

Classification of *in vivo* autofluorescence spectra using support vector machines

WuMei Lin

Xin Yuan

Hong Kong University of Science & Technology
Department of Electrical & Electronic Engineering
Clear Water Bay
Kowloon
Hong Kong
China

Powing Yuen

William I. Wei

University of Hong Kong
Queen Mary Hospital
Division of Otorhinolaryngology
Hong Kong

Jonathan Sham

University of Hong Kong
Department of Clinical Oncology
Hong Kong

PengCheng Shi

Jianan Qu

Hong Kong University of Science & Technology
Department of Electrical & Electronic Engineering
Clear Water Bay
Kowloon
Hong Kong
China
E-mail: eequ@ust.hk

Abstract. An algorithm based on support vector machines (SVM), the most recent advance in pattern recognition, is presented for use in classifying light-induced autofluorescence collected from cancerous and normal tissues. The *in vivo* autofluorescence spectra used for development and evaluation of SVM diagnostic algorithms were measured from 85 nasopharyngeal carcinoma (NPC) lesions and 131 normal tissue sites from 59 subjects during routine nasal endoscopy. Leave-one-out cross-validation was used to evaluate the performance of the algorithms. An overall diagnostic accuracy of 96%, a sensitivity of 94%, and a specificity of 97% for discriminating nasopharyngeal carcinomas from normal tissues were achieved using a linear SVM algorithm. A diagnostic accuracy of 98%, a sensitivity of 95%, and a specificity of 99% for detecting NPC were achieved with a nonlinear SVM algorithm. In a comparison with previously developed algorithms using the same dataset and the principal component analysis (PCA) technique, the SVM algorithms produced better diagnostic accuracy in all instances. In addition, we investigated a method combining PCA and SVM techniques for reducing the complexity of the SVM algorithms. © 2004 Society of Photo-Optical Instrumentation Engineers.
[DOI: 10.1117/1.1628244]

Keywords: fluorescence spectroscopy; support vector machines; tissue diagnosis.

Paper 03006 received Jan. 22, 2003; revised manuscript received May 19, 2003 and Jul. 10, 2003; accepted for publication Jul. 15, 2003.

1 Introduction

Cancer is a leading cause of death worldwide. Over the past two decades, the promise that optical techniques may offer new medical diagnostic tools for detecting cancerous tissues has stimulated a great deal of research. Autofluorescence spectroscopy is one of the primary research areas in which rapid progress has been achieved. It has been demonstrated that technologies utilizing the information carried in the endogenous fluorescence of tissue (autofluorescence) can be particularly suitable for noninvasive diagnosis of malignant and preclinical lesions because the characteristics of autofluorescence are directly linked to tissue biochemistry and architecture.^{1–3} The key step in spectroscopic diagnosis of tissue is to build a robust algorithm that extracts characteristic features from autofluorescence spectral signals and correlates these features with tissue pathology. The autofluorescence spectral signal is a multivariate function of wavelengths. To minimize the loss of clinically useful information in autofluorescence signals, a multivariate statistical method should be used to recognize the spectral characteristics for the development of diagnostic algorithms.

A variety of multivariate statistical algorithms have been successfully utilized in the spectroscopic diagnosis of tissue. For example, O'Brien et al.⁴ used principal component analysis (PCA) and multivariate linear regression (MLR) methods for guided angioplasty using fluorescence spectroscopy. Ramanujam et al.⁵ developed PCA algorithms for improving the

diagnosis capability of fluorescence spectroscopy in the detection of cervical intraepithelial neoplasia. Wang et al.⁶ and Eker et al.⁷ used the partial least-squares (PLS) method to build algorithms for classifying the autofluorescence spectral signals recorded from normal and cancerous tissues in the oral cavity and larynx. In nonlinear approaches, Tumer et al.⁸ reported on a classifier using radial basis function networks to diagnose early cervical lesions. Rovithakis et al.⁹ used artificial neural networks to analyze autofluorescence and to discriminate pathological from normal peripheral vascular tissue.

In previous studies, we investigated the *in vivo* autofluorescence of nasopharyngeal carcinoma (NPC) and normal tissue.^{10,11} NPC is common among Southeast Asians and may occur at any age. Successful treatment is possible when NPC is detected in its early stages. White-light endoscopy is the currently available detection method, but it produces poor diagnostic accuracy for flat or small lesions and the identification of tumor margins in advanced stages of NPC. The common practice for diagnosing subclinical tumors in the high-risk group is through random endoscopic biopsy. However, only 5.4% of patients from the high-risk group have been shown to have asymptomatic NPC in random biopsies of the nasopharynx. The results of our investigation have demonstrated that *in vivo* autofluorescence spectra of NPC and normal nasopharyngeal tissues are different and that the difference can be used to discriminate NPC from normal tissue.^{10,11}

Simple algorithms using the autofluorescence signals in a few wavelength bands and algorithms based on the PCA method have been developed for diagnosing NPC.

