Journal of
Biomedical Optics

SPIEDigitalLibrary.org/jbo

Study of support vector machine and serum surface-enhanced Raman spectroscopy for noninvasive esophageal cancer detection

Shao-Xin Li
Qiu-Yao Zeng
Lin-Fang Li
Yan-Jiao Zhang
Ming-Ming Wan
Zhi-Ming Liu
Hong-Lian Xiong
Zhou-Yi Guo
Song-Hao Liu

SPIE

Journal of Biomedical Optics 18(2), 027008 (February 2013)

Study of support vector machine and serum surface-enhanced Raman spectroscopy for noninvasive esophageal cancer detection

Shao-Xin Li,[a,b*] Qiu-Yao Zeng,[d*] Lin-Fang Li,[d] Yan-Jiao Zhang,[c] Ming-Ming Wan,[a] Zhi-Ming Liu,[a] Hong-Lian Xiong,[a] Zhou-Yi Guo,[a] and Song-Hao Liu[a]
[a]South China Normal University, School of Information and Optoelectronic Science and Engineering, Guangzhou 510631, China
[b]Guangdong Medical College, School of Information Engineering, Dongguan 523808, China
[c]Guangdong Medical College, School of Basic Medicine, Dongguan 523808, China
[d]Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China and Department of Clinical Laboratory, Guangzhou, 510060, China

Abstract. The ability of combining serum surface-enhanced Raman spectroscopy (SERS) with support vector machine (SVM) for improving classification esophageal cancer patients from normal volunteers is investigated. Two groups of serum SERS spectra based on silver nanoparticles (AgNPs) are obtained: one group from patients with pathologically confirmed esophageal cancer ($n = 30$) and the other group from healthy volunteers ($n = 31$). Principal components analysis (PCA), conventional SVM (C-SVM) and conventional SVM combination with PCA (PCA-SVM) methods are implemented to classify the same spectral dataset. Results show that a diagnostic accuracy of 77.0% is acquired for PCA technique, while diagnostic accuracies of 83.6% and 85.2% are obtained for C-SVM and PCA-SVM methods based on radial basis functions (RBF) models. The results prove that RBF SVM models are superior to PCA algorithm in classification serum SERS spectra. The study demonstrates that serum SERS in combination with SVM technique has great potential to provide an effective and accurate diagnostic schema for noninvasive detection of esophageal cancer. *© 2013 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JBO.18.2.027008]

Keywords: surface-enhanced Raman spectroscopy; support vector machines; esophageal cancer; serum.

Paper 12725R received Nov. 8, 2012; revised manuscript received Dec. 29, 2012; accepted for publication Jan. 17, 2013; published online Feb. 6, 2013.

1 Introduction

Esophageal cancer is one of the common malignant tumors worldwide with an annual incidence of about 300 thousand cases and a five-year survival rate of 10% to 25%. Early diagnosis and treatment is important to improve the survival rate.[1] However, there are many disadvantages in conventional esophageal cancer screening methods such as white-light endoscopy and biopsy. For example, endoscopy check depends on visual identification of gross morphological tissue changes. It is difficult to detect subtle early neoplastic changes, and the examination results are relevant to the skill of physician. Excision biopsy is invasive and impractical for a high-risk patient with multiple suspicious lesions. Therefore, it is urgently desirable to develop a noninvasive means to early diagnose esophageal cancer.

In recent years, the Raman spectroscopy technique has received a great deal of interest in the biomedical field.[2–4] It has been applied to differentiate normal and malignant tissues of various body sites, include breast, bladder, lung, prostate, cervix, skin, etc.[5–8] Raman spectroscopy is a molecular vibration spectral method discovered by the Indian scientist Raman C.V. in 1928. It can provide fingerprinting type information about the structure and conformation of macromolecules such as proteins, lipids and nucleic acids. Compared with the fluorescence spectroscopy and infrared absorption spectroscopy, Raman spectroscopy has many advantages.[4] For instance, there is no photobleaching in Raman scattering, and Raman spectral peaks are narrow.

However, two disadvantages of Raman spectroscopy make it difficult to practical application in clinic diagnosis. One of drawbacks is that the efficiency of Raman scattering is very low due to its extremely small cross-section (10^{-30} to 10^{-25} cm^2); the other is that a strong fluorescence background of biological samples make it difficult to extract from the original signal. With the discovery of surface-enhanced Raman spectroscopy (SERS) in 1974 by Fleischman et al., Raman spectroscopy technique acquired a rapid development.[9] Single molecule absorbing onto a single silver nanoparticle has been successful probed by SERS. This facilitates the application of SERS technology in the detection of biological materials such as DNA, RNA, and proteins. The most recent reports show that SERS has been used for target detection of tumor markers in the blood or on the cell surface by the immunoassay approaches.[10–12]

Blood samples are ideal disease screening materials for noninvasive diagnosis. They are rich in proteins, fates, cholesterol, etc. At the early stage of cancer, these components will undergo subtle changes which can be revealed by SERS. Chen R. group has researched several types of patients include gastric cancer, nasopharyngeal carcinoma, colon tumor with blood SERS. They successfully distinguish cancer patients from normal volunteers with sensitivity and specificity of 90% or more.[13–15]

*These authors contributed equally to this work.

