Tobacco Consumption Induced Changes in the Healthy Oral Mucosa and its Effect on Differential Diagnosis of Oral Lesions – A Clinical *In Vivo* Raman Spectroscopic Study

Hemant Krishna[1], Sidramesh Muttagi[2], Pranav Ingole[2], Pankaj Chaturvedi[2] and Shovan Kumar Majumder[1,*]

[1]*Homi Bhabha National Institute, Raja Ramanna Centre for Advanced Technology, Indore- 452013, India*

[2]*Department of Head and Neck Surgery, Tata Memorial Hospital, Mumbai-400012, India*

Abstract: *Objective*: To investigate tobacco consumption induced changes in the *in vivo* Raman spectra of oral mucosa of healthy volunteers and to study its effect on the differential diagnosis of oral lesions.

Materials and Methods: The clinical *in vivo* study involved 28 healthy volunteers and 171 patients having malignant and potentially malignant lesions of the oral cavity. Twenty of the healthy volunteers had habits of either smoking and/or of chewing tobacco while the rest did not have any tobacco consumption habits. The *in vivo* Raman spectra were measured using a compact and portable near-infrared Raman spectroscopic system. A probability based multi-class diagnostic algorithm, developed for supervised classification, was employed to classify the whole set of measured tissue Raman spectra into various categories.

Results: It was found that the Raman spectra of healthy volunteers with tobacco consumption habits could be separated from the spectra of those without any habit of tobacco consumption with an accuracy of over 95%. Further, it was found that exclusion of the spectral data of the oral cavity of the healthy volunteers from the reference normal database considerably improved the overall classification accuracy (92.3% as against 86%) of the algorithm in separing the oral lesions from the normal oral mucosa.

Conclusion: The results of the clinical study demonstrate the potential of Raman spectroscopy in screening tobacco users who are at an increased risk of developing dysplasia or malignancy. Further, the results also show that for accurate discrimination of oral lesions based on their Raman spectra, the reference normal database should exclude spectral data of tobacco using healthy subjects.

Keywords: *In vivo* Raman spectroscopy, tobacco consumption induced changes, oral mucosa, probability based multivariate diagnostic algorithm, multi-class classification of oral lesions, maximum representation and discrimination feature (MRDF), sparse multinomial logistic regression (SMLR).

1. INTRODUCTION

Oral cancer, mainly attributed to the practice of consuming tobacco in various forms [1-2], is a major health problem in India and other South-Asian countries [3-4]. Although the oral cavity is easily accessible to inspection, but due to lack of awareness and inadequate health care facility, oral cancer is most often detected at an advanced stage when the treatment is less successful thereby leading to high morbidity and mortality [5-7]. The current gold standard for clinical diagnosis of oral cancer is biopsy and subsequent histopathologic examination [7]. The process is both invasive and time-consuming. A real-time diagnostic method to enable non-invasive monitoring of oral cavity in individuals with suspicious oral lesions is, thus, an urgent current need. Recent research has demonstrated the applicability of near-infrared Raman spectroscopy as a promising alternate tool for oral cancer diagnosis [8-14]. However, most of the early investigations [15-23] were primarily limited to either *ex-vivo* studies [15-21] or *in vivo* studies on animal models [22-23]. Perhaps the necessity of long data collection time owing to the weaker Raman signals precluded its use for *in vivo* studies on human oral cavity. With the availability of improved detectors and spectroscopic systems, it is now possible to acquire good quality tissue Raman spectrum in a clinically acceptable data collection time (< 5 s) and several publications reporting *in vivo* studies have already appeared [24-28].

The first *in vivo* study was reported by Guze *et al*. [24] who measured Raman spectra in the higher wave number (1800 -3000 cm^{-1}) region for *in vivo* characterization of the human oral cavity in healthy volunteers and correlated the spectral variability of the observed C-H stretch bands near 3000 cm^{-1} to the different degrees of keratinization of the oral mucosa. This was followed by the *in vivo* study by the Bergholt *et al*. [25] who also acquired Raman spectra from the

*Address correspondence to this author at the Homi Bhabha National Institute, R & D Block-A1, Raja Ramanna Centre for Advanced Technology, Indore 452 013, India; Tel: 91-731-2488437; Fax: 91-731-2488425;
E-mail: shkm@rrcat.gov.in, shovan.k.majumder@gmail.com

ISSN: 1927-7210 / E-ISSN: 1927-7229/16

healthy volunteers but in the conventional fingerprint region (800 -1800 cm^{-1}) for evaluating the applicability of the approach for characterization of their oral cavity. The measured Raman spectra showed considerable variability, which they correlated to the variations in anatomical locations of the interrogated tissue sites within the oral cavity. The applicability of *in vivo* Raman spectroscopy for differential diagnosis of oral lesions was reported by Singh *et al*. [26] who used a commercial Raman spectrometer to measure *in vivo* Raman spectra from patients already identified of having maliganancy of oral buccal mucosa and showed that Raman spectroscopy in combination with a PCA-LDA based diagnostic algorithm could delineate malignant from the uninvolved nomal tissue sites as well as the premalignat lesions appearing in the contralateral buccal mucosa of the same set of patients with an accuracy of up to 87% . In a contemporaneous study, Sahu *et al.* [27] explored the potential of Raman spectroscopy in differentiating tobacco associated pathological changes of buccal mucosa from aging related physiological changes and showed that though there were aging related changes in the Raman spectra but that did not have any influence on the classification of lesions. They also carried out an *ex-vivo* study [28] of the exfoliated cells obtained from the oral cavity of normal healthy controls with and without tobacco habits, disease controls and tumor patients and showed that these could be separated with varying accuracies with respect to cytology based on the measured Raman spectra of these tissue categories. Recently, our group reported a full-scale clinical study [29] on the comprehensive evaluation of the efficacy of *in vivo* Raman spectroscopy for differential detection of oral lesions in the whole of the oral cavity. Further in an independent study [30], we have also investigated the anatomical variability of *in-vivo* Raman spectra of the oral cavity of healthy volunteers and showed that the diagnostic algorithm with anatomy matched spectral data as input led to considerable improvement in the overall accuracy of classification of the oral lesions.

The common general objective of all the above mentioned studies [15-29] was to evaluate the potential of *in vivo* Raman spectroscopy as a tool for an improved diagnosis of oral cancer. However, no attempt was made to systematically study the effect of tobacco consumption on the oral cavity of healthy volunteers which otherwise do not have any disease of the oral cavity. We report, in this paper, the results of an *in vivo* Raman spectroscopic study carried out to characterize the tobacco consumption induced changes on the healthy oral mucosa of individuals having no history of any disease of the oral cavity and also investigate the effect of the tobacco consumption habits of healthy individuals (with the measured Raman spectra of their oral cavity chosen as the normal control) on the outcome of the diagnostic algorithm employed for the differential diagnosis of various oral lesions. We also evaluated the applicability of *in vivo* Raman spectroscopy in separating tobacco users from the non-users. This is important because there is enough epidemiological evidence that long term exposure to tobacco causes alteration in normal mucosa and also is one of the significant etiological factors for the development of oral cancers and pre-cancers [1-2]. The oral lesions investigated belonged to any of the three histopathologic categories: 1) oral squamous cell carcinoma (OSCC), 2) oral leukoplakia (OLK), and 3) oral sub-mucosal fibrosis (OSMF). The diagnostic algorithm was a probability based multivariate statistical algorithm capable of direct multi-class classification of different oral tissue pathologies. It was also found that the overall multi-class classification accuracy of the algorithm was considerably improved (92.3% as against 86%) when the normal database comprised spectral data devoid of the set of spectra of healthy individuals having tobacco consumption habits.