The goal of this study is to explore a new technique for developing diagnostic algorithms with improved accuracy for classifying autofluorescence from NPC and normal tissue *in vivo*. Specifically, we present the development and evaluation of new classifiers for autofluorescence using linear and nonlinear support vector machine (SVM) algorithms, the most recent advance in the field of pattern recognition. The possibility of building a simple algorithm and improving its diagnostic performance with a combination of PCA and SVM methods is investigated.

2 Materials and Methods

2.1 Support Vector Machine Algorithms

A support vector machine is a new method for classifying multivariate data. It was first proposed by Vapnik^{12,13} and successfully extended by a number of other researchers in recent years.^{14–16} It is based on the principle of minimization of structural risk in constructing an optimally separating hyperplane that separates different classes of data. When the separating boundary is nonlinear, SVM maps the sample data with specific kernel functions to a higher dimensional feature space to linearize the boundary and generate the optimal separating hyperplane. Compared with other multivariate statistical methods, SVM was designed to be particularly effective in developing a reliable classifier from a training set with a small sample size.¹³ In addition, no assumption about the statistical properties of the classified data is made when developing the SVM classification algorithm.

SVM has become a rapidly emerging technique in the classification of data in the past 5 years. In particular, the applications of SVM techniques have been recently extended to the processing of biological and medical data. For example, SVM has been successfully applied in cellular protein studies and gene selection for cancer classification,^{17–19} image processing of digital mammograms, and computer-aided detection of lesions in computer tomography (CT),^{20,21} detection of cancer using microarray expression data and signatures,^{22–24} and the detection and screening of diseased tissue based on feature screening.^{25,26}

The detailed theory of the SVM method is described in Refs. 12–16. Briefly, SVM maps the sample data linearly or nonlinearly to a high-dimensional feature space. The hyperplane optimally separating two classes of data is given as an expansion on a small number of sample data in the training set known as support vectors that are always the closest to the optimal hyperplane. The support vectors correspond to the training samples that are most difficult to classify. Mathematically, a hyperplane is defined by $w \times x + b = 0$ in the feature space of sample data, where w is the norm to the hyperplane and b is a plane constant. Given a set of labeled training samples $\{x_i, y_i\}$, $i = 1, \dots, l$, $y_i \in \{-1, +1\}$, $x_i \in R^n$. Here l and n are the number and dimension of the sample vectors, respectively. Suppose that a hyperplane can separate the positive samples ($y_i = +1$) from the negative samples ($y_i = -1$) in the feature space of sample data; then $y_i(w \times x_i + b) - 1 \geq 0$. The margin of a separating hyperplane, defined

as the sum of the shortest distances from the positive and negative samples to the hyperplane, equals $2/\|w\|$. The SVM algorithm is simply looking for an optimal hyperplane with the maximal separating margin by

$$\text{minimizing } \frac{\|w\|^2}{2}, \text{ subject to } y_i(w \times x_i + b) - 1 \geq 0. \quad (1)$$

When two classes of training samples are not linearly separable, the constraint is relaxed to allow misclassification: $y_i(w \times x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$. The optimization problem becomes

$$\text{minimizing } \frac{\|w\|^2}{2} + C \times \sum \xi_i, \\ \text{subject to } y_i(w \times x_i + b) \geq 1 - \xi_i, \quad (2)$$

or

$$\text{maximizing } \sum \alpha_i - \frac{1}{2} \sum \alpha_j y_i y_j (x_i, x_j), \\ \text{subject to } \sum \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, \quad (3)$$

where C is a parameter chosen by the user to define the cost of constraint relaxation. A larger C corresponds to a higher penalty for the assigned errors. Each α_i is a Lagrange multiplier corresponding to a sample (x_i, y_i) in the training set. The optimization process determines support vectors that are the training sample vectors with nonzero Lagrange multipliers. A linear SVM classifier can then be constructed using these support vectors as

$$f(x) = \text{sgn}(w^T x + b) \quad (4)$$

or

$$f(x) = \text{sgn} \left[\sum_{i=1}^m \alpha_i y_i (x_i \times x) + b \right], \quad \alpha_i > 0, \quad (5)$$

where m is the number of support vectors and x is the sample to be classified.