Address all correspondence to: Zhou-Yi Guo, South China Normal University, School of Information and Optoelectronic Science and Engineering, Guangzhou 510631, China. Tel: 86-020-85211428; Fax: 86-020-85216052; E-mail: ann@scnu.edu.cn

0091-3286/2013/$25.00 © 2013 SPIE

However, the differences of Raman spectroscopy between normal and pathologic tissues are usually tiny that it is difficult to differentiate them with direct means. The powerful and robust spectral data processing algorithms are much needed to extract effective diagnostic information. Multivariate statistical techniques such as principal component analysis (PCA), linear discriminant analysis (LDA) and artificial neural networks (ANNs) have been successfully used to develop diagnostic algorithms.[5,15,16] For example, Feng et al., use PCA-LDA technique to distinguish serum SERS of nasopharyngeal cancer, they acquire sensitivity 90.7% and specificity 100%.[13]

Another powerful multivariate technique, support vector machine (SVM), which was introduced by Vapnik and Burges, has attracted great attention due to its capability of representing nonlinear features.[17,18] The SVM technique has now been applied to classify spectral data for tissue diagnose. For instance, Huang et al., used PCA-SVM techniques to classify multiclass Raman spectra from different types of pathological colonic tissues.[19] Lin et al., implemented linear and nonlinear SVM methods for the classification of autofluorescence spectroscopy from nasopharyngeal carcinomas and normal tissues with diagnostic accuracy higher than that of PCA-LDA.[20] To date, the application of SVM to distinguish serum SERS for noninvasive cancer detection has not yet been reported. This study aims to explore the potential of SVM technique combining with serum surface-enhanced Raman spectroscopy (SERS) to detect esophageal cancer. Three multivariate statistic analysis methods including principal components analysis (PCA), conventional SVM (C-SVM) and conventional SVM combination with PCA (PCA-SVM) methods were implemented to classify the same spectra dataset. The diagnostic performance of all SVM models were exhaustively optimized and evaluated by leaving one out cross validation method.

2 Materials and Methods

2.1 Synthesis of AgNPs

AgNPs was synthesized following the reported method.[21] First, 18 mg of $AgNO_3$ were added to 100 ml water under strong stirring. When the solution was boiling, 2 ml of citrate sodium (1 wt.%) was slowly added into the solution. Then the solution was boiled for 40 min and cooled in ambient conditions. The AgNPs was characterized by an absorption maximum at 424 nm in Fig. 1. A transmission electron microscopy (TEM) photograph of the prepared silver colloid was showed in the inserted picture.

2.2 Preparation of Human Serum Samples

Serum samples were collected from 61 individuals consisting of 31 healthy volunteers and 30 esophageal cancer patients who were confirmed clinically with histopathology. In these 30 cancerous patients, 24 cases were mid-cancers (II, III stage), 3 cases early cancers (I stage) and 3 cases advanced cancers (IV stage). All patients were from Sun Yat-sen University Cancer Center and signed an informed consent to permit collection of blood prior to research. After 12 h of overnight fasting, a single 3 ml peripheral blood samples were obtained from the study subjects between 7:00 and 8:00 A.M. Serum was obtained by extracting supernatant from blood samples centrifuged at 3500 rpm for 5 min.

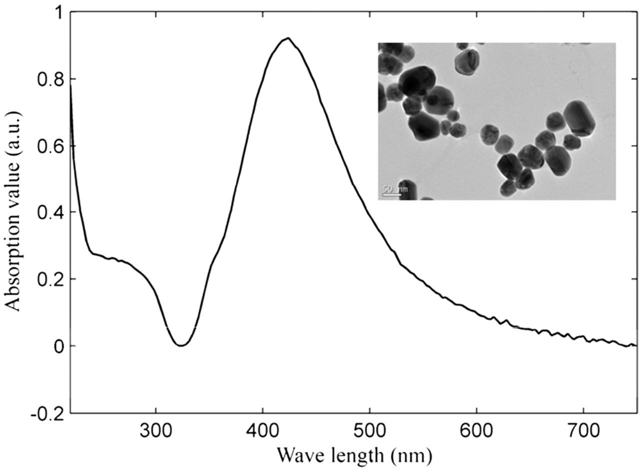

Fig. 1 UV/visible absorption spectrum of the Ag colloid. The absorption maximum is located at 424 nm. The inserted picture is the TEM micrograph of Ag colloidal surface.

Before SERS measurement, 20 μL silver colloidal nanoparticles were mixed with 20 μL serum. The mixture was stirred with the pipette tip and then incubated 1 h at room temperature. Next, a drop of this mixture was transferred onto an aluminum plate and dry naturally 1 h for SERS measurement.

2.3 SERS Measurements and Data Preprocessing

The Raman spectroscopy was recorded with a confocal Raman microscopy (Renishaw, inVia, United Kingdom) in the range of 600 to 1800 cm^{-1} with a spectral resolution about 1 cm^{-1} under a 785 nm diode laser excitation. The power of laser exposed on sample is about 0.5 mw with a spot diameter about 5 μm. The spectra were collected in back-scattered geometry using a Leica DM2500 microscope equipped with objective 20×; The software package WIRE 3.2(Renishaw) was employed for spectral acquisition and analysis. Each Raman spectra was accumulated two times with an integration time of 10 s. Each sample was collected three spectra and then these three spectra were averaged. All data were collected under the same conditions.

A fifth-order polynomial was employed to fit the autofluorescence background, and then this polynomial was subtracted from original spectra. In order to compare the changes of spectral shapes and relative peak intensities among different serum samples, the area normalized of spectra under the curve was employed. Vancouver Raman algorithm was used to spectra smoothed and baseline correct.[22] It is an automated autofluorescence background subtraction algorithm based on modified multipolynomial fitting.

2.4 Support Vector Machine

Support vector machine (SVM) is a relatively young multivariate data classification method and was first proposed by Vapnik. It is based on the principal of minimization of structural risk by the appropriate choice of function subset and discriminant function, ensuring the actual risk of learning machines to a minimum. Thus, it is an excellent learning machine with optimal classified ability and generalization ability. SVM has been applied successfully in many fields, such as face recognition, text categorization, gene selection, and so on.[20,23–25] Compared with other multivariate statistical methods, SVM has many advantages: first, it is a powerful way to classify a small size of datasets; second, it can give reproducible solutions; third, it has the ability to deal with class boundaries with complex conditions by replacing the kernel functions.

