A Computer Aided Diagnosis System for Breast Cancer

Alireza Osareh[1] and Bita Shadgar[2]

[1] Computer Engineering Department, Shahid Chamran University, Ahvaz, Iran

[2] Computer Engineering Department, Shahid Chamran University, Ahvaz, Iran

Abstract

Breast cancer is one of the most prevalent cancers, ranking second only to lung cancer and is the most prevalent form of cancer among worldwide women. Each year, about 1,000,000 women would be newly diagnosed with breast cancer and over 500,000 women died from breast cancer every year. In this paper, a computer-aided diagnosis (CAD) framework for breast cancer is developed using application of supervised machine learning techniques to the classification of cancerous /non-cancerous data. Here, we attempt to explore several different feature selection and extraction techniques and combine the optimal feature subsets with various learning classification methods such as K-nearest neighbors, probabilistic neural networks and support vector machines classifiers. To evaluate the generalization ability of the proposed system for distinguishing the benign and malignant cases, 2 benchmark FNAB and gene microarray datasets are utilized. The best overall accuracy for breast cancer diagnosis is achieved equal to 98.80% and 96.33% respectively using support vector machines classifier models against two widely used breast cancer benchmark datasets.

Keywords: *Signal-to-noise, False positive, K-nearest neighbors, Probabilistic neural networks, Support vector machines*

1. Introduction

Cancer begins with uncontrolled division of one cell, which results in a visible mass named tumor. Tumor can be benign or malignant. Malignant tumor grows rapidly and invades its surrounding tissues causing their damage. Breast cancer is a malignant tissue beginning to grow in the breast. The symptoms of breast cancer include breast mass, change in shape and dimension of breast, differences in the color of breast skin, breast aches and gene changes etc.

Breast cancer is the second leading cause of death for women all over the world and more than 8% women will suffer this disease during their lifetime [1]. According to the report of the World Health Organization, about 1,000,000 women would be newly diagnosed with breast cancer and over 500,000 women died from breast cancer every year [1]. It is estimated that the incidence of this disease will increase getting along with the damaging of environment in the future.

In 2008, there were reported approximately 182,460 newly diagnosed cases and 40,480 deaths in the United States [2]. Since the causes of breast cancer still remain unknown, early detection is the key to reduce the death rate (40% or more). The earlier the cancers are detected, the better treatment can be provided. However, early detection requires an accurate and reliable diagnosis which should also be able to distinguish benign and malignant tumors. A good detection approach should produce both low false positive (FP) rate and false negative (FN) rate. Previously, the most effective modality for detecting and diagnosing is mammography [2].

Although breast cancer incidence has increased over the past decade, breast cancer mortality has declined among women of all ages [3]. This favorable trend in mortality reduction may relate to improvements made in breast cancer treatment and the widespread adoption of mammography screening. However, it is well known that expert radiologists can miss a significant proportion of abnormalities [4]. In addition, a large number of mammographic abnormalities turn out to be benign after biopsy.

Conventional methods of monitoring and diagnosing the diseases rely on detecting the presence of particular signal features by a human observer. Due to large number of patients in intensive care units and the need for continuous observation of such conditions, several computer aided-diagnosis approaches for automated diagnostic systems have been developed in the past ten years to attempt to solve this problem. Such techniques work by transforming the mostly qualitative diagnostic criteria into a more objective quantitative feature classification problem [5].

Fig. 1 shows the various stages followed for the design of a classification system. As it is apparent from the feedback arrows, these stages are not independent. On the contrary, they are interrelated and, depending on the results, one may go back to redesign earlier stages in order to improve the overall performance.

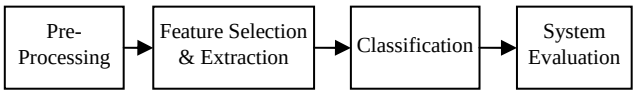


Fig. 1 Different steps of a typical CAD system for cancer detection

Medical diagnostic decision support systems have become an established component of medical technology. The main concept of the medical technology is an inductive engine that learns the decision characteristics of the diseases and can then be used to diagnose future patients with uncertain disease states.