If the separating boundary is nonlinear in the feature space of the data sample and the linear SVM classifier cannot separate the training data well, improved classification results may be obtained using a nonlinear SVM method. Assuming the nonlinear separating boundary can be linearized in a higher dimensional feature space using a mapping process: $\Phi: R^n \rightarrow H: R^{n+v} \Rightarrow x \rightarrow \Phi(x)$, a linear SVM classifier can be constructed in the new feature space, H . Here, Φ is the mapping function and v is the increased dimension of H space. The inner product of two mapped vectors in H space can be efficiently computed using kernel functions $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ in the feature space of the sample data. Note that only the inner products between the sample data in the training set are needed to calculate the Lagrange multipliers and to identify the support vectors. Also, only the inner products between the support vectors and the sample to be classified are needed in the SVM classifier. Thus it is not necessary to use the explicit mapping function to identify support vectors and form the classifier in the new feature space of in-

creased dimensions. The power of the nonlinear SVM method is that with a known kernel function, the classification of nonlinearly separable samples can be simply performed in the feature space of the sample data.

The nonlinear SVM classifier can be constructed as

$$f(x) = \text{sgn} \left[\sum_{i=1}^m \alpha_i y_i K(x_i, x) + b \right], \quad \alpha_i > 0. \quad (6)$$

The kernel functions used in constructing nonlinear SVM classifiers are the polynomial function, the radial basic function (RBF), and the neural network function, etc.^{13–15} One of the most used kernel functions in the reported work is the RBF kernel defined as $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$. In this work, the RBF kernel was used to construct nonlinear SVM algorithms to classify the *in vivo* autofluorescence of normal nasopharyngeal tissue and nasopharyngeal carcinoma.

2.2 In Vivo Autofluorescence Spectra

The autofluorescence spectral data used in this study are the same as these used for the development of multiple-wavelength band ratio algorithms and PCA algorithms in our previous work.¹¹ A total of 216 autofluorescence spectral signals were measured *in vivo* from 85 nasopharyngeal carcinoma lesions and 131 normal tissue sites of 59 subjects during routine nasal endoscopy. The instrumentation of the measurement system was described in detail in a previous paper.¹⁰ The excitation light was a mercury arc lamp filtered with a band-pass filter with a bandwidth from 390 to 450 nm. The autofluorescence was excited primarily by the strong spectral lines of the mercury at 404.66 and 435.84 nm, which share the same upper initial energy level, 1S_0 . The relative intensities between the two strong spectral lines then remain constant because the intensities are determined by their relative transition probabilities and are independent of the operation conditions of the light source.

The autofluorescence and reflection of the excitation light signals were collected from examined tissue by a Karl Storz nasal endoscope and separated by a dichotic mirror with a cut-on wavelength at 470 nm. The fluorescence signals collected by the endoscope were conducted to a multichannel spectrometer with seven optical fibers that were evenly distributed in the image plane of the endoscope. The spectrally dispersed autofluorescence signals were recorded using an ICCD camera in the wavelength range from 470 to 680 nm, with an interval of 0.37 nm. In order to reduce the dimensions of raw data, the spectra were smoothed using moving-window filtering and the wavelength resolution was reduced to 1 nm. The final dimension of each spectral signal before further processing was 211.

The autofluorescence signals were collected from endoscopically normal and abnormal sites for each subject. Two autofluorescence spectra were recorded from each site in two different angles because the characteristics (intensity and line shape) of autofluorescence signals are affected by the illumination and collection geometry. Normally, a total of four spectra were collected from one subject except for a few cases in which it was difficult to maneuver the endoscope to measure fluorescence signals from the same tissue site at different angles. Biopsy specimens were taken from tissue sites from

which the autofluorescence signals were collected. The histological examinations of the biopsy samples, which served as a gold standard for labeling normal tissue with -1 and for labeling cancerous lesions with $+1$, were performed by experienced pathologists. The *in vivo* autofluorescence spectroscopy measurements were conducted in the Department of Otorhinolaryngology at Queen Mary Hospital, The University of Hong Kong. This study was approved by the university's ethics committee.

3 Results and Discussion

Typical raw autofluorescence spectra collected from nasopharyngeal carcinomas and normal tissues *in vivo* are shown in Figs. 1(a) and 1(b). As can be seen, the intensities of the raw signals vary across a wide range because of variations in measurement conditions (including illumination power, separation of endoscope from tissue, and incident or emission angles over the imaged tissue surface) from individual to individual and from measuring site to site for each individual. To eliminate the variations caused by the measurement conditions, each spectral signal was normalized to its area before further processing. The normalized spectra from nasopharyngeal carcinomas and normal tissues are shown in Figs. 1(c) and 1(d).