For linearly separable binary classification sample sets, SVM finds an optimal hyperplane to maximize the margin between them. When the sample sets are nonlinear nonseparable, SVM maps the sample data to a higher dimensional feature space to linearize the boundary of sample sets by specific kernel functions. The three most frequently used kernel functions are:

$$\text{Linear}: \; K(x_i, x_j) = x_i \cdot x_j + 1, \tag{1}$$

$$\text{Polynomial kernel}: \; K(x_i, x_j) = (x_i \cdot x_j + 1)^d, \tag{2}$$

$$\text{Gaussian radial basis function (RBF)}: \; K(x_i, x_j) = \exp\left[\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right], \tag{3}$$

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

To get a support vector machine classifier with good generalization ability, choice of an appropriate kernel function which projects data to the feature space is critical. For a given kernel function, optimization of the corresponding parameter is also important, such as the polynomial order d in the polynomial kernel and Gaussian width σ in the Gaussian RBF kernel. Once the data are mapped into the feature space an infinite number of separating hyperplanes may exist, creating the risk of overfitting the hyperplanes to the given data points. The overfit hyperplane can perfectly separate training data, but it poorly predicts unseen data. To overcome this problem, a penalty factor C is introduced to allow some training data to be misclassified, the higher C the lower misclassification rate.

In this work, conventional SVM technique (C-SVM) and conventional SVM combining with PCA method (PCA-SVM) are to be assessed. For every type of SVM technique, linear and RBF kernel functions are employed to build two classes of diagnostic algorithms. All diagnostic algorithms are optimized with grid search and evaluated by leave one sample out cross validation method, which involves using one sample held out from dataset as the validation data, and the remain samples as the training data, this process is repeated such that each sample is used once as the validation data. The optimization criterion is to maximize overall diagnostic accuracy obtained from leave one out cross validation methods.

The optimized algorithms developed from the training set are used to classify the withheld spectra measured from one subject. This procedure is repeated until all 61 mean spectra measured from 61 subjects are classified. The overall diagnostic accuracy of a particular algorithm is calculated based on the classifications of the withheld spectra over 61 rounds of cross validation. In order to compare the performance of SVM models, the same dataset is used for the investigation of PCA. The leave one sample out cross validation is executed to test the performance of PCA. The LIBSVM toolbox 3.1 created by Chang and Lin is used for SVM classifications. All the procedure is implemented with MATLAB language.

3 Results

We have measured the regular Raman spectra and surface-enhanced Raman spectra of serum sample come from the same esophageal cancer patient in order to assess the silver colloid enhancement effects on the serum Raman scattering. Figure 2(a) displays SERS spectra of the serum sample from an esophageal cancer patient by mixing the serum with silver colloid at a 1∶1 proportion, Fig. 2(b) the regular Raman spectra of serum sample from the same patient without the Ag colloid. It is clearly shows that there is almost no Raman peak observed for the regular Raman spectra of serum sample without silver colloid, while there is a huge enhanced for the SERS spectra with Ag colloid. The dramatic increase in many dominant vibration bands indicates there is a strong interaction between AgNPs and serum. Because of this interaction, biochemical substances of serum sample absorb closely on the surface of silver particles, resulting in a tremendous enhanced of Raman scattering.

A total of 61 serum SERS spectra were obtained, in which 30 Raman spectra were from cancerous patients and 31 from normal subjects. Figure 3 shows normalized average SERS spectra ±1 standard deviations of esophageal cancer and normal serum in the range from 600 to 1800 cm^{-1}. The shade area represents the standard deviations. Primary Raman peaks are observed in normal and cancerous esophageal serum at 683 cm^{-1} (C-S twist), 722 cm^{-1} (C-H bending adenine, coenzyme), 858 cm^{-1} (C-C stretch of proline ring, ring breathing of tyrosine), 915 cm^{-1} (C-C stretch of proline ring, glucose), 1020 cm^{-1} (C-H stretch of phenylalanine), 1130 cm^{-1} (C-N stretch, D-mannos), 1219 cm^{-1} (C C_6H_5 phenylalanine, tryptophan), 1314 cm^{-1} (CH_3CH_2 twisting collagen/lipids), 1346 cm^{-1} (CH_3CH_2 wagging, tryptophan adenine, guanine), 1445 cm^{-1} (CH_2 bending, collagen/lipids), 1585 cm^{-1} (C C bending, phenylalanine, acetoacetate, riboflavin). The strongest peaks are at 1219, 1346, 1445, and 1585 cm^{-1}.[13,14,26,27] The spectral differences between esophageal cancer and normal serum are clearly displayed in Fig. 3(c). The distinct differences imply that there is an enormous potential to diagnosis esophageal cancer with serum SERS technique.

We employed linear and RBF kernel functions to develop C-SVM diagnostic algorithms for classification serum SERS

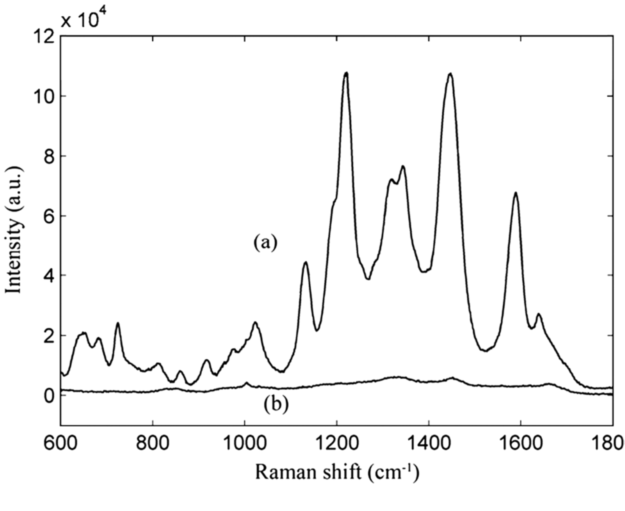

Fig. 2 (a) SERS spectra of the serum sample from a patient with esophageal cancer obtained by mixing the serum with Ag colloid at a 1∶1 proportion. (b) The regular Raman spectra of serum sample from the same patient without the Ag colloid.

