

Optical Spectroscopy and Imaging for Diagnosis of Oral Cavity Neoplasia

By

Hemant Krishna

Enrolment No. : PHYS03201304002

**Raja Ramanna Centre for Advanced Technology,
Indore-452013, India**

A thesis submitted to the

Board of Studies in Physical Sciences

In partial fulfillment of requirements

for the Degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE



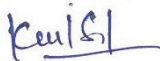
August, 2018

Homi Bhabha National Institute¹

Recommendations of the Viva Voce Committee

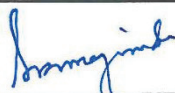
As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by Hemant Krishna entitled "Optical Spectroscopy and Imaging for Diagnosis of Oral Cavity Neoplasia" and recommend that it may be accepted as fulfilling the thesis requirement for the award of Degree of Doctor of Philosophy.

Chairperson – Prof. K. S. Bindra



Date: 18-09-2019.

Guide / Convener – Prof. S. K. Majumder



Date: 18/09/19

Examiner- Prof. Asima Pradhan



Date: 3/10/19

Member 1- Prof. S. R. Mishra

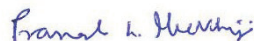
Date: S R Mishra
18-9-2019

Member 2- Prof. A. Moorti



Date: 18/9/2019

Member 3- Prof. P. K. Mukhopadhyay



Date: 18/9/2019

Member 4- Prof. C. Mukherjee



Date: 18/9/19

External Member - Prof. K. Divakar Rao



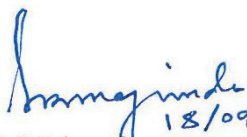
Date: 18/9/19

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to HBNI.

I hereby certify that I have read this thesis prepared under my direction and recommend that it may be accepted as fulfilling the thesis requirement.

Date: 18/09/19

Place: Indore


S. K. Majumder
18/09/19

¹ This page is to be included only for final submission after successful completion of viva voce.

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Hemant Krishna

DECLARATION

I, hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree / diploma at this or any other Institution / University.

Hemant Krishna

List of Publications arising from the thesis

Journals

1. “Range-Independent Background Subtraction Algorithm for Recovery of Raman Spectra of Biological Tissue”, **H. Krishna**, S. K. Majumder, and P. K. Gupta, *Journal of Raman Spectroscopy*, **Dec 2012**, 43(12), 1884-1894.
2. “Anatomical variability of *in vivo* Raman spectra of normal oral cavity and its effect on oral tissue classification”, **H. Krishna**, S. K. Majumder, P. Chaturvedi, and P. K. Gupta, *Biomedical Spectroscopy and Imaging*, **Aug 2013**, 2(3), 199-217.
3. “*In vivo* Raman spectroscopy for detection of oral neoplasia: A pilot clinical study”, **H. Krishna**, S. K. Majumder, S. Muttagi, P. Chaturvedi, and P. K. Gupta, *Journal of Biophotonics*, **Sep 2014**, 7(9), 690-702.
4. “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”, **H. Krishna**, S. Muttagi, P. Ingole, P. Chaturvedi and S. K. Majumder, *Journal of Analytical Oncology*, **Aug 2016**, 5(3), 110-123.

Manuscripts to be submitted:

1. “An optical spectroscopy based point-of-care diagnostic device for automated screening of oral neoplasia”, **H. Krishna**, and S. K. Majumder, *To be submitted*.
2. “A handy fluorescence imaging tool for improved visual assessment of oral cavity lesions”, **H. Krishna**, and S. K. Majumder, *To be submitted*.

Conferences

1. “Optical Spectroscopy for Oral Cancer Diagnosis”, **H. Krishna**, P. Ingole, P. Chaturvedi, D. Dharkar, S. K. Majumder and P. K. Gupta, *Proceedings of 21st Meeting of International Society for laser Surgery and Medicine*, Indore, **2015**, Aug., 19-22.
2. “Detection of Oral Cancer- The Photonics Way”, **H. Krishna**, and S. K. Majumder, *SMC Bulletin ISSN: 2394-5087*, **2015**, 6(2), 47-51.
3. “Development of a liquid-crystal tunable filter based hyper-spectral imaging system and a software interface for automated acquisition and analysis of spectral images”, **H. Krishna**, and S. K. Majumder, *Proceedings of DAE-BRNS National Laser Symposium-25*, KIIT University, Bhubaneswar, **2016**, Dec., 20-23.
4. “Studies on fluorescence photo-bleaching of urine for diagnosis of oral cancer”, **H. Krishna**, and S. K. Majumder, ” *Proceedings of DAE-BRNS National Laser Symposium-26*, BARC Mumbai, **2017**, Dec., 20-23.

5. “Development of image processing algorithm for better identification of neoplastic lesions in human oral cavity based on their fluorescence images”, **H. Krishna**, and S. K. Majumder, *Proceedings of eBBT*, IIT-Indore, **2018**, Jan. ,5-6.

Others

1. “Anatomical Variability of in-vivo Raman spectra of Oral cavity and its effect on tissue classification”, **H. Krishna**, and S. K. Majumder, *RRCAT Newsletter*, **2013**, 26(2), 21.
2. “Influence of tobacco consumption on the Raman Diagnosis of oral neoplasia”, **H. Krishna**, and S. K. Majumder, *RRCAT Newsletter*, **2016**, 29(2), 20.

Hemant Krishna

Inspired by the love
my wife Shikha,
and my daughter Deekshita

This thesis is dedicated to
my beloved parents

ACKNOWLEDGEMENTS

I owe special debt of gratitude to many people who contributed to the completion of most significant scientific accomplishment in my life.

I express my heartfelt gratitude and deep sense of indebtedness to my research guide Dr. S. K. Majumder for his continuous encouragement, valuable guidance and constant support in carrying out the research work. He not only suggested the current topic as a possible PhD project work but also taught me the art of experiments and science behind data analysis. The various conceptual and fruitful discussions with him from time to time have helped me strengthen my understanding of fundamentals in the field of Optical diagnosis. Also, I am highly thankful to him for his patience while going through the whole of the thesis page by page.

I am extremely grateful to Dr. Pankaj Chaturvedi, Dr. Muttagi Sidhharamesh, Dr. Pranav Ingole and the nursing staff of the Head and Neck Surgery Department, Tata Memorial Hospital, Mumbai for their active support and cooperation in carrying out clinical studies. I am highly thankful to Dr. S.S Nayyar (Raman Caner Foundation, Indore), Dr Noseen Kanchwala (MY Hospital, Indore) and Dr. Ameya (IIMS, Jodhpur), for their involvement and help in the clinical studies.

I am highly thankful to Laser Controls & Instrumentation Division, RRCAT, Indore, specially Shri V.P. Bhanage, Shri P.P. Deshpande, Shri S. Maskawade, and Shri A. Ansari for their support in electronic instrumentation and software implementations.

I am happy to acknowledge the support of Shri R.K. Gupta, Shri A.M. Kher, Shri S. Saleem, and Shri K. Panneerselvam, Laser Components Design & Fabrication Section, RRCAT, Indore for fabrication work involved in development of the systems.

It is my pleasure to thank Dr. K S. Bindra, Associate Director, Laser Group, RRCAT for his guidance and support during my Ph.D work. I take this opportunity to express my

sincere thanks to the Doctoral Committee members Dr. S. R. Mishra, Dr. A. Moorti, Dr. P. K. Mukhopadhyay, Dr. C. Mukherjee and Dr. K. Diwakar Rao for guidance and support during my Ph.D. work.

I would like to thank Shri Shankar V. Nakhe, Director Laser Group, RRCAT for his keen interest and constant support to this project work.

I would like to thank all colleagues of Laser Biomedical Applications Section, especially Dr. Khageswar Sahu Shri. Khan Mohd Khan, and Shri Vishwamita Sharama for their support and cooperation.

This project work would be impossible, were it not for much friendly support available within reach. For this I express my warm sense of thanks to all my friends for their moral support.

I would like to convey my deepest regards to my parents, brother and daughter for their love and support. Finally, I want to express my greatest appreciation to my wife Shikha for love, sacrifice and all the support during the completion of the thesis.

RRCAT, Indore

Hemant Krishna

August 2018

Summary	v
List of figures.....	vii
List of tables.....	xv
 Chapter 1	 1
Optical Spectroscopy and Imaging for <i>In-Vivo</i> Diagnosis of Oral Cancer.....	1
1.1 Introduction	1
1.2 Basics of Optical Diagnosis	2
1.3.1 Diffuse Reflectance	3
1.3.2 Fluorescence	4
1.3.3 Raman Scattering	5
1.3 Status Report on Optical Diagnosis of Oral Cancer	5
1.4 Specific Aims	9
 Chapter 2	 11
<i>In-Vivo</i> Raman Spectroscopy of Human Oral Cavity: Clinical Requirements	
2.1 Introduction	11
2.2 Development of Portable Raman System for Clinical <i>In-Vivo</i> Studies	12
2.3 Fluorescence Background Subtraction - Retrieval of Tissue Raman Signatures from the Measured Raw Spectra	13
2.3.1 Framework of the Range Independent Algorithm (RIA)	16
2.3.2 Validation of the Algorithm on Mathematical Phantom Raman Spectrum	18
2.3.3 Setting of Algorithm Parameters	22
2.3.4 RIA verses Polynomial Fitting based Algorithms – Performance Comparison	26
2.3.5 Performance Evaluation on Raw Spectra with Different Signal to Background Ratios	29
2.3.6 Performance Evaluation on Raw Spectra with Different Background Profiles	31

2.3.7 Validation of RIA on <i>In-Vivo</i> Raman Spectrum of Human Tissue	32
2.4 Data Analysis	35
2.4.1 Diagnostic Algorithm	35
2.4.2. Standard Error Confidence Interval	42
2.4.3. Pillai's V Measure	43
2.4.4. Multiclass Receiver Operating Characteristic Analysis	43
2.5 Summary	44
Chapter 3	47
<i>In-Vivo Raman Spectroscopy for Detection of Oral Neoplasia</i>	
3.1 Introduction	47
3.2 Clinical Protocol for the Studies	48
3.3 Study Design and Spectral Measurements	49
3.4 Data Pre-processing and Analysis	50
3.5 Results	52
3.6 Discussions	59
3.7 Summary	63
Chapter 4	65
Tobacco Consumption Induced Changes in the Healthy Oral Mucosa and its Effect on Differential Diagnosis of Oral Lesions	
4.1 Introduction	65
4.2 Study Design	66
4.3 Results	67
4.4 Discussions	78
4.5 Summary	82
Chapter 5	83
Anatomical Variability of <i>In-Vivo</i> Raman Spectra of Normal Oral Cavity and its Effect on Oral Tissue Classification	
5.1 Introduction	83

5.2 Materials and Methods	85
5.2.1 Study Design	85
5.2.2 Data Analysis	86
5.3 Results	89
5.4 Discussions	100
5.5 Summary	103
Chapter 6	105
A Comparison of Fluorescence and Raman Spectroscopy for Clinical Diagnosis of Oral Neoplasia	
6.1 Introduction	105
6.2 Materials and Methods	106
6.2.1 Clinical Spectroscopy Systems	106
6.2.2 Spectroscopic Measurements	107
6.2.3 Data Processing and Analysis	108
6.3 Results and Discussions	108
6.4 Summary	114
Chapter 7	115
Optical Spectroscopy and Imaging Based Point-of-Care Diagnostic Devices for Automated Screening of Oral Neoplasia	
7.1 Introduction	115
7.2 The Optical Spectroscopy based Point-of-Care Diagnostic Device	117
7.2.1 Device Hardware	117
7.2.2 Device Software	119
7.2.3 Methods of Data Analysis for Prediction of Diagnosis	125
7.2.4 Preparing the Device as a Stand-Alone Tool for Real-Time Diagnosis in a Clinical Setting	130
7.2.5 Point-of-Care Device as a Stand-Alone Tool for Real-Time Diagnosis in a Clinical Setting	134
7.3 Fluorescence Imaging based Point-of-Care Device for Improved Visual Assessment of Oral Cavity Lesions	136

7.3.1 Device Hardware and Software	137
7.3.2 Clinical Validation of the Device	139
7.4 Summary	142
Chapter 8	143
Summary and Future Perspectives	
References	151

SUMMARY

The thesis primarily focuses on the evaluation of the potential of optical spectroscopy and imaging for the diagnosis of human oral cavity cancer. The highlights of the work carried out as part of the thesis are as follows:

Successful use of Raman spectroscopy for tissue diagnostic applications requires an appropriate algorithm that can faithfully retrieve weak tissue Raman signals from the intense background of the measured raw Raman spectra. In the most widely used iterative polynomial fitting based algorithms, the lineshape and intensity of the extracted Raman spectra are found to depend on the range of the wavenumbers selected for the fitting. We have developed an iterative smoothening based novel background subtraction algorithm that does not depend on the selection of the range of wavenumbers of the raw Raman spectrum unlike the iterative polynomial fitting algorithms.