2. MATERIALS & METHODS

2.1. Instrumentation

In vivo Raman spectra were measured using an in-house assembled, compact and portable Raman spectroscopic system described earlier [32]. In brief, the system has all its components accommodated into a 32''suitacase. A 785 nm diode laser (Crysta Laser, Reno, NV) serves as the illumination source. The light is delivered to the target tissue using a bifurcated fiber-optic probe (Visionex Inc., Warner Robins, GA) consisting of a central 400-µm-core-diameter fused-silica illumination fiber surrounded by seven 300-µm fused-silica beam-steered collection fibers. The distal end of the illumination fiber has an in-line band-pass filter for rejection of signals generated in the fibers themselves. The distal end of the collection fiber has an in-line notch filter for rejection of the elastically scattered illumination light. The collection fibers are aligned linearly and imaged on to a 200 µm entrance slit of an imaging spectrograph (Andor Shamrock SR-303i, Belfast, Northern Ireland) coupled with a thermoelectrically cooled (-70^{0}C), back-illuminated, deep-depletion charge-coupled-device (CCD) camera (Andor DU420A-BR-DD, Belfast, Northern Ireland). The

system can acquire good quality tissue Raman spectra with ~20 cm^{-1} resolution and ~ 50:1 signal to noise ratio for an integration time of ~ 5 seconds.

2.2. Clinical Measurements

The *in vivo* study was conducted at the Tata Memorial Hospital (TMH), Mumbai with the approval of the TMH Ethical Committee. All the patients undergoing routine medical examination of the oral cavity at the Out Patient Department (OPD) of TMH were recruited in the *in vivo* study regardless of gender or race. The final eligibility of each patient was determined by the participating doctor based on the medical condition of the patient such that patient care was not compromised.

All the spectral measurements were performed by the participating head and neck surgeon (PI) using a standard protocol which was maintained for all individuals in this study. Prior to recording spectra from an individual, the fiber-optic probe was disinfected with CIDEX (Johnson and Johnson, India), washed with PBS and cleaned dry with a piece of sterilized cotton. The mucosal surface was wiped with sterile gauge to remove any saliva, blood or betel quid incrustations accumulated at the tissue surface. The probe tip was also wiped dry between consecutive measurements from different tissue sites in an individual. For recording the *in vivo* Raman spectra, the tip of the fiber-optic probe was placed in gentle contact with the tissue surface and it was ensured that none of the patients or the normal volunteers complained of the probe being painful. The overhead room lights in the OPD room were turned off temporarily during spectral acquisition to minimize the contribution of the ambient light in the acquired spectra.

The study involved 171 patients (with oral lesions) enrolled for medical examination of the oral cavity at TMH and 28 healthy volunteers with no history of any disease of the oral cavity. The details the distribution of the number of individuals and the spectral measurements from the various sites of their oral cavity are summarized in Table **1**. Of the 28 healthy volunteers 20 had habits of either smoking and/or chewing tobacco and the rest were tobacco non-users. The tobacco using healthy volunteers had habits of either smoking or chewing tobacco for more than 5 years. Biopsies were taken subsequent to acquisition of spectra from the oral cavity sites suspected of being malignant or potentially malignant. However, as per the terms of the approval from the Ethical Committee of the hospital, no biopsies were available from the investigated sites of the patients with oral submucous fibrosis (OSMF) and the diagnosis of this condition was based on clinical findings only. The biopsy samples were fixed in formalin and were examined later by an experienced pathologist who was blinded to the results of the optical spectra. The histopathology report was considered as the “gold standard”. All the Raman spectra were categorized in accordance with their histological identities and grouped into OSCC, OLK, OSMF, or normal oral squamous tissue. Informed consent was obtained from each patient as well as the normal volunteers who participated in this study. Age, sex, and details of smoking habit (if any) were also recorded for all subjects included in the study. The age variations for OSCC, OSMF, OLK and healthy volunteers were ~35±11, 51±13, 53 ±14 and 44±10 years respectively. The overall ratio of male to female population was ~6:1.

2.3. Pre-processing of Data

Prior to Raman spectral measurements from the oral cavity of a subject, the wavenumber axis was calibrated using the excitation laser line and the Raman spectra measured from acetaminophen, and

Table 1: The Distribution of the Number of Patients and Healthy Volunteers and the Spectra Measured from the Different Sites of their Oral Cavity into Various Categories

No. of Patients	No. of Sites	Category
113	316	Oral Squamous Cell Carcinoma (OSCC)
25	94	Oral Submucous Fibrosis (OSMF)
33	105	Oral Leukoplakia (OLK)
20	204	Normal with tobacco consumption habit (N:WTH)
8	83	Normal without any tobacco consumption habit (N:WOTH)

naphthalene standards. The signals from various pixels of the CCD were binned along the vertical axis to create a single spectrum for each measurement. Prior to any further processing, the spectrum was truncated to include only the region of 950 - 1750 cm^{-1}. A sequence of steps was then executed on this binned, truncated spectrum following the procedure described earlier [31]. First, the spectrum was corrected for the system spectral response by using a NIST traceable calibration lamp (LS-1, Ocean optics, Inc., Dunedin, FL) after removal of the dark signal. In the next step the artifacts in the measured tissue spectrum (due to laser-induced artifacts generated in the fiber-optic probe) were removed. This was done by recording the spectrum of the backscattered light from a roughened aluminium block and then iteratively subtracting this spectrum scaled by a range of different intensities till the optimal ratio for background removal is reached. The resultant spectrum with the lowest standard deviation of the residual between the data and the model fit was used for fiber background removal. Following removal of fiber artifacts, the spectrum was binned along the wavenumber axis in 3.5 cm^{-1} intervals and noise smoothened using a second-order Savitzky–Golay filter. Following noise removal, the remaining data were background subtracted using the range-independent background subtraction algorithm (RIA) [31] to retrieve the weak tissue Raman spectrum. The method uses a model based on modified iterative smoothing of the measured Raman spectrum in such a manner that the high-frequency Raman peaks are gradually eliminated finally leaving the underlying broad baseline which can be subtracted from the raw spectrum to yield the true Raman signal. Each background-subtracted tissue Raman spectrum was normalized with respect to its mean spectral intensity across all the Raman bands.

2.4. Data Analysis

The *in vivo* Raman spectra measured from the tissue sites belonging to the oral cavity of healthy volunteers as well as of patients with oral lesions were analyzed employing different multivariate statistical methods whose brief description is as follows.

Pillai's V

In order to quantify the differential amongst the measured Raman spectra belonging to different pathology classes as well between the spectra of healthy tobacco users and the non-users, a metric, called Pillai's V [32], was calculated. In brief, Pillai's V, used in multivariate analysis of variance (MANOVA), is a statistical measure of the amount of separation between the samples belonging to multiple classes and is the trace of the matrix defined by the ratio of between-group variance (B) to total variance (T). The Pillai's V trace is given by:

$$V = trace(BT^{-1}) = \sum_{i=1}^{h} \frac{\lambda_i}{\lambda_i + 1} \quad (1)$$

where λ_i is the i^{th} eigenvalue of the $W^{-1}B$ in which W is the within-group variance and h is the number of factors being considered in MANOVA, defined by h = c-1, c being the number of classes. A high Pillai's V means a higher amount of separation between the samples of classes, with the between-group variance being relatively large compared to the total variance.

Standard Error (SE) Confidence Interval

To identify the region of spectral differences between the Raman spectra of healthy volunteers with (N:WTH) and without tobacco consumption habits (N:WOTH), standard error (SE) confidence intervals [10] were utilized. The SE was calculated at each wavenumber as:

$$SE\ (\lambda) = \sqrt{\frac{\sigma^2_{N:WTH}(\lambda)}{n_{N:WTH}} + \frac{\sigma^2_{N:WOTH}(\lambda)}{n_{N:WOTH}}} \quad (2)$$

Here, σ^2 is the pooled variance of the intensities at each wavenumber of each tissue type and n_{Type} is the number of tissue spectra included in that particular tissue type. The SE was then multiplied by appropriate t-values based on total degrees of freedom and a predefined confidence level to produce a confidence interval. Difference spectra between two tissue types were overlaid on the confidence interval to qualitatively identify statistically significant spectral differences. The portion of the difference spectra outside the confidence interval represents the region of the spectral differences with predefined statistical significance.

Multi-Class Diagnostic Algorithm

In order to analyze the diagnostic content of the *in vivo* Raman spectra measured from the oral cavity of the different subjects and quantitatively predict their class-memberships, probability-based multivariate statistical diagnostic algorithms capable of direct multi-class classification [33] was developed. The development of the algorithms [33] consisted of two steps: 1) extraction of diagnostic features from the spectra using nonlinear maximum representation and

discrimination feature [34] (MRDF) and 2) probabilistic classification based on linear sparse multinomial logistic regression [35] (SMLR) for classifying the nonlinear features into corresponding tissue pathologies.

Given a set of input data comprising Raman spectra with a given dimensionality from different pathology classes, nonlinear MRDF [33-34] aims to find a set of nonlinear transformations on the input data that optimally discriminate between the different classes in a reduced dimensionality space. It uses nonlinear transforms that are polynomial mappings of the input data and computes K<<D nonlinear transformation vectors, Φ_K, from D-dimensional (where D is the number of wavenumbers over which spectra were recorded) spectral data of oral tissues, such that the projections of the input data on Φ_K from the different tissue categories are statistically well separated from each other.

Classification with SMLR [33,34] is a probabilistic multi-class model based on sparse Bayesian machine-learning framework of statistical pattern recognition. The central idea of SMLR is to separate a set of labeled input data into its constituent classes by predicting the posterior probabilities of their class-membership. It computes the posterior probabilities using a multinomial logistic regression model and constructs a decision boundary that separates the data into its constituent classes based on the computed posterior probabilities following Bayes' rule. Classification of a given set of input data x is based on the vector of posterior probability estimates yielded by the SMLR algorithm and a class is assigned to each dataset (transformation of the original spectrum) for which its posterior probability is the highest.

Normalized spectra were used as inputs to the developed algorithms as described previously by Majumder *et al*. [33]. All analyses were performed using leave-one-individual-out cross validation. In this method, the set of spectra from an individual was held out from the full set of spectral data of N individuals and the training of the algorithm was performed using all the spectra of the remaining N-1 individuals. The algorithm thus trained was then used to predict the class-memberships of the withheld spectra from the excluded individual. This was repeated N number of times (until the spectra of all the N number of individuals were exhausted) each time excluding a different individual for the purpose of validation and re-training the algorithm using spectra of the rest of the individuals. Since the training set data remained completely independent of the test data in each of the N loops (as the set of spectra from an individual was never a part of both the training and the validation sets simultaneously), the validation was statistically unbiased. Each spectrum was classified to the predicted class membership (pathology) with the highest posterior probability.

Multiclass Receiver-Operating Characteristic (ROC) Analysis

Quantification of the performance of an algorithm is one of the most critical tasks in machine learning based classification. Although a variety of methods like volume under the surface, global performance index, Matthews correlation coefficient, confusion entropy etc. have been used for that purpose, we have used the Hand and Till measure (HTM) which is the most widely used technique for quantifying the performance of a multi-class classification algorithm [36]. Its major attraction is that it is objective and does not require any subjective input from the user. HTM is based on multi-class Receiver Operating Characteristic (ROC) approximation and extends the Area Under the Curve (AUC), one of the most popular measures for binary classifiers to multiclass tasks. In practice, it calculates the average of the AUCs on the pair-wise binary problems derived from the multi-class problems to generate a scalar summary of the algorithm's performance. Given 'c' number of pathology classes, overall performance of a multi-class diagnostic algorithm is taken as the average of pairwise area under the ROC curves between c(c−1)/2 pairs of classes and given by Hand and Till measure (HTM) [37] as:

$$HTM = \frac{2}{c(c-1)} \sum_{i<j} AUC(i,j) \quad (3)$$

Where, AUC is the area under the two-class ROC curve involving classes 'i' and 'j'. The summation is calculated over all pairs of distinct classes, irrespective of order. Similar to the two-class case, the closer the HTM equals to 1, the more accurate the corresponding diagnostic algorithm is.

3. RESULTS

Figure **1** shows the mean, normalized Raman spectra of normal oral mucosa of the healthy volunteers with and without tobacco consumption habits. Each spectrum is the average over the respective number of tissue sites interrogated in the

corresponding case. The error bars represent ±1 standard deviation. It is apparent from the figure that subtle but significant differences exist in peak intensities between the two cases indicating biochemical differences inherent in the two different normal tissue types. The percentage variation ($\sigma/\bar{x}$) in the spectral intensities from the different measurement sites was observed to lie in the range of ~15%-35% over the respective number of tissue sites included in the two cases for all the measured spectra. Here, $\bar{x}$ is the mean intensity value from different measurement sites of one category and σ is one standard deviation. For illustrating the spectral differences between the tissue types, mean difference spectrum obtained by subtracting one mean spectrum from the other is shown in Figure **2**. The gray band shows the confidence interval calculated by multiplying SE with a t-value corresponding to 95% confidence intervals and the degrees of freedom equal to number of spectral measurements (corresponding to each category) minus the number of tissue categories. A number of significantly different Raman bands are observed for the tobacco non-users with respect to the tobacco users. For example, the intensities of the Raman bands in the wavenumber regions of 1244-1272 cm^{-1}, 1297-1313 cm^{-1}, 1434-1456 cm^{-1} and 1643-1672 cm^{-1} are found to be considerably higher in the spectra of tobacco non-user as compared to those of tobacco users indicating changes in the collagen and lipid contribution.

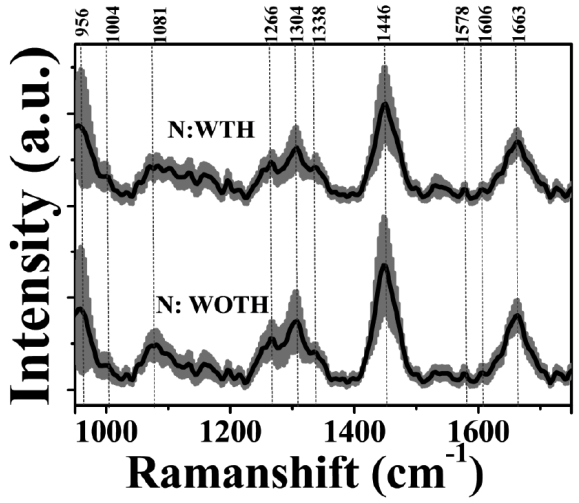

Figure 1: Mean, normalized Raman spectra of the oral mucosa of healthy volunteers with tobacco consumption habit (N:WTH) and without any tobacco consumption habit (N:WOTH). The error bars (gray) represent ±1 standard deviation.

In order to quantify the above differences in the Raman signatures of healthy volunteers with and without tobacco consumption habits, the MRDF-SMLR based multi-class classification algorithm was applied in binary mode on the spectral data sets. Table **2** lists the results of leave-one-subject-out supervised classification in the form a confusion matrix. One can see excellent discrimination between the tissue types with classification accuracy of over 95%, which further confirms the statistical significance of the spectral differences observed between the two. In addition to assigning class labels, the diagnostic algorithm also yielded posterior probabilities of the measured tissue sites belonging to each oral tissue category. The posterior probabilities are indicative of the certainty of classification and they are plotted for all the different tissue sites included in each tissue category. Figure **3** illustrates the computed posterior probabilities for the oral tissue sites investigated for the healthy tobacco non-users and the tobacco users. One can see that while more than ~93% of the tissue sites are having posterior probabilities of greater than 80% of belonging to either of the two categories, less than 5% of the tissue sites in either of the categories are having exceedingly low posterior probabilities (<30%) of belonging to their respective category. This is not quite unusual considering the individual variation in tobacco absorption, metabolism and excretion along with the fact that the grouping of an individual with a healthy oral cavity into either of the two categories of tobacco users and non-users was based on self reported questionnaire.

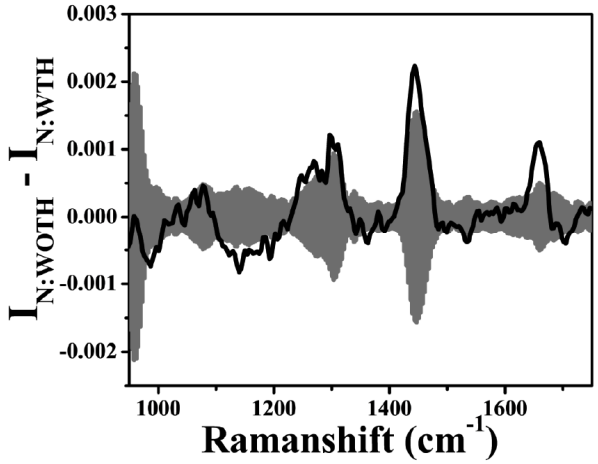

Figure 2: Mean difference spectra showing statistical differences between oral tissue Raman spectra of healthy volunteers without any tobacco consumption habit (N:WOTH) and with tobacco consumption habit (N:WTH). Gray bands indicate 95% confidence intervals of the difference determined by standard error confidence intervals.

Figure **4** shows the mean Raman spectra of the different oral lesions investigated. The spectra are

Table 2: Confusion matrix displaying classification of the Raman spectra of normal oral tissue sites into two classes: normal without any tobacco consumption habit (N:WOTH) and normal with tobacco consumption habit (N:WTH). The classification results were obtained by using the MRDF-SMLR diagnostic algorithm in leave-one-subject-out cross validation mode

	Raman Classification	
Tissue Category	**N:WOTH**	**N:WTH**
N:WOTH (n=83)	95.2%	4.8%
N:WTH (n=204)	4.9%	95.1%

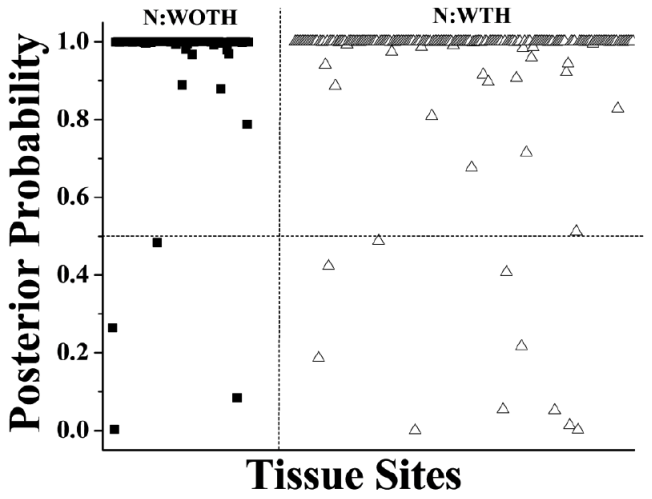

Figure 3: Posterior probabilities of being classified as normal without any tobacco consumption habit (N:WOTH) and normal with tobacco consumption habit (N:WTH).

averaged over all the tissue sites interrogated in the corresponding lesions. For the sake of comparison, the spectra of the normal oral mucosa of healthy volunteers with and without tobacco consumption habits are also shown in the same figure. The error bars (one standard deviation) represent the variability of Raman spectral signatures across the different tissue categories. The Raman bands appearing in the spectra of the different oral tissue types reveal that the pathologic spectra can largely be separated based on protein and lipid related Raman features. For instance, the intensity of 1004, 1213, 1338, 1578 and 1606 cm^{-1} Raman bands, believed to be due to proteins, [8-16,18-19] were found to be higher for malignant tissues as compare to normal. On the other hand, the lipid-specific Raman peaks at ~1081, 1266, 1304, 1446, and 1663cm^{-1} were found to be stronger in normal. In contrast, the differences in the Raman spectra of potentially malignant with respect to normal are seen to be around 1081, 1304, 1450 and 1663 cm^{-1} Raman peaks indicating an increased tendency of the potentially malignant tissues to show keratinization as compared to normal.

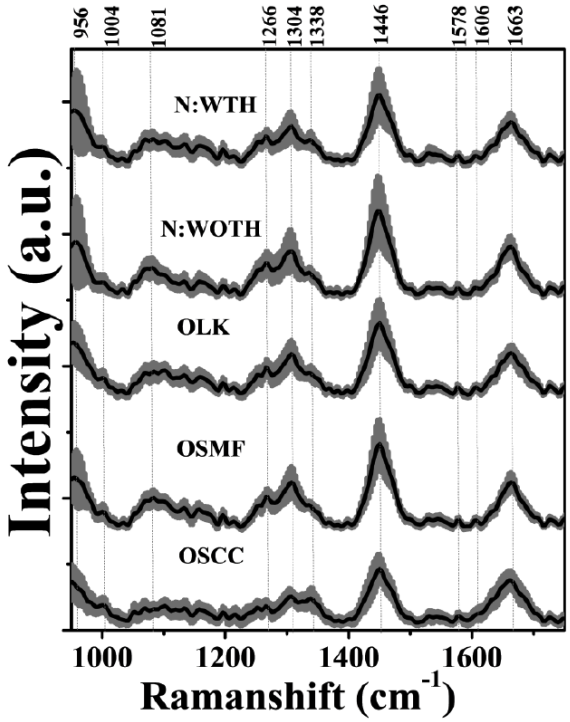

Figure 4: Mean, normalized Raman spectra of OSCC (n=316), OSMF (n=94), OLK (n=105), N:WOTH (n=83), and N:WTH (n=204). The error bars (gray) represent ±1 standard deviation.

In order to investigate whether the normal oral tissue sites of healthy volunteers with tobacco consumption habits could be separated from the rest of the tissue types in multi-class discrimination platforms, the probabilistic multi-class classification algorithm was applied on the full set of spectra belonging to the following tissue categories; Case-I: normal with tobacco consumption habit (N:WTH), normal without any tobacco consumption habit (N:WOTH), and potentially malignant (consisting of spectra of OSMF and OLK tissue sites pooled together), Case-II: N:WTH, N:WOTH, potentially malignant (PM) and malignant (comprising spectra of OSCC tissue sites) and Case-III: N:WTH, N:WOTH, OSCC, OSMF and OLK. Tables **3-5** show the classification results in the form of confusion matrices displaying comparison of actual or pathological with that of Raman spectroscopic diagnosis for the whole set of spectra. In all the cases, the classification results were obtained based on leave-one-subject-out cross validation of the respective data sets. A look at the tables clearly reveals that the oral tissue sites of the tobacco using healthy volunteers can be separated in all the cases with an accuracy of

classification of ~80%. Figures **5a**, **b** and **c** illustrate the posterior probabilities computed by the MRDF-SMLR algorithm for the measured tissue spectra of each tissue class of belonging to that particular class for different cases investigated. It is worth noting that majority of the misclassified sites of the healthy tobacco users fall into either malignant or potentially malignant categories indicating a possibility of mucosal alterations at the interrogated locations.

Table 3: Confusion matrix displaying classification of the Raman spectra of oral tissue sites into three classes: normal without any tobacco habit (N:WOTH), normal with tobacco habit (N:WTH), and potentially malignant (PM; consisting of spectra of OSMF and OLK tissue sites pooled together). The classification results were obtained by using the MRDF-SMLR diagnostic algorithm in leave-one-subject-out cross validation mode

Tissue Category	Raman Classification		
	N:WOTH	N:WTH	PM
N:WOTH (n=83)	92.8%	0%	7.2%
N:WTH (n=204)	1.0%	92.1%	6.9%
PM (n=199)	8.6%	4.0%	87.4%

Table 4: Confusion matrix displaying classification of the Raman spectra of oral tissue sites into four classes: normal without any tobacco consumption habit (N:WOTH), normal with tobacco consumption habit (N:WTH), potentially malignant (PM) and malignant (OSCC). The classification results were obtained by using the MRDF-SMLR diagnostic algorithm in leave-one-subject-out cross validation mode

Tissue Category	Raman Classification			
	N:WOTH	N:WTH	PM	OSCC
N:WOTH (n=83)	85.6%	4.8%	3.6%	6.0%
N:WTH (n=204)	3.4%	82.4%	9.8%	4.4%
PM (n=199)	2.0%	7.6%	84.9%	5.5%
OSCC (n=316)	1.6%	4.1%	5.7%	88.6%

In order to investigate the effect of the tobacco induced variability in the Raman spectra of the oral cavity of healthy volunteers on the outcome of supervised classification, the probabilistic multi-class diagnostic algorithm was applied on two sets of spectral data; Set-I: OSCC, OSMF, OLK and pooled set of spectra of tissue sites of healthy volunteers with and without tobacco consumption habit (N:ALL) and Set-II: OSCC, OSMF, OLK and spectra of tissue sites of healthy volunteers with no tobacco consumption habit (N:WOTH). In both the cases, the common task of the algorithm was to classify the measured tissue Raman spectra into four different tissue categories: "normal", OSCC, OSMF and OLK. Tables **6-7** show the confusion matrices listing classification results corresponding to Set-I and Set-II respectively. It is apparent from the tables that the overall classification accuracy is significantly improved in the case of Set-II where the spectra of tissue sites of healthy volunteers with tobacco habits are excluded from the data of normal category. Table **8** shows the Pillai's V values obtained for the two sets of Raman spectra. One can see higher values of Pillai's V for the Set-II indicating a larger separation between the four tissue categories in this case.

Figures **6a** and **b** illustrate the posterior probabilities computed by the MRDF-SMLR algorithm for the measured tissue spectra of each tissue category of belonging to that particular category for the two different cases: Case-1 when the pooled spectral data of tobacco users and non-users was considered as the reference normal, and Case-2 when the set of spectra of tobacco users was excluded from the spectral data of reference normal. It is apparent from the figures that while more than ~82% (on an average) of the correctly classified tissue sites in different tissue categories have a posterior probability >0.80 for Case-2 (spectra of tobacco users excluded from normal), the corresponding fraction is reduced to ~72%, when the set of spectra of the tobacco users was included in the normal database.

The ROC analyses of the classification results provided a quantitative evaluation of the overall performance of the diagnostic algorithm for the two different classification cases. Table **8** lists the HTM values obtained for the two cases. While the estimated HTM value for Case-1 is seen to be 0.95, for Case-2 the corresponding value is seen to be 0.99. It is important to mention here that the HTM value is a quantitative measure of the gross performance of an algorithm and the HTM for an ideal diagnostic algorithm will have a value of 1.

4. DISCUSSIONS

There has been no systematic study, thus far, on the use of *in vivo* Raman spectroscopy for exploring the tobacco consumption induced changes in the oral mucosa of healthy volunteers. Further, no reports exist

Table 5: Confusion matrix displaying classification of the Raman spectra of oral tissue sites into five classes: normal without any tobacco consumption habit (N:WOTH), normal with tobacco consumption habit (N:WTH), OSCC, OSMF and OLK. The classification results were obtained by using the MRDF-SMLR diagnostic algorithm in leave-one-subject-out cross validation mode

Tissue Category	Raman Classification				
	N:WOTH	N:WTH	OSCC	OSMF	OLK
N:WOTH (n=83)	91.6%	4.8%	2.4%	1.2%	0%
N:WTH (n=204)	2.4%	79.9%	9.8%	6.4%	1.5%
OSCC(n=316)	1.3%	6.0%	87.7%	0.9%	4.1%
OSMF (n=94)	0%	9.6%	2.1%	86.2%	2.1%
OLK(n=105)	1%	5.7%	4.8%	2.8%	85.7%

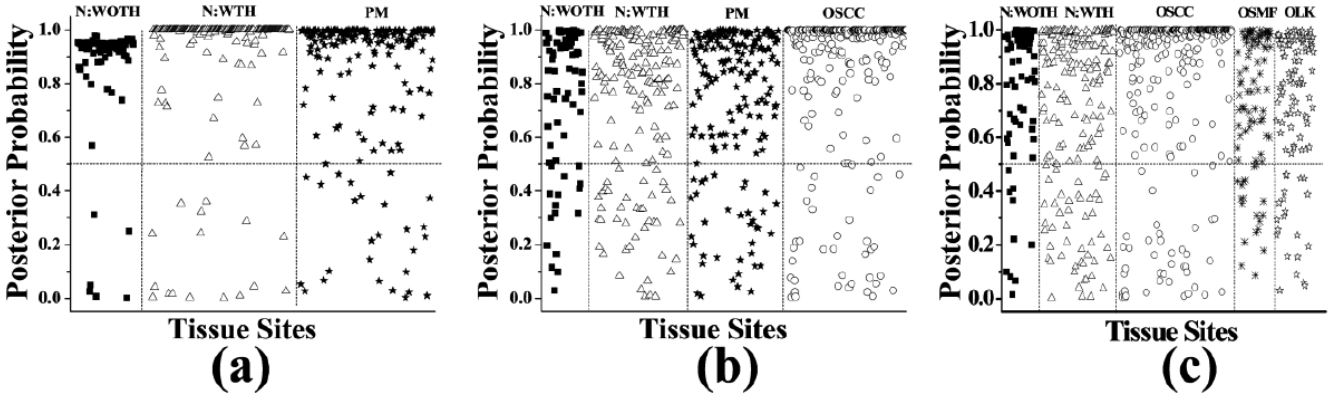

Figure 5: Posterior probabilities for the Raman spectra of the oral tissue sites of being classified as: (**a**) normal without any tobacco consumption habit (N:WOTH), normal with tobacco consumption habit (N:WTH), and potentially malignant (PM; consisting of spectra of OSMF and OLK tissue sites pooled together), (**b**) N:WOTH, N:WTH, PM and malignant (comprising spectra of OSCC tissue sites), and (**c**) N:WTH, N:WOTH, OSCC, OSMF and OLK.

Table 6: Confusion matrix displaying classification of the Raman spectra of oral tissue sites into four classes: OSCC, OSMF, OLK and pooled set of spectra of tissue sites of healthy volunteers with and without tobacco consumption habit (N:ALL). The classification results were obtained by using the MRDF-SMLR diagnostic algorithm in leave-one-subject-out cross validation mode

	Raman Classification			
Tissue Category	N:ALL	OSCC	OSMF	OLK
N:ALL (n=287)	85.0%	8.4%	3.8%	2.8%
SCC (n=316)	8.2%	88.6%	0.3%	2.9%
SMF (n=94)	13.8%	0%	85.1%	1.1%
LPK (n=105)	13.3%	4.8%	0%	81.9%

on investigating whether the outcome of a spectroscopic diagnosis depends on inclusion or exclusion of the spectral data (of the oral cavity) of healthy individuals with tobacco consumption habits in the reference normal database. The present study is aimed at addressing to this crucial issue. Another important goal of the present study is to investigate the feasibility of using Raman spectroscopy for separating tobacco consuming healthy individuals (with no history of any disease of the oral cavity) from those healthy individuals who do not have any history of either

Table 7: Confusion matrix displaying classification of the Raman spectra of oral tissue sites into four classes: OSCC, OSMF, OLK and spectra of tissue sites of healthy volunteers with no tobacco consumption habit (N:WOTH). The classification results were obtained by using the MRDF-SMLR diagnostic algorithm in leave-one-subject-out cross validation mode

	Raman Classification			
Tissue Category	N:WOTH	OSCC	OSMF	OLK
N:WOTH (n=83)	96.4%	2.4%	1.2%	0%
OSCC (n=316)	1.6%	89.9%	2.5%	6.0%
OSMF (n=94)	1.1%	0%	96.8%	2.1%
OLK (n=105)	1.0%	4.8%	1.9%	92.3%

Table 8: The values of Pillai's V and the Hand-Till measure (HTM) for four-class receiver operating characteristic (ROC) analysis of the classification results. While, Set-I corresponds to OSCC, OSMF, OLK and N-ALL (i.e. pooled set of spectra of the oral tissue sites of healthy volunteers with and without tobacco consumption habits), Set-II corresponds to OSCC, OSMF, OLK and N:WOTH (i.e. spectra of the oral tissue sites of healthy volunteers without any tobacco consumption habit)

	Measure	
Classification Set	**Pillai's V**	**HTM**
Set-I	0.78	0.95
Set-II	0.96	0.99

tobacco consumption or any other oral diseases. This is required because recent advances in the understanding of oral cancer have enough evidence to suggest that several cytological and molecular changes occur in the visibly normal oral mucosa several years before frank cancers develop [5-7], and tobacco, a genotoxic as well as a local irritant, is a key contributor to that [1-2,38]. In fact, it has been observed that more than 75-90% of patients who develop dysplasia or malignancy of the oral cavity have a history of long-term tobacco use [39]. Thus monitoring of the oral cavity of otherwise healthy individuals at regular intervals and identifying individuals already at risk for oral cancer and its precursors has the potential to improve early detection, providing the opportunity to intervene when treatment is most effective.

Traditionally, identifying a tobacco consuming individual and quantifying his tobacco exposure for recognizing the person's risk for oral neoplasia, is based on certain personal information such as the number of cigarettes (or bidi) smoked a day, the amount (or weight) of tobacco products chewed, the duration and frequency of tobacco consumption etc. which are generally obtained from a self reported questionnaire. Owing to the subjective nature of the furnished information and also due to the intrinsic variability that may exist in the absorption, metabolism and excretion of tobacco in different individuals, the self reported methodology is not expected to provide a reliable estimate of the actual tobacco exposure. In order to overcome this problem, recently, certain biomarkers have been proposed for assessing or monitoring the tobacco exposure and also evaluating the efficacy of measures designed to control the ill effects of tobacco. For example, cotinine present in serum, saliva or urine, has been reported to be one of the most specific and sensitive biomarkers of tobacco exposure over a short time (few days) and its levels have been found to positively correlate to the risks of some tobacco related diseases [40]. Similarly, hair nicotine has been shown to be a valid and reliable measure of long term exposure that has the potential to be readily applied in epidemiological studies [41]. Though each of these methods has been shown to be useful in certain situations and research aimed at improving the accuracy of these approaches is ongoing, a common limitation of these methods is the requirement of a series of processing steps which are both time and chemicals consuming. Thus, there remains an important need for alternative methods that can help objectively separate tobacco users from the non-users without requiring their personal feedback. The present study clearly demonstrates that Raman spectroscopy has the ability to do that job and it can be used as an alternate tool to non-invasively distinguish tobacco users not only from the non-users but also

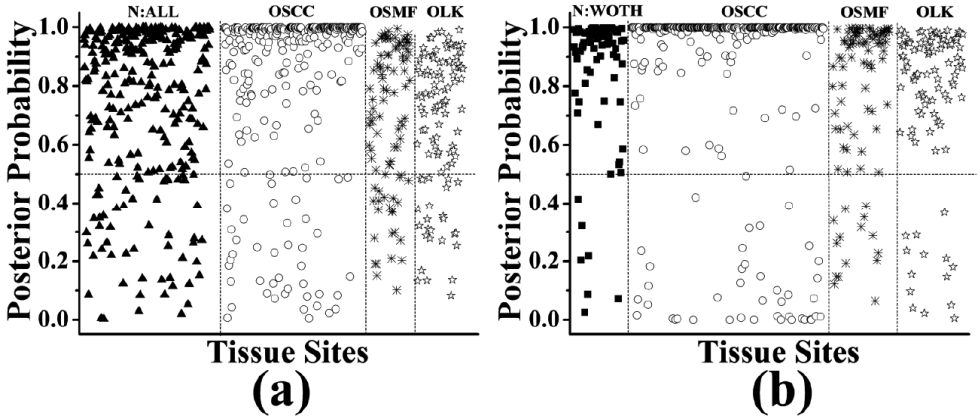

Figure 6: Posterior probabilities for the Raman spectra of the oral tissue sites of being classified as: (**a**) N:ALL (spectra of tobacco users and non-users put together), OSCC, OSMF, and OLK, and (**b**) N:WOTH (spectra of tobacco users excluded), OSCC, OSMF, and OLK.

from other premalignant and malignant lesions of the oral cavity. For example, one can see from the results that while tobacco users can be separated from the tobacco non-users with an accuracy of over 95% in the binary classification mode (Table **2**), the accuracy is ~80% when it comes to separating these from the OSCC, OSMF and OLPK along with the tobacco non-users in the multi-class classification platform (Tables **5**).

The primary basis of the Raman spectroscopic identification of the tobacco users is the various changes in the spectral signatures of the oral cavity mucosa caused by the consumption of tobacco. In Figure **1** one can clearly see that the tobacco exposed and non-exposed oral tissues exhibit notable changes in the intensities of protein and lipid related spectral features. For example, the Raman peaks at 1243, 1299, 1447, and 1655 cm^{-1} associated with collagen were found to have larger intensities for the spectra of the healthy oral cavities without any tobacco exposure as compared to those having tobacco exposure. This probably can be attributed to two reported facts. First, nicotine in tobacco is believed to inhibit the growth of fibroblasts and their production of fibronectin and collagen [1, 42] and second, consistent use tobacco is known to affect the surface epithelium in the oral cavity resulting in thickening of the epithelium (white lesion) thereby reducing the subsurface collagen contribution [43].

While the findings of the binary classification revealed that the healthy volunteers grouped into tobacco users and non-users (based on the feedback of the reported questionnaire) could indeed be separated from each other with excellent accuracy (Table **2**) based on the Raman spectra measured from their oral cavity, it was also relevant to find the significance of these inter-normal spectral variations towards multi-class classification of the different oral tissue pathologies. One can see from the classification results listed in the confusion matrices (Table **6** and **7**) that the classification with spectral data of tobacco users excluded from the reference normal provides an improved overall classification accuracy of ~92% as against an accuracy of ~86% when the pooled set of spectra of tobacco users as well as non-users was taken into account during classification. The spectra belonging to OLK, OSMF and healthy volunteers are seen to be correctly classified with ~82%, 85% and 85%, accuracies respectively, when the reference normal includes the spectra of tobacco users. However the situation is seen to drastically improve with the corresponding accuracies improving to ~ 92%, 97% and 96% when the spectra of tobacco consuming healthy volunteers are excluded from the spectral data of reference normal. In the case of OSCC too, the classification accuracy is seen to improve when the spectra of tobacco users are excluded from the database of normal control during multi-class classification, but the improvement is found to be less (only a little over ~1%) in comparison with other categories. All these observations are further supported by the multi-class ROC analyses that resulted in an HTM value 0.99 for classification with spectra of tobacco users excluded from the reference normal database and 0.95 for the case when pooled spectral data of both the categories was taken into consideration. The reason for this improvement in the classification accuracy can be understood from Table **8** which lists the values of the Pillai's V before and after exclusion of tobacco users. It is seen that Pillai's V value corresponding to the classification with tobacco users excluded from the reference normal is larger than its value obtained for the pooled spectral data (without excluding the spectra of tobacco users). One may note that Pillai's V is a quantitative measure of the separation between different pathology classes and a large Pillai's V value means large amount of separation between different pathology classes [33].

We have used a multi-class diagnostic algorithm based on MRDF and SMLR for separating the oral lesions of different pathologies from the normal oral mucosa based on their measured Raman spectra. The major advantage of using the MRDF-SMLR algorithm is that it is inherently polychotomous allowing one to simultaneously classify spectral data into more than two classes without the need to train and heuristically combine multiple dichotomous classifiers. It is a non-linear algorithm that uses higher order correlations and is capable of providing improved discrimination because of its built-in capability to separate classes which are not linearly separable in the original input data space [33]. Another important advantage of the algorithm is that being based on a Bayesian framework, it is able to predict the posterior probability of class-membership of the investigated tissue sites. This idea is demonstrated in Figures **5-6** where the predicted posterior probabilities for the different oral tissue sites of being classified into their respective pathology classes are plotted. The availability of this quantitative information during tissue discrimination would allow the clinician to reassess those sites that are classified with higher relative uncertainty.

Compared to non-probabilistic classification schemes like SVM [44], nearest-mean classifier (NMC), principal components analysis (PCA) [45], linear discriminant analysis (LDA) [46], or PCA-LDA etc. where sites having a diagnostic score below a certain threshold would be classified as normal, in the probabilistic scheme the sites showing lower probability than that for "absolute normal" may be further interrogated if the objective is to not to miss any abnormal sites, as may be required for accurate screening of the oral cavity.

It is important to mention here that the performance of a diagnostic algorithm also depends on the prototype spectral data included in the training set and the detection of any abnormality is possible only by comparing with a benchmark normal that serves as the reference. The more exact the benchmark normal, the better is the possibility of accurate detection. Thus, consideration of spectral data of healthy oral cavities with no history of any disease of the oral cavity as the reference normal is expected to better assess the "normal-ness" of the contralateral uninvolved region of the oral cavity (of a patient) that is assumed to be normal based on visual assessment. In contrast, the inclusion of spectral data of the visually uninvolved regions in the reference normal, might increase the number of false negatives, since these regions might have some sub-visual changes due to the field effect of the disease (though no histopathological confirmation is possible) and thus may not be truly normal. Due to this reason the spectral data from the normal squamous tissue sites of the healthy volunteers instead of that from the tissue sites of normal appearing mucosa in the contralateral uninvolved region of the oral cavity of patients were used as the reference normal database for the diagnostic algorithm employed in the present study.

The results of the present study definitely prove the hypothesis that consumption of tobacco causes detectable spectral changes in the otherwise healthy oral mucosa of an individual. In terms of clinical diagnosis using Raman spectroscopy, this signifies that one should incorporate spectral data of only the healthy individuals, who do not have any history of tobacco consumption, in the training set as reference normal to have an improved performance of the algorithm (i.e. more accurate diagnosis) for the test set. However, it should be noted that the present study was based on spectra from a limited number of individuals assumed to be representative of the entire patient population. The patient selection criteria as well as the limited number of spectra in each tissue category might influence the classification results obtained in this study. Further, the intrinsic variability of tissue Raman spectra that might result due to other factors like the influence of anatomy, age, gender etc. also might affect the classification performance. Therefore, further clinical studies in a larger patient population, which are already in progress, will be used to address these issues and validate the classification estimates presented here.

5. CONCLUSIONS

To conclude, a clinical study was carried out to characterize the variability of the *in vivo* Raman spectra of the oral cavity of healthy volunteers with and without any tobacco consumption habits and investigate the effect of inclusion and exclusion of the spectral data of tobacco users in the reference normal database on the performance of a probabilistic multi-class diagnostic algorithm employed to discriminate malignant and potentially malignant oral lesions from the healthy oral mucosa. It was found that the tobacco consuming healthy volunteers could be separated from those without any habit of tobacco consumption with an accuracy of over 95% based on the Raman spectra measured from their oral cavity. This indicates the potential of Raman spectroscopy to detect preclinical changes (in the apparently normal mucosa) that could serve as predictor of increased risk of dysplasia or malignancy in an individual. Exclusion of the spectral data of the oral cavity of the healthy volunteers from the reference normal database was found to provide an overall classification accuracy of ~92% as against an accuracy of ~86% obtained in the case of pooled spectral set of the oral cavity of tobacco users and non-users. The results demonstrate the necessity of a reference normal spectral database of only tobacco non-users for training a diagnostic algorithm in order to make an accurate prediction of the pathology of the target tissue.

ACKNOWLEDGEMENTS

The authors would like to thank the nursing staff of the Head and Neck Surgery Department, Tata Memorial Hospital, Mumbai for their help and cooperation. They would also like to thank Dr. P. K. Gupta for his active support.

REFERENCES

[1] Sham A, Cheung LK, Jin LJ, Corbet EF. The effects of tobacco use on oral health. Hong Kong Med J 2003; 9: 271-7.

[2] Boffetta P, Hecht S, Gray N, Gupta P, Straif K. Smokeless tobacco and cancer. Lancet Oncol 2008; 9: 667-75. http://dx.doi.org/10.1016/S1470-2045(08)70173-6

[3] Rastogi T, Devesa S, Mangtani P, *et al.* Cancer incidence rates among South Asians in four geographic regions: India, Singapore, UK and US. Int J Epidemiol 2008; 37: 147-60. http://dx.doi.org/10.1093/ije/dym219

[4] American Cancer Society. Global cancer facts & figures. 3rd ed. Atlanta: American Cancer Society; 2015. Available from: http://www.cancer.org/research/cancerfactsstatistics/global

[5] El-Mofty S. Early detection of oral cancer. Egypt J Oral Maxillofac Surg 2010; 1: 25-31.

[6] Garg P, Karjodkar F. Catch them before it becomes too late-oral cancer detection. Report of two cases and review of diagnostic AIDS in cancer detection. Int J Prev Med 2012; 3: 737-41.

[7] Epstein JB, Zhang L, Rosin M. Advances in the diagnosis of oral premalignant. J Can Dent Assoc 2002; 68: 617-21.

[8] Mahadevan-Jansen A. Raman Spectroscopy: From Benchtop to Bedside. In: Vo-Dinh T editor. Biomedical photonics handbook, Washington DC: CRC Press, 2003; Chapter 30. http://dx.doi.org/10.1201/9780203008997.ch30

[9] Chen P, Shen A, Zhou X, Hu J. Bio-Raman spectroscopy: A potential clinical method assisting in disease diagnosis. Anal Methods 2011; 3: 1257-69. http://dx.doi.org/10.1039/c1ay05039g

[10] Lieber CA, Majumder SK, Ellis DL, Billheimer DD, Mahadevan-Jansen A. In-vivo non melanoma skin cancer diagnosis using Raman micro spectroscopy. Lasers Surg Med 2008; 40: 461-7. http://dx.doi.org/10.1002/lsm.20653

[11] Haka AS, Volynskaya Z, Gardecki JA *et al. In vivo* margin assessment during partial mastectomy breast surgery using Raman spectroscopy. Cancer Res 2006; 66: 3317-22. http://dx.doi.org/10.1158/0008-5472.CAN-05-2815

[12] Teh SK, Zheng W, Ho KY, Teh M, Yeoh KG, Huang Z. Diagnostic potential of near-infrared Raman spectroscopy in the stomach: differentiating dysplasia from normal tissue. Br J Cancer 2008; 98: 457-65. http://dx.doi.org/10.1038/sj.bjc.6604176

[13] Bergholt MS, Zheng W, Lin K *et al.* Combining near infrared-excited autofluorescence and Raman spectroscopy improves in-vivo diagnosis of gastric cancer. Biosens Bioelectron 2011; 26: 4104-10. http://dx.doi.org/10.1016/j.bios.2011.04.005

[14] Stone N, Kendall CA. Raman spectroscopy for early cancer detection, diagnosis and elucidation of disease specific biochemical changes. In: Pavel M, Morris M D editors. Emerging Raman Applications and Techniques in Biomedical and Pharmaceutical Fields, Berlin: Springer 2010; p. 315-46. http://dx.doi.org/10.1007/978-3-642-02649-2_13

[15] Malini R, Venkatakrishna K, Kurien J, *et al.* Discrimination of normal, inflammatory, premalignant, and malignant oral tissue: A Raman spectroscopy study. Biopolymers 2006; 81: 179-93. http://dx.doi.org/10.1002/bip.20398

[16] Li Y, Wen ZN, Li LJ, Li ML, Gao N, Guo YZ. Research on the Raman spectral character and diagnostic value of squamous cell carcinoma of oral mucosa. J Raman Spectrosc 2010; 41: 142-47.

[17] Guze K, Short M, Zeng H, Lermana M, Sonis S. Comparison of molecular images as defined by Raman spectra between normal mucosa and squamous cell carcinoma in the oral cavity. J Raman Spectrosc 2011; 42: 1232-9. http://dx.doi.org/10.1002/jrs.2838

[18] Sunder NS, Rao NN, Kartha VB, Ullas G, Kurien J. Laser Raman spectroscopy: A novel diagnostic tool for oral cancer. J Orofac Sci 2011; 3: 15-9.

[19] Deshmukh A, Singh SP, Chaturvedi P, Krishna CM. Raman spectroscopy of normal oral buccal mucosa tissues: study on intact and incised biopsies. J Biomed Opt 2011; 16: 127004. http://dx.doi.org/10.1117/1.3659680

[20] Devpura S, Thakur JS, Dethi S, Naik VM, Naik R. Diagnosis of head and neck squamous cell carcinoma using Raman spectroscopy: Tongue tissue. J Raman Spectrosc 2012; 43: 490-6. http://dx.doi.org/10.1002/jrs.3070

[21] Su L, Sun YF, Chen Y, Chen *et al.* Raman spectral properties of squamous cell carcinoma of oral tissues and cells. Laser Phys 2012; 22: 311-6. http://dx.doi.org/10.1134/S1054660X12010185

[22] Schut TCB, Witjes MJH, Sterenborg HJCM, *et al.* In-vivo detection of dysplastic tissue by Raman spectroscopy. Anal Chem 2000; 72: 6010-8. http://dx.doi.org/10.1021/ac000780u

[23] Oliveira AP, Bitar RA, Silveria L, Zangaro RA, Martin AA. Near-Infrared Raman spectroscopy for oral carcinoma diagnosis. Photomed Laser Surg 2006; 24: 348-53. http://dx.doi.org/10.1089/pho.2006.24.348

[24] Guze K, Short M, Sonis S, Karimbux N, Chan J, Zeng H. Parameters defining the potential applicability of Raman spectroscopy as a diagnostic tool for oral disease. J Biomed Opt 2009; 14: 014016. http://dx.doi.org/10.1117/1.3076195

[25] Bergholt MS, Zheng W, Huang Z. Characterizing variability in in-vivo Raman spectroscopic properties of different anatomical sites of normal tissue in the oral cavity. J Raman Spectrosc 2012; 43: 255-62. http://dx.doi.org/10.1002/jrs.3026

[26] Singh SP, Deshmukh A, Chaturvedi P, Krishna CM. *In vivo* Raman spectroscopic identification of premalignant lesions in oral cavity. J Biomed Opt 2012; 17: 105002. http://dx.doi.org/10.1117/1.JBO.17.10.105002

[27] Sahu A, Deshmukh A, Ghanate AD, Singh SP, Chaturvedi P, Krishna CM. Raman spectroscopy of oral buccal mucosa: A study on age-related physiological changes and tobacco-related pathological changes. Technol Cancer Res Treat 2012; 11: 529-41. http://dx.doi.org/10.7785/tcrt.2012.500304

[28] Sahu A, Tawde S, Venkatesh P, *et al.* Raman spectroscopy and cytopathology of oral exfoliated cells for oral cancer diagnosis, Anal Methods 2015; 7: 7548-59. http://dx.doi.org/10.1039/C5AY00954E

[29] Krishna H, Majumder SK, Muttagi S, Chaturvedi P, Gupta PK. *In vivo* Raman spectroscopy for detection of oral neoplasia: A pilot clinical study. J Biophotonics 2014; 7: 690-702. http://dx.doi.org/10.1002/jbio.201300030

[30] Krishna H, Majumder SK, Chaturvedi P, Gupta PK, Anatomical variability of in-vivo Raman spectra of normal oral cavity and its effect on oral tissue classification. Biomed Spectrosc Imaging 2013; 2: 199-217.

[31] Krishna H, Majumder SK, Gupta PK. Range-independent background subtraction algorithm for recovery of Raman spectra of biological tissue. J Raman Spectrosc 2012; 43: 1884-94. http://dx.doi.org/10.1002/jrs.4127

[32] Cohen J. Statistical power analysis for the behavioral sciences. 2nd edn. New Jersey: Lawrence Erlbaum Associates; 1988.

[33] Majumder SK, Gebhart SC, Johnson MD, Thompson R, Lin WC, Mahadevan-Jansen A. A probability-based spectroscopic diagnostic algorithm for simultaneous discrimination of brain tumor and tumor margins of normal brain tissue. Appl Spectrosc 2007; 61: 548-57. http://dx.doi.org/10.1366/000370207780807704

[34] Talukder A. Nonlinear feature extraction for pattern recognition applications. PhD Thesis, Pennsylvania: Carnegie Mellon University, 1999.

[35] Krishnapuram B, Cari L, Figueiredo MAT. Sparse multinomial logistic regression: Fast algorithums and generalization bounds. IEEE Trans Pattern Anal Machine Intell 2005; 27: 957-68. http://dx.doi.org/10.1109/TPAMI.2005.127

[36] Majumder SK, Keller MD, Boulos FI, Kelley MC, Mahadevan-Jansen A. Comparison of autofluorescence, diffuse reflectance, and Raman spectroscopy for breast tissue discrimination. J Biomed Opt 2008; 13: 054009. http://dx.doi.org/10.1117/1.2975962

[37] Hand DJ, Till RJ. A simple generalization of the area under the ROC curve for multiclass classification problems. Mach Learn 2001; 45: 171-86. http://dx.doi.org/10.1023/A:1010920819831

[38] Khanna A, Gautam DS, Mukherjee P. Genotoxic effects of tobacco chewing. Toxicol Int 2012; 19: 322-6. http://dx.doi.org/10.4103/0971-6580.103683

[39] Doll R, Peto R, Wheatley K, Gray R, Sutherland I. Mortality in relation to smoking: 40 years' observation on male British doctors. Br Med J 1994; 309: 901-11. http://dx.doi.org/10.1136/bmj.309.6959.901

[40] Wall MA, Johnson J, Jacob P, Benowitz NL. Cotinine in the serum, saliva, and urine of nonsmokers, passive smokers, and active smokers. Am J Public Health. 1988; 78: 699-701. http://dx.doi.org/10.2105/AJPH.78.6.699

[41] Al-Delaimy WK, Hair as a biomarker for exposure to tobacco smoke. Tob Control 2002; 11: 176-82. http://dx.doi.org/10.1136/tc.11.3.176

[42] Tipton DA, Dabbous MK. Effects of nicotine on proliferation and extracellular matrix production of human gingival fibroblasts *in vitro*. J Periodontol 1995; 66: 1056-64. http://dx.doi.org/10.1902/jop.1995.66.12.1056

[43] Taybos G. Oral changes associated with tobacco use. Am J Med Sci 2003; 326: 179-82. http://dx.doi.org/10.1097/00000441-200310000-00005

[44] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20: 273-297. http://dx.doi.org/10.1007/BF00994018

[45] Jolliffe IT. Principal Component Analysis. 2nd ed. New York: Springer; 2002.

[46] Duda RO, Hart PE,Stork DG. Pattern Classification. 2nd ed. New York: Wiley 2001.

Received on 19-04-2016 Accepted on 16-05-2016 Published on 10-08-2016

http://dx.doi.org/10.6000/1927-7229.2016.05.03.4