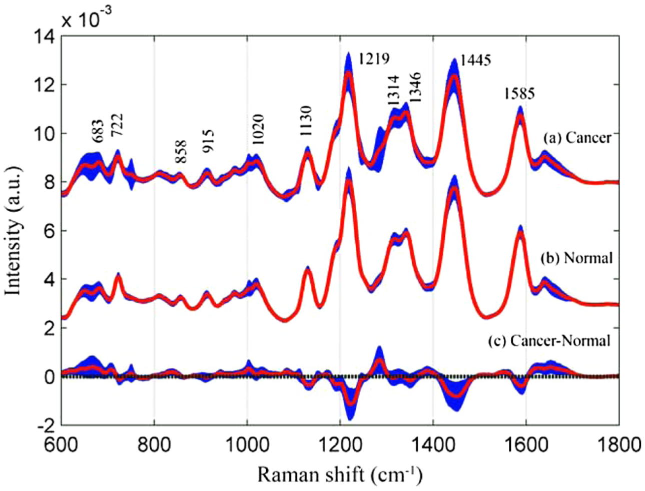

Fig. 3 Normalized mean SERS spectra of 30 esophageal cancer and 31 normal serum sample. (a) Cancer, (b) normal, (c) difference spectra (cancer-normal, the difference spectra intensity is enlarged five times for clear display), shade area represents the standard deviations.

spectra of esophageal cancer patients from normal volunteers. The linear C-SVM algorithm needs to be optimized to search the optimal parameter C, which gives the best trade-off between the training error and generalization ability. The search range for parameter C is performed from 2^{-10} to 2^5 with step of power of two. The overall diagnostic accuracy as a function of parameter $\log_2 C$ is shown in Fig. 4. It is displayed that the largest diagnostic accuracy of 73.3% is acquired with the parameter C value at 2^{-8}. Table 1 shows the classification results of the serum SERS spectra using the leave one out cross validation with the linear C-SVM algorithm at parameter $C = 2^{-8}$. A diagnostic sensitivity of 73.3% and specificity of 71.0% can be obtained.

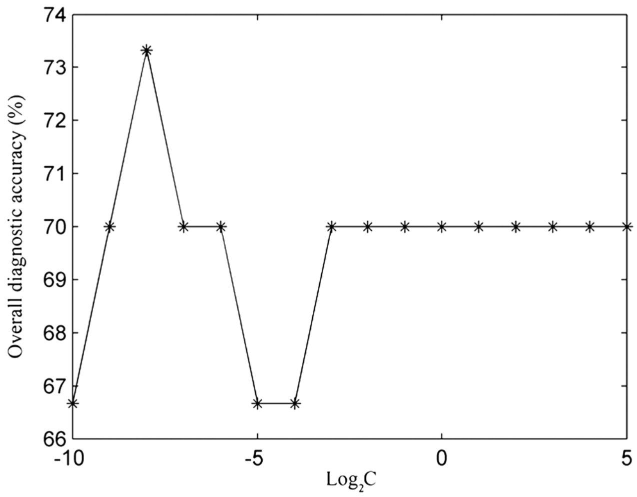

Fig. 4 Dependence of classification accuracy on parameter C for a linear SVM algorithm.

Table 1 Results of classification of serum SERS with different algorithms.

	C-SVM		PCA-SVM		
Algorithm	Linear	RBF	Linear	RBF	PCA-LDA
Specificity (%)	71.0	83.9	77.4	86.7	77.4
Sensitivity (%)	73.3	83.3	76.7	83.3	76.7
Accuracy (%)	72.1	83.6	77.0	85.2	77.0

In the development of RBF kernel C-SVM algorithm, the parameter C and Gaussian width σ are to be optimized to build the efficient classifier. In this study the grid search is implemented to exhaustively search optimal parameters by trying various pairs of parameters. The range of C is from 2^{-10} to 2^{15} and Gaussian width σ from 0 to 2^{-15}. Figure 5 shows the three-dimensional (3-D) map of overall diagnostic accuracy as a function of parameter C and Gaussian radial width σ. It is clearly exhibits that a smaller C or σ gives lower diagnostic accuracy. The largest diagnostic accuracy of 80.0% locates at $C = 2^8$ and $\sigma = 2^{-14}$. Table 1 lists the cross validation results of serum SERS spectra at the highest overall diagnostic accuracy in the RBF C-SVM. The diagnostic sensitivity of 83.3% and specificity of 83.9% are achieved for differentiation serum SERS spectra between esophageal cancers and normal subjects in the using the RBF C-SVM algorithm.

In this study, the range of Raman spectra is from 600 to 1800 cm^{-1} with 1108 variables, such a high-dimensional spectra would lead to a very complex calculation and time-consuming in the application of C-SVM. In addition, the spectra also contain many redundant data and noise. All these limit the efficiency of SVM technique. It is significant to reduce the dimensions of the spectral data by PCA technique to simplify the implementation of the SVM algorithm and to improve the performance.