In the SVM learning and testing procedure, each spectrum was treated as a vector with 211 dimensions and labeled according to the result of the histological examination. Specifically, 131 normal samples were labeled as -1 and 85 cancerous samples were labeled as $+1$. SVM^{light} (version 5.00), an implementation of support vector machines in C language,²⁷ was used to process the normalized autofluorescence data. Diagnostic algorithms based on linear and nonlinear SVMs were developed. The performance of the SVM algorithms was evaluated with the leave-one-case-out cross-validation. In the cross-validation procedure, spectral samples recorded from one subject were held out from the whole dataset and the remaining samples were treated as a training set for developing the SVM diagnostic algorithms to classify the withheld autofluorescence spectra. The linear SVM algorithms developed from the training set were optimized by exhaustively searching the optimal value of parameter C . The optimization criterion was to minimize the number of misclassified samples or maximize the classification accuracy in the training set. In the exhaustive searching process, the initial value of C was set to the default C provided by SVM^{light} that was calculated with $[Average(x, x)]^{-1}$. Here, x consists of the sample vectors in the training set. Thus the initial value of parameter C was related to the statistical property of the training dataset. In the nonlinear approach, an RBF kernel was used to construct the diagnostic algorithm. Tuning the value of parameter γ of the RBF kernel, the optimal value of parameter C was determined by exhaustive searching with each selected value of γ . The criterion for optimizing the algorithms in the training set was the same as that used in linear SVM method.

The optimized algorithms developed from the training set were used to classify the withheld spectral samples measured from one subject. This procedure was repeated until all 216 samples collected from 59 subjects were classified. It should be noted that new algorithms were constructed using the remaining spectra recorded from 58 subjects as a training set in

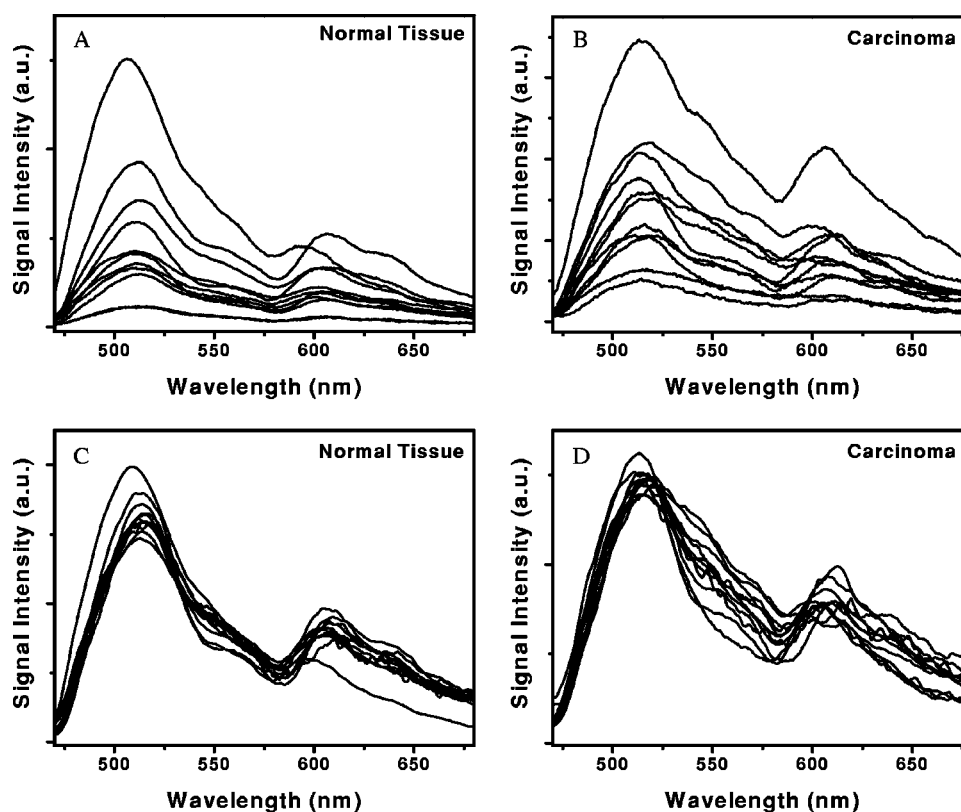


Fig. 1 Autofluorescence spectra of nasopharyngeal carcinoma (NPC) and normal tissue. (a) Raw spectra from normal tissue. (b) Raw spectra from NPC. (c) Normalized spectra from normal tissue. (d) Normalized spectra from NPC.

every round of cross-validation. The overall classification accuracy, sensitivity, and specificity of a particular algorithm were calculated based on the classifications of the withheld samples over 59 rounds of cross-validation.

The results of the classifications of all the *in vivo* autofluorescence signals using linear and nonlinear SVM algorithms are summarized in Table 1. As a reference, the best results in classifying the same dataset using the PCA algorithm reported in Ref. 11 are also included in the table. The results demonstrate that both the linear SVM algorithm and nonlinear SVM algorithms performed better than the PCA algorithm. The linear SVM and RBF SVM algorithms were constructed with 38 and 70 sample vectors (or support vectors), respectively. Figure 2 shows how the classification accuracy of a linear SVM algorithm depends on the parameter C in a wide-range search for optimal C . The algorithm was optimized by choosing the

Table 1 Results of classification of autofluorescence with different algorithms.

