

Table 3. Mean and standard deviation obtained from independent t-test.

Attribute	Group Statistics					Difference between means of two classes
	Benign = 0 Malignant = 1	N	Mean	Std. Deviation	Std. Error Mean	
V1	0	458	0.296	0.1674	0.0078	0.424
	1	241	0.72	0.2429	0.0156	
V2	0	458	0.133	0.0908	0.0042	0.524
	1	241	0.657	0.272	0.0175	
V3	0	458	0.144	0.0998	0.0047	0.512
	1	241	0.656	0.2562	0.0165	
V4	0	458	0.136	0.0997	0.0047	0.419
	1	241	0.555	0.321	0.0207	
V5	0	458	0.212	0.0917	0.0043	0.318
	1	241	0.53	0.2452	0.0158	
V6	0	458	0.1413	0.12176	0.00569	0.618
	1	241	0.7593	0.31263	0.02014	
V7	0	458	0.21	0.108	0.005	0.388
	1	241	0.598	0.2274	0.0146	
V8	0	458	0.129	0.1059	0.0049	0.457
	1	241	0.586	0.3351	0.0216	
V9	0	458	0.106	0.0502	0.0023	0.153
	1	241	0.259	0.2558	0.0165	

Table 4. Feature set number and corresponding feature(s).

Set number	Feature set
1	{Bare Nuclei}
2	{Bare Nuclei, Uniformity of cell size}
3	{Bare Nuclei, Uniformity of cell size, Uniformity of Cell Shape}
4	{Bare Nuclei, Uniformity of Cell Size, Uniformity of Cell Shape, Normal Nucleoli}
5	{Bare Nuclei, Uniformity of Cell Size, Uniformity of Cell Shape, Normal Nucleoli, Clump Thickness}
6	{Bare Nuclei, Uniformity of Cell Size, Uniformity of Cell Shape, Normal Nucleoli, Clump Thickness, Marginal Adhesion}
7	{Bare Nuclei, Uniformity of Cell Size, Uniformity of Cell Shape, Normal Nucleoli, Clump Thickness, Marginal Adhesion, Bland Chromatin}
8	{Bare Nuclei, Uniformity of Cell Size, Uniformity of Cell Shape, Normal Nucleoli, Clump Thickness, Marginal Adhesion, Bland Chromatin, Single Epithelial Cell Size}
9	{Bare Nuclei, Uniformity of Cell Size, Uniformity of Cell Shape, Normal Nucleoli, Clump Thickness, Marginal Adhesion, Bland Chromatin, Single Epithelial Cell Size, Mitoses}

Thus, the top 4 or top 5 dominance-based features are effectively able to classify the WBCD sample barring the extraction and summation of all the feature values. Table 9 displays the values of γ and DP to justify the assertion and validate the effectiveness of ANN classifier. A comparison of present work's outcome with some of literatures in terms of classification accuracy for breast cancer detection is presented in Table 10.

