Journal of Biomedical Optics 10(2), 024034 (March/April 2005)

Support vector machine for optical diagnosis of cancer

S. K. Majumder
N. Ghosh
P. K. Gupta
Centre for Advanced Technology
Biomedical Applications Section
Indore 452013, India
E-mail: shkm@cat.ernet.in

Abstract. We report the application of a support vector machine (SVM) for the development of diagnostic algorithms for optical diagnosis of cancer. Both linear and nonlinear SVMs have been investigated for this purpose. We develop a methodology that makes use of SVM for both feature extraction and classification jointly by integrating the newly developed recursive feature elimination (RFE) in the framework of SVM. This leads to significantly improved classification results compared to those obtained when an independent feature extractor such as principal component analysis (PCA) is used. The integrated SVM-RFE approach is also found to outperform the classification results yielded by traditional Fisher's linear discriminant (FLD)-based algorithms. All the algorithms are developed using spectral data acquired in a clinical *in vivo* laser-induced fluorescence (LIF) spectroscopic study conducted on patients being screened for cancer of the oral cavity and normal volunteers. The best sensitivity and specificity values provided by the nonlinear SVM-RFE algorithm over the data sets investigated are 95 and 96% toward cancer for the training set data based on leave-one-out cross validation and 93 and 97% toward cancer for the independent validation set data. When tested on the spectral data of the uninvolved oral cavity sites from the patients it yielded a specificity of 85%. [DOI: 10.1117/1.1897396]

Keywords: diagnostic algorithm; support vector machine; recursive feature elimination; Fisher's linear discriminant; principal component analysis; squamous cell carcinoma; oral cancer.

Paper 03136 received Nov. 19, 2003; revised manuscript received Jun. 15, 2004; accepted for publication Aug. 27, 2004; published online Apr. 26, 2005.

1 Introduction

Diagnosis of cancer at an early stage is important for effective management of the disease. Recently optical spectroscopy has received considerable attention for noninvasive, *in situ*, near-real-time diagnosis of cancer.[1–5] For diagnosis, it exploits subtle changes in the spectra of tissue as tissue transforms from normal to malignant. Central to optical diagnosis is a diagnostic algorithm that can best extract the diagnostic features from the tissue spectra and accurately correlate them with the tissue histopathology. Most of the algorithms reported for optical diagnosis of cancer[6–18] use traditional multivariate statistical techniques such as Fisher's linear discriminant analysis,[1,2,6–9] partial least-squares (PLS) analysis,[10] singular value decomposition[11] (SVD), principal component analysis[12–15] (PCA), etc. These classical linear techniques have the advantage of providing closed-form expressions that lead to simplicity in their design. However, they extract information from only the second-order correlation in the data and ignore higher order correlations, which could also be useful for improved discrimination.[16] Use of nonlinear techniques[16] is receiving attention for the purpose of development of algorithms since these could exploit higher order correlation. Artificial neural networks (ANNs) provide an array of nonlinear algorithms for feature extraction and classification[16,17] and have also been used recently for laser-induced fluorescence (LIF) diagnosis of oral leukoplakia,[18] cervical precancer,[19] and atherosclerotic plaques[20] with excellent discrimination results. Van Staveren et al.[18] demonstrated the use of multilayer ANN-based algorithms for autofluorescence detection of oral leukoplakia. The diagnostic algorithms based on ensembles of radial basis function (RBF) neural networks developed by Tumer et al.[19] could identify cervical precancer more accurately when compared to their previous multivariate statistical algorithms. Rovithakis et al.[20] developed a higher order neural (HON)-network-based diagnostic algorithm and demonstrated its use for LIF detection of atherosclerotic plaques with excellent discrimination results. Apart from ANN-based algorithms, use of other state-of-the-art statistical pattern recognition techniques has also been reported recently.[21,22] For example, Agrawal et al.[21] used wavelet transforms and showed that features extracted from the polarized autofluorescence spectra of breast tissues through this transforms could serve as good discrimination indices. We recently showed that a nonlinear diagnostic algorithm based on the theory of maximum representation and discrimination feature (MRDF) can provide much improved diagnostic performance as compared to that based[22] on linear PCA.

Address all correspondence to Dr. Shovan K. Majumdar, Center for Advanced Technology, Laser Program/Gov of India, R&D Block-D, Indore 452013 India.

1083-3668/2005/$22.00 © 2005 SPIE

Another powerful recent approach for statistical pattern recognition based on machine learning is the theory of support vector machine (SVM), originally developed by Vapnik[23] and Burges.[24] SVMs have already received tremendous attention in a wide variety of classification problems[24–29] and are being actively pursued for various theoretical extensions.[30–32] The possibility of using SVMs for developing diagnostic algorithms is also attracting attention.[33,34] While Palmer et al.[33] used a linear SVM classifier for classifying autofluorescence and diffuse reflectance spectra of breast tissues *in vitro*, Lin et al.[34] classified *in vivo* autofluorescence spectra from nasopharingeal tissues by using both the linear and the nonlinear SVM classifier with RBF kernel. In the reports of both the groups, the tissue spectra were dimensionally reduced by applying linear PCA algorithms prior to using the SVM approach for classification. Lin et al.[34] showed that the classification performance of an SVM classifier trained on the full spectral data was comparable to that obtained with the classifier trained on the diagnostically relevant principal components only. Their combined PCA-SVM approach was reported to have reduced computational complexity.

In this paper, we report, the use of an SVM for both feature extraction and classification jointly by integrating the approach of recursive feature elimination[35] (RFE) in the framework of an SVM (Refs. 23 and 24). RFE is a new technique developed recently by Guyon et al.[35] for extracting an optimal subset of nested features relevant for classification from a set of data with a vast number of features. Since RFE performs feature extraction using a performance criterion set by the classifier, the use of the integrated framework of SVM and RFE is expected to lead to a better classification performance compared to that with the use of an independent feature extractor such as PCA. We developed both linear and nonlinear SVM-based diagnostic algorithms using spectral data acquired in a clinical *in vivo* LIF study conducted on patients being screened for cancer of the oral cavity and normal volunteers. Although, in this paper, we focus on binary classification, i.e., cancerous versus normal, it can be easily extended to a multiclass classification using various approaches,[36] thereby enabling one to classify spectral data into more than two classes comprised of patients with various kinds of lesions of the oral cavity, for example, leukoplakia, erythroplakia, etc. in addition to cancerous and noncancerous lesions. In this paper, however, we focus on classifying spectral data of cancerous and normal tissue. We also compare the diagnostic efficacy of the SVM-based algorithms with that based on PCA and Fisher's linear discriminant (FLD) using the same spectral data set. The algorithms based on SVM-RFE as well as SVM alone provide significantly improved diagnostic performance as compared to that based on both PCA and FLD in discriminating the cancerous tissue sites of the oral cancer patients from the healthy squamous tissue sites of normal volunteers as well as the uninvolved tissue sites of the patients with cancer of the oral cavity.

2 Materials and Methods

In vivo autofluorescence spectra were recorded using a N_2-laser-(337-nm)-based portable fluorimeter reported earlier.[15,22] It comprises a sealed-off pulsed N_2 laser, a spectrograph (Acton Research Corporation, Acton, MA, USA), an optical fiber probe, and a gateable intensified CCD (ICCD) detector (4 Quik 05A, Stanford Computer Optics, Inc., Berkely, CA, USA). The spectral data acquisition was computer controlled. The autofluorescence spectra were recorded with the tip of the fiber optic probe placed in contact with the tissue surface. From each site, spectra were recorded in the 375- to 700-nm spectral range. During each measurement of tissue fluorescence, a reference spectrum was also acquired simultaneously from the phosphor-coated tip of an additional fiber illuminated with N_2 laser radiation leaking from the other end of the N_2 laser cavity. The peak of this reference spectrum was used to normalize the acquired tissue spectra and thus account for the observed pulse-to-pulse variation of the N_2 laser power. The intensity of fluorescence from each tissue site is reported in this calibrated unit.