Algorithm	Sensitivity (%)	Specificity (%)	Accuracy (%)
Linear SVM	94	97	96
Nonlinear SVM (RBF)	95	99	98
PCA (Ref. 11)	93	95	94

value of C that maximized classification accuracy for the leave-one-case-out cross-validation.

In the development of nonlinear SVM algorithms, an optimal C that maximizes classification accuracy can be determined for each selected value of the RBF parameter γ . It was found that the classification accuracy was not sensitive to parameter γ . In a wide range of values for C and γ , many sets of C and γ could be found to yield the same classification accuracy. Figure 3 shows the optimal sets of C and γ that pro-

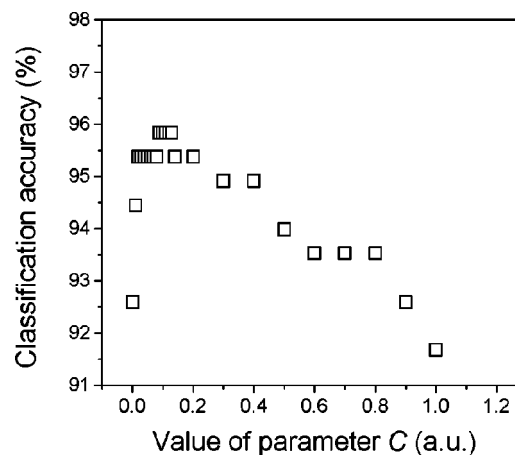


Fig. 2 Dependence of classification accuracy on parameter C for a linear SVM algorithm.

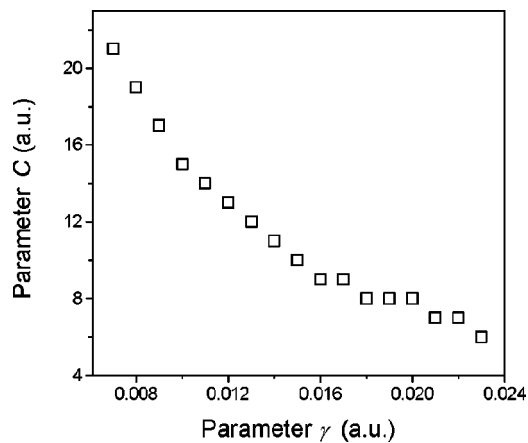


Fig. 3 Optimal sets of parameters C and γ obtained in the development of a nonlinear SVM algorithm using an radial basic function (RBF) kernel. The C and γ sets shown in the figure produced the same and maximal accuracy.

duced the maximal classification accuracy in a training set. For both linear and nonlinear SVM algorithms, the classifiers generated from the training sets over 59 rounds of cross-validations were stable and almost the same. It should be noted that optimal parameters (C and γ) determined by the procedure of leave-one-case-out cross-validation may result in an optimistic bias.

As can be seen, the nonlinear SVM produced a sensitivity and specificity higher than the linear SVM. This indicates that the boundary separating autofluorescence from nasopharyngeal carcinoma and that from normal tissue is not linear. The spectral signals of some cancerous tissues cannot be clearly separated from those of normal tissues in the feature space of the spectral dataset. The linearization of the separating boundary by mapping the original spectral data to a high dimensional space in the nonlinear SVM algorithms improves the diagnostic accuracy.

The subjects enrolled in this study were all clinically suspicious patients exhibiting a nasopharyngeal abnormality endoscopically. Endoscopic diagnosis of nasopharyngeal carcinoma is based on the fact that most of tumors show marked racial and geometric variation. The biopsy must be taken from an endoscopically identified tumorous tissue site for histological analysis. However, this criterion generates a significant number of false positives, especially when the lesion is relatively flat. A clinical investigation showed that the sensitivity and false positive rate of endoscopy with a single biopsy for detecting nasopharyngeal carcinoma was about 85 and 15%, respectively.²⁸ Here, the false positive rate is defined as the ratio of the number of false positives over the total number of endoscopically identified tumorous tissue sites. In this study it was found that 16 biopsies taken from 59 subjects showed false positives. The false positive rate was about 27%. In contrast to endoscopic diagnosis, only two false positives were produced by the autofluorescence spectroscopy method. The sensitivity of the fluorescence method with the SVM algorithm was up to 95%. Thus autofluorescence spectroscopy combined with conventional nasal endoscopy can substantially improve diagnostic accuracy.

Finally, it also should be pointed out that the lesions studied in this work were all invasive. Unlike other well-investigated organ sites, such as the lung, colon, and cervix, dysplasia and carcinoma *in situ* in the nasopharynx are rarely reported.^{29–32} Over a period of 10 months of collection of *in vivo* autofluorescence spectra, we did not find a lesion of nasopharyngeal dysplasia and carcinoma *in situ*.

