Comparison of the Performance of Linear Multivariate Analysis Methods for Normal and Dyplasia Tissues Differentiation Using Autofluorescence Spectroscopy

Shou Chia Chu, Tzu-Chien Ryan Hsiao, Jen K. Lin, Chih-Yu Wang, and Huihua Kenny Chiang*, *Member, IEEE*

***Abstract*—We compared the performance of three widely used linear multivariate methods for autofluorescence spectroscopic tissues differentiation. Principal component analysis (PCA), partial least squares (PLS), and multivariate linear regression (MVLR) were compared for differentiating at normal, tubular adenoma/epithelial dysplasia and cancer in colorectal and oral tissues. The methods' performances were evaluated by cross-validation analysis. The group-averaged predictive diagnostic accuracies were 85% (PCA), 90% (PLS), and 89% (MVLR) for colorectal tissues; 89% (PCA), 90% (PLS), and 90% (MVLR) for oral tissues. This study found that both PLS and MVLR achieved higher diagnostic results than did PCA.**

***Index Terms*—Colorectal tissue, light-induced autofluorescence, multivariate linear regression, oral tissue, partial least squares, principal component analysis.**

I. Introduction

LIGHT-INDUCED autofluorescence (LIAF) measurement is an optical technique that is based on the principle that intrinsic tissue fluorophores absorb ultraviolet light and fluoresce at longer wavelengths. Previous researchers have shown that low-power light radiation is capable of inducing autofluorescence from tissues without causing damage, and which can be implemented in real time during a clinical examination [1]. In general, the different stages of pathological development of tissue may influence the composition of intrinsic tissue fluorophores and lead to changes in the LIAF spectra. Presently, many various types of LIAF measurement systems have been developed by researchers for identifying different stages of the pathological development of a number of surface cancers, including oral cancer [2], nasopharyngeal carcinoma [3] esophageal carcinogenesis [4] gastrointestinal cancer [5], colorectal neoplasms [6], skin tumors [7], and cervical precancer [8].

Human tissues contain intrinsic fluorophores, such as collagen, nicotinamide adenine dinucleotide (NADH), and flavin adenine dinucleotide (FAD). The fluorescence spectra of these intrinsic tissue fluorophores are broadband and overlapping with each other. In addition, the measured fluorescence spectra can be affected by scattering and absorption effects while the fluorescence propagates through the tissue. Therefore, differentiating among the different stages of cancerous tissue development on the basis of autofluorescence spectra is a challenging research objective. Many researchers utilize three multivariate autofluorescence spectroscopic methods—principal component analysis (PCA), partial least squares (PLS), and multivariate linear regression (MVLR)—for the differentiation of different developmental stages of biologic tissue.

PCA is a widely used multivariate spectroscopic method [3], [9]–[11]. By using the singular value decomposition method, PCA decomposes the entire fluorescence spectra into a linear set of eigenvalues and corresponding eigenvectors. Ramanujam *et al.* [10], [11] adopted the PCA method for differentiating human squamous intraepithelial lesions (SILs) from normal squamous epithelia and inflammation tissues, and also for discriminating high-grade SILs, non-high-grade SILs, and non-SILs of cervix tissue. Chang *et al.* [3] compared the performances of PCA, two-wavelengths analysis, and three-wavelengths analysis for classifying the autofluorescence spectra of nasopharyngeal carcinomas. Their results showed that PCA could achieve a higher diagnostic accuracy for the detection of nasopharyngeal carcinoma because it takes advantage of the diagnostic information carried across the entire fluorescence spectra.

PLS is another widely used multivariate spectroscopic method [2], [9], [12]–[14]. The PLS method, which is based on factor analysis, describes the linear relationship between the multivariate data set and the desired output variables. This method extracts a set of factors that account for most variance of the measured spectra, and thus the related information of different stages of cancer tissue development can be extracted from the multivariate data set by using fewer variables. Wang and Chiang [2], [13], [14] applied the PLS method in the LIAF spectroscopic analysis of oral squamous cell carcinoma for discrimination of different stages of cancer tissue development.

O'Brien *et al.* [15] proposed the use of MVLR for developing a spectral classification algorithm for fluorescence-guided laser angioplasty. Their results indicated that this method classified atherosclerotic and normal aorta with an 89% accuracy. Schomacker *et al.* [16] successfully implemented MVLR for identifying the different stages of pathological development

Manuscript received September 28, 2005; revised March 26, 2006. This work was supported in part by the National Science Council, R.O.C. under Contract NSC 90-2736-L-010-003 and Contract NSC 91-2736-L-010-002. *Asterisk indicates corresponding author.*

S. C. Chu is with the Institute of Biomedical Engineering, National Yang-Ming University, Taipei, Taiwan R.O.C. (e-mail: d49004003@ym.edu.tw).

T.-C. R. Hsiao is with Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

J. K. Lin is with Division of Colon and Rectal Surgery, Veterans General Hospital, Taipei, Taiwan, R.O.C.

C.-Y. Wang is with Department of Biomedical Engineering, I-Shou University, Kaohsiung, Taiwan, R.O.C.

*H. K. Chiang is with the Institute of Biomedical Engineering, National Yang-Ming University, Taipei, Taiwan R.O.C. He is also with Department of Education and Research, Taipei City Taipei Hospital, Taipei, Taiwan R.O.C. (e-mail: hkchiang@ym.edu.tw).

Digital Object Identifier 10.1109/TBME.2006.883643

of colorectal tissue. The MVLR method includes two main steps: 1) linear decomposition of each fluorescence spectrum through the fluorescence spectra of each of the intrinsic tissue fluorophores; 2) multivariate linear regression of the coding values through use of the factor scores of each of the intrinsic tissue fluorophores. Their results demonstrated that the linear combination of the spectra of intrinsic tissue fluorophores (collagen, NADH, and FAD) and the absorption spectrum of hemoglobin provided the physical composition of the measured fluorescence spectra of different stages of pathological development of colorectal tissue.

PLS, PCA, and artificial neural network neural network (ANN) algorithms have been commonly employed in multivariate spectroscopic analysis. The ANN technique has several variations in algorithm, such as the widely used back propagation network (BPN) algorithm. A previous study by Hsiao and Chiang *et al.* have shown that the PLS, PCA, and BPN algorithms are all based on a common three-layered computation architecture, which is composed of an input layer, a hidden layer, and an output layer [12]. These layers are associated with each other by connecting weights. All these algorithms can reach reasonable spectrum analysis accuracy in general cases. However, PLS and PCA can obtain a global minimum result, while BPN often converges to a local minimum result [12].

