Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines

EFFENDI WIDJAJA[1], WEI ZHENG[2] and ZHIWEI HUANG[2]

[1]Institute of Chemical and Engineering Sciences, Singapore 627833; [2]Bioimaging Laboratory, Department of Bioengineering, Faculty of Engineering, National University of Singapore, Singapore 117576

Received September 18, 2007; Accepted November 21, 2007

Abstract. The ability of combining near-infrared (NIR) Raman spectroscopy with support vector machines (SVM) for improving multi-class classification between different histopathological groups in tissues was evaluated in this study. A total of 105 colonic tissue specimens from 59 patients including 41 normal, 18 hyperplastic polyps and 46 adenocarcinomas were used for this purpose. A rapid-acquisition dispersive-type NIR Raman system was utilized for tissue Raman spectroscopic measurements at 785-nm laser excitation. A total of 817 tissue Raman spectra were acquired and subjected to principal components analysis (PCA) for SVM-based multi-class classification, in which 324 Raman spectra were from normal, 184 from polyps and 309 from adenocarcinomatous colonic tissue. Two types of SVM (i.e., C-SVM and ν-SVM) with three different kernel functions (linear, polynomial and Gaussian radial basis function (RBF) in combination with PCA were used to develop effective diagnostic algorithms for classification of Raman spectra of different colonic tissues. The performance of various SVM-based algorithms was evaluated and compared using a leave-one-out, cross-validation method. The results showed that in the C-SVM classification, the maximum overall diagnostic accuracy of 99.3, 99.4 and 99.9% can be achieved using the linear, polynomial and RBF kernels, respectively; while in the ν-SVM classification, the maximum overall diagnostic accuracy of 98.4, 98.5 and 99.6% can be obtained using the linear, polynomial and RBF kernels, respectively. All the polyps can be identified from normal and adenocarcinomatous tissue using the C-SVM algorithms. The RBF C-SVM algorithm was proven to be the best classifier for providing the highest diagnostic accuracy (99.9%) for multi-class classification. This study demonstrates that NIR Raman spectroscopy in combination with a powerful SVM technique has great potential for providing an effective and accurate diagnostic schema for cancer diagnosis in the colon.

Correspondence to: Dr Zhiwei Huang, Bioimaging Laboratory, Department of Bioengineering, Faculty of Engineering, National University of Singapore, 9, Engineering Drive 1, Singapore 117576
E-mail: biehzw@nus.edu.sg

Key words: adenocarcinoma, colonic cancer, near-infrared, optical diagnosis, polyp, principal components analysis, Raman spectroscopy, support vector machines

Introduction

Gastrointestinal (GI) malignancies continue to be the second leading cause of cancer-related deaths in developed countries (1). In Singapore, colorectal carcinoma has become the most commonly diagnosed malignancy and is the second most common cause of cancer death (2). Reducing mortality rates of GI cancer relies on the early detection and immediate removal of pre-malignant lesions [e.g. dysplasia, carcinoma *in situ* (CIS)] (1). However, using the conventional screening tool, such as a white-light endoscope, it is difficult to identify early neoplasia or subtle lesions (e.g. flat adenomas) in the GI tract. This is because routine endoscopy simply relies on the observation of gross morphological changes of tissues, leading to the poor diagnostic sensitivity. Excisional biopsies currently remain the standard approach for cancer diagnosis, though this method is invasive and impractical for mass screening of high-risk patients with multiple suspicious lesions.

In the past decade, optical spectroscopic methods such as Raman spectroscopy have been comprehensively investigated for cancer and precancer diagnosis and evaluation (3-12). Raman spectroscopy measures inelastic light scattering and is a vibrational spectroscopic technique that can provide specific spectroscopic fingerprints based on the molecular composition and structures of biological tissues (3,4,8-10). Furthermore, near-infrared (NIR) Raman spectroscopy holds significant advantages over other vibrational spectroscopy techniques in that water exhibits very low absorption at the working wavelength range and tissues exhibit far less autofluorescence compared with visible light excitation. Less water absorption makes it easy to detect other tissue components and results in deeper light penetration into the tissue. As such, NIR Raman spectroscopy has received great interest for *in vivo* and *in vitro* diagnosis of malignancies in a number of organs including the colon (3,5-15). NIR Raman spectroscopy studies showed that tissue Raman spectral features could be used to correlate with the molecular and structural changes associated with neoplastic transformations (5,7,10,12), demonstrating the feasibility of NIR Raman spectroscopy technique for early cancer detection. However, NIR Raman signals are inherently weak. There are only about 10^{-8} of the incident photons that are inelastically scattered from tissue to generate unique Raman spectroscopic patterns. In addition, Raman spectral differences are usually subtle with apparently spectral overlapping and variations in intensity between different tissue types (5,7,8,13-18). Therefore, the powerful and robust spectral data processing and sophisticated diagnostic algorithms are much needed to best extract the most diagnostically significant Raman spectral features in order to accurately correlate them with the tissue histopathology. Multivariate statistical techniques [e.g. principal components analysis (PCA), linear discriminant analysis (LDA), artificial neural network (ANN) and fuzzy C-means] (6,7,12,14,16-19) have been successfully utilized in developing effective diagnostic algorithms for spectroscopic diagnosis of cancers. For example, employing PCA-LDA techniques, a high diagnostic accuracy (>90%) can be achieved for identifying Raman spectra of cancer from normal tissue in the colon (17).

Another powerful multivariate technique, support vector machines (SVM), which was based on the machine learning approach and originally developed by Vapnik (19,20) and Burges (21), has attracted great attention due to its capability of representing non-linear relationships and producing models that generalize well in classifying the unseen data (21-26). The SVM technique has now emerged as an efficient approach to the classification of spectral data for tissue diagnosis (27-30). For instance, Lin *et al* (27) used linear and non-linear SVM to differentiate *in vivo* autofluorescence spectra of nasopharyngeal carcinoma (NPC) from normal tissue with diagnostic accuracy being higher than that using PCA-LDA. Palmer *et al* (28) used a linear SVM classifier for identifying autofluorescence and diffuse reflectance spectra of breast cancer tissues *in vitro*. Majumder *et al* (29) used linear and non-linear SVM to classify fluorescence spectra of malignant tissue from normal tissue in the oral cavity. Our group also applied both linear and non-linear SVM in identifying the Raman spectra of cancer from normal tissue in the larynx with diagnostic accuracy >95% (30). To date, studies of SVM for spectroscopic diagnosis of cancerous tissue are still very limited and most efforts are focusing on the binary-class problems (27-29). In addition, SVM has not yet been applied to the classification of NIR Raman spectroscopy for cancer diagnosis in detail. In this study, we explored the SVM technique for multi-class classification of NIR Raman spectra acquired from different pathological groups of colonic tissues *in vitro*. The conventional SVM (C-SVM) and the modified SVM (ν-SVM) approaches (see support vector mechanisms) with three different kernel functions (linear, polynomial and radial basis function (RBF) in combination with PCA were implemented to develop diagnostic algorithms for effective tissue diagnostic in the colon. The diagnostic performances of various SVM-based algorithms developed using the two approaches were evaluated in an unbiased manner using the leave-one-out cross-validation method.