The study involved 13 normal volunteers with no history of the disease of the oral cavity and 16 patients selected from those enrolled for medical examination of the oral cavity at the outpatient department (OPD) of the Government Cancer Hospital, Indore. Informed consent was obtained from each patient as well as the normal volunteers who participated in this study. The patients included in this study had no history of malignancy and were suspected on visual examination by the concerned physician of having early cancer of the oral cavity. For these patients, biopsies were taken from the suspected areas subsequent to acquisition of spectra. Only those patients for whom histopathological diagnosis was squamous cell carcinoma (SCC), grade I, were included in this study. *In vivo* autofluorescence spectra were acquired from a total of 171 tissue sites from patients, of which 83 were SCC and the rest were uninvolved squamous tissue. Spectra were also recorded from 154 sites from healthy squamous tissue of normal volunteers. In each patient, the normal tissue sites interrogated were from the contralateral apparently uninvolved region of the oral cavity. On an average, five spectra from the cancerous tissue sites and four spectra from the uninvolved tissue sites were recorded. In normal volunteers, on an average, 10 spectra were recorded from the healthy squamous tissues. Each site was treated separately and classified via the diagnostic algorithm developed.

2.1 *Spectral Data*

Each tissue fluorescence spectrum consisted of 717 intensity values (corresponding to 717 pixels of the ICCD) spanning the wavelength range of 375 to 700 nm. The autofluorescence spectra recorded from different cancerous and contralateral normal sites of the oral cavity of a patient are shown in Figs. 1(a) and 1(b), respectively. The considerable site-to-site variation in the spectra is apparent. The differences in the spectra from some of the normal and cancerous tissue sites are not that apparent, because they are masked by the large intrapatient and interpatient variability in the intensity and line shapes. While some of this variation may represent intrinsic variation in tissue fluorescence, the variable nature of the contact of the probe with the tissue surface in a clinical situation will also add to the variation. It is pertinent to note that in the *in vitro* studies on oral cavity tissues,[37] where the variability due to the nature of contact of probe with tissue surface is expected to be minimal, a percentage variation $(\sigma/\bar{x})$ of ~30% was observed in the spectrally integrated intensities

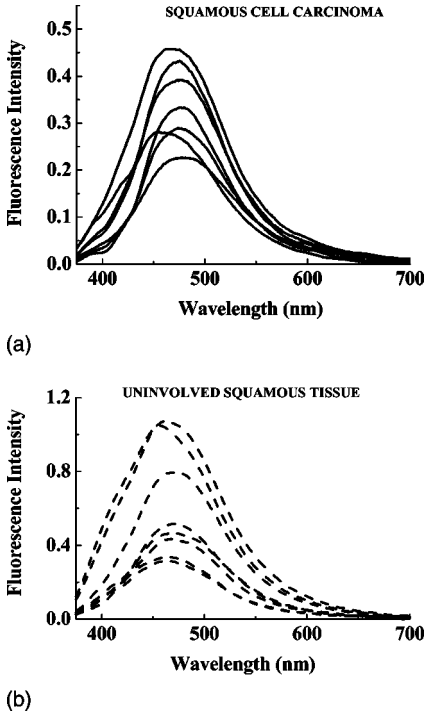

Fig. 1 N_2-laser-excited autofluorescence spectra recorded from (a) squamous cell carcinoma tissue sites (solid line) and (b) uninvolved tissue sites (dashed line) of the same patient.

(ΣI) from different sites of normal or cancerous tissues over the total sample size investigated. Here, $\overline{x}$ is the mean of ΣI values from different sites of one category and σ is the standard deviation. In comparison, in this *in vivo* study, the percentage variation ($\sigma/\overline{x}$) in (ΣI) was ~60%. To ensure good discrimination, it is necessary to minimize these variations that may obscure the intercategory differences. To minimize the inter- and intrapatient variability, a two-step procedure for preprocessing of the raw spectral data was adopted. In the first step, the mean spectrum over all the healthy squamous tissue sites of the normal volunteers was calculated and subtracted from the spectrum of each tissue site of the oral cavity of patients as well as of normal volunteers. Since mean subtraction displays the differences in the spectra of the diseased with respect to the mean spectra of the healthy squamous tissue, it is expected to lead to enhancement of spectral differences between the two diagnostic categories. Next, the resultant spectrum of each category was normalized with respect to the standard deviation of the spectra of that category. This normalization is expected to remove from the spectra the influence of scatter in the spectral intensity by making the standard deviation of the spectra of each diagnostic category equal to unity. Indeed, mean subtraction followed by normalization of the spectra with respect to their respective standard deviations made the spectral differences between the two diagnostic categories much more apparent. Figure 2 shows the spectra for cancerous and uninvolved sites of the oral cavity of the same patient after preprocessing. Note here that the differences in the preprocessed spectra from cancerous and contralateral uninvolved tissue sites of the same patient are generally more distinct[12,37] as compared to the differences when preprocessed spectra from similar tissue sites of all the

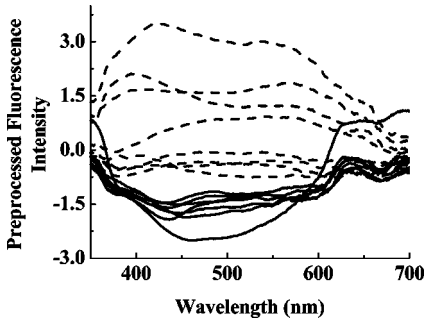

Fig. 2 Preprocessed autofluorescence spectra from squamous cell carcinoma tissue sites (solid line) and from uninvolved squamous tissue sites (dashed line) of the oral cavity of the same patient.

patients are considered as a whole. Various earlier reports on measurements of tissue fluorophores[38] as well as tissue parameters[39] also demonstrate this effect. Figure 3 shows the preprocessed spectra from cancerous and contralateral normal tissue sites of four patients chosen at random. It is evident from the figure that the interpatient differences in the preprocessed spectra do not appear to be that prominent in comparison with the intrapatient differences shown in Fig. 2.

2.2 *Algorithm Development*

For the development of the diagnostic algorithm, the entire set of preprocessed spectral data from the SCC tissue sites of the patients and the healthy squamous tissue sites of the normal volunteers was randomly split into two groups: the training data set and the validation data set, ensuring that both sets contain roughly equal number of spectral data from each histopathologic category. The purpose of the training data set was to develop and optimize the diagnostic method, and the purpose of validation set was to prospectively test its accuracy in an unbiased manner. The random assignment was carried out to ensure that not all the spectral data from a single individual were contained in the same data set. Next, the preprocessed spectral data of the training set were used as inputs for the development of the diagnostic algorithms.

The performance of a diagnostic algorithm depends on the prototype spectral data included in the training set. The more exactly the prototype data represent the different disease categories to be discriminated, the better will be the accuracy expected in the performance of the algorithm. The general

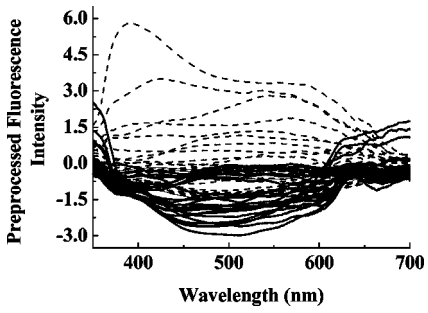

Fig. 3 Preprocessed autofluorescence spectra from squamous cell carcinoma tissue sites (solid line) and from uninvolved squamous tissue sites (dashed line) of the oral cavity of four patients chosen at random.

practice is to use spectral data of uninvolved tissue sites surrounding the cancerous tumor of patients as the normal database for development of diagnostic algorithms.[12,15] However, the normal appearing region surrounding a cancerous tumor of a patient might have some biochemical changes due to the field effect of malignancy,[40] particularly at the advanced stage of the disease. This was believed to be the reason for obtaining reduced classification performance in our earlier *in vivo* studies.[15] In this study, although the patients included were reported to have earlier stages (grade I) of squamous cell carcinoma, we investigated the use of two separate normal databases for the development of diagnostic algorithms. In one, we took as normal database, the spectral data of contralateral uninvolved tissue sites of patients, while in the second, we took the spectral data of healthy squamous tissue sites of normal volunteers who had no history of any disease of oral cavity. Our initial results showed that use of spectral data of normal volunteers gave slightly improved (by ~5 to 7%) classification performance. Therefore, for subsequent development of diagnostic algorithms we considered, as the normal database, the spectral data from the healthy squamous tissue sites of the normal volunteers and avoided use of spectral data from tissue sites of normal-appearing mucosa in the contralateral uninvolved region of the oral cavity of patients.

