We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



9

Improving Prostate Cancer Classification: A Round Robin Forward Sequential Selection Approach

Sabrina Bouatmane[1], Ahmed Bouridane[2,3],
Mohamed Ali Roula[4] and Somaya Al-Maadeed[5]
[1]*Département d'Electronique, Faculté des Sciences de l'Ingénieur, Université de Jijel, Jijel*
[2]*Departemt of Computer Science, King Saud University, Riyadh*
[3]*School of Computing, Engineering and Information Sciences*
Northumbria University at Newcastle, Pandon Building
[4]*Faculty of Advanced Technology, University of Glamorgan, Pontypridd*
[5]*Department of Computer Science & Engineering Qatar University, Doha*
[1]*Algérie*
[2]*Saudi Arabia*
[3,4]*UK*
[5]*Qatar*

1. Introduction

Over the last decade prostate cancer has become one of the most common cancer in male population with an estimated 1.37 million people diagnosed and 200,000 annual death rate worldwide (Stewart & Kleihues, 2003). Biopsies are often advised after a Prostate Specific Antigen (PSA) test reveals high levels of PSA in the blood which usually indicate high risks of Prostatic Carcinoma (PCa). The biopsy is needed because high PSA levels can also be caused by other benign conditions like Benign Prostatic Hyperplasia (BPH) (Kronz, Westra & Epstein, 1999)

Biopsy of the prostate, usually stained by Hematoxylin and Eosin (H&E) technique, is the key step for confirming the diagnosis of malignancy and grading treatment. By viewing the microscopic images of biopsy specimens, pathologists can determine the histological grades. In December 1999, a study of more than 6,000 patients by Johns Hopkins researchers found that up to two out of every 100 people who come to larger medical centers for treatment, following a biopsy, are given a diagnosis that is "totally wrong". The results suggested that second opinion pathology examinations not only prevent errors, but also save lives and money. Human assessment is time consuming and very subjective due to inter- and intra-observer variations. At present, most diagnosis of cancer is still done by visual examination of radiological images, microscopy of biopsy specimens, direct observation and so on. These views are typically interpreted in a qualitative manner by clinicians trained to classify abnormal features such as structural irregularities. A more quantitative and reproducible approach for analyzing images is highly desired. Therefore, how to develop a more

objective computer-aided technique to automatically and correctly classify prostatic carcinoma is the goal of this research study. The aim here is to use automatic classifiers as a diagnosis aid along with human expertise by applying image processing and computer vision techniques to perform quantitative measurements of relevant features that can discriminate between different types of tissues that occur in biopsies. In the case of the prostate gland, four major classes of tissues have to be recognized and labeled by the pathologist (Figure 1 shows some samples of each class):

1. *stroma*: STR (normal muscular tissue);
2. *benign prostatic hyperplasia*: BPH (a benign condition);
3. *prostatic intraepithelial neoplasia*: PIN (a precursor state for cancer);
4. *prostatic carcinoma*: PCa (abnormal tissue development corresponding to cancer).

Numerous investigations have been carried out using different approaches such as morphology, texture analysis, and others for the classification of prostatic samples (Bartels et al., 1998; Clark et al., 1987). The Gleason grading system (Gleason & Tannenbaum, 1977) is a well known method. In this grading system, the prostate cancer can be classified into five tumor grades represented by a number ranging from 1 to 5 with five being the worst grade possible (O'Dowd et al., 2001). Tabech et al. proposed (Tabesh et al., 2005) an automatic two-stage system for prostate cancer diagnosis with the Gleason grading. The color, morphometric and texture features are extracted from prostate tissue images in their system. Then, linear and quadratic Gaussian classifiers were used to classify images into tumor/non tumor classes and further categorized into low/high grades for cancer images. Huang et al. proposed (Huang & Lee, 2009) two feature methods based on fractal dimension to analyze the variations of intensity and texture complexity in the regions of interest. Each image can be classified into an appropriate grade by using Bayesian, KNN, and support vector machine (SVM) classifiers, respectively. Leave on out and k fold cross-validation procedures were used to estimate the correct classification rates.

However, all these studies have been performed using a color space that is limited either to gray-level images, or to the standard RGB channels. In both cases, the color sampling process results in a loss of a considerable amount of spectral information, which may be extremely valuable in the classification process. High throughput liquid crystal tunable filters (LCTF) have recently been used in pathology, enabling a complete high resolution optical spectrum to be generated at every pixel of a microscope image. Studies suggest that multispectral images can capture relevant data not present in conventional RGB images. In (Liu et al., 2002) the authors used a large set of multispectral texture features for the detection of cervical cancer. In (Barshack et al., 1999), spectral morphometric characteristics were used on specimen of breast carcinoma cells stained with haematoxylin and eosin (H&E). Their analysis showed a correlation between specific patterns of spectra and different groups of breast carcinoma cells. Larsh et al. (Larsh et al., 2002) suggested that multispectral imaging can improve the analysis of pathological scenes by capturing patterns that are transparent both to the human eye and the standard RGB imaging.

In (Boucheron et al., 2007) Boucheron et al a comparison is performed between multispectral and RGB data for nuclei classification of breast tissue. Using SVM classifiers, the authors have concluded that multispectral bands do not contain much more discriminatory spectral information than the RGB bands for nuclei classification. However, the research was concerned with the classification of single pixels and it was limited to the classification of nuclei of histological breast images. Masood & Rajpoot (Masood & Rajpoot, 2008) present a study based on the comparison of two approaches: 3D spectral/spatial analysis

and 2D spatial analysis. They have compared the results using a textural analysis on single hyperspectral band against 3D spectral spatial analysis of histological colon images. However, the classification features were not extracted from multispectral data but rather from segmented 2D images obtained from multispectral data. Roula et al have described a novel approach, in which additional spectral data is used for the classification of prostate needle biopsies (Roula, 2002, 2003) which reduced overall error rate from 11.6% to 5.1%.

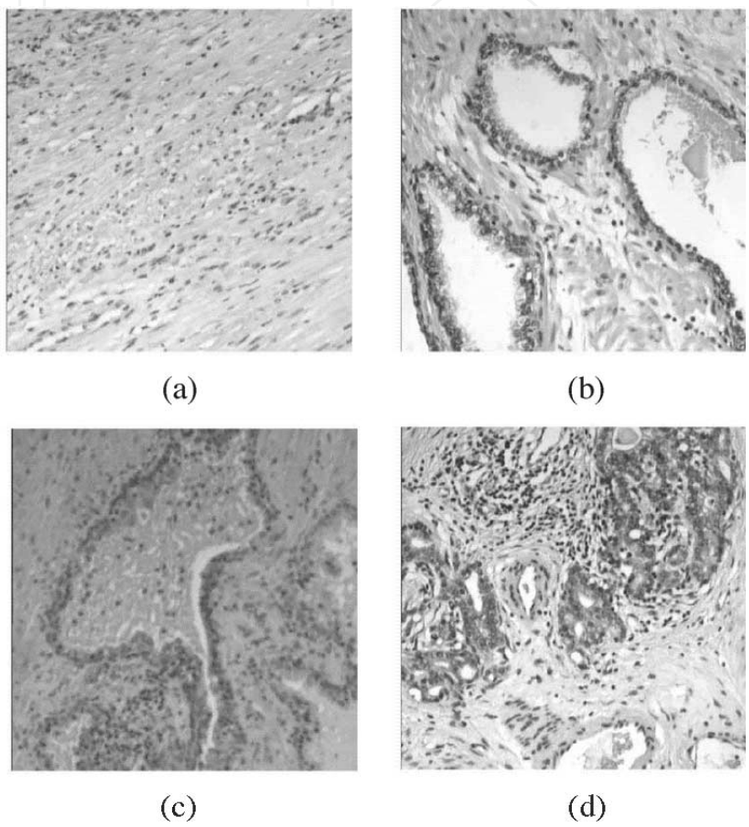


Fig. 1. Images showing representative samples of the four classes. (a) Stroma. (b) BPH. (c) PIN. (d) PCa.