In order to improve the accuracy of diagnosis and evaluate the prognostic risk, a number of CAD approaches have been proposed for breast cancer diagnosis and prognostic risk evaluation. For instance, Butler et al. applied Bays classifiers combined with feature selection to diagnose breast cancer, which reached 90% accuracy by using X-Ray scatter images [6]. Song et al. adopted artificial neural network by using ultrasound image of breast to predict breast cancer and got 95% sensitivity and 76.5% specificity [7]. Abonyi and Szeifert applied supervised fuzzy clustering technique, and obtained 95.57% accuracy [8]. Setiono got 98.1% overall accuracy by using neuro-rule method [9].

All above studies demonstrate that CAD is capable of improving the radiologist's performance. In this paper, several state of the arts machine learning classifier models, such as support vector machines (SVM), K nearest neighbors (KNN) and probabilistic neural networks (PNN) are combined with feature selection/extraction algorithms to discriminate between benign tumors and breast cancer patients.

2. Materials and Methods

2.1 Datasets

Breast cancer is a malignant tumor that has developed from cells of the breast. Although scientists know some of the risk factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity) that increase a woman's chance of developing breast cancer, they do not yet know what causes most breast cancers or exactly how some of these risk factors cause cells to become cancerous. Research is under way to learn more and scientists are making great progress in understanding how certain changes in DNA can cause normal breast cells to become cancerous.

In this work, we utilized 2 publicly available benchmark datasets. The dataset of fine needle aspirate of breast lesions (dataset I) contains 692 specimens of fine needle aspirates of breast lumps (FNAB), including 235 positive samples (malignancy) and 457 negative samples (benign). All of the specimens were confirmed by open biopsy and each sample includes 11 features which consist of patient age and ten attributes of the cell. The observations of the cellular attributes were all made by a consultant pathologist.

In addition to dataset I, a second dataset comprised of gene microarrays comes from references [10] and [11] was also exploited. It contains 295 microarrays, 115 belong to the "good-prognosis" class and the remaining 180 belong to the "poor-prognosis" class. Indeed, each sample contains a 70 gene prognosis profile.

A typical microarray holds spots representing several thousand to several tens of thousands of genes or ESTs (expressed sequence tags). After hybridization the microarray is scanned and converted into numerical data. Finally the data should be normalized. The purpose of this step is to counter systematic variation (e.g. difference in labeling efficiency for different dyes, compensation for signal spill over from neighboring spots) and to allow a comparison between different microarrays [18].

2.2 Feature Extraction and Selection

Feature extraction and selection are important steps in breast cancer detection and classification. An optimum feature set should have effective and discriminating features, while mostly reduce the redundancy of features pace to avoid "curse of dimensionality" problem. The "curse of dimensionality" suggests that the sampling density of the training data is too low to promise a meaningful estimation of a high dimensional classification function with the available finite number of training data. For some advanced classification methods, such as artificial neural network and support vector machine, the dimension of feature vectors not only highly affects the performance of the classification, but also determines the training time of the algorithm. Thus, how to extract useful features and make a good selection of the features is a crucial task for CAD systems.

Feature extraction, linearly or nonlinearly, transforms the coordinate system of the original variables. The most well-

known feature extraction technique is principal component analysis (PCA). PCA performs on the symmetric covariance matrix or symmetric correlation matrix, and solves the Eigen values and Eigenvectors of the matrix. PCA is good at reducing the high dimensional correlated features into low dimensional features. The feature vector of the auto-covariance coefficients can be optimized by PCA effectively.

Generally, algorithms for feature selection can be categorized into two classes: filter and wrapper. Filter approach (such as Relief algorithm [12]) selects features using a pre-processing step and does not take into account the bias of induction algorithms. On the contrary, to search for a good subset of the features, wrapper approach uses the induction algorithm as a part of the evaluation function. As the wrapper approach has obvious advantages over filter approach, especially for complex feature dataset, these techniques have more applications in breast cancer detection [13]. For example, [13] applied a wrapper approach (linear stepwise feature selection) to a feature set composed of 15 sonographic features of breast cancer and found that the two most significant features were the average orientation of gray level gradients along the margin and depth-to-width ratio.

Signal-to-Noise Ratio: A feature saliency measures provide a filter method to measure the relative usefulness of features and a means to rank the features. Signal-to-noise ratio is a value that uses the signal to compare with other background noise. Usually, it is simple and capable of fast ranking and filtering features for classifiers [14]. The definition of signal to noise ratio (*SNR*) for two classes is formulated as:

$$SNR_i = \frac{|\mu_P(i) - \mu_N(i)|}{\sigma_P(i) + \sigma_N(i)} \quad (1)$$

where SNR_i is the value of saliency metric for the i-th feature; $\mu_N(i)$ and $\mu_P(i)$ are the averages of the i-th feature in class N and class P respectively; $\sigma_N(i)$ and $\sigma_p(i)$ are the standard deviations of the i-th feature in class N and class P respectively.