We have developed a portable Raman spectroscopy system for clinical *in-vivo* studies. The system was used to carry out *in-vivo* studies at Tata Memorial Hospital (TMH), Mumbai with the approval of TMH ethical committee. The *in-vivo* Raman spectra were recorded from oral cavity of ~ 200 individuals. The different tissue sites investigated belonged to either of the four histopathologic categories: 1) squamous cell carcinoma (OSCC), 2) sub-mucosal fibrosis (OSMF), 3) leukoplakia (OLK) and 4) normal squamous tissue. All the measured *in-vivo* tissue spectra were analyzed for differential diagnosis of oral lesions using a probability based multi-class diagnostic algorithm. The best classification accuracy was observed to be 85%, 89%, 85% and 82% in classifying the oral tissue spectra into four different pathology classes- normal, OSCC, OSMF and OLK respectively, on the basis of leave-one-individual-out cross-validation, with an overall accuracy of 86%. Further, the influence of tobacco habits on the Raman characteristics of healthy oral mucosa was investigated. It was found that exclusion of the spectral data of the healthy volunteers with tobacco habits from the reference

normal database considerably improved the overall classification accuracy (92 % as against 86%) of the algorithm in separating the oral lesions from the normal oral mucosa. We have also characterized the variability of the measured Raman spectra recorded from the different anatomical sites of the oral cavity of healthy volunteers and showed that they could be mainly grouped into four distinct anatomical clusters. Further analysis showed that when inter-anatomical variability was taken into account and the performance of the algorithm was checked on the spectra in each of the anatomical groups, the overall classification accuracy was found to improve by 7%.

A side-by-side comparison of the relative performances of fluorescence and Raman spectroscopy was carried out for *in-vivo* discrimination of various oral tissue pathologies in a clinical setting. The results showed that while the simultaneous multi-class classification accuracy for the Raman was significantly better than the fluorescence (86 % as against 77%), the binary (normal vs. abnormal) classification accuracy for both the systems was comparable (> 90%).

Based on these results, we developed two point-of-care devices for screening of oral human oral cavity, one based on optical spectroscopy, and the other based on optical imaging. The optical spectroscopic device (named OncoDiagnoScope) is a tablet computer based, compact and portable instrument capable of providing real-time diagnostic feedback about the oral tissue under interrogation. The device was validated on patients with oral neoplasia in various hospitals and cancer screening camps and found to detect cancer with an accuracy of over 90%. The other point-of-care device (named Vision Enhancement Module) is a handy fluorescence imaging tool for real-time, non-contact and *in-situ* imaging of fluorescence from human oral cavity intended for improved visual assessment of the oral cavity. Using this instrument, regions of oral lesions can be better identified against the healthy oral tissues based on their natural characteristics in response to light.

LIST OF FIGURES

2.1 (a) A photograph and (b) a schematic of the portable clinical Raman spectroscopy system for <i>in-vivo</i> Raman measurements.....	13
2.2 Flow chart of the Range-Independent Algorithm (RIA) for background subtraction..	16
2.3 Mathematically generated synthetic Raman spectrum, (a) the background, (b) the Raman signal simulated with twelve Lorentzian peaks on a null baseline, and (c) Raman spectrum mimicking a typical raw tissue Raman spectrum.	20
2.4 Pictorial demonstration of the working of the RIA. (a) The raw Raman spectrum (bold line) and truncated spectrum in the range of interest (fine line with an offset), (b) truncated spectrum with linear extension (bold line) and added Gaussian peaks (fine line with an offset), (c) the background curves at the end of 10 th , 100 th and final round of iterations, and (d) the recovered Raman signal (gray line), synthetic Raman signal (black dashed line), and residual (black fine line). The final estimated background is also shown by fine line with an offset for clarity in panel (c).	21
2.5 The Raman signal recovered by the RIA from the synthetic raw Raman spectrum (a) for three different heights of the added Gaussian peaks: (i) normal (light gray line) equal to the maximum of the ordinate value of the raw spectrum, (ii) sub-normal equal to one hundredth of normal (dark gray line), and (iii) above-normal equal to ten times normal (black dashed line), (b) for five different full-width-at-half-maximum (FWHM) values of the added Gaussian peaks, and (c) for five different spans of the moving point average (MPA) filter, where N is the number of data points included in the span.	24

2.6 Raw synthetic Raman spectrum and the Raman signals recovered from it using (b) the RIA, (c) the ModPoly, and (d) the I-ModPoly for three different spectral ranges: (i) range-1 corresponding to $800-1800\text{ cm}^{-1}$, (gray bold line) (ii) range-2 corresponding to $980-1580\text{ cm}^{-1}$ (black fine line), and (iii) range-3 corresponding to $1150-1750\text{ cm}^{-1}$ (black dashed line).	28
2.7 Raw synthetic Raman spectra generated with signal to background ratio (SBR) of (a) 0.01, (b) 0.1, and (c) 1.0. The recovered Raman signal (black line) using the RIA and the residual (gray line) for the corresponding SBR values are shown in panels (d) , (e) , and (f) respectively. The residuals are shown with an offset for clarity in the corresponding panels.	30
2.8 Raw synthetic Raman spectra with (a) exponential, (b) Gaussian, and (c) sigmoidal backgrounds. The Raman signals recovered by the RIA (black line) from the respective raw spectra are shown in the panels (d) , (e) , and (f) respectively. The corresponding residuals (gray line) are shown with an offset.	31
2.9 (a) The experimentally measured raw <i>in-vivo</i> Raman spectra, and (b) the Raman spectra following background subtraction using the RIA for oral (upper row), and skin (lower row) tissues. For the sake of clarity the spectra are shown with an offset.	32
2.10 The Raman spectra recovered from the experimentally measured raw Raman spectrum (shown in Fig. 2.10a) of (a) oral tissue, and (b) skin tissue using the RIA, the ModPoly, and the I-ModPoly for three different spectral ranges: (i) range-1 corresponding to $950-1750\text{ cm}^{-1}$, (gray bold line) (ii) range-2 corresponding to $980-1520\text{ cm}^{-1}$ (black fine line), and (iii) range-3 corresponding to $1200-1700\text{ cm}^{-1}$ (black dashed line).	33

2.11 Flow chart for the implementation of the probability based multivariate diagnostic algorithm.	39
3.1 Mean, normalized Raman spectra of OSCC (n=316), OSMF (n=94), OLK (n=105), and normal (n=287) oral tissue sites. The error bars (gray) represent ± 1 standard deviation.	52
3.2 Mean difference spectra showing statistical differences between different pathologies and normal oral tissue spectra. Gray bands indicate the 90% confidence intervals of the difference determined by standard error confidence intervals.	53
3.3 Posterior probabilities for being classified as normal, OSCC, OSMF, and OLK for the Raman spectra of the oral tissue sites interrogated. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.	57
3.4 Posterior probabilities for being classified as normal, potentially malignant (OSMF & OLK grouped together), and malignant (OSCC) for the Raman spectra of the oral tissue sites interrogated. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.	57
3.5 Posterior probabilities for being classified as (a) normal and malignant (OSCC), (b) normal and potentially malignant (OSMF and OLK grouped together), and (c) normal and abnormal (OSCC, OSMF and OLK grouped together) for the Raman spectra of the oral tissue sites interrogated. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.	58

4.1 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.	67
4.2 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.	68
4.3 Posterior probabilities of being classified as normal without any tobacco consumption habit (N:WOTH) and normal with tobacco consumption habit (N:WTH).	70
4.4 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.	71
4.5 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.	74
4.6 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.	77

5.1 Mean, normalized Raman spectra of the different anatomical sites of oral cavity of healthy volunteers. The error bars (gray) represent ± 1 standard deviation.	90
5.2 Ratio of within-class to between-class separation as a function of number of clusters. The minimum represents the optimum number of clusters (=4).	91
5.3 Mean, normalized Raman spectra of the different anatomical sites of normal healthy tissues belonging to the four different anatomical clusters (AC): (i) AC-I: outer lip and lip vermillion border; (ii) AC-II: buccal; (iii) AC-III: hard palate, and (iv) AC-IV: dorsal, lateral, ventral tongue and soft palate.	93
5.4 Mean pair-wised difference spectra showing statistical differences between the four anatomical clusters. Gray bands indicate the 95% confidence intervals of the difference determined by standard error confidence intervals.	94
5.5 Mean, normalized Raman spectra of malignant, potentially malignant and normal oral tissue sites. The error bars (gray) represent ± 1 standard deviation.	95
5.6 Posterior probabilities computed by the MRDF-SMLR algorithm for the measured tissue spectra of each tissue class of belonging to that particular class (a) when anatomical clustering was considered and (b) when anatomical clustering was not considered. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.	99
6.1 (a) Schematic, and (b) setup of N ₂ laser based fluorescence spectroscopy system.	106

6.2 Mean (a) autofluorescence spectra, and (b) Raman spectra of human oral tissues belonging to different pathology categories. Spectra are plotted with one standard deviation to represent inter-patient variability.	109
6.3 Posterior probabilities for being simultaneous multiclass classification as normal, OSCC, OSMF and OLK for the (a) fluorescence, and (b) Raman spectra of the human oral cavity. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.	112
6.4 Posterior probabilities for being classified as normal and abnormal (OSCC, OSMF, and OLK) for the (a) fluorescence, and (b) Raman spectra of the human oral cavity. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.	113
7.1 OncoDiagnoScope (a) developed system, and (b) schematic.	118
7.2 The graphic user interface of the OncoDiagnoScope indented to be used for configuration of the system.	120
7.3 The graphic user interface of the OncoDiagnoScope system for the real-time display of tissue optical spectra, documentation and diagnosis. The total time for data acquisition and analysis was under 7 s.	121
7.4 The graphic user interface of the OncoDiagnoScope indented to be used for the offline review and analysis of the saved data.	124

7.5 Mean, normalized (a) fluorescence and (b) diffuse reflectance spectra of OSCC, OSMF, OLK, and normal oral tissue sites. The error bars represent ± 1 standard deviation at that wavelength.	132
7.6 (a) Vision Enhanced Module (VEM) and (b) Graphic User Interface (GUI) software developed for VEM.	138
7.7 Representative fluorescence images of difference oral tissue pathologies as acquired using Vision Enhanced Module (VEM).	141
7.8 Demarcated area of the suspected tissue sites present in the corresponding fluorescence images of the Fig. 7.7.	141

LIST OF TABLES

2.1 The parameters of Lorentzian function used for constructing the synthetic Raman spectrum. Here, r_0 is the position of the Lorentzian peak, ω_0 is the full-width at half-maximum (FWHM) and A_0 is the total area under the curve above the baseline.	19
2.2 A comparison of the correlation of the mathematically simulated Raman peaks with the Raman peaks recovered by the RIA, the ModPoly and the I-ModPoly algorithms from the mathematically generated raw Raman spectrum over different spectral ranges.	29
3.1 Histopathological distribution of tissue sites included in the clinical pilot study.	50
3.2 Confusion matrix displaying classification of the Raman spectra of oral tissue sites into four classes: normal, OSCC, OSMF, and OLK using the MRDF-SMLR based diagnostic algorithm.	54
3.3 Confusion matrix displaying results of classification of the Raman spectra of oral tissue sites into three classes: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC) using the MRDF-SMLR based diagnostic algorithm.	55
3.4 The classification results of the Raman spectra of oral tissue sites into two classes using the MRDF-SMLR based diagnostic algorithm. Here the spectra belonging to OSCC are referred to as “Malignant”, those belonging to OSMF and OLK pooled	

together are referred to as “Potentially Malignant”, and the spectra belonging to OSCC, OSMF and OLK pooled together are referred to as “Abnormal”.	56
3.5 The results of ROC analyses for binary, three-class and four-class classifications using the MRDF-SMLR based diagnostic algorithm.	59
4.1 Histopathological distribution of tissue sites included in the study.	66
4.2 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.	69
4.3 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.	73
4.4 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.	73
4.5 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 using the	

MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.	74
4.6 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.	75
4.7 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.	76
4.8 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).	77
5.1 Histopathological distribution of tissue sites included in the <i>in-vivo</i> study.....	86
5.2 The number of spectra of different anatomical locations comprising the four different clusters resulted from Fuzzy c-means cluster analysis.	92
5.3 Results of unsupervised classification in the form of confusion matrix displaying the comparison of predicted membership with the actual membership of the four anatomical clusters. DT: dorsal tongue, LT: lateral tongue, VT: ventral tongue, SP:	

soft palate, OL: outer lip, LVB: lip vermillion border, HP: hard palate and BM: buccal mucosa. N: total number of tissue sites belonging in a group.	92
5.4 Confusion matrices displaying results of classification of the Raman spectra of oral tissue sites, into three classes: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC) using the MRDF-SMLR based diagnostic algorithm for different anatomical clusters.	96
5.5 Confusion matrix displaying results of anatomy matched overall classification of the Raman spectra of oral tissue sites into three classes: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC) using the MRDF-SMLR based diagnostic algorithm. Each element of the matrix was obtained by computing the sum of the corresponding elements of the matrices listed in Table-5.4 and dividing the sum by the total sum of the elements over the corresponding rows multiplied by 100.	97
5.6 Confusion matrix displaying results of classification of the Raman spectra of oral tissue sites into three classes: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC) using the MRDF-SMLR based diagnostic algorithm without considering anatomical clustering.	97
5.7 Pillai's V and HTM values for the set of Raman spectra belonging to different pathology classes before and after anatomical clustering. While Pillai's V is a quantitative measure of the separation between different pathology classes, HTM is a measure of the performance of a diagnostic algorithm.	98
6.1 Confusion matrices displaying the results of simultaneous multi-class classification of the fluorescence and Raman spectra of oral tissue sites into four classes: normal,	

OSCC, OSMF, and OLK using the non-linear MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.	111
6.2 Confusion matrices displaying the results of binary classification of the fluorescence and Raman spectra of oral cavity into normal and abnormal tissue using the non-linear MRDF-SMLR based diagnostic algorithm. Here the spectra belonging to OSCC, OSMF and OLK pooled together are referred to as “Abnormal”.	113

Chapter 1

Optical Spectroscopy and Imaging for In-Vivo Diagnosis of Oral Cancer

1.1. Introduction

Oral cancer is one of the most common cancers in India and other South-Asian countries [1], [2]. India tops in the prevalence of oral cancer in the world, and use of tobacco and pan masala are known to account for this large incidence of oral cancer [3]–[7]. More than 35% of Indians (48% males and 20% females) above the age of 15 years either smoke or use tobacco in any form [8] and nearly 0.9 million tobacco-related deaths occur in India annually as compared to 1.5 million worldwide [1], [9]. Although the oral cavity is easily accessible to inspection, patients with oral cancer most often present themselves at an advanced stage when treatment is less successful thereby leading to high morbidity and mortality [10], [11]. Early detection of oral cancer remains the best way to ensure patient survival and quality of life [12]. Currently, the most widely used method for screening of oral cancer is visual examination of the oral cavity [13]. The process, being subjective, often fails to satisfactorily detect changes in oral mucosa associated with early cancers or pre-cancerous alterations that generally precede invasive cancers. Although, over the years, several new approaches like the

use of brush cytology, toluidine blue staining, Lugol iodine and Methylene blue solution, reflectance visualization after acetic acid application, and illumination with a chemiluminescent light source have been proposed to address the limitations of the conventional oral examination [14]–[16], a high rate of false positives are generally associated with these methods. Thus, there is an urgent current need for alternative diagnostic methods that can enhance the visual assessment of oral lesions and particularly help discriminate malignant and potentially malignant lesions from the normal oral mucosa.