2.3 *SVM*

SVMs are powerful tools for data classification. The central idea of an SVM is to separate classes with a surface that maximizes the margin between them.[24] The formulation of the technique relies on the theory of uniform convergence in probability and associated structural risk minimization (SRM) principle[23] to minimize the structural risk, i.e., the probability of misclassifying yet-to-be-seen patterns for a fixed but unknown probability distribution of the data. The mathematical formulation and associated theoretical background of SVM have been detailed in Vapnik's book[23] as well as in several literature sources.[24,25,31,32] In the following, we briefly discuss the basic ideas of SVM for the purpose of our description.

Given a set of N-dimensional (N being the number of wavelengths over which spectra were recorded) LIF spectral data of cancerous and normal squamous tissue sites labeled by $\lambda \in \{-1,+1\}$ with $\lambda=+1$ referring to cancer and $\lambda=-1$ referring to normal, the task of an SVM is to separate this set of binary labeled input data into its constituent classes. A simple way to build a binary classifier is to construct a hyperplane (decision boundary) in the N-dimensional input space that separates class members (positive examples) from nonmembers (negative examples) considered as points in that space. A look at the LIF spectral data (see Figs. 1 and 3) would show that because of considerable intercategory overlap, there exists no separating hyperplane in the input space that successfully separates the positive from negative examples. One approach to solve this inseparability problem is to map the data from the input space into a higher dimensional feature space through an *a priori* chosen nonlinear mapping and construct a separating hyperplane that is linear in that space, but is nonlinear with respect to the input space.[24] However, the technical difficulty involved in mapping the training set data to a higher dimensional space for classification is the computational burden.[24] Furthermore, artificially separating the data in this way exposes the learning system to the risk of finding trivial solutions that may overfit the data.[24] This means that there may exist infinitely many hyperplanes that can successfully separate the training set data, but perform miserably on unseen (test) data points.

The SVM is developed to simultaneously sidestep both these difficulties. It avoids overfitting by choosing an optimal separating hyperplane (OSH) in the feature space (from among the many) that maximizes the width of the margin between the classes, i.e., the empty area around the decision boundary defined by the distance to the nearest training data points of either class.[24] The OSH also minimizes the risk of misclassifying not only the data points in the training set (i.e., empirical risk minimization) but also the yet-to-be-seen data points of the test set for a fixed but unknown probability distribution of the data thereby following the SRM principle.[23] The approach of SRM equips the SVM with a greater ability to generalize, which in turn leads to significantly improved classification performance as compared to the traditional techniques that follow only the empirical risk minimization principle to minimize the mean-squared error over the training data set.

The location of the OSH in the feature space is specified by real-valued weights on the training set data points.[24] Those training set data points that lie far away from the OSH do not participate in its specification and therefore receive weights of zero. Only the training set data points that lie close to the decision boundary between the classes receive nonzero weights.[23,24] These training set data points are called support vectors,[24] since only these points define the classification boundary and removing them would change the location of the OSH. It has also been shown by Vapnik[23,41] and Burges[24] that if the training data points must be separated without errors by an OSH, the expected error rate on an unseen data point is bounded by the ratio of the number of support vectors to the number of training data points. Since the ratio is independent of the dimension of the problem, obtaining a small set of support vectors can guarantee a good generalization performance of an SVM classifier.

Another important advantage of the SVM approach is that it avoids the computational burden of explicitly mapping the input data to the higher dimensional feature space (via nonlinear mapping $\phi: R_0 \rightarrow F$ from input space R_0 to the feature space F) without ever explicitly performing the mapping, since neither the SVM learning algorithm nor the SVM decision function must represent explicitly the input data points in the feature space $\phi(x)$ and only use dot products between such points $\langle \phi(x), \phi(y) \rangle$ in the feature space.[24] This is done simply by defining a function $K(x,y)=\langle \phi(x), \phi(y) \rangle$ that plays the role of dot product in the feature space. The function $K(x,y)$ is called the kernel function[24] and is termed legitimate only if it obeys Mercer's theorem.[23,24] The use of this kernel function enables the SVM to operate efficiently in a nonlinear high-dimensional feature space without being adversely affected by the dimensionality of that space.

Computationally, the algorithm proceeds by calculating in the final step the two-class decision function defined by an SVM classifier:

$$D(x)=\mathrm{sign}\left[\sum_{\forall x_i \in S} \alpha_i \lambda_i K(x_i,x)+\alpha_0\right], \qquad (1)$$

where $K(x_i, x)$ is the kernel function of a new data point x (to be classified) and a set of training data points x_i, S is the set of support vectors (a subset of training set), and $\lambda_i = \pm 1$ is the label of training data points x_i and $\alpha_i \geq 0$ are the Lagrange multipliers for OSH.

For the LIF spectra data that contain considerable class overlap, the maximum-margin (or the hard-margin) SVM approach may not be able to find any separating hyperplane at all.[24] This problem is addressed by using a soft margin that allows some training data points to fall on the wrong side of the separating hyperplane.[24] Therefore, completely specifying an SVM, in this case, also requires specifying additional parameters that provide the magnitude of the penalty for violating the soft margin. These parameters, along with others, are determined during the training phase of the SVM algorithm by solving a quadratic optimization problem given by[23]

$$\min_\alpha \left(\alpha^T \mathbf{\Lambda K \Lambda} \alpha + C \sum_j \epsilon_j \right), \tag{2}$$

under the constraint: $\lambda_j D(x_j) \geq 1 - \epsilon_j$, $\forall x_j$ in the training set, where $\mathbf{\Lambda}$ is a diagonal matrix containing the labels λ_j, and the matrix $\mathbf{K}$ stores the values of the kernel function $k(x_i, x)$ for all the training data points belonging to both the classes. The set of slack variables ϵ_j in Eq. (2) allow for the class overlap, controlled by the penalty weight $C > 0$. This parameter C, called the regularization parameter, basically controls the trade-off between the largest margin and lowest number of errors. For $C = \infty$, no class overlap is allowed. During optimization, the values of α_i become 0 for the majority of training data points, except for the support vectors that comprise only a small subset of the total number of training data points and are only finally needed for separating class members from nonclass members. This property allows the SVM to classify new data points efficiently, since the majority of the training data points can be safely ignored.

2.4 *Selection of Kernels*

The selection of an appropriate kernel function is very important, since it defines the feature space in which the training set data points are classified. As long as the kernel function is legitimate, i.e., it obeys the Mercer's theorem,[23,24] an SVM will operate correctly even if the designer does not know exactly what features of the training data are being used in the kernel-induced feature space. This kernel function must be chosen *a priori* and it determines the type of the SVM classifier. Given a set of support vectors, x_i and a data point x (to be classified), the simplest kernel that can be used is just the dot product in the input space: $K(x_i, x) = x_i, x + 1$, resulting in a linear classifier. When this dot product kernel is used, the feature space is essentially the same as the N-dimensional input space, and the SVM will define a linear OSH in this space. Raising the kernel to higher powers yields nonlinear kernels that are polynomial separating surfaces of higher degrees in the input space. In general, nonlinear kernels, such as $K(x_i, x) = (x_i, x + 1)^d$ result in a d'th-order polynomial SVM classifier. Similarly, use of Gaussian RBF results in an RBF kernel: $K(x_i, x) = \exp(-\|x_i - x\|^2)/2\sigma^2)$, where σ is the width of the Gaussian.

