Comparative evaluation of spectroscopic models using different multivariate statistical tools in a multicancer scenario

A. D. Ghanate,[a] S. Kothiwale,[a] S. P. Singh,[a] Dominique Bertrand,[b] and C. Murali Krishna[a]
[a]Chilakapati Lab, ACTREC, Navi Mumbai, India 410210
[b]Institut National de la Recherche Agronomique (INRA), 44316 Nantes cedex 3, France

Abstract. Cancer is now recognized as one of the major causes of morbidity and mortality. Histopathological diagnosis, the gold standard, is shown to be subjective, time consuming, prone to interobserver disagreement, and often fails to predict prognosis. Optical spectroscopic methods are being contemplated as adjuncts or alternatives to conventional cancer diagnostics. The most important aspect of these approaches is their objectivity, and multivariate statistical tools play a major role in realizing it. However, rigorous evaluation of the robustness of spectral models is a prerequisite. The utility of Raman spectroscopy in the diagnosis of cancers has been well established. Until now, the specificity and applicability of spectral models have been evaluated for specific cancer types. In this study, we have evaluated the utility of spectroscopic models representing normal and malignant tissues of the breast, cervix, colon, larynx, and oral cavity in a broader perspective, using different multivariate tests. The limit test, which was used in our earlier study, gave high sensitivity but suffered from poor specificity. The performance of other methods such as factorial discriminant analysis and partial least square discriminant analysis are at par with more complex nonlinear methods such as decision trees, but they provide very little information about the classification model. This comparative study thus demonstrates not just the efficacy of Raman spectroscopic models but also the applicability and limitations of different multivariate tools for discrimination under complex conditions such as the multicancer scenario. [DOI: 10.1117/1.3548303]

Keywords: Raman spectroscopy; principal component analysis; factorial discriminant analysis; partial least square discriminant analysis; decision tree.

Paper 10292RR received May 31, 2010; revised manuscript received Jan. 3, 2011; accepted for publication Jan. 4, 2011; published online Feb. 28, 2011.

1 Introduction

Cancer is considered to be one of the most life threatening diseases, and a major public health problem in both developed and developing countries. According to a report, a total of 1,479,350 new cancer cases and 562,340 deaths from cancer were estimated to occur in the United States alone in 2009, corresponding to more than 1,500 deaths per day. The prostate, lung, colon, and rectum are the most common cancers in males, while breast, lung, and bronchus are the most common cancers in females.[1] Late detection of cancerous lesions is considered to be the major cause of death due to cancer. Early detection is limited by the fact that cancers most often are asymptomatic and may not appear in diagnostic images. Histopathological diagnosis, based on visual examination to identify morphological peculiarities that are indicative of disease, is the gold standard for cancer diagnosis. But this approach can be subjective, time consuming, prone to interobserver disagreement, and may not provide prognostic information to a clinician, especially for precancerous conditions.

Detection of neoplastic changes by optical spectroscopy techniques such as Fourier transform infrared (FTIR), Raman, and fluorescence has been actively pursued in recent times. Early transformation from a healthy to a diseased state is attributed to biochemical changes within the tissue. Since optical spectroscopic techniques are very sensitive to biochemical changes, they are being pursued as potential alternatives or adjuncts to conventional diagnostic methodologies. The efficacy and potential applications of optical spectroscopic techniques in cancer diagnosis has been well demonstrated. References 2–4 present the latest topical reviews on biomedical applications of LIF, FTIR, and Raman. Another attractive feature of optical spectroscopic methods is that data are amenable to multivariate statistical tools. Various statistical techniques have been successfully deployed to exploit subtle differences in the spectral profile for diagnosis. Principal component analysis (PCA) and its derivatives such as linear discriminate analysis (LDA) and logistic regression, are some of the widely used methods. Among optical spectroscopic methods, Raman spectroscopy can provide detailed information about the chemical composition of the tissue under study and, since progression of a disease is accompanied by chemical changes, it can provide the physician with valuable information for diagnosing a disease. The most important attribute of Raman spectroscopy is its *in vivo* applications, since light can be delivered and collected rapidly via optical fibers which can be incorporated into catheters, endoscopes, cannulas, and needles.[5–8] Raman microspectroscopy, which

Address all correspondence to: Dr. C. Murali Krishna, Scientific Officer 'E' and Principal Investigator, Chilakapati Lab ACTREC, Tata Memorial Centre, Sector '22', Kharghar, Navi Mumbai – 410210, India. Tel: +91–22-2740 5039; E-mail: mchilakapati@actrec.gov.in; pittu1043@gmail.com.

1083-3668/2011/16(2)/025003/9/$25.00 © 2011 SPIE

facilitates high spatial resolution down to 1 μm, is often used for in-depth analysis of selected regions, while conventional or macro-Raman spectroscopy is better suited for diagnostic applications especially for *in situ* or *in vivo* conditions because the probing area is larger in this mode (20 to 100 μm). Moreover, the findings of conventional spectroscopy of *ex vivo* samples can be extrapolated, to a reasonable extent, to *in vivo* or *in situ* conditions. Hence, conventional Raman studies of *ex vivo* samples are often used as exploratory approaches before taking up eventual *in vivo* or *in situ* evaluations. We have carried out extensive Raman spectroscopic investigations of *ex vivo* oral, cervical, breast, stomach, colon, and ovarian cancer tissues.[9–18] Our studies have demonstrated the feasibility of classifying normal, premalignant, inflammatory, and malignant conditions in the above-stated cancers. We have also verified spectroscopic diagnostic models of normal and pathological oral, breast, and cervix cancers over large data from certified and single blinded samples.