PCA is a mathematical tool that reduces the dimensions of dataset using an orthogonal transformation to convert the observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs). The transformation is such that the first principal component has the largest possible variance, and each succeeding component has the highest variance possible under the constraint that it be orthogonal to the preceding components. Therefore the principal components are normally arranged in the order of their contributions to the variance of entire dataset. Most of the information carried in the dataset is distributed in

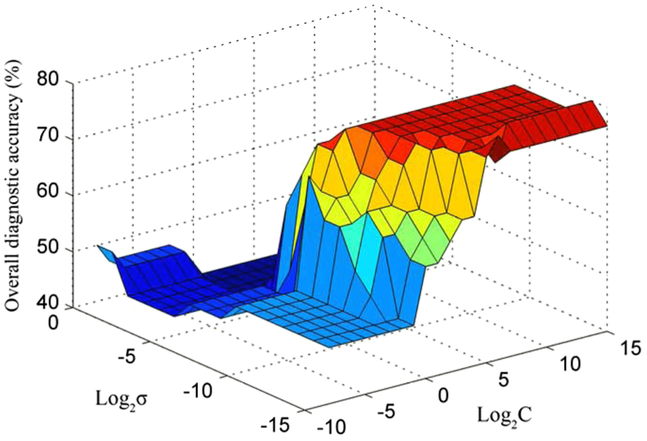

Fig. 5 3-D map of overall diagnostic accuracy as a function of parameter C and Gaussian radial width σ using the RBF C-SVM algorithm.

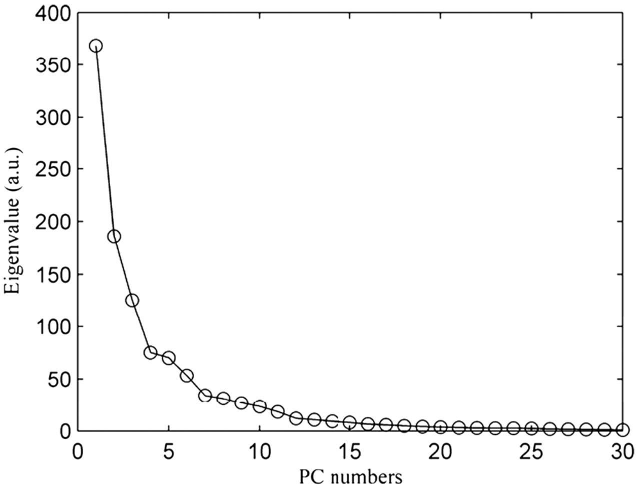

Fig. 6 Eigenvalues of principal components contributed to the total variance of all serum SERS spectra.

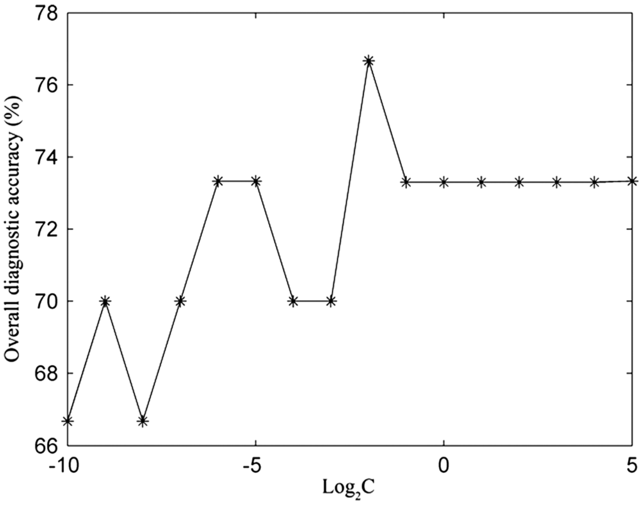

Fig. 7 Dependence of classification accuracy on parameter C for a linear PCA-SVM algorithm.

first few principal components, and the contributions of the rest of principal components are negligible. By this orthogonal transformation, the dimensions of dataset can be significantly reduced without losing important information.

When PCA is employed to process Raman spectra, it transforms Raman shift into a set of PC spectra. Each PC loading spectra is a combination of original Raman spectra. Figure 6 shows the contributions of eigenvalues of each principal component to the total variance of all serum SERS spectra. As shown in the figure, the eigenvalues drop off rapidly with increasing PC numbers, and the first few PCs retain the maximum variance of the data. For instance, the first two PCs account for 50.1% of the total variance; the first five PCs account for 74.6%; the first 10 PCs account for 89.8%; and the first 20 PCs account for 97.4%. In order to comparison the classification performance of SVM algorithm and PCA technique, PCA combination with linear discriminant analysis (PCA-LDA) is used to classify the same Raman dataset. The classification accuracy of 77.0% is obtained with the first 20 PCs scores. This result is lower than that of RBF kernel C-SVM algorithm.

Linear and RBF kernel functions are employed to develop PCA-SVM algorithm. The first 20 projection scores of serum SERS on PC loadings are used to build dataset for developing PCA-SVM models. All the parameters of PCA-SVM models need to be optimized to build efficient classifier. For linear PCA-SVM algorithm, the range of parameter C is set from 2^{-10} to 2^5. Figure 7 exhibits dependence of classification accuracy on parameter C. It is found that largest overall diagnostic accuracy of 76.7% is obtained with the parameter C of 0.25. The cross validation results of serum SERS spectra at the parameter $C = 0.25$ in the linear PCA-SVM are listed in Table 1. The diagnostic sensitivity of 76.7% and specificity of 77.4% are obtained.

In the RBF kernel PCA-SVM classification, grid search approach was implemented to find the optimum values of the parameter C and Gaussian width σ. The range of parameter C and σ is set from 2^{-10} to 2^{15} and from 0 to 2^{-15}. Figure 8 shows the 3-D map of overall diagnostic accuracy as a function of parameter C and Gaussian radial width σ. It is clearly exhibit that a smaller C gives lower classification accuracy. The largest overall diagnostic accuracy of 83.3% locates at $C = 2^3$ and $\sigma = 2^{-10}$. Based on the optimal values of C and σ, the diagnostic sensitivity of 83.3% and specificity of 86.7% are achieved in Table 1 for differentiation serum SERS spectra between normal volunteers and esophageal caner patients.