Sequential forward selection: In this study, we propose using a wrapper approach based on sequential forward selection (SFS) [15], a classic and well-known hill-climbing, deterministic search algorithm which starts from an empty subset of genes. It sequentially selects genes, one at a time, until no further improvement is achieved in the evaluation function value. As another advantage with respect to filter approaches, our wrapper approach does not need to fix a specific number of features to train the final classifier, and the number of features that induce the final classification model is selected by the search component inserted in the own wrapper procedure.

Our wrapper approach estimates, by the leave one-out cross-validation (LOOCV) procedure [16], the goodness of the classifiers using only the feature subset found by the SFS search procedure. Thus, the breast cancer datasets are projected maintaining the only values of the selected features and the class variable for all samples: the goodness of the proposed feature subsets, using the specific classifier, is estimated by the explained LOOCV technique over this projected dataset, which only includes the features selected by the SFS search procedure and the class of the samples.

2.3 Classification Models

After the features have been extracted and selected, they are input into a classifier to categorize the input samples into lesion/non-lesion or benign/malignant classes. Here, 3 supervised learning algorithms are utilized to build models to perform classification, namely, SVM, KNN and probabilistic neural network (PNN) [10].

Support Vector Machines: SVM was proposed by Vapnik et al., [17] based on the statistical learning theory and structural risk minimization, which was extensively used as an effective algorithm to deal with classification and regression problems. This method separates the classes of input patterns with the maximal margin hyperplane. This hyper plane is constructed as [17]:

$$f(x) = \langle w, x \rangle + b \quad (2)$$

where x is the feature vector, w is the vector that is perpendicular to the hyper plane, and $b\|w\|^{-1}$ specifies the offset from the beginning of the coordinate system. To benefit from non-linear decision boundaries, the separation is performed in a *feature space F*, which is introduced by a nonlinear mapping φ of the input patterns. This mapping is defined as follows:

$$\langle \varphi(x_1), \varphi(x_2) \rangle = K(x_1, x_2) \qquad \forall x_1, x_2 \in \mathrm{X} \quad (3)$$

for some kernel function K(·,·). The kernel function represents the non-linear transformation of the original feature space into the F. However, to guarantee that the resultant hyper-plane separates the classes; the following constraints must be satisfied:

$$y_i.(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1,..n \quad (4)$$

where $y_i \in \{-1,1\}$ denotes the class label corresponding to the input pattern $\boldsymbol{x}_i$. The variables ξ_i are utilized to allow for the training of the classifier on linearly non-separable classes. The slack variables must be penalized in the minimization term. Consequently, learning of the SVM classifier is equivalent to solving a minimization problem with the objective function of the form:

$$\min_{w \in X} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{5}$$

The penalty C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. This approach is called soft margins. Using the Lagrange multiplier technique, we can transform this optimization problem to a dual form:

$$\min_{\alpha \in R^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j . K(x_i, x_j) \tag{6}$$

subject to:

$$0 \le \alpha_i \le C \qquad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{7}$$

In above formulation, the $\alpha = \{\alpha_1, \alpha_2, ...\alpha_n\}$ is the vector of Lagrange multipliers. The Lagrange multipliers that solve the Equation (6) can be used to compute the decision function:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \tag{8}$$

where

$$b = y_i - \sum_{j=1}^{n} \alpha_j y_j K(x_j, x_i) \tag{9}$$

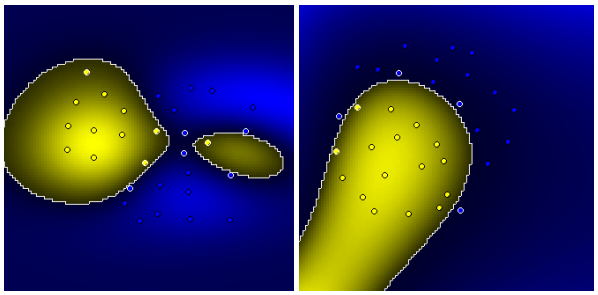
There are many kernels that can be used such as Gaussian radial basis functions (RBF):

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{10}$$

where $\sigma > 0$ is a constant that defines the kernel width. Another kernel function is the polynomial (of degree d):

$$K(x_i, x_j) = (1 + x_i . x_j)^d \tag{11}$$

where $d > 0$ is a constant that defines the kernel order. The associated parameters, order d or Gaussian σ are determined within the training phase. Fig. 2, demonstrates examples of nonlinear binary classification using SVM with RBF (σ =10) and polynomial (d = 5) kernels respectively.