Although PCA, PLS, and MVLR have yielded results with promising sensitivity and specificity for cancer tissue classification, these methods are quite different in mathematical and physical features. These methods also differ in their discriminate analytic ability in the multivariate spectroscopic analysis. Eker *et al.* [17] compared the PCA and PLS methods for laryngeal fluorescence spectroscopic analysis. Their results showed that PLS performed as well as PCA. O'Brien *et al.* [1] compared the MVLR and PCA methods for fluorescence-guided laser angioplasty and concluded that PCA emphasized a few distinct spectral features and enabled more accurate classification than MVLR. Since these methods have been widely used in the classification of cancerous tissues, a comparison of their performances and discriminate abilities is very useful for providing essential information about the advantages of each methods for other cancer tissue diagnosis.

In this study, the main focus is to compare the respective performances of the above-mentioned three linear multivariate spectroscopic methods when used in the diagnosis of different stages of pathological development of colorectal and oral tissues. To achieve this objective, the first step is to obtain the fluorescence spectra of three samples of colorectal and oral tissues in three different pathological states: normal state, tubular adenoma/epithelial dysplasia state, and cancerous state; and then a three-layered multivariate architecture is adopted for the analysis of the data from the fluorescence spectra. The number of hidden nodes was set according to the number of principal components (PCs), and which was determined based on the minimum prediction error sum of squares (PRESS) from cross-validation (CV) analysis.

II. Material and Methods

Subjects: A fluorescence spectrometer (Aminco-Bowman Series II, Thermo Spectronic, Waltham, MA) with a 150-W Xenon lamp was used for tissue LIAF measurements. A total of 70 colorectal tissue specimens were provided from the Veterans General Hospital, Taipei, Taiwan, and 96 oral tissue specimens were got from forty-eight male Syrian hamsters. The surgical tissues were stored in a $-70\,^{\circ}\mathrm{C}$ refrigerator immediately, and were stored until the time of fluorescent measurement. There were 28 normal, 20 tubular adenoma, and 22 cancer tissues in colorectal tissues; 48 normal, 30 epithelial dysplasia, and 18 cancer in oral tissues.

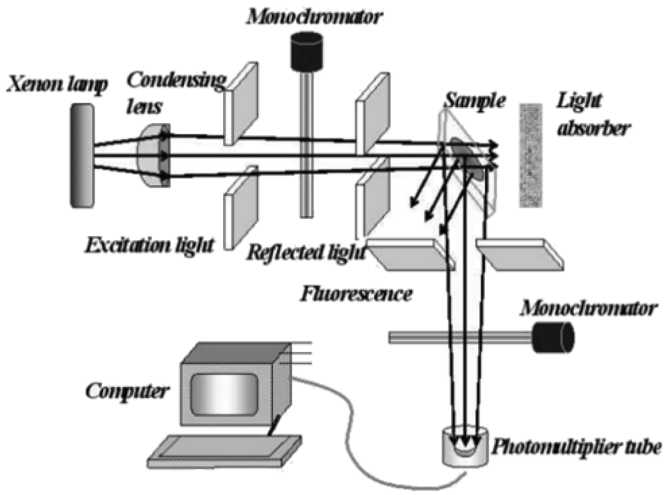


Fig. 1. A schematic diagram. Showing the internal structure of the LIAF measurement system.

In this study, all tissues were frozen in 5 min immediately after surgical operation. A preliminary experiment that was conducted to monitor the autofluorescence alterations between the fresh and frozen samples showed less than 3% change in the intensities between two kinds of samples. According to Schomaker *et al.* the fluorescence intensity of NADH decays with a time constant of 118 min after resection [16]. Since the tissues were frozen within 5 min after surgical operation, the time that elapsed between the resection and the freezing of samples might not affect the data significantly.

All specimens (5×5 mm) were set on a metal plate and mounted on a custom-designed sample holder so that a specimen's surface was facing the excitation beam at a 37° incident angle. The fluorescence was collected at 90° normal to the excitation light, as Fig. 1. Such an arrangement can effectively reduce the collection of the excitation light reflected from the surface of the specimen [2]. The emission spectrum was measured from 350 to 580 nm with the excitation wavelength set at 330 ± 5 nm [with a 10-nm full-width at half-maximum (FWHM)]. All measured spectra were area-normalized; the original fluorescence spectrum was divided by the integrated area under the entire spectrum. After the LIAF measurement, the specimens were embedded in 10% neutral formalin and sent to the pathology department for diagnosis. Two pathologists, who were blinded to the fluorescence spectra results, reviewed the specimens. The pathology report for each specimen was recorded and used for further analysis.

Three-Layered Multivariate Architecture: In the analysis of spectral data, a three-layered multivariate architecture, which is composed of an input layer, a hidden layer, and an output layer, is adopted for comparing the respective physical features

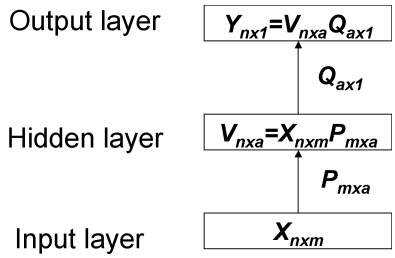


Fig. 2. The three-layered multivariate architecture of the PCA, PLS, and MVLR methods in multivariate spectroscopic analysis.

of the three linear multivariate spectroscopic methods—PCA, PLS, and MVLR (Fig. 2), i.e.,

$$\begin{aligned} V_{nxa} &= X_{nxm}P_{mxa} \\ Y_{nx1} &= V_{nxa}Q_{ax1} \end{aligned} \tag{1}$$

where the matrices X_{nxm}, V_{nxa}, and Y_{nx1} represent the variables from the input layer, the hidden layer, and the output layer, respectively; the matrix P_{mxa} represents the connective weights between the input layer and the hidden layer; and the matrix Q_{ax1} represents the connective weights between the hidden layer and the output layer. In multivariate spectroscopic analysis, the input-layer variables can be treated as the fluorescence spectra, i.e.,

$$X_{nxm} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{2}$$

where the element x_{nm} denotes the intensity of fluorescence spectra of the n^{th} sample at the m^{th} wavelength. The output-layer variables can be coded in predefined values for representing the different stages of pathological development of tissue, for example "2" for normal tissue, "1" for tubular adenoma/epithelial dysplasia tissue, and "0" for cancerous tissue). The development stages were determined by the pathology of each specimen.