The strength of optical spectroscopy methods relies in its objective diagnosis. But the fact remains that the outcomes are not reliable unless both the input dataset and discrimination methodology are robust enough. So far, specificity or applicability evaluations of a spectroscopic model were limited to a given cancer, e.g., standard sets of oral, normal, and pathological conditions were always evaluated by test oral tissue spectra. In the present study, we have evaluated the specificity of already developed Raman spectroscopic models in healthy and cancerous conditions of the breast, oral cavity, cervix, larynx, and colon tissues using different multivariate statistical tools. Concomitant discrimination of the nature of the tissue, as well as healthy versus cancerous conditions, is considerably tough, allowing for a fair comparison of different discriminant methods. Comparative evaluation of discrimination efficiency has been performed for linear unsupervised methods such as PCA and supervised methods such as factorial discriminant analysis (FDA), partial least square discriminant analysis (PLSDA), and a nonlinear supervised method based on decision trees. The findings of the study are discussed in this paper.

2 Materials and Methods

In our earlier studies, we have developed and evaluated the specificity of Raman spectroscopic models for breast, oral, cervix, colon, and larynx cancers. Raman spectra were recorded with an instrument assembled by us, details of which are described elsewhere.[9–16] In brief, this setup consists of a diode laser (SDL-8530/785 nm, 100 mW) as excitation source and HR 320 spectrograph (600 gr/mm blazed at 900 nm) coupled to a Spectrum One liquid nitrogen cooled CCD that is used for recording the Raman signal. Unwanted signals from excitation source are filtered by a holographic filter (HLBF 785.0, Kaiser Optics). A notch filter (HSPF-5812, Kaiser Optics) is used to remove the Rayleigh scattering. An integration time of 30 s per accumulation with 20 accumulations are the spectral acquisition parameters. Samples are kept moist with saline during measurement, and these conditions are maintained constant for all the measurements.

In the present study, randomly selected 223 spectra from spectral models of the breast (normal 25, malignant 21), oral cavity (normal 22 , malignant 24), cervix (normal 21, malignant 31), larynx (normal 25, malignant 28), and colon (normal 11, malignant 15) were analyzed in a common spectral range of 1200 to 1700 cm^{-1}. These spectra were a part of the spectral models which were tested and verified by us.[9–16] Before proceeding for any analysis, these spectra are required to be converted into a digitized spectral matrix form. For this, consider x_i to be a vector representing a digitized spectrum with p elements (in this study, with spectral range of 1200 to 1700 cm^{-1}, spectra were interpolated to have common 800 digitized point absorbance values). If we have n such spectra (in this study, $n = 223$), we can build a matrix X, dimensioned ($n\,X\,p$), in which an element x_{ij} is the absorbance of the i'th spectrum, for the j'th digitized point. All spectral data used in this study were transformed by the standard normal variate technique (SNV) to reduce within-class variability.[19] After SNV transformation, the 223 spectra together were treated as training samples for generating a classification model with the help of multiple multivariate tools. The efficiency of discrimination of various statistical tools was evaluated using a blinded training dataset.

3 Data Analysis

3.1 *Linear Discriminant Methods*

PCA was performed using the algorithm implemented in commercially available GRAMS 32 (Galactic Corporation, USA). As in our earlier studies, PCA was performed using three different approaches.[9–18] In the first approach, spectra from different tissues were pooled together and scores of factors were checked for discrimination in the unsupervised mode. In the second approach, multiple discriminating parameters were used to give a better and more objective diagnosis. For this, the spectral model of different tissue types was used as the standard calibration dataset. This standard set was subjected to PCA to derive parameters such as scores, the spectral residuals, and the Mahalanobis distance. When a test spectrum (blinded training sample spectrum) and any standard set are of the same class, the values of these parameters for the test samples will fall in the range for that of the standard set and vice versa. In the third approach, match/mismatch tables were computed using the "limit test" approach. In this methodology, the test spectra are matched against standard calibration sets with the above-mentioned discrimination parameters. If the values of the given spectrum fall within a specific set limits, the spectra show YES or PASS, otherwise NO or FAIL.

FDA[20] and PLSDA[21–23] methods were employed using the SAISIR[24] Package which is an open source code. Since a stepwise discriminant analysis method along with verification has been employed for generating classification models (SAISIR documentation) for all of these methods, a criterion of 95% classification efficiency was used for selecting a minimum number of factors for generating models.

FDA models were built using PCA scores generated for the training data set. To remove eigenvalues which provide negligible contribution in data variation, a total of 30 PCA factors were used for generating PCA scores. These scores have been used for developing FDA models for different number of factors. These models were tested against blinded training samples for calculating the classification efficiency.

PLSDA models were generated using different number of partial least square discriminant (PLS) factors. Classification of blinded training samples was done on the basis of maximum index of predicted response variables (class membership variable) for all class types using PLS models.

3.2 *Nonlinear Discriminant Models*

We have implemented decision tree models based on classification and regression tree (CART),[25] J48 (based on C4.5 algorithm)[26] and random forest method (RFM)[27] algorithms. This analysis was carried out using algorithms from Weka.[28] Models were developed and blinded training samples were used for testing the capability of these models for efficient classification.

4 Results and Discussion

4.1 *Spectral Features*

The mean Raman spectra of normal and malignant tissues are shown in Fig. 1. Strong vibrational modes of lipids can be seen in normal breast tissue spectra at 1267, 1301, 1440, 1654, and 1746 cm^{-1} as compared to prominent vibrational modes of proteins in malignant breast tissue spectra indicated by broad and strong amide I at 1650 cm^{-1}, δCH2 bend at around 1450 cm^{-1}, and broad peaks in the amide III 1200 to 1350 cm^{-1} region [Fig. 1(a)]. In general, this trend holds true for the spectra of oral, colon, and larynx tissues [Figs. 1(b), 1(d), and 1(e)]. Mean spectra of normal cervical tissue shows abundance of structural proteins such as collagen and elastin indicated by vibrational modes at 1245, 1267, 1385, 1633, and 1671 cm^{-1}, whereas spectral features in malignant cervical tissue suggests the presence of noncollagenous proteins and lipids at 1305, 1331, and 1645 cm^{-1} [Fig. 1(c)]. In addition to this, sharper amide I and III peaks and redshift in 1450 cm^{-1} can also be seen in malignant spectra in contrast to broad amide I peaks at 1384 and 1269 cm^{-1} of normal cervical tissue spectra.[9–16]

As explained above, even on a cursory glance, the spectral features of normal and malignant conditions in a given tissue-type show very prominent differences. These features can be easily exploited to bring out classification of a normal and a malignant condition in any particular tissue type. But when we consider the spectral features of all the tissue types (breast, oral cavity, colon, cervix, and larynx) investigated in this study, great similarity can be seen among spectra, especially in normal conditions. Hence, it is a big challenge to classify an individual condition in the complex situation as exists in our study. We have explored the feasibility of classifying them by different discrimination algorithms based on PCA, FDA, PLSDA, and decision tree models.