The performance of the RBF kernel PCA-SVM algorithm using different numbers of principal components scores was investigated. The results are displayed in Fig. 9. When the first five PCs are used, the classification accuracy is 73.7%. This result is higher than that of linear SVM algorithm with full spectra (72.1%). With the PCs adding up to 15, the classification accuracy increases to maximum 85.2%. When PCs continue to add, the classification accuracy of RBF PCA-SVM algorithm does not further increase. This illustrates that the first five 5 PCs capture the information of linearly separable samples, and the PCs from 5 to 15 seize the information of nonlinearly separable samples. With the PCA combination with nonlinear SVM methods, the number of variables is reduced from 1108 to 20. This shows that the PCA-SVM method considerably simplifies calculations without sacrificing diagnostic accuracy.

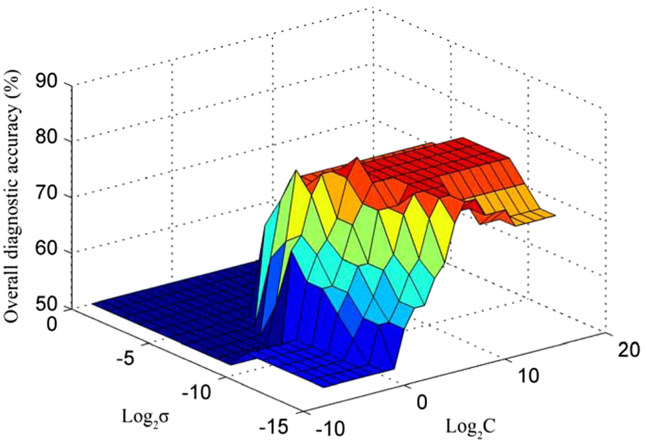

Fig. 8 3-D map of overall diagnostic accuracy as a function of the parameter C and Gaussian radial width σ using the RBF PCA-SVM algorithm.

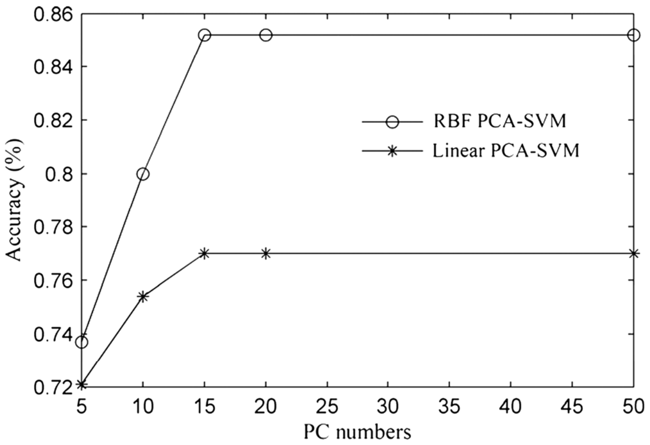

Fig. 9 Performance of PCA-SVM algorithm using principal component scores.

4 Discussion

Raman spectroscopy is a unique noninvasive detection technique that can acquire abundant structural feature and composition information of biomacromolecule. It may become a promising clinical diagnostic tool by probing subtle changes of biomolecule relevant to tissue pathology. Raman scattering signal is very weak, but with the appearance of SERS techniques, the applications of Raman spectroscopy in biomedical field are greatly advanced.[9,10,28,29] In this study we have measured serum SERS of normal volunteers and esophageal cancerous patients by Ag colloid. The dramatic increase in many dominant bands in Fig. 2 implies there is a great potential to diagnose the cancer by serum SERS technique. The specific differences of SERS spectra between normal subjects and cancerous patients are observed in Fig. 3. It reflects the changes of biochemical constitutes related to malignant transformation. For instance, a considerable decrease of 1445 cm^{-1} (CH_2 bending, collagen/lipids) in cancerous serum illustrates that the proportion of fat content is greatly reduced. The likely reason is that amplified cancerous cells consume a lot of fat, resulting in decrease of lipid molecules. This found is agreement with other reports in many cancer tissues.[30] The significant increase of Raman bands from 1620 to 1670 cm^{-1} (C O stretch, amide I, β-helix proteins) indicates a higher concentration of proteins in the β-pleated sheet conformation for esophageal cancer, which also had been found by Maziak et al., and Bergholt et al.[31]

The Raman spectral profiles of cancerous and normal tissues are similar except there are subtle differences of some spectral peaks. It is difficult to distinguish cancerous tissue from normal subjects by directly contrast characteristic Raman peak. The multivariate statistical analysis method is desirable to extract a lot of useful information from Raman spectra. PCA is a conventional multivariate statistical analysis method which converts high dimensional Raman spectra into several unrelated variables without loss of valid information by orthogonal transformation. It has been widely applied to the medical diagnosis of Raman spectroscopy.