PCA Method: PCA is a widely used multivariate spectroscopic analytic tool. This method transforms the original variables of the fluorescence spectra into a smaller set of linear combinations of PCs that account for most of the variance of the original data set. By the PCA method, the first step is to decompose the fluorescence spectra by using the singular value decomposition [18], i.e., $X_{nxm} = O_{nxa}S_{axa}P_{axm}^T = V_{nxa}P_{axm}^T$, where the superscript T denotes the transposition of the matrix, a represents the number of eigenvalues, and $a \leq \min(n, m)$. The matrices $S_{axa}^T S_{axa}$ and P_{mxa} are the eigenvalues and corresponding eigenvectors of the square covariance matrix $X_{mxn}^T X_{nxm}$. The diagonal matrix S_{axa} is the singular value matrix of X_{nxm} with positive or zero elements, i.e., $s_{ij} = 0$ if $i \neq j$. The matrices $O_{nxa} = [O_1, O_2, \ldots, O_a]$ and $P_{mxa} = [P_1, P_2, \ldots, P_a]$ are column-orthogonal matrices, in which the columns represent unit vectors, i.e., $O_i^T O_j = P_i^T P_j = 0$ if $i \neq j$; and $O_i^T O_i = P_i^T P_i = 1$ if $i = j$. In addition, the matrix P_{mxa} can be treated as the PCs of matrix X_{nxm}.

In the three-layered multivariate architecture, the matrix P_{mxa} represents the connective weights between the input layer and the hidden layer. The factor scores of PCs, $V_{nxa} = [V_1, V_2, \ldots, V_a] = O_{nxa}S_{axa} = (O_{nxa}S_{axa}P_{axm}^T)P_{mxa} = X_{nxm}P_{mxa}$, can be calculated and denoted as the hidden-layer variables. The connective weights Q_{ax1} between the hidden layer and the output layer can be determined from the matrices V_{nxa} (hidden layer) and Y_{nx1} (output layer) by using the least squares method. Hence, the PCA method can reduce the fluorescence spectra into the factor scores of PCs as well as constructs the connective weights between the factor scores of PCs and the stages of pathological development of tissue.

PLS Method: PLS is the other commonly employed method of multivariate spectroscopic analysis, which describes the maximum variance between the fluorescence spectra X_{nxm} and the known diagnostic results Y_{nx1}. The connective weights, P_{mxa} and Q_{ax1}, and the factor scores, V_{nxa}, can be derived from the matrices X_{nxm} and Y_{nx1} based on the following equations:

$$\begin{aligned} X_{nxm} &= V_{nxa}P_{axm}^T + E \\ &= V_1P_1^T + V_2P_2^T + \cdots + V_aP_a^T + E_{nxm} \\ Y_{nx1} &= V_{nxa}Q_{ax1} + F \\ &= V_1Q_1 + V_2Q_2 + \cdots + V_aQ_a + F_{nx1} \end{aligned} \tag{3}$$

where a represents the number of PLS regression factors, and the matrices E_{nxm} and F_{nx1} represent the residual matrices after the extraction of PLS regression factors. The structure of the PLS method can also be represented as a three-layered multivariate architecture. The matrix P_{mxa} represents the connective weights between the input layer and the hidden layer, and the matrix Q_{ax1} represents the connective weights between the hidden layer and the output layer. The factor scores, $V_{nxa} = X_{nxm}P_{mxa}$, can be denoted as the hidden-layer variables, and then a is the number of hidden nodes [12].

The PLS procedure includes the following steps:

1) setting the matrix Y_{nx1} as the temporal scores matrix V_i;
2) computing the values of the connective weights P_i^T based on the matrices X_{nxm} and Y_{nx1} by using the least squares method with the vector scaled to length 1;
3) estimating the values of the scores matrix V_i with the use of the matrices X_{nxm} and P_i;
4) computing the values of the loading weight Q_i based on the matrices V_i and Y_{nx1} by using the least squares method;
5) determining the values of the residual matrices E_{nxm} and F_{nx1} after the steps (a)–(d) have been repetitively executed for a number of times equal to a, where

$$E_{nxm} = X_{nxm} - \sum_{i=1}^{a} V_iP_i^T \quad \text{and} \quad F_{nx1} = Y_{nx1} - \sum_{i=1}^{i} V_iQ_i.$$

The PLS factor scores V_i are determined for describing the most variance of the multivariate fluorescence spectra X_{nxm} as well as by correlating the known diagnostic results Y_{nx1}. Therefore, the relationship between the matrices X_{nxm} and Y_{nx1} can be represented by the connective weights P_{axm}^T and Q_{ax1} and

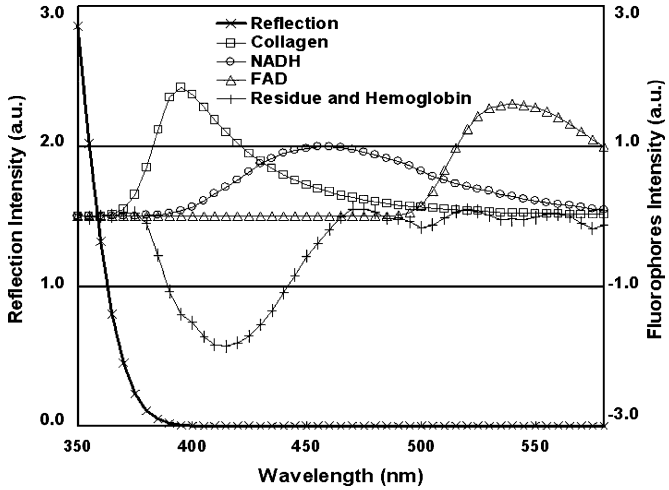


Fig. 3. The fluorescence spectra of collagen, NADH, and FAD fluorophores, and part of the reflection spectra of excitation light, and the spectra of residue (including hemoglobin).

the factor scores V_{nxa} by making $||E_{nxm}||$ and $||F_{nx1}||$ as small as possible (the $||\ ||$ symbol denotes Euclidean normalization).

Magnitude of PLS Factor Scores: By applying the eigenvalues/eigenvectors concept of the PCA method, the magnitude of PLS factor scores can be derived. Based on the PLS analysis, the factor scores V_{nxa} can be rewritten as follows:

$$\begin{aligned} V_{nxa} &= [V_1, V_2, \ldots, V_a] \\ &= \left[\frac{V_1}{||V_1||}, \frac{V_2}{||V_2||}, \ldots, \frac{V_a}{||V_a||}\right] \\ &\quad \times \begin{bmatrix} ||V_1|| & 0 & \ldots & 0 \\ 0 & ||V_2|| & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & ||V_a|| \end{bmatrix} \end{aligned} \tag{4}$$

where the scalar $||V_a||$ and vector $V_a/||V_a||$ are the nth magnitude and unit vector of the factor scores in the PLS method: $||V_a|| = \sqrt{V_a^T V_a}$. Therefore, (3) and (4) can be rewritten as $X_{nxm} \approx O_{nxa} S_{axa} P_{axm}^T$. Because the matrices O_{nxa} and P_{mxa} are the column-orthogonal matrices, the square covariance matrix $V_{axn}^T V_{nxa} = S_{axa}^T S_{axa}$ can be also treated as the eigenvalues of the square covariance matrix $X_{mxn}^T X_{nxm}$ in the PLS method.