We used linear as well as both the nonlinear (polynomial and RBF) kernels for the development of SVM diagnostic algorithms with the *in vivo* LIF spectral data. The selection of optimal values of the order d in the polynomial kernel and the width σ in the Gaussian RBF kernel is an optimization problem, where the possible values that the parameters can have is a finite set, and the cost function is defined by the application. We chose the cost function as the misclassification error in the training set data obtained with the leave-one-out cross-validation estimate. If the total number of misclassified samples was the same at more than one d or σ values, then the value at which the total number of cancerous samples misclassified was minimum was selected. For selecting optimal d value for the polynomial kernel, the polynomial SVM was trained on the full spectral data of the training set with the polynomial kernel raised to different degrees d selected from a set of d values ranging from 1 to 4 with increments of 1. The optimal value of d was chosen to be the one that gave the highest leave-one-out cross-validation classification performance. We restricted the set of d values up to 4, since for d values larger than 4, the learning algorithm was found to have convergence problems with the given data set. Since the kernel is learned from the data at hand during training of the algorithm, it appears that the polynomial kernel at higher values of d became a ''bad kernel'' for the given data. In other words, it means that the kernel matrix perhaps no longer remained positive-definite and became diagonal during learning from the given spectral data probably due to the generation of a large number of irrelevant features in the kernel-induced feature space. Similarly, the optimal value of σ was selected using an exhaustive search method. The RBF-SVM classifier was trained on the full set of spectral data of the training set for the different σ values selected from a set of σ values ranging from 0.1 to 1000 with increments of 0.1 for σ values between 0 to 1, with increments of 1 for σ values between 1 to 20, with increments of 5 for σ values between 20 to 100, and with increments of 50 for σ values between 100 to 1000. Optimal value of σ was the one that gave the least leave-one-out cross-validation error.

2.5 *FLD*

Given a set of input data comprising of LIF spectral data from cancerous and normal tissues with a given dimensionality, the FLD (Ref. 42) aims to project this data onto a line and performs classification in this 1-D space. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class. This defines the Fisher's criterion, which is maximized over all linear projections, w:

$$J(w) = \frac{|\mu_1 - \mu_2|^2}{S_1^2 + S_2^2},$$

where μ represents the mean, S^2 represents the variance, and the subscripts denote the two classes. Maximizing this criterion yields a closed-form solution that involves the inverse of a covariance-like matrix.

2.6 *Feature Selection: RFE*

For each preprocessed LIF spectral data consisting of 717 intensity values we have in the input data space 717 features representing intensities at different wavelengths. It is often

necessary, while designing a classifier, to select a subset of diagnostically relevant features from the vast number available. It is important because throwing away irrelevant features (i.e., the features that do not assist in classification) reduces the risk of overfitting and decreases computational complexity.[16] At the same time, limiting the number of features has the associated risk of reducing the expected classification performance by introducing a bias.[43] The objective of any feature selection exercise is to select optimal number of features, using which the performance of the classifier is as good as if not better than that using all the features. The selection of an optimal subset of features from a set of features can be carried out by using an appropriately designed performance measure to evaluate their ability to classify the samples[16] (e.g., cancer versus normal). A brute-force search of the best combination of features (combination of 2, 3, or more number of features of 717 features) that results in best classification performance is impractical, because the number of possible feature combinations will be prohibitively large for such large set of features. One approach[43] is to train the classifier itself with the full set of features, compute some feature ranking criteria (e.g., the weights of the classifier) to evaluate how well an individual feature contribute to the classification, rank the features based on the criteria, and then use a fixed number of top-ranked features to finally classify the data. This method has an important drawback in that if some of the features (say, the least-ranked ones) are removed and the process is repeated, the resulting ranking of the remaining features differs from their previous ranking obtained without removing any of the features. Therefore, use of this approach may not provide an optimal subset (for classification) from the full set of features. This problem has been very effectively addressed by the recursive feature elimination (RFE) method, proposed recently by Guyon et al.[35] In this approach, feature ranking is carried out through a recursive procedure. Given the preprocessed spectral data with a full set of features (i.e., 717 intensity values at the different wavelengths), the algorithm starts by training the classifier using all the features as input, assesses the relative importance of the features in the classifier by computing the feature ranking criteria (e.g., the weights of the classifier), eliminates the least important feature corresponding to the smallest ranking criterion, and lists the index corresponding to this feature in a feature-ranked list initialized for this purpose. The classifier is again trained with the remaining set of surviving features as input, the least important feature corresponding to the lowest ranking criterion is again eliminated, and the index corresponding to this feature is added to the previous feature-ranked list. This procedure of training the classifier, computing the feature ranking criteria, and feature elimination is carried out recursively to update the feature-ranked list at each iteration until all 717 features of the original spectral data have been assessed. Thus, at the end of the iterative loop, one gets, as the output, a feature-ranked list. After having prepared the list, the next task of RFE algorithm is to decide on the subset of optimal number of features required for best classification. For that, the different numbers of top-ranked features are selected to form a series of different feature subsets (starting with the full set) and the performance of the classifier is assessed iteratively with these selected subsets of features to determine the optimal subset. The series of different feature subsets formed are basically nested $F_1 \subset F_2 \subset \cdots \subset F$, which means that the selected subset of l features is included in the subset of $l+1$ features. Clearly, the previous method of feature ranking is computationally equivalent to the first iteration of RFE. Thus, RFE provides a ranked list of features indicative of feature subset ranking, as opposed to feature ranking. This means that the features that are top ranked (i.e., eliminated last) are not necessarily the ones that are individually most relevant. Only taken together, they are relevant for classification.

In our case, we used both SVM and FLD to select an optimal subset of features using RFE. While the feature-ranking criterion used for SVM-RFE was

$$w_r = \sum_{\forall x_i \in S} \alpha_i \lambda_i \mathbf{K}(x_i, x),$$

for FLD-RFE it is

$$w_r = \mathbf{S}_w^{-1}(\mu_1 - \mu_2),$$

where $\mathbf{S}_w$ is the within-class scatter matrix and α, λ, $\mathbf{K}$, and μ are as defined previously. For computational reasons, we removed several features at a time. We started with all 717 features. At the end of the first iteration, all the features were ranked and the bottom half closest to half of the total number of features was eliminated. Similarly, at each subsequent iteration, we eliminated close to half of the remaining features. We thus obtained a total of 11 nested subsets of features of increasing informative density from the whole set of features. The 11 subsets are composed of 717, 350, 175, 80, 45, 20, 15, 10, 5, and 1 feature, respectively. The quality of these subsets of features was assessed by training the four classifiers (one FLD and the three SVMs) at each iteration stage.

2.7 *Analysis of Algorithm Performance*

The performance of a diagnostic algorithm depends on how accurately the algorithm separates the set of data being tested into the different classes. The relative performance of the different diagnostic algorithms was assessed by carrying out a receiver-operating characteristic[44] (ROC) analysis of the corresponding classification results. An ROC curve was generated corresponding to each diagnostic algorithm for the validation data set by plotting the true positive rate (sensitivity) as a function of the false positive rate (1-specificity) as the classification threshold was varied. An ROC curve provides a visual comparison of the trade-off between sensitivity and specificity of a diagnostic test. The closer the curve follows the left-hand border and the top border of the ROC space, the better is the performance of the diagnostic algorithm. Similarly, the closer the curve comes to the 45-deg diagonal of the ROC space, the less is its accuracy. To quantify the performance measure of the different algorithms, the areas under the different ROC curves were estimated. An area of 1 represents an ideal diagnostic algorithm, while an area of 0.5 represents a worthless one. The closer the area is to 1, the more accurate is the corresponding diagnostic algorithm.

3 Results

Table 1 lists the diagnostic results obtained with a linear SVM classifier trained on the spectral data set corresponding to raw spectra and preprocessed spectra with the full set of spectral

Table 1 Classification results obtained with the linear SVM classifier and the conventional nearest mean classifier (NMC) using the data set corresponding to the unprocessed raw spectra and the preprocessed spectra with full set of spectral features.

		Training Data Set		Validation Data Set		
Spectral Data	Classifiers	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Data Set I Specificity (%)	Data Set II Specificity (%)
Raw spectra	SVM	81	94	78	96	74
	NMC	74	58	83	58	66
Preprocessed spectra	SVM	86	91	88	92	77
	NMC	81	65	80	58	55

Sensitivity and specificity values in the training set data represent leave-one-out cross-validation values.

features (i.e., $N=717$ intensity values). For comparison's sake, the classification results yielded by a conventional NMC on the same data sets are also listed in the same table. An NMC is based on least Euclidean distance of the test features from the means of the prototype features of the corresponding tissue types in the training set. The sensitivity and specificity values for the training set data were obtained on the basis of leave-one-out cross-validation. It is evident from the table that the sensitivity and specificity values in the training and the validation data sets are much improved with the preprocessed spectral data as compared to the unprocessed raw spectral data. Therefore, we extended the subsequent exercise on algorithm development only with the preprocessed spectral data sets.