SVM is another multivariate statistical analysis technique presented recently. It can process binary classification problem with nonlinear boundary by mapping to a higher dimensional space. The main advantages are that it can efficiently classify the small samples, regardless of the distribution nature of samples. In this work, we introduce C-SVM and PCA-SVM with linear and RBF kernel to differentiate serum SERS spectra of esophageal cancer patients from that of normal subjects. The performances of these diagnostic algorithms based on SVM techniques are comprehensively evaluated by leave one sample out cross validation method. To compare with conventional multivariate statistic methods, PCA-LDA was employed to classify the same dataset. The results listed in Table 1 show that the classification accuracy of RBF kernel SVM techniques (83.6%) is better than that of PCA-LDA methods (77.0%). Similar results are also proven by many other groups.[17,27,32] Possible reason maybe attribute to the fact that SVM utilizes nonlinear correlations for spectra classification.[17]

In the course of implementing SVM algorithm, it is important to select a proper parameter to acquire maximum diagnostic accuracy. In this study, grid search and leave one out cross validation method are employed to optimize the parameter C and Gaussian width σ for the RBF kernel SVM algorithm. The range of C is from 2^{-10} to 2^{15} and Gaussian width σ from 0 to 2^{-15} with step of power of two. We obtain the maximum overall diagnostic accuracy of 80.0% for the RBF C-SVM algorithm at C of 2^9 and σ of 2^{-14}, and 83.3% for the RBF PCA-SVM algorithm at C of 2^4 and σ of 2^{-11}. In order to confirm the optimization results, the local optimization around the optimal point is implemented further. For RBF C-SVM algorithm, the range of C is from 2^8 to 2^{10} and σ from 2^{-13} to 2^{-15} with step of $2^{0.2}$; for RBF PCA-SVM algorithm, the rang of C is from 2^2 to 2^4 and σ from 2^{-9} to 2^{-11} with step of $2^{0.2}$. The results reveal that there is no further increase for RBF C-SVM algorithm. This demonstrates that the optimal methods are effective for the study.

The present study shows that the maximum diagnostic accuracy of linear C-SVM algorithm is 72.1%, which is lower than that of linear PCA-SVM algorithm, 77.0%. The results indicate that the linear PCA-SVM algorithm is superior to the linear C-SVM algorithm in classification serum SERS spectra. The main reason may be that the linear C-SVM algorithm uses the entire spectra of 1108 variables, wherein there is some redundant information which affects the classification results. On the other hand, the linear PCA-SVM algorithm only employs 20 variables extracted from Raman spectra by PCA technique, which can refine important information contained in spectra and reduce redundant information. Similar results appear also in RBF kernel SVM algorithm. The largest overall diagnostic accuracy of RBF kernel C-SVM algorithm and RBF kernel PCA-SVM algorithm is 83.6% and 85.2%, respectively, indicating more redundant information contained in RBF C -SVM models than in RBF PCA -SVM.

The performance of PCA-SVM algorithm using different numbers of PCs is investigated in Fig. 9. The overall diagnostic accuracy of RBF kernel PCA-SVM models is higher than that of linear PCA-SVM models in the same number of PCs, indicating the existence of nonlinear boundary between cancerous and normal serum SERS spectra. The overall diagnostic accuracy of 72.1% for linear kernel PCA-SVM models in the first five PCs is equal to that of linear C-SVM. This manifests that the first five PCs contain all of linear classification information for classifying linearly separable samples.

The data space of C-SVM models contains entire spectra with 1108 variables. Such a high-dimensional data space will inevitable lead to computational complexity and time-consuming in optimizing and implementing SVM algorithm. For the purpose of simplifying implementation of SVM algorithm, the PCA method is introduced to reduce dimension of data space. With the combination of PCA and SVM techniques, the dimensions of serum Raman spectra are dramatically reduced from 1108 variables down to 20 variables, thus the computation is drastically decreased. This simplification is particularly important for applications that require rapid processing of a large amount of multivariate data, such as in real-time multispectral imaging and optical processing systems.

As can be found from Table 1, the nonlinear SVM produces a diagnostic accuracy higher than the linear SVM, especially using RBF kernel which yielded an overall diagnostic accuracy higher than the linear SVM in C-SVM and PCA-SVM models. This reveals a fact that the optimal separating hyperplane of serum SERS is nonlinear. Hu and Lin find that RBF kernel is the most reasonable choice in SVM due to its simplicity and ability to models data of arbitrary complexity.[33] In fact, the linear kernel is a special case of nonlinear kernel. Our study and other reports confirm this conclusion.[20]

It is reported that the accuracy of endoscopic ultrasonography in the determination of the T stage of esophageal cancer is approximately 65% to 90%, with an average sensitivity of 75% and specificity of 70%.[1] In this study, we obtain a sensitivity of 83.3% and specificity of 86.7% by classifying serum SERS spectra from normal volunteers and esophageal cancer patients with RBF PCA-SVM technique. However, there are some deficiencies for this preliminary study. For example, the data set are too small, needing to expand the sample numbers. The research of the variability of sample is absent. Maybe the SERS spectra of a patient's serum have been altered at different times. These factors need to be further studied in our future work.

5 Conclusion

In conclusion, the C-SVM methods and PCA-SVM methods are successfully implemented for the classification of serum SERS spectra from normal volunteers and esophageal cancer patients. A number of effective diagnostic models based on C-SVM and PCA-SVM techniques with different kernel functions are developed and the diagnostic performances are comprehensively evaluated and compared. The PCA-SVM methods can considerably simplify the complexity of calculation without sacrificing the performance of the algorithm. The RBF PCA-SVM algorithm is superior to PCA-LDA algorithm in classification serum SERS spectra. Serum SERS combining with SVM has great potential to provide an effective and accurate diagnostic means for noninvasive esophageal cancer detection.

Acknowledgments

The authors would like to acknowledge the financial support of the Dongguan Science and Technology Project (201010815207), the National Natural Science Foundation of China (61275187, 60778047), Specialized Research Fund for the Doctoral Program of Higher Education of China (20114407110001), the Natural Science Foundation of Guangdong Province (9251063101000009) and the Cooperation Project in Industry, Education and Research of Guangdong province and Ministry of Education of China (2011A090200011).