The major problem arising in using multispectral data is high-dimensional feature vector size (> 100). The number of training samples used to design the classifier is small relative to the number of features. For such a high dimensionality problem, pattern recognition techniques suffer from the well-known curse-of-dimensionality (Jain et al., 2000): keeping the number of training samples limited and increasing the number of features will eventually result in badly performing classifiers. One way to overcome this problem is to reduce the dimensionality of the feature space. While a precise relationship between the number of training samples and the number of features is hard to establish, a combination of theoretical and empirical studies has suggested the following rule of thumb regarding the ratio of the sample size to dimensionality: the number of training samples per class should be greater than or equal to five times the features used (Dash & Liu, 1997). For example, if we have a feature vector of dimension 20, then we need at least 100 training samples per class to design a satisfactory classifier.

Another way to reduce the dimensionality of the feature space is by using feature selection methods. The term feature selection refers to the selection of the best subset of the input

feature set. These methods used in the design of pattern classifiers have three goals: (1) to reduce the cost of extracting the features, (2) to improve the classification accuracy, and (3) to improve the reliability of the estimation of the performance, since a reduced feature set requires less training samples in the training procedure of a pattern classifier (Jain et al., 2000). Feature selection produces savings in the measuring features (since some of the features are discarded) and the selected features retain their original physical interpretation.

In previous papers (Bouatmane et al., 2007), we addressed the high input dimensionality problem by selecting the best-subset of sequential forward selection SFS followed by a classification using a nearest neighbour classifier (1NN) technique. Although, this approach produced results superior to previously reported methods (Roula, 2002,2003) , the classification accuracy can be further improved by decomposing this multiclass problem into a number of simpler two-class problems. In this case, each subproblem can be regarded separately and solved using a suitable binary classifier. The outputs of this collection of classifiers can then be combined to produce the overall result for the original multiclass problem. In this paper, we propose a Round-Robin (RR) classification algorithm using a sequential forward selection/nearest neighbor (SFS/1NN) classifier to improve the classification accuracy. Round Robin classification is a technique which is suitable for use in multiclass problems. The technique consists of dividing the multiclass problem into an appropriate number of simpler binary classification problems (Furnkranz, 2002). Each binary classifier is implemented as an SFS/1NN classifier, and the final outcome is computed using majority voting technique. A key characteristic of this approach is that, in a binary class, the classifier attempts to find the features that only distinguish that particular class. Thus, different features are selected for each binary classifier, resulting in an overall increase in the classification accuracy. In contrast, in a multiclass problem, the classifier tries to find those features that distinguish all classes at once.

The remainder of this chapter is organized as follows: Sect. 2 gives a description of the dataset used including texture and structural features. Section 3 is concerned with feature selection problem and describes the RR approach followed by the probability estimate for the classifier outputs and error estimation. Sect. 4 describes the image acquisition and dataset. Sect. 5 gives the results obtained and their analysis and discussion including a performance comparative study. Sect. 6 analyses the features selected and sect. 7 gives ROC curves and finally sect. 8 gives a summary of the chapter.

2. Images features

Over the last years, the most prolific and promising works in the area of cancer classification have been in the area of texture analysis of the nucleus (Tabesh et al., 2005; Liu et al., 20). This is not surprising since pre-cancerous abnormalities are manifested in visual and subvisual changes in cell characteristics. In fact, it is generally believed that the initial signs of cell neoplasia appear in the nucleus. Because nuclear chromatin and its spatial arrangement can be viewed as a type of texture and whether tissue samples are examined at low, medium or high magnification, texture is a key element in the differentiation between normal and malignant tissue patterns.

However, texture features are not sufficient to classify all the groups. The complex structures present in BPH, PIN and also PCa need a higher level description. Thus, structural features, based on segmentation, have been computed for different spectral bands and consolidated in a large feature vector. The features used are described in the following subsections.

2.1 Texture feature

To identify prostatic patterns, texture features are needed as a discriminative measurement for the samples. Haralick (Haralick, 1979) assumed that texture information is sufficiently identified by a matrix indexed by grey levels and where the elements represent the frequency of having two defined grey levels separated by a defined distance in a defined direction. This matrix is called grey level co-occurrence matrix (GLCM):

$$co(i, j, d, \theta) = \alpha \tag{1}$$

The above equation means that there are α pairs of pixels having i and j respectively, as grey levels and separated by the cylindrical co-ordinate $[d, \theta]$. The values of d, for which the GLCM is computed, depend on the nature of the texture. Small d values are suitable for fine textures, whereas larger distances are needed to measure coarse textures.

For an image of 256 grey levels (Ng=256), there would be 65536 feature elements to use as a measure for the texture. Therefore, the direct use of the co-occurrence matrix is computationally intensive and as such is not practical. Instead, the texture features are represented by deriving some more meaningful measurements. A set of features was proposed by Haralick to characterise the homogeneity, the coarseness, the periodicity and the linearity of textures. These features are defined as follows:

Angular Second Moment

$$ASM = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i, j)^2 \tag{2}$$

Contrast or difference moment

$$CON = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - j)^2 P(i, j) \tag{3}$$

Dissimilarity

$$DIS = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} |i - j| P(i, j) \tag{4}$$

Correlation

$$COR = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left[(i - \mu_x)(j - \mu_y) P(i, j)^2 \right] / (\sigma_x \sigma_y) \tag{5}$$

Where $\mu_x, \mu_y, \sigma_x, \sigma_y$ are the means and the variances of the row sums and column sums of the co-occurrence matrix respectively.

Entropy or randomness

$$ENT = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i, j) \log P(i, j) \tag{6}$$

Inverse difference Moment:

$$IDM = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{P(i, j)}{(1 + (i - j)^2)} \tag{7}$$

2.2 Structural features

The use of texture features alone is not sufficient to capture the complexity of the patterns in prostatic neoplasia. Although, the classification of stroma is relatively simple because of its homogenous nature at low resolution, BPH and PCa present more complex structures, as both can contain glandular areas and nuclei clusters as well. The glandular areas are smaller in regions exhibiting PCa while the nuclear clusters are much larger. The PIN pattern is an intermediate state between the BPH and PCa. It appears that accurate classification requires the quantification of these differences. Segmenting the glandular and the nuclear areas can achieve this quantification, as the glandular areas are lighter compared to the surrounding tissue, while the nuclear clusters are darker (Larsh et al., 2002; Roula et al., 2002).
Figure 2 summarises the segmentation scheme. From the segmented images 1 and 2, two features, f1 and f2 can be computed

$$f_1 = N / W^2 \tag{8}$$

$$f_2 = G / W^2 \tag{9}$$

Where G and N are the number of pixels segmented as glandular area and classified as nuclear area, respectively. W is the size of the analysis window. These two features allow the quantification of how much nuclear clusters and glandular areas are present in the samples.