Optical spectroscopy, in recent years, has been suggested and validated as a powerful alternate tool that can help solve all these problems. Its major attraction stems from its ability to provide biochemical and morphological information about the tissue in a minimally invasive or non-invasive manner [17-52]. Most of the work on optical diagnosis of cancer has been performed using fluorescence spectroscopy [17-31], diffuse reflectance or elastic scattering spectroscopy [32]–[36] and Raman spectroscopy [37-52]. Of these, fluorescence spectroscopy is the one that has been most widely investigated for the detection of oral lesions [21], [22], [24]–[31] followed by diffuse reflectance spectroscopy either alone or in combination with fluorescence spectroscopy [32]–[34], [36]. Only in recent years Raman spectroscopy has joined the league [38], [42], [45]–[51]. However, the applications of Raman spectroscopy in this area are still at the early stages of development [37], [39]–[41], [51], [52].

1.2 Basics of Optical Diagnosis

When light is incident on a tissue, the light photons either get reflected from the surface of the tissue (specular reflection) or get transmitted through it. The transmitted photons while propagating inside the tissue either get absorbed (‘absorption’) or emerge out of the surface after undergoing many scattering events (‘diffuse reflectance’) or transmit through the

boundaries ('transmission'). The absorbed photon may be reemitted as fluorescence light. Scattering of the photon can be either elastic (without change in the wavelength) or inelastic Raman scattering (with change in the wavelength). The underlying hypothesis of optical diagnosis of cancer is the fact that during disease progression, the changes in optical properties of tissue can be interrogated by these optical means such as fluorescence, diffuse reflectance, Raman and their combinations using the visible, infrared and near infrared light. A brief description of the underlying hypothesis on use of these techniques for tissue diagnostic is presented here.

1.2.1 Diffuse Reflectance

Diffuse reflectance spectroscopy measures light that undergoes multiple elastic scattering events within the tissue before emerging out of the tissue surface, thereby providing morphologic information about the tissue [32]–[34]. It bears the signature of absorption by various chromophores present within the tissue, thus providing some amount of biochemical information about these absorbing chromophores. The light in the ultra violet (UV) spectral region below 400 nm is strongly absorbed by the DNA base pairs and structural proteins (collagen and elastin), co-enzymes and lipids, whereas the visible and near infrared (NIR) light is mainly absorbed by different forms of hemoglobin (oxygenated and deoxygenated), porphyrins, melanin and water present in the tissue. The measured diffuse reflectance is also sensitive to the changes in the sub-cellular architecture (associated with disease transformation), which generally get reflected in the observed morphological parameters (such as nuclear to cytoplasm ratio, mitochondrial size and density etc.) used for histological assessment of tissues [19]. It is also sensitive to the ultra-structural features that are beyond the limit of Abbe resolution for optical microscopy, and thus provides information which otherwise is not possible to obtain from conventional histology [19].

1.2.2 Fluorescence

In fluorescence, the chromophores present in the target tissue following excitation to a higher electronic state on absorption of light of short wavelengths (typically UV or blue), de-excite to their ground electronic state with reemission light of higher wavelengths (i.e. lower energy). For fluorescence diagnostic applications, these chromophores are either endogenously present in the tissue or can be administered exogenously to enhance fluorescence. Tissue contains several endogenous fluorescent bio-molecules such as structural proteins (collagen and elastin), coenzymes (NADH, FAD), aromatic amino acids (tryptophan, tyrosine, and phenylalanine) and porphyrins [53]. The {excitation; emission} maxima pairs for these fluorophores are known to be at {280 nm; 350 nm} for tryptophan, {275 nm; 300 nm} for tyrosine, {260 nm; 280 nm} for phenylalanine, {325 nm; 400 nm} for collagen and elastin, {351 nm; 460 nm} for NADH, {450 nm; 535 nm} for FAD and {400-450 nm; 630 and 690 nm} for porphyrins [18], [53], [54]. By measuring the UV-induced fluorescence of native tissue (often called “autofluorescence”) it should, in principle, be possible to learn about the relative concentrations and redox states of such compounds and in turn to learn about the biochemical state of the tissue. However, interpretation of tissue autofluorescence is complicated by the intrinsic scattering and absorption properties of the tissue, rendering autofluorescence measurements significantly more complicated than measurements of fluorophores in dilute solution [32], [53]. Despite these difficulties, several studies have shown that fluorescence spectroscopy can be used for clinical diagnosis of cancers [22], [55]. Either single-point or imaging measurements have been performed. While the single-point spectroscopy is used to get the detailed spectral information from one localized tissue site at a time, the spectral imaging method provides a comprehensive spectroscopic information from a larger area of tissue surface [35], [55].

1.2.3 Raman Scattering

Raman scattering is an inelastic scattering process which probes the vibrational energy levels of molecules and specific peaks in the Raman spectrum denote particular chemical bonds or bond groups of the molecules. In fact, Raman scattering from a biological sample provides more comprehensive information about the “molecular fingerprint” of the biomolecules present in the sample as compared to the information obtained using fluorescence or diffuse reflectance [44], [50], [52]. Four important components of biological tissues that contribute to the Raman spectra are proteins, lipids, nucleic acids and water [47], [52]. Because of the chemical specificity of Raman scattering, it has the ability to discern the subtle biochemical changes associated with disease transformation thus making it particularly suited for diagnostic applications. However, early attempts at measuring *in-vivo* Raman spectra were difficult because of the fluorescent nature of the biological tissues and the limitations related to light sources and detectors [43], [44], [52]. With continued improvements in detector technologies over last two decades, it has now become possible to acquire good quality tissue Raman spectrum in a clinically acceptable data collection time [43], [44], [52]. Further, portable, powerful and stable diode lasers emitting wavelengths in the "optical window" of tissue, such as 785 nm and 830 nm, at which they generate minimal background fluorescence while penetrating fairly deeply into tissue are now readily available.

1.3. Status Report on Optical Diagnosis of Oral Cancer

The different modalities of the optical spectroscopy such as diffuse reflectance, fluorescence, combined diffuse reflectance and fluorescence and Raman spectroscopy have been extensively studied for *ex-vivo* and *in-vivo* diagnosis of oral cancer. We present an overview of some representative studies to indicate the present state-of-art of optical spectroscopy and imaging for oral cancer diagnosis.

Some of the earliest works on optical diagnosis were carried out using fluorescence spectroscopy by the MIT group [19], [27] as well as our group [28] to interpret the observed differences in the UV-induced fluorescence in oral tissue of different pathological states. The fluorescence spectra recorded with excitation wavelengths of 337, 365, and 410 nm have been demonstrated to produce the greatest separation of neoplastic and dysplastic tissue from normal oral tissue [27]. Using 337 nm as the excitation wavelength, our group also showed that the autofluorescence intensity was considerably less for malignant as compared to normal oral tissue [29]–[31], [56]. The decrease in the fluorescence intensity was attributed to the alterations in tissue architecture or biochemical composition during neoplastic progression of oral tissue [27], [56]. The translational research on fluorescence spectroscopy for *in-vivo* oral cancer detection include several small clinical studies which have further validated the fact that the fluorescence intensity from healthy oral mucosa is generally greater than that from the abnormal one [21], [22], [26], [33]. Algorithms developed on the basis of the differences in fluorescence spectra could discriminate healthy mucosa from dysplastic and cancerous tissue with high sensitivity and specificity [22], [29]–[31]. Everything put together, these findings suggest that fluorescence can be utilized to develop simple and objective tools for *in-vivo* identification of oral neoplasia [21], [29], [51]. Recently, the U.S. Food and Drug Administration approved an autofluorescence imaging device for early detection of oral neoplasia [25]. The commercial device, marketed as the VELscope® (LED Dental, Inc., White Rock, BC, Canada), has a blue/violet light (400 – 460 nm wavelengths) for illuminating oral tissue and optical filters (long pass and notch filters) to enable doctors directly visualize fluorescence in the oral cavity [25]. The VELscope and other fluorescence imaging devices [57] proposed in the literature rely on qualitative observations for detecting and delineating neoplastic oral lesions. These instruments, therefore, require well-defined and

standardized image interpretation criteria, and appropriate user training for reliable screening of the oral cavity to detect the presence of abnormal lesions.

Diffuse reflectance spectroscopy, which provides a method to examine the changes in the elastic scattering and absorption properties of tissue, has also been reported to be a promising tool for detection of premalignant and malignant changes in the oral tissues [34], [36]. For example, while an increased scattering is reported in case of hyperplasia and hyperkeratosis [58], a decrease in scattering and oxygen saturation values along with an increase in blood content and scattering slope have been demonstrated in malignant transformation of oral tissues [36]. Mallia *et al.* [34] have conducted a clinical trial and used the ratio of diffuse reflectance intensities at oxygenated hemoglobin absorption dips at 545 and 575 nm as the parameter for the diagnosis of oral cancer. Several clinical *in-vivo* studies have used fluorescence either alone or in combination with diffuse reflectance spectroscopy with very promising results [32], [33]. Veld *et al.* [32] have used various normalization methods to correct fluorescence spectra for intrinsic scattering and absorption properties of the tissue with the help of measured diffuse reflectance, and compared the diagnostic potential of fluorescence, diffuse reflectance and corrected fluorescence and combination of these. Combination of autofluorescence and diffuse reflectance is reported to produce improved performance for distinguishing benign from dysplastic and malignant oral lesions, suggesting complementary nature of fluorescence and diffuse reflectance for improved diagnosis [32].

Earlier report on the use of Raman spectroscopy for diagnosis of oral cancer were limited to *ex-vivo* studies [59]–[65] or *in-vivo* studies on animal models [66], [67]. Later, technological improvements in detectors and light sources have motivated several groups [44], [47] including ours [20] to start *in-vivo* Raman studies on the human oral cavity to investigate its efficacy for differential diagnosis of various oral tissue pathologies. The first

in-vivo study was reported by Guze *et al.* [48], who measured Raman spectra in the higher wave number ($1800 - 3000 \text{ cm}^{-1}$) 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.* [47], who also acquired Raman spectra from the healthy volunteers but in the conventional fingerprint region ($800 - 1800 \text{ cm}^{-1}$) 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.* [45], who used a commercial Raman spectrometer to measure *in-vivo* Raman spectra from patients already identified of having malignancy of oral buccal mucosa and showed that Raman spectroscopy in combination with diagnostic algorithm could delineate malignant from the uninvolved normal tissue sites as well as the premalignant lesions appearing in the contra lateral buccal mucosa of the same set of patients with an accuracy of up to 87%. In a concurrent study [49] they investigated the influence of aging related physiological changes on the differential detection of malignant, premalignant and normal buccal mucosa 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 [49]. If the results prove to be good enough particularly for detecting pre-cancerous and early cancerous changes, then Raman spectroscopy will stand the chance to be a promising alternate tool that can improve diagnostic specificity and sensitivity and will be worth tracking in coming years.

1.4. Specific Aims

The objective of the present work is to carry out *in-vivo* studies on the use of optical spectroscopy and imaging for the diagnosis of human oral cavity neoplasia.

The specific aims of the present dissertation include:

1. To evaluate the applicability of *in-vivo* Raman spectroscopy for discrimination of oral lesions from health oral tissues in a clinical setting and also to investigate the effects of tobacco consumption and anatomical variability on the measured Raman spectral signatures as well as the outcome of differential diagnosis of the oral lesions.
2. To carry out a comparative evaluation of the relative diagnostic performances of *in-vivo* fluorescence and Raman spectroscopy for differential diagnosis of various lesions of human oral cavity
3. To develop optical spectroscopy and imaging based point-of-care diagnostic devices capable of real time, non-invasive detection of oral lesions.