MVLR Method: The MVLR method includes two main steps: 1) linear decomposition of each fluorescence spectrum through use of the fluorescence spectra of each of the intrinsic tissue fluorophores (Fig. 3); 22) multivariate linear regression of the coding values by using the factor scores of each of the intrinsic tissue fluorophores. In the multivariate spectroscopic analysis, the measured fluorescence spectra X_{nxm} consist of a linear combination of the following elements: 1) part of the excitation light source (E) of reflected and scattered light from the front surface; 2) the autofluorescence spectra of the intrinsic tissue fluorophores of tissues. Collagen (C), NADH (N), and FAD (F) are the main intrinsic tissue fluorophores. The mathematical representation of the linear combination is

$$\hat{X}_i = e_i E + c_i C + n_i N + f_i F \tag{5}$$

where e_i, c_i, n_i, and f_i represent the factor scores of the spectral vector E, C, N, and F when $||X_i - \hat{X}_i||$ is as small as possible. Because all spectra are normalized, these factor scores also represent the fractional percentages of their respective spectra. The difference between "X" and "$\hat{X}_i$" was minimized by fitting "X" and the measurement spectrum "$\hat{X}_i$", over 350–360 nm (spectral vector E), over 370 to 380 nm (spectral vector C), over 470–480 nm (spectral vector N), and over 550–560 nm (spectral vector F), but neglecting the range from 380–470 nm, where the hemoglobin absorption range is and a good fit is not expected. To quantify the hemoglobin re-absorption effect, $h_i = 1 - (X_i(\lambda = 425\ \text{nm}))/(\hat{X}_i(\lambda = 425\ \text{nm}))$, was defined as the factor score of the hemoglobin absorption factor at 425 nm wavelength [16].

In the three-layered multivariate architecture, the factor scores, e_k, c_k, n_k, f_k, and h_k, represent the hidden-layer variables. The connective weights between the hidden and output layers, Q_{ax1}, are determined from the factor scores e_i, c_i, n_i, f_i, h_i (hidden layer), and Y_{nx1} (output layer) by using the least squares method. Hence, the MVLR method decomposes each spectrum to obtain the scores of the spectral vector E, C, N, F, and H and constructs the connective weights of these scores and the stages of pathological development of tissue.

CV Technique: To evaluate effectively the diagnostic performances of PCA, PLS, and MVLR in differentiating normal, tubular adenoma (epithelial dysplasia), and cancerous tissues, a CV technique is adopted. In this study, a leave-one-out CV method was used. Each sample of the tissues is divided into n groups with each group consisting of only one specimen. One group was kept apart for prediction purposes, and all the other groups were used for calibration. The analytical model was then established by using the calibration set, and the prediction result was obtained from the prediction set. Next, the other group was used as the prediction set, and the steps were repeated. The procedure was continued until all n groups had been used for prediction. This leave-one-out CV method yielded a reasonable estimate of the predictive ability of the statistical algorithms.

III. Results and Discussion

Measured Spectra: Fig. 4 illustrates the averaged spectra of the three different stages of pathological development of colorectal tissue; the error bar represents one standard deviation. The normalized fluorescence spectra (Y axis) were plotted against the emission wavelength (X axis). The position of sample and range of illuminate often affect intensity of spectra so normalizing the fluorescence spectrum removes the absolute intensity information. Therefore, the analysis of normalized fluorescence spectra emphasizes information on the spectral shape instead of the intensity of the fluorescence spectra. One tail-like decay spectrum (350–370 nm) and two peaks near 385 and 470 nm appear in Fig. 3. The tail-like decay spectrum represents part of the spectrum of the excitation light source reflected and/or scattered from the surface of the tissue, and the first peak (near 385 nm) decreases whereas the second peak (near 470 nm) increases with the different stages of pathological development of tissue (normal, tubular adenoma, and cancer). In general, these different stages may influence the composition of intrinsic tissue fluorophores and lead to changes in the

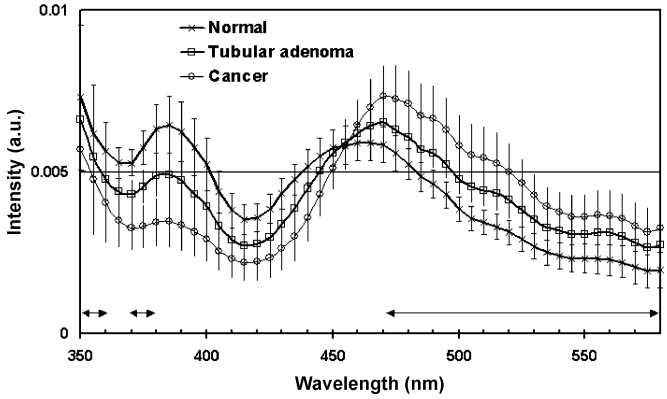


Fig. 4. The three normalized colorectal fluorescence spectra, with one standard deviation. The symbol "↔" denotes the selective range of the MVLR method: 350–360 nm for component E, 370–380 nm for component C, and 470–580 nm for components N and F.

fluorescence spectra. The change of two peaks may be due to the relative decrease in collagen and the increase in NADH with the different stages, from normal state to cancerous state.

The fluorescence spectra of three types of intrinsic tissue fluorophores—collagen, NADH, and FAD—were measured by using the same fluorescence spectrometer. All samples were commercial grade (collagen: Taiwan Salt Biotech. Factory, Tainan, Taiwan; NADH and FAD: Sigma Chemical Co., St. Louis, MO). To simulate the tail-like spectrum of the excitation light reflected and/or scattered from the tissue surface, the reflected light from the surface of BaSO4 was measured while it was being excited at 330 $\pm$ 5 nm (with a 10-nm FWHM). Fig. 4 plots the normalized fluorescence spectra of these fluorophores and the tail-like spectrum of the excitation light. The peaks of the fluorescence spectra of collagen, NADH, and FAD were near 395, 455, and 545 nm, respectively. The residual spectrum was obtained from the averaged residual spectrum between each of the measured spectra (X_i) and a linear spectral combination of intrinsic tissue fluorophores $(\hat{X}_{nxm})$.