(a) (b)

Fig. 2 Proposed beam former SVM decision surfaces given in (a) by a RBF classifier, and in (b) by a polynomial, where the support vectors are indicated by a white ring around the sample points.

The parameters σ and d represent how sparse and easily separable the data are in the feature space, and thus, they affect the complexity of the resulting SVM classifier.

K-Nearest Neighbor: KNN classifier is one of the simplest and oldest methods for performing general, nonparametric classification. In this model, the distances between the test sample and all the other samples in the training set is first measured. Then, the class of the test sample is assigned according to a simple majority vote over the labels of its K nearest neighbors.

Probabilistic Neural Network: PNN was proposed by Specht in 1988 [18]. It is designed to improve the performance of conventional neural networks in which long computation times are required. PNN replaces the sigmoid activation function often used in neural networks with a statistically derived exponential function. The PNN is an extension of what is probably the simplest possible classifier i.e., find the training sample closest to the test sample and assign it the same class. A single PNN is capable of handling multiclass problem. This is opposite to the so-called one-against-the rest or one-per class approach taken by some classifiers, such as the SVM, which decompose a multiclass classification problem into dichotomies and each chotomizer has to separate a single class from all others [19]. The architecture of a typical PNN is as shown in Fig. 3.

The PNN architecture is composed of many interconnected processing units or neurons organized in successive layers. The input layer unit does not perform any computation and simply distributes the input to the neurons in the pattern layer.

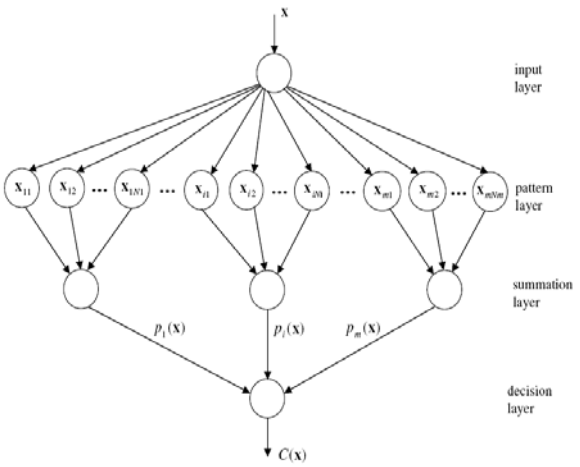


Fig. 3 A typical PNN architecture

On receiving a pattern x from the input layer, the neuron x_{ij} of the pattern layer computes its output as:

$$\varphi_{ij}(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left[-\frac{(x-x_{ij})^T(x-x_{ij})}{2\sigma^2}\right] \quad (12)$$

where d denotes the dimension of the pattern vector x, σ is the smoothing parameter and x_{ij} is the neuron vector. The summation layer neurons compute the maximum likelihood of pattern x being classified into C_i by summarizing and averaging the output of all neurons that belong to the same class:

$$p_i(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \frac{1}{N_i} \sum_{j=1}^{N_i} \exp\left[-\frac{(x-x_{ij})^T(x-x_{ij})}{2\sigma^2}\right] \quad (13)$$

where N_i denotes the total number of samples in class C_i. If the a priori probabilities for each class are the same, and the losses associated with making an incorrect decision for each class are the same, the decision layer unit classifies the pattern x in accordance with the Bayes's decision rule based on the output of all the summation layer neurons:

$$C(x) = \arg\max(p_i(x)), \quad i = 1,2,...,m \quad (14)$$

where $C(x)$ denotes the estimated class of the pattern x and m is the total number of classes in the training samples.