Chapter 2

In-Vivo Raman Spectroscopy of Human Oral Cavity: Clinical Requirements

2.1 Introduction

In recent years, Raman spectroscopy has demonstrated its potential as a promising new tool for tissue analysis [39], [40], [44], [51], [52], [68]. It has the intrinsic ability to discern subtle biochemical changes associated with neoplastic transformation thus making it particularly suited for diagnosis of cancer [44], [52]. However, the use of *in-vivo* Raman spectroscopy in a clinical setting has earned relatively less attention because of the necessity of large data collection time owing to the much weaker Raman signal. With continued improvements in detectors and spectroscopic systems, it has now become possible to acquire good quality tissue Raman spectrum with a reasonable integration time [37], [39], [40], [52], [69]. In this chapter, we will describe the development of a portable near infrared (NIR) Raman system that is capable of recording Raman spectra from human tissues in a clinically acceptable data collection time and can be used for *in-vivo* clinical studies in patients. Further, another challenging problem in using Raman spectroscopy for all tissue applications is to extract

rather weak tissue Raman signals from the measured raw spectra having orders of magnitude stronger background fluorescence, the primary reason why most researchers in the field have moved to the NIR wavelengths for excitation [44], [52]. We have developed a novel scheme for fluorescence background removal that overcomes the shortcomings of the existing methods based largely on polynomial fittings and yields in a faithful recovery of Raman signal for biological tissues from the measured raw spectrum. The method is based on modified iterative moving point average smoothing of the measured raw spectra. A notable feature of the method is that it is insensitive to the choice of the fitting range which allows one to obtain range-independent and artifact-free tissue Raman signal with zero baseline. The details of this method will be presented in this chapter. Finally, a brief description of the various statistical methods used for data analyses and diagnostic algorithms employed for differential diagnosis will also be presented.

2.2 Development of Portable Raman System for Clinical *In-Vivo* Studies

To measure *in-vivo* Raman spectra from human oral cavity a compact and portable Raman spectroscopic system was assembled in-house. A photograph and a schematic of the system are shown in Figs 2.1a and b respectively. The portable clinical system has all its sub-systems (diode laser, fiber-optic probe, spectrograph and CCD camera) accommodated into a 32'' suitcase. A 785 nm diode laser (Crysta Laser, Reno, NV) is used to deliver excitation light (~80 mW) to the target tissue using a fiber-optic probe (Visionex Inc., Warner Robins, GA) consisting of a central 400- μm -core-diameter fused-silica excitation fiber surrounded by seven 300- μm fused-silica beam-steered collection fibers. While the distal ends of the collection fibers have in-line notch filters for rejection of the excitation light, a band-pass filter sitting at the tip of the excitation fiber blocks the signals generated in the fiber itself and allows only the laser line to pass through. 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°C), back-illuminated, deep-depletion charge-coupled-device (CCD) camera (Andor DU420A-BR-DD, Belfast, Northern Ireland). The system is able to acquire good quality tissue Raman spectra with signal to noise ratio 50:1 for an integration time of less than 5 seconds. The overall resolution of the system is $\sim 20 \text{ cm}^{-1}$.

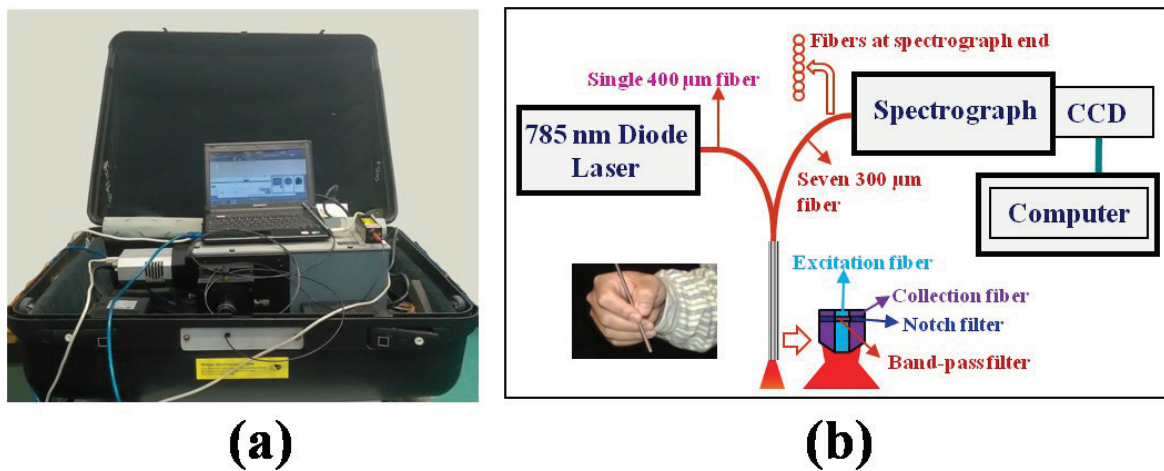


Fig. 2.1: (a) A photograph and (b) a schematic of the portable clinical Raman spectroscopy system for in-vivo Raman measurements.

2.3 Fluorescence Background Subtraction - Retrieval of Tissue Raman Signatures from the Measured Raw Spectra

The first and foremost requirement to be able to use Raman spectroscopy for tissue analysis is an appropriate technique that can faithfully extract rather weak Raman signals from the broad, orders of magnitude stronger, background believed to be arising from fluorescence and/or the scattering tail of the laser line [39], [40], [44], [69]. Over the years a variety of methods of varying rigor have been proposed for correcting the background and isolating the weak Raman signal [70-95]. One approach to minimize background interference is to acquire

Raman spectra at two or more slightly shifted [70]–[73] or continuously modulated excitation wavelength [74] and use mathematical processing [75] to retrieve the actual Raman spectra from the set of measured spectra. The approach [70]–[74] exploits the fact that the broad background is not affected by a change of excitation wavelength but the Raman spectrum is frequency shifted. Another approach [76]–[80] is to exploit the differences in times-scales of response of the Raman and the associated background signal and use time-gating for filtering out the Raman signal. However, both of these approaches require incorporating some modifications in the layout of the conventional dispersive Raman systems routinely employed for tissue diagnosis. This, in general, adds to the bulk, sophistication and in turn, cost of the systems which may not be desired in view of its end use as a clinical diagnostic tool. Moreover, another bottleneck that they commonly share is the long data acquisition time that makes their use almost impractical in a clinical situation.

Another popular approach is to make use of methods based on mathematical post-processing of the measured raw spectra for offline extraction of the Raman signal. Various mathematical techniques used for this purpose include manual estimation by visual inspection [81], full-matrix methods such as orthogonal signal correction (OSC) [82], multiplicative signal correction (MSC) [83], frequency-domain filtering such as Fourier Transform [70], Wavelet transforms [84]–[86] and rolling cycle filtering [87], derivative preprocessing such as computing first and second-order derivatives [88], [89], and polynomial fitting [69], [90]–[92]. Though each of these methods has been shown to be useful in certain situations, polynomial fitting [69], [90]–[95] is the most popular and widely used method among the Raman community dealing with tissue diagnostic applications. In this method, the background is estimated with the help of a single polynomial function whose order is selected based on empirical experience so as to best represent the contour of the background to be subtracted from the measured spectrum. However, the line shapes of the extracted Raman

spectra are found to vary significantly with the change in the order of the polynomial function. In order to overcome this limitation, Lieber and Mahadevan-Jansen [69] proposed a modified multi-polynomial fit (ModPoly) based iterative algorithm that was found to largely reduce the dependence of the spectra on the polynomial order. The algorithm was further improved by Zhao *et al.* [94] who added a peak-removal procedure during the first iteration and a statistical method to take into account the effect of the signal to noise ratio on the fitting. The algorithm was found to substantially improve background correction particularly for the spectra with high noise or less intense Raman peaks [94]. The significant advantages that both the algorithms offer are that they are automated, simple to use, quick and most importantly preserve to a large extent the spectral contours and intensities of the extracted Raman bands [69], [94]. Although the modified polynomial based algorithms are superior to most of the remaining background subtraction algorithms [95] in their ability for simple and effective background reduction and are extensively used, they have one major shortcoming. The method is sensitive to the choice of the spectral region to be used in the fit. Thus selection of different start and stop wavenumbers leads to Raman spectra of significantly different lineshapes and intensities. This effect is the most prominent if any of these wavenumbers happens to be in a Raman active region particularly at the position of a Raman peak or on the leading or trailing slope of a Raman band. Moreover, both the algorithms, generically being polynomial based, still suffer, albeit to a less extent, from their dependence on the order of the polynomials used in the iterative fit.

In order to address these problems, we have developed a novel background subtraction algorithm, called Range Independent Algorithm (RIA), which avoids or further minimizes the shortcomings of the modified polynomial based methods while facilitating faithful representation of the extracted Raman spectrum for robustness and ease of use in a

clinical situation. The algorithm was found to provide good range independence and retrieval of all the Raman peaks by efficiently subtracting the associated intense background.

2.3.1 Framework of the Range Independent Algorithm (RIA)

The underlying basis of the RIA is iterative smoothing of the measured raw 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.

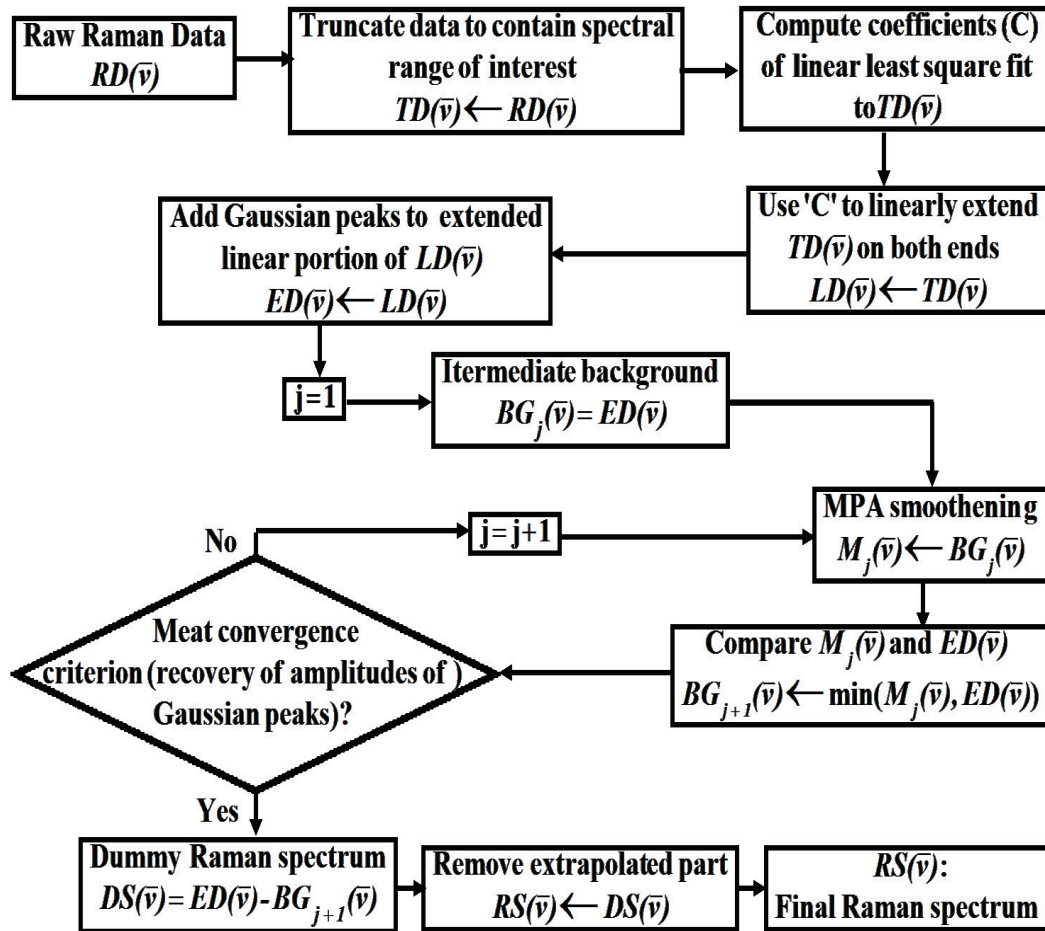


Fig. 2.2: Flow chart of the Range-Independent Algorithm (RIA) for background subtraction.

A detailed layout of the RIA is shown in Fig.2.2. As a preparatory step, the algorithm first truncates the measured raw Raman spectrum to contain the spectral range between the selected wavenumbers (i.e. the start and the stop wavenumbers) over which the Raman peaks are intended to be extracted. This truncated spectrum serves as the input to the subsequent computational steps comprising the RIA. The first step is to perform a linear least-square fit (using a polynomial of degree one) of the input spectral data over the selected range of wavenumbers to derive the coefficients of linear regression. The regression coefficients thus obtained are then used to extrapolate two sets of linear data (i.e. straight lines) to be appended at the two ends of the raw spectral data to extend the curve linearly beyond the selected portion of the spectral range. Next, two Gaussian peaks (of appropriate height and width) one on each side are added to the extended linear portion of the spectrum with the intention that these Gaussian peaks serve as the reference to be used for selecting the criterion of convergence for the subsequent iterative procedure. The peaks are positioned on the abscissa at a distance from the respective end points of the selected spectral range in such a way that these do not interfere with the ordinate values within the Raman spectral range. The details about the selection of the various parameters (like height, width, position of the Gaussian peaks etc. including others) related to the algorithm development are given in the “setting of algorithm parameter” section (2.3.3). Following the linear extension of the truncated raw spectrum outside the spectral region of interest and adding two Gaussian peaks to the extended linear part on the two sides, the whole of the extended spectrum (i.e. the raw Raman spectrum plus sidewise linear extensions with the Gaussian peaks) is subjected to modified iterative smoothing based on moving point averaging (MPA). The MPA is actually a zero-order Savitzky-Golay filter [96] and is equivalent to a low-pass filter which tends to filter out the high-frequency components of a signal leaving the low-frequency baseline intact [97]. It smoothes a set of data points by replacing each data point with the average of a pre-defined

number of neighboring data points called the span. The modified iterative MPA works based on the following principle: at the end of each round of iteration the ordinates of the smoothed data (referred to as intermediate background) are compared with those of the extended spectrum and the smaller of the ordinate values of the two over the entire range of abscissa are concatenated to generate a modified curve (i.e. the current intermediate background) to be MPA smoothed during the next round of iteration. The process of smoothing by MPA, subsequent comparison of the intermediate background (i.e. the smoothed data) and the extended spectrum (i.e. the raw data), and concatenation of smaller-of-the-two ordinate values is iterated until the amplitudes of the Gaussian peaks added to the raw data are recovered on subtracting the intermediate background (at that round of iteration) from the raw data. The concatenated curve constructed with the smaller-of-the-two ordinate values after the final round of iteration is truncated to include only the region between the start and stop wavenumbers selected at the outset. This truncated curve represents the background which is associated with the experimentally measured raw Raman spectrum and is then subtracted from the raw spectrum to obtain the true Raman spectrum with zero background.