The dimension of spectral data is generally high. For instance, the autofluorescence spectrum used in this study has 211 dimensions. The high dimension of data space may cause complexity in optimization and implementation of the SVM algorithm because computation of all the inner products between the sample and support vectors in a high-dimensional feature space is complicated and time-consuming. This limits the use of the SVM method in applications that require fast or even real-time data processing. We investigated the possibility of simplifying the implementation of the SVM algorithm and improving its performance by reducing the dimensions of the spectral data using the PCA method.

Principal component analysis is a mathematical tool that reduces the dimensions of a dataset to a set of informative principal components (PCs) that account for most of the variance of the original dataset. The first PC accounts for as much of the variability in the dataset as possible, and each succeeding component accounts for as much of the residual variability as possible. Therefore the PCs are normally arranged in the order of their contributions to the variance of entire dataset. In principle, PCA is an operation that rotates the coordinates of the original data to form new coordinates using the PCs. Most of the information carried in the data is distributed in the first few PCs of the new coordinates. The contributions from the rest of the PC coordinates are negligible. By presenting the original data in new PC coordinates formed with a few informative PCs, the dimensions of the data can be significantly reduced without losing important information. The implementation of SVM algorithms in a data space with a much lower dimension should be more efficient and consume less time.

When PCA is used to process autofluorescence spectra, it transforms wavelengths, the original spectral variables, into a set of PC spectra. Each original spectrum is a combination of the PC loading spectra that are orthogonal to each other. The PCs with negligible contributions to the variance of the dataset are eliminated. The dimensions of the dataset for developing the diagnostic algorithm can then be significantly reduced without losing useful information. Figure 4 shows the contribution of each PC to the variance of the 216 autofluorescence spectra. The PCs were calculated with a MATLAB-based PCA program. As shown in the figure, the first two PCs account for 86.4% of the total variance; the first five PCs account for 97%; the first ten PCs account for 98.5%; and the first twenty PCs account for 99.5%.

Linear and nonlinear SVM algorithms were developed using the projection scores of the autofluorescence spectra on PC loadings. The performance of the linear SVM algorithms with different numbers of PCs was compared with the results obtained from the spectral data in wavelength space shown in Table 1. We found that the linear SVM algorithm with the first two PCs produced a sensitivity of 94% and a specificity of 97% in differentiating NPC from normal tissue. This accuracy is identical to that achieved by the linear SVM algorithm in the wavelength space of 211 dimensions. The performance of

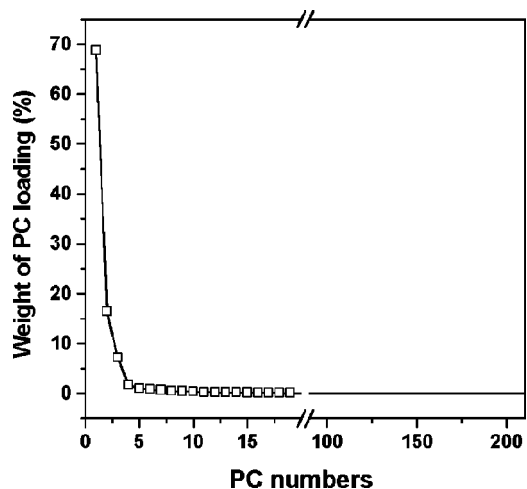


Fig. 4 Contributions of principal components to the total variance of 216 spectral data.

the linear SVM algorithm was not improved by adding more PCs. This demonstrates that the first two PCs captured all the information necessary for classifying the linearly separable samples. It indicates that linear SVM algorithms developed in two-dimensional PC space are equivalent to those developed from the autofluorescence spectral data. The complexity of the linear SVM algorithm using two PC scores was tremendously reduced because the PCA method decreased the dimension of the data from 211 to 2. Also, the combined PCA and linear SVM methods produced a classification accuracy that was better than that of the PCA algorithms reported in our previous study.¹¹

The results of implementing nonlinear SVM with PC scores are summarized in Table 2. The results with the first two PCs are identical to those of the linear SVM algorithm. The classification accuracy begins to increase by adding up to six PCs. The maximal accuracy is achieved when the first nine PCs are used. This result is identical with the best result obtained from the nonlinear SVM algorithm using the spectral data in the wavelength space of 211 dimensions. The performance of the nonlinear SVM algorithm was not improved further by adding more than nine PCs. This indicates that the first two PCs contribute to the classification of linearly separable samples, and the PCs from six to nine capture the major information needed to classify the nonlinearly separable samples. With the combined PCA and nonlinear SVM meth-

Table 2 Performance of nonlinear SVM algorithms using principal component scores.