Materials and methods

Instrumentation. The instrument used for tissue Raman spectroscopic studies has been described in detail elsewhere (31). Briefly, this system consists of a 785-nm diode laser (maximum output: 300 mW, SDL Inc., San Jose, CA), a transmissive imaging spectrograph (HoloSpec-f/2.2-NIR, Kaiser Optical Systems Inc., Ann Arbor, MI) with a volume phase technology (VPT) holographic grating (HSG-785-LF, Kaiser Optical Systems Inc.), an NIR-optimized back-illuminated, deep-depleted charge-coupled device (CCD) detector (LN/CCD-EEV 1024x256, QE ≥75% at 900 nm, Princeton Instruments, Trenton, NJ) and a specially designed fiber optic Raman probe that can effectively eliminate interference from fiber-optic fluorescence and silica Raman signals (8). The 785-nm laser is coupled to a 100-μm core diameter fiber (NA=0.22) and the fiber is connected to the Raman probe via an SMA connector. Tissue NIR Raman signals collected by the probe are fed into the transmissive spectrograph and the holographic grating disperses the incoming light onto the liquid-nitrogen-cooled CCD detector controlled by a PC. The tissue Raman spectra associated with autofluorescence background are displayed on the computer screen in real-time and can be saved for further analysis. The system acquired Raman spectra over the wavenumber range of 800-1800 cm^{-1} and each spectrum was acquired within 5 sec with light irradiance of 1.56 W/cm^2. The spectral resolution of the system is 4 cm^{-1}. All wavelength-calibrated spectra were also corrected for the wavelength-dependence of the system using a standard lamp (RS-10, EG&G Gamma Scientific, San Diego, CA).

Colonic tissue samples. A total of 105 colonic tissue samples were collected from 59 patients who underwent partial colectomy or biopsies with clinically suspicious lesions or histopathologically proven malignancies of the colon. All patients preoperatively signed an informed consent permitting the investigative use of the tissues and this study was approved by the Ethics Committee of the National Healthcare Group of Singapore. After biopsies or surgical resections, tissue samples were immediately sent to the laboratory for Raman measurements. After spectral measurements, the tissue samples were fixed in 10% formalin solution and then submitted back to the hospital for a histopathological examination. A total of 817 tissue Raman spectra from different sites of colonic tissue samples were acquired, in which 324 Raman spectra were from 41 normal, 184 from 18 hyperplastic polyps (benign) and 309 from 46 adeno-carcinomatous colonic tissue. Note that to reduce the spectral measurement errors in this study, each tissue Raman spectrum obtained for tissue classification was the average spectrum of 3 repeated Raman measurements on the same tissue location and for each tissue sample, a number of Raman spectra from different locations (3 to 10 locations, depending on the tissue size) of the same tissue sample were acquired and considered as different Raman spectra for multi-class classification.

Support Vector Machines (SVM). Support vector machines (SVM) algorithm was first introduced by Vapnik (19) and Burges (21) and has proven successful in many applications, such as object recognition (24), face detection (23) and text categorization (22,25). With the following advantages, SVM has become an efficient approach in the classification of spectral data (27-30,32): i) SVM creates a reliable classifier, particularly useful for working with a small or limited size of datasets, as it is based on structural risk minimization (SRM) that reduces the risk of data over-fitting; ii) SVM gives reproducible solutions when the same parameters of classifiers are used and iii) SVM has the ability to draw class boundaries with complex conditions by replacing the kernel functions.

The selection of an appropriate kernel function is critical in group classification in SVM, as the function defines the feature space whereby the training data points are classified. The kernel function maps the input vector onto a higher dimensional space such that a better hyperplane can be obtained with minimal classification errors. The Kernel function is chosen as a priori to determine the type of SVM classifiers. The three most commonly used kernel functions are:

(a) Linear: $K(x_i, x_j) = x_i \cdot x_j + 1$ (1)

(b) polynomial kernel of degree p: $K(x_i, x_j) = (x_i \cdot x_j + 1)^p$ (2)

(c) Gaussian radial basis function (RBF): $K(x_i, x_j) = \exp \frac{-\|x_i - x_j\|^2}{2\sigma^2}$ (3)

where x_i and x_j are the two generic sample data vectors.

The selection of the optimal values of the parameters, such as polynomial order p in the polynomial kernel, the radial width σ in the Gaussian RBF kernel and C in all these three types of classifiers, is an optimization problem. As for the cost function of this optimization chosen, we employed the overall diagnostic accuracy obtained using the leave-one-out cross-validation procedure. With the use of the Kernel function, the non-linear decision boundary can be found in the input space and then SVM can be implemented as long as the kernel function obeys the Mercer's theorem (19,22).

One notes that the conventional SVM techniques (termed as C-SVM) are only designed for solving the problem with binary classification. An extension of the SVM technique for multi-class classification is more robust and useful in a clinical setting, in which different types of pathological tissues can be studied and classified rapidly. Hence, in this study, we incorporated the one-against-one strategy (OAO) (25-33), one of the most appropriate approaches dealing with classification problems into SVM for multi-class classification. The central idea of OAO is to construct one binary classifier first for every pair of distinct classes in SVM and then all the binary classifiers with a total number of M (M-1)/2 (where M is the number of classes involved) are generated. When each binary classification B_{ij} is performed in SVM, the model is trained by setting the samples from class ωi as positive and the samples from class ω_j as negative. For a testing sample χ, if the classifier B_{ij} puts χ into class ωi, then the score for class ωi is added by one. Otherwise, the score for class ω_j is increased by one. After each of the M (M-1)/2 binary classifiers is assigned to its score, the final decision in the OAO strategy will be made based on the principle of the 'winner-takes-all', in which χ is assigned to the class with the largest number of scores. It should be noted that the conflict of decision may occur when the same score is obtained from the two different classes and in this case, the scheme by choosing the class that has the highest prior probability will be executed to overcome such an ambiguity in SVM.

On the other hand, to simplify the implementation of the conventional SVM technique using the parameter C, modifications on SVM had been proposed by Scholkopf *et al* (34-36) in which a new parameter of ν ($\in [0, 1]$) was introduced to replace the parameter of $C \in [0, \infty]$ in the conventional SVM. The modified SVM algorithm (termed as ν-SVM) is simply looking for an optimal hyperplane with the maximal separating margins. The parameter ν directly determines the number of support vectors, while the number of support vectors gives a leave-one-out generalization bound. Thus, we employed both the conventional SVM (C-SVM) and the modified SVM (ν-SVM) techniques in order to evaluate their performances for the multi-class classification of Raman spectra between different pathological colonic tissues.

Data preprocessing. Under the 785-nm laser excitation, the raw spectra acquired from colonic tissue in the 800-1800 cm^{-1} range represented a combination of prominent tissue autofluorescence, very weak tissue Raman scattering signals, and noise. Thus, the raw spectra were preprocessed by adjacent 5-point smoothing to reduce noise. To extract the Raman features contained in the measured spectral data, a fifth-order polynomial (8) was found to be optimal for fitting the broad autofluorescence background in the noise-smoothed spectrum and this polynomial was then subtracted from the raw spectrum to yield the tissue Raman spectrum alone. Each of the background-subtracted Raman spectra was normalized to the integrated area under the curve from 800-1800 cm^{-1} to enable a better comparison of the spectral shapes and relative peak intensities among the different tissue samples and histopathological groups.

Statistical analysis. The high dimension of Raman spectral space (each Raman spectrum ranging from 800-1800 cm^{-1} with a set of 544 intensities) will result in computational complexity and inefficiency in optimization and implementation of the SVM algorithms. As such, PCA was performed on tissue Raman dataset to reduce the dimension of Raman spectral space while retaining the most diagnostically significant information for tissue classification. To eliminate the influence of inter and/or intra-subject spectral variability on PCA, the entire spectra were standardized so that the mean of the spectra was zero and the standard deviation of all the spectral intensities was one. Mean centering ensures that the principal components (PCs) form an orthogonal basis (37,38). The standardized Raman data sets were assembled into data matrices with wavenumber columns and individual case rows. Thus, PCA was performed on the standardized spectral data matrices to generate PCs comprising of a reduced number of orthogonal variables that accounted for most of the total variance in the original spectra. Each loading vector is related to the original spectrum by a variable called the PC score, which represents the weight of that particular component against the basis spectrum. PC scores reflect the differences between different classes. The diagnostically significant PC scores are used as an input for the development of SVM algorithms for multi-class classification.

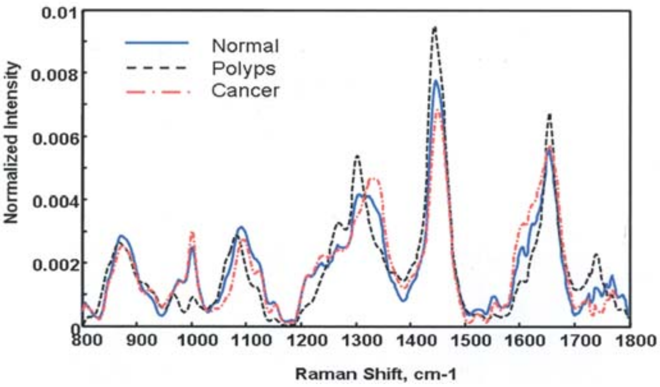

Figure 1. Mean NIR Raman spectra from normal (n=324), polyps (n=184) and adenocarcinoma (n=309) colonic tissues, respectively.

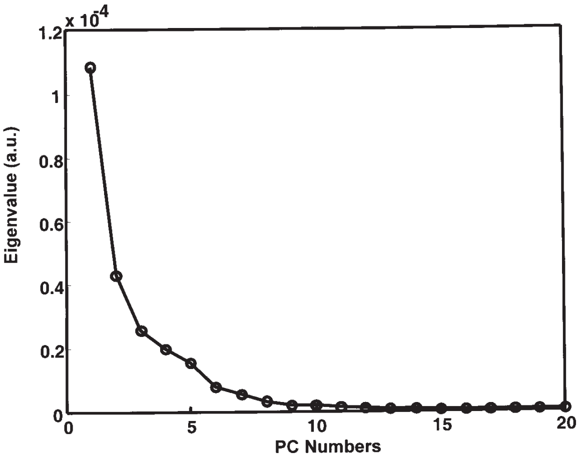

Figure 2. Eigenvalues of principal components contributed to the total variance of all Raman spectral data.

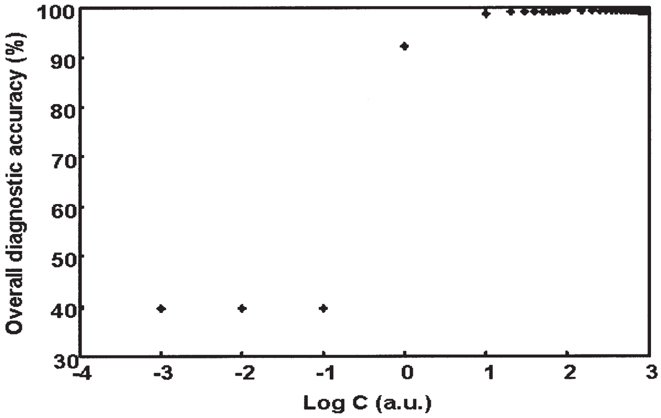

Figure 3. Dependence of overall diagnostic accuracy on the parameter *C* using the linear C-SVM algorithm.

Leave-one-out cross-validation. For a limited number of datasets on Raman spectra obtained from different sites of colonic tissues in this study, a leave-one spectrum-out cross-validation method (38,39) was used to evaluate the performance of the SVM algorithms for multi-class tissue classification in an unbiased fashion. In each round of cross-validation, one spectrum is held out from the entire dataset and assigned as a testing set for the classifier developed using the remaining tissue spectra assigned as a training set. This process is repeated until all the withheld spectra in the dataset are validated and then the overall diagnostic accuracy, sensitivity and specificity of the particular SVM algorithms were calculated accordingly.

Computation. All Raman spectral data analyses were performed in MATLAB environment (MATLAB 6.5; The MathWorks Inc., Natick, MA). In-house programs written in MATLAB were used for Raman data preprocessing, principal components analysis and leave-one-out cross validation, while the SVM MATLAB toolbox (40) was used for SVM classifications. A personal computer with a 3GHz Pentium IV processor and 522 MB of RAM was used for all computations.

Results

Fig. 1 shows the normalized mean of NIR Raman spectra from normal, benign and malignant tumors, respectively. Primary Raman peaks at around 875, 1002, 1090, 1267, 1320, 1445, 1605, 1655 and 1740 cm^{-1} and can be consistently observed in both normal and abnormal colonic tissues, with the strongest signals at 1320, 1445 and 1655 cm^{-1}. It is observed that the significant differences in Raman spectral features (e.g. peak intensities, peak positions and spectral shapes) exist in different tissue types, reflecting the changes of biochemical compositions and structures of pathological tissue in the colon (5,17,41).

In this study, multivariate statistical techniques, such as principal components analysis (PCA) and support vector machines (SVM), that utilized the entire Raman spectrum to determine the most diagnostically significant features (factors), were employed to develop and evaluate the diagnostic algorithms for improving tissue diagnosis and classification. All the 817 Raman spectra ranging from 800 to 1800 cm^{-1} measured from three types of colonic tissues were subjected to PCA and then SVM for algorithm development. PCA was used to reduce the dimension of variables in Raman spectra, while still keeping the diagnostically useful spectral features (or components) for tissue diagnosis. Upon PCA calculation, orthogonal principal component (PC) scores and loadings were generated for algorithm development using SVM.

Fig. 2 illustrates the eigenvalues of each PC contributed to the total variance of 817 Raman spectral data. It is observed that the eigenvalues drop off rapidly with increasing PC numbers and the first few PCs retain the maximum variance of the data being considered. For instance, the first PC accounts for the largest variance within the spectral data sets, which is ~41.4% of the total variance, whereas the successive PCs describe the features that contribute to progressively smaller variances. The first five PCs account for 81.31% of the total variance; the first 10 PCs account for 88.8%; the first 15 PCs account for 90.6%; and the first 20 PCs collectively account for 91.6%. The variance of each individual PC (weight of PC loading or scores) starting from PC 18th and above has been much <0.2%. Hence, only the first 18 PC scores accounting for 91.4% were selected in this study as the main spectral features for multi-class classification using both C-SVM and

Table I. Classification results of Raman prediction of the three pathological groups in the colon (C-SVM).

Pathology and Classification	Raman Prediction									Total
	Linear C-SVM[a]			Polynomial C-SVM[b]			RBF C-SVM[c]			
	Normal	Polyp	Cancer	Normal	Polyp	Cancer	Normal	Polyp	Cancer	
Normal	324	0	0	324	0	0	324	0	0	324
Polyp	0	184	0	0	184	0	0	184	0	184
Cancer	6	0	303	5	0	304	1	0	308	309
Sensitivity (%)	98.8	100.0	100.0	98.9	100.0	100.0	99.8	100.0	100.0	
Specificity (%)	100.0	100.0	98.1	100.0	100.0	98.4	100.0	100.0	99.7	

Classification results of Raman prediction of the three pathological groups in the colon obtained using the leave-one-out cross-validation method in C-SVM with different kernel functions. [a]The linear C-SVM (C=80) as a diagnostic algorithm; [b]the polynomial C-SVM as a diagnostic algorithm (taken from overall diagnostic accuracy at 99.4%); [c]the RBF C-SVM as a diagnostic algorithm (taken from overall diagnostic accuracy at 99.9%).

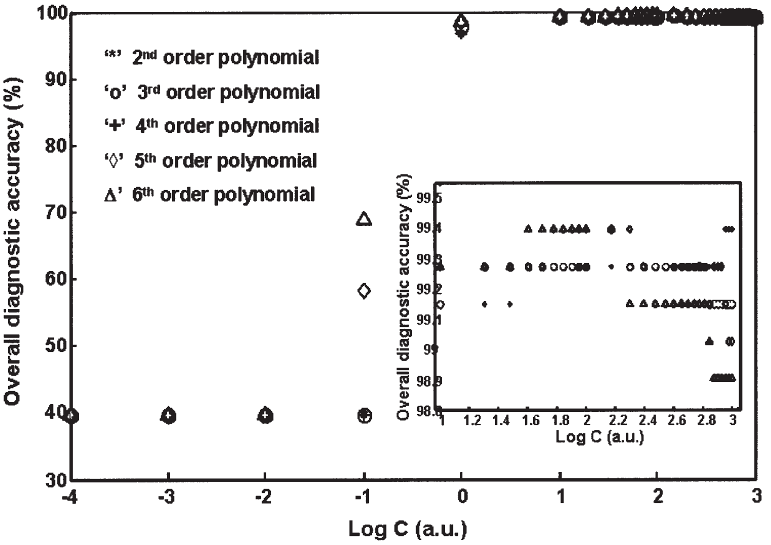

Figure 4. Dependence of overall diagnostic accuracy on the parameter *C* under different orders or polynomials using the polynomial C-SVM algorithm.

ν-SVM techniques. Note that no further improvements on diagnostic accuracy were found when more PCs were loaded into SVM for classification (data not shown).

We used the linear and nonlinear (polynomial and RBF) kernels for the development of SVM diagnostic algorithms with the Raman spectral dataset from colonic tissue. The linear C-SVM algorithm developed from the training dataset needs to be optimized by exhaustively searching for the optimal parameter *C*, which gives the best trade-off between the training error and the generalization ability. We tried to maximize the overall classification accuracy as an optimization criterion for obtaining the optimal *C*. A wide-range search for optimal *C* was performed from 0.0001 to 1000 and the best overall diagnostic accuracy was obtained at 99.3% for the *C* values ranging from 80 to 750. The overall diagnostic accuracy as a function of parameter log (*C*) in the aforementioned range is shown in Fig. 3. It is seen that when *C* is <0.1, the overall diagnostic accuracy is <40%, indicating that the training error is larger when the parameter *C* is smaller. With increasing *C* values, the overall diagnostic accuracy increases accordingly (Fig. 3) and when the *C* values reach up to 800, the accuracy starts to drop since the generalization ability becomes weaker.

Table I shows the classification results of the three different pathological groups (normal, polyp and cancer) of colonic tissue obtained using the leave-one-out cross-validation with the linear C-SVM algorithm at parameter *C*=80. A diagnostic sensitivity of 98.8, 100.0 and 100.0%; and specificity of 100.0, 100.0 and 98.1%, respectively, can be reached for differentiation between normal, polyps and cancers in the colon using the linear C-SVM algorithm.

In the development of polynomial C-SVM algorithms, the optimal *C* values that gave the maximum overall diagnostic accuracy were exhaustively sought for different orders of polynomial kernel. Fig. 4 shows the results of overall diagnostic accuracy as a function of parameter *C* for the five different orders (e.g. 2nd, 3rd, 4th, 5th and 6th) of polynomial. We found that these different orders of polynomial kernels yielded the maximum overall diagnostic accuracy of 99.4% (see inset in Fig. 4) with the parameter *C* ranging from 40 to 200 and therefore, the maximum diagnostic accuracy obtained was not very sensitive to the orders of polynomials using the

Table II. The maximum overall diagnostic accuracies obtained by exhaustively searching for the optimal combinations of the parameters C and σ in the RBF C-SVM.

Number	Parameter C	Maximum overall diagnostic accuracy
1	0.01	39.7 (σ = 0.1)
2	0.1	98.8 (σ = 30)
3	1	99.9 (σ = 16)
4	5	99.9 (σ = 15)
5	10	99.9 (σ = 18)
6	20	99.9 (σ = 19)
7	30	99.9 (σ = 20)
8	40	99.9 (σ = 25)
9	50	99.9 (σ = 25)
10	60	99.9 (σ = 25)
11	70	99.9 (σ = 25)
12	80	99.9 (σ = 25)
13	90	99.9 (σ = 25)
14	100	99.9 (σ = 25)
15	200	99.9 (σ = 55)
16	300	99.9 (σ = 55)
17	400	99.9 (σ = 55)
18	500	99.9 (σ = 55)
19	600	99.9 (σ = 55)
20	700	99.9 (σ = 55)
21	800	99.9 (σ = 55)
22	900	99.9 (σ = 55)
23	1000	99.9 (σ = 55)

Note that at each particular parameter *C*, there are a number of σ values yielding the repetitive maximum overall diagnostic accuracy, though only the first σ values giving the maximum overall diagnostic accuracy are listed.

polynomial C-SVM algorithm. The classification results evaluated at the highest overall diagnostic accuracy (99.4%) are listed in Table Ib. A diagnostic sensitivity of 98.9, 100.0 and 100.0%; and specificity of 100.0, 100.0 and 98.4%, respectively, can be obtained for differentiation between normal, polyps and cancers in the colon using the polynomial C-SVM algorithm.

The implementation of RBF kernel as C-SVM classifier requires the user to specify not only the parameter *C* but also the radial width σ. A proper selection of these two parameters is critical in yielding maximum classification accuracy. Although a number of optimization schemes have been introduced in the literature (42), there is no consensus yet in the SVM technique. In this study, we utilized a conventional grid-search on *C* and σ, in which the combination of these two parameters giving the highest overall diagnostic accuracy was selected as the winner (43). To search for the optimal Gaussian radial width σ, the RBF C-SVM classifier was trained for the different σ values ranging from 0 to 1000, with different step sizes chosen according to the σ values: i) a step size of 0.1 for σ values between 0 to 1; ii) 1 for σ values between 1 to 20; iii) 5 for σ values between 20 to 100 and iv) 50 for σ values between 100 to 1000. In the mean time, the training for these various σ values was performed at various regularization values of the parameter *C* ranging from 0.01 to 1000. The grid-search in the RBF C-SVM yielded many optimal combinations of the parameters *C* and σ with the maximum overall diagnostic accuracy of 99.9% for the three-class classification of colonic tissues (Table II).

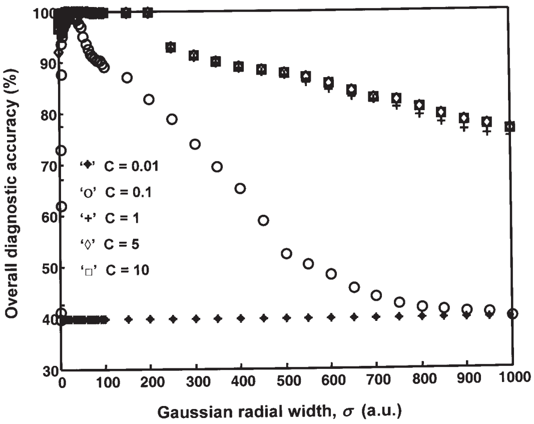

Figure 5. Dependence of overall diagnostic accuracy on the parameter-Gaussin radial width σ at different values of the parameter *C* using the RBF C-SVM algorithm.

Fig. 5 shows the results of overall diagnostic accuracy as a function of Gaussian radial width σ, at various values of the parameter *C* in the RBF C-SVM. The smaller values of parameter *C* (e.g. ≤0.01) gave a poor overall diagnostic accuracy (40% only). When the C value increases to 1, the diagnostic accuracy is improved significantly and the maximum diagnostic accuracy of 99.9% can be reached when the *C* value is >1 at various values of the parameter σ using the RBF C-SVM algorithm. Table Ic lists the cross-validation results of the pathological groups of colonic tissues evaluated at the highest overall diagnostic accuracy (99.9%) in the RBF C-SVM. A diagnostic sensitivity of 98.9, 100.0 and 100.0%; and specificity of 100.0, 100.0 and 99.7%, respectively, can be achieved for differentiation between normal, polyps and cancers in the colon using the RBF C-SVM algorithm.

As mentioned above in SVM, ν-SVM is more frequently used in data classification since it is more intuitive and efficient in dealing with the optimization of the parameter ν which ranges from 0 to 1 only, as compared to the C parameter optimization which ranges from 0 to infinity in the C-SVM. In the linear ν-SVM, only the parameter ν which controls the number of support vectors needs optimizing. One notes that there are no established guidelines in the SVM to search for optimal values of ν, the linear ν-SVM classifiers in this study were trained with different ν values (e.g. 0.001, 0.01, 0.1, 0.5 and 1). The leave-one-out cross-validation results show that the linear ν-SVM does not depend on the ν parameter since all these ν values produce similar overall classification accuracy at 98.4% (Fig. 6) and the diagnostic sensitivity of 97.6, 100.0 and 99.8%; and specificity of 99.7, 100.0 and 96.1%, respectively, can be achieved for differentiation between normal, polyps and cancerous colonic tissues using the linear ν-SVM algorithm with the parameter ν of 1 (Table IIIa).

In the polynomial ν-SVM, in order to find the optimal values of the order *p* of polynomial and the parameter ν, the

Table III. Classification results of Raman prediction of the three pathological groups in the colon (ν-SVM).

Pathology and Classification	Raman Prediction									Total
	Linear ν-SVM[a]			Polynomial ν-SVM[b]			RBF ν-SVM[c]			
	Normal	Polyp	Cancer	Normal	Polyp	Cancer	Normal	Polyp	Cancer	
Normal	323	0	1	323	0	1	323	0	1	324
Polyp	0	184	0	0	184	0	1	183	0	184
Cancer	12	0	297	12	0	297	1	0	308	309
Sensitivity (%)	97.6	100.0	99.8	97.6	100.0	99.8	99.6	100.0	99.8	
Specificity (%)	99.7	100.0	96.1	99.7	100.0	96.1	99.7	99.5	99.7	

Classification results of Raman prediction of the three pathological groups in the colon obtained using the leave-one-out cross-validation method in ν-SVM with different kernel functions: [a]the linear ν-SVM as a diagnostic algorithm (ν=1); [b]the polynomial ν-SVM as a diagnostic algorithm (ν=1, p=4); [c]the RBF ν-SVM as a diagnostic algorithm (ν=1, σ=30).

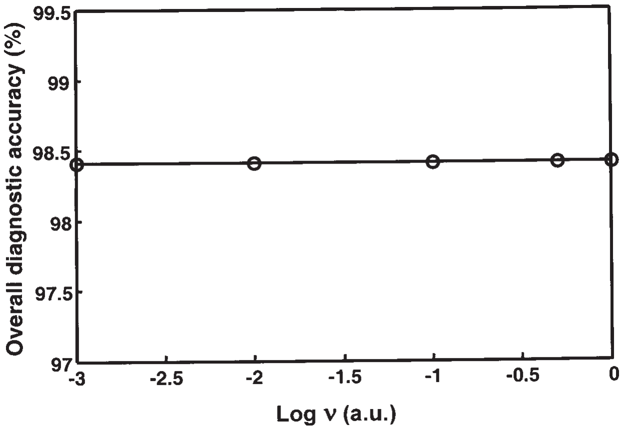

Figure 6. Dependence of overall diagnostic accuracy on the parameter ν using the linear ν-SVM algorithm.

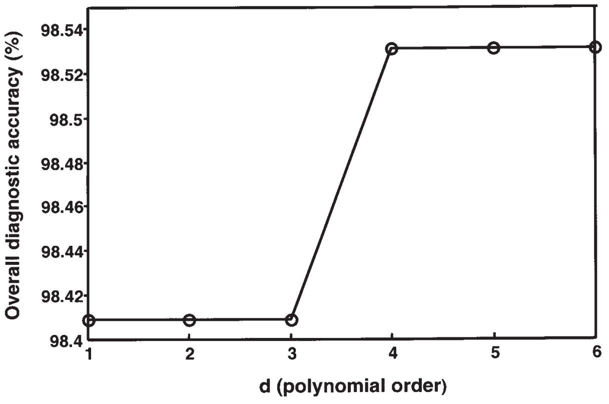

Figure 7. Dependence of the overall diagnostic accuracy on the order of the polynomial kernel using the polynomial ν-SVM algorithm.

non-linear SVM classifier was trained using various ν values (e.g. 0.001, 0.01, 0.1, 0.5 and 1) at a particular order of polynomial. Fig. 7 shows the obtained overall diagnostic accuracy as a function of the order *p* of polynomial using the polynomial ν-SVM. It was found that the overall classification accuracy at a fixed polynomial order did not depend on ν values. When the 3rd or 4th order of polynomial was employed at various ν values for classification, the overall diagnostic accuracy remained at 98.4 and 98.5%, respectively. Therefore, a fixed value of the parameter ν at 1 was then selected to exhaustively search for the optimal polynomial order of *p*. We found that the performance of classification of Raman spectra between the three histopathological groups of colonic tissues was only slightly improved even if a higher order *p* of polynomial (e.g. 4th order or above) was used as the kernel function in the polynomial ν-SVM algorithm (Fig. 7). A diagnostic sensitivity of 97.6, 100.0 and 99.8%; and specificity of 99.7, 100.0 and 96.1%, respectively, can be reached for differentiation between normal, polyps and cancer in the colon using the polynomial ν-SVM with the parameter ν of 1 and the polynomial order *p* of 4 (Table IIIb).

In the RBF ν-SVM classification, an approach similar to the polynomial ν-SVM above was taken to find the optimum value for the parameter ν in the RBF-SVM algorithm. A fixed radial width σ of 1 at various values of the parameter ν (e.g. 0.001, 0.01, 0.1, 0.5 and 1) was used to train the RBF ν-SVM classifier. We found that all these ν values yield similar overall diagnostic accuracy at 98.6%. This result again confirmed that the parameter ν did not have much impact on the RBF ν-SVM classification for the current Raman spectral dataset from colonic tissues. On the other hand, to search for the optimal Gaussian radial width σ, RBF ν-SVM classifier was trained for the different σ values chosen from a set of σ values ranging from 0 to 1000, with different step sizes chosen according to the σ values: i) a step size of 0.1 for σ values between 0 to 1; ii) 1 for σ values between 1 to 20; iii) 5 for σ values between 20 to 100 and iv) 50 for σ values between 100 to 1000. The training for various σ values was performed using a fixed parameter ν at 1. Fig. 8 shows the obtained overall diagnostic accuracy as a function of the Gaussian radial width σ using the RBF ν-SVM. The most optimum overall diagnostic accuracy of 99.6% can be achieved at σ values ranging from 30 to 150. Based on these optimal values of σ, for instance, σ = 30, a diagnostic sensitivity of 99.6, 100.0 and 99.8%; and a specificity of 99.7, 99.5 and 99.7%, respectively, can be obtained for differentiation between normal, polyps and cancer colonic tissue using the RBF ν-SVM with the parameter ν of 1 (Table IIIc).

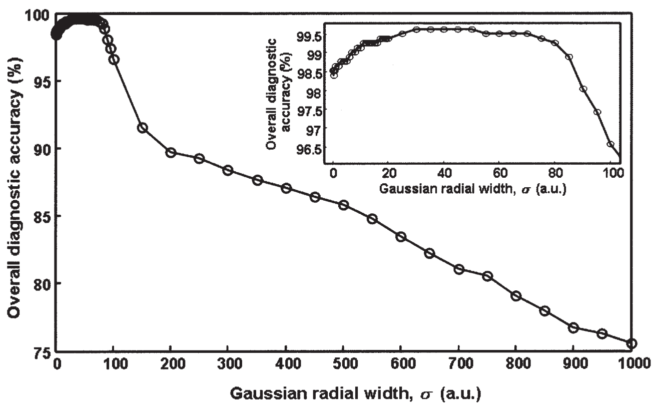

Figure 8. Dependence of the overall diagnostic accuracy on the parameter-Gaussian radial width σ using the RBF ν-SVM algorithm.

For comparison purposes, the cross-validation testing based on leave-one-sample-out was performed for all SVM models (linear, polynomial and RBF for both C- and ν-SVM) as described above. As such, a new dataset for multi-class classification was formed by re-arranging the 817 Raman data. Raman spectra measured from various locations in one sample were averaged and the average spectrum was assigned as independent data. With this new representation, a total of 105 independent colonic Raman spectra from 105 samples (41 normal, 18 polyps and 46 malignant tumors) were generated and used to re-evaluate all SVM models (data not shown). Results showed that when we used the first 13 PCs as inputs for SVM modeling, all models gave overall diagnostic accuracy of 100%, except that the linear ν-SVM produced overall diagnostic accuracy of 98.2%. Hence, the overall diagnostic accuracies obtained using the leave-one-sample-out cross-validation method are similar or even superior to the results using the leave-one-spectrum-out cross-validation.

Discussion

One of the challenges of applying Raman spectroscopy in medical diagnostics is the interference of strong autofluorescence background that underlies the weak tissue of Raman spectra. Under the 785-nm laser excitation in this study, the autofluorescence background intensity of colonic tissues can be 8 to 30 times more intense than the weak Raman signal in the spectral region of 800 to 1800 cm^{-1} (i.e. 838-915 nm) probed by the macro-Raman fiber probe (8,31). That is, tissue Raman signal only constitutes a very small fraction (3 to 12%) of overall measured spectral signals (over 88% is tissue autofluorescence). If autofluorescence background was left untreated, it can dominate the Raman spectra and make analysis of tissue biomolecules and biochemistry impossible. Therefore, the autofluorescence contribution must be minimized in order to resolve and analyze the Raman signals from tissues. As such, different from previous work that directly applied SVM on micro-Raman spectra of single cells, which displayed a relatively less fluorescent background contribution, without further data processing (44,45), the preprocessing approach on macro-Raman data acquired from colonic tissues is critical in this study in removing the autofluorescence interference while enhancing very weak tissue Raman signals for better tissue differentiation using various models of SVM.

With the combined PCA and SVM techniques, the dimensions of tissue Raman spectra were dramatically reduced from 544 variables (Raman spectral data in wavelength or wavenumber space) down to 18 variables (PC scores) to significantly simplify the computational complexities for the development of effective SVM algorithms for multi-class classification of colonic tissues without sacrificing diagnostic accuracy. Further studies showed that the performance of various SVM algorithms were not significantly improved by adding more PCs, indicating that the first 18 PCs used already contained most of the diagnostically significant information for effective tissue classification in the current dataset. These results demonstrate that high diagnostic performances of multi-class classification can be achieved using the SVM-based algorithms, though there is a slight variation in diagnostic accuracy between the linear, polynomial and RBF kernels as well as a slight variation in the performance between C-SVM and ν-SVM techniques using the same kernel but with different parametric settings (Figs. 3-8). The performance of leave-one-out cross-validation on the classification of colonic Raman spectra using the C-SVM and ν-SVM algorithms are generally very good, with an overall diagnostic accuracy of >98% being achieved. For instance, in the C-SVM classification, the maximum overall diagnostic accuracy for linear, polynomial and RBF kernels are 99.3, 99.4 and 99.9%, respectively. In the ν-SVM classification, the maximum overall diagnostic accuracy for linear, polynomial and RBF kernels are 98.4, 98.5 and 99.6%, respectively. All the benign tissues (e.g. hyperplastic polyps) can be identified from normal and cancerous tissue using the C-SVM technique. The high diagnostic accuracy of SVM-based algorithms appears to be due to the fact that SVM utilizes higher order correlations for tissue classification (29).

Essentially, the ν-SVM algorithms are not very different from C-SVM algorithms, however, with the introduction of ν parameter to effectively control the number of support vectors, the tedious procedure for selecting regularization constant *C* for data classification can be avoided. In the ν-SVM, ν parameter has values ranging from 0 to 1 only and is convenient and intuitive for specifying a fraction of data points which are allowed to be errors (46). This is different from C-SVM that utilizes the parameters such as parameter *C* that does not have any intuitive interpretation for classification (22,27). Our studies showed that the cross-validation results did not depend on the ν values for any types of the kernel functions used. This indicated that in the ν-SVM classification, the selection of the values of the parameter ν was not critical and the diagnostic performance only depended on the order *p* of polynomial for polynomial kernel as well as the radial width σ for RBF kernel. Furthermore, the classification accuracies of ν-SVM on colonic Raman data are only slightly lower than those of C-SVM for all types of kernel functions (Tables I and III). The main advantage of ν-SVM classification is that the parameter optimization, especially the selection of radial width σ of RBF kernel function is straightforward and more efficient compared to the operation of the C-SVM and the computational complexity of ν-SVM algorithms for multi-class classification can be significantly reduced when the ν values do not give much impact on diagnostic performance (Fig. 6).

It was evident that the diagnostic accuracy on colon Raman spectral data can be further improved using the nonlinear kernel functions, especially using RBF kernel which produced a sensitivity and specificity higher than the linear SVM in C-SVM and ν-SVM. This also suggests that the optimal separating hyperplane for Raman data of multi-class tissues is not linear. Hu and Lin (33) found that RBF kernel was the most reasonable choice in SVM due to its simplicity and ability to model data of arbitrary complexity. In fact, the linear kernel is only part of the RBF kernel and the number of hyperparameters which influence the complexity of model selection for RBF kernel is much lower compared to the polynomial kernel (33). Although parameter optimization in RBF C-SVM is more tedious (including regularization constant *C* and Gaussian radial width σ), we found that the RBF C-SVM classifier yielded the best classification performance among the other SVM algorithms for differentiation between normal, polyps and malignant tissues (overall diagnostic accuracy 99.9%, with only one cancer tissue being misclassified). The best performance of multi-class classification on colonic Raman spectral data using the RBF C-SVM algorithm in this study is consistent with the results reported by Lin *et al* (27) who used RBF kernel in SVM for binary group classification on autofluorescence spectra from *in vivo* nasopharyngeal tissues.

In conclusion, the combined PCA-SVM techniques were successfully implemented for accurate multi-class classification of NIR Raman spectra from different types of pathological colonic tissues. A number of effective diagnostic algorithms based on C-SVM and ν-SVM techniques with different kernel functions were developed and the relative diagnostic performances of the algorithms were comprehensively evaluated and compared. The radial basis function (RBF) C-SVM algorithm has proven to be the best classifier for providing the highest diagnostic accuracy (~99%) for differentiation between different histopathological groups of colonic tissue. Therefore, the methodology developed for multi-class classification of tissue Raman spectra using SVM is useful for tackling the histopathological grouping problems faced in clinical settings. NIR Raman spectroscopy in combination with SVM has the potential to provide an effective and accurate diagnostic means for cancer diagnosis in the colon.

Acknowledgements

This study was supported by the BioMedical Research Council, the Academic Research Fund from the Ministry of Education, the Institute of Chemical and Engineering Sciences and the Faculty Research Fund from the National University of Singapore.

References

1. Ferrandez A and DiSario JA: Colorectal cancer: screening and surveillance for high-risk individuals. Expert Rev Anticancer Ther 3: 851-862, 2003.
2. Chia KS, Seow A, Lee HP and Shanmugaratnam KS: Cancer incidence in Singapore 1993-1997. Report No. 5. Singapore Cancer Registry, Singapore, 2000.
3. Frank CJ, McCreery RL and Redd DC: Raman spectroscopy of normal and diseased human breast tissues. Anal Chem 67: 777-783, 1995.
4. Mahadevan-Jansen A and Richards-Kortum R: Raman spectroscopy for the detection of cancers and precancers. J Biomed Opt 1: 31-70, 1996.
5. Shim MG, Wong LKS, Marcon NE and Wilson BC: *In vivo* near-infrared Raman spectroscopy: demonstration of feasibility during clinical gastrointestinal endoscopy. Photochem Photobiol 72: 146-150, 2000.
6. Stone N, Stavroulaki P, Kendall C, Birchall M and Barr H: Raman spectroscopy for early detection of laryngeal malignancy: preliminary results. Laryngoscope 110: 1756-1763, 2000.
7. Bakker Schut TC, Witjes MJ, Sterenborg HJ, Speelman OC, Roodenburg JL, Marple ET, Bruining HA and Puppels GJ: *In vivo* detection of dysplastic tissue by Raman spectroscopy. Anal Chem 72: 6010-6018, 2000.
8. Huang Z, McWilliams A, Lui H, McLean D, Lam S and Zeng H: Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. Int J Cancer 107: 1047-1052, 2003.
9. Caspers PJ, Lucassen GW and Puppels GJ: Combined *in vivo* confocal Raman spectroscopy and confocal microscopy of human skin. Biophys J 85: 572-580, 2003.
10. Gniadecka M, Wulf HC, Nielsen OF, Christensen DH and Hercogova J: Distinctive molecular abnormalities in benign and malignant skin lesions: studies by Raman spectroscopy. Photochem Photobiol 66: 418-423, 1997.
11. Mizuno A, Kitajima H, Kawauchi K, Muraishi S and Ozaki Y: Near-infrared Fourier transform Raman spectroscopic study of human brain tissues and tumors. J Raman Spectrosc 25: 25-29, 1994.
12. Mahadevan-Jansen A, Mitchell MF, Ramanujam N, Malpica A, Thomsen S, Utzinger U and Richards-Kortum R: Near-infrared Raman spectroscopy for *in vitro* detection of cervical precancers. Photochem Photobiol 68: 123-132, 1998.
13. Lau DP, Huang Z, Lui H, Man CS, Berean K, Morrison MD and Zeng H: Raman spectroscopy for optical diagnosis in normal and cancerous tissue of the nasopharynx - preliminary findings. Lasers Surg Med 32: 210-214, 2003.
14. Lau DP, Huang Z, Lui H, Morrison MD, Shen L and Zeng H: Raman Spectroscopy for Optical Diagnosis in the Larynx - Preliminary Findings. Lasers Surg Med 37: 192-200, 2005.
15. Huang Z, Lui H, McLean DI, Korbelik M and Zeng H: Raman spectroscopy in combination with background near-infrared autofluorescence enhances the *in vivo* assessment of malignant tissues. Photochem Photobiol 81: 1219-1226, 2005.
16. Huang Z, Lui H, Chen XK, McLean DI and Zeng H: Raman spectroscopy of *in vivo* cutaneous melanin. J Biomed Opt 9: 1198-1205, 2004.
17. Molckovsky ALM, Wong KS, Shim MG, Marcon NE and Wilson BC: Diagnostic potential of near-infrared Raman spectroscopy in the colon: differentiating adenomatous from hyperplastic polyps. Gastrointest Endosc 57: 396-402, 2003.
18. Ryder AG, O'Connor GM and Glynn TJ: Quantitative analysis of cocaine in solid mixtures using Raman spectroscopy and chemometric methods. J Raman Spectrosc 31: 221-227, 2000.
19. Vapnik VN: Statistical Learning Theory, Wiley, New York, 1998.
20. Cortes C and Vapnik VN: Support vector networks. Mach Learn 20: 273-297, 1995.
21. Burges CJC: A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2: 121-167, 1998.
22. Hearst M: Using SVMs for text categorisation. IEEE Intelligent Systems 13: 18-28, 1998.
23. Osuna E, Freund R and Girosi F: Training support vector machines: an application to face detection. Proc IEEE Conf on Computer Vision and Pattern Recognition, pp130-136, 1997.
24. Scholkopf B: Support Vector Machines - a practical consequence of learning theory. IEEE Intelligent Systems 13: 29-39, 1998.
25. Joachims T: Text characterization with support vector machines. Technical Report LS VIII Number 23, University of Dortmund, 1997.
26. Bonneville M, Meunier J, Bengio Y and Soucy JP: Support vector machines for improving the classification of brain pet images. Proc SPIE 3338: 264-273,1998.
27. Lin WM, Yuan X, Yuen P, Wei WI, Sham J, Shi PC and Qu J: Classification of *in vivo* autofluorescence spectra using support vector machines. J Biomed Opt 9: 180-186, 2004.

28. Palmer GM, Zhu C, Breslin TM, Xu F, Gilchrist KW and Ramanujam N: Comparison of multiexcitation fluorescence and diffuse reflectance spectroscopy for the diagnosis of breast cancer. IEEE Trans Biomed Eng 50: 1233-1242, 2003.
29. Majumder SK, Ghosh N and Gupta PK: Support vector machine for optical diagnosis of cancer. J Biomed Opt 10: 024034, 2005.
30. Widjaja E, Zheng W and Huang Z: Classification of ENT tissue using near-infrared Raman spectroscopy and support vector machines. Proc SPIE 5862: 25-30, 2005.
31. Huang Z, Zeng H, Hamzavi I, McLean DI and Lui H: Rapid near-infrared Raman spectroscopy system for real-time *in vivo* skin measurements. Opt Lett 26: 1782-1784, 2001.
32. Zomer S, Brereton RG, Carter JF and Eckers C: Support vector machines for the discrimination of analytical chemical data: application to the determination of tablet production by pyrolysis - gas chromatography-mass spectrometry. Analyst 129: 175-181, 2004.
33. Hu CW and Lin CJ: A Comparison of methods for multi-class support vector machines. IEEE Trans Neural Netw 13: 415-425, 2002.
34. Scholkopf B, Smola AJ, Williamson RC and Bartlett PL: New support vector algorithms. Neural Comput 12: 1207-1245, 2000.
35. Martin DR, Fowlkes CC and Malik J: Learning to detect natural image boundaries using brightness and texture. In Advances in Neural Information Processing Systems 14: 1352-1389, 2002.
36. Chang CC and Lin CJ: Training ν-support vector classifiers: theory and algorithms. Neural Comput 13: 2119-2147, 2001.
37. Devore JL: Probability and statistics for engineering and the science. Brooks/Cole, Pacific Grove, 1992.
38. Lachenbruch P and Mickey RM: Estimation of error rates in discriminant analysis. Technometrics 10: 1-11, 1968.
39. Dillion RW and Goldstein M: Multivariate analysis: methods and applications. John Wiley and Sons, New York, 1984.
40. Chang CC and Lin CJ: LIBSVM - a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm
41. Huang Z, Zheng W and Colin S: Near-infrared Raman spectroscopy for colonic cancer diagnosis. Proc SPIE 5862: 25-29, 2005.
42. Chapelle O, Vapnik V, Bousquet O and Mukherjee S: Choosing multiple parameters for support vector machines. Machine Learn 46: 131-159, 2002.
43. Hu CW, Chang CC and Lin CJ: A practical guide for support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
44. Rosch P, Harz M, Peschke KD, Ronneberger O, Burkhardt H, Schule A, Schmauz G, Lankers M, Hofer S, Thiele H, Motzkus HW and Popp J: On-line monitoring and identification of bioaerosols. Anal Chem 78: 2163-2170, 2006.
45. Rosch P, Harz M, Peschke KD, Ronneberger O, Burkhardt H and Popp J: Identification of single eukaryotic cells with micro-Raman spectroscopy. Biopolymers 82: 312-316, 2006.
46. Nadeau C and Bengio Y: Inference for the generalization error. Advances in Neural Information Processing Systems 12, MIT Press, 2000.