3. Results and Discussion

3.1 Feature Ranking based on SNR Filter

SNR was employed to find out the predominant features and filter the irrelevant features for classification. The 11 features of dataset I were ranked by using SNR index. The ranking result is shown in Table 1. It can be seen that feature numbers 9, 7 and 8 are ranked as the top 3 by using SNR criterion, corresponding to 'Necrotic epithelial cells', 'Nuclear pleiomorphism' and 'Nuclear size', respectively.

This is consistent with the result of other researcher that 'Nuclear size' has great significance to distinguish the benign from malignant breast tumor [20]. The ranking results indicated that the above 3 features contain more informative and important information than other features for distinguishing between benign tumor and breast cancer. It supplies a valuable clue for cytopathologist to pay more attention to these factors in their clinical breast tumor diagnoses.

Table 1: Feature ranking results of dataset I by using SNR criterion.

Feature ranking Method	Feature rank										
	1	2	3	4	5	6	7	8	9	10	11
SNR	9	7	8	4	1	3	2	10	11	5	6

3.2 Classification Results

Three evaluation terms i.e., sensitivity (*Sen*), specificity (*Spe*) and overall accuracy (*O*) [21] were introduced to estimate the performance of classifiers. They are defined as follow:

$$Sen = TP/(TP + FN) \quad (15)$$

$$Spe = TN/(TN + FP) \quad (16)$$

$$O = (TP + TN)/(TP + FN + TN + FP) \quad (17)$$

where TP and TN are the number of samples which are correctly identified as positives or negatives by the classifier in the test set, respectively and FN and FP represent the numbers of samples corresponding to those cases as they are mistakenly classified as benign or malignant, respectively.

Considering imbalanced positive and negative samples in the datasets, another appropriate quantity for evaluating the classification accuracy of imbalanced positive and negative samples is the Matthews Correlation Coefficient *MCC*, which is given as follows [14]:

$$MCC = \frac{TP.TN - FN.FP}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \quad (18)$$

Obviously, the scope of the *MCC* is within the range of [-1, 1]. The larger the *MCC* value, the better the classifier performance.

In this study, machine learning approaches including SVM, KNN and PNN were applied to diagnose breast cancer using dataset I (using top 3 ranked features) and evaluate the prognostic risk of recrudescence and metastasis through dataset II when either feature subset selection or feature extraction, by using LOOCV.

Training algorithm of the SVM, based on quadratic programming, incorporates several optimization techniques such as decomposition and caching. The quadratic programming problem in the SVM was solved by using the MATLAB optimization toolbox. For the implementation of the SVMs with the RBF kernel functions (SVM-RBF), one has to assume a value for σ. The optimal σ can only be found by systematically varying its value in the different training from the training data file with an assumed σ value. After the support vectors have been found and SVM constructed, the model was applied to compute the misclassification rate. The σ value was varied between 0.1 and 0.9, at interval of 0.1. The σ = 0.5 resulted in the minimum misclassification rate was thus chosen.

The generalization ability of the SVM is controlled by two different factors: the training error rate and the capacity of the learning machine measured by its Vapnik-Chervonenkis dimension. The smaller the VC dimension of the function set of the learning machine, the larger the value of training error rate. In this work, the tradeoff between the complexity of decision rule and training error rate was controlled by changing a parameter *C* in the SVM classifiers. The SVMs were trained for different *C* values until to have the best result. The best result was obtained for *C* = 15 in the testing procedure. Moreover, in this case the number of support vectors in the SVMs training was found to be 25. Indeed, the polynomial kernel (SVM-Poly) was used where the optimum parameters were obtained as *d* = 6, *C* = 8.5 and the number of support vectors were found equal to 32.

In KNN classification, the number of neighbors, i.e. K needs to be pre-defined. A reasonable and practical approach would be to use trial and error to identify K such that it gives the lowest misclassification error rate. We performed such an experiment with different K values ranging from 1 to 7 (K is chosen to be odd to avoid ties) and found K = 5 as optimum value.

There was an outstanding issue associated with the PNN concerning network structure determination that is determining the network size, the locations of pattern layer neurons as well as the value of the smoothing parameter. The PNN had 20 pattern layer neurons, two summation layer neurons; each corresponds to one of two classes and one output layer neuron to make a two-class Bayesian decision.

The objective was to select representative pattern layer neurons from the training samples. The output of a summation layer neuron becomes a linear combination of the outputs of pattern layer neurons. Subsequently, an orthogonal algorithm was used to select pattern layer neurons. As in the SVM training, the smoothing parameter r was determined based on the minimum misclassification rate computed from the partial evaluation data set. The minimum misclassification rate was attained at σ = 0.1.