4.2 *Linear Discriminant Methods*

As mentioned earlier, three approaches were employed for spectral data analysis. In the first approach, spectra of breast, oral, cervix, larynx, and colon were pooled for PCA and scores of the factors were explored for discrimination in an unsupervised manner. The results indicate that spectra of any given normal

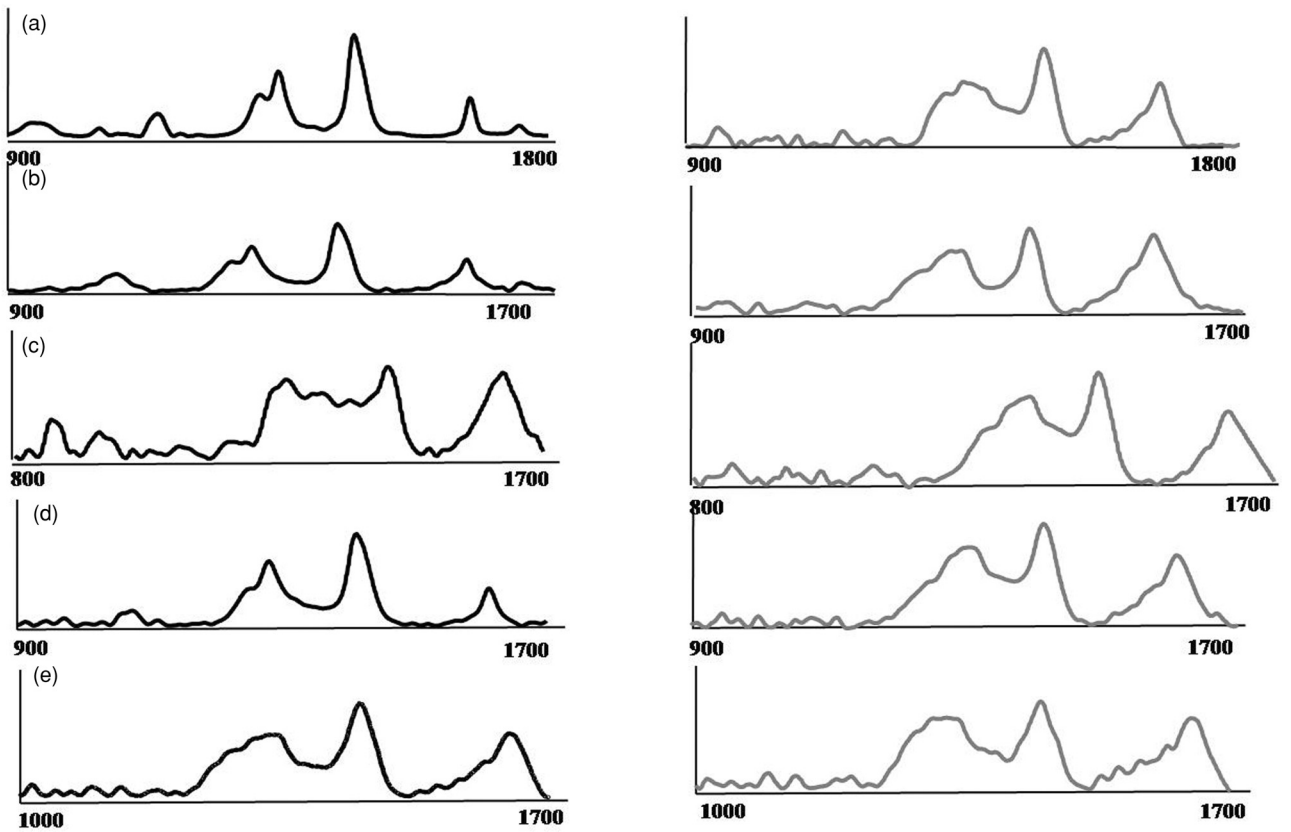

Fig. 1 Mean Raman spectra of normal (black) and malignant tissues (gray). (a) breast, (b) oral, (c) cervix, (d) larynx, (e) colon.

Table 1 Representative best cases of limit test analysis.

	Breast Nor (25)	Breast Malig (21)	Cervix Norm (21)	Cervix Malig (31)	Colon Normal (11)	Colon Malig (15)	Larynx Norm (25)	Larynx Malig (28)	Oral Normal (22)	Oral Malig (24)
Breast Norm	Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y,	N,N,N,N, N,N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,	N,N,N,N, N,N,N,N, N,N,N	N,N,N,N, N,N,N,N, N,N,N,	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N	N,N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N,N, N
Oral Mal	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,	N,N,N,N, N,N,N,N, N,N,N,	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N	Y,Y,Y,Y,Y, Y,Y,Y,Y, Y,Y,Y,Y,Y,Y Y,Y,Y,Y,Y, Y,Y,Y,Y

tissue are exclusive and segregate from their malignant counterpart. But, when we consider findings in the broader prospective, i.e., across different origin, a huge overlap was observed among clusters (data not shown). In supervised analysis, PCA score, spectral residuals, and Mahalanobis distance were used as discriminating parameters. *M* distance is measured in terms of standard deviations from the mean of the training samples. It is a statistical measure of proximity of two spectra.[29,30] Once again, as in the case of unsupervised PCA, this analysis did not yield segregation for all the cancers (data not shown). In the third approach, the limit test was performed. As an example, the typical limit test results of breast normal and oral cavity malignant models are shown in Table 1. Results suggest that the spectral models of breast normal, oral cavity malignant were very specific. In this case, the spectra of breast normal, oral malignant only match (Y) and the rest do not match (N) the standard model of breast normal and oral malignant, respectively. Limit test results of cervix normal and cervix malignant models are shown in Table 2. In this analysis, in addition to the spectra from cervix normal, cervix malignant (11/31), colon malignant (1/15), breast malignant (10/21), and larynx malignant (11/28), spectra also matched the cervix normal standard model. In the case of the cervix malignant model, besides cervix malignant, spectra of cervix normal (13/21), colon malignant (15/15), colon normal (11/11), breast malignant (15/21), oral malignant (2/23), larynx malignant (28/28) and larynx normal (8/25) were matching the cervix malignant standard model. Similarly, we have also observed cross-matching for a few other models. A concise account on the findings of the limit test for all spectroscopic models is given in Table 3. This study reveals the high sensitivity of the limit test but this approach suffers from poor specificity, as is seen for a few spectral models.