Comparing the PCs of PCA and PLS Methods: In the three-layered multivariate architecture, the connective weights between the input and hidden layers are the PCs of the input-layer variables. Because the PCA method employs only the input-layer variables for the mathematical derivation of PCs and the PLS method employs the input- and output-layer variables for the mathematical derivation of PCs, it is necessary to compare the statistical meaning of PCs in both the PCA and PLS methods. Fig. 5 illustrates the four leading PCs of 70 colorectal through use of the PCA and PLS methods. In colorectal tissue, the correlation coefficient (*Pearson's* r) [19] of the first PCs between the PCA and PLS methods is 0.998. Therefore, both of the first PCs of the PCA and PLS methods have a significant correlation. This statistical result indicates that the first PCs of the PCA method also correlate the relationship between the input- and output-layer variables. In addition, the mathematical derivation of the PCs of the PCA method employs only the input-layer variables; and hence, it can be concluded that the input-layer variables (350–580 nm range) correlated to the output-layer variables. Therefore, the normalized LIAF spectra correlated to the different stages

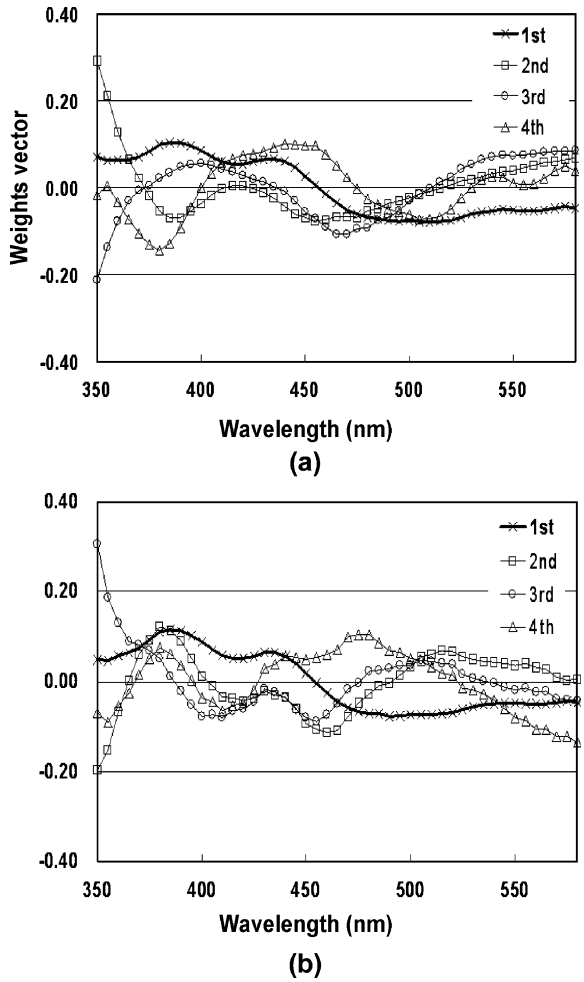


Fig. 5. The four leading principal components of the colorectal fluorescence spectra as determined by the (a) PCA and (b) PLS methods.

of pathological development of tissue. The normalized LIAF spectra from 350–580 nm, which is excited by light source, are the major representative spectra of the colorectal tissues.

From the first PCs of the LIAF spectra of the colorectal tissue, it can be observed that a waveform segment of positive intensity extends within the range from 370 to 400 nm indicative of a relative decrease in collagen with cancer development, and a waveform segment of negative intensity extends within the range from 450 to 500 nm indicative of a relative increase in NADH with cancer development. Hence, this result suggests that the first PCs in colorectal tissue captured the difference in the spectral data between the regions of 370–400 nm and 450–500 nm and correspond to the autofluorescence range of two intrinsic tissue fluorophores, collagen, and NADH, respectively.

Comparing the Factor Scores of PCA and PLS Methods: In the three-layered multivariate architecture, the hidden-layer variables are the factor scores of PCs. Given that Ramanujam *et al.* [10] showed that eigenvalues could be used to describe most of the variance of the original data set in the PCA method. Wang and Chiang [2] showed that the larger factor scores represented more information in the PLS method. It is worth noting that the factor scores of PCs can be used to compare the calibration abilities of the PCA and PLS methods. Therefore,

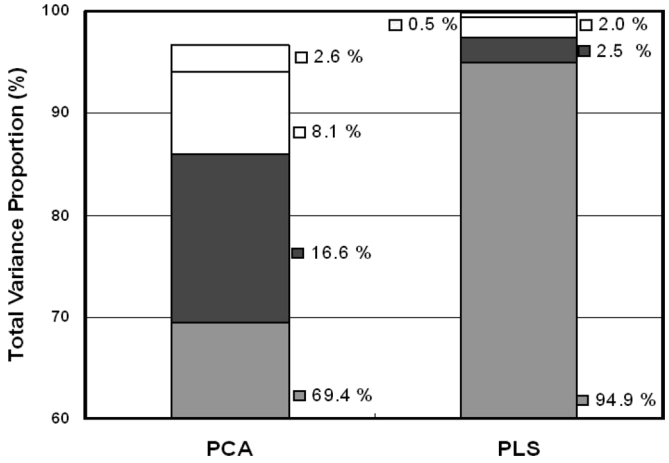


Fig. 6. Summation of the four leading total variance proportion in the colorectal fluorescence spectra, as determined by the PCA and PLS methods.

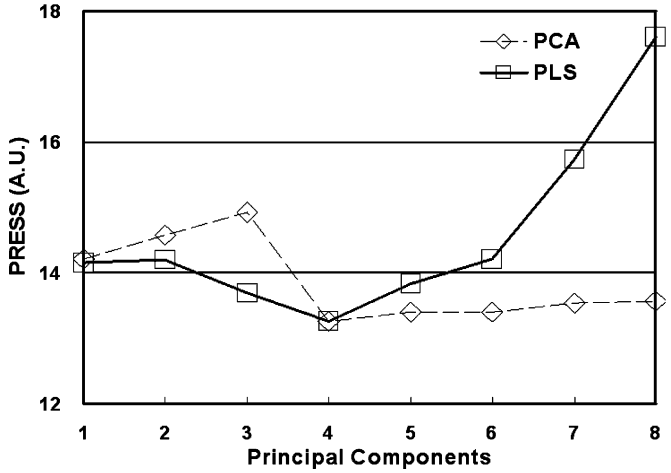


Fig. 7. The $PRESS$ values of the PCA and PLS methods of analyzing the colorectal fluorescence spectra, according to CV analysis.

to compare the respective calibration abilities of PCA and PLS, the following total variance proportion (TVP) is adopted:

$$\mathrm{TVP(i)} = \frac{s_i^2}{\sum_{j=1}^{n} s_j^2} \qquad \text{for the PCA metod}$$

and

$$\mathrm{TVP(i)} = \frac{\|V_i\|^2}{\sum_{j=1}^{n} \|V_j\|^2} \qquad \text{for the PLS method.}$$

Fig. 6 shows the four leading TVPs of colorectal fluorescence spectra for each of the two methods. The summation in colorectal spectra of $\mathrm{TVP}(1) - \mathrm{TVP}(4)$ of the PCA method is over 96%; whereas the summation in colorectal is over 99% for the PLS method and the PCA method. This result indicates that most of the variance of the fluorescence spectra can be described by using the four leading PCs. In addition, the 94.9% $\mathrm{TVP}(1)$ of the PLS method is much larger than the 69.4% $\mathrm{TVP}(1)$ of the PCA method in colorectal tissue.

In summary, the PLS method is likely capable of obtaining more variance of the original data set because this method links the input- and output-layer variables for the calculation of the hidden-layer variables, whereas PCA links only the input layer variables for the calculation of the hidden-layer variables.

Optimal Number of Principle Components(PCs) Using PCA and PLS Methods: The PRESS value of the PCA and PLS methods in CV analysis is calculated to determine the optimal number of hidden nodes, i.e.,

$$\mathrm{PRESS} = \sum_{i=1}^{n} \sum_{j=1}^{n-1} (y_{ij} - \hat{y}_{ij})^2$$

where y_{ij} and $\hat{y}_{ij}$ denote the desired and predicted values on the jth spectrum in the i^{th} CV analysis process. Fig. 7 shows that the location of the lowest PRESS value corresponds to four PCs in colorectal tissue in both the PCA and the PLS methods.

TABLE I
COLORECTAL TISSUE SPECTROSCOPIC DIAGNOSTIC USING CV ANALYSIS

Method	Prediction	Pathology			Accuracy	
		C	P	N	subgroup	group average
PCA	C	20	4	0	91%	85%
	P	2	15	3	75%	
	N	0	1	25	89%	
PLS	C	19	1	0	86%	89%
	P	3	18	2	90%	
	N	0	1	26	93%	
MVLR	C	19	1	0	86%	89%
	P	3	17	1	85%	
	N	0	2	27	96%	

C=cancer; P=tubular adenoma; and N=normal.

In the three-layered multivariate architecture, the lowest PRESS value indicates the optimal number of hidden nodes [12], [20]. The choice of an insufficient number of hidden nodes can lead to a higher PRESS value in the CV calculation process. Likewise, the choice of a greater-than-needed number of hidden nodes will lead to an overfit and result in a higher PRESS value. Hence, Fig. 7 indicates that the leading PCs in tissues of the PCA and PLS methods are the main PCs of fluorescence spectra in the research.

Diagnostic Accuracy: Table I illustrates the diagnostic accuracy of the spectroscopic analysis of colorectal tissue by using the PCA, PLS, and MVLR methods in CV analysis. The PCA method achieved an accuracy of 91% for cancerous tissues, 75% for tubular adenoma tissues, and 89% for normal tissues; the PLS method achieved 86%, 90%, and 93%; and the MVLR method achieved 86%, 85%, and 96%. Based on these facts, it can be concluded that PLS and MVLR obtained better group-averaged predictive diagnostic accuracy than PCA. In addition, notably both the PLS and MVLR methods achieved better diagnostic results in the tubular adenoma group (90% and 85%) than the PCA method.

The PCA method obtained a group-averaged predictive diagnostic accuracy in colorectal tissue, not too much worse than the PLS and MVLR methods, because the measured fluorescence spectra, 350–580 nm, which is excited by light source, are the major representative spectra of the tissues. The PLS method is a better method than the MVLR method based on the fact

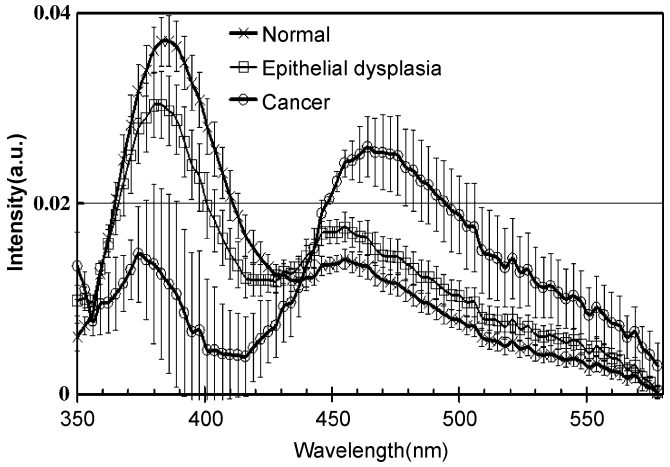


Fig. 8(a). The three normalized oral fluorescence spectra, with one standard deviation.

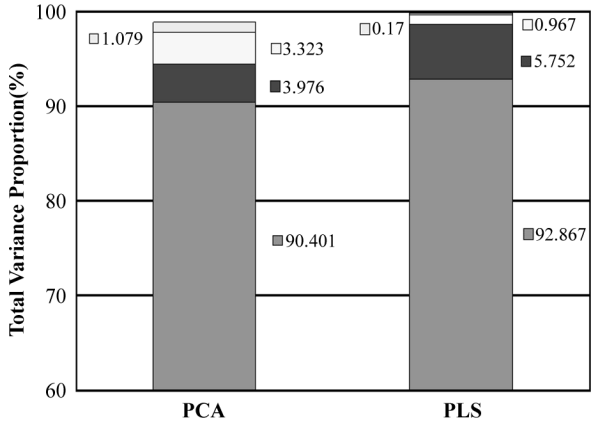


Fig. 8(b). Summation of the four leading total variance proportion in the oral fluorescence spectra, as determined by the PCA and PLS methods.

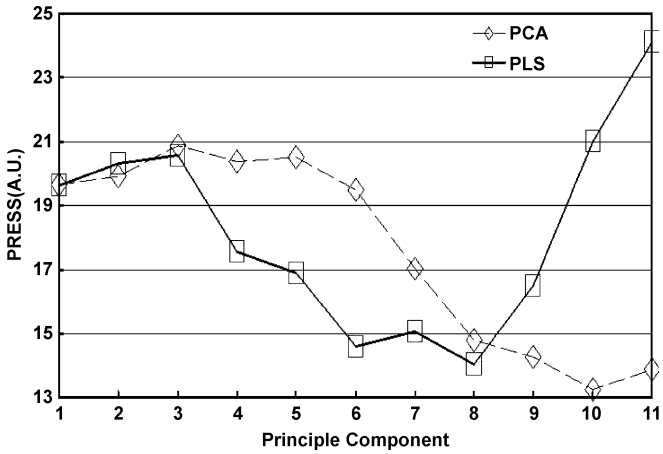


Fig. 8(c). The PRESS values of the PCA and PLS methods of analyzing the oral fluorescence spectra based on CV analysis.

that the PLS method requires no prior knowledge of the fluorescence spectra of intrinsic tissue fluorophores and is unnecessary to select each of the fluorophores' spectral range for the spectroscopic analysis.

Implementation on Oral Tissue Study: In order to further evaluate the respective performances of the three methods-PCA, PLS, and MVLR, these methods were also implemented on previous oral tissue study for comparison.

Fig. 8(a) illustrates normalized autofluorescence spectra of different stages of pathological development of oral tissue.—normal, epithelial dysplasia, and cancer tissue. In general, the fluorescence spectra characteristics of oral tissues are very similar to those of colon tissue. In normal and epithelial dysplasia tissues, the collagen peak is higher than NADH peak. In cancer tissues, the collagen peak is lower than the NADH peak. Collagen peak intensity decreases and NADH peak increases with the development of cancerous tissue. It can be observed that the hemoglobin absorption peak appears at 420 ± 5 nm.

Fig. 8(b) shows the contribution of factor score [TVP(1) $\sim$(4)] of the PLS and PCA methods. The respective TVP(1) values of the PLS and PCA methods in oral tissue are 92% and 90%; the summation the four leading TVPs of the two methods in oral tissue is over 99%. This result indicates that most important variance of the fluorescence spectra can be well described by the four leading PCs in PCA and PLS methods.

The PRESS values of the PCA and PLS methods analyze the oral fluorescence spectra. Fig. 8(c) shows that the lowest PRESS can be achieved by using eight PCs in PLS and ten PCs in PCA.

The data shown in Table II can be compared to evaluate the diagnostic performance of the PCA, PLS and MVLR methods, which shows that the group-averaged CV accuracy is 89% for PCA, 90% for PLS, and 90% for MVLR. Although the overall performance of these three methods were very close to each other; PLS achieved the best diagnostic results for normal tissues stages; MVLR achieved better diagnostic results for cancer and epithelia dysplasia than PLS and PCA methods did. It is worth knowing that PCA provided best diagnostic accuracy for cancer (c) diagnosis. However, in clinics, the diagnostic accuracy for precancer/early cancer tissues (P) is more important than for cancer tissues because cancer tissues are more easily to be recognized under traditional examinations.

TABLE II
ORAL TISSUE SPECTROSCOPIC DIAGNOSTIC USING CV ANALYSIS

Method	Prediction	Pathology			Accuracy	
		C	P	N	subgroup	group average
PCA	C	18	6	1	100%	
	P	0	24	6	80%	89%
	N	0	0	42	88%	
PLS	C	17	4	1	94%	
	P	1	26	4	87%	90%
	N	0	0	43	90%	
MVLR	C	18	3	1	100%	
	P	0	27	9	90%	90%
	N	0	0	38	79%	

C=cancer; P=epithelial dysplasia; and N=normal.

IV. CONCLUSIONS

In this study, three widely used linear multivariate autofluorescence spectroscopic methods—PCA, PLS, and MVLR—were examined and compared for evaluation of their utilization values in cancer diagnosis. By the multivariate spectroscopic analysis, the following facts can be concluded: the connective weights between the input and hidden layers are equivalent to the PCs of fluorescence spectra; the hidden-layer

variables are equivalent to the factor scores of PCs; and the connective weights between the hidden and output layers are equivalent to the connective weights between the factor scores and the coding values of different stages of pathological development of tissue.

The diagnostic accuracy of these three methods were compared for the fluorescence spectroscopic analysis of colorectal and oral samples. The results obtained from use of these three widely used linear multivariate analysis methods showed great promise for fluorescence spectroscopic analysis in the differentiation of biological tissues between different stages of colorectal and oral carcinogenic development.

Both PLS and MVLR methods reached a high accuracy performed better in differentiating the normal and dyplasia group than did PCA; this differentiation capability is important for early cancer detection. In addition, from comparisons of the calibration abilities of the PCA and PLS methods, it can be concluded that PLS, which benefits from the cross-correlation of the input-layer and output-layer variables, is capable of obtaining a larger variance than could PCA. Therefore, a better overall diagnostic accuracy could be reached by PLS.

REFERENCES

[1] C. R. Kapadia, F. W. Ctruzzola, K. M. O'Brien, M. L. Stetz, R. Enriquez, and L. I. Deckelbaum, "Laser induced fluorescence spectroscopy of human colonic mucosa," *Gastroenterology*, vol. 99, pp. 150–157, 1990.

[2] C. Y. Wang, C. T. Chen, C. P. Chiang, S. T. Young, S. N. Chow, and H. K. Chiang, "Partial least-squares discriminant analysis on autofluorescence spectra of oral carcinogenesis," *Appl. Spectrosc.*, vol. 52, no. 9, pp. 1190–1196, 1998.

[3] H. P. Chang, J. N. 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.*, vol. 56, no. 10, pp. 1361–1367, 2002.

[4] R. Glasgold, M. Glasgold, H. Savage, J. Pinto, R. Alfano, and S. Schantz, "Tissue autofluorescence as an intermediate endpoint in NMBA-induced esophageal carcinogenesis," *Cancer Lett.*, vol. 82, no. 1, pp. 33–41, 1994.

[5] H. Stepp, R. Sroka, and R. Baumgartner, "Fluorescence endoscopy of gastrointestinal diseases: basic principles, techniques, and clinical experience," *Endoscopy*, vol. 30, no. 4, pp. 379–386, 1998.

[6] R. L. Prosst and J. Gahlen, "Fluorescence diagnosis of colorectal neoplasms: A review of clinical applications," *Int. J. Colorectal Dis.*, vol. 17, no. 1, pp. 1–10, 2002.

[7] R. Cubeddu, A. Pifferi, P. Taroni, A. Torricelli, G. Valentini, F. Rinaldi, and E. Sorbellini, "Fluorescence lifetime imaging: an application to the detection of skin tumors," *IEEE J. Sel. Topics Quantum Electron.*, vol. 5, no. 4, pp. 923–929,, Jul.-Aug. 1999.