2.3.2 Validation of the Algorithm on Mathematical Phantom Raman Spectrum

For the validation of the RIA a mathematical phantom Raman spectrum mimicking a typical raw tissue Raman spectrum was generated. The phantom Raman spectrum comprised a series of Lorentzian peaks on a null baseline (with a distribution similar to that seen in Raman spectra of human tissue):

$$y_R = \sum_{i=1}^N \frac{2A_{0i}}{\pi} \frac{\omega_{0i}}{4(r - r_{0i})^2 + \omega_{0i}^2} \dots\dots\dots (2.1)$$

where, r_{0i} is the position of the peak, ω_{0i} is the bandwidth of the peak at the full-width at half-maximum (FWHM), and A_{0i} is the total area under the curve above the baseline. The Raman signal was simulated using twelve peaks with intensity and positions similar to that seen in a tissue Raman spectra [94]. The details of the parameters used for construction of simulated data are given in Table 2.1.

Table 2.1: *The parameters of Lorentzian function used for constructing the synthetic Raman spectrum. Here, r_0 is the position of the Lorentzian peak, ω_0 is the full-width at half-maximum (FWHM) and A_0 is the total area under the curve above the baseline.*

r_0	ω_0	A_0
856	31	990
944	23	720
1002	6	240
1027	20	600
1071	30	960
1123	79	240
1259	24	1050
1265	26	1350
1302	16	1110
1341	14	570
1444	17	3150
1652	21	2820

The baseline background (y_F) was generated using a fifth-order polynomial and added to the simulated Raman data. The baseline background is given by:

$$y_F = 6.1716 \times 10^2 - 9.2437 \times 10^{-1} r + 1.169 \times 10^{-3} r^2 - 7.3381 \times 10^{-7} r^3 + 1.9267 \times 10^{-10} r^4 - 1.6268 \times 10^{-14} r^5 \quad \dots (2.2)$$

where, y_F represents the background intensity, and r represents the Raman shift in cm^{-1} . The coefficients were chosen such that the shape of baseline background was comparable to the baseline of a raw tissue Raman spectrum. The phantom Raman spectrum with the above parameters is shown in Fig.2.3.

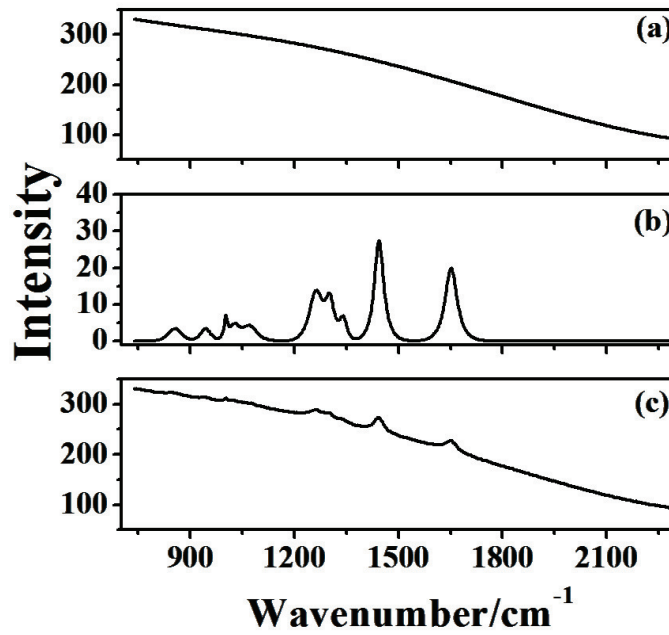


Fig. 2.3: Mathematically generated synthetic Raman spectrum, (a) the background, (b) the Raman signal simulated with twelve Lorentzian peaks on a null baseline, and (c) Raman spectrum mimicking a typical raw tissue Raman spectrum.

Figs. 2.4a-d provide a pictorial demonstration of the working of the range-independent algorithm applied on the mathematically generated raw Raman spectrum. Fig. 2.4a shows the synthetic Raman spectrum from which the true Raman bands (i.e. the Lorentzians added in there) are to be extracted. The spectrum was first truncated to include the spectral region of interest between 800 cm^{-1} and 1800 cm^{-1} and is shown in the same figure. The first step of the algorithm was to linearly extend the truncated spectrum beyond

the truncation points on both sides and add Gaussian peaks one on each side. Fig. 2.4b shows the truncated spectrum with linear extensions and added Gaussian peaks. The whole of the extended spectrum was then subjected to the modified iterative MPA smoothing procedure till the amplitude of the added Gaussian peaks were recovered on subtracting the intermediate background (at that round of iteration) from the raw data.

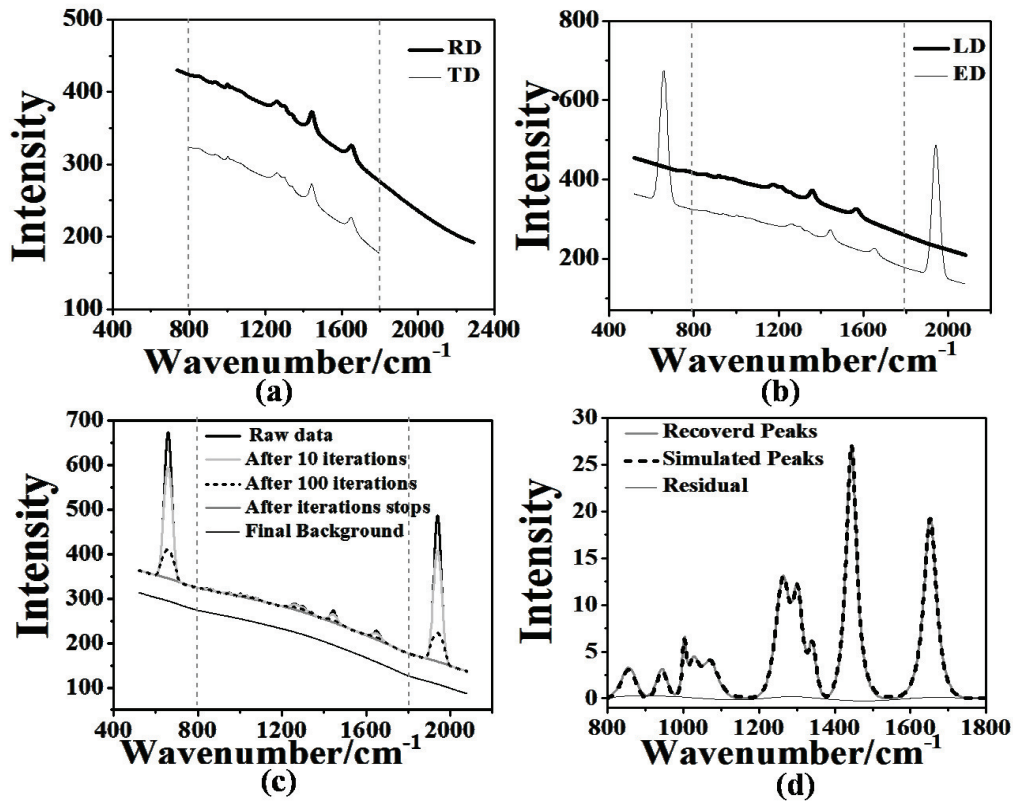


Fig. 2.4: Pictorial demonstration of the working of the RIA. (a) The raw Raman spectrum (bold line) and truncated spectrum in the range of interest (fine line with an offset), (b) truncated spectrum with linear extension (bold line) and added Gaussian peaks (fine line with an offset), (c) the background curves at the end of 10th, 100th and final round of iterations, and (d) the recovered Raman signal (gray line), synthetic Raman signal (black dashed line), and residual (black fine line). The final estimated background is also shown by fine line with an offset for clarity in panel (c).

Fig. 2.4c demonstrates the background curves at the end of ten, hundred and final round of iterations. The final background spectrum is also shown in the same figure. This background curve was then subtracted from the extended raw spectrum and re-truncated to generate the desired Raman signatures in the selected spectral range. The extracted Raman spectrum is shown in Fig. 2.4d. The true Raman spectrum along with the residual (true minus extracted) is also shown in the same figure. It is evident from the Fig. 2.4d that the RIA has retrieved all the Raman peaks with an excellent efficacy. The computed coefficient of correlation (R^2) was found out to be 0.99.

2.3.3 Setting of Algorithm Parameters

Successful use of an algorithm requires appropriate selection of a number of different parameters used for its design. The important first parameter for an iterative algorithm is to decide on a criterion of convergence most suitable for that algorithm. In the case of the RIA the iterative MPA smoothing was terminated only when the amplitudes of the two mathematically generated Gaussian peaks added on the two sides of the linear extensions of the truncated raw spectrum were recovered following subtraction of the intermediate background from the raw data. However, the construction of a Gaussian peak [98] requires two parameters to be supplied as input: its amplitude and standard deviation (σ) (or FWHM $\sim 2.345\sigma$). The selection of optimal values for these parameters is an optimization problem, where the possible values that the parameters can have is a finite set, and the cost function is defined by the application. For the RIA the cost function was chosen as the accuracy with which the Raman bands could be recovered from the mathematically generated raw Raman spectrum. For selecting optimal peak height, the RIA was run on the mathematical raw spectrum for three different peak heights: (i) normal, which was equal to the maximum of the ordinate values of the truncated raw spectrum over the selected range of wavenumber, (ii)

sub-normal, whose value ranged from three-fourth to one-hundredth of the normal height, and (ii) above-normal whose value ranges from one and half to hundred times the normal height. The optimal height was chosen to be the one that could recover the Raman bands with the maximum accuracy. It was found that while for peak heights less than the normal there was significant variability in the intensity and distribution of the extracted Raman bands throughout the spectral range, for heights equal to or taller than the normal there was no change in the extracted Raman spectra. This is shown in Fig. 2.5a where the extracted Raman spectra are displayed for three different peak heights, normal, sub-normal (one-hundredth of normal) and above-normal (ten times normal). The reason for the observed variability in the extracted Raman spectra for sub-normal peak heights are due to the fact in such situations the convergence criterion for the iterative MPA i.e. the recovery of the amplitudes of the added Gaussian peaks is met before the recovery of all the Raman peaks is completed. Whereas for normal and above-normal peak heights, the convergence criterion is attained on or after all the Raman peaks are recovered. Thus the height of the Gaussian peaks was chosen as equal to the maximum of the ordinate values of the truncated raw spectrum over the selected wavenumber range.

Similarly, the optimum FWHM of the Gaussian peaks was also based on empirical observations and was selected using an exhaustive search method. The RIA was applied on the mathematically generated raw Raman spectrum for the different FWHM values selected from a set of FWHM values ranging from 5 cm^{-1} to 100 cm^{-1} with increments of 5 for FWHM values between 5 to 50 and with increments of 10 for FWHM values between 50 to 100. It was found that while for FWHM values less than 20 cm^{-1} or below there was significant variability in the recovered Raman spectra throughout the spectral range, for FWHM values equal to or more than 20 cm^{-1} there was no change in the extracted Raman spectra. This is shown in Fig. 2.5b where the extracted Raman spectra are displayed for five different FWHM

values, 10 cm^{-1} , 15 cm^{-1} , 20 cm^{-1} , 50 cm^{-1} , and 100 cm^{-1} . It is clear from the figure that while the recovered spectra significantly vary for FHHMs below 20 cm^{-1} , there is no variation seen in the spectra above 20 cm^{-1} . The variability in the recovered Raman spectra at lower values of the FWHM of the Gaussians can be ascribed to the suboptimal estimation of the background by the RIA. The smaller the widths of the Gaussians, the quicker is its recovery and if the FHHM happens to be much smaller than that of the actual Raman peaks, the iterative smoothing operation will be terminated (as the convergence criterion will be satisfied) much before the Raman peaks are recovered fully. Similarly, at very large FWHM, the number of iterations will keep on continuing in order to satisfy the criterion of convergence even after the Raman peaks are recovered and this will force the RIA to keep on smoothing the data thereby pushing down the baseline even further.