The overall accuracies for LOOCV by using dataset I are shown in Table 2. As it can be seen, the optimum overall accuracies of SVM-Poly, SVM-RBF, KNN (K = 5) and PNN were achieved equal to 97.09%, 98.80%, 96.37% and 97.23% respectively. All of the overall accuracies of SVM with 2 kernel functions are superior to those of KNN and the SVM-RBF provided the highest accuracy (98.80%).

Table 2: LOOCV results of dataset I by using the original features.

Classifier	***Sen* (%)**	***Spe* (%)**	***O*(%)**	***MCC***
SVM-Poly	95.19	99.20	97.09	0.922
SVM-RBF	95.45	99.63	98.80	0.936
KNN	94.06	92.59	93.37	0.890
PNN	92.86	97.15	97.23	0.928

In the second phase, PCA was applied to the dataset II. The output of the PCA was a set of eigenvectors and Eigen values, with the Eigen values representing the amount of variance over the whole dataset for each vector. The first two principal modes contained 62.4% of the total features variance, i.e. 49.7% + 12.7%. A central issue in PCA is choosing the number of principal components to be retained. Here, we conducted another experiment and projected the dataset II onto the subspace of the principal components which accounted for more than 95% of the total variance. The first 10 Eigen values contained 95.1% of total variance.

Then, we developed a SVM-RBF based classifier in reduced 10-dimensional feature space (retaining 95.1% of the total variance). Having trained the SVM-RBF classifier in this subspace, the generalization performance was then measured. The constructed classifier achieved 95.01% overall accuracy (96.11% sensitivity and 93.20% specificity) and an *MCC* value equal to 0.923 respectively.

Considering each sample of dataset II contains 70 features, SFS wrapper-based technique was employed to identify the predominant features and filter the irrelevant features to

further improve the prediction accuracy and save computational time. In this way, we aimed to choose the most informative and accurate subsets of features towards our breast cancer diagnosis task. Among the initial 70 features, 25 features were selected as the best subset. Having done that, the selected features were input into our classifiers to categorize the samples into cancerous/not-cancerous classes.

The overall accuracies for LOOCV by using dataset II are shown in Table 3. As it can be seen, the optimum overall accuracies of SVM-Poly, SVM-RBF, KNN (K = 7) and PNN were achieved equal to 95.00%, 96.33%, 88.45% and 93.39% respectively. All of the overall accuracies of SVMs with 2 kernel functions were superior to those of KNN and PNN the SVM-RBF provided the highest accuracy. Indeed, value of *MCC* for SVM-RBF achieved 0.944 and was again superior to those of other classifiers. In fact, SVM-RBF has shown excellent performance and outperforms other classifiers to distinguish the abnormal cases from normal ones.

Table 3: LOOCV results of dataset II by using the wrapper-based technique

Classifier	***Sen* (%)**	***Spe* (%)**	***O*(%)**	***MCC***
SVM-Poly	94.34	92.20	95.00	0.911
SVM-RBF	96.85	93.11	96.33	0.944
KNN	89.28	85.03	88.45	0.871
PNN	91.27	95.54	93.39	0.898

4. Conclusions

In this paper, we have investigated the issues of breast cancer diagnosis and prognostic risk evaluation of recrudescence and metastasis by using 3 well-known classifiers i.e., SVM, KNN, PNN. These classifiers were combined with SNR feature ranking method; SFS feature selection and PCA feature extraction based on FNAB dataset I and gene microarrays dataset II, respectively.

Feature ranking and filtering supplied the informative and important features to classify breast tumor. It provides the physicians a valuable clue to pay more attention to these relevant features in their clinical breast tumor diagnosis. Feature ranking and filtering also improved the evaluation performance to the prognostic risk of recrudescence and metastasis.

The best overall accuracy for breast cancer diagnosis and prognostic risk of recrudescence and metastasis was achieved equal to 98.80% for dataset I by using a fine-tuned SVM-RBF classifier.

On the other hand, the PCA could retain more than %95 of whole features variance for dataset II by using the first 10 Eigen values (out of 70) of the covariance matrix. The optimum SVM-RBF classifier in this reduces 10-dimensional feature space achieved 95.01% overall accuracy. Indeed to investigate the effectiveness of feature selection methods, SFS wrapper-based technique was utilized. This technique in turn, chose the most prominent 25 features of dataset II out of 70. Again in this reduced feature space and among different constructed classifiers SVM-RBF provided the highest overall accuracy equal to %96.33.