[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.*, vol. 45, no. 8, pp. 953–961, Aug. 1998.

[9] H. Martens and T. Næs, "Methods for calibration," in *Multivariate Calibration*. New York: Wiley, 1996, ch. 3, p. 97.

[10] N. Ramanujam, M. F. Mitchell, A. Mahadevan, S. Thomsen, A. Malpica, T. Wright, N. Atkinson, and R. Richards-Kortum, "Development of a multivatirate statistical algorithm to analyze human cervical tisue fluorescence spectra acquired *in vivo*," *Lasers Surg. Med.*, vol. 19, pp. 46–62, 1996.

[11] N. Ramanujam, M. F. Mitchell, A. Mahadevan-Jansen, S. L. Thomsen, G. Staerkel, A. Malpica, T. Wright, N. Atkinson, and R. Richards-Kortum, "Cervical precancer detection using a multivariate statistical algorithm based on laser-induced fluorescence spectra at multiple excitation wavelengths," *Photochem. Photobiol.*, vol. 64, no. 4, pp. 720–735, 1996.

[12] T. C. R. Hsiao, C. W. Lin, and H. K. Chiang, "Weights initialization of the backpropagation network by the partial least squares methods," *Neurocomputing*, vol. 50, no. 1, pp. 237–247, 2003.

[13] C. Y. Wang, C. T. Chen, C. P. Chiang, S. T. Young, S. N. Chow, and H. K. Chiang, "A probability-based multivariate statistical algorithm for auto-fluorescence spectroscopic identification of oral carcinogenesis," *Photochem. Photobiol.*, vol. 69, no. 4, pp. 471–477, 1999.

[14] C. Y. Wang, T. Tsai, H. C. Chen, S. C. Chang, C. T. Chen, and C. P. Chiang, "Autofluorescence spectroscopy for *in vivo* diagnosis of DMBA-induced hamster bucccal pouch precancers and cancers," *J. Oral Pathol. Med.*, vol. 32, pp. 18–24, 2003.

[15] K. M. O'Brian, A. F. Gmitro, G. R. Gindi, M. L. Stetz, F. W. Cutruzzola, L. I. Laifer, and L. I. Deckelbaum, "Development and evaluation spectral classification algorithms for fluorescence guided laser angioplasty," *IEEE Trans. Biomed. Eng.*, vol. 36, no. 4, pp. 424–431,, Apr. 1989.

[16] K. T. Schomacker, J. K. Frisoli, C. C. Compton, T. J. Flotte, J. M. Richter, N. S. Nishioka, and T. F. Deutsch, "Ultraviolet laser-induced fluorescence of colonic tissue: basic biology and diagnostic potential," *Lasers Surg. Med.*, vol. 12, pp. 63–78, 1992.

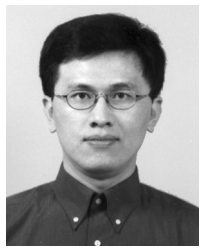
[17] C. Eker, R. Rydell, K. Svanberg, and S. Andersson-Engels, "Multivariate analysis of laryngeal fluorescence spectra recorded *in vivo*," *Lasers Surg. Med.*, vol. 28, pp. 259–266, 2001.

[18] K. I. Diamantaras and S. Y. Kung, "Principal component analysis," in *Principal Component Neural Networks: Theory and Applications*. New York: Wiley, 1996, ch. 3, p. 55.

[19] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Statistical description of data," in *Numerical Recipes in C: The Art of Scientific Computing*. New York: Cambridge Univ. Press, 1992, ch. 14, p. 636.

[20] M. Stone, "Cross-validatory choice and assessment of statistic predictionm," *J. Roy. Statist. Soc., Ser. B*, vol. 36, pp. 111–147, 1974.

Shou Chia Chu was born in Taipei, Taiwan, in 1975. He received the B.S. degree in physics from the National Tsing-Hua University, Taiwan, in 1999 and the M.S. degree from the Institute of Biomedical Engineering, at National Yang-Ming University, Taipei, in 2001. He is working towards the Ph.D. degree in the Institute of Biomedical Engineering, the National Yang-Ming University, Taipei. His research interests include LIAF spectroscopy and LIAF Image.



Tzu-Chien Ryan Hsiao was born in Taoyuan Taiwan, in 1971. He received the M.S. and Ph.D. degrees from the Institute of Physics at the National Sun Yat-Sen University and Institute of Biomedical Engineering at National Yang-Ming University, Taiwan, in 1996 and 2003, respectively.

In 2003, he became an Assistant Professor with the Institute Biomedical Engineering at I-Shou University, and now he is an Assistant Professor of Department of Computer Science, National Chiao Tung University (since 2006) and also a deputy director of biomedical research and development division of Hsinchu Biomedical Science Park. His research interests include neural networks, virtual instrumentation, and multivariate spectral analysis.

Jen K. Lin received the M.D. degree from Taipei Medical University, Taipei, Taiwan, 1977 and the Ph.D. degree from the Institute of Clinical Medicine from National Yang-Ming University, Taipei, in 1993.

In 1979, he joined Department of Surgery, Veterans General Hospital, Taipei, Taiwan R.O.C. He is now Chief of the Division of Colon and Rectal Surgery, Veterans General Hospital and Professor of Surgery in National Yang-Ming University Taipei, Taiwan R.O.C.

Chih-Yu Wang was born in Taiwan, R.O.C., on July 1, 1966. He received the B.S. and the M.S. degrees in control engineering from the National Chiao-Tung University, Hsin-Chu, Taiwan, 1989 and 1991, respectively. He received the Ph.D. degree in biomedical engineering from National Yang Min University, Taipei, Taiwan, 1998.

In 1979, he joined the faculty of the Department of Biomedical Engineering, I-Shou University, where he is currently an Associate Professor. His research interests include biophotonics, optical imaging, virtual biomedical instruments (VBI), and system integration.

Huihua Kenny Chiang (S'91–M'92) received the B.S. degree in electrical engineering from the National Tsing-Hua University, Taiwan, 1982. He received the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1987 and 1991, respectively.

In 1992, he was a Research Scientist at the Georgia Tech Research Institute, GA.. In 1993, he joined the Institute of Biomedical Engineering, the National Yang-Ming University, Taipei, Taiwan, as an Associate Professor, and now is Professor of the Institute of Biomedical Engineering (since 1999). His current research interests include biophotonics, optical diagnostic techniques, and ultrasound signal processing.