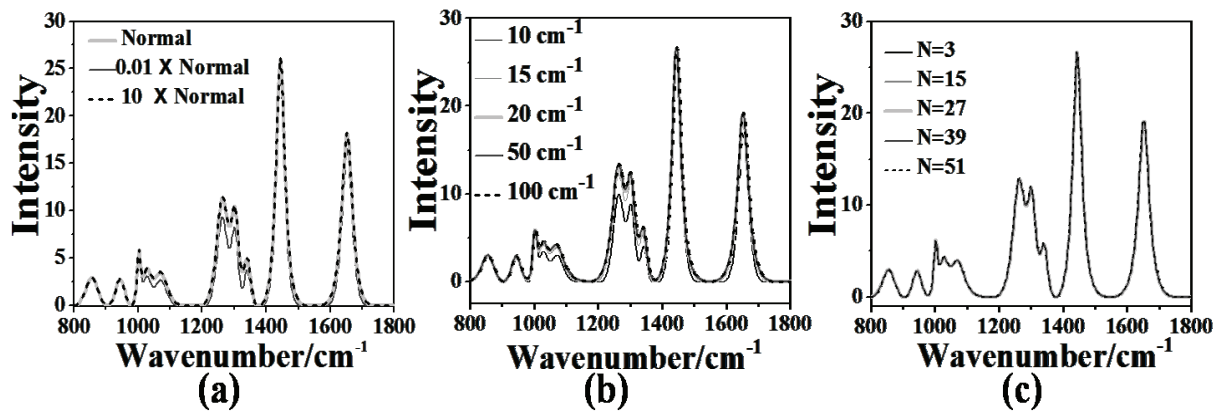


Fig. 2.5: The Raman signal recovered by the RIA from the synthetic raw Raman spectrum (a) for three different heights of the added Gaussian peaks: (i) normal (light gray line) equal to the maximum of the ordinate value of the raw spectrum, (ii) sub-normal equal to one hundredth of normal (dark gray line), and (iii) above-normal equal to ten times normal (black dashed line), (b) for five different full-width-at-half-maximum (FWHM) values of the added Gaussian peaks, and (c) for five different spans of the moving point average (MPA) filter, where N is the number of data points included in the span.

In addition to the height and FWHM of the Gaussian peaks, the position to where the peaks should be added on the extended linear portion of spectral data was also required to be decided for the RIA. This is important because of two factors: first, the RIA works on the principle of iterative MPA which replaces the intensity of a given point by the average of the nearby intensity points and second, the peaks are Gaussian which means that it extends up to infinity before its value comes to zero. Since, in practice the value of a Gaussian can be considered to be significantly small (~ 0.0001) at a distance of $\sim 4.3\sigma$ away from its peak, addition of peaks at least at a distance of 4.3σ away from the end points of the truncated spectrum is expected to have no influence on the outcome of the MPA filter thereby eliminating the chance of any distortion to be induced by the RIA in the recovered Raman signatures.

Central to the RIA is the modified iterative smoothing of a measured raw Raman spectrum by an MPA filter. The MPA filter attempts to estimate the average of the distribution of each intensity value of the spectrum and this estimation is based on a specified number of neighboring intensity values called the span. Thus prior to using the RIA it is required to specify the span (of the MPA) which determines the window of neighboring data points to be included in the smoothing calculation for each data point. This window moves across the spectral data set as the smoothed intensity value is calculated for each predictor intensity value. Since the objective of the RIA is to estimate the low frequency baseline by smoothing the high-frequency Raman peaks through successive iterations, the choice of the span or the number of the neighboring data points is not very critical in this case. This was found by applying the RIA on the mathematically generated raw Raman spectrum with different values of span ranging from 3 to 51 in steps of 4. Fig. 2.5c shows the recovered Raman spectrum for different values of the span of the MPA. While a smaller span was found to require an increased number of iterations to retrieve the Raman peaks, a larger span was

found to do the reverse. However, a very large span value (not shown in figure) was found to introduce changes in the line shape of the extracted Raman spectra. This is because in such a situation the Gaussian peaks start interfering with the MPA estimate of the smoothed intensity in the spectral range of interest leading to artifacts.

2.3.4 RIA verses Polynomial Fitting based Algorithms – Performance Comparison

The background subtracted Raman spectra are generally found to have distortions or artifacts particularly at the end points. This is primarily contributed by the algorithms used for background subtraction. The distortions vary from spectrum to spectrum and often turn out to be a nuisance in the subsequent quantitative analysis of the spectral data. The effect is more pronounced in the case where the start or the end wavenumber falls in the Raman active regions. One recommended way to minimize this distortion is to choose non-Raman regions for selecting the start and the end wavenumbers. However, this involves user intervention and is clearly a disadvantage for its automated use. The novelty of the RIA is that it does not lead to any distortions in the recovered Raman spectra irrespective of whether the start or the end wavenumber falls in the Raman active regions. Fig. 2.6a shows the mathematically generated Raman spectrum whereas Fig. 2.6b shows the recovered spectrum after background subtraction with the RIA for the three different spectral ranges: (i) range-1 corresponding to $800\text{--}1800\text{ cm}^{-1}$, (ii) range-2 corresponding to $980\text{--}1580\text{ cm}^{-1}$, and (iii) range-3 corresponding to $1150\text{--}1750\text{ cm}^{-1}$ respectively. For comparison sake, the Raman spectra recovered from the same raw mathematical spectrum by the ModPoly [69] and the I-ModPoly [94] algorithms for the same set of wavenumber ranges are shown in Figs. 2.6c and 2.6d respectively. It is clear from the figure that almost no distortion effect is induced by the RIA in any of the spectral ranges and hardly any artifacts are seen at the end points of the recovered spectra. In

the RIA, the end point distortions are avoided by use of linear extrapolation of the raw data beyond the selected end points on the abscissa and then carrying out iterative smoothing of the whole of the extended spectrum using modified MPA. Since in iterative modified MPA the original intensity of the end point of the truncated spectrum is replaced at each iteration by the average of the nearby intensity points equal in number from either side (the spectrum side as well as the linear extension side), the end point of the extracted Raman signal is forced to remain placed in its original position thereby minimizing the distortion effect. Instead of a linear extension, use of non-linear extensions employing non-linear least-square regression with polynomial function was also tried. However, non-linear extensions could not eliminate the distortions in a consistent manner and were seen to be dependent on the polynomial order and selection of start and stop wavenumbers. This is not unexpected because a non-linear curve tends to always follow the pattern of intensity variation of the curvature immediately following the truncated spectrum and therefore expected to change significantly in Raman-active as compared to non-Raman region.

It is also clear from Figs. 2.6(c-d) that when the modified polynomial based algorithms are used to recover Raman spectra for the different spectral ranges, it leads to significant changes in the line shape of the Raman bands and their intensities. In contrast, when the RIA is applied on the same raw spectrum over the three different spectral ranges, Raman spectra of almost identical spectral contours and intensities over the entire range of wavenumbers are obtained. Table 2.2 provides a summary of the performance of the RIA in comparison with the ModPoly and I-ModPoly algorithms over different spectral ranges. It is evident from the table that while for the RIA the computed value of R^2 does not change and remains equal to 0.99 irrespective of the spectral ranges over which the Raman peaks are recovered, it varies widely across the spectral ranges for the modified polynomial based

algorithms being poorer particularly when the start or the stop wavenumber lies in the Raman active regions.

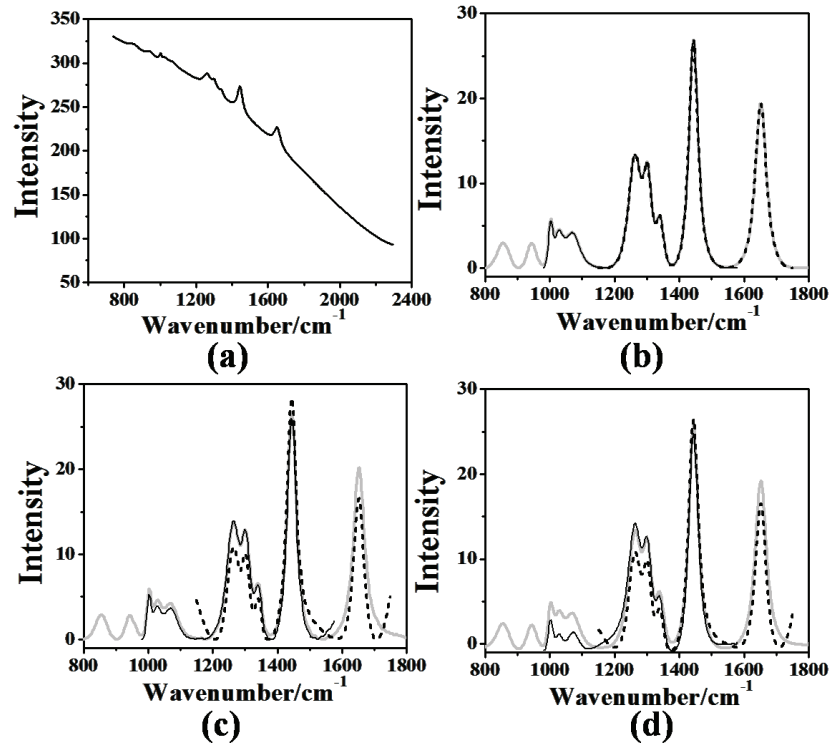


Fig. 2.6: (a) Raw synthetic Raman spectrum and the Raman signals recovered from it using (b) the RIA, (c) the ModPoly, and (d) the I-ModPoly for three different spectral ranges: (i) range-1 corresponding to $800\text{-}1800\text{ cm}^{-1}$, (gray bold line) (ii) range-2 corresponding to $980\text{-}1580\text{ cm}^{-1}$ (black fine line), and (iii) range-3 corresponding to $1150\text{-}1750\text{ cm}^{-1}$ (black dashed line).

The range-independence of RIA is due to two reasons: first, it takes care of the end points of a spectrum by tending to anchor it to their original position with the use of linear extrapolation combined with MPA. Second, since it is based on iterative smoothing which keeps on flattening the high frequency Raman bands riding a broad envelop to finally mimic the low-frequency contour of this envelop, the slope of the computed background curves remains the same in the common spectral region irrespective of the fitting range used for

background subtraction. In contrast, in the case of modified polynomial based methods, since they use a given order polynomial function for the fit, the slope of the fitting curve varies with the fitting range selected for the extraction of the Raman spectra. This leads to changes in the slope of the computed background in the common region thus causing variability in the line shapes of the recovered Raman spectra.

Table 2.2: *A comparison of the correlation of the mathematically simulated Raman peaks with the Raman peaks recovered by the RIA, the ModPoly and the I-ModPoly algorithms from the mathematically generated raw Raman spectrum over different spectral ranges.*

Spectral Range	R² value between simulated and extracted Raman signal		
	RIA	ModPoly	I-ModPoly
800-1800 cm ⁻¹	0.99	0.96	0.95
980-1580 cm ⁻¹	0.99	0.98	0.89
1150-1750 cm ⁻¹	0.99	0.88	0.81

2.3.5 Performance Evaluation on Raw Spectra with Different Signal to Background Ratios

A background subtraction algorithm is desired to accurately recover the Raman signal from the raw Raman spectrum irrespective of the signal-to-baseline (SBR) values. By SBR we mean the ratio of the intensity of the tallest Raman peak to its baseline height i.e. $SBR = (\text{signal maximum} - \text{signal minimum}) / (\text{baseline maximum} - \text{baseline minimum})$. However, in practice, it is often seen that the performance of the algorithm varies with SBR faring worse for its lower values. In order to analyze the performance of the RIA with respect to SBRs, it was applied on the mathematically generated raw Raman spectra of various SBRs

ranging from high to relatively much lower values. Figs. 2.7(a-c) show synthetic Raman spectra generated with three different SBRs: SBR=0.01 (Fig. 2.7a), SBR=0.1 (Fig. 2.7b) and SBR=1.0 (Fig. 2.7 c). The Raman spectra after processing by the RIA are shown in Figs. 2.7(d-f).

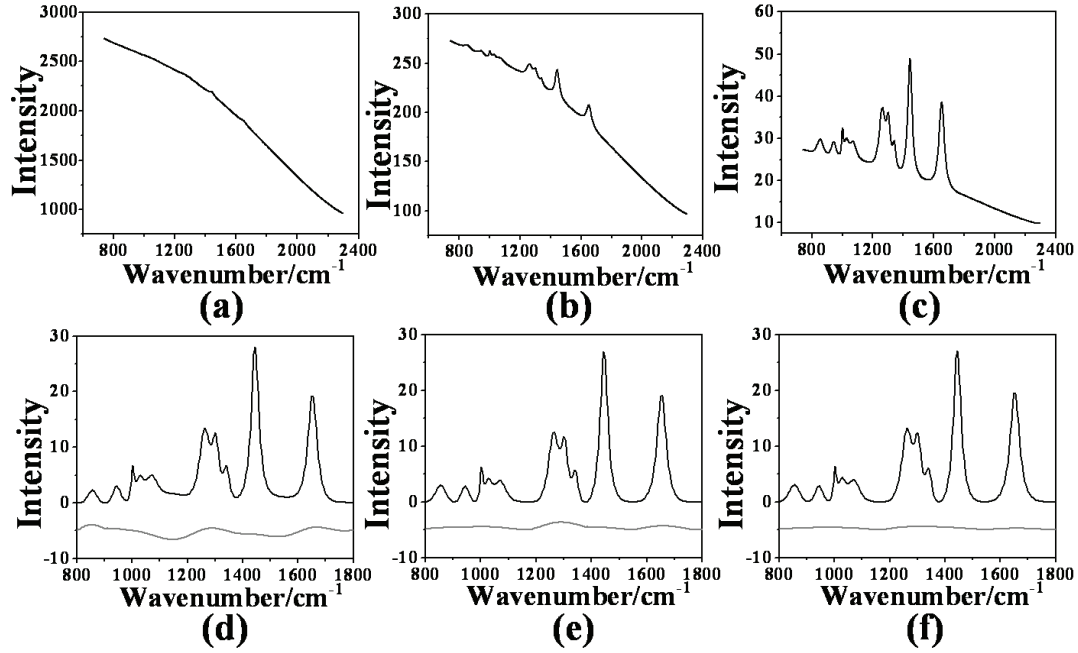


Fig. 2.7: Raw synthetic Raman spectra generated with signal to background ratio (SBR) of (a) 0.01, (b) 0.1, and (c) 1.0. The recovered Raman signal (black line) using the RIA and the residual (gray line) for the corresponding SBR values are shown in panels (d), (e), and (f) respectively. The residuals are shown with an offset for clarity in the corresponding panels.