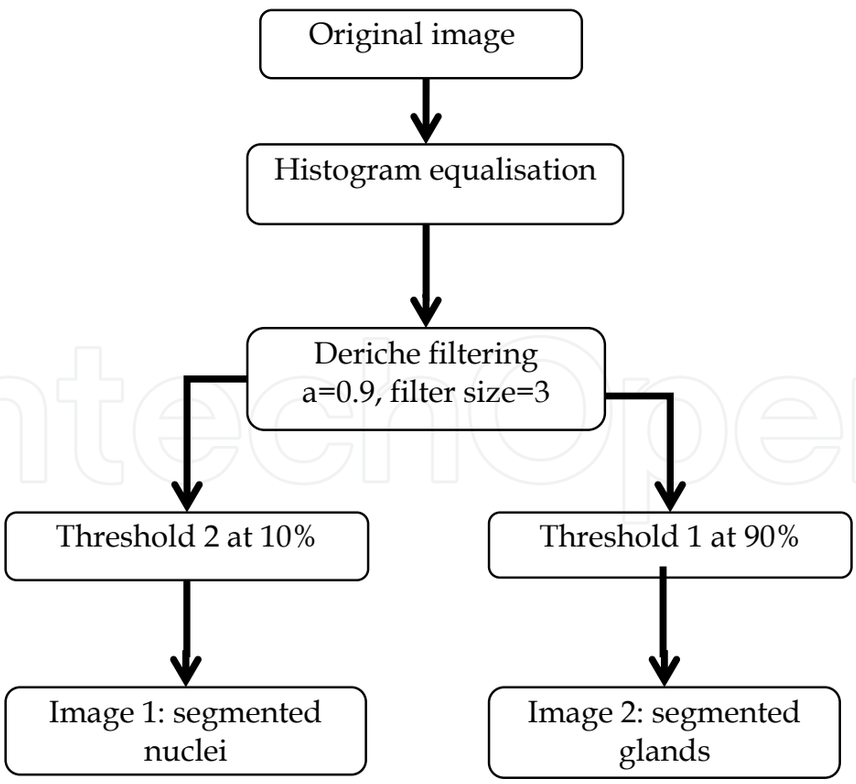


Fig. 2. Segmentation of nuclei and glandular areas

3. Classification of prostate cancer using round robin approach

3.1 Feature selection problem

As discussed in Section 1, the major problem arising from multispectral data is related to the feature vector size. Typically, with 16 bands and 8 features in each band, the feature vector size is 128. For such a high dimensionality problem, pattern recognition techniques suffer from the well-known curse-of-dimensionality problem: keeping the number of training samples limited and increasing the number of features will eventually result in badly performing classifiers (Jain et al., 2000; Jimenez, & Landgrebe, 1998).

PCA (a well-known unsupervised feature extraction method) has been used by Roula et al. on the large resulting feature vectors to reduce its dimensionality to a manageable size. In their work, Roula et al. used PCA and a linear discrimination function on significant PCA components for the classification.

Another technique to reduce the dimensionality of the feature space is by using feature selection methods. The term feature selection refers to the selection of the best subset of the input feature set. This results in a feature selection producing a smaller set of features (since some of the features are discarded) with the selected features retaining their original physical interpretation. This feature selection problem can be viewed as a multiobjective optimization problem since it involves minimizing the feature subset while maximizing classification accuracy.

Mathematically, the feature selection problem can be formulated as follows: Suppose Y is an original feature vector with cardinality n, $X \subseteq Y$, J(X) is the selection criterion function for the new feature vector X. The goal is to optimize J(X). The choice of an algorithm for selecting the features from an initial set depends on n. The feature selection problem is said to be of small scale, medium scale, or large scale accordingly as n belongs to the intervals [0,19], [20,49], or $[50,+\infty]$, respectively (Duda et al., 2001; Kudo & Sklansky, 2000).

Generally, feature selection algorithms have two components: a selection algorithm that generates proposed subsets of features and attempts to find an optimal subset; and an evaluation algorithm which determines how 'good' a proposed feature subset is, by returning some measure of goodness to the selection algorithm. However, without a suitable stopping criterion the feature selection process may run exhaustively or forever through the space of subsets. Stopping criteria can be: (i) whether addition (or deletion) of any feature does not produce a better subset; and (ii) whether an optimal subset according to some evaluation function is obtained. Ideally, a feature selection method searchs through the subsets of features, and tries to find the best one among all the competing candidate subsets according to some evaluation function. However, this procedure is exhaustive as it tries to find only the best one. It may be too costly and practically prohibitive, even for a medium-sized feature set size. Other methods based on heuristic or random search methods; attempt to reduce computational complexity by compromising performance (Davies & Russell, 1994).

In (Dash & Liu, 1997) different feature selection methods are categorized into two broad groups (i.e., filter and wrapper) depending on the type of classification algorithm used for the selection of the subset. For example, the filter methods do not require a feedback from the classifier and estimate the classification performance by some indirect assessments, such as distance measures which reflect how well the classes separate from each other. On the other hand, the wrapper methods are classifier-dependent. Based on the classification accuracy, the methods evaluate the "goodness" of the selected feature subset directly, which should intuitively lead to a better performance. Currently, many experimental results reported so far use the wrapper methods.

In this work, an SFS algorithm, which is simple and empirically successful, is proposed for feature selection. It starts with an empty subset of features and performs a hill-climbing deterministic search. At each iteration, a feature not yet selected is individually incorporated in the subset to calculate a criterion. Then the feature which yields the best criterion value is included in the new subset. This iteration will not be stopped until no improvement of the criterion value is achieved. SFS is used as a wrapper approach, therefore the criterion employed to carry out the search is based on error estimation by the selected features using 1NN classifier. In addition, we propose another scheme in which the multiclass problem is addressed using Round Robin (RR) classification approach where the classification problem is decomposed into a number of binary classes. The key point is that it is then possible to design simpler and more efficient binary classifiers as will be demonstrated in the next Section.

3.2 Round robin method

The RR or pairwise class binarization transforms a c-class problem into $c(c-1)/2$ two-class problems i, j with one for each set of classes i, j ($i = 1, \ldots, c-1$, $j = i+1, \ldots, c$). A binary classifier for problem i, j is trained with examples of classes i and j, whereas examples of classes $k \neq i, j$ are ignored for this problem (Furnkranz, 2002). Figure 3 illustrates a multiclass (four-class) learning problem where one classifier (SFS/1NN classifier in this study) separates all classes. Figure 4 shows Round Robin learning with $c(c-1)/2$ classifiers. For a four-class problem, the Round Robin trains six classifiers, one for each pair of classes. Each class is trained using a feature selection algorithm based on the SFS/1NN classifier.

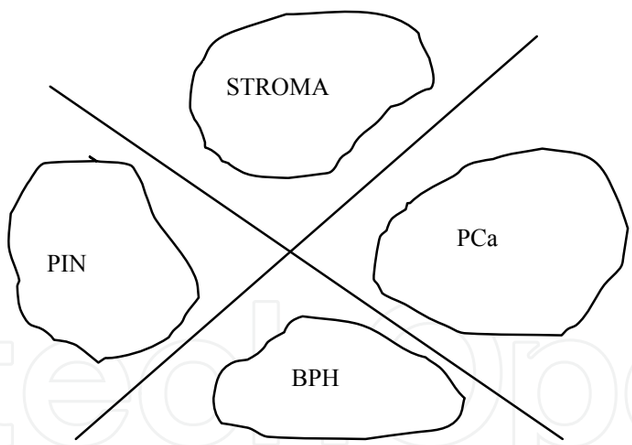


Fig. 3. Multiclass learning

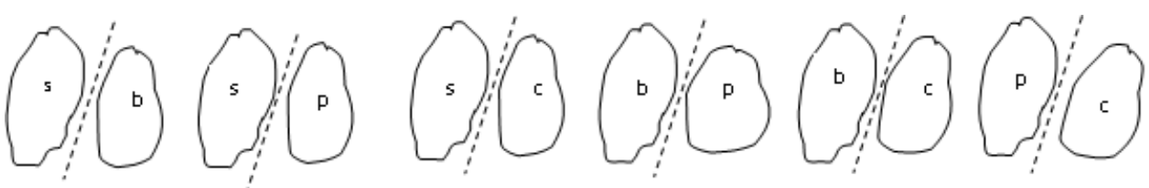


Fig. 4. Round Robin learning. p: PIN. c: PCa. b: BPH. s: STR.