Table 2 Representative worst cases of limit test analysis.

	Breast Nor (25)	Breast Malig (21)	Cervix Norm (21)	Cervix Malig (31)	Colon Norm (11)	Colon Malig (15)	Larynx Norm (25)	Larynx Malig (28)	Oral Normal (22)	Oral Malig (24)
Cervix Norm	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,	N,N,N,N, Y,Y,Y,Y,N, N,N,N,N, Y,Y,Y,N, Y,Y,Y,N	Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y Y	N,N,N,N, Y,Y,Y,N,N, Y,N,Y,Y, Y,N,N,N, N,Y,Y,N,Y, Y,N,N,N, N,N,N,N, N	N,N,N,N, N,N,N,N, N,N,N	N,N,N,N, N,N,N,N, N,N,Y,N, N,N,N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,	N,N,Y,Y,N, N,Y,N,N, Y,Y,Y,Y,Y, N,Y,N,N, N,Y,Y,N,N, N,N,N,N, N	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N	N,N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N
Cervix Mal	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N	N,N,N,N, N,N,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y	N,N,N,N, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,N,N, N,N	Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y	Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y	Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y	N,N,N,N, Y, Y,Y,Y,Y,Y, Y,Y,N,N,N, N,N,N,N,N	Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y,Y,Y, Y,Y,Y	N,N,N,N, N,N,N,N, N,N,N,N, N,N,N,N, N,N	N,N,N,N,N, N,N,N,N, N,Y,Y,N,N, N,N,N,N,N, N,N,N,N, N

Table 3 Limit test analysis (diagonal elements are true positive predictions and ex-diagonal elements are false positive predictions).

PCA	CeM	CeN	CoM	CoN	BrM	BrN	OrM	OrN	LaM	LaN
CeM	31	13	15	11	15	0	2	0	28	8
CeN	11	21	1	0	10	0	0	0	11	0
CoM	0	0	15	0	1	0	0	0	0	0
CoN	1	0	0	11	2	0	0	0	7	0
BrM	3	0	15	0	21	0	0	0	2	0
BrN	0	0	0	0	0	25	0	0	0	0
OrM	0	0	0	0	0	0	24	0	0	0
OrN	0	0	0	0	0	17	9	22	0	4
LaM	17	0	3	3	2	0	0	0	28	0
LaN	0	0	0	0	0	20	0	4	0	25

The abnormal findings of the limit test could be attributed to the spectral range that was used for analysis. As is well known, spectra at several ranges need to be explored to bring out classification which often varies from one cancer to another. In our earlier studies, we have used different spectral ranges to bring out classification among normal and malignant classes. For example, short spectral range of 1400 to 1700 cm^{-1} and 1200 to 1800 cm^{-1} were used to bring out classification among breast and oral tissues, respectively. Longer spectral ranges of 800 to 1800 cm^{-1}, 900 to 1750 cm^{-1}, and 1000 to 1800 cm^{-1} gave good classification for cervix, larynx, and colon tissues.[9–16] But, in the present study, we have restricted analyses to a common range of 1200 to 1700 cm^{-1}. This could be a limitation in applicability of this approach. Therefore, we have also explored other multivariate methods such as FDA, PLSDA, and decision trees to bring out the classification.

4.3 *Factorial Discriminant Analysis*

The PCA scores calculated using spectral models [Fig. 2(a)] were further processed using the linear discrimination function, FDA. FDA is extremely useful whenever there are grounds to postulate the existence of a number of groups (categories) into which the samples may be classified, and one has to look for the best discrimination and for a quantitative evaluation of the differences between these groups. In FDA, sample cases are attributed to the group on the basis of classification probability of each spectrum, computed from the distance in discriminant space of PCA scores between the spectrum and the centroid of the nearest class. FDA aims to find out a small number of generalized variables (or factors) that can describe most of the variances and correlations of the initial variables (in this case, Raman shift wave number), reducing the dimension of the measurement space without a loss of information.

In this study, the first 20 factors were used for FDA which provides a classification efficiency of 95.07% [Fig. 2(b)]. Figure 3 shows that the 10 groups of samples are clearly separated except a few cases where some overlaps are seen. The summarized class prediction results are shown in Table 4. Except for a few samples, most of the datasets do not cross-match with other classes (e.g., cervical normal, colon normal, breast malignant, breast normal, larynx malignant). Most overlapping is seen in cervix malignant, where two samples have been misclassified as cervix normal and three samples as larynx malignant. Other mismatches include: one colon malignant sample misclassified as breast malignant, one oral malignant sample as larynx malignant, two oral normal samples as oral malignant, and two larynx normal samples as breast normal. The overall classification efficiency of this methodology is 95.07%, i.e., 212 of 223 spectra are classified correctly.

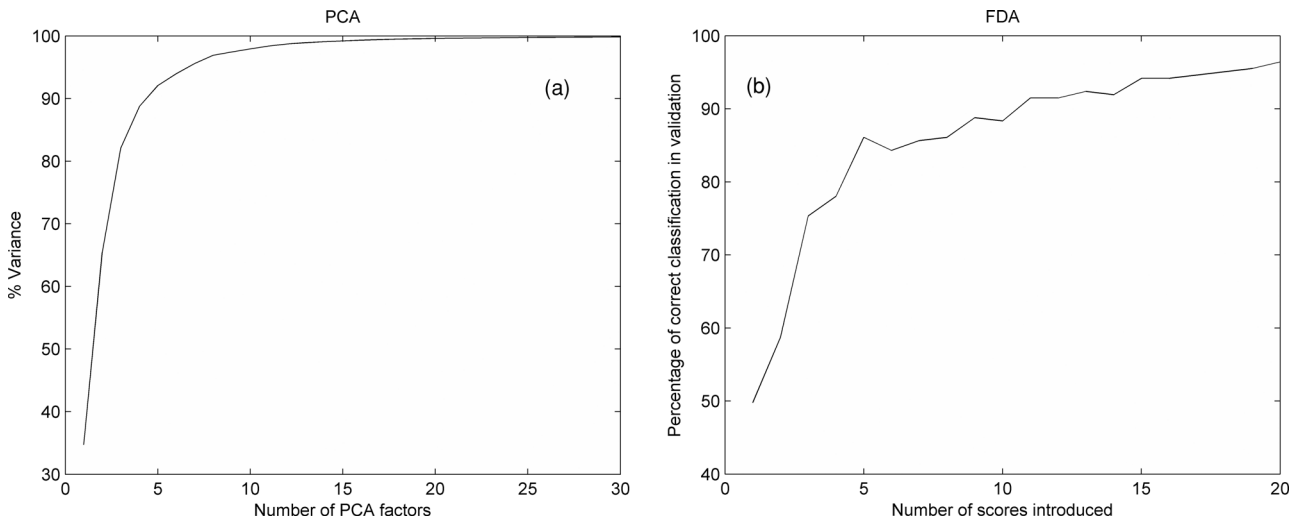

Fig. 2 (a) Cumulative percent variance contribution of PCA factors used for FDA. (b) Representation of variation of correct classification percentage in validation data set against number of scores used.

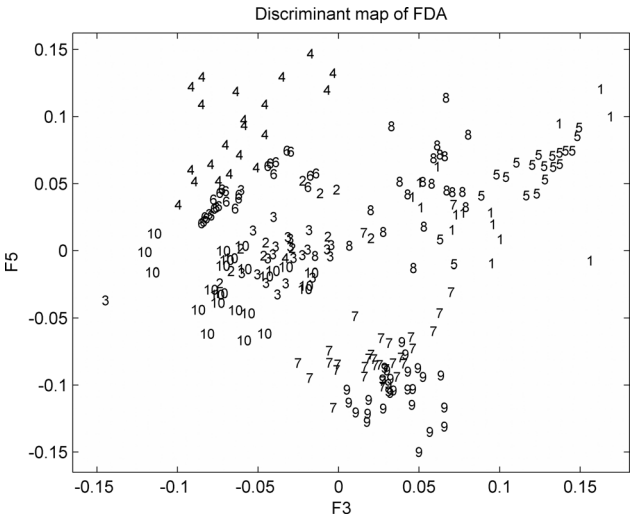

Fig. 3 Scatter plot for factorial discriminant analysis: (1) colon malignant; (2) colon normal; (3) oral malignant; (4) oral normal; (5) breast malignant; (6) breast normal; (7) cervix malignant; (8) cervix normal; (9) larynx malignant; (10) larynx normal.

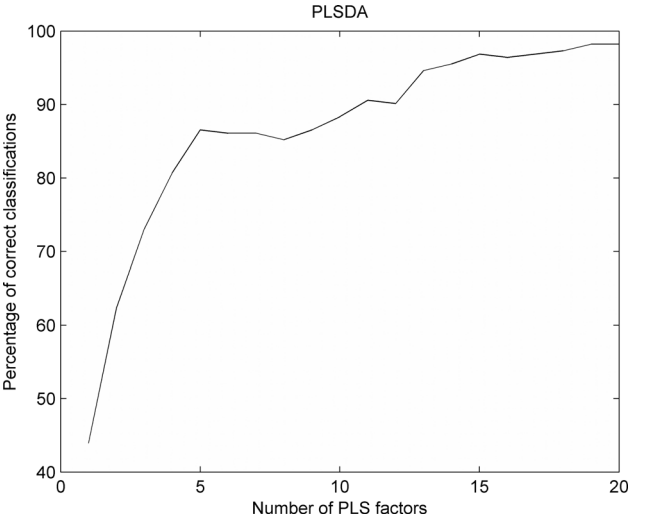

Fig. 4 Representation of variation of percentage of correct classification against number of PLS factors used.

4.4 *Partial Least Square Discriminant Analysis*

PCA-based models are based on the variance within the data set, irrespective of the fact that the variance may or may not be useful for separating classes. Therefore we have explored such a method which captures variance that is useful in separating classes and ignores variance within the class. PLS is one such multivariate regression method which establishes a relationship between one or more dependent variables (class membership of sample cases) and a group of descriptors (group of spectra representing intensity values at different wavenumbers). These are modeled simultaneously, to find the latent variables (LVs) in descriptors that will predict the LVs in dependent variables. In contrast to PCA, which works to explain maximum variation between descriptors, PLSDA explains maximum separation between defined response variables based on descriptors. Since PLS factors are computed hierarchically, the first factor contributes to the maximum variability in data relevant for classification, while the last factors are mostly responsible for random variations and experimental errors. Hence the optimal number of PLS factors, i.e. those modeling information in descriptor useful to predict the response variable but avoiding overfitting, is determined on the basis of the classification efficiency of the model developed using these factors (Fig. 4). In this analysis, we have chosen the first 14 PLS factors depicting 95.5% of classification efficiency and the PLSDA model has been developed and validated by predicting class membership for blinded training samples. The findings of PLSDA results are shown in Table 4 wherein samples from the class comprising of cervical normal, colon normal, breast normal, oral malignant, and larynx malignant shows no cross-matching with other classes. But 10 instances of misclassification were seen. Five samples of cervix malignant were misclassified—two as cervix normal and three as larynx malignant. Two samples of oral normal were misclassified as oral malignant, one colon malignant sample as breast malignant, one breast malignant sample as colon malignant, and one larynx normal as breast normal. Overall, 213 out of 223 spectra are classified correctly, indicating that the classification efficiency of this methodology is 95.52%.

Table 4 Results obtained using linear multivariate analysis methods (total number of misclassified instances observed using different discriminant methods is shown using roman numbers, while each misclassified instance type is mentioned in parentheses).

	No. of misclassified instances (class of misclassified instance)	
Original class (number of spectra)	FDA	PLS-DA
CeM (31)	V (2 CeN, 3 LaM)	V (2 CeN, 3 LaM)
CeN (21)	–	–
CoM (15)	I (BrM)	I (BrM)
CoN (11)	–	–
BrM (21)	–	I (CoM)
BrN (25)	–	–
OrM (24)	I (LaM)	–
OrN (22)	II (2 OrM)	II (2 OrM)
LaM (28)	–	–
LaN (25)	II (2 BrN)	I (BrN)
Classification efficiency	95.07% (212/223)	95.52% (213/223)

4.5 *Decision Trees*

In view of the complex multicancer scenario, the input data possesses a nonlinear structure. In order to map spectral

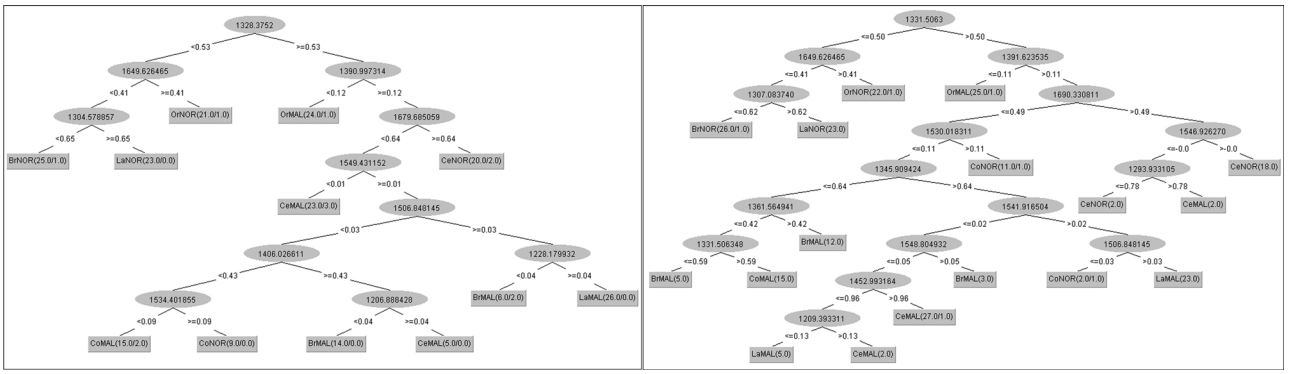

Fig. 5 Decision trees for CART (a) and J48 (b) showing discrimination map for spectra of different tissue types. Ellipses represent intermediate nodes and rectangles final nodes (leaves). The condition for splitting is represented in the branches connecting the nodes. The number of samples classified as a particular class is shown in the rectangle.

instances correctly in such a nonlinear feature space, we have explored nonlinear discrimination algorithms like the decision tree. Decision trees are classifiers that predict class labels for data instances. As mentioned earlier, the performance of multivariate analysis depends on the size of the training dataset. The presence of a large number of variables might mislead the classification and affect classification efficiency. Hence, decision tree algorithms follow selective utilization of important variables (in this case, Raman shift wave number) in data for binary discrimination. This is essentially based on a series of if–then statements that, when applied to an instance (in this case, a spectra) in a data set, results in its classification. Decision trees classify instances by sorting them down the tree from the data mining root node to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test rule for some attribute (in this case, Raman shift wave number) of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. Thus, decision trees selectively utilize important variables (Raman shift wave number) in data for binary discrimination.

The algorithms that are used for constructing decision trees work by choosing a variable at each step that is the next best variable to use in splitting the set of items. "Best" is defined by how well the variable splits the set into subsets that have the same value of the target variable. Different algorithms use different formulas for measuring best (e.g., entropy function, gini index, and classification error method). These formulas are applied to each candidate subset, and the resulting values are combined (e.g., averaged) to provide a measure of the quality of the split.

Table 5 Results obtained using nonlinear multivariate analysis methods (total number of misclassified instances observed using different discriminant methods is shown using roman numbers, while each misclassified instance type is mentioned in parentheses).

	No. of misclassified instances (class of misclassified instance)		
Original class (number of spectra)	CART	J48	RFM
CeM (31)	III (CeN, CeN, CoM)	I (CoN)	–
CeN (21)	I (CeM)	I (CoN)	–
CoM (15)	–	–	–
CoN (11)	II (BrM, BrM)	–	–
BrM (21)	I (OrM)	I (CeM)	–
BrN (25)	–	–	–
OrM (24)	–	–	–
OrN (22)	I (OrM)	I (OrM)	–
LaM (28)	II (CeM, CeM)	–	–
LaN (25)	II (BrN, OrN)	II (BrN, OrN)	–
Classification efficiency	94.618% (211/223)	97.31% (217/223)	100% (223/223)

Table 6 A summary of the efficiency of all the classifiers used in the study.

Method	Efficiency (%)
Limit test	
Best case (Breast Normal)	100
Worst case (Cervix Malignant)	71
FDA	95.07
PLS-DA	95.52
CART	94.618
J48	97.3094
RFM	100

The algorithm used for CART utilizes gini index $(1 - \sum_j p_j^2)$ as a classification criterion, where p_j represents the proportion of instances from different classes at each node. The tree generated through CART is shown in Fig. 5(a). CART led to a high number of misclassifications. Following are the details of misclassification: three samples of cervix malignant were misclassified, two as cervix normal and one as colon malignant, one sample of cervix normal was misclassified as cervix malignant, two colon normal as breast malignant, one oral normal as oral malignant, two larynx malignant as cervix malignant, one larynx normal as breast normal, and one larynx normal as oral normal. In this case, the total misclassifications are 12 (Table 5). The overall success rate is 94.6%, i.e., 211 spectra of 223 spectra are correctly classified. But the point to be noted here is that, although the number of misclassifications is quite similar to that obtained for FDA or PLSDA, they spanned across many classes of samples, indicating poor sensitivity of this methodology.

The other decision tree method J48 utilizes entropy function $(\sum_j -p_j \log_2 p_j)$ as classification criteria, where p_j again represents the proportion of instances from different classes at each node. In this case, misclassified instances included: one sample of cervix malignant misclassified as colon normal, one cervix normal as colon normal, one breast malignant as cervix malignant, one oral normal as oral malignant, one larynx normal as breast normal, and one larynx normal as oral normal. The tree generated through J48 is shown in Fig. 5(b). Decision trees made using the J48 method perform better than CART as it accurately classifies 217 of 223 samples (Table 5). Like CART analysis, J48 also shows poor specificity as it misclassifies a number of classes, but these results are comparable with results of other discrimination methods such as FDA and PLSDA as shown in Table 4.

The results for RFM suggest that efficiency of decision trees increases significantly by aggregation. RFM uses bagging or boosting which allows accurate classification of data. In this study, the RFM model is built with 10 constituent trees. The results using this model which has no misclassified instances as shown in Table 5. Trees are not shown here for RFM, as it generates a forest of trees to classify different classes.

4.6 *Comparative Note*

The main aim of this study is to evaluate the performance of various multivariate statistical tools with a view to provide robust diagnostic results. Hence, dealing with a complex problem like discrimination of multiple-cancer dataset, should provide a more reliable merit about performance efficiency of various discriminant algorithms. We have compared the performance of different discrimination methods to evaluate the specificity of Raman spectral models of normal and malignant tissues of five different types. A concise summary of the results is shown in Table 6. In the case of linear discrimination models, PCA shows a wide range of classification efficiency with minima of 71% (cervix malignant) and maxima of 100% efficiency (breast normal, oral malignant). All the other linear methods show a classification efficiency of ≥95%. For nonlinear discrimination models, variable efficiency is observed when different algorithms are used to build decision trees. The accuracy of decision trees can be improved using aggregation. In fact we have obtained 100% efficiency using the Random forest method.

We have not considered simple accuracy percentage as the sole criterion for deciding the efficacy of a method. The number of classes accurately predicted is a weightier criterion as compared to simple accuracy percentage. Since FDA and PLSDA both misclassify five classes as compared to seven in the case of CART (Table 5), CART is considered to be less efficient as compared to FDA and PLSDA even though the number of misclassified instances are comparable. The classification results for J48 *vis a vis* FDA and PLSDA are fairly comparable.

It seems simple classifiers belonging to discrimination methods work well as compared to complex methods like decision trees (CART and J48) as conclusions reached in our study are based on very small datasets by machine learning standards. As more data become available we expect to observe an improvement in the performance of CART and J48 in prediction accuracy, but the ability to classify different classes may not improve any further.

In addition to accuracy, there are other factors which contribute to the merits of a given classifier. These include simplicity and insights gained into predictive structure of the data. Though discrimination methods such as FDA or PLSDA show satisfactory results, they are unable to handle interactions between different variables. Also these are black boxes giving very little insight into the structure of the data. By contrast, decision trees are able to exploit and reveal the relationship between variables. Trees are easy to interpret and provide information about the relationship between predictor variables and responses by performing stepwise selection of variables.

All these discriminant methods have advantages as well as disadvantages. But application of multiple discriminant methods (both linear and nonlinear) utilizing their significant features to develop a single discrimination model will help improve classification. For example, PCA scores of factors can be fed first to a decision tree algorithm as a training dataset. The decision tree model for discrimination thus developed can be used to enumerate the factors participating in the model. These factors contribute to the variation in data that is important for classification. PCA scores of these enlisted factors can be used further to process FDA. The model thus generated will be more robust because of supervised selection of PCA scores instead of selection

based only on variance, irrespective of their role in classification. Though this method overrides the problem of over-fitting caused by PCA factors selection for the FDA model, it suffers from the fact that the model generated lacks linearity which is present in the spectral data. Many such prospects are available to utilize the spectral data in a wiser manner.

In conclusion, this study supports the applicability of Raman spectroscopic models in two different aspects: (1) for efficient discrimination of tissues even in a multicancer scenario thus strengthening the confidence regarding the power of Raman spectroscopy for objective classification with the help of multivariate tools, and also (2) for development of a single platform spectral library which can be utilized for diagnostic purposes instead of following different spectral models developed individually, thus simplifying the process.

Acknowledgments

We would like to acknowledge our clinical collaborators Dr. Jacob Kurien, Dr. Stanley Mathew, Dr. M. S. Vidyasagar, and Dr. Pralhad Kustagi for providing samples and clinical support in the evaluation of results. We would also like to acknowledge the funding agencies DST (Project No. SP/S2/L04/2001), DAE-BRNS (Project No. 2003/34/17/BRNS/1903), and ICMR (Project No. 5/13/ 23/2003-NCD-III) for funding the projects related to oral, cervical, and breast cancer, respectively. Our sincere thanks to Dr. K. Kalyan Kumar, Dr. M. V. P. Chowdhary, Dr. Mahidhar Kodali, and Ms. Malini who recorded spectra and developed standard models for different cancers. Editorial makeover of the manuscript by Dr. A. Bagwe, SCOPE Cell, ACTREC is gratefully acknowledged.

References

1. A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer Statistics 2009," *CA Cancer J. Clin.* **59**, 225–249 (2009).
2. C. Kendall, M. Isabelle, F. Bazant-Hegemark, J. Hutchings, L. Orr, J. Babrah, R. Baker, and N. Stone, "Vibrational spectroscopy: a clinical tool for cancer diagnostics," *Analyst* **134**(6), 1029–1045 (2009).
3. A. Nijssen, S. Koljenovic, T. C. Bakker Schut, P. J. Caspers, and G. J. Puppels, "Towards oncological application of Raman spectroscopy," *J. Biophoton.* **2**(1–2), 29–36 (2009).
4. K. Moghissi, M. R. Stringer, and K. Dixon, "Fluorescence photodiagnosis in clinical practice," *Photodiagn. Photodyn. Ther.* **5**(4), 235–237 (2008).
5. R. Manoharan, Y. Wang, and M. S. Feld, "Histochemical analysis of biological tissues using Raman spectroscopy," *Spectrochimic Acta, Part A* **52**, 215–249 (1996).
6. E. B. Hanlon, R. Manoharan, T. W. Koo, K. E. Shafer, J. T. Motz, M. Fitzmaurice, J. R. Kramer, I. Itzkan, R. R. Dasari, and M. S. Feld, "Prospects for *in vivo* Raman spectroscopy," *Phys. Med. Biol.* **45**(2), R1–59 (2000).
7. Y -K. Min, T. Yamato, E. Kahada, T. Ito, and H. Hamaguchi, "Evaluation of pancreatic cancer with Raman spectroscopy in a mouse model," *J. Raman Spectrosc.* **36**, 73–76 (2005).
8. H. H. Mantsch, L. P. Choo-Smith, and R. A. Shaw, "Vibrational spectroscopy and medicine: an alliance in the making," *Vib. Spectrosc.* **30**, 31–41 (2002).
9. R. Malini, K. Venkatakrishna, J. Kurien, K. M. Pai, L. Rao, V. B. Kartha, and C. M. Krishna, "Discrimination of normal, inflammatory, premalignant, and malignant oral tissue: a Raman spectroscopy study," *Biopolymers* **81**(3), 179–193 (2006).
10. K. Maheedhar, B. M. Vadhiraja, P. Kushtagi, M. S. Vidyasagar, D. J. Fernandes, V. B. Kartha, and C. Murali Krishna, "Evaluation and validation of Raman spectroscopic diagnostic methodology of cervix cancers by blind studies," *Intl. J. Rad. Oncol. Biol. Phys.* **69**, S400–S400 (2007).
11. C. Murali Krishna, N. B. Prathima, B. M. Vadhiraja, R. Malini, D. J. Fernandes, M. S. Vidyasagar, and V. B. Kartha, "Raman spectroscopy studies for diagnosis of cancer in human uterine cervix," *Vib. Spectrosc.* **41**, 136–141 (2006).
12. M. V. Chowdary, K. K. Kumar, J. Kurien, S. Mathew, and C. M. Krishna, "Discrimination of normal, benign, and malignant breast tissues by Raman spectroscopy," *Biopolymers* **83**(5), 556–569 (2006).
13. C. Murali Krishna, J. Kurein, S. Mathews, K. K. Kumar, and M. V. P. Chowdary, "Raman spectroscopy of breast tissues," *Expert Rev. Mol. Diagn.* **8**, 149–166 (2008).
14. K. K. Kumar, M. V. P. Chowdary, S. Mathew, C. Murali Krishna, and J. Kurien, "Raman spectroscopic diagnosis of breast cancers: Evaluation of models," *J. Raman Spectrosc.* **39**, 1276–1282 (2008).
15. M. V. Chowdary, K. K. Kumar, K. Thakur, A. Anand, J. Kurien, C. M. Krishna, and S. Mathew, "Discrimination of normal and malignant mucosal tissues of the colon by Raman spectroscopy," *Photomed. Laser Surg.* **25**(4), 269–274 (2007).
16. P. Pujary, K. Maheedar, C. Murali Krishna, and K. Pujary, "Classification of Normal and Malignant laryngopharyngeal Tissues by Raman Spectroscopy: A Pilot study," UICC World Cancer Congress, Geneva, Switzerland (2008).
17. K. K. Kumar, A. Anand, M. V. P. Chowdary, K. J. Kurien, C. Murali Krishna, and S. Mathew, "Discrimination of normal and malignant stomach mucosal tissues Raman spectroscopy: A pilot study," *Vib. Spectrosc.* **44**, 382–387 (2007).
18. K. Maheedhar, R. A. Bhat, R. Malini, N. B. Prathima, P. Keerthi, P. Kushtagi, and C. Murali Krishna, "Diagnosis of ovarian cancer by Raman spectroscopy: a pilot study," *Photomed. Laser Surg.* **26**, 83–90 (2008).
19. R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Correction to the description of standard normal variate (SNV) and de-trend (DT) transformations in practical spectroscopy with applications in food and beverage analysis," *J. Near Infrared Spectrosc.* **1**, 185–186 (1993).
20. D. Pena, "Multivariate data analisis" (in Spanish), McGraw Hill-Interamericana de Espana S.A.U. (2002).
21. S. Chevallier, D. Bertrand, A. Kohler, and P. Courcoux, "Application of PLS-DA in multivariate image analysis," *J. Chemom.* **20**, 221–229 (2007).
22. R. A. Eisenbeis and R. B. Avery, *Discriminant Analysis and Classification Procedures: Theory and Applications*, Heath, Lexington, MA (1972).
23. D. Kleinbaum, L. Kupper, and K. Muller, *Applied Regression Analysis and Other Multivariate Methods*," 2nd ed., PWS-Kent, Boston (1988).
24. D. Bertrand, "SAISIR (2010). Package of function for chemometrics in the MATLAB (R) environment" Unité Biopolymères, Interactions, Assemblages, INRA, Nantes, France, http://easy-chemometrics.fr/index.htm.
25. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "*Classification and Regression Trees*," CA, Wadsworth, Belmont (1984).
26. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publ., San Mateo, CA (1993).
27. L. Breiman, "Random Forests," *Mach. Learn.* **45**(1), 5–32 (2001).
28. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.* **11**(1), 10–18 (2009).
29. H. L. Mark, "Use of Mahalanobis distances to evaluate sample preparation methods for near-infrared reflectance analysis," *Anal. Chem.* **59**(5), 790–795 (1987).
30. PLSplus/IQ User's Guide, Galactic Industries Corporation, Salem, New Hampshire (1999).