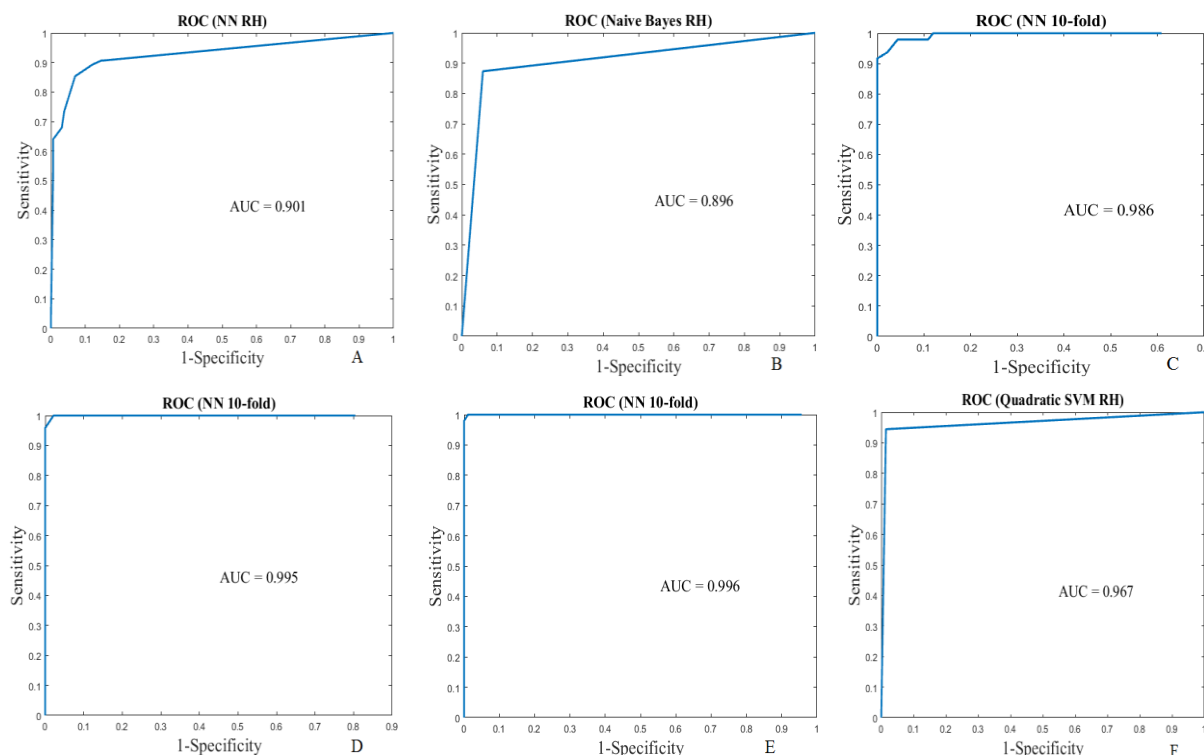


Figure 4. Receiver operating characteristics curves. **(A)** ANN and **(B)** Naïve Bayes classifiers with RH protocol for feature set #1. **(C)**, **(D)** and **(E)** ANN classifier for feature sets #4, #5 and #9 with 10-fold cross validation protocol. **(F)** Quadratic SVM classifier with RH protocol for feature set #5.

Table 5. Classification results for feature set #4.

Parameters	Validation Protocol	Naïve Bayes	ANN	Linear SVM	Quadratic SVM	Cubic SVM	SVM RBF (Coarse)	SVM RBF (Medium)	SVM RBF (Fine)
Sensitivity (%)	RH	96.5	96	93.9	93.4	94.4	92.4	93.9	98.5
	5-fold	95.8	96.7	93.8	93.3	94.2	93.4	95	96.7
	10-fold	95.8	97.9	93.4	93.4	94.2	93.3	95	96.3
Specificity (%)	RH	95	97.2	97	97.9	97.9	97.4	96.2	95.6
	5-fold	95.4	97.8	97.2	96.9	97.2	97.2	96.5	96.1
	10-fold	95.2	99.3	97.2	96.9	97.2	97.2	96.1	95.4
Accuracy (%)	RH	95.5	96.8	95.9	96.3	96.7	95.7	95.4	96.6
	5-fold	95.6	97.4	96	95.7	96.1	95.8	96	96.3
	10-fold	95.4	98.9	95.9	95.7	96.1	95.9	95.7	95.7
AUC	RH	0.957	0.966	0.954	0.956	0.962	0.949	0.950	0.971
	5-fold	0.956	0.972	0.955	0.951	0.957	0.953	0.958	0.964
	10-fold	0.955	0.986	0.953	0.951	0.957	0.953	0.955	0.958

The highest value(s) for the particular data division protocol for all classifiers are shown in bold.

Table 10 shows that the present work (dominance-based filtering approach with ANN classifier) outperformed most of the reported studies on WBCD database (except Chen et al. (2011) and Dora et al. (2017), both got 100% accuracy with holdout data division protocol, which is less recognized in research fields and statistically not very adequate). This ensures the generality of the proposed mechanism for all other kinds of classification tasks.

Table 6. Classification results for feature set #5.

Parameters	Validation Protocol	Naïve Bayes	ANN	Linear SVM	Quadratic SVM	Cubic SVM	SVM RBF (Coarse)	SVM RBF (Medium)	SVM RBF (Fine)
Sensitivity (%)	RH	97	94.3	95.4	95.7	94.9	93.2	97.2	95.4
	5-fold	97.5	95.8	95	95.4	96.3	94.2	97.5	97.5
	10-fold	97.5	99.2	95.9	95	95.8	95.4	97.1	97.5
Specificity (%)	RH	95.1	96.7	96.7	97.7	97.5	97	96	96.6
	5-fold	95.6	98.5	97.4	97	96.7	97.2	96.5	96.1
	10-fold	95.7	99.8	97.2	96.7	96.7	96.9	96.5	96.1
Accuracy (%)	RH	95.7	95.8	96.3	97	96.6	95.7	96.4	96.2
	5-fold	96.3	97.6	96.6	96.4	96.6	96.1	96.9	96.6
	10-fold	96.3	99.6	96.7	96.2	96.4	96.4	96.7	96.6
AUC	RH	0.960	0.955	0.961	0.967	0.962	0.951	0.966	0.960
	5-fold	0.966	0.972	0.962	0.962	0.965	0.957	0.970	0.968
	10-fold	0.966	0.995	0.965	0.959	0.963	0.962	0.968	0.968

The highest value(s) for the particular data division protocol for all classifiers are shown in bold.

Table 7. Classification results for feature set #9.

Parameters	Validation Protocol	Naïve Bayes	ANN	Linear SVM	Quadratic SVM	Cubic SVM	SVM RBF (Coarse)	SVM RBF (Medium)	SVM RBF (Fine)
Sensitivity (%)	RH	96.5	97.4	96.7	94.2	97.5	95.9	96.2	99.2
	5-fold	97.1	96.7	95.8	95	95.4	96.3	96.3	96.3
	10-fold	97.1	99.2	95.9	96.3	94.2	96.3	96.7	96.7
Specificity (%)	RH	95	98.8	97.1	97	96.3	97	97.2	89.9
	5-fold	95.2	99.1	97.4	97.4	97.2	97.1	97.2	97.6
	10-fold	95.2	99.8	97.2	97.4	97.2	96.9	96.9	96.9
Accuracy (%)	RH	95.5	98.4	97	96	96.7	96.6	96.9	93.1
	5-fold	95.8	98.3	96.8	96.6	96.6	96.8	96.9	97.1
	10-fold	95.9	99.7	96.7	97	96.1	96.7	96.8	96.8
AUC	RH	0.957	0.981	0.969	0.956	0.969	0.965	0.967	0.946
	5-fold	0.961	0.979	0.966	0.962	0.963	0.967	0.967	0.969
	10-fold	0.961	0.996	0.965	0.968	0.957	0.966	0.968	0.968

The highest value(s) for the particular data division protocol for all classifiers are shown in bold.

Table 8. Comparative chart of ANN classifier with 10-fold cross validation protocol for feature sets #4, #5 and #9 along with percentage errors.

Parameters	Feature set	Feature set	Feature set	% Error from set #9	
	#4	#5	#9	Set #4	Set #5
Sensitivity (%)	97.9	99.2	99.2	1.3	0
Specificity (%)	99.3	99.8	99.8	0.5	0
Accuracy (%)	98.9	99.6	99.7	0.8	0.1
AUC	0.986	0.995	0.996	1	0.1

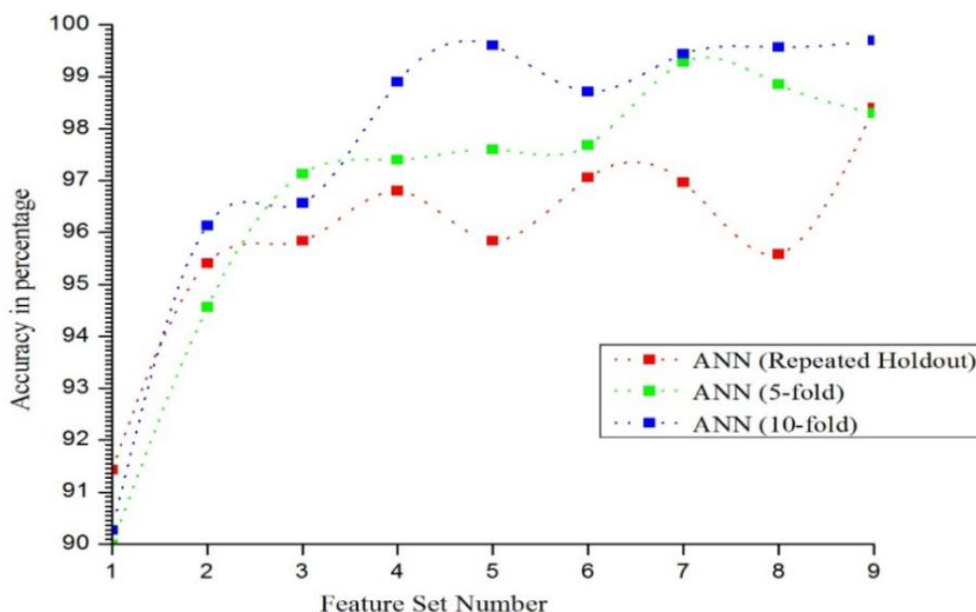


Figure 5. Flow of accuracy of ANN classifier for all feature sets (feature set #1 to #9) for RH, 5-fold and 10-fold data division protocols.

Table 9. ANN classifier performances with 10-fold cross validation protocol in terms of Youden's index and Discriminatory power for feature sets #4, #5 and #9.

Parameters	Feature set #4	Feature set #5	Feature set #9
Youden's Index (γ)	0.972	0.99	0.99
Discriminatory Power (DP)	4.849	6.082	6.082

Table 10. Comparison of present work with past literatures.

Author(s)	Method	Feature selection method	Highest Acc. (%)
Polat and Gunes (2007)	LS-SVM	Not used	98.53
Sahan et al. (2007)	k-NN	AIS Algorithm	99.14
Yeh et al. (2009)	DPSO	Statistical Analysis	98.71
Subhasini et al. (2009)	RBFNN	Not used	96.56
Chen et al. (2011)	SVM	RS Algorithm	100
Marcano et al. (2011)	AMMLP	Not used	99.26
Inan et al. (2013)	ANN	PCA	99.29
Zheng et al. (2014)	SVM	K-means	97.38
Bhardwaj&Tiwari (2015)	GONN	Not used	99.26
Aličković&Subasi (2017)	Rot. Forest	GA	99.48
Dora et al. (2017)	GNRBA	CCA	100
Present Work	ANN	Dominance-based	99.70

CONCLUSIONS AND FUTURE SCOPE

This work reports the impact of dominance-based filtering approach on performances of major state-of-the-art classifiers used in machine learning paradigm. WBCD database was utilized for RH, 5-fold and 10-fold cross validation protocols using sensitivity, specificity, accuracy and AUC parameters. ANN has emerged as the best classifier with classification accuracies of 98.9% for four dominant features (feature set #4), 99.6% for five dominant features (feature set #5) and 99.7% for all features (feature set #9) subjected to the 10-fold

crossvalidation protocol. It is interesting to note that the accuracies and AUCs conceived by ANN classifier for four and five abridged features of dominance-based filtering technique are almost equivalent to the values for all nine features (feature set #9) with less than 2% error.

There are some limitations of this work which can be looked out in future by researchers. As this empirical study was done in WBCD (original) dataset consists of narrow feature space ($n = 9$), the reliability of the research can be increased by introducing a live breast-cancer dataset with a large number of feature vectors. Also, though dominance-based filtering technique is computationally fast, the feature rank is algorithm liberated. This can be prevented by adding an algorithm-relative block with proposed tactic for rank calculation. Furthermore, the proposed approach can be applied for other cancer datasets to test its generalized performance capabilities.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc Natl Acad Sci.* 1990;87(23):9193–6.
2. Abonyi J, Szeifert F. Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognit Lett.* 2003;24(14):2195–207.
3. Chen TC, Hsu TC. A gas based approach for mining breast cancer pattern. *Expert Syst Appl.* 2006;30(4):674–81.
4. Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. *Digit Signal Process.* 2007;17(4):694–701.
5. 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):415–23.
6. Yeh W-C, Chang W-W, Chung YY. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Syst Appl.* 2009;36(4):8204–11.
7. Subashini TS, Ramalingam V, Palanivel S. Breast mass classification based on cytological patterns using rbfn and svm. *Expert Syst Appl.* 2009;36 (3):5284–90.
8. Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Syst Appl.* 2011;38(7):9014–22.
9. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Syst Appl.* 2011;38(8):9573–79.
10. Zanyat EA. Support vector machines (svms) versus multilayer perception (mlp) in data classification. *Egypt Informatics J.* 2012;13(3):177–83.
11. Ren J. Ann vs. svm: which one performs better in classification of mcs in mammogram imaging. *Knowledge-Based Syst.* 2012;26:144–53.
12. Inan O, Uzer MS, Yılmaz N. A new hybrid feature selection method based on association rules and pca for detection of breast cancer. *Int J Innov Comput.* 2013;9(2):727–39.
13. Chen CH. A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. *Appl Soft Comput J.* 2014;20:4–14.
14. 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):1476–82.
15. Bhardwaj A, Tiwari A. Breast cancer diagnosis using genetically optimized neural network model. *Expert Syst Appl.* 2015;42(10):4611–20.
16. Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and rotation forest. *Neural Comput Appl.* 2017;28(4):753–63.
17. Dora L, Agrawal S, Panda R, Abraham A. Optimal breast cancer classification using gauss–newton representation based algorithm. *Expert Syst Appl.* 2017;85:134–45.
18. Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. *Expert Syst Appl.* 2015;42(15–16):6184–95.

19. Mangasarian O. L. and Wolberg W. H.: Cancer diagnosis via linear programming. *SIAM News*, 1990 September; 23(5): pp 1 & 18.
20. Uci Machine Learning Repository: Breast Cancer Data Set [Internet]. [place unknown: publisher unknown] [cited 2019 Apr 25]. Available from: <https://archive.ics.uci.edu/ml/datasets/breast+cancer>.
21. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. 3rd ed. USA: Morgan Kaufmann Publ. Elsevier; 2012. p.745.
22. Haykin S. *Neural Networks and Learning Machines*. Pearson Educ.2009;3:146-8.
23. Dash M, Liu H. Feature selection for classification. *Ida Elsevier Intell Data Anal*. 1997;1(97):131–56.
24. Fisher R. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd; 1925.p. 239.
25. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell*. 1997;97(1–2):273–324.
26. Suto J, Oniga S, Sitar PP, editors. Comparison of wrapper and filter feature selection algorithms on human activity recognition. 6th International Conference On Computers Communications and Control(Icccc); 2016; Oradea:IEEE; 2016, p. 124–9.
27. Theodoridis S, Koutroumbas K. *Pattern Recognition*. USA: Elsevier, 2009. p. 961.
28. Lewis D. D., “Naive(Bayes)at forty: The independence assumption in information retrieval,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1998 April, 1398, pp. 4–15.
29. Meena K, Malarvizhi N. An efficient human identification through multimodal biometric system. *Brazilian Arch Biol Technol*. 2016;59(2):1–10.
30. Cortes C, Vapnik V. Support-Vector networks. *Mach Learn*. 1995;20(3):273–97. doi:10.1023/a:1022627411411networks.
31. Duda RO, Hart PE, Stork DG. *Pattern Classification*. USA: Wiley-Interscience, 2000. p. 680.
32. Byun H, Lee SW. Applications of support vector machines for pattern recognition: A Survey. In: *Lecture Notes In Computer Science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. Canada: Springer Verlag; 2002.2388.p. 213–36.
33. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational Learning Theory*. USA: Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92); 1992. p. 144–52.
34. Hagan MT, Demuth HB, Beale MH. *Neural Network Design*. USA: Martin Hagan. 2014. p. 800..
35. Krishnamurthy S, Narasimhan G, Rengasamy U. Early and accurate model of malignant lung nodule detection system with less false positives. *Brazilian Arch Biol Technol*. 2018, pp 1-12.
36. Youden WJ. Index for Rating Diagnostic Tests. *Cancer*. 1950;3(1):32–5.
37. Blakeley DD, Oddone EZ, Hasselblad V, Simel DL, Matchar DB. Noninvasive carotid-artery testing - A metaanalytic review. *Ann Intern Med*. 1995;122(5):360–7.



© 2018 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>).