References

1. M. R. Wong, "Esophageal cancer: a systematic review," *Curr. Probl. Cancer* **24**(6), 298–373 (2000).
2. M. S. Bergholt et al., "*In vivo* diagnosis of esophagus cancer using image-guided raman endoscopy and biomolecular modeling," *Technol. Cancer Res. Treat.* **10**(2), 103–112 (2011).
3. E. B. Hanlon et al., "Prospects for *in vivo* Raman spectroscopy," *Phys. Med. Biol.* **45**(2), 1–59 (2000).
4. M. I. C. Kendall et al. "Vibrational spectroscopy: a clinical tool for cancer diagnostics," *Analyst* **134**(6), 1029–1045 (2009).
5. Y. Hu et al., "Classification of normal and malignant human gastric mucosa tissue with confocal Raman microspectroscopy and wavelet analysis," *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **69**(2), 378–382 (2008).
6. S. Duraipandian et al., "*In vivo* diagnosis of cervical precancer using Raman spectroscopy and genetic algorithm techniques," *Analyst* **136**(20), 4328–4336 (2011).
7. Z. Huang et al., "Near-infrared Raman spectroscopy for optical diagnosis of lung cancer," *Int. J. Cancer* **107**(6), 1047–1052 (2003).
8. A. Beljebbar et al., "Identification of Raman spectroscopic markers for the characterization of normal and adenocarcinomatous colonic tissues," *Crit. Rev. Oncol. Hematol.* **72**(3), 255–264 (2009).
9. D. Cialla et al., "Surface-enhanced Raman spectroscopy (SERS): progress and trends," *Anal. Bioanal. Chem.* **403**(1), 27–54 (2012).
10. S. Abalde-Cela et al., "Surface-enhanced Raman scattering biomedical applications of plasmonic colloidal particles," *J. R. Soc. Interface* **7**(Suppl. 4), S435–S450 (2010).
11. X. Qian et al., "*In vivo* tumor targeting and spectroscopic detection with surface-enhanced Raman nanoparticle tags," *Nat. Biotechnol.* **26**(1), 83–90 (2007).
12. C. T. Nguyen et al., "Detection of chronic lymphocytic leukemia cell surface markers using surface enhanced Raman scattering gold nanoparticles," *Cancer Lett.* **292**(1), 91–97 (2010).
13. S. Feng et al., "Nasopharyngeal cancer detection based on blood plasma surface-enhanced Raman spectroscopy and multivariate analysis," *Biosens. Bioelectron.* **25**(11), 2414–2419 (2010).
14. J. Lin et al., "A novel blood plasma analysis technique combining membrane electrophoresis with silver nanoparticle-based SERS spectroscopy for potential applications in noninvasive cancer detection," *Nanomed. Nanotechnol. Biol. Med.* **7**(5), 655–663 (2011).
15. S. Feng et al., "Gastric cancer detection based on blood plasma surface-enhanced Raman spectroscopy excited by polarized laser light," *Biosens. Bioelectron.* **26**(7), 3167–3174 (2011).
16. M. Gniadecka et al., "Diagnosis of basal cell carcinoma by Raman spectroscopy," *J. Raman Spectrosc.* **28**(2–3), 125–129 (1997).
17. S. K. Majumder, N. Ghosh, and P. K. Gupta, "Support vector machine for optical diagnosis of cancer," *J. Biomed. Opt.* **10**(2), 024034 (2005).
18. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.* **2**(2), 121–167 (1998).
19. E. Widjaja, W. Zheng, and Z. Huang, "Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines," *Int. J. Oncol.* **32**(3), 653–662 (2008).
20. W. Lin et al., "Classification of *in vivo* autofluorescence spectra using support vector machines," *J. Biomed. Opt.* **9**(1), 180–186 (2004).
21. Y. F. W. Ren and E. Wang, "A binary functional substrate for enrichment and ultrasensitive SERS spectroscopic detection of folic acid using graphene oxide/Ag nanoparticle hybrids," *ACS Nano* **5**(8), 6425–6433 (2011).
22. H. L. J. Zhao, D. I. Mclean, and H. Zeng, "Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy," *Appl. Spectros.* **61**(11), 1225–1232 (2007).
23. L. V. Ganyun et al., "Fault diagnosis of power transformer based on multilayer SVM classifier," *Electric Power Syst. Res.* **74**(1), 1–7 (2005).
24. M. A. F. Abdat, Y. Guermeur, and W. Blondel, "Hybrid feature selection and SVM-based classification for mouse skin precancerous stages diagnosis from bimodal spectroscopy," *Opt. Express* **20**(1), 228–244 (2012).
25. R. F. E. Osuna and F. Girosi, "Training support vector machines: an application to face detection," in *Proc. CVPR'97*, Puerto Rico, IEEE Computer Society, Washington, DC (1997).
26. J. W. Chan et al., "Micro-Raman spectroscopy detects individual neoplastic and normal hematopoietic cells," *Biophys. J.* **90**(2), 648–656 (2006).
27. N. Stone et al., "Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers," *J. Raman Spectrosc.* **33**(7), 564–573 (2002).

28. K. K. Maiti et al., “Development of biocompatible SERS nanotag with increased stability by chemisorption of reporter molecule for *in vivo* cancer detection,” *Biosens. Bioelectron.* **26**(2), 398–403 (2010).
29. L. Jiang et al., “Raman reporter-coated gold nanorods and their applications in multimodal optical imaging of cancer cells,” *Anal. Bioanal. Chem.* **400**(9), 2793–2800 (2011).
30. M. S. Bergholt et al., “*In vivo* diagnosis of gastric cancer using Raman endoscopy and ant colony optimization techniques,” *Int. J. Cancer* **128**(11), 2673–2680 (2011).
31. D. E. Maziak et al., “Fourier-transform infrared spectroscopic study of characteristic molecular structure in cancer cells of esophagus: an exploratory stud,” *Cancer Detect. Prevent.* **31**(3), 244–253 (2007).
32. M. Sattlecker et al., “Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics,” *Analyst* **135**(5), 895–901 (2010).
33. C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE Trans. Neural Netw.* **13**(2), 415–425 (2002).