Number of Principal Components	Number of Support Vectors	Sensitivity (%)	Specificity (%)
2	86	94	97
6	38	94	98
9	52	95	99
210	60	95	99

ods, the dimensions of the data were reduced from 211 to 9. It is important to note that the numbers of support vectors used in the SVM algorithms with six and nine PCs are also less than those of SVM algorithms using spectral data in wavelength space. This demonstrates that the PCA method can substantially simplify SVM algorithm without sacrificing diagnostic accuracy.

4 Conclusions

Linear and nonlinear SVM methods were successfully implemented for the classification of autofluorescence signals from nasopharyngeal carcinomas and normal tissues. The results demonstrate that autofluorescence spectroscopy with an SVM algorithm can achieve high diagnostic accuracy in differentiating nasopharyngeal carcinoma from normal tissue. Compared with the previously developed algorithms using the same dataset and the principal component analysis technique,¹¹ the SVM algorithms produced better diagnostic accuracy in all instances. In principle, a statistical comparison is desirable to evaluate the performance of different algorithms. However, such an analysis requires an accurate estimate of the statistical properties of a dataset to draw a reliable conclusion. This cannot be done when the size of the dataset is small, such as that used in this study.

Combined PCA and SVM methods were investigated. It was found that SVM and combined PCA and SVM algorithms produced the same diagnostic accuracy. PCA can substantially reduce the complexity of an SVM algorithm without sacrificing the performance of the algorithm. The simplification of the algorithm is particularly important for applications that require rapid processing of a large amount of multivariate data, as in real-time multispectral imaging and optical processing systems.^{33,34} It is an interesting finding that SVM can help to identify the PCs that carry the information headed for classifying linearly or nonlinearly separable samples, which provide important statistical properties of multivariate data samples. Finally, it should be noted that using leave-one-out cross-validation is common for small datasets, but independent training-validation and test sets are more desirable for a study with a large volume of samples.

Acknowledgment

The authors acknowledge support from the Hong Kong Research Grants Council through grants HKUST6052/00M and HKUST6025/02M.

References

1. R. Richards-Kortum and E. Sevick-Muraca, "Quantitative optical spectroscopy for tissue diagnosis," *Annu. Rev. Phys. Chem.* **47**, 555–606 (1996).
2. G. A. Wagnieres, W. M. Star, and B. C. Wilson, "In vivo fluorescence spectroscopy and imaging for oncological applications," *Photochem. Photobiol.* **68**, 603–632 (1998).
3. N. Ramanujam, "Fluorescence spectroscopy of neoplastic and non-neoplastic tissues," *Neoplasia* **2**, 89–117 (2000).
4. K. M. O'Brien, A. F. Gmitro, G. R. Gindi, M. L. Stetz, F. W. Cutruzzola, L. I. Laifer, and L. L. Deckelbarm, "Development and evaluation of classification algorithms for fluorescence guided laser angioplasty," *IEEE Trans. Biomed. Eng.* **36**, 424–431 (1989).
5. N. Ramanujam, M. F. Mitchell, A. Mahadevan, S. Thomsen, A. Malpica, T. Wright, N. Atkinson, and R. Richards-Kortum, "Development of a multivariate statistical algorithm to analyze human cervical tissue," *Lasers Surg. Med.* **19**, 46–62 (1996).

