Optical pathology using oral tissue fluorescence spectra: classification by principal component analysis and *k*-means nearest neighbor analysis

Sudha D. Kamath
K. K. Mahato
Center for Laser Spectroscopy
KMC Life Sciences Center
Manipal Academy of Higher Education
Manipal 576 104 India
E-mail: kkmahato@gmail.com

Abstract. The spectral analysis and classification for discrimination of pulsed laser-induced autofluorescence spectra of pathologically certified normal, premalignant, and malignant oral tissues recorded at a 325-nm excitation are carried out using MATLAB@R6-based principal component analysis (PCA) and *k*-means nearest neighbor (*k*-NN) analysis separately on the same set of spectral data. Six features such as mean, median, maximum intensity, energy, spectral residuals, and standard deviation are extracted from each spectrum of the 60 training samples (spectra) belonging to the normal, premalignant, and malignant groups and they are used to perform PCA on the reference database. Standard calibration models of normal, premalignant, and malignant samples are made using cluster analysis. We show that a feature vector of length 6 could be reduced to three components using the PCA technique. After performing PCA on the feature space, the first three principal component (PC) scores, which contain all the diagnostic information, are retained and the remaining scores containing only noise are discarded. The new feature space is thus constructed using three PC scores only and is used as input database for the *k*-NN classification. Using this transformed feature space, the centroids for normal, premalignant, and malignant samples are computed and the efficient classification for different classes of oral samples is achieved. A performance evaluation of *k*-NN classification results is made by calculating the statistical parameters specificity, sensitivity, and accuracy and they are found to be 100, 94.5, and 96.17%, respectively.
[DOI: 10.1117/1.2437738]

Keywords: oral tissue; laser-induced fluorescence; principal component analysis; *k*-nearest neighbor.

Paper 06182R received Jul. 5, 2006; revised manuscript received Oct. 7, 2006; accepted for publication Oct. 23, 2006; published online Mar. 2, 2007.

1 Introduction

Cancers of the oral cavity represent a major health problem, as indicated by their high incidences in many parts of the world.[1] This cancer is one of the 10 most common cancers all over the world with more than 500,000 new cases projected world wide annually.[2] In India, there is a high incidence of oral cancer accounting for as much as 50% of all cancers.[3] It predominantly occurs among males with a male:female ratio ranging from 2 to 10 for various intraoral sites. The incidence and mortality from oral cancer have either been stable or increasing over the past three decades in several countries, with the increase being more striking particularly among young men in the western and eastern European regions.[4] According to the Indian cancer registry,[4–6] malignant tumors of the lip, oral cavities, and pharynx are the most commonly cited group of cancers. Between 90 and 95% of all oral cancers arise from the cells that line the mouth. Despite the easy accessibility of the oral cavity for examination, there is no satisfactory mechanism to adequately screen and detect premalignant changes and early lesions in the upper aerodigestive tract.[7,8] Most oral lesions are usually ignored and are not treated until they are advanced or have metastasized. At this advanced stage, the effectiveness of chemotherapy, radiotherapy, and surgery or combinations of these modalities have been disappointing. Moreover, successful therapy of oral cancer has been significantly hindered by the subsequent development of secondary tumors, which leads to treatment failure and death. Therefore, early detection of neoplastic changes in the oral cavity may be the best approach to improve results of therapy and survival rates.

Oral cancer is one among the few human cancers with a vast potential for prevention. Several forms of tobacco use (chewing, smoking, reverse smoking) and alcohol drinking

Address all correspondence to: K. K. Mahato, Center for Laser Spectroscopy, KMC Life Sciences Center, Manipal Academy of Higher Education, Manipal, 576 104, India; Tel: +91-820-2922526; Fax: +91-820-2571919, 2570062; E-mail: kkmahato@gmail.com

1083-3668/2007/12(1)/014028/9/$25.00 © 2007 SPIE

are most causative factors for oral cancer.[6] Because of the widespread usage of tobacco, the incidence of oral cancer is quite high in developing countries. Researchers in oral cancer agree that the early detection of oral carcinoma greatly increases the probability of cure with minimum impairment and deformity. A recent review,[6] pointed out that adults above the age of 40 years should undergo regular oral cancer examinations as part of their regular oral and dental health checkups. Routine oral visual examinations can help in early detection of oral cancer, enabling interventions that contribute to reduced morbidity and/or improved survival. Primary prevention of oral cancer, which involves reducing the exposure to tobacco, betel quid, and alcohol, has been shown to be effective in controlling the oral cancer incidence rates. Secondary prevention involves screening for the early detection.[4,7–9] However, the facilities for early detection of oral lesions in developing countries, in terms of well-equipped clinics with qualified clinicians and pathologists, are sparse and awareness about screening for early detection is almost nil. A technique that is fast; requires only minimally trained technicians to operate; uses relatively reasonable, portable equipment; and that can objectively evaluate in a community screening program the state of an oral lesion as inflammatory, premalignant or malignant, can speed up the identification of cases that require professional medical/hospital examination for follow-up. The laser-induced fluorescence (LIF) method with statistical/mathematical data processing fulfills these requirements.

All oral lesions are not necessarily premalignant or malignant.[7,8] The development of a noninvasive and accurate method for real-time screening and diagnosis of oral cavity lesions would have great potential to improve early detection of neoplastic changes.

LIF spectroscopy and fluorescence imaging have been studied as potential noninvasive diagnostic tools for differentiating normal and neoplastic oral tissues.[10] Compared with several other cancer diagnosis tools, the autofluorescence technique has a high specificity and sensitivity for discrimination between the diseased and nondiseased tissue, and also has the advantage of being noninvasive and producing a real-time diagnosis.[11] Recently, fluorescence spectroscopy has been extended to the medical field to characterize various metabolic and pathological changes at the cellular and tissue level because it is the most sensitive method for monitoring minor changes in the structure and microenvironment of a fluorophore.[12] The technique involves illumination of tissue with monochromatic light and recording the fluorescence spectrum. LIF can utilize the naturally occurring tissue fluorophores (autofluorescence) or exogenous fluorophores.[13–24] Autofluorescence techniques offer a number of advantages, including avoidance of potential side effects of added dyes or drugs. The fluorescence spectral profile of the tissues provides information not only about their architecture (epithelial thickness), and organization, but also about their metabolic state and concentrations of fluorescing molecules, which can be correlated to histological changes. The alterations in tissue architecture (epithelial thickness) and cellular composition induced by processes such as dysplasia and inflammation are reflected in variations in the optical properties of human tissue. Any alterations in tissue architecture (epithelial thickness) that inhibit the ability of excitation photons to reach the natural fluorophores or of the fluorescence emission photon to escape from the tissue and be detected by the spectroscopic system would affect the fluorescent signature. Additionally, changes in the concentration or form of natural fluorophores such as collagen would alter the emitted fluorescence. Thus, by measuring the fluorescence signal from a tissue, which is different for different epithelial thicknesses and cellular compositions, information concerning a disease condition can be obtained.[25] The spectroscopy technique has the added advantage that it can be repeatedly applied *in situ*/*in vivo* without the deleterious effects of repeated biopsy.