It revealed that classifier and kernel function selection are necessary to get the best results. The study suggests that SVM models and especially RBF kernels may be further developed to be a potential practical methodology for clinical assistant breast cancer diagnosis by providing the physicians with the immediate second opinion. Such a CAD system can also help the inexperienced physicians to avoid misdiagnosis.

References

[1] D. Parkin, “Epidemiology of cancer: global patterns and trends” *Toxicology Letters*. vol. 5, pp. 102-103, 1998.

[2] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray and M.J. Thun, “Cancer statistics” *A Cancer Journal for Clinicians*. vol. 58, pp. 71–96, 2008.

[3] F. Bray, P. McCarron, and D. M. Parkin, “The changing global patterns of female breast cancer incidence and mortality” *Breast Cancer Res.* vol. 6, pp. 229-239, 2004.

[4] R. L. Birdwell, D. M. Ikeda, K. D. O’Shaughnessy, and E. A. Sickles, “Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection” *Radiology*. vol. 219, pp. 192-202, 2001.

[5] N. Guler, E. Ubeyli and I. Guler, “Recurrent neural network employing Lyapunov exponents for EEG signals classifications” *Expert systems with Applications*. vol. 29, pp. 506-514, 2005.

[6] S. Butler and G. Webb, “A case study in feature invention for breast cancer diagnosis using X-Ray scatter images” *Lecture Notes in Artificial Intelligence*. Vol. 2903, pp. 677-685, 2003.

[7] J. Song and S. Venkatesh, “Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses” *Academic Radiology*. vol. 12, pp. 487-496, 2005.

[8] J. Abonyi and F. Szeifert “Supervised fuzzy clustering for the identification of fuzzy classifiers” *Pattern Recognition Letters*. vol. 24, pp. 2195-2207, 2003.

[9] R. Setiono, “Generating concise and accurate classification rules for breast cancer diagnosis” *Artificial Intelligence in Medicine*. vol. 18**,** pp. 205-219, 2000.

[10] L. Hadjiiski and B. Sahiner, “Advances in computer-aided diagnosis for breast cancer. *Curr. Opin. Obstet. Gynecol*. vol. 18, pp. 64-70, 2006.

[11] L. Veer and H. Dai, “Gene expression profiling predicts clinical outcome of breast cancer” *Nature.* vol. 415, pp. 530-536, 2002.

[12] R. Kohavi and G.H. John, "Wrappers for feature subset selection" *Artificial Intelligence*, vol. 97, pp. 273–324. 1997.

[13] K. Horsch, A.F. Ceballos, M.L. Giger, I.R. Bonta, Z. Huo, C.J. Vyborny, E.R. Hendrick and L. Lan, "Optimizing feature selection across a multimodality database in computerized classification of breast lesions" *Progress in Biomedical Optics and Imaging*, vol. 3, pp. 986–992, 2002.

[14] Q. Yuan, C. Cai, H. Xiao, X. Liu, Y. Wen, "Diagnosis of Breast Tumours and Evaluation of Prognostic Risk by Using Machine Learning Approaches" *Communications in Computer and Information Science*, vol. 2, pp. 1250-1260, 2007.

[15] A. Choudhary and M. Brun, "Genetic test bed for feature selection" *Bioinformatics*. vol. 22, pp. 837-842, 2006.

[16] J. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. John-Wiley, 1973.

[17] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[18] D. Specht, "Probabilistic neural networks for classification, mapping or associative memory" *in Proc. IEEE Int. Conf. Neural Network,* pp. 525-532, 1988.

[19] E. Ubeyli, "Comparision of different classification algorithms in clinical decision-making, *Experts Systems*, vol. 24, no. 1, pp. 17-31, 2007.

[20] W. Wolberg and W. Street, "Computerized breast cancer diagnosis and prognosis from fine-needle aspirates" *Arch. Surg*. vol. 130, pp. 511-516, 1995.

[21] D. Delen and G. Walker, "Predicting breast cancer survivability: a comparison of three data mining methods" *Artificial Intelligence in Medicine*, vol. 34, pp. 113-127, 2005.