6. C. Y. Wang, C. T. Chen, C. P. Chiang, S. T. Young, S. N. Chow, and H. K. A. Chiang, "Partial least-squares discriminant analysis on autofluorescence spectra of oral carcinogenesis," *Appl. Spectrosc.* **52**, 1190–1196 (1998).
7. C. Eker, R. Rydell, K. Svanberg, and S. Andersson-Engels, "Multivariate analysis of laryngeal fluorescence spectra recorded *in vivo*," *Lasers Surg. Med.* **28**, 259–266 (2001).
8. K. Tumer, N. Ramanujam, J. Ghosh, and R. Richards-Kortum, "Ensembles of radial basis function networks for spectroscopic detection of cervical precancer," *IEEE Trans. Biomed. Eng.* **45**, 953–961 (1998).
9. G. A. Rovithakis, M. Maniadakis, M. Zervakis, G. Filippidis, G. Zacharakis, A. N. Katsamouris, and T. G. Papazoglou, "Artificial neural networks for discriminating pathologic from normal peripheral vascular tissue," *IEEE Trans. Biomed. Eng.* **48**, 1088–1097 (2001).
10. J. Y. Qu, P. W. Yuen, Z. J. Huang, D. Kwong, J. Shan, S. J. Lee, W. K. Ho, and W. I. Wei, "Preliminary study of *in vivo* autofluorescence of nasopharyngeal carcinoma and normal tissue," *Lasers Surg. Med.* **26**, 432–440 (2000).
11. H. P. Chang, J. Y. Qu, P. W. Yuen, J. Sham, D. Kwong, and W. I. Wei, "Light-induced autofluorescence spectroscopy for detection of nasopharyngeal carcinoma *in vivo*," *Appl. Spectrosc.* **56**, 1361–1367 (2002).
12. V. Vapnik, *Estimation of Dependencies Based on Empirical Data* Springer-Verlag, New York (1982).
13. V. Vapnik, *The Nature of Statistical Learning Theory*, Chap. 4, Springer-Verlag, New York (1995, 2000).
14. C. Burges, "A tutorial on support vector machine for pattern recognition," *Data Min. Knowl. Discov.* **2**, 121–167 (1998).
15. T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*, MIT Press, Cambridge, MA (1999).
16. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Advances in Neural Information Processing Systems (NIPS)* **13**, pp. 668–674 (2001).
17. S. J. Hua and Z. R. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics* **17**, 721–728 (2001).
18. Z. Yuan, K. Burrage, and J. S. Mattick, "Prediction of protein solvent accessibility using support vector machines," *Proteins* **48**, 566–570 (2002).
19. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Behav. Ecol. Sociobiol.* **46**, 389–422 (2002).
20. A. Bazzani, A. Bevilacqua, D. Bollini, R. Brancaccio, R. Campanini, N. Lanconelli, A. Riccardi, and D. Romani, "An SVM classifier to separate false signals from microcalcifications in digital mammograms," *Phys. Med. Biol.* **46**, 1651–1663 (2001).
21. S. B. Gokturk, C. Tomasi, B. Acar, C. F. Beaulieu, D. S. Paik, R. B. Jeffrey, J. Yee, and S. Napel, "A statistical 3-D pattern processing method for computer-aided detection of polyps in CT colonography," *IEEE Trans. Med. Imaging* **20**, 1251–1260 (2001).
22. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics* **16**, 906–914 (2000).
23. S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Natl. Acad. Sci. U.S.A.* **98**, 15149–15154 (2001).
24. B. Fritz, B. F. Schubert, G. Wrobel, C. Schwaenen, S. Wessendorf, M. Nessler, C. Korz, R. J. Rieker, K. Montgomery, R. Kucherlapati, G. Mechttersheimer, R. Eils, S. Joos, and P. Lichter, "Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma," *Cancer Res.* **62**, 2993–2998 (2002).
25. S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *J. Biomed. Inform.* **34**, 28–36 (2001).
26. K. L. Chan, T. W. Lee, P. Sample, M. H. Goldbaum, R. N. Weinreb, and A. T. J. Sejnowski, "Comparison of machine learning and traditional classifiers in glaucoma diagnosis," *IEEE Trans. Biomed. Eng.* **49**, 963–974 (2002).
27. T. Joachims, "SVM^{light} Support Vector Machine," <http://svmlight.joachims.org/>.
28. D. Kwong, J. Nicholls, W. I. Wei, D. Chua, J. Sham, P. W. Yuen, A. Cheng, C. C. Yau, P. Kwong, and D. Choy, "Correlation of endoscopic and histologic findings before and after treatment for nasopharyngeal carcinoma," *Head Neck* **23**, 34–41 (2001).
29. C. W. Chan, J. M. Nicholls, J. S. Sham, P. Dickens, and D. Choy, "Nasopharyngeal carcinoma *in situ* in nasopharyngeal carcinoma," *J. Clin. Pathol.* **45**, 898–901 (1992).
30. J. Nicholls, J. J. Sham, M. H. Ng, and D. Choy, "In-situ carcinoma adjacent to recurrent nasopharyngeal carcinoma. Evidence of a new growth?" *Pathol. Res. Pract.* **189**, 1067–1070 (1993).
31. F. Cheung, S. W. Pang, F. Hioe, K. N. Cheung, A. Lee, and T. K. Yau, "Nasopharyngeal carcinoma *in situ*," *Cancer* **83**, 1069–1073 (1998).
32. M. W. Pak, K. F. To, Y. M. D. Lo, L. Y. S. Chan, J. H. M. Tong, K. W. Lo, and C. A. van Hasselt, "Nasopharyngeal carcinoma *in situ* (NPCIS)—pathologic and clinical perspectives," *Head Neck* **24**, 989–995 (2002).
33. M. P. Nelson and M. L. Myrick, "Fabrication and evaluation of a dimension-reduction fiber optic system for chemical imaging applications," *Rev. Sci. Instrum.* **70**, 2836–2844 (1999).
34. Jianan Y. Qu, Hanpeng Chang, and Shenming Xiong, "Optical processing of light induced autofluorescence for characterization of tissue pathology," *Opt. Lett.* **26**, 1268–1270 (2001).