There are very few[22,23,26–31] systematic studies on the effectiveness of discriminating between premalignant and malignant conditions in oral tissue by LIF. One aim of our studies was to see whether this could be done successfully. The results discussed in the following show that LIF could successfully discriminate among normal, premalignant [oral submucous fibrosis (OSMF)], and malignant situations.

OSMF is an insidious chronic disease affecting any parts of the oral cavity and sometimes the pharynx. Although occasionally preceded by and/or associated with vesicle formation, it is always associated with a juxtaepithelial inflammatory reaction followed by a fibroelastic change of the lamina propria, with epithelial atrophy leading to stiffness of the oral mucousa and causing trismus and inability to eat.[32,33]

To identify a better classifier technique for the classification of oral data for optical pathology, comparative evaluation of different classifier techniques using same set of spectral data is necessary. Keeping this in mind, in the present analysis using MATLAB principal component analysis (PCA)-based *k*-means nearest neighbor (*k*-NN) classifier technique, we used the same set of spectral data that have been used in earlier analysis using PCA and artificial neural network (ANN) classifier techniques[27] for its comparative evaluation. The main purpose of our studies is to address these aspects in screening for early detection of oral malignancy by LIF.

In our laboratory, we studied the fluorescence spectra of normal as well as different stages of malignant oral tissues and developed fluorescence spectroscopy techniques for optical pathology.[18,21,27] In this study, we have classified the fluorescence spectra recorded at 325-nm excitation from pathologically certified normal, premalignant, and malignant oral tissue using the PCA-based *k*-NN technique. These spectral samples are the same as those used in Ref. 27. The input feature space selection for *k*-NN classification is done using the PCA technique and a cluster-space classification is developed, which implements the *k*-NN method efficiently and makes it feasible for dimensional data classification.[34–36] We used the nearest neighbor (NN) classification technique, which classifies an unknown sample to a class that has the most "similar" or "nearest" sample point in the training set of data. The results in this study show that the discrimination of normal, premalignant, and malignant oral conditions can be achieved quite successfully by LIF. The results obtained in this study are presented and discussed in this paper.

2 Materials and Methods

2.1 *Sample Collection and Handling*

Biopsied normal, premalignant, and malignant oral tissue samples were obtained from the college of Dental Surgery and

Table 1 Sample details.

Spectrum No.	Sample Type	Mean Age	Histopathology	Patient Habit	Spectroscopy
1–20	Normal, standard set	60±8.5	Uninfected area-normal epithelial cell	No tobacco habit	Normal
21–40	Malignant, standard set	65±5	Cancer of buccal mucosa-squamous cell carcinoma	Chewing tobacco and smoking	Malignant
41–60	Premalignant, standard set	61±10.5	Oral submucous fibrosis	Smoking cigarettes	Premalignant
61–100	Normal, test set	60±8.5	Uninfected area-normal epithelial cell	No tobacco habit	Normal
101–137	Malignant, test set	65±5	Cancer of buccal mucosa-squamous cell carcinoma	Chewing tobacco and smoking	Malignant
138–143	Premalignant, test set	61±10.5	Oral submucous fibrosis	Smoking cigarettes	Premalignant

Sai Baba Cancer Hospital, KMC Manipal Academy of Higher Education (MAHE), India. Tissues from uninvolved areas from the same subjects were used as healthy controls. A mirror image of each sample was fixed in 10% neutral buffered formalin and was sent for histopathological certification. The samples were kept moist with saline (pH=7.4), and spectra were recorded within half an hour of tissue removal. Samples were mounted on a quartz plate and fluorescence measurements were carried out with 325-nm laser excitation. Whenever necessary, the tissue samples were snap frozen in liquid nitrogen and stored at −80°C for later analysis. Trial runs showed that the spectra remain unchanged at least for 2 h after biopsy, if kept in saline.[18] For each sample, spectra were recorded from several points, separated by few hundred micrometer to few millimeters, and were treated as independent data because of the inhomogeneous nature of the tissue, and in addition, to explore the feasibility in surgical boundary demarcation. Details of the sample preparation and spectral recording are mentioned elsewhere.[18–21] No particular care was taken regarding the orientation of the tissue under study. Biopsied oral tissue subjects used in this study were of different sizes, having their surface area (approximately) in the range of 15 to 25 mm^2 and of thickness in the range 2 to 5 mm. The sample details are given in Table 1.

2.2 *Experimental Setup*

A block diagram of the experimental setup for recording the fluorescence spectra is shown in the Fig. 1. The excitation used in this study is 325 nm at a 10-Hz repetition rate with energy per pulse of 100 to 200 μJ from an Nd-YAG/MOPO/FDO source (Spectra Physics, Quanta Ray, Model: PRO 230 10, MOPO SL). The fluorescence emission from the tissue samples was collected by fiber optic probe, details of which are given elsewhere.[18] In brief, we used a typical homemade seven-fiber probe with a central fiber for excitation and the surrounding six fibers for collection of fluorescence. The individual fibers (Thorlabs FG 200-UEP Fiber type) were of 240 μm diameter with a numerical aperture of 0.25. The excitation end of the fiber was held perpendicular to the tissue surface under study and the distance between the tip of the fiber and the tissue surface was maintained to 2 to 3 mm, as optimized for better fluorescence intensity. To adjust the distance between the tissue surface and the fiber tip and to position the excitation at different sites of a tissue sample, the excitation end of the fiber was mounted on an *XYZ* micrometer translational stage. The output of the fiber was coupled into an imaging spectrograph (Spectra Pro 150 spectrograph 300 g/mm, 300-nm blazed grating) equipped with an Andor intensified CCD (ICCD) system (Andor Technology, Northern Ireland, ICCDBH-501-25F-01) for spectral recording. A time delay of 500 ns provided by a DG 535 delay and gate generator (Stanford Research Systems) was used to compensate for the cable and other delay times. The total fluorescence after the delay was recorded with an optimized 75-ns gate.[18] All spectra were recorded with a slit width of 250 μm (5-nm bandpass) and with an average of 100 laser pulses (10 Hz).

2.3 *Data Analysis*

In this study we used 143 fluorescence spectra (60 normal, 26 premalignant, and 57 malignant) from 19 normal, 5 premalig-

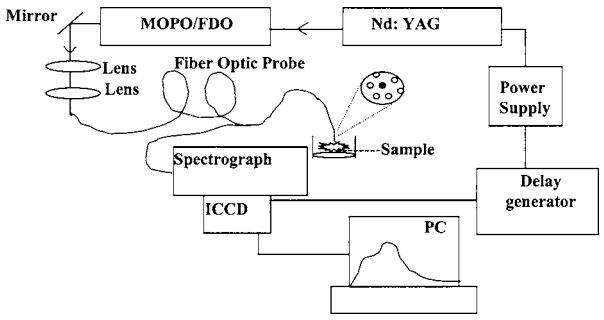

Fig. 1 Schematic of the experimental setup for LIF measurement.

nant, and 19 malignant oral tissue samples. These spectral samples are the same samples, as are used in Ref. 27 for PCA and ANN analysis. In this study, these 143 spectra are treated as independent data for the following reasons. First, in epithelial cancers (before it has penetrated the underlying tissue), often a tissue specimen can have normal and neoplastic regions adjacent to each other.[30,31,37,38] The pathologist, therefore, examines several sites on a biopsy sample and even if only one site shows as abnormal, the sample is considered as abnormal. Second, for *in vivo* applications in screening, surgical boundary demarcation, and optically guided biopsy it is necessary to identify the exact site where malignancy is suspected. The variations in the spectra from site to site will be the deciding factor in this case and they must be treated independently. Third, in epithelial cancers, the abnormal proliferation begins near the basal lamina (dysplasia), which then advances into the epithelial region (carcinoma *in situ*), and later, in the invasive stage, penetrates into the connective tissue region. We have shown[38] by Raman spectroscopy that in biopsied tissue samples of oral cancer, the subepithelial region showed very similar spectra for normal and malignant samples, and only the epithelial spectra discriminated between the two very well for many samples. This means that unless one is sure of the region (epithelial, subepithelial, or connective tissue) one is looking at in a biopsy sample, there can be both normal and malignant types of spectra from adjacent sections of the same sample. This possibility is further supported by the result that in none of the spectra of normal samples (oral, breast, cervix, and ovary systems that we have studied) recorded from different sites of the same specimen, did we observe any anomaly, since normal samples are normal at all sites. Only malignant samples have shown this anomalous behavior,[39] with some sites giving typical normal spectra and other sites of the same sample showing typical malignant spectra, since these samples have the possibility for both types of sites existing together. In view of these points, when there is no *a priori* information available for a sample it is necessary to examine many sites on the sample and treat each site as independent. This is necessary even when one is guided by a visual examination by the clinician for biopsy because of possible errors such as past pointing.[37–44] In this study, we used MATLAB@R6-based PCA and (*k*-NN) analysis to discriminate among normal, premalignant, and malignant samples.

The PCA method is a linear feature transformation technique that preserves the variance of the feature space. This is a classical method, usually employed for the dimensionality reduction of a feature space.[43] When many of the variables in a multidimensional data set have a significant correlation, PCA successfully captures nearly all the information of the original data set in the first few principal components (PCs). PCA mainly has three effects. First, it transforms the data by defining new variables, termed PCs, which are statistically uncorrelated. Second, it orders the resulting PCs in such a way that those possessing larger variances are arranged first. And third, in many problems, the first few components capture most of the discrimination capability. In such cases, only the first few components must be considered.

PCA is applicable for data analysis problems in several domains where the data elements are significantly correlated. In PCA, the PCs are arranged in the order of their contribution to the variance of the entire spectral data set. The first PC accounts for as much of the variability in the data set as possible, and each succeeding component accounts for the next largest amount of variation and is independent of the previous PC. Each new PC is orthogonal to the previous PC and contains the maximum information unexplained by the previous one. Thus, PCA processing can dimensionally reduce the variables of the original spectral data into a small number of informative PCs that fully describe the variations in the spectral data within the limitations of noise. For doing PCA, the six features mean, median, standard deviation, energy, maximum intensity, and the spectral residuals are extracted from each spectrum and MATLAB@R6 (Refs. 45 and 46) based PCA is performed on the reference database. Applying PCA to the feature space matrix, the original data are transformed into a set of PC scores. Since most of PCs account mainly for noise and do not provide diagnostic information, it is necessary to identify a small group of informative PCs to build the algorithms for classification of oral spectra of different conditions (normal, malignant, and premalignant). Including a minimal number of PCs reduces the complexity of a diagnostic algorithm. The contribution of each PC to the total variance of spectral data is proportional to its eigenvalue. High-order PCs often account for less than 1% of the total variation and represent mostly noise[44] only. In this paper, PCs that have a variation of more than 1% variance in the spectral data are considered to be informative PCs.

In PCA, first the calibration standards for each class of the samples (normal, premalignant, and malignant) are prepared using pathologically certified normal, premalignant, and malignant, samples. In this case, we randomly chose 20 spectra each from the normal, premalignant, and malignant groups of spectra and the calibration sets were prepared. Once the calibration standards were established, any new sample could be classified as normal, premalignant, or malignant by comparing it with the standard models for normal, premalignant, and malignant obtained from pathologically certified samples.

The standard calibration sets were optimized by removing outliers from the model set using cluster analysis and selecting an optimum number of factors (PCs) required for the dimensional data analysis.[47,48] Including outlier samples in the training set introduces a bias to the final model.[47,48] The cluster plot for detecting outliers in the calibration model is generally plotted between the PC scores of the first two PCs because the first two PCs represent the plane of best fit through the data.

After the PCs with diagnostic information were identified, a *k*-NN method was used to build the algorithm for classification of the oral tissue spectra of different pathological conditions (normal, malignant, and premalignant). To construct algorithms for *k*-NN classification, only those PCs that have significantly different projection scores for normal, malignant, and premalignant tissues were selected. The spectral samples used in *k*-NN analysis are the same set of normal, premalignant and malignant samples as used in PCA.

Nearest neighbor methods provide an important data classification tool for recognizing object classes in pattern recognition domains.[34–36,49,50] This method classifies an unknown sample to that class having most "similar" or "nearest" sample point in the training set of data. The nearest sample was found by using concept of distances or metrics known as

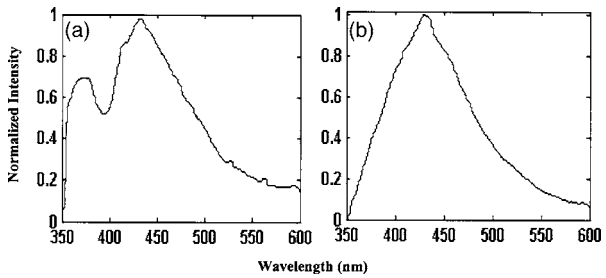

Fig. 2 Typical background subtracted, smoothened, and normalized fluorescence spectra of (a) normal and (b) malignant oral tissue.

Table 2 Eigenvalues and percent variance values of a standard set of 60 (20 normal, 20 malignant, and 20 premalignant).

Factor No.	Eigenvalues	Total Percent Variance	Commulative Variance (%)
1	6.5524	95.79	95.79
2	0.1913	2.79	98.58
3	0.0934	1.36	99.94

Euclidean distance. Euclidean distance is the square root of sum of the square of the x distance plus the square of the y distance for a given pair of cases x and y. Once the Euclidean distance is identified, the k-NN classifies an unknown sample to that class having most similar or nearest sample point in the training set of data. Nearest neighbor methods can detect a single or multiple numbers of nearest neighbors.[36,49–51] In this study, the single nearest neighbor method was used to classify oral lesions including different pathological conditions (normal, malignant, and premalignant).

3 Results and Discussion

A typical plot of normal and malignant oral tissue fluorescence spectra recorded at a 325-nm excitation is shown in Fig. 2. As mentioned in the data analysis section, in this study 143 fluorescence spectra (60 normal, 26 premalignant, and 57 malignant) from 19 normal, 5 premalignant, and 19 malignant oral tissue samples were used and discrimination analysis was performed using the MATLAB@R6 algorithm-based PCA and k-NN analysis.

3.1 *PCA*

In this classification problem, to reduce the dimensionality, we used six features, each from 143 sample spectra (60 calibration + 83 test) on which PCA was performed. Twenty spectra each, randomly chosen from certified normal, premalignant, and malignant tissue samples were combined to discern the best approach to prepare calibration sets in the three classes. Six features of 60 training samples belonging to the normal, premalignant, and malignant groups were first pooled together and PCA was performed on this reference database. Since only a few PCs are required to express the large amount of spectral data, the eigenvalues drop to significant quantities after the first few. The spectra of a given set of samples may have contributions only from a limited number of factors, and except for the first few factors, many of the eigenvalues may be close to zero with no practical contribution to the spectra. A simple plot of eigenvalues can also be used to informally indicate the number of significant components. In our analysis, the eigenvalues, the contribution to total variance, the eigenvectors, and other parameters were used to determine the significant number of factors required for a correct analysis.

Figures 3(a) and 3(b), respectively, show the eigenvalues for the factors and total percentage variance (i.e., total percentage contribution to the variation spectra with increasing number of factors) of 60 standard oral samples (20 normal, 20 malignant, and 20 premalignant). The plots indicate that three factors are sufficient to explain 99.94% of the variance in the total data set. And thus the feature vector of length 6 could be reduced to three components. With many types of data sets, it is common for the first few PCs to possess most of the variance of the original data. Table 2 shows the variance associated with first three PCs. We notice that first PC itself covers over 95.79% of variance. From Table 2, we can see that the eigenvalues decrease very rapidly and are almost zero after three factors. Therefore, in the present analysis, the first three PCA components, which show difference in total percent variance values among normal, premalignant, and malignant cases, were retained for further analysis and remaining higher order PCs were discarded. Thus, the feature space, a 6 $\times$ 143 matrix, was reduced dramatically to a 3 $\times$ 143 matrix,

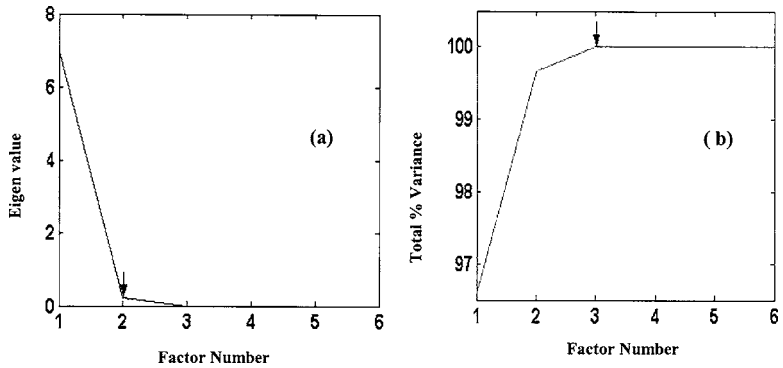

Fig. 3 Plots of calculated (a) eigenvalues and (b) total percent variance of a PCA decomposition of the 60 oral tissue (20 normal, 20 malignant, and 20 premalignant) spectral data.

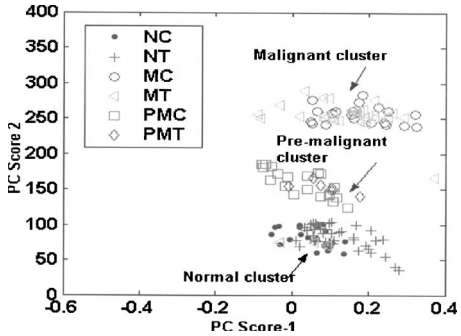

Fig. 4 Cluster/scatter plot in log mode between the scores of first two PCs of 143 spectra of oral normal, malignant, and premalignant samples: NC, normal calibration; NT, normal test; MC, malignant calibration; MT, malignant test; PMC, premalignant calibration; PMT, premalignant test.

and hence classification became computationally more efficient.

Figure 4 shows a cluster/scatter plot obtained by plotting the scores of the first two PCs of 143 samples [(20+40) normal, (20+37) malignant, and (20+06) premalignant]. We clearly see from the plot that all samples diagnosed as normal, malignant, and premalignant by pathological examination in the calibration set clustered in three distinct regions. This shows that the calibration set of samples used in this study are almost free of outliers. Also, when all test normal, premalignant, and malignant samples are used, almost all normal and premalignant samples are clustered around their respective calibration set centroid vectors. Though some of the malignant and normal test samples are overlapped, malignant samples are scattered more compared to normal and premalignant samples. Only three test malignant samples overlapped with the normal and premalignant regions. The specificity and sensitivity of this technique is thus found to be quite good.

3.2 *k-NN Classification*

NN methods can detect a single or multiple numbers of NNs. A single NN method is primarily suited for recognizing data where we have sufficient confidence in the fact that class distributions are nonoverlapping and the features used are discriminatory. In most practical applications, however, the data distributions for various classes are overlapping and more than one NNs are used for majority voting.[36,49–52] The NN method works well with data where features are statistically independent. From the score plot (Fig. 4), it is evident that the data points of training set of 60 spectra (20 normal, 20 malignant, and 20 premalignant) are clustered in three distinct groups without overlapping. Thus, in this study, we used the single NN method to classify the oral lesions with different pathological conditions (normal, malignant, and premalignant).

3.2.1 *k-NN classifier*

This is a classification scheme used to determine the class of a given sample by its feature space. This is a variant of the NN technique in which a prototype sample is computed from the reference set and a given test sample is classified as belonging to the class of the closest prototype. Here, the prototype sample is computed as the mean of feature vector of the reference set belonging to a particular class. This prototype vector is known as the centroid vector.

In cluster analysis, clusters are generated that maximize the distance between the centers of clusters; a centroid is the value for all the objects in the cluster.[34–36,49–52] The centroid method requires a full set of coordinates to be presented for all of the objects to be classified. It calculates the centroid coordinates of each cluster, then the Euclidean distances between each pair of centroids. The pair with the least distance is merged before proceeding to the next iteration.

3.2.2 *Feature selection for k-NN classification*

As already explained, the NN method works better with data where features are statistically independent. Feature selection involves determining which subset of features best distinguishes among the various object types. For an improvement in the data analysis, in this study we proposed a *k*-NN model that is a slight modification of the standard *k*-NN rule. Usually, in traditional *k*-NN classification, *k* NNs of a test document are computed first. Then the similarities of this document are assigned to the most similar class (as measured by the aggregate similarity). A major drawback in *k*-NN is that it uses all features equally in computing similarities. This can lead to poor similarity measures and classification errors, when only a small subset of the data is useful for classification. By keeping these points in mind, we slightly modified the proposed *k*-NN classification technique and achieved a higher degree of accuracy in classification by using only the independent features in the classification process. The steps involved in the modified *k*-NN classification model are as follows:

1. The six features extracted from each oral tissue spectrum were mean, median, standard deviation, spectral residual, energy, and maximum intensity.
2. Thus, from 143 oral tissue samples (60 training + 83 test), we had a feature space matrix of dimensions 6×143. On this data set, we performed PCA to reduce the dimensionality. Using PCA, the feature vector of length 6 was reduced to three components. Outliers were detected and removed, and a number of factor selection techniques (explained in the PCA section) were performed on this PCA-transformed data set for further data analysis.
3. Out of six PCs, only the first three PCs accounted for over 99.94% of the total data variance, which demonstrates that the significant data variation can be described by a few informative PCs. Thus, these three PCs were retained and the higher order PCs that mainly accounted for random noise, and did not contain diagnostic information were discarded.
4. The PC scores of the first three PCs have had the maximum variance in the data set were used as the input feature space for the *k*-NN classification.
5. The PC scores of the calibration (which were used as the standard set in PCA) set of 60 oral tissues (20 normal, 20 malignant, and 20 premalignant) were used as the calibration feature space for the *k*-NN model and the PC scores of the remaining 83 samples (40 normal, 37 malignant, and 6 premalignant) were used as test samples, and classification was achieved.

6. For any test sample, the three spectral distances (i.e., Euclidean distances), corresponding to "normal centroid," "malignant centroid," and "premalignant centroid" were estimated according to the relations

$$SD_n = \sqrt{\sum_{i=1}(EB_i^t - EB_i^{nc})^2}, \quad (1)$$

$$SD_m = \sqrt{\sum_{i=1}(EB_i^t - EB_i^{mc})^2}, \quad (2)$$

$$SD_{pm} = \sqrt{\sum_{i=1}(EB_i^t - EB_i^{pmc})^2}, \quad (3)$$

where EB_i^t denotes ith feature value of the test sample; EB_i^{nc}, EB_i^{mc}, and EB_i^{pmc}, respectively, denote centroid vectors for the normal, malignant, and premalignant groups; and SD_n, SD_m, and SD_{pm}, respectively, represent the spectral distances of the normal, malignant, and premalignant samples.

3.2.3 *Steps involved in k-means NN classification*

The input is a test sample vector and centroid vectors for normal, malignant, and premalignant samples. The output is the classified sample (normal, malignant, and premalignant). The method is as follows

Step 1. Compute the distance between test sample and normal centroid vector (SD_n) according to Eq. (1). Similarly, compute the distances between test sample and the malignant centroid vector (SD_m) and the premalignant centroid vector (SD_{pm}) using Eqs. (2) and (3), respectively.

Step 2. Classify the test sample as "normal" if $SD_n < SD_m$ and $SD_n < SD_{pm}$. Similarly, classify the sample as "malignant" if $SD_m < SD_n$ and $SD_m < SD_{pm}$. If $SD_{pm} < SD_n$ and $SD_{pm} < SD_m$, classify it as premalignant. This means that the class with the smallest distance from test data is declared the winner, and the test pattern is allocated to this class.

Step 3. Repeat the preceding steps for all given samples to obtain the classified groups.

Step 4. Classification ends.

3.2.4 *Classification results using k-means NN technique*

Once the k-NN was trained with input feature values of 20 normal, 20 malignant, and 20 premalignant samples, the technique was ready to predict any unknown new data. A MATLAB-based program was executed to predict any new data. The classification of test samples was made based on the algorithm already mentioned. In this study, we have features from the same 40 normal, 37 malignant, and 6 premalignant oral tissue spectra as used in PCA. When this classification technique was used on calibration model, all 60 calibration samples (20 normal, 20 malignant, and 20 premalignant) were classified to their respective groups (as clustered in three distinct regions of cluster plot of PC scores). In our analysis using the *k*-NN technique, all 40 normal test spectra were classified as normal, 34 out of 37 malignant spectra were classified as malignant, and all 6 test premalignant spectra were classified as premalignant. The specificity, sensitivity, and accuracy obtained in this classification technique were 100, 94.5, and 96.17%, respectively.

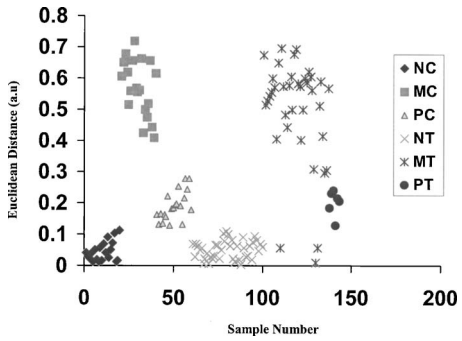

Fig. 5 Plot of the Euclidean distance against sample number for the 143 samples (20+40 normal, 20+37 malignant, and 20+06 premalignant): NC, normal calibration; MC, malignant calibration; PC, premalignant calibration; NT, normal test; MT, malignant test; and PT, premalignant test.

Figure 5 shows the plot of the Euclidean distance against sample number for the 143 samples (20+40 normal, 20+37 malignant, and 20+06 premalignant). In Fig. 5, the "normal centroid" is taken as the reference point for plotting Euclidean distances of all samples. From the plot, it is clear that the Euclidean distances of premalignant samples lie in between the Euclidean distances of normal and malignant samples. The calibration set of samples (normal, premalignant, and malignant) cluster in three distinct groups without any overlap. The test normal and premalignant samples also cluster along with the corresponding calibration set samples, showing 100% discrimination of these samples. Three of the malignant test samples were overlapped with the normal calibration set cluster, and this may be due to from the normal sites of the malignant samples. A closer observation of Fig. 5 shows that only a very small number of malignant samples fall outside the general range of malignant species, indicating the probability of malignant samples being in the respective cluster is about 97% and finding them out of the cluster is only 3%. Thus, the plot clearly shows discrimination of the different classes of oral spectral samples used in this study.

The *k*-NN performance parameters are shown in Table 3 and are compared with the performance parameters of other analyses done on the same set of spectral data using the PCA and ANN analysis.[27] The performance parameters, specificity,

Table 3 Performance evaluation of *k*-NN classification and comparison with performance of other conventional techniques PCA and ANN.

Classifier	Specificity (%)	Sensitivity (%)	Accuracy (%)	Reference
k-NN	100	94.5	96.17	Current study
PCA	100	92.9	96.5	Biopolymers (Ref. 27)
ANN	100	96.5	98.3	Biopolymers (Ref. 27)

sensitivity, and accuracy in PCA match/no match analysis were 100, 92.9, 96.5%, respectively, whereas in case of ANN analysis, these parameters were 100, 96.5, 98.3%, respectively.[27] In this study, with the MATLAB PCA-based *k*-NN analysis, the specificity, sensitivity, and accuracy are found to be 100, 94.5, and 96.17%, respectively. When compared in terms of sensitivity, ANN and MATLAB-based *k*-NN techniques are found slightly better than the conventional PCA (match/mismatch) technique. And when compared in terms of accuracy, PCA and *k*-NN analyses were similar in nature, whereas ANN analysis is found to be better than PCA and *k*-NN analyses.

As already noted, we showed that by forming standard calibration sets of pathologically certified samples of normal, premalignant, and malignant tissue samples one can match a new sample with these standard sets. This leads to a more reliable diagnosis and makes it possible to do periodic screening, monitoring of effectiveness of therapy, surgical boundary check, and early detection of recurrence, if any, without the necessity of repeated biopsy. The preparation of the standard sets requires only a central well-equipped hospital facility. Once the standard sets are ready, they can be made available to any clinical setup even in small hospitals. With a relatively inexpensive instrument, the screening process can thus be carried out across large cross sections of the rural population, for whom such screening will not be easily available otherwise. To the best of our knowledge, none of the earlier workers have done this *k*-NN analysis on oral tissue fluorescence spectra.

The great advantage of the optical method combined with statistical data analysis is that cases that show anomalous behavior can be reevaluated with a second scan with more sampling points. Furthermore, because optical methods do not require tissue removal, they may provide better tools for experts to direct diagnostic biopsies, screen for early detection, and determine tumor margins.

4 Conclusions

The characterization of native fluorescence spectroscopic properties of oral tissue spectra in different pathological conditions such as normal, malignant, and premalignant were studied using the PCA-based *k*-NN technique. The detailed statistical analysis (MATLAB PCA-based *k*-NN technique) of the fluorescence recorded data at 325 nm excitation enabled us to extract the diagnostic information from the measured native fluorescence spectra of both normal and diseased subjects. To be more precise, in this technique, an unknown sample is classified into a diagnostic category (group) by the computer program, which minimizes the rate of misclassification. Our results suggest that the presented classification technique achieves objective discrimination among normal, premalignant, and malignant oral tissues with high specificity and sensitivity. On the basis of these observations, it is obvious that the presented technique discriminates oral malignancy effectively and hence it can be used as an alternative or complementary technique to the other existing conventional methods of disease diagnosis. The small time required to acquire and analyze the fluorescence spectra together with the high rates of success prove that the method is very attractive for real-time applications.

As a preliminary investigation, we studied a few cases in each group. In this regard, more detailed investigations on a large population consisting of normal, malignant, and premalignant subjects are essential to improve specificity and sensitivity of the presented technique for the characterization of various pathological conditions. Furthermore, such investigations with large groups of experimental subjects will also be useful for the development of a statistical database and user-friendly diagnostic algorithm that could facilitate fast screening of patients for early detection of malignant and nonmalignant subjects. Overall, our results indicate great promise for fluorescence-spectroscopy-based detection of carcinoma in oral tissues, which would be a great improvement in guiding biopsies and diagnosing and classifying tissues in different pathological conditions.

Acknowledgments

The authors are thankful to the Manipal Academy of Higher Education, India, for the use of different facilities to carry out this study, and to Mr. N. Subramanya and Mr. B. K. Manjunath for recording some of the spectra used here. Our thanks are also due to Prof. Keerthilatha M. Pai, College of Dental Sciences, MAHE, Manipal, and Dr. Satadru Ray, Surgical Oncology, KMC, MAHE, Manipal, for providing the oral tissues used in this study, to Dr. B. R. Krishnanand for pathology support, and to Dr. V. B. Kartha, Senior Scientist, Retired, Centre for Laser Spectroscopy, MAHE, Manipal, for his fruitful suggestions.

References

1. W. Chih-Yu and T. Tsuimin, "PLS-ANN based classification model for oral submucous fibrisis and oral carcinogenesis," *Lasers Surg. Med.* **32**, 318–326 (2003).
2. D. Saranath, "*Cancers of the oral cavity*," in *Carcinogenecity Testing, Predicting and Interpreting Chemical Effects*, K. T. Kitchin, Ed., pp. 653–675, Marcel Dekker, New York (1992).
3. P. N. Notani, "Global variation in cancer incidence and mortality," *Curr. Sci.* **81**, 465–474 (2001).
4. R. Shankarnarayan, "Health care auxiliaries in the detection and prevention of oral cancer," *Oral Oncol.* **33**, 149–154 (1997).
5. National cancer registry programme. Biennial report 1988–1989, "Indian Council of Medical Research," New Delhi (1992).
6. L. P. Lauren, "The effectiveness of community-based visual screening and utility of adjunctive diagnostic aids in the early detection of oral cancer," *Oral Oncol.* **39**, 708–723 (2003).
7. J. B. Epstein, L. Zhang, and R. Miriam, "Advances in the diagnosis of oral premalignant and malignant lesions," *J. Can. Dent. Assoc.* **68**, 616–621 (2002).
8. M. Partridge, "Oral cancer: 2. Clinical presentation and use of new knowledge about the biology of cancer to establish why tumors may recur," *Dent. Update* **27**, 288–294 (2000).
9. B. K. Joseph, "Oral cancer: prevention and detection," *Med. Princ. Pract.* **11**(Suppl. 1), 32–35 (2002).
10. G. D. C. De Veld, M. J. H. Witjes, H. J. C. M. Sterenborg, and J. L. N. L. Roodenburg, "The status of *in vivo* autofluorescence spectroscopy and imaging for oral oncology," *Oral Oncol.* **41**, 117–131 (2005).
11. R. Nirmala, J. X. Chen, K. Gossage, R. K. Rebecca, and B. Chance, "Fast and noninvasive fluorescence imaging of biological tissues *in vivo* using a fly-spot scanner," *IEEE Trans. Biomed. Eng.* **48**, 1034–1041 (2001).
12. S. Madhuri, N. Vengadesan, P. Aruna, D. Koteeswaran, P. Venkatesan, and S. Ganesan, "Native fluorescence spectroscopy of blood plasma in the characterization of oral malignancy," *Photochem. Photobiol.* **78**, 197–204 (2003).

13. C. T. Chen, H. K. Chiang, S. N. Chow, C. Y. Wang, Y. S. Lee, J. C. Tsai, and C. P. Chiang, "Autofluorescence in normal and malignant human oral tissue and in DMBA-induced hamster buccal pouch carcinogenesis," *J. Oral Pathol. Med.* **64**, 470–474 (1998).
14. H. K. Chiang, S. N. Chow, C. Y. Wang, Y. S. Lee, J. C. Tassi, and C. P. Chiang, "Autofluorescence in normal and malignant human oral tissue and in DMBA-induced hamster buccal pouch carcinogenesis," *J. Oral Pathol. Med.* **64**, 470–474 (1998).
15. J. K. D. Dhingra, D. F. Perrault, K. McMillan, E. Rebeiz Elie, S. Kabani, R. Manoharan, I. Itzkan, M. S. Feld, and S. M. Shapshay, "Early diagnosis of upper aerodigestive tract cancer by autofluorescence," *Arch. Otolaryngol. Head Neck Surg.* **122**, 1181–1186 (1996).
16. A. G. Water, R. Jacob, R. Ganeshappa, B. Kemp, A. K. El-Naggar, J. L. Palmer, G. Clayman, M. F. Mitchell, and R. Richards Kortum, "Noninvasive diagnosis of oral neoplasia based on fluorescence spectroscopy and native tissue autofluorescence," *Otolaryngol.-Head Neck Surg.* **124**, 1251–1258 (1998).
17. S. K. Majumder and P. K. Gupta, "Synchronous luminescence spectroscopy for oral cancer diagnosis," *Lasers Life Sci.* **9**, 143–152 (2000).
18. B. K. Manjunath, K. Jacob, C. Muralikrishna, M. S. Chidananda, K. Venkatakrishna, and V. B. Kartha, "Autofluorescence of oral tissue for optical pathology in oral malignancy," *J. Photochem. Photobiol., B* **73**, 49–58 (2004).
19. K. Venkatakrishna, J. Kurien, M. P. Keerthilatha, C. Murali Krishna, G. Ullas, and V. B. Kartha, "Optical pathology of oral tissue—a Raman spectroscopic diagnostic method," *Curr. Sci.* **80**, 101–105 (2001).
20. K. Venkatakrishna, M. P. Keerthilatha, C. Murali Krishna, J. Kurien, G. Ullas, and V. B. Kartha, "LC-LIF for early detection of oral cancer," *Curr. Sci.* **84**, 551–557 (2003).
21. V. B. Kartha, J. Kurien, M. P. Keerthilatha, R. Lakshmi, L. Rai, K. K. Mahato, C. Muralikrishna, and C. Santhosh, "Diagnosis at the molecular level: analytical laser spectroscopy for clinical applications," in Satoshi Kaneko, editor. *Research Signpost*, S. Kaneko, Ed., pp. 153–221, Trivandrum, India (2005).
22. G. A. Wagnieres, W. M. Star, and B. C. Wilson, "*In vivo* fluorescence spectroscopy and imaging for oncological application," *Photochem. Photobiol.* **68**, 603–632 (1998).
23. G. M. Palmer, C. L. Marshek, K. M. Vrotsos, and N. Ramanujan, "Optimal method for fluorescence and diffuse reflectance measurement of tissue biopsy samples," *Lasers Surg. Med.* **30**, 191–200 (2002).
24. R. J. Lakowicz, *Principles of Fluorescence Spectroscopy*, 2nd ed., pp. 63–66, Kluwer Academic/Plenum, New York (1999).
25. A. Gillenwater, R. Jacob, and R. Richards-Kortum, *Head Neck* **20**, 556–562 (1998).
26. M. G. Muller, T. A. Valdez, I. Georgakoudi, V. Backman, C. Fuentes, S. Kabani, N. Laver, Z. Wang, C. W. Boone, R. R. Dasari, S. M. Shapshay, and M. S. Feld, "Spectroscopic detection and evaluation of morphologic and biochemical changes in early human oral carcinoma," *Cancer* **97**, 1681–1692 (2003).
27. G. S. Nayak, D. K. Sudha, M. P. Keerthilatha, A. Sarkar, R. Satadru, J. Kurien, L. D'Almeida, B. R. Krishnanand, C. Santhosh, V. B. Kartha, and K. K. Mahato, "Principal component analysis (PCA) and artificial neural network (ANN) analysis of oral tissue fluorescence spectra. Classification of normal premalignant and malignant pathological conditions, *Biopolymers* **82**, 152–166 (2006).
28. T. Tsai, H. M. Chen, C. Y. Wang, J. C. Tsai, C. T. Chen, and C. P. Chiang, "In-vivo Autofluorescence spectroscopy of oral premalignant and malignant lesions: distortion of fluorescence intensity by submucous fibrosis," *Lasers Surg. Med.* **33**, 40–47 (2003).
29. D. C. G. de Veld, M. Skurichina, M. J. H. Witjes, P. W. D. Robert, J. C. M. Dick Sterenborg, W. M. Star, and J. N. Roodenburg, "Autofluorescence characteristics of healthy oral mucosa at different anatomical sites," *Lasers Surg. Med.* **32**, 367–376 (2003).
30. B. M. Mehimann, K. Rick, H. Stepp, G. Grevers, R. Baumgartner, and A. Leunig, "Autofluorescence imaging and spectroscopy of normal and malignant mucosa in patients with head and neck cancer," *Lasers Surg. Med.* **25**, 323–334 (1999).
31. C. Hanpeng, Y. Q. Jianan, P. Yuen, J. Sham, D. Kwong, and I. W. William, "Light-induced autofluorescence spectroscopy for detection of naopharyngeal carcinoma *in vivo*," *Appl. Spectrosc.* **56**, 1361–1367 (2002).
32. W. G. Shafer and C. A. Waldron, *Cancer* **36**, 1021–1028 (1975).
33. L. Wynder, I. J. Bross, and R. M. Feldman, *Cancer* **10**, 1300 (1957).
34. H. P. Kriegel and M. Schubert, "Classification of websites as sets of feature vectors," in *Proc. IASTED Int. Conf., Database and Applications*, Innsbruck, Austria (2004).
35. D. C. G. de Veld, M. Skurichina, M. J. H. Witjes, P. W. D. Robert, D. J. C. M. Sterenborg, W. M. Star, and J. L. N. Roodenburg, "Autofluorescence characteristics of healthy oral mucosa at different anatomical sites," *Lasers Surg. Med.* **32**, 367–376 (2003).
36. W. Simon, "Fast k-NN classification for multichannel image data," *Pattern Recogn. Lett.* **17**, 713–721 (1996).
37. K. Vinay, S. C. Ramzi, and L. R. Stanley, *Basic Pathology*, 5th ed., Figure of Cervical Cancer, p. 189, A PRISM Indian Edition (1992).
38. C. Murali Krishna, G. D. Sockalingum, J. Kurien, R. Lakshmi, L. Venteo, M. Pluot, M. Manfait, and V. B. Kartha, "Micro-Raman spectroscopy for optical pathology of oral squamous cell carcinoma," *Appl. Spectrosc.* **58**, 107–114 (2004).
39. M. S. Chidananda, K. Satyamoorthy, R. Lavanya, A. P. Manjunath, and V. B. Kartha, "Optical diagnosis of cervical cancer by fluorescence spectroscopy technique," *Int. J. Cancer* **119**, 139–145 (2006).
40. M. F. Mitchell, "Accuracy of colposcopy. Consult," *Obstet. Gynecol. (N.Y., NY, U. S.)* **6**, 70–73 (1994).
41. G. Javaheri and M. D. Fejgin, "Diagnostic value of colposcopy in the investigation of cervical neoplasia," *Am. J. Obstet. Gynecol.* **137**, 588–594 (1980).
42. P. Mitra, M. Sushmita, and K. P. Sankar, "Staging of cervical cancer with soft computing," *IEEE Trans. Biomed. Eng.* **47**, 934–940 (2000).
43. I. T. Joliffe, *Principal Component Analysis*, Springer-Verlag, New York (1986).
44. B. S. Qiue, W. M. An, Q. Tong, Y. Q. Gao, L. Q. Cheng Cai, X. Ma, and J. M. Zhu, "Differentiating Alzheimer's diseases using artificial neural network analysis of *in vivo* proton MR spectroscopy," *Proc. Intl. Soc. Mag. Reson. Med.* 8, 292 (2000).
45. P. Rudra, *Getting Started with MATLAB 5*, Oxford University Press, New Delhi (2000).
46. P. J. William, *Introduction to MATLAB 6 for Engineers*, McGraw-Hill, Singapore (2001).
47. D. W. Scott, *Outlier Detection and Clustering by Partial Mixture Modeling*, Physics-Verlag/Springer (2004).
48. P. Legendre, "Reply to J. M. Preston and T. L. Kirlin, Acoustic seabed classification: improved statistical method," *Can. J. Fish. Aquat. Sci.* **60**, 1301–1305 (2003).
49. B. Xu and C. Fang, "Clustering analysis for cotton trash classification," *Text. Res. J.* **69**, 656–662 (1999).
50. G. Guodong and Z. Li, "Stan content-based audio classification and retrieval by support vector machines," *IEEE Trans. Neural Netw.* **14**, 209–215 (2003).
51. C. Hanpeng, Y. Qu Jianan, Y. Powing, J. Sham, D. Kwong, and I. W. William, "Light-induced autofluorescence spectroscopy for detection of nasopharyngeal carcinoma *in vivo*," *Appl. Spectrosc.* **56**, 1361–1368 (2002).
52. S. Z. Li, "Content-based classification and retrieval of audio using the nearest feature line method," *IEEE Trans. Speech Audio Process.* **8**, 619–625 (2000).