It is apparent from the figures that in all the situations the Raman peaks have been faithfully retrieved without any distortions. However, when the performance of the RIA was further analyzed for even smaller values of SBR, it was found that Raman peaks could be faithfully retrieved without any distortions up to an SBR value of ≥ 0.005 . With SBR below ~ 0.005 , the tiny modulations in the line-shape of the background itself was found to lead to distortions and the emergence of false peaks in the recovered spectrum thereby limiting the performance of the algorithm. It is pertinent to mention here that the RIA is primarily

designed for use with tissue Raman spectra and although tissue Raman spectra typically display low SBR (as compared to a Raman active biochemical), the SBR values are found to be always $> \sim 0.05$ in the range of the integration times (~ 2 -5 sec) normally used in a clinical study. A tissue Raman spectrum with SBR below ~ 0.01 is not considered as a good quality spectrum.

2.3.6 Performance Evaluation on Raw Spectra with Different Background Profiles

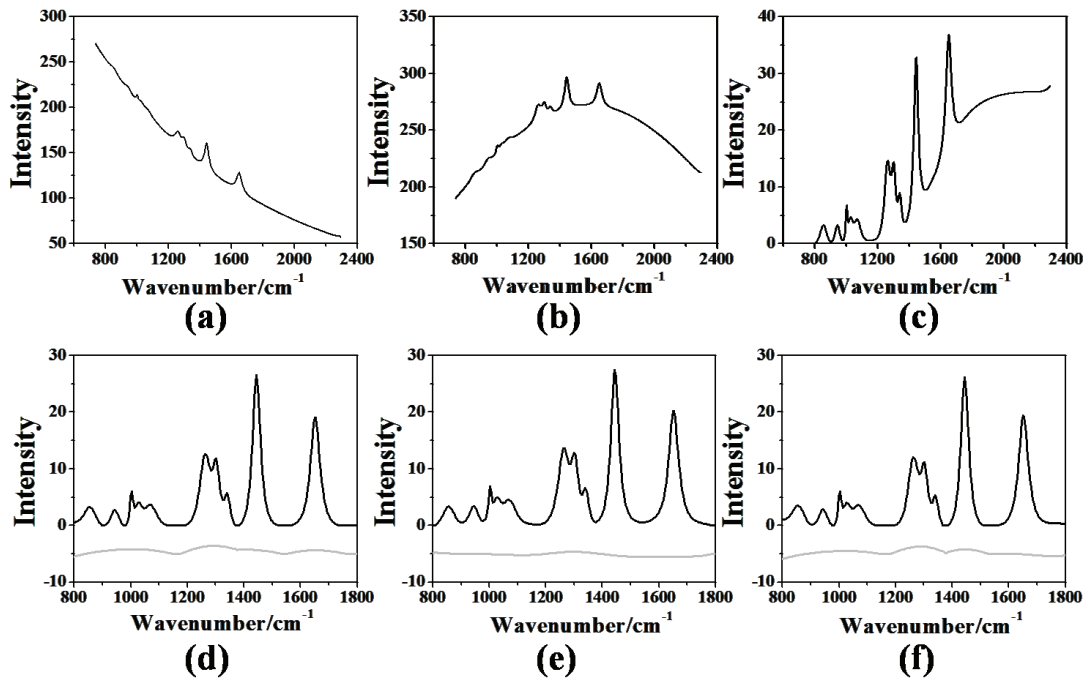


Fig. 2.8: Raw synthetic Raman spectra with (a) exponential, (b) Gaussian, and (c) sigmoidal backgrounds. The Raman signals recovered by the RIA (black line) from the respective raw spectra are shown in the panels (d), (e), and (f) respectively. The corresponding residuals (gray line) are shown with an offset.

In order to investigate the RIA's performance on a variety of dissimilar baselines (that might be obtained in practical situations), three more types of baselines were generated: (i) an

exponential baseline, (ii) a Gaussian baseline, and (iii) a sigmoidal baseline. These baselines along with the twelve Lorentzians simulated earlier were used to generate three raw Raman spectra which were further processed by the RIA. Fig. 2.8 shows the raw Raman spectra (a, b and c) and the recovered Raman spectra (d, e and f) following processing by the RIA. It can be seen from the figure that in all the three cases there is complete removal of the Raman peaks with almost no obvious distortions or artifacts. Thus the RIA seems to estimate the baseline almost independent of its curvature.

2.3.7 Validation of RIA on *In-Vivo* Raman Spectrum of Human Tissue

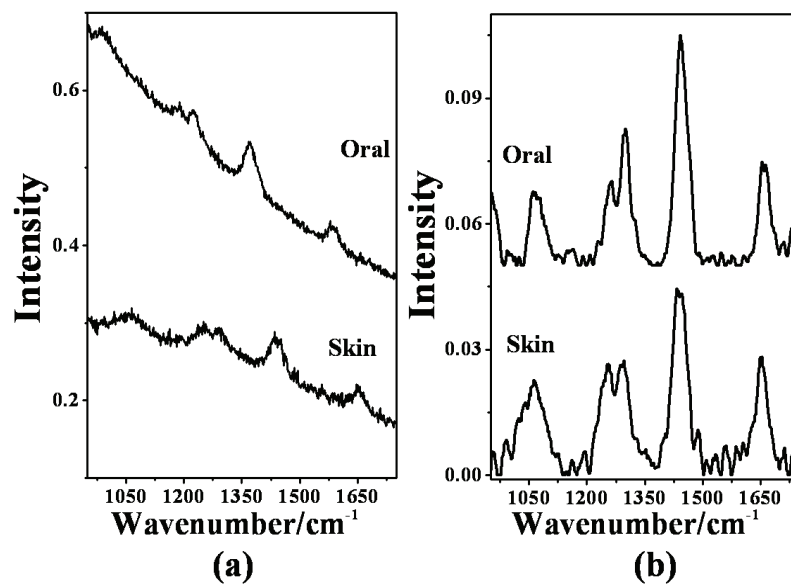


Fig. 2.9: (a) The experimentally measured raw in-vivo Raman spectra, and (b) the Raman spectra following background subtraction using the RIA for oral (upper row), and skin (lower row) tissues. For the sake of clarity the spectra are shown with an offset.

Tissue spectra typically display low signal-to-noise ratios and an intense background thereby always posing challenge to a background subtraction algorithm. Thus, in order to test the efficacy of the RIA in recovering characteristic tissue Raman peaks, it was used to process

raw Raman spectra measured *in-vivo* from human oral cavity and skin tissues. Fig. 2.9 shows the measured Raman spectra (Fig. 2.9a) of these *in-vivo* human tissues and the extracted Raman spectra (Fig. 2.9b) after processing with the RIA. One may see that the narrow Raman bands peaking around 1010 (phenylalanine: protein), 1060-1160 (nucleic acids, lipids, glycogen), 1240-1360 (amide III: proteins, glycogen), 1440 (proteins, lipids), and 1650-1680 cm^{-1} (amide I: proteins, lipids), characteristics of a typical tissue Raman spectrum [44], [52], [91], [99], [100] are preserved demonstrating the effectiveness of the RIA in subtracting background and retrieving Raman signatures.

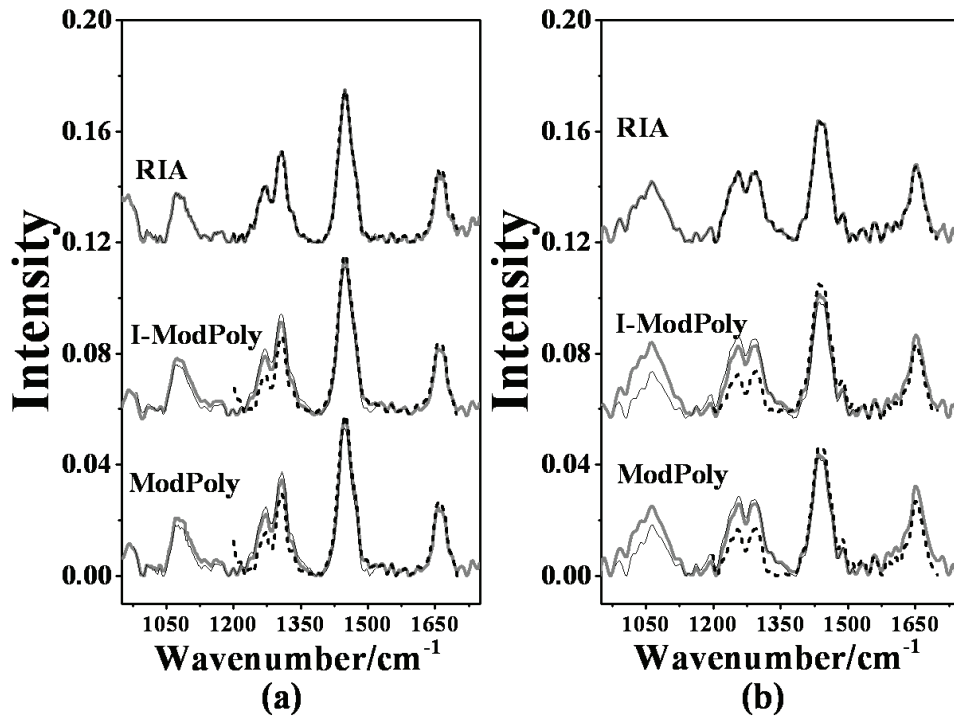


Fig. 2.10: The Raman spectra recovered from the experimentally measured raw Raman spectrum (shown in Fig.2.10a) of (a) oral tissue, and (b) skin tissue using the RIA, the ModPoly, and the I-ModPoly for three different spectral ranges: (i) range-1 corresponding to 950-1750 cm^{-1} , (gray bold line) (ii) range-2 corresponding to 980-1520 cm^{-1} (black fine line), and (iii) range-3 corresponding to 1200-1700 cm^{-1} (black dashed line).

Figs. 2.10a and b demonstrate the Raman spectra, recovered using the RIA, from the measured *in-vivo* raw Raman spectra (shown in Fig. 2.9a) of oral cavity and skin respectively for three different spectral ranges. For comparison sake, the corresponding Raman spectra extracted using the modified polynomial based methods are also shown in the same figures. It is apparent from the figures that all the methods can retrieve the Raman peaks normally expected to be present within the selected spectral range. However, the Raman spectra extracted using the modified polynomial based methods show significant changes in the intensities and line shapes of the Raman bands for different spectral ranges similar to what was seen earlier in the case of mathematically generated Raman spectrum. In contrast, the RIA does not show any range dependence and the extracted Raman spectra of all the samples show almost identical line shapes and intensities without any artifacts over all the selected ranges of wavenumbers.

For various applications in general and tissue diagnostic applications in particular, it is often required to background subtract a large number of Raman spectra measured experimentally. In such situations it is always desirable to have a background subtraction algorithm that can be completely automated without any user intervention. In this regard the RIA is suitable for full automation, since it does not need any parameters to be optimized during the course of its implementation. Another factor that may become important while considering a background subtraction algorithm for processing a large number of spectra is the computation time required to process a given spectrum. It is desirable to have the computation time as small as possible for successful use of the automated algorithm in practice. Since the major computational load of RIA (like the other iterative methods) lies in the iterative MPA operation of order $O(N)$ complexity, N being the number of data points to be smoothed, the computation time is primarily dependent on the number of iterations. The actual time taken by the RIA to complete background subtraction was computed by applying

it on a series of raw Raman spectra measured from the oral cavity and it was found that depending on the raw data at disposal, the RIA takes ~300-500 iterations for completing background subtraction requiring an average computation time of <500ms per spectrum. It is important to note here that *in-vivo* clinical Raman spectroscopy studies typically use an integration time of ~2-5 s to obtain good quality raw tissue Raman spectrum. Thus, an additional processing time of around a second per spectrum will not result in a significant change in data acquisition time (from a patient) and thus may not be a serious issue

2.4 Data Analysis

Subsequent to background subtraction of the measured raw spectra from tissues, the next major task is to extract and quantify the diagnostic information from these background subtracted spectra and classify them into different tissue pathologies. The different methods of data analysis employed for that purpose are described in this section.

2.4.1 Diagnostic Algorithm

Successful use of Raman spectroscopy for tissue diagnostic applications requires an appropriate diagnostic algorithm that can discriminate between various categories of tissues based on the recorded Raman spectra of tissues. A probability based multivariate statistical algorithm capable of simultaneously classifying spectral data into multiple (more than two) classes was employed for that purpose [101]. The algorithm [101] consists of two steps: i) extraction of diagnostically relevant spectral information through nonlinear maximum representation and discrimination feature (MRDF) [102] and ii) probabilistic classification via sparse multinomial logistic regression (SMLR) [103]. The brief discussion about the details of diagnostic algorithm [101] is presented here.