Figures 4 and 5 demonstrate the leave-one-out cross-validation error as a function of the degrees (d) of the polynomial kernel and the widths (σ) of the Gaussian RBF kernel, respectively. From Fig. 4 it is clear that the leave-one-out error is the minimum for $d=2$, and therefore, we used polynomial kernel of degree 2 for training the polynomial SVM classifier for algorithm development. Figure 5 shows that leave-one-out error is the minimum at more than one σ value (e.g., at $\sigma=50$, 75, and 100). However, for the σ value of 100 the total number of cancerous samples misclassified was the minimum and therefore, $\sigma=100$ was used as the width of the RBF kernel for subsequent training of the RBF SVM classifier.

To train an SVM algorithm one must supply *a priori* the value of the regularization parameter C to the learning algorithm. Since no established guideline exists in the SVM methodology[23,24] as to what should be the optimal value of C, the linear and nonlinear SVM classifiers were trained with different values of C ($C=1$, 10, and ∞). It was found that the classifier with $C=\infty$ gave the best generalized classification performance, i.e., the total misclassification error over the training (leave-one-out cross-validation) and the independent validation data sets was the least. Therefore, for subsequent feature subset selection with the RFE algorithm we trained each of the SVM classifiers with $C=\infty$ at each iteration stage. To evaluate the diagnostic contribution of each selected subset of features at each iteration stage of the RFE algorithm, we set the cost function as the total number of samples misclassified by the classifier in the independent validation set as well as in the training set data with leave-one-out cross-validation. The optimal subset of features was the one for which the total number of misclassified samples was the minimum. If the total number of misclassified samples was the same for more than one feature subsets, then the feature subset for which the total number of cancerous samples misclassified was minimum, was selected.

The total misclassification error for the 11 nested subsets of features was determined with the SVM RFE method. The results for linear, polynomial, and RBF kernels are shown in Figs. 6 to 8, respectively. It is evident from the figures that while the misclassification error is the minimum for the linear SVM classifier trained with the subset of 45 features ranked

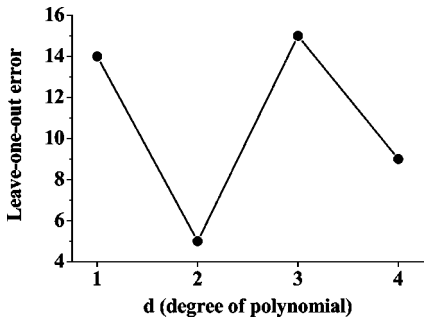

Fig. 4 Leave-one-out cross-validation error in the training set data as a function of the degree of the polynomial kernel for the polynomial SVM classifier.

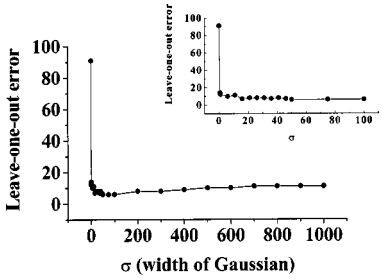

Fig. 5 Leave-one-out cross validation error in the training set data as a function of the width of the Gaussian radial basis function for the RBF SVM classifier.

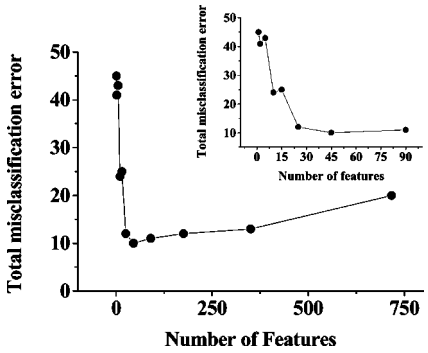

Fig. 6 Total misclassification error in the training and the validation data set as a function of the 11 nested subsets of features obtained using SVM RFE algorithm with a linear kernel.

by the SVM RFE algorithm, for both the polynomial SVM and the RBF SVM the respective misclassification errors are the minimum with the subset of 350 features ranked by the respective SVM RFE algorithms. The diagnostic performances of the SVM RFE algorithms with linear, polynomial, and RBF kernels are listed in Tables 2 to 4. The sensitivity and specificity values for the training set data represent the leave-one-out cross-validation values.

Similarly, for the development of the RFE algorithm with the FLD classifier, it was trained on the training set data and tested on the training set (leave-one-out cross-validation) as well as on both the independent validation data sets at each iteration stage. Figure 9 shows the total misclassification error as a function of the 11 nested subsets of features obtained with the FLD RFE method. The figure clearly shows that the misclassification error is the minimum with the subset of 45 features ranked by the FLD RFE algorithm. Table 5 summarizes the diagnostic results obtained with the FLD RFE algorithm. Here also, the sensitivity and specificity values for the training set data represent the leave-one-out cross-validation values.

Table 6 lists the sensitivity and specificity values for the training and the validation data sets obtained using the FLD as well as the three SVM algorithms with the full spectral features as well as with the optimal subset of features selected using the respective RFE algorithms. For comparison's sake, the sensitivity and specificity values obtained using the linear PCA-based algorithm as well as the linear SVM algorithm trained with the diagnostically relevant principal components

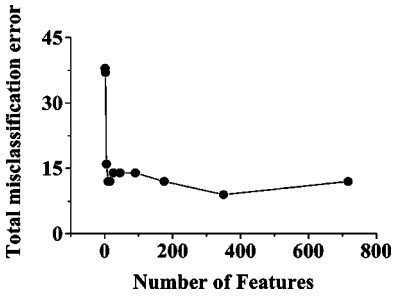

Fig. 7 Total misclassification error in the training and the validation data set as a function of the 11 nested subsets of features obtained using the SVM RFE algorithm with a polynomial kernel.

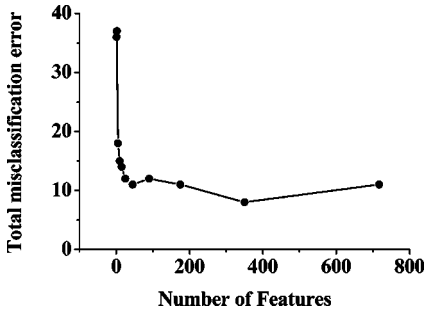

Fig. 8 Total misclassification error in the training and the validation data set as a function of the 11 nested subsets of features obtained using the SVM RFE algorithm with an RBF kernel.

(PCs) are listed in Table 7. PCA of preprocessed spectra resulted in six PCs that collectively accounted for 99.5% of the total variance of the spectral data. Of the six PCs, only four (PC 1, PC 3, PC 4, and PC 5) were found to have significantly different ($p<0.001$) values for SCC and normal squamous tissue. Therefore, these four PCs, which together accounted for 79% of the total variance (PC 1 accounting for 70%, PC 3 for 6%, and PC 4 for 2% of the total variance, and PC 5 accounting the remaining 1%) were used for subsequent classification.

Figure 10 shows the ROC curves generated for the different diagnostic tests based on SVMs and FLD. To quantify the accuracy of the tests, the areas under the curves were also estimated. Table 8 lists the area under the curve values for the ROC curves corresponding to linear, polynomial, and RBF SVM diagnostic algorithms.

4 Discussion

In Table 1, we summarized the diagnostic performance of the SVM classifier and the nearest mean classifier. For both the classifiers, the diagnostic results were obtained using the preprocessed spectral data and the unprocessed raw spectral data. The results clearly indicate that SVM outperforms the NMC for both data sets. The superior classification performance of the SVM classifier originates from the built-in capability of the SVM approach to separate classes that are not linearly separable in the original parametric space.[24] The advantage of the two-step preprocessing of the raw spectral data, as described in the previous sections, is also apparent from the table.

The diagnostic performances of the SVM-based and the linear-PCA-based algorithms over the training and the two independent validation data sets are listed in Table 7. For SVM-based algorithms, classification results were obtained for two cases. In one case, SVM was used as a classifier with PCA, providing the diagnostically relevant features (SVM PCA), and in the second case, SVM was used for classification with the full set of spectral features as well as for both feature extraction and classification jointly using the SVM-RFE approach. A perusal of the table shows that the SVM-based algorithms have resulted in significantly improved classification performance as compared to that obtained with the PCA-based algorithms. Further, in view of the previous work by Lin et al.[34] using the SVM PCA, our results show that the integrated SVM-RFE approach gives considerably improved diagnostic performance as compared to the SVM-PCA algo-

Table 2 Classification results of the linear SVM-based diagnostic algorithm for the training and the validation data sets with the 11 subsets of features selected through the RFE method.