The objects are then classified by applying a combination rule on the set of decisions. One strategy is to use voting where the object is labeled to the class with the highest number of votes. When classifying an unknown new sample, each classifier (1NN in this case) determines to which of its two classes the sample is more likely to belong. In this case, we are faced with the possibility of ties. To avoid these ties, a probability estimate value for each classification has to be used.

In pattern recognition, 1NN is one of the simplest and most widely used algorithms. Given a query sample x, a 1NN algorithm determines the closest neighbor of x in the training nodes using some distance metric (e.g. Euclidean distance in our study) and predicts the class label of the nearest node. In contrast to other statistical classifiers, 1NN needs no model to fit. This property simplifies the structure of the training process by avoiding model training, thus training with a 1NN classifier only requires selecting the appropriate features.

For the sake of probability estimates, probabilistic outputs of the classifier were required rather than label prediction. For 1NN, objects are assigned to the class of the nearest object in the training set. Posterior probabilities are estimated by comparing the nearest neighbor distances for all classes (Duin & Tax, 1998). A RR ensemble converts a c-class problem into a series of two-class problems by creating one classifier for each pair of classes. New items are classified by submitting them to the c(c-1)/2 binary predictors. The final prediction is achieved by a majority voting. The probability of a query q belonging to a class c can be calculated as follows (Grimaldi et al., 2003) (equation 10):

$$p(c|q) = \frac{\sum_{m \in M} p_m(c|q).1_{(mc=c)}}{\sum_{m \in M} p_m(c|q)} \tag{10}$$

where M is the set of ensemble members, mc is the class predicted by m and Pm(c|q) is the posterior probability given by ensemble predictor m (the binary classifiers). If m does not involve class c, then Pm(c|q) = 0. The probability estimates of the binary classifiers will be combined using the maximum rule; therefore the instances are assigned to the class with the maximum output given by equation (10). Clearly, RR is a problem decomposition technique. However, there are some aggregation benefits as each class is focused on by c-1 classifiers.

3.3 Error estimation

Given a small set of samples, appropriate strategies for learning and testing become very critical to avoid over-fitting. Leave-one-out (LOO) and k-fold cross-validation are two popular error estimation procedures to reduce bias in machine learning and testing problems especially with small sample size (sss) (Jain et al., 2000). The procedure of LOO method is to take one out of n observations and use the remaining n-1 observations as the training set for deriving the parameters of the classifier. The classifier is then used to classify the removed observation. This process is repeated for all n observations in order to obtain the estimation of the classification accuracy. In the case of k-fold cross-validation method, the entire sample set is randomly partitioned into k disjoint subsets of equal size, where n is the total number of samples in the entire set. Then, k -1 subsets are used to train the classifier and the remaining subset is used to test for accuracy estimation. This process is repeated for all distinct choices of k subsets and the average of correct classification rates is calculated. Notice that k -fold cross-validation is reduced to LOO if k=n.

When referring to the performance of a classification model, we are interested in the model's ability to correctly predict or separate the classes. When looking at the errors made by a classification model, the confusion matrix used in this paper gives the full picture. The confusion matrix shows how accurate the predictions are made by the model. The rows correspond to the known class of the data, i.e. the labels in the data while the columns correspond to the predictions made by the model. The value of each of element in the matrix is the number of predictions made with the class corresponding to the column, for example, with the correct value as represented by the row. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made.
Accuracy is the overall correctness of the model and is calculated as the sum of the correct classifications divided by the total number of classifications. Precision is a measure of the accuracy provided that a specific class has been predicted. It is defined by:

$$Precision = \frac{tp}{tp + fp} \tag{11}$$

where tp and fp are the numbers of true positive and false positive predictions for the considered class. Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is also commonly called sensitivity, and corresponds to the true positive rate and can be written as:

$$Recall = Sensitivity = \frac{tp}{tp + fn} \tag{12}$$

where tp and fn are the numbers of true positive and false negative predictions for the considered class. tp+ fn is the total number of test examples of the considered class.

4. Sample preparation, image acquisition and datasets description

Entire tissue samples were taken from prostate glands. Sections 5-μm thick were extracted and stained using the widely used H&E stains. These samples were routinely assessed by two experienced pathologists and graded histologically as showing STR, BPH, PIN, and PCa. From these samples, whole subimage sections were captured using a classical microscope and CCD camera. An LCTF (VARISPECTM) was inserted in the optical path between the light source and a CCD camera. The LCTF has a bandwidth accuracy of 5 nm. The wavelength is controllable through the visible spectrum (from 400 to 720 nm). This allowed for the capture of multispectral images of the tissue samples by using different spectral frequencies. Figure 5 shows a prostatic tissue sample viewed at different magnification.
In order to offset any bias due to the different range of values for the original features, the input feature values are normalized over the range [1,11] using equation (13) (Raymer et al., 2000). Normalizing the data is important to ensure that the distance measure allocates equal weight to each variable. Without normalization, the variable with the largest scale will dominate the measure:

$$x'_{i,j} = \left(\frac{x_{i,j} - \min_{k=1\ldots n} x(k,j)}{\max_{k=1\ldots n} x(k,j) - \min_{k=1\ldots n} x(k,j)} \times 10 \right) + 1 \tag{13}$$

where x_j^i is the jth feature of the ith pattern, $x_{i,j}^t$ is the corresponding normalized feature and n is the total number of patterns.

The data were taken from a total of 10 different patients with typically 3-6 biopsies per patient (from different areas in the prostate) and 8-12 images were taken from each samples (from different areas in the image). The dataset consists of textured multispectral images taken at 16 spectral channels (from 500 to 650 nm) (Roula et al., 2002). Five hundred and ninety-two different samples (multispectral images) of size 128 × 128 have been used to carry out the analysis. The samples are examined at low power (40 x objective magnifications) by the two highly experienced independent pathologists and labelled into four classes: 165 cases of Stroma, 106 cases of BPH, 144 cases of PIN, and 177 cases of PCa.

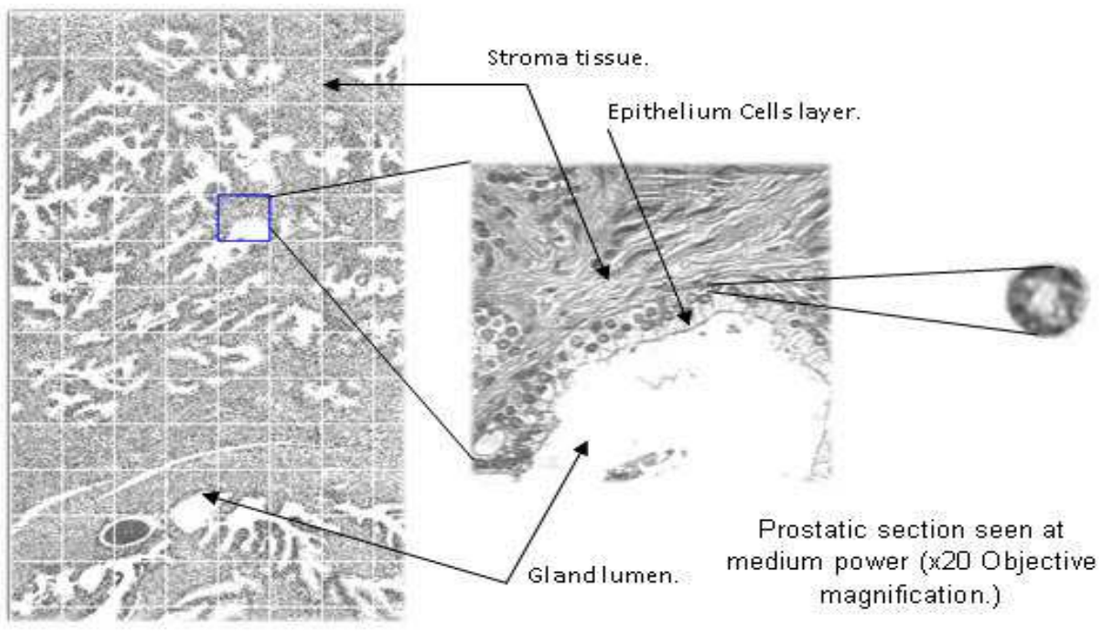


Fig. 5. Prostatic tissue sample viewed at low and medium magnifications

5. Experiments and discussion

The assessment of the classification performance has been made using three procedures: 4-fold cross-validation, 10 cross-validation and leave-one-out (LOO) which was applied patient-wise. To obtain a k-fold cross-validation estimate of the classification performance, the dataset was randomly split into k sets of a roughly equal size. Splitting was carried out such that the proportion of samples per class was roughly equal across the sets. Each run of the k-fold cross-validation algorithm consisted of a classifier design on k-1 dataset subsets (training) while testing was performed on the remaining subset. The optimal feature subset for each cross-validation run was determined as the subset with the highest LOO accuracy estimate on the corresponding training set.

The first aim was to determine the optimum number of features to obtain the best achievable classification performance. Therefore, the feature selection algorithm SFS described in section 3 with a 1NN classifier was used. Figure 6 shows the results obtained using LOO error estimation. The curve representing the results from the feature selection shows a strong increase in performance for small subsets followed by slight increase up to medium sized subsets. Large subsets cause a drop in the recognition rate.
For k-fold cross-validation the results show that using SFS with different training sets does not yield identical feature subsets. This is illustrated by the diagram in Figure 7 which shows the fraction of how often a feature was selected divided by the total number of simulations using 4 cross-validation method. One can see that the selected features originate from different spectral bands.

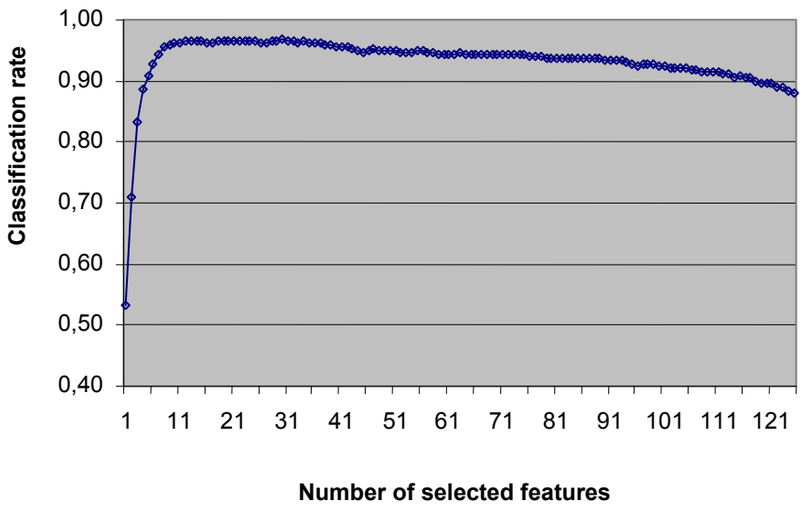


Fig. 6. Recognition rate of SFS algorithm

The accuracies of the selected features subsets are given in Table 1. The combination of the binary classifiers' results generated by proposed Round Robin algorithm is performed using two methods: the voting rule using the resulted classes (Tahir & Bouridane, 2006) and the maximum probability obtained using equation (10). For all the cross validation estimations, the RR SFS/1NN with the maximum probability gives the best classification accuracy. As shown in Table 1, RRSFS algorithm using LOO error estimation achieves the lowest error rate. The overall classification error has been reduced from 3.37% to 0.17%. To gain an insight into the classification of different classes of prostate cancer, the confusion matrix of the multiclass SFS/1NN and the proposed Round Robin learning using SFS/1NN are also given. Table 2 depicts the results using the LOO error estimation where Table 3 gives the corresponding results using 4 cross-validations. Note that in all the cases, BPH and PIN classes present the highest error rate in terms of classification but the use of Round Robin algorithm reduces significantly the error rate in these classes.

Bagging is a general method of combining classifiers that can be applied to any base method. It is a relatively simple idea: n datasets are created by sampling the patterns with replacement from the original training set. Each of the n datasets has the same number of patterns as the original training set. A classifier is then trained on each dataset by combining the outputs using simple voting. Bagging has obtained impressive error reductions with decision trees such as CART (Breiman, 1996) and C4.5 (Freund & Schapire, 1996; Quinlan, 1996) on a wide range of datasets.

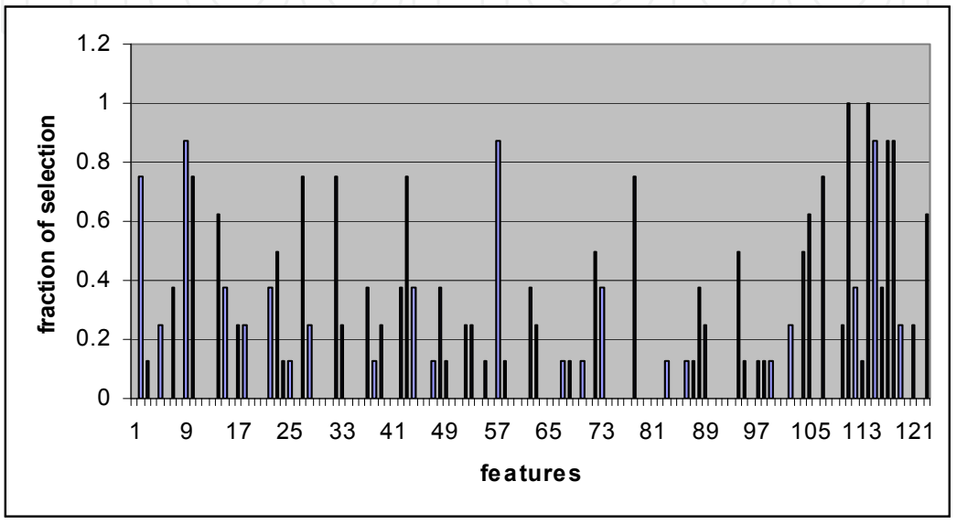


Fig. 7. Subsets yielded by application of the SFS from 4 cross-validations

	4 cross-validation error estimation %	10 cross validation error estimation %	Leave-one-out error estimation %
1NN classifier	13.34	12.18	12
SFS/1NN classifier	10.22	7.41	3.37
Round robin SFS/1NN (voting rule)	10.98	9.62	2.87
Round robin SFS/1NN (maximum probability rule)	8.91	7.26	0.17

Table 1. Comparison of error classification rate

In Boosting, the classifiers in the ensemble are trained serially, with the weights on the training instances set adaptively according to the performance of the previous classifiers. If

the classifier does not directly support weighted instances, this can be simulated by sampling from the training set with a probability proportional to an instance weight. The main idea is that the classification algorithm should concentrate on the difficult instances.

SFS/1NN multiclass learning						Round Robin SFS/1NN learning				
Classified as:	BPH	PCa	PIN	Stroma	Error (%)	BPH	PCa	PIN	Stroma	Error (%)
BPH	101	0	0	5	4.71	106	0	0	0	0
PCa	1	174	2	0	1.69	0	177	0	0	0
PIN	0	2	137	5	4.86	0	0	143	1	0.69
Stroma	5	0	0	160	3.03	0	0	0	165	0
overall					3.37					0.17

Table 2. Classification Error by multiclass and round robin learning using SFS/1NN and loo error estimation

Round Robin SFS/1NN learning						SFS/1NN multiclass learning				
Classified as:	BPH	PCa	PIN	Stroma	Error (%)	BPH	PCa	PIN	Stroma	Error (%)
BPH	96	0	0	10	9.43	93	3	2	8	12.26
PCa	1	164	8	4	7.43	2	163	11	1	7.90
PIN	0	13	129	2	10.41	3	8	122	11	15.27
Stroma	8	1	5	151	8.48	5	1	3	156	5.45
overall					8.91					10.22

Table 3. Classification error by multiclass and round robin learning using SFS/1NN and 4cross-validation

C4.5			Nearest neighbor			
C4.5	Bagging	Boosting	NN	Bagging	Boosting	RR-SFS
91.6	93.2	95.4	88.0	89.2	88.1	99.83

Table 4. Classification accuracy (%) using various ensemble techniques

Table 4 shows the comparison between the RR-SFS/1NN versus Bagging and AdaBoost. Decision Tree (C4.5) (Quinlan, 1993) and NN classifiers are used as base classifiers for bagging and boosting.

Unfortunately, bagging and boosting are unable to improve the classification accuracy when an NN classifier is used as a base classifier (Yongguang et al., 2004). This fact is clearly seen from Table 6 where the classification accuracy is degraded while using AdaBoost, and only minor improvements are achieved when using bagging. However, the classification accuracy is improved by using bagging and boosting when C4.5 is used as base classifier. Furthermore, it is clear from the table that the proposed Round Robin ensemble technique using TS/1NN has outperformed both bagging and boosting ensemble-design techniques.

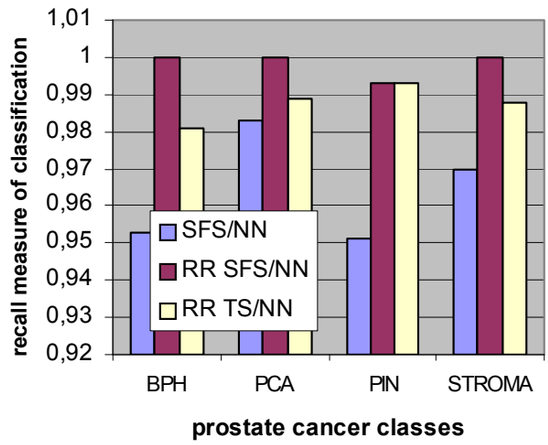
A key characteristic of the proposed Round Robin approach is that different features are captured and used for each binary classifier in the four-class problem, thus producing an overall increase in the classification accuracy. In contrast, in a multiclass problem, the classifier tries to find those features that distinguish all four-classes at once. Furthermore, the inherent curse-of-dimensionality problem, which arises in a multispectral data, is also resolved by the RR SFS/1NN classifiers since each classifier is trained to compute and use only those features that distinguish its own binary classes.

Table 5 shows the number of features used by the ensemble of binary classifiers. Different numbers of features have been used by the various binary classifiers producing an overall increase in the classification accuracy. Fc represents those features that are common in two or more different binary classifiers. The total number of features in the proposed Round Robin technique is comparable with the multiclass SFS/1NN with lower error rate, but the number of features used by each binary classifier is smaller than that used in other methods. Consequently, multispectral data is better utilized by using a Round Robin technique since the use of more features means more information is captured and used in the classification process. Furthermore, simple binary classes are also useful for analyzing features and are extremely helpful for pathologists in distinguishing various patterns such as BPH, PIN, STR, and PCa.

	Feature selection method		Features used
	SFS/1NN	Multi-class	13
1	SFS/1NN	Binary-class (stroma-Bph)	4
2		Binary-class (Stroma-Pin)	1
3		Binary-class (stroma-PCa)	4
4		Binary-class (Bph-Pin)	1
5		Binary-class (Bph-PCa)	1
6		Binary-class (Pin-PCa)	4
	SFS/1NN	Round Robin	$\sum_{i=1}^{6} F - F_c = (15-3) = 12$

Table 5. Number of Features Used By Different Classifiers

Figure 8 shows the results of Recall and Precision measures for different algorithms including the results of Round Robin tabu search RR TS/1NN (Tahir & Bouridane, 2006). From the graphs presented one can observe that for both Precision and Recall, the values of RR SFS/1NN are very high for different classes of prostate cancer. In addition, one can notice from equations (11) and (12) that the values for FP and FN tend to zero when the Precision and Recall tend to 100%. Thus, the false positives and especially false negatives are almost null with our approach. This clearly demonstrates the efficiency of our proposed RR technique.



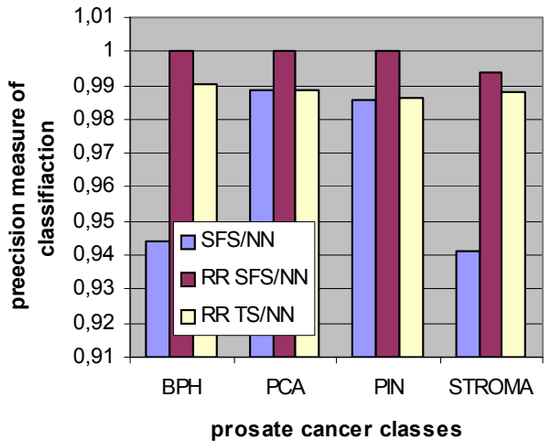


Fig. 8. Precision and recall measures of classification

6. Analysis of the selected features

Very often, it is interesting to know if the difference in the mean values for a given feature between two groups is accidental or due to an inherent difference between the groups regarding a specific feature. For example, the mean of a given image feature can be numerically different for normal and cancer prostate cells. But does this difference reflect a

real physical dissimilarity between the two groups or is it due to those specific samples? And in the case where it is a real physical difference, what is the level of confidence when making such statements?

In this work, Student t-test (Montgomery, 1997) was used as a statistical test of significance for mean difference of each class pair for all the selected features. Table 6 and 7 show the selected features for SFS/1NN classifier and RR SFS/1NN method using LOO, respectively. The asterisk (*) in the table shows that this feature exhibits a significant difference in means for all group pairs of cancer (Stroma, BPH, PIN, PCa) with 95 % confidence (p-value<0.05) while (**) shows confidence in difference of means higher than 99% (p-value<0.01). For the round robin method, t-test was run only for the binary classes.

In Table 6, for 9 out of the 13 features selected, the p-value exhibits values lower than 0.05, i.e. yields confidence levels in difference between groups >95%. Three features exhibit a confidence level in the mean difference superior to 99%.

It was observed that dissimilarity, inverse difference moment, entropy and contrast are the texture features selected. They are all measures of homogeneity of grey level texture. This indicates that prostatic tissues display a clear visual difference in terms of texture. Consequently, neighboring pixels were more likely to have larger grey level differences for different grades of malignancy. Note that the features which have not asterisks exhibit significant difference in means, but not for all the pairs of classes.

Rank	Selected features	Spectral band
1	Inverse difference moment *	13
2	structural (f2) *	9
3	Structural (f1)	8
4	Dissimilarity *	3
5	Contrast	7
6	Structural (f1) **	7
7	Structural (f2)	5
8	Dissimilarity	11
9	Contrast *	4
10	Entropy **	8
11	Contrast **	6
12	Inverse difference moment	8
13	Structural (f1) *	11

Table 6. Selected features by SFS/1NN Classifier

Binary classifiers	Selected features	Spectral band
(Stroma-Bph) classifier	Structural (f2)**	11
	Dissimilarity **	15
	Dissimilarity **	13
	Angular second moment **	13
(Stroma-Pin) classifier	Contrast **	15
(Stroma-PCa) classifier	Structural (f2) **	9
	Structural (f1)**	3
	Inverse difference moment **	5
	Dissimilarity **	4
(Bph-Pin) classifier	Structural (f1)**	14
(Bph-PCa) classifier	Structural (f1)**	14
(Pin-PCa) classifier	Inverse difference moment **	10
	Contrast **	2
	Dissimilarity **	4
	Structural (f2)**	9

Table 7. Selected features By RR-SFS/1NN Classifier

For the RR SFS/1NN method, all the features presented in Table 7 exhibit a confidence level in mean difference superior to 99% for all the binary classes. This can be explained by the fact that the RRSFS/1NN method selects the features that distinguish only that class. In contrast, in multiclass SFS/1NN, the classifier tries to find those features that distinguish all classes at once.

The presence of structural features can be observed, especially to discriminate BPH and PCa from the other classes. This is because BPH is first characterised by a conspicuous glandular presence. For the PCa, this is due to the predominance of nuclei clusters and the total absence of glands, which makes it easy to detect using the structural features. We note also that the texture features selected are all measures of homogeneity. Contrast is selected alone by Stroma-Pin classifier without the structural features since the glands are totally absent from Stroma. PIN, which is an intermediate state between PCa and BPH may or may not contain lumen glands. Correlation is totally absent from the two tables thus indicating that correlation is a poor discriminant feature. It can be concluded that the joint use of texture and structural features is an efficient method to classify all groups together.

Finally, it is important to see the impact of the multispectral dimension on the classification; the features selected in both methods are from different bands. This shows that the satisfactory results obtained previously are not only due to the adequate choice of features but to the contribution of the multispectral information which characterizes the different classes.

7. ROC curve

ROC curve (receiving operating characteristic) analysis has been widely used as a method for medical decisions making. It is a plot of false positive rate (X-axis) versus true positive rate (Y-axis) of a binary classifier. ROC is commonly used for visualizing and selecting classifiers based on their performance. The true positive rate (TPR) is defined as the ratio of the number of correctly classified positive cases to the total number of positive cases. The false positive rate (FPR) is defined as the ratio of incorrectly classified negative cases to the total number of negative cases (Fawcett, 2003).

ROC curves help researchers focus on classification rules with low false positive rates, which are most important for early detection of cancer.

The diagonal line y = x corresponds to a classifier which predicts a class membership by randomly guessing it. Hence, all useful classifiers must have ROC curves above this line.

We assume that one of the classes is the class of interest and the objects labeled in this class will be called 'positive'. This achieved by considering the BPH as the negative diagnosis while Pca and PIN form the positive diagnosis outcome.

The classifier gives a continuous valued output given by equation (10) which is cut at a certain threshold. All objects for which the classifier output exceeds the threshold are labeled as positive while the remaining results are labeled as negative. By varying the threshold value from the minimum to the maximum value of the classifier output, one can construct a ROC curve. Figure 9 illustrates the ROC curves obtained with the two methods RRSFS/1NN and SFS/1NN using 4 cross-validation and an independent test set. The test set is obtained by splitting the dataset onto two equal sets, training set and test set. For cross-validation, given the test sets generated from 4 cross-validation, we can simply merge the instances together by their assigned scores into one large test set and we then plot the result.

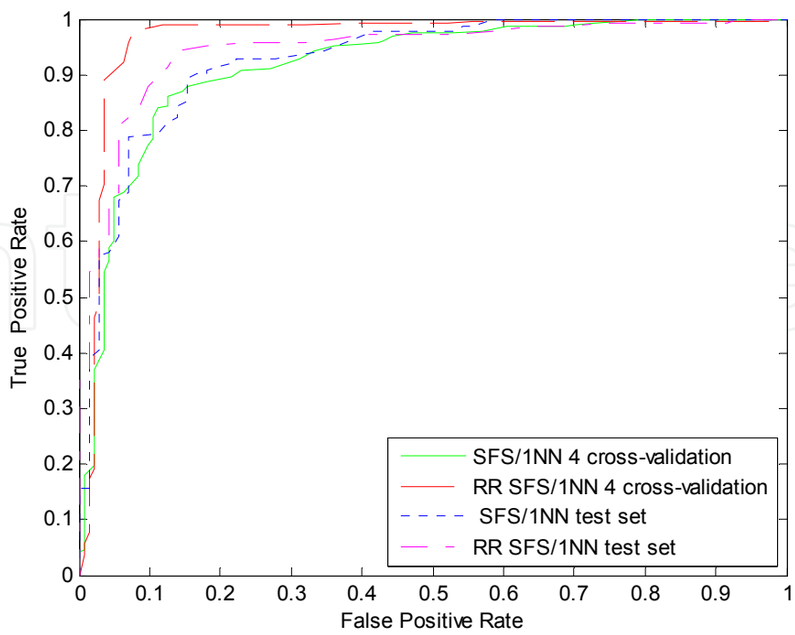


Fig. 9. ROC curves for SFS/1NN and RR-SFS/1NN classifier

The results are comparable or better than those obtained in other recent studies (Taher & Bouridane, 2006); this further demonstrates that our new proposed Round Robin technique results in an improved ability to distinguish cancer prostate tissues from healthy ones. It is clear from the figure that RRSFS/1NN algorithm performs better than simple SFS/1NN with high TPR rate.

8. Conclusion

In this chapter, a Round Robin SFS/1NN algorithm is proposed for the classification of prostate needle biopsies using multispectral imagery. To achieve this, a set of features was computed over a wide range of visible wavelength and the results have indicated a significant increase in the classification accuracy with Round Robin technique with high TPR. A key characteristic of the proposed Round Robin approach is that different features are used for each binary classifier from multispectral images, thus producing an overall increase in the classification accuracy. In contrast, in a multiclass problem, the classifier tries to find only those features that distinguish all classes at once. RR SFS/1NN has also demonstrated the effectiveness of some texture and structural features to make difference between different classes which can be helpful for the pathologist. Finally, the algorithm is generic and can be used for different datasets from other pattern recognition areas.

9. References

Barshack, I.; Kopolovic, J.; Malik, Z. & Rothmann, C. (1999)Spectral morphometric characterization of breast carcinoma cells. *Brit. J. Cancer*, vol. 79, no. 9–10, pp. 1613–1619.

Bartels, P. H et al. (1998). Nuclear chromatin texture in prostatic lesions: IPIN and adenocarcinoma. *Anal. Quant. Cytol. Histol*., vol. 20, no. 15,pp. 389–396.

Bouatmane, S.; Nekhoul, B.; Bouridane, A. & Tanougast C. (2007). Classification of Prostatic Tissues using Feature Selection Methods. *IFMBE Proceedings*, vol 16, pp 843-846.

Boucheron, L. ;Bi Z.; Harvey, N.; Manjunath, B. & Rimm, D. (2007). Utility of multispectral imaging for nuclear classification of routine clinical histopathology imagery. *BMC Cell Biology*, 8(Suppl. 1):S8.

Breiman, L. (1996). Bagging predictors. *Machine Learning,* vol. 24 , 123-140.

Clark T. D.; Askin, F. B. & Bagnell, C. R (1987). Nuclear roundness factor: A quantitative approach to grading in prostate carcinoma, reliability of needle biopsy tissue, and the effect of tumor stage on usefulness. *Prostate*. vol. 10, no. 3, pp. 199–206.

Dash, M. Liu, H. (1997). Feature Selection for Classification. *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131- 156.

Davies, S. & Russell, S. (1994). NP-completeness of searches for smallest possible feature sets. *Proc. AAAI Fall Symp Relevance*, pp.37–39.

Duda, R. O. Hart, P. E. & Stork, D.G. (2001). *Pattern Classification*. Hoboken, NJ: Wiley-Interscience.

Duin, R.P.W. & Tax, D.M.J. (1998). Classifier conditional posterior probabilities. *Lecture Notes in Computer Science*, vol. 1451, Springer, Berlin, 611-619.

Fawcett, T. (2003). ROC graphs: Notes and practical considerations for researchers. *Tech Report HPL-2003-4, HP Laboratories*.

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. Machine Learning, Proceedings *of the Thirteenth International Conference*, pp. 325-332.

Furnkranz, J. (2002). Round robin classification. *J. Mach. Learn. Res*., vol. 2,pp. 721–747.

Gleason, D. F. & Tannenbaum, M. (1977). The veteran's administration cooperative urologic research group: Histologic grading and clinical staging of prostatic carcinoma, in *Urologic Pathology: The Prostate*. Philadephia, PA: Lea Febiger, , pp. 171–198.

Grimaldi, M.; Cunningham, P. & Kokaram, A. (2003). An evaluation of alternative feature selection strategies and ensemble techniques of classifying music. *Workshop in Multimedia Discovery and Mining* at ECML/PKDD.

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proc. Of the IEEE,* vol. 67, pp.786-804.

Huang, P.W. & Lee, C.H. (2009) automatic classification for pathological prostate images based on fractal analysis, *IEEE transactions on medical imaging*, VOL. 28, NO. 7, pp.1037–1050.

Jain, A. K.; Duin, R. P. W. & Mao, J. *(2000). Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Mach. Intell*., vol. 22, no. 1, pp. 4–37.

Jimenez, L. O. & Landgrebe, D. A. (1998). Supervised classification in high dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev*., vol. 28, no. 1, pp. 39–54.

Kronz, J. D; Westra, W. H & Epstein, J. I. (1999). Mandatory second opinion Surgical Pathology at a Large Referral Hospital, Cancer, vol. 86, no. 11, pp. 2426-2435.

Kudo, M. & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognit*., vol. 33, pp. 25–41.

Larsh, P.; Cheriboga, L. Yee, H. & Diem, M. (2002). Infrared spectroscopy of humans cells and tissue: Detection of disease. *Technol*. *Cancer Res. Treat*., vol. 1, no. 1, pp. 1–7.

Liu, Y.; Zahoa, T. & Zhang, J. (2002) Learning multispectral texture features for cervical cancer detection. *Proc IEEE Int. Symp. Biomed. Imaging*, Washington, DC, pp. 169–172.

Masood, K.& Rajpoot N. (2008). Colon biopsy classification Annals of the BMVA Vol. 2008, No. 4, pp 1–16.

Montgomery, D. (1997). *Design and analysis of experiments*. John Wiley & Son, 4th Ed.

O'Dowd, G. J.; Veltri, R. W.; Miller, M. C. & Strum, S. B.(2001). The Gleason score: A significant biologic manifestation of prostate cancer aggressiveness on biopsy, *Prostate Cancer Res. Inst.: PCR Insights*, vol. 4, no. 1, pp. 1–5.

Quinlan, J. R. (1996). Bagging, Boosting, and C4.5. *Proceedings of the Thirteenth National Conference on Arti_cial Intelligence*, 725-730.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann,.

Raymer, M. L. et al. (2000)Dimensionality reduction using genetic algorithms. IEEE Trans. Evol. Comput., vol. 4, no. 2, pp. 164–171.

Roula, M. A.; Diamond, J.; Bouridane, A.; Miller, P. & Amira, A. (2002). A multispectral computer vision system for automatic grading of prostatic neoplasia. *Proc. IEEE Int. Symp. Biomed. lmaging* , pp. 193–196.

Roula, M.; A. Bouridane, A. & Miller, P. (2003) A quadratic classifier based on multispectral texture features for prostate cancer diagnosis. *Proc. 7th Int. Symp. Signal Process. Appl*., Paris, France, pp. 37–40.

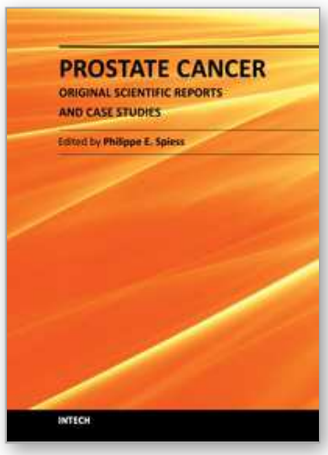
Stewart, B. W. & Kleihues, P. (2003). World Cancer Report World Health Organization, *International Agency for Research on Cancer.*

Tabesh, A.; Kumar, V.; Pang, H.; Verbel, D. Kotsianti, A.; Teverovskiy M. & Saidi, O. (2005). Automated prostate cancer diagnosis and Gleason grading of tissue microarrays, *Proc. SPIE Med. Imag*., vol. 5747, pp.58–70.

Tahir, M.A. & Bouridane, A. (2006). Novel Round-Robin Tabu Search Algorithm for Prostate Cancer Classification and Diagnosis Using Multispectral Imagery. *IEEE transactions on information technology in biomedicine*, Vol. 10, No. 4.

Yongguang, B. Ishii, N. & Du, X. (*2004*). Combining multiple k-nearest neighbour classifiers using different distance functions. *Lectures Notes in Computer Science (LNCS 3177), 5th Int. Conf. Intell. Data Eng. Autom. Learn.*, U.K.

IntechOpen



Prostate Cancer - Original Scientific Reports and Case Studies
Edited by Dr. Philippe E. Spiess

ISBN 978-953-307-342-2
Hard cover, 238 pages
Publisher InTech
Published online 21, November, 2011
Published in print edition November, 2011

This book encompasses three sections pertaining to the topics of cancer biology, diagnostic markers, and therapeutic novelties. It represents an essential resource for healthcare professionals and scientist dedicated to the field of prostate cancer research. This book is a celebration of the significant advances made within this field over the past decade, with the hopes that this is the stepping stone for the eradication of this potentially debilitating and/or fatal malignancy.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Sabrina Bouatmane, Ahmed Bouridane, Mohamed Ali Roula and Somaya Al-Maadeed (2011). Improving Prostate Cancer Classification: A Round Robin Forward Sequential Selection Approach, Prostate Cancer - Original Scientific Reports and Case Studies, Dr. Philippe E. Spiess (Ed.), ISBN: 978-953-307-342-2, InTech, Available from: http://www.intechopen.com/books/prostate-cancer-original-scientific-reports-and-case-studies/improving-prostate-cancer-classification-a-round-robin-forward-sequential-selection-approach

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IntechOpen

IntechOpen