Lasers in Surgery and Medicine 36:323–333 (2005)

Relevance Vector Machine for Optical Diagnosis of Cancer

Shovan K. Majumder, PhD,* Nirmalya Ghosh, MTech, and Pradeep K Gupta, PhD
Biomedical Applications Section, Centre for Advanced Technology, Indore 452013, India

Background and Objectives: A probability-based, robust diagnostic algorithm is an essential requirement for successful clinical use of optical spectroscopy for cancer diagnosis. This study reports the use of the theory of relevance vector machine (RVM), a recent Bayesian machine-learning framework of statistical pattern recognition, for development of a fully probabilistic algorithm for autofluorescence diagnosis of early stage cancer of human oral cavity. It also presents a comparative evaluation of the diagnostic efficacy of the RVM algorithm with that based on support vector machine (SVM) that has recently received considerable attention for this purpose.

Study Design/Materials and Methods: The diagnostic algorithms were developed using in vivo autofluorescence spectral data acquired from human oral cavity with a N_2 laser-based portable fluorimeter. The spectral data of both patients as well as normal volunteers, enrolled at Out Patient department of the Govt. Cancer Hospital, Indore for screening of oral cavity, were used for this purpose. The patients selected had no prior confirmed malignancy and were diagnosed of squamous cell carcinoma (SCC), Grade-I on the basis of histopathology of biopsy taken from abnormal site subsequent to acquisition of spectra. Autofluorescence spectra were recorded from a total of 171 tissue sites from 16 patients and 154 healthy squamous tissue sites from 13 normal volunteers. Of 171 tissues sites from patients, 83 were SCC and the rest were contralateral uninvolved squamous tissue. Each site was treated separately and classified via the diagnostic algorithm developed. Instead of the spectral data from uninvolved sites of patients, the data from normal volunteers were used as the normal database for the development of diagnostic algorithms.

Results: The diagnostic algorithms based on RVM were found to provide classification performance comparable to the state-of-the-art SVMs, while at the same time explicitly predicting the probability of class membership. The sensitivity and specificity towards cancer were up to 88% and 95% for the training set data based on leave-one-out cross validation and up to 91% and 96% for the validation set data. When implemented on the spectral data of the uninvolved oral cavity sites from the patients, it yielded a specificity of up to 91%.

Conclusions: The Bayesian framework of RVM formulation makes it possible to predict the posterior probability of class membership in discriminating early SCC from the normal squamous tissue sites of the oral cavity in contrast to dichotomous classification provided by the non-Bayesian SVM. Such classification is very helpful in handling asymmetric misclassification costs like assigning different weights for having a false negative result for identifying cancer compared to false positive. The results further demonstrate that for comparable diagnostic performances, the RVM-based algorithms use significantly fewer kernel functions and do not need to estimate any hoc parameters associated with the learning or the optimization technique to be used. This implies a considerable saving in memory and computation in a practical implementation. Lasers Surg. Med. 36:323–333, 2005.

Key words: diagnostic algorithm; oral cancer; posterior probability; relevance vector machine (RVM); squamous cell carcinoma (SCC); support vector machine (SVM)

INTRODUCTION

Recent research has demonstrated the applicability of optical spectroscopic technology for non-invasive, in situ, near-real time diagnosis of cancer [1–4]. The approach requires a suitable diagnostic algorithm that can best classify the measured spectra from an unknown tissue by using a stored database of spectra of tissues of known histopathologic classification. Over the years, a variety of diagnostic algorithms of varying rigor have been developed for optical diagnosis of cancer [5–30]. Most of the earlier algorithms are based on empirically selected indices like absolute or normalized fluorescence intensities [5–11], ratio of intensities at selected pairs of emission wavelengths [12–16], or ratio of integrated intensities over appropriately chosen wavelength bands [17]. Recent efforts are directed towards using statistical pattern recognition techniques [18–30] to exploit the entire spectral information content of the full range of spectral data for extracting the best diagnostic features and accurately classifying them into corresponding histopathologic categories. Although traditional linear techniques like principal component analysis (PCA), Fisher's linear discriminant (FLD), etc. [18–23] have been used for this purpose; use of sophisticated, state-of-the-art techniques [24–30] is receiving increasing attention for their superior performance. These include artificial neural network

*Correspondence to: Dr. Shovan K. Majumder, Biomedical Applications Section, R & D Block-D, Centre for Advanced Technology, Indore 452 013, India. E-mail: shkm@cat.ernet.in
Accepted 7 February 2005
Published online 11 April 2005 in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/lsm.20160

(ANN) [24–26], wavelet transforms [27], maximum representation and discrimination feature (MRDF) [28], and more recently support vector machine (SVM) [29,30]. Amongst all these, SVM, in particular, is the best suited for this kind of supervised classification problems [31]. The central idea of SVM is to map a set of input data to a high-dimensional feature space through a kernel function and separate classes in the kernel induced feature space with a maximum margin hyperplane that maximizes the minimum distance from the hyperplane to the closest input data points [32]. In general, the hyperplane corresponds to a non-linear decision boundary in the input space and depends only on a subset of the original input data called the support vectors [31]. The formulation of the technique relies on the theory of uniform convergence in probability and associated structural risk minimization (SRM) principle [32]. Palmer et al. [29] have used a linear SVM classifier for classifying in vitro autofluorescence and diffuse reflectance spectra of breast tissues and reported excellent classification results. Lin et al. [30] have used SVM to classify nasopharingeal tissues based on features extracted using linear PCA of in vivo autofluorescence spectra from nasopharingeal tissues and demonstrated significantly improved classification performance of combined SVM-PCA algorithm as compared to that based on linear PCA alone.

Although SVMs and other state-of-the-art techniques have been very successful in correctly identifying the class membership of a tissue from its recorded spectra, a major drawback of all these approaches is that they cannot provide a posterior probability of classification of the tissue to different classes. Such classification is particularly important in the context of asymmetric misclassification costs where the misclassification cost associated with some classes (false negative for cancer) may be significantly higher than that of others (false positive for cancer). Therefore, in clinical settings, the posterior probabilities of class membership need to be explicitly computed in order to handle asymmetric misclassification costs in a principled theoretical framework. The goal of the present study is to report, for the first time to our knowledge, the application of the theory of relevance vector machine (RVM) [33], a recent Bayesian machine-learning framework of statistical pattern recognition, for development of a probability based diagnostic algorithm for autofluorescence diagnosis of cancer. The in vivo autofluorescence spectral data recorded from the oral cavity of patients (with oral cancer) as well as of normal volunteers were used for this purpose. Both linear and non-linear RVMs were used for development of algorithms. The algorithms were compared with that developed using equivalent SVMs based on the same spectral data set. RVM-based algorithms not only showed diagnostic performance comparable to SVM, but also provided a principled estimate of posterior probabilities of class membership. Further, while the SVMs required an a priori estimation of a regularization parameter, the RVMs did not need to estimate any hoc parameters associated with the learning or the optimization technique to be used [33]. This considerably speeded up the training phase of algorithm due to lack of necessity to perform cross-validation over ad-hoc parameters that is wasteful both of data and computation.

MATERIALS AND METHODS

In vivo autofluorescence spectra were recorded using a N_2 laser (337 nm) based portable fluorimeter reported earlier [23,28]. It comprised a sealed-off pulsed N_2 laser, a spectrograph (Acton Research Corporation, 15 Discovery Way, Acton, MA), an optical fiber probe, and a gateable intensified CCD detector (4 Quik 05A, Stanford computer optics, Inc., Berkeley, CA). The spectral data acquisition was computer controlled. From each site, spectra were recorded in the 375–700-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, India. Informed consent was obtained from each patient as well as the normal volunteers who participated in this study. A medical history was also obtained from them noting their age, sex, and habits related to smoking. It was observed that the ratio of male to female population was $\sim$2 and the mean age was 46 ± 12 years ranging from a minimum of 24 years to a maximum of 70 years. 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. From these patients, biopsies were taken from the suspected areas subsequent to acquisition of spectra. Only those patients were included in this study for whom histopathological diagnosis was squamous cell carcinoma (SCC), Grade-1. 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.

During recording of the in vivo autofluorescence spectra, the tip of the fiber-optic probe was placed in gentle contact with the tissue surface and it was ensured that none of the patients or the normal volunteers complained of the probe being painful. The spectra were always recorded by a single

person to minimize variations of probe pressure induced by personal measuring styles. This is expected to reduce the resulting site-to-site variability in the measured spectra that might obscure the intercategory spectral differences to be exploited for diagnosis. However, it is pertinent to note that a detailed study has been carried out recently by Nath et al. [34] to investigate the effect of probe pressure on cervical tissue fluorescence and it has been shown that variation in probe pressure does not significantly affect the diagnostic results.

Spectral Data

The autofluorescence spectra recorded from different cancerous and contralateral normal sites of the oral cavity of a patient are shown in Figure 1a,b, respectively. The considerable site-to-site variation in intensity and line shape of the spectra is apparent. While some of this variation represents intrinsic variation in tissue fluorescence, the variable nature of the contact of the probe with the tissue surface in the clinical situation has also added to the variation. It is important to note that in contrast to our earlier in vitro studies on oral cavity tissues [8], where a percentage variation in the spectrally integrated intensities (ΣI) from different sites of normal or cancerous tissues was only ~30%, the percentage variation in ΣI observed in the present in vivo study was ~60% over the total patient size investigated. In order to ensure good discrimination, it was necessary to minimize these variations that might obscure the intercategory differences. In order to do that, 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 displays the spectra for cancerous and uninvolved sites of the oral cavity of the same patient after preprocessing. However, it is pertinent to emphasize here that, the differences in the preprocessed spectra from cancerous and contralateral uninvolved tissue sites of the same patient are generally more distinct [20,23] as compared to the differences when preprocessed spectra from similar tissue sites of all the patients are considered as

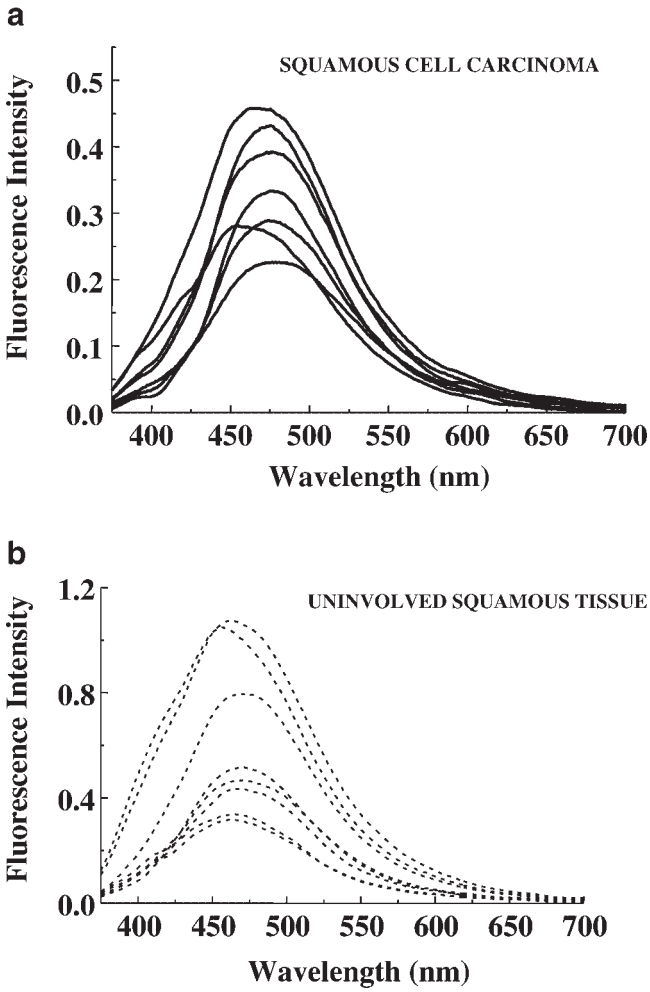

Fig. 1. N_2 laser-excited autofluorescence spectra recorded from (**a**) squamous cell carcinoma tissue sites (solid line) and (**b**) uninvolved tissue sites (dashed line) of the same patient.

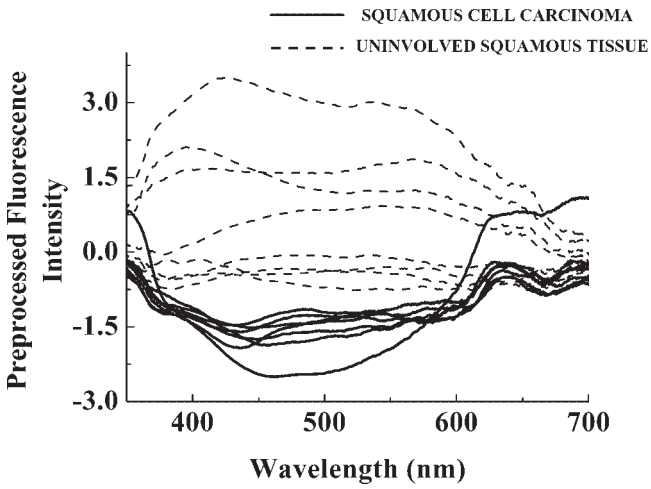

Fig. 2. Preprocessed autofluorescence spectra from squamous cell carcinoma tissue sites (solid line) and from uninvolved squamous tissue sites (dashed line) of the oral cavity of the same patient.

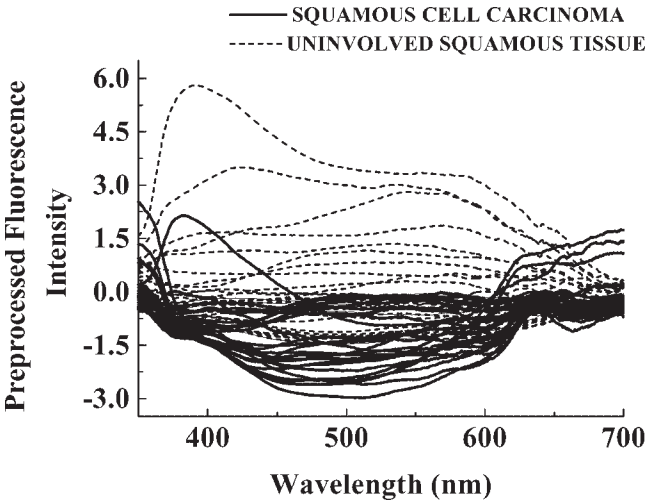

Fig. 3. Preprocessed autofluorescence spectra from squamous cell carcinoma tissue sites (solid line) and from uninvolved squamous tissue sites (dashed line) of the oral cavity of four patients chosen at random.

a whole. Figure 3 displays the pre-processed spectra from cancerous and contralateral normal tissue sites of four patients chosen at random. It is evident from the figure that the inter-patient differences in the preprocessed spectra do not appear to be that prominent in comparison with the intra-patient differences shown in Figure 2. It also justifies the need to further develop a sophisticated diagnostic algorithm for classification.

Splitting of Spectral Data: Training Set and Validation Set

Prior to 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 were randomly split into two groups: training data set and 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. In order to address this issue, in the present study, the use of two separate normal databases in the training set was investigated, the cancer database keeping same for both. In one, the spectral data of contralateral uninvolved tissue sites of patients were taken as the normal database, while in the second, the spectral data of healthy squamous tissue sites of normal volunteers were considered as the normal database. The validation data set was identical in both the cases and comprised spectral data from cancerous tissue sites of patients and healthy squamous tissue sites of normal volunteers. Our initial results showed that use of spectral data of normal volunteers in the training set gave improved classification performance in the validation data set with an increase of $\sim$5% in sensitivity and of $\sim$7% in specificity. However, when the spectral data of uninvolved tissue sites were used for validation, the specificity values were observed to rather decrease by $\sim$6–10%. This is not surprising, because in the case of statistical decision theoretic approach for supervised classification, the classifier learns the necessary information for future classification from the given training set data at hand. If the data in the training set are not true representatives of the future test set data, classification errors are inevitable [35]. In the present context, it means that spectral data from many of the uninvolved tissue sites of patients assumed to be normal (based on visual examination, since no histopathological confirmation was possible) during the training phase might not be truly normal due to the field effect of malignancy [36]. In contrast, this possibility did not exist for the squamous tissue sites from normal volunteers who had no history of any disease of oral cavity. Due to this reason, for subsequent development of diagnostic algorithms, the spectral data from the healthy squamous tissue sites of the normal volunteers were used as the normal database in the training set instead of that from the tissue sites of normal appearing mucosa in the contralateral uninvolved region of the oral cavity of patients.

Development of Diagnostic Algorithm

Given the d-dimensional (d being the number of wavelengths over which spectra were recorded), training set data of laser-induced fluorescence (LIF) spectra belonging to cancerous and normal squamous tissue sites, the task of a diagnostic algorithm is to separate this set of input data into its constituent classes, and also to predict the true class-membership of a tissue spectrum that is not a part of the training set. A simple way to build a classifier is to construct a hyperplane (decision boundary) in the d-dimensional input space that separates class members from non-members considered as points in that space. A look at the LIF spectral data (see Fig. 3) would show that because of considerable intercategory overlap, there exists no separating hyperplane in the input space that successfully separates the cancerous from the normal spectra. 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 non-linear-mapping and construct a separating hyperplane that is linear in that space, but is non-linear with respect to the input space [31,35]. However, the technical difficulty involved in mapping the training set data to a higher-dimensional space for classification is twofold [31]: one is the computational burden and the other is the possible risk of finding trivial solutions that may overfit the data, that is, there may exist infinitely many hyperplanes that can successfully separate the training set data, but may perform miserably on unseen (test) data points. Although

several approaches are being pursued to simultaneously sidestep both these difficulties, the methodologies based on SVM developed by Vapnik [31,32] and more recently RVM developed by Tipping [33] are the two most successful approaches that have become widely established to date.

Support Vector Machine (SVM)

An SVM [31] avoids overfiiting by choosing an optimal separating hyperplane (OSH) in the feature space (from among the many) that maximizes the width of the margin between classes thereby following the SRM principle [32] of statistical learning and makes predictions based on a function of the form

$$f(x) = \sum_{i=1}^{N} w_i K(x, x_i) + w_0 \tag{1}$$

where $K(x,x_i)$ is a kernel function effectively defining one basis function for each data point in the training set and $\{w_i\}$ are the model weights reflecting the importance of the training set data points, which specify the location of the OSH in the feature space. Those training set data points that lie far away from the OSH do not participate in its specification and therefore receives weights of zero. Only the training set data points that lie close to the decision boundary between the classes receive non-zero weights [31,32]. These training set data points are called "support vectors" [31], since only these points define the classification boundary and removing them would change the location of the OSH. The introduction of kernel functions in the SVM framework enables one to define the feature space implicitly and thus overcomes the problem of computational burden of explicitly mapping the input data to the higher-dimensional feature space via non-linear mapping [31]. However, in order to qualify as a legitimate kernel, the kernel function must have to satisfy Mercer's condition, that is, it must need to be a continuous symmetric kernel of a positive integral operator [31,32]. As long as the kernel function is legitimate, 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.

Relevance Vector Machine (RVM)

Unlike an SVM, an RVM [33,37,38] is a fully probabilistic model (based on Bayesian maximum a priori (MAP) estimation framework) whose objective is to separate the set of input data into its constituent classes by predicting the posterior probabilities of their class-membership. The prediction is based on a decision function identical in functional form to the SVM as shown in Equation (1). Thus, training an RVM essentially involves estimation of appropriate values of the weights (w) associated with the kernel functions. A preference for a sparse representation is encoded in the RVM by defining the prior distribution over the weights (w) as a zero-mean Gaussian distribution:

$$p(w|\alpha) = \prod_{i=0}^{N} N\left(w_i \middle| 0, \alpha_i^{-1}\right) = \prod_{i=0}^{N} \sqrt{\frac{\alpha_i}{2\pi}} \exp\left(-\frac{\alpha_i w_i^2}{2}\right) \tag{2}$$

where the $N+1$ parameters $\{\alpha_i\}_{i=0,N} > 0$ (N being the size of the training set data) are the inverse variances that control the width of the Gaussian distributions over the corresponding weight. The "RVM trick" is to use α as variable parameters and to infer their values from the data. It is, therefore, necessary to provide additional hyperpriors over the values of these priors. For the hyperprior, the RVM formulation uses a non-informative prior, implying that before the inference operation we have no knowledge of what the parameters are likely to be. This form of prior is known as Automatic Relevance Determination (ARD) prior [33]. Generally, a Gamma distribution, with its parameters chosen to make it essentially flat over a wide range of "reasonable" values of α, is used as the non-informative hyperpriors over α. The introduction of an individual hyperparameter with an ARD prior (for every weight) is the key feature of the RVM formulation, and ultimately makes it possible to achieve sparsity in practice: during the iterative optimization process of the learning phase, many of the α_i are driven to very large values so that the associated posterior probabilities of the corresponding weights (w) close to zero become extremely high, implying that the corresponding model weights w_i can be effectively pruned out. Those training vectors that are associated with non-zero weights correspond to the most relevant training data points since they capture the data's underlying distribution. These are called 'relevance vectors' (motivated by the principle of automatic relevance determination) and represent 'prototypical' examples of respective classes [33].

The advantage of the RVM approach is that unlike in the case of SVM, there is no restriction on the choice of the kernel functions [33]. The kernel function is simply viewed as a basis function. Its choice determines the type of the RVM classifier and also defines the feature space in which the training set data points are classified. Given a set of training data points 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: $K(x_i,x) = x_i \cdot x + 1$, resulting in a linear classifier. Similarly, use of Gaussian radial basis functions results in a radial basis function (RBF) kernel: $K(x_i, x) = \exp(-\|x_i - x\|^2)(/2\sigma^2)$, where σ is the width of the Gaussian. Both these kernels are Mercer kernels [31] and they were specifically chosen in the present study because the objective was to compare the diagnostic efficacy of a RVM classifier with that of its SVM counterpart.

The optimal value for the width σ in the Gaussian RBF kernel is decided by optimizing the cost function defined for the application. The misclassification error obtained with leave-one-out cross validation of the training set data was used as the cost function. In cases where the total number of misclassified samples was the same for more than one σ value, the value for σ, for which the total number of cancerous samples misclassified was minimum was chosen as the optimal value. Both the RBF-RVM and the RBF-SVM classifiers were trained on the spectral data of the training set for the different σ values selected from a set of σ values ranging from 0.1 to 1,000 with increments of 0.1 for σ values between 0 and 1, with increments of 1 for σ values

between 1 and 20, with increments of 5 for σ values between 20 and 100 and with increments of 100 for σ values between 100 and 1,000. Optimal value of σ was the one that gave the least leave-one-out cross validation error.

Analysis of Algorithm Performance

In order to critically evaluate the relative performance of the diagnostic algorithms developed using RVM and SVM formulations, a receiver-operating characteristic (ROC) curve [39] corresponding to of each them was generated for the validation data and an ROC analysis was carried out for their corresponding classification results. The rationale behind this is the fact that an ROC curve, being a plot of the true positive rate (sensitivity) as a function of the false positive rate (1-specificity) for varying classification thresholds, provides a qualitative comparison of the trade-off between sensitivity and specificity of a diagnostic test. Further, the area under the ROC curve is indicative of the accuracy with which an algorithm can separate a set of data being tested into the different classes thereby providing a quantitative performance measure of the algorithm. 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 [39]. Similarly, the closer the area equals to 1, the more accurate is the corresponding diagnostic algorithm [39].

RESULTS AND DISCUSSIONS

Table 1 lists the sensitivity and specificity values for the training and the validation data sets obtained using linear RVM as well as SVM classifiers. For comparison sake, the classification results yielded by a conventional nearest-mean classifier on the same data sets are also listed in the same table. A nearest mean-classifier 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 both the RVM and the SVM outperform the nearest mean classifier for both the data sets. The superior classification performance of the RVM and SVM classifier originates from the built-in capability of these approaches to separate classes, which are not linearly separable in the original parametric space [31,33].

Figure 4 demonstrates the leave-one-out cross validation error as a function of the widths (σ) of the Gaussian RBF kernel for the RVM and the SVM classifiers. From the figure, it is clear that while the leave-one-out error is the minimum for $\sigma = 26$ for the RBF-RVM classifier, it is minimum at more than one σ values (e.g., at $\sigma = 50$, 75, and 100) for the RBF SVM classifier. However, for the σ value of 100, the total number of cancerous samples misclassified was the minimum. Therefore, for subsequent algorithm development with the RBF-SVM classifier, $\sigma = 100$ was chosen as the width of the RBF kernel, whereas $\sigma = 26$ was used as the width for subsequent training of the RBF-RVM classifier.

In order to train an SVM algorithm, one needs a priori estimation of a regularization parameter (associated with the learning) C controlling the trade-off between the training error and width of the margin between classes. Since there exists no established guideline in the SVM methodology [31,32] for determining the optimal value of C, a cross-validation procedure was employed on training the non-linear SVM classifier with different values of C ($C = 1$, 10, 100, and ∞). The classifier with $C = \infty$ was found to give the best generalized classification performance, that is, the total misclassification error over the training (leave-one-out cross validation) and the independent validation data sets was the least. It is clearly a disadvantage since the procedure is wasteful both of data and computation. In contrast, the RVM approach does not need any additional parameters like "C" to set beforehand, apart from the need to choose the type of the kernel and associated parameters [33].

Table 2 lists sensitivity, specificity, false negative, and false positive values yielded by the RVM as well as the SVM based diagnostic algorithms for the training (on the basis of leave-one-out cross validation) and the validation data sets. It is evident from the table that both the RVM as well as the SVM classifiers with RBF kernel have outperformed the respective linear ones. However, the diagnostic performance of the RVM-based algorithms is seen to be largely comparable to that of the SVM based ones with SVM based ones providing marginally improved perfor-

TABLE 1. Classification Results Provided by the Linear RVM and SVM Classifiers and the Conventional Nearest Mean Classifier (NMC)

	Training data set		Validation data set		
			Data set-I		Data set-II
Classifiers	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Specificity (%)
RVM	84	93	86	96	91
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.

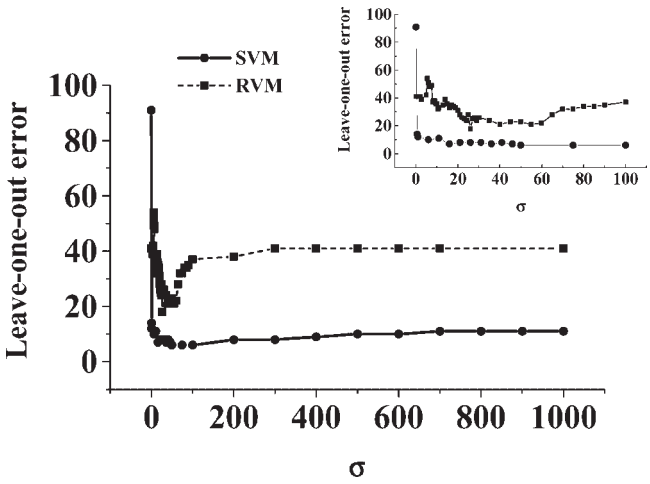

Fig. 4. Leave-one-out cross validation error in the training set data as a function of the width of the Gaussian radial basis function for the RBF-RVM and the RBF-SVM classifiers.

mance in some cases. The receiver-operating-characteristic (ROC) analysis of the classification results (Fig. 5 and Table 3) provides a more critical evaluation. Figure 5 shows that both the ROC curves corresponding to the SVM- and RVM-based algorithms are very close to the point of ideal performance (i.e., the upper left-hand corner). This is further supported by the observations of similar values of the area under the ROC curves (Table 3) corresponding to the algorithms based on RVM and SVM.

Although the diagnostic performance of SVMs are similar to that of RVMs, SVM suffers from the major limitation in that it makes explicit classifications and cannot provide a quantitative estimate for the confidence with which a site is classified in a specific group (normal or malignant, in the present case). This problem has been addressed to very recently by attempting a probabilistic prediction for the SVM classification through a post-processing strategy by fitting, a posteriori, a sigmoid function to the fixed SVM output [40]. However, this approximate probability has turned out to be very different from the true posterior probability of classification [33]. In contrast, the RVM approach, being based on Bayesian formulation, predicts posterior probability of class membership in a principled manner [33]. Figure 6 plots the posterior probabilities predicted by the RBF-RVM algorithm for the spectra of tissue sites comprising the two independent validation data sets of being classified as SCC. Such probabilistic feedback always facilitates separation of 'inference' and 'decision' [41] and would prove to be extremely useful in practical situations to judiciously compensate for asymmetric misclassification costs (which nearly always apply in real applications) and varying class proportions, attempt to improve performance by rejection of the more ambiguous data points, and explore the possibility of the fusion of outputs with other probabilistic sources of information before applying decision criteria. A further advantage of the RVM formulation as probabilistic generalized linear model is that it can be extended to

TABLE 2. Classification Results of the RVM- and SVM-Based Diagnostic Algorithms for the Training Data Set and the Two Independent Validation Data Sets

	Training set data				Validation set data					
	(42 SCC tissue sites from patients and 77 healthy squamous tissue sites from normal volunteers)				Data set-I (41 SCC tissue sites from patients and 77 healthy squamous tissue sites from normal volunteers)				Data set-II (88 uninvolved tissue sites from patients)	
Diagnostic algorithm	Sensitivity (%)	False negative (%)	Specificity (%)	False positive (%)	Sensitivity (%)	False negative (%)	Specificity (%)	False positive (%)	Specificity (%)	False positive (%)
Linear RVM (no. of RV = 9)	84	16	93	7	86	14	96	4	89	11
RBF-RVM (no. of RV = 10)	88	12	95	5	91	9	95	5	91	9
Linear SVM (no. of SV = 30)	86	14	91	9	86	14	92	8	77	23
RBF-SVM (no. of SV = 27)	93	7	96	4	93	7	95	5	82	18

Sensitivity and specificity values in the training set data represent leave-one-out cross validation values. The number of relevance vectors (RV) and the support vectors (SV) generated by the respective classifiers are also listed in the table. Sensitivity is the ratio of the number of diseased tissue sites correctly diagnosed to the total number of tissues sites with the disease in question, and specificity is the ratio of normal tissue sites correctly diagnosed to the total number of normal tissue sites investigated. False negative is the ratio of the number of disease tissue sites misdiagnosed to the total number of tissue sites with the disease in question, and false positive is the ratio of the normal tissue sites misdiagnosed to the total number of normal tissue sites investigated.

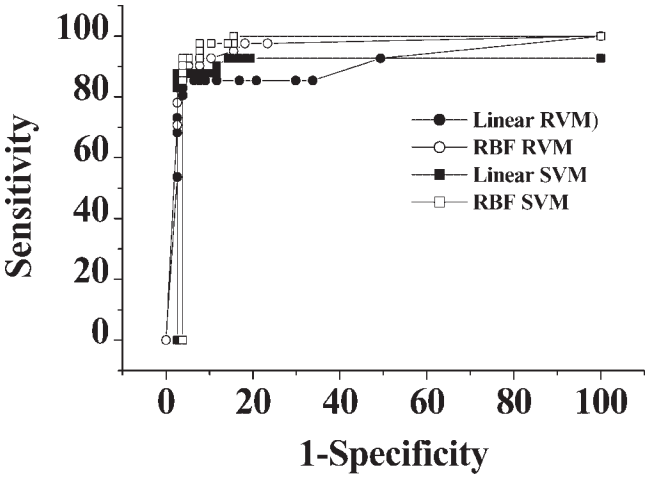

Fig. 5. Receiver-operating characteristic (ROC) curves for RVM and SVM based diagnostic algorithms.

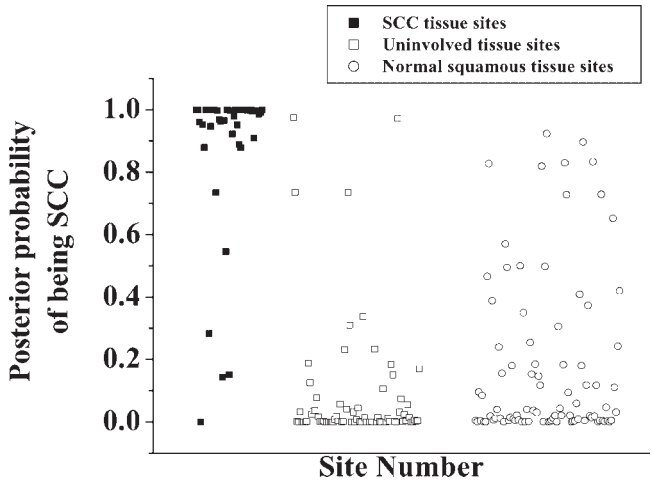

Fig. 6. Posterior probabilities of being classified as squamous cell carcinoma (SCC) for the spectra of tissue sites comprising the independent validation data sets.

the multiple-class case in a straightforward and principled manner [33], without the need to train and heuristically combine multiple dichotomous classifiers, as is standard practice for the SVM [42]. This would facilitate a rapid classification of spectral data simultaneously into more than two classes in situations, where one would deal with patients with various kinds of lesions of oral cavity, for example, leukoplakia, erythroplakia, etc. in addition to cancerous and non-cancerous.

An important task during development of any statistical algorithm for supervised classification is to evaluate the generalized classification ability of the algorithm. One should ideally consider single spectrum per individual to ensure complete independence of the full set of spectral data for that purpose. However, since it requires participation of enormously large number of individuals, difficult to arrange in many practical situations, one is left with no option but to use limited spectral data at disposal for algorithm development. The independence, in such cases, is approximated by following either the holdout method, where the available data is split into two subsets, one for training and other for testing; or the leave-one-out-method, where training is performed using $N-1$ samples (N being the size of the data set) and test is carried out only on the excluded sample [35]. Both these methods have been followed in the present study. For the holdout method, since there are no good guidelines available on how to divide the available data into training and test sets [35], the full set of spectral data was randomly split into training and validation subsets as was also done by Ramanujam et al. [20] while developing diagnostic algorithms for autofluorescence diagnosis of cervical precancer. Thus, classification results obtained in the present study can be considered reliable in predicting future classification performance. Moreover, the mathematical formulation of both the SVM and the RVM approaches is such that they are robust enough to generalize well on previously unseen data [32,33] despite getting trained on a set of data that is limited in size and have some sort of correlation. This is evident from a look at the diagnostic results in Table 4, where performances of the algorithms have been listed for two separate cases. In one, the training data set comprised spectral data from nine patients and seven normal volunteers and the validation data set comprised spectral data from the remaining seven patients and six normal volunteers. In the other, the training and the validation data sets comprised randomly split spectral data from the 16 patients and the 13 normal volunteers. It is apparent from the table that the performances of the algorithms for the two cases are comparable. The key mechanism responsible for such generalized classification ability of the algorithms is the sparsity of representation of their respective decision boundaries [31,33]. This means that in instead of memorizing the full set of training data, both the algorithms require only a small subset of training data (relevance vectors in the case of RVM and support vectors in the case of SVM) that contain the underlying classification information required to correctly classify previously unseen data points not part of the training data.

In fact, the appealing feature of the RVM approach is that it leads to models that are significantly sparser than the corresponding SVMs [33], while sacrificing little if anything in the accuracy of prediction. This is evident from the Table 2 where one can see that RVM-based algorithm utilized significantly fewer (9 for linear and 10 for RBF RVM) relevance vectors (i.e., kernel functions with non-

TABLE 3. Area Under the ROC Curve Values Corresponding to the Four Diagnostic Algorithms Tested on the Validation Data Set

	Linear RVM	RBF RVM	Linear SVM	RBF SVM
Area under the ROC curve	0.90	0.96	0.90	0.96

TABLE 4. Classification Results of the RVM- and the SVM-Based Diagnostic Algorithms Trained on Data Set Comprising Spectral data From the Cancerous and Normal Squamous Tissue Sites of Nine Patients and Seven Normal Volunteers and Validated Over Spectral Data From the Remaining Eight Patients and Six Normal Volunteers

		Training set data		Validation set data	
Diagnostic algorithms	Trained on	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
RBF-RVM	Data set comprising spectral data from	86	97	85	97
RBF-SVM	SCC and normal squamous tissue sites from nine patients and seven normal volunteers	90	96	87	96
RBF-RVM	Data set comprising randomly split	88	95	91	95
RBF-SVM	spectral data from SCC and normal squamous tissue sites of all the 17 patients and 13 normal volunteers	93	96	93	95

For comparison's sake, the corresponding results of the algorithms over the training and the validation data sets comprised randomly split spectral data from the cancerous tissue sites of the 16 patients and the healthy squamous tissue sites of the 13 normal volunteers are also provided in the same table. Sensitivity and specificity values in the training set data represent leave-one-out cross validation values.

zero weights) as compared to the number of support vectors (30 for linear and 27 for RBF SVM) utilized by the SVM based algorithm. From the algorithm point of view, it implies a better ability to generalize on previously unseen data points. Figure 7 displays the relevance vectors associated with the RVM algorithm with RBF kernel. These represent those prototypical tissue spectra of the training set that are finally needed for classifying the spectra of the validations data sets. It is clear from the figure that out of 10 relevance vectors generated by the RVM, 6 correspond to the cancerous and the rest correspond to the normal squamous tissues.

The classification error probability, the ultimate performance measure of a statistical classifier for supervised classification, strongly depends on the mathematical formulation of the classifier. In practice, the classification error is estimated from all the available data that are split into training and test sets [43]. The best classifier is the one that can provide a reliable estimate of classification error (in predicting future classification performance) independent of the sizes of both the training as well as the test set data. This requires minimizing the expected risk [44] while formulating the classifier. Unfortunately, the expected risk cannot be minimized directly, since the underlying probability distribution of the data is unknown. Therefore, most of the traditional statistical classifiers compute a stochastic approximation of the expected risk, called the empirical risk, based on the available information of the training data points and give conditions on the learning mechanism that ensures that asymptotically ($N\rightarrow\infty$, N being the size of the training set data) the empirical risk will converge toward the expected risk [44]. They minimize the mean squared error over the training data set for this purpose. Now, since the size of the training data set is finite, the classification error probability of the classifiers designed using this finite set is always higher than the corresponding asymptotic error probability and it decreases as N increases [35]. Further, for small sizes of the test data large deviations are possible and the estimates can be unreliable [35]. However, both the SVM and RVM approaches efficiently overcome this drawback. They incorporate a different induction principle (SRM principle) of statistical learning theory [32] and minimize the structural risk, that is, 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. This equips them with a greater ability to estimate the error probability that is close to the expected risk and consequently they can accurately classify test data of any size. It becomes evident from Table 2 where one can see that the leave-one-out cross validation estimates of false negative and false positive values over the training set data

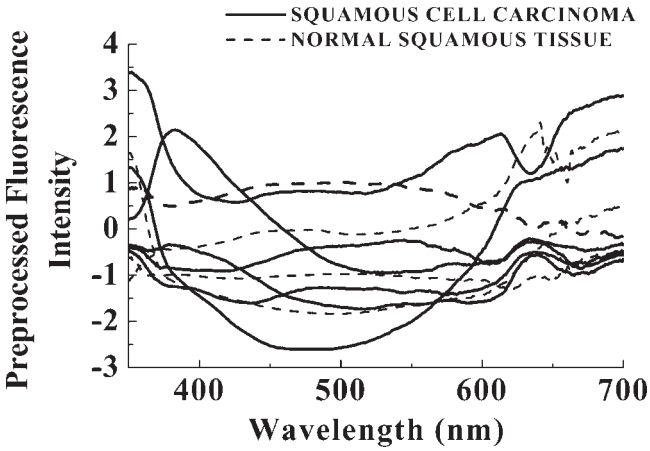

Fig. 7. Relevance vectors generated by the non-linear RVM algorithm with RBF kernel.

are comparable to that over the validation data set of size 118.

Another major concern of immense practical utility for automated diagnosis is to have the computation time and memory usage of the algorithm (during numerical implementation) as small as possible. Since the computational bottleneck of both the SVM and the RVM algorithms (like almost all kernel methods) lies in the matrix inversion operation of order $O(N^3)$ complexity and $O(N^2)$ memory storage [31,33], N being the number of training data points, the problem seems to be intractable for data sets with very large size ($N\sim$ few thousands). However, this issue has been significantly mitigated by use of certain computational tricks in the numerical implementation of the algorithms [31,33] thereby making it feasible to train both the algorithms on several thousands of data point within reasonable time scale ($\sim$several minutes). Since the size of the training set included in our study was only 119, the time taken and the memory used for training both the SVM and the RVM algorithms in our case were only few seconds and few kilobytes, respectively. However, the necessity to perform cross-validation for choosing the optimal trade-off parameter C made the actual training time for the SVM algorithms extended to several minutes in practical implementation. In contrast, this disadvantage of extended training time was offset by the lack of necessity to perform cross-validation over any nuisance parameter in the RVM algorithms.

It is also pertinent to note here that the development of diagnostic algorithms described above was based on spectral data from patients who belonged to high-risk population (were suspected 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 the present 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.

CONCLUSIONS

The application of the theory of RVM for developing diagnostic algorithm for optical diagnosis of cancer is reported. Both linear and non-linear RVM algorithms have been developed and compared with that based on equivalent SVMs using spectral data acquired in a clinical in vivo LIF study conducted on patients being screened for cancer of oral cavity and normal volunteers. The Bayesian framework of RVM formulation makes it possible to predict the posterior probability of class membership in discriminating early SCC from the normal squamous tissue sites of the oral cavity in contrast to dichotomous classification provided by the non-Bayesian SVM. The results further demonstrate that for comparable diagnostic performances, the RVM-based use significantly fewer kernel functions and do not need to estimate any hoc parameters associated with the learning or the optimization technique to be used. This implies a considerable saving in memory and computation in a practical implementation.

ACKNOWLEDGMENTS

The authors 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 fluorimeter installed at the Government Cancer Hospital, Indore.

REFERENCES

1. Wagnieres GA, Star WM, Wilson BC. In vivo fluorescence spectroscopy and imaging for oncological applications. Photochem Photobiol 1998;68:603–632.
2. Servick-Muraca E, Richards Kortum R. Quantitative optical spectroscopy for tissue diagnosis. Annu Rev Phys Chem 1996;47:556–606.
3. Ramanujam N, Fluorescence spectroscopy of neoplastic and non-neoplastic tissues. Neoplasia 2000;2(1):1–29.
4. Mahadevan-Jansen A, Richards-Kortum R. Raman spectroscopy for the detection of cancers and precancers. J Biomed Opt 1996;1(1):31–70.
5. Cothren RM, Sivak MV, Dam JV, Petras RE, Fitzmaurice M, Crawford JM, Wu J, Brennan JF, Rava R, Manoharan R, Feld MS. Detection of dysplasia at colonoscopy using laser induced fluorescence: A blinded study. Gastrintest Endosc 1996;44(2): 168–176.
6. Lin WC, Toms S, Motamedi M, Jansen ED, Mahadevan-Jansen A. Brain tumor demarcation using optical spectroscopy; an in-vitro study. J Biomed Opt 2000;5(2):214–220.
7. Gupta PK, Majumder SK, Uppal A. N_2 laser excited autofluorescence spectroscopy for human breast cancer diagnosis. Lasers Surg Med 1997;21:417–422.
8. Majumder SK, Uppal A, Gupta PK. Autofluorescence spectroscopy of tissues from human oral cavity for discriminating malignant from normal. Lasers Life Sci 1999;8:211–227.
9. Kapadia CR, Cutruzzola FW, O'Brien KM, Stetz ML, Enriquez R, Deckelbaum LI. Laser-induced fluorescence spectroscopy of human colonic mucosa: Detection of adenomatous transformation. Gastroenterology 1990;99:150–157.
10. Marchesini RM, Brambilla M, Pignoli E, Bottiroli G, Croce AC, Fante MD, Spinelli P, Palma SD. Light-induced fluorescence spectroscopy of adenomas, adenocarcinomas and non-neoplastic mucosa in human colon: In-vitro measurements. J Photochem Photobiol B Biol 1992;14:219–230.
11. Schomacker KT, Frisoli JK, Compton CC, Flotte TJ, Richter JM, Nishioka NS, Deutsch TF. Ultraviolet laser-induced fluorescence of colonic tissue: Basic biology and diagnostic potential. Lasers Surg Med 1992;12:63–78.
12. Koenig F, McGovern FJ, Althausen AF, Deutsch TF, Schomacker KT. Laser induced autofluorescence diagnosis of bladder cancer. J Urol 1996;156(5):1597–1601.
13. Yang Y, Katz A, Celmer EJ, Zurawska-Szczepaniak M, Alfano RR. Optical spectroscopy of benign and malignant breast tissues. Lasers Life Sci 1987;7(2):115–127.
14. Yang Y, Tang GC, Bessler M, Alfano RR. Fluorescence spectroscopy as photonic pathology method for detecting colon cancer. Lasers Life Sci 1995;6(4):259–276.
15. Brewer M, Utzinger U, Silva E, Gershenson D, Bast RC, Follen M, Richards-Kortum R. Fluorescence spectrsocopy for in-vivo characterization of ovarian tissue. Lasers Surg Med 2001;29:128–135.
16. Dhingra JK, Perrault DF, McMillan K, Rebeiz EE, Kaban S, Manoharan R, Itzkan I, Feld MS, Shapshay SM. Early diagnosis of upper aerodigestive tract cancer by auto fluorescence.. Arc Otolaryngol and Head Neck Surg 1996; 122(11):1181–1186.
17. Majumder SK, Uppal A, Gupta PK. In-vitro diagnosis of human uterine malignancy using N_2 laser-induced autofluorescence spectroscopy. Current Science 1996;70(9):833–836.
18. Wang CY, Chen CT, Chiang CP, Young ST, Chow SN, Chiang HK. A probability-based multivariate statistical algorithm

for autofluorescence spectroscopic identification of oral carcinogenesis. Photochem Photobiol 1999;69(4):471–477.
19. Atkinson EN, Mitchell MF, Ramanujam N, Richards-Kortum R. Statistical techniques for diagnosing CIN using fluorescence spectroscopy: SVD and CART. J Cell Biochem Supplement 1995;23:125–130.
20. Ramanujam N, Follen-Mitchell M, Mahadevan-Jansen A, Thomson S, Staerkel G, Malpica A, Wright T, Atkinson N, Richards-Kortum R. Cervical precancer detection using multivariate statistical algorithm based on laser-induced fluorescence spectra at multiple excitation wavelengths. Photochem Photobiol 1996;64:720–735.
21. Zuluaga A, Utzinger U, Durkin H, Fuchs A, Gillenwater A, Jacob R, Kemp B, Fan J, Richards-Kortum R. Fluorescence excitation-emission matrices of human tissue: A system for in-vivo measurement and method of data analysis. Appl Spectrosc 1999;53:302–311.
22. Heintzelmann D, Utzinger U, Fuchs H, Zuluaga A, Gossage K, Gillenwater AM, Jacob R, Kemp B, Richards-Kortum R. Optimal excitation wavelengths for in-vivo detection of oral neoplasia using fluorescence spectroscopy. Photochem Photobiol 2000;72(1):103–113.
23. Majumder SK, Mohanty SK, Ghosh N, Gupta PK, Jain DK, Khan F. A pilot study on the use of autofluorescence spectroscopy for diagnosis of the cancer of human oral cavity. Curr Sci 2000;79(8):1089–1094.
24. van Staveren HJ, van Veen RL, Speelman OC, Witjes MJ, Star WM, Roodenburg JL. Classification of clinical autofluorescence spectra of oral leukoplakia using an artificial neural network: A pilot study. Oral Oncol 2000;36(3):286–293.
25. Tumer K, Ramanujam N, Ghosh J, Richards-Kortum R, Ensembles of radial basis function networks for spectroscopic detection of cervical pre-cancer. IEEE Trans BME 2001;45(8):953–961.
26. Rovithakis GA, Maniadakis AN, Zervakis M, Fillipidis G, Zakarakis G, Katsamouris AN, Papazoglou TG, Artificial neural networks for discriminating pathologic from normal peripheral vascular tissue. IEEE Trans BME 2001;48(10): 1088–1096.
27. Agrawal N, Gupta S, et al. Wavelet transform of breast tissue fluorescence spectra—A technique for diagnosis of tumors. IEEE J Selected Topics in Quantum Electronics 2003;9(2): 154–161.
28. Majumder SK, Ghosh N, Kataria S, Gupta PK. Nonlinear pattern recognition for laser-induced fluorescence diagnosis of cancer. Lasers Surg Med 2003;33:48–56.
29. Palmer GM, Zhu C, Breslin TM, Xu F, Gilchrist KW, Ramanujam N. Comparison of multiexcitation fluorescence and diffuse reflectance spectroscopy for the diagnosis of breast cancer (March 2003). IEEE Trans Biomed Eng 2003; 50(11):1233–1242.
30. Lin WM, Yuan X, Yuen P, Wei WI, Sham J, Shi PC, Qu J. Classification of in-vivo autofluorescence spectra using support vector machines. J Biomed Opt 2004;9(1):180–186.
31. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 1998;2(2):121–167.
32. Vapnik VN, Statistical learning theory. Hoboken: John Wiley and Sons; 1998.
33. Tipping ME. Sparse Bayesian learning and relevance vector machine. J Machine Learn Res 2001;1:211–244.
34. Nath A, Rivoire K, Chang S, Cox D, Atkinson EN, Follen M, Rischards-Kortum R. Effect of probe pressure on cervical fluorescence spectroscopy measurements. J Biomed Opt 2004; 9(3):523–533.
35. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: A review. IEEE Trans Pattern Anal 2000;22(1):4–37.
36. Brookner CK, Utzinger U, Staerkel G, Richards-Kortum R, Mitchell MF. Cervical fluorescence of normal women. Lasers Surg Med 1999;24:29–37.
37. Berger JO, Statistical decision theory and Bayesian analysis, Second edition. New York: Springer; 1985.
38. MacKay DJC. The evidence framework applied to classification networks. Neural Comput 1992;4(5):720–736.
39. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–298.
40. Sollich P. Probabilistic methods for support vector machines. In: Solla SA, Leon TK, Muller RK, editors. Advances in neural information processing systems, Vol. 12. Cambridge: MIT Press; 2000. pp 349–355.
41. Duda RO, Hart PE. Pattern classification and scene analysis. Hoboken: John Wiley; 1973.
42. Weston J, Watkins C. Multi-class support vector machines. Technical Report 1998, CSD-TR-98_04, Department of Computer Science, Royal Holloway, University of Lonsdon, Egham, TW20 0EX, UK.
43. Hand DJ. Recent advances in error rate estimation. Pattern Recognit Letts 1986;4(5):335–346.
44. Muller KR, Mika S, Gunnar R, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw 2001;12(2):181–201.