		Training Data Set			Validation Data Set		
Diagnostic Algorithm	Number of Features	Sensitivity (%)	Specificity (%)	Misclassification Error (%)	Sensitivity (%)	Specificity (%)	Misclassification Error (%)
	(All)717	86	91	11	88	92	9
	350	93	96	5	88	94	8
	175	95	97	3	88	92	9
	90	98	97	2	88	92	9
	45	**98**	**97**	**2**	**88**	**94**	**8**
Linear SVM	25	98	97	2	88	91	10
	15	86	94	9	74	88	16
	10	86	96	8	78	88	15
	5	64	83	24	71	86	19
	2	69	87	19	63	86	22
	1	62	83	24	68	84	21

Sensitivity, specificity, and the misclassification error in the training set data are reported based on leave-one-out cross-validation. The row with bold figures indicates the optimal feature subset.

rithm developed based on the same data set. This is not surprising because PCA, which is basically an independent feature extractor,[16] extracts features by projecting the input data into a new feature space of lower dimensionality through a linear transformation matrix. PCA optimizes the transformation matrix by finding the largest variations in the original input space,[16,42] thereby reducing the dimension of the original data by optimally representing the data in the form of a

Table 3 Classification results of the polynomial SVM-based diagnostic algorithm for the training and the validation data sets with the 11 subsets of features selected through the RFE method.

		Training Data Set			Validation Data Set		
Diagnostic Algorithm	Number of Features	Sensitivity (%)	Specificity (%)	Misclassification Error (%)	Sensitivity (%)	Specificity (%)	Misclassification Error (%)
	717	93	97	4	90	94	8
	350	**93**	**100**	**2**	**90**	**95**	**7**
	175	90	97	5	93	94	7
	90	88	97	6	90	94	8
	45	88	97	6	90	92	8
Polynomial SVM	25	90	96	6	90	92	8
	15	90	100	3	88	92	9
	10	90	99	4	90	92	8
	5	93	94	6	90	90	10
	2	62	94	18	63	91	19
	1	60	91	21	66	92	17

Sensitivity, specificity, and the misclassification error in the training set data are reported based on leave-one-out cross-validation. The row with bold figures indicates the optimal feature subset.

Table 4 Classification results of the RBF-SVM-based diagnostic algorithm for the training and the validation data sets with the 11 subsets of features selected through the RFE method.

		Training Data Set			Validation Data Set		
Diagnostic Algorithm	Number of Features	Sensitivity (%)	Specificity (%)	Misclassification Error (%)	Sensitivity (%)	Specificity (%)	Misclassification Error (%)
	717	93	96	5	93	95	6
	350	**95**	**96**	**4**	**93**	**97**	**4**
	175	90	95	7	90	99	4
	90	90	96	6	90	96	6
	45	95	97	3	88	95	8
RBF SVM	25	95	96	4	88	94	8
	15	95	99	2	83	91	12
	10	90	97	5	83	94	10
	5	86	95	8	85	92	10
	2	57	96	18	56	94	19
	1	64	94	17	59	94	19

Sensitivity, specificity, and the misclassification error in the training set data are reported based on leave-one-out cross-validation. The row with bold figures indicates the optimal feature subset.

few PCs (which are linear combinations of the original data). However, the PCs do not ensure any class-discriminatory information. The drawback of an independent feature extraction algorithm such as PCA is that its optimization criterion is different from the classifier's minimum classification error criterion,[16] which can cause inconsistency between feature extraction and classification stages of a diagnostic algorithm and, consequently, may degrade the performance of classifiers. This problem is overcome by pursuing the integrated approach of SVM and RFE, since RFE performs feature extraction by selecting the diagnostically relevant input variables while using the performance criterion set by the classifier itself.[35] Further, note also that computational complexity is also not reduced in the SVM-PCA approach, because the SVM classification operation does not depend on the dimensionality of the feature space, which can be even infinite.[24] Perhaps the SVM-PCA approach could be little faster, but at the cost of classification performance. However, the speed aspect should not matter when developing an off-line diagnostic algorithm where the main focus is the accuracy and simplicity. Speed requirements can also be taken care of by the SVM-RFE approach, which practically performs dimension reduction through feature selection.

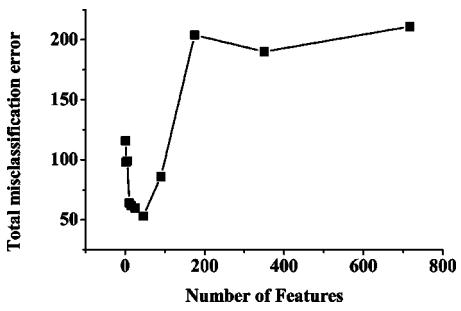

Fig. 9 Total misclassification error in the training and the validation data set as a function of the 11 nested subsets of features selected through RFE for FLD classifier.

In Table 6, we list the diagnostic performance of the FLD and the three SVM algorithms with the full spectral features as well as with the optimal subset of features selected using the respective RFE algorithms. A perusal of the table shows that SVM-based algorithms provide significantly improved diagnostic performance as compared to FLD. While the difference in diagnostic performance is particularly large for full set of spectral features, it is reduced for an optimal subset of features (selected by the respective RFE algorithms). The ROC analysis of the classification results provides a more critical evaluation. Figure 10 shows that while all three ROC curves corresponding to the SVM-based algorithms are very close to the point of ideal performance (i.e., the upper left-hand corner), the ROC curve corresponding to the FLD-based algorithm is quite far away from the ideal point. This is further supported by the observations of significantly higher values of the area under the ROC curves (Table 8) corresponding to the SVM-based algorithms as compared to that based on FLD with the performance of the RBF-SVM algorithm being the best.

The large improvement in diagnostic performance of SVM-based algorithms as compared to that based on classical FLD, appears to be due to the fact that while FLD extracts information from only the second-order correlations in the input spectral data[42] (covariance matrix) to enhance the class-discriminatory information, the SVMs use higher order

Table 5 Classification results of the FLD-based diagnostic algorithm for the training and the validation data sets with the 11 subsets of features selected through the RFE method.

		Training Data Set			Validation Data Set		
Diagnostic Algorithm	Number of Features	Sensitivity (%)	Specificity (%)	Misclassification Error (%)	Sensitivity (%)	Specificity (%)	Misclassification Error (%)
	717	60	44	96	37	48	115
	350	45	44	111	61	60	79
	175	45	53	102	46	52	102
	90	79	94	27	71	70	59
	45	**83**	**99**	**18**	**73**	**92**	**35**
FLD	25	76	96	28	76	92	32
	15	76	92	32	80	90	30
	10	81	96	23	68	91	41
	5	69	83	48	66	83	51
	2	67	82	51	71	82	47
	1	71	74	55	76	73	61

Sensitivity, specificity, and the misclassification error in the training set data are reported based on leave-one-out cross-validation. The row with bold figures indicates the optimal feature subset.

correlations.[24] Note also that FLD optimizes the transformation matrix by finding the largest ratio of between-class variation and within-class variation when projecting the original input data to a feature space of lower dimension.[42] Thus, it considers the squared separation between the means of each class and, therefore, is not expected to perform well on nonsymmetric data such as the LIF spectral data that may have multiple clusters per class.[45] This follows because when input data has multiple clusters per class it might so happen that the mean for a class of two clusters can lie close to the mean of another class.

From the viewpoint of pattern recognition, the task of tissue classification based on LIF spectral data is a pattern classification problem, and the feature vector for classification comprises the measured intensities corresponding to the different pixels (of the detector) that specify the dimension of the

Table 6 Classification results of all the diagnostic algorithms for the training data set and the two independent validation data sets with the full and optimal subsets of features selected through the RFE method.

		Training Data Set		Validation Data Set		
Diagnostic Algorithm	Number of Features	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Data Set I Specificity (%)	Data Set II Specificity (%)
FLD	Full	60	44	37	48	67
	Optimal subset	83	99	73	92	80
Linear SVM	Full	86	91	88	92	77
	Optimal subset	98	97	88	94	85
Polynomial SVM	Full	93	97	90	94	85
	Optimal subset	93	100	90	95	86
RBF SVM	Full	93	96	93	97	82
	Optimal subset	95	96	93	97	85

Sensitivity and specificity values in the training set data represent leave-one-out cross-validation values.

Table 7 Classification results for the training data set and the two independent validation data sets obtained with PCA-based algorithms and linear-SVM-based algorithms.

		Training Data Set		Validation Data Set		
Diagnostic Algorithm	Number of Features	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Data Set I Specificity (%)	Data Set II Specificity (%)
PCA+NMC	Four PCs	83	66	80	58	56
SVM-PCA	Four PCs	69	90	76	91	71
SVM	Full	86	91	88	92	73
SVM-RFE	Optimal subset	98	97	88	94	85

For the PCA-based algorithm, the diagnostically relevant PCs were classified using the NMC. For the SVM-based algorithm, classification results were obtained for two cases. In one case, SVM was used as a classifier with PCA providing the diagnostically relevant features (SVM-PCA), and in the second case, SVM was used for classification with the full set of spectral features as well as for both feature extraction and classification jointly using the SVM-RFE approach. Sensitivity and specificity values in the training set data represent leave-one-out cross-validation values.

features. If working directly with all these spectral features whose dimension is much higher ($N=717$ in this case) as compared to the size ($\sim$119 in this case) of the training samples, the classifier might suffer from the so-called ''curse of dimensionality,''[16] causing it to have poor generalization in classification performance. The use of RFE not only helps in choosing an optimal subset of features that are relevant for classification, but also reduces the feature dimension by solving the ''curse of dimensionality'' problem. This is evident from the observed large improvements in the classification performance of the diagnostic algorithm based on FLD for the optimal subset as compared to the full set of spectral features, leading to an increase of 29.5 and 49.5%, respectively, in the average sensitivity and specificity values. Note, however, that the diagnostic algorithms based on SVM are not too sensitive to the selection of optimal subset of features. For example, while for the linear SVM, the average sensitivity and specificity values improve by 6 and 4%, respectively, by going from the full set to the optimal subset of spectral features, no change in sensitivity and only a 2% increase in specificity values was observed for the polynomial SVM. For the RBF SVM, the resulting improvements in the average sensitivity and specificity values were found to be by 1 and 3.5%, respectively. This highlights the built-in capability of the SVM to sidestep overfitting to a large extent, despite the fact that it was trained on a set of training data where the number of features is large compared to the size of training patterns[24] (e.g., the present set of LIF spectral data). Further, an added advantage of using RFE is that it directly becomes clear which spectral regions dominate the classification problem, in contrast to using PCA for dimension reduction, where one must perform a component-loading operation to get back the spectral regions of interest. For example, the optimal subset of 350 features selected by the SVM RFE with RBF kernel correspond to wavelengths that span nearly the entire spectral region going from 355 to 700 nm.

Also note here that although the standard SVMs are designed for binary classification,[23,24] multiclass classification problems could be solved either by directly constructing a multiclass SVM classifier[36] or by using voting scheme methods based on combining many basic binary SVM decision functions.[36] For example, one-against-all decomposition is the most commonly used voting scheme method. In this case, the classification problem to k classes is decomposed to k binary SVM decision functions $f_m(x)$, $m \in K=\{1,...,k\}$, where the decision function $f_m(x)$ separates training data of the m'th class from the training patterns of other classes. The classification of a pattern x is performed according to maximal value of the functions $f_m(x)$, i.e., the label of x is computed as $\arg\max_{m\in K} f_m(x)$. The development SVM algorithms for multiclass classification is underway in our group. Here, the potential of the SVM to simultaneously classify spectral data into more than two classes comprising patients with various kinds of lesions of the oral cavity, for example, leukoplakia, erythroplakia, etc., in addition to cancerous and noncancerous tissues will be explored.

Note also here that the development of diagnostic algorithms described here was based on spectral data from patients who belonged to a high-risk population (were suspected

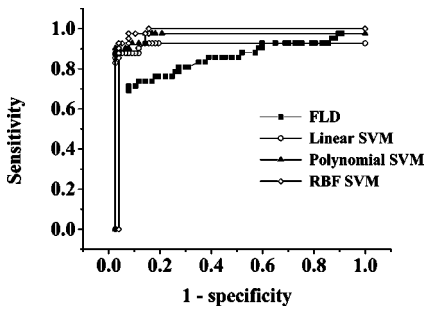

Fig. 10 ROC curves for different diagnostic algorithms based on SVMs (linear, polynomial, and RBF) and FLD.

Table 8 Area under the ROC curve values corresponding to the four diagnostic algorithms tested on the validation data set with optimal subset of features.

	FLD	Linear SVM	Polynomial SVM	RBF SVM
Area under the ROC curve	0.71	0.90	0.94	0.96

of having SCC on visual examination). This patient selection criteria might influence the sensitivity and specificity values obtained in this study. However, the motivation for this work was to compare the relative performance of the different types of diagnostic algorithms using the same spectral data set from the same population of patients. The patient selection criterion is unlikely to influence this comparison.

5 Conclusions

The application of an integrated framework of the SVM and RFE for discrimination of early squamous cell carcinoma from the normal squamous tissue sites of the oral cavity was reported. The flexibility of the framework of the SVM-RFE algorithm makes it convenient to conduct feature extraction and classification jointly, leading to improved classification results. Both linear- and nonlinear-SVM-based diagnostic algorithms were developed using spectral data acquired in a clinical *in vivo* LIF study conducted on patients being screened for cancer of the oral cavity and normal volunteers. The relative diagnostic performances of the algorithms have been evaluated and also compared with that of the classical FLD and PCA-based algorithms. The results show significantly improved classification performance of the integrated SVM-RFE algorithms as compared to both FLD and PCA-based algorithms.

Acknowledgments

The authors would like to thank Mr. S. Shyam Sunder, Mr. C. Rajan, and Mr. A. G. Bhujle for their contribution to the development of the clinical system and for several fruitful discussions, and Dr. M. S. Gujral and Dr. S. K. Kataria for providing the clinical LIF spectral data acquired using the N_2-laser-based portable flurimeter installed at the Government Cancer Hospital, Indore.

References

1. G. A. Wagnieres, W. M. Star, and B. C. Wilson, ''In vivo fluorescence spectroscopy and imaging for oncological applications,'' *Photochem. Photobiol.* **68**, 603–632 (1998).
2. E. Servick-Muraca and R. Richards Kortum, ''Quantitative optical spectroscopy for tissue diagnosis,'' *Annu. Rev. Phys. Chem.* **47**, 556–606 (1996).
3. N. Ramanujam, ''Fluorescence spectroscopy of neoplastic and non-neoplastic tissues,'' *Neoplasia* **2**(1), 1–29 (2000).
4. A. Mahadevan-Jansen and R. Richards-Kortum, ''Raman spectroscopy for the detection of cancers and precancers,'' *J. Biomed. Opt.* **1**(1), 31–70 (1996).
5. I. J. Bigio and I. R. Mourant, ''Ultraviolet and visible spectroscopies for tissue diagnostics: fluorescence spectroscopy and elastic scattering spectroscopy,'' *Phys. Med. Biol.* **42**, 803–814 (1996).
6. C. R. Kapadia, F. W. Cutruzzola, K. M. O'Brien, M. L. Stetz, R. Enriquez, and L. I. Deckelbaum, ''Laser-induced fluorescence spectroscopy of human colonic mucosa: detection of adenomatous transformation,'' *Gastroenterology* **99**, 150–157 (1990).
7. R. M. Marchesini, M. Brambilla, E. Pignoli, G. Bottiroli, A. C. Croce, M. D. Fante, P. Spinelli, and S. D. Palma, ''Light-induced fluorescence spectroscopy of adenomas, adenocarcinomas and non-neoplastic mucosa in human colon: in-vitro measurements,'' *J. Photochem. Photobiol., B* **14**, 219–230 (1992).
8. K. T. Schomacker, J. K. Frisoli, C. C. Compton, T. J. Flotte, J. M. Richter, N. S. Nishioka, and T. F. Deutsch, ''Ultraviolet laser-induced fluorescence of colonic tissue: Basic biology and diagnostic potential,'' *Lasers Surg. Med.* **12**, 63–78 (1992).
9. S. K. Majumder, A. Uppal, and P. K. Gupta, ''In-vitro diagnosis of human uterine malignancy using N_2 laser-induced autofluorescence spectroscopy,'' *Curr. Sci.* **70**(9), 833–836 (1996).
10. C. Y. Wang, C. T. Chen, C. P. Chiang, S. T. Young, S. N. Chow, and H. K. Chiang, ''A probability-based multivariate statistical algorithm for autofluorescence spectroscopic identification of oral carcinogenesis,'' *Photochem. Photobiol.* **69**(4), 471–477 (1999).
11. E. N. Atkinson, M. F. Mitchell, N. Ramanujam, and R. Richards-Kortum, ''Statistical techniques for diagnosing CIN using fluorescence spectroscopy: SVD and CART,'' *J. Cell Biochem. Suppl.* **23**, 125–130 (1995).
12. N. Ramanujam, M. Follen-Mitchell, A. Mahadevan-Jansen, S. Thomson, G. Staerkel, A. Malpica, T. Wright, N. Atkinson, and R. Richards-Kortum, ''Cervical precancer detection using multivariate statistical algorithm based on laser-induced fluorescence spectra at multiple excitation wavelengths,'' *Photochem. Photobiol.* **64**, 720–735 (1996).
13. A. Zuluaga, U. Utzinger, H. Durkin, A. Fuchs, A. Gillenwater, R. Jacob, B. Kemp, J. Fan, and R. Richards-Kortum, ''Fluorescence excitation-emission matrices of human tissue: a system for in-vivo measurement and method of data analysis,'' *Appl. Spectrosc.* **53**, 302–311 (1999).
14. D. Heintzelmann, U. Utzinger, H. Fuchs, A. Zuluaga, K. Gossage, A. M. Gillenwater, R. Jacob, B. Kemp, and R. Richards-Kortum, ''Optimal excitation wavelengths for in-vivo detection of oral neoplasia using fluorescence spectroscopy,'' *Photochem. Photobiol.* **72**(1), 103–113 (2000).
15. S. K. Majumder, S. K. Mohanty, N. Ghosh, P. K. Gupta, D. K. Jain, and F. Khan, ''A pilot study on the use of autofluorescence spectroscopy for diagnosis of the cancer of human oral cavity,'' *Curr. Sci.* **79**(8), 1089–1094 (2000).
16. A. K. Jain, R. P. W. Duin, and J. Mao, ''Statistical pattern recognition: a review,'' *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(1), 4–37 (2000).
17. A. K. Jain, J. Mao, and K. M. Mohiuddin, ''Artificial neural networks: a tutorial,'' *Computer*, **29**(3), 628–633 (1996).
18. H. J. van Staveren, R. L. van Veen, O. C. Speelman, M. J. Witjes, W. M. Star, and J. L. Roodenburg, ''Classification of clinical autofluorescence spectra of oral leukoplakia using an artificial neural network: a pilot study,'' *Oral Oncol.* **36**(3), 286–293 (2000).
19. K. Tumer, N. Ramanujam, J. Ghosh, and R. Richards-Kortum, ''Ensembles of radial basis function networks for spectroscopic detection of cervical pre-cancer,'' *IEEE Trans. Biomed. Eng.* **45**(8), 953–961 (2001).
20. G. A. Rovithakis, A. N. Maniadakis, M. Zervakis, G. Fillipidis, G. Zakarakis, A. N. Katsamouris, and T. G. Papazoglou, ''Artificial neural networks for discriminating pathologic from normal peripheral vascular tissue,'' *IEEE Trans. Biomed. Eng.* **48**(10), 1088–1096 (2001).
21. N. Agrawal, S. Gupta, Bhawna, A. Pradhan, K. Viswanath, and P. K. Panigrahi, ''Wavelet transform of breast tissue fluorescence spectra—a technique for diagnosis of tumors,'' *IEEE J. Sel. Top. Quantum Electron.* **9**(2), 154–161 (2003).
22. S. K. Majumder, N. Ghosh, S. Kataria, and P. K. Gupta, ''Nonlinear pattern recognition for laser-induced fluorescence diagnosis of cancer,'' *Lasers Surg. Med.* **33**, 48–56 (2003).
23. V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York (1998).
24. C. J. C. Burges, ''A tutorial on support vector machines for pattern recognition,'' *Data Min. Knowl. Discov.* **2**(2), 121–167 (1998).
25. C. Cortes and V. N. Vapnik, ''Support vector networks,'' *Mach. Learn.* **20**, 273–297 (1995).
26. M. S. Schmidt and H. Gish, ''Speaker identification via support vector classifiers, in *Proc. 21st IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP-96)*, pp. 105–108, Atlanta, GA (1996).
27. E. Osuna, R. Freund, and F. Girosi, ''Training support vector machines: an application to face detection,'' in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 130–136 (1997).
28. T. Joachims, ''Text charactrization with support vector machines,'' Technical Report LS VIII Number 23, University of Dortmund (1997).
29. M. Bonneville, J. Meunier, Y. Bengio, and J. P. Soucy, ''Support vector machines for improving the classification of brain pet images,'' *Proc. SPIE* **3338**, 264–273 (1998).
30. S. S. Keerthi, S. K. Shevade, C. Bhattacharya, and K. R. K. Murthy, ''A fast iterative nearest point algorithm for support vector machine classifier design,'' *IEEE Trans. Neural Netw.* **11**, 124–136 (2000).
31. L. Mangasarian, ''Generalized support vector machines,'' in *Advances in Large Margin Classifiers*, A. Smola, P. Barltlett, B.

Scholkopf, and D. Schuurmans, Eds., MIT Press, Cambridge, MA (2000).
32. L. Mangasarian and D. R. Musicant, ''Lagrangian support vector machines,'' *J. Mach. Learn. Res.* **1**, 161–177 (2001).
33. G. M. Palmer, C. Zhu, T. M. Breslin, F. Xu, K. W. Gilchrist, and N. Ramanujam, ''Comparison of multiexcitation fluorescence and diffuse reflectance spectroscopy for the diagnosis of breast cancer (March 2003),'' *IEEE Trans. Biomed. Eng.* **50**(11), 1233–1242 (2003).
34. W. M. Lin, X. Yuan, P. Yuen, W. I. Wei, J. Sham, P. C. Shi, and J. Qu, ''Classification of in-vivo autofluorescence spectra using support vector machines,'' *J. Biomed. Opt.* **9**(1), 180–186 (2004).
35. I. Guyon, J. Weston, S. Barnhill, and V. N. Vapnik, ''Gene selection for cancer classification using support vector machines,'' *J. Mach. Learn. Res.* **46**, 389–422 (2002).
36. J. Weston and C. Watkins, ''Multi-class support vector machines,'' Technical Report CSD-TR-98_04, Department of Computer Science, Royal Holloway, University of Lonsdon, Egham, TW20 0EX, UK (1998).
37. S. K. Majumder, A. Uppal, and P. K. Gupta, ''Autofluorescence spectroscopy of tissues from human oral cavity for discriminating malignant from normal,'' *Lasers Life Sci.* **8**, 211–227 (1999).
38. A. Uppal and P. K. Gupta, ''Measurement of NADH concentration in normal and malignant human tissues from breast and oral cavity,'' *Biotechnol. Appl. Biochem.* **37**, 45–50 (2003).
39. N. Ghosh, S. K. Mohanty, S. K. Majumder, and P. K. Gupta, ''Measurement of optical transport properties of malignant and normal human breast tissues,'' *Appl. Opt.* **40**, 176–184 (2001).
40. C. K. Brookner, U. Utzinger, G. Staerkel, R. Richards-Kortum, and M. F. Mitchell, ''Cervical fluorescence of normal women,'' *Lasers Surg. Med.* **24**, 29–37 (1999).
41. V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, New York (1982).
42. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 2nd. ed., pp. 340–377, Prentice-Hall International, Englewood Cliffs, NJ (1988).
43. L. Hermes and J. M. Buhmann, ''Feature selection for support vector machines,'' in *Proc. 15th International Conf. on Pattern Recognition*, Vol. 2, pp. 716–719 Barcelona, Spain (2000).
44. C. E. Metz, ''Basic principles of ROC analysis,'' *Semin. Nucl. Med.* **8**, 283–298 (1978).
45. A. Talukder and D. Casasent, ''A closed-form neural network for discriminatory feature extraction from high dimensional data,'' *Neural Networks* **14**, 1201–1218 (2000).