Maximum representation and discrimination feature (MRDF):

The non-linear MRDF [102] is a feature extraction procedure that aims to compute $K \ll D$ set of nonlinear transformation vectors, $\Phi_K = [\phi_1, \phi_2, \dots, \phi_K]^T$, from D -dimensional (where D is the number of wavenumbers over which spectra were recorded) spectra, such that the projections ‘ y ’ of the input data $x = [x_1, x_2, \dots, x_D]^T$ (intensities corresponding to wavenumbers of the spectra) on Φ_K from the different tissue classes are statistically well separated from each other [101].

$$y = \Phi_K^T x = [y_1, y_2, \dots, y_K]^T \quad \dots\dots\dots (2.3)$$

For high dimensional data, i.e. the dimensionality ‘ D ’ is larger than the size ‘ S ’ of the data, generally a two stage MRDF with restricted polynomial transform at each stage is used [101], [102]. In the first stage, the input data from each tissue type are raised to the power p' i.e. $x_{p'} = [x_1^{p'}, x_2^{p'}, \dots, x_D^{p'}]^T$ and subjected to a transform Φ'_M to produce the first stage output features $y'_M = \Phi'^T_M x_{p'}$ in the feature space of reduced dimension $M \leq S \ll D$. Here, S is size of the tissue type which has minimum number of spectra. The transform Φ'_M is chosen so as to maximize a discrimination measure:

$$E'_d = \sum_{m=1}^M \phi_m^T (R_{p'}) \phi_m \quad \dots\dots\dots (2.4)$$

$$\text{with, } R_{p'} = \sum_{i=1}^{L-1} \sum_{j=i+1}^L R_{ijp'} = \sum_{i=1}^{L-1} \sum_{j=i+1}^L \frac{1}{S_i S_j} (x_{ip'} - x_{jp'}) (x_{ip'} - x_{jp'})^T \quad \dots\dots\dots (2.5)$$

$R_{p'}$ is basically a vector outer-product difference matrix and is calculated from higher order correlation ($x_{p'}$) of input data x . L is the number of classes present in a given classification task, and S_i and S_j are the number of spectra corresponding to the i^{th} and j^{th}

class, respectively. The dominant eigenvectors of R_p are the solutions of Φ'_M and are used to get the first stage output features $y'_M = \Phi'^T_M x_p$.

In the second stage, the reduced M dimensional output features y'_M for each tissue type are further raised to the power p i.e. $y'_{Mp} = [y'^p_1, y'^p_2, \dots, y'^p_M]$ and subject to second transform to yield the final output features $y_k = \Phi'_M y'_{Mp}$ in the nonlinear feature space of dimension K ($K \leq M$). The transform Φ_K is chosen so as to maximize a new discrimination measure:

$$E_d = \sum_{k=1}^K \frac{\phi_k^T (R_p) \phi_k}{\phi_k^T (C_p) \phi_k} \dots\dots\dots (2.6)$$

$$\text{where, } R_p = \sum_{i=1}^{L-1} \sum_{j=i+1}^L R_{ijp} = \sum_{i=1}^{L-1} \sum_{j=i+1}^L \frac{1}{S_i S_j} (y'_{iMp} - y'_{jMp})(y'_{iMp} - y'_{jMp})^T \dots\dots\dots (2.7)$$

$$\text{and, } C_p = \sum_{i=1}^L C_{ip} = (\sum_{i=1}^L E(y'_{iMp} y'^T_{iMp}) - \mu_i \mu_i^T) \dots\dots\dots (2.8)$$

where, C_p is the summation of intra-class covariance matrices and μ_i being the mean of class 'i' in the reduced M -dimensional feature space. The final transformations ' $\Phi_K = [\phi_1, \phi_2, \dots, \phi_K]^T$ ', are the ' K ' dominant eigenvectors of $(\sum_{i=1}^L C_{ip})^{-1} (\sum_{i=1}^{L-1} \sum_{j=i+1}^L R_{ijp})$. Since the nonlinearities introduced in the two stages were different (p' in 1st stage and p in 2nd stage), this is expected to produce more general nonlinear transforms on the input spectral data leading to improved separation of the final nonlinear features for the tissue types in the new feature space [101].

Sparse Multinomial Logistic Regression (SMLR):

The SMLR [103] based classifier separates a set of labeled input data into its constituent classes by predicting the posterior probabilities of their class membership [101]. The training of the classifier involves estimation of a total L (number of classes) set of sparsity promoting model weights $w = [w^{(1)^T}, \dots, w^{(L)^T}]^T$ based on maximum a posteriori (MAP) framework using the training data and their true class labels. If many of the training set data points are irrelevant in making the decision boundaries for the class prediction, the vector w is said to be sparse and many of its entries are exactly zero [101], [103]. To promote sparsity in w , the SMLR method chooses a Laplacian prior on w which means that $p(w) \propto \exp(-\lambda \|w\|_1)$, where $p(w)$ is a prior distribution on w , $\|w\|_1 = \sum_i |w_i|$ denotes the l_1 -norm and λ is a tunable regularization parameter needed to be optimized during the training phase of the algorithm [103]. The details of the SMLR method for estimation of w are described elsewhere [101], [103]. In the context of the classification of spectral data used in this thesis work, the goal of SMLR is to predict the posterior probabilities of a spectrum belonging to each of the L classes given L sets of feature weights, one for each class [101]. If $y = [y^{(1)}, y^{(2)}, \dots, y^{(L)}]^T$ is a 1-of- L class encoding and if $w^{(i)}$ is the model weight reflecting the importance of the training set data point associated with class i , then the probability that a given spectrum x belongs to class i is given by [101]:

$$P(y^{(i)} | x, w) = \frac{\exp(w^{(i)^T} x)}{\sum_{i=1}^L \exp(w^{(i)^T} x)} \quad \dots\dots\dots 2.9$$

Subsequently, the class is assigned to a given tissue spectrum whose posterior probability of belonging to that particular class is the highest according to the Bayesian rule of classification [101], [103].

Implementation of the Diagnostic Algorithm

The flow chart for the implementation of the probability based multivariate diagnostic algorithm is shown in Fig. 2.11. For a given classification task, the input spectral data were normalized using to the scheme described elsewhere [101], [102] and then subjected to feature extraction by the MRDF prior to classification by the SMLR. The MRDF reduces dimensionality of the high-dimensional spectral data and results in a set of few non-linear output features that contained the maximum class discriminatory information. The optimal number of these features and the optimum p' and p are decided by employing a cross-validation procedure and are identified as the values that yields the least misclassification error using nearest mean classifier [102]. The optimal number of features usually varies from 15 to 20 depending on the classification task. This optimal number of output nonlinear features was used as input to the SMLR for subsequent classification. The SMLR classifier is trained with different values of the regularization parameter λ (varying between 1 and $1e-5$) and the optimum value of λ is chosen to be the one that gives the highest classification accuracy.

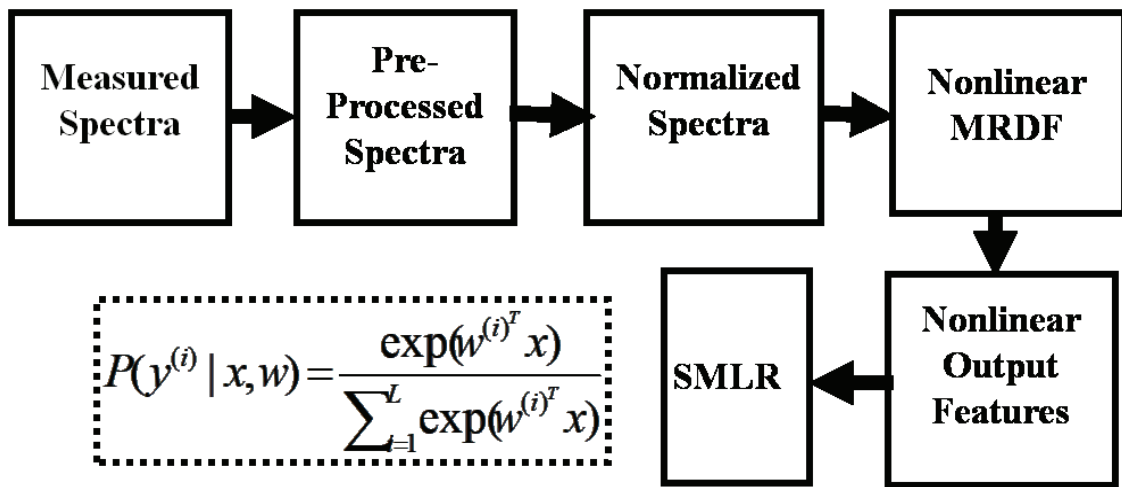


Fig. 2.11: Flow chart for the implementation of the probability based multivariate diagnostic algorithm.

In order to have a statistically unbiased estimate of the generalized classification ability of the MRDF-SMLR algorithm (i.e. how correctly the algorithm can classify previously unseen data), it is necessary to ensure that the validation set data should neither contain any of the pieces of the training set data with which the algorithm was trained and optimized, nor any part of the data of a subject whose remaining part remained in the training set. The generalized classification ability of the algorithm is evaluated, for a given classification task, by adopting the method of leave-one-subject-out cross validation. In this method, for N total number subjects (patients and healthy volunteer) for a given classification task, spectra of $N-1$ of them are used to train the algorithm and the set of spectra of the remaining subject is left excluded (and therefore not used by the algorithm during training) as the validation set. This is repeated N times (until the spectra of all the withheld subjects were classified) each time excluding a different subject for the purpose of validation and retraining the algorithm using spectra of the rest of the subjects. Since the training set data remains completely independent of the test data in each of the N loops (as the set of spectra from a subject is never a part of both the training and the validation sets simultaneously), the validation is statistically unbiased. Further, since the spectral data of each patient in the validation set has proper representation in the training set which contains the full set of spectral data of patients and healthy volunteers, the information necessary for classifying the spectra of the test subject is appropriately learnt by the algorithm (from the training set data) and thus the prediction made by it was balanced and unbiased.

For a given classification task, MRDF-SMLR algorithm computes the posterior probabilities of the different interrogated tissue sites of belonging to various tissue classes. Subsequently, the class is assigned to a given tissue spectrum whose posterior probability of belonging to that particular class is the highest according to the Bayesian rule of classification [101], [103]. The predictive accuracies of the diagnostic algorithm used for different

classification tasks are calculated with respect to histopathology as the gold standard of reference.

It is important to mention here that the nonlinear MRDF-SMLR based diagnostic algorithm has the ability to both compress the large amount of data obtained with each Raman spectrum as well as retain only the diagnostically relevant portions of the spectra in this compression. In contrast, the most widely used algorithm for tissue diagnosis employs principal components analysis (PCA) [104] that creates only a single model for a dataset and compresses the data in decreasing degrees of shared variance. Since the diagnostically relevant features in the Raman spectra are very small in comparison with the shared spectral content, it is imperative that these features be retained. Further, the conventional diagnostic algorithms based on PCA or linear discriminant analysis (LDA) being linear in nature cannot perform well on non-symmetric data that may have multiple clusters per class and also suffer from the limitation of extracting information from only the second order correlation in the data [101], [105], [106]. In contrast, the nonlinear MRDF-SMLR uses higher order correlations and therefore can provide improved discrimination [101] because of its built-in capability to separate classes which are not linearly separable in the original input data space.

In a clinical scenario, depending on the requirement, the emphasis may be on either screening or screening and diagnosis both. While the task of screening requires delineation of abnormalities without going for further evaluation, the diagnosis seeks to know the pathological status of the detected abnormality. The capability of the algorithm to operate in both binary as well as multi-class classification platforms certainly adds to the utility of the technique in a clinical situation where it can be opted for either screening and/or diagnosis. Another important feature 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. In contrast, the PCA-LDA algorithm generally results in a numeric-valued output score to

represent the degree to which a given observation belongs to a particular class. There are many advantages of having probability estimates. First, the availability of probability estimates allows one to follow a systematic approach for selecting the appropriate threshold for classification. Therefore, instead of requiring to choose the threshold in an ad-hoc manner as in the PCA-LDA algorithm, the underlying Bayesian risk model in SMLR takes into account the relative cost of misclassifying normal tissue sites into abnormal and vice-versa.

2.4.2 Standard Error Confidence Interval

Prior to applying the MRDF-SMLR diagnostic algorithm, which mathematically transforms the spectra into a new feature-space thereby making it impossible to relate the diagnostically relevant features with the original wavenumber space, an initial analysis needs to be performed to qualitatively determine the statistically significant differences between the tissue spectra. Though these results are not used in the diagnostic algorithm, they allow exploration of the responsible mechanisms for differential diagnosis. To identify the region of the spectral differences between the pathologic and normal tissue spectra a statistical analysis is performed based on standard error (SE) confidence intervals [99]. The variance of the intensity at each wavenumber is first calculated for tissue spectra belonging to each pathology class. The composite variance (σ^2) of the spectra at each wavenumber is calculated as:

$$\sigma_{lesions}^2(\lambda) = \frac{\sum_i \sigma(\lambda)_i^2 (df)_i}{\sum_i (df)_i} \dots\dots\dots (2.10)$$

where, σ^2 is the variance of the intensity at each wavenumber λ for each lesion type ‘ i ’ and df corresponds to the degrees of freedom for each pathology (= number of tissue specimens-1). The SE of the pooled lesion spectra and normal oral tissue spectra is then calculated at each wavenumber as:

$$SE(\lambda) = \sqrt{\frac{\sigma_{\text{normal}}^2(\lambda)}{n_{\text{normal}}} + \frac{\sigma_{\text{lesions}}^2(\lambda)}{n_{\text{lesions}}}} \dots\dots\dots (2.11)$$

where, n is the number of tissue spectra included in the particular tissue type. The SE is then multiplied by the appropriate t-values based on the total degrees of freedom and a predefined confidence level to produce a confidence interval. Difference spectra for the abnormal with respect to the normal are overlaid on these confidence intervals to qualitatively identify statistically significant spectral differences.

2.4.3 Pillai's V Measure

In order to quantify the amount of separation in the Raman spectra between the different pathologic categories of tissue, a method called Pillai's V [107], is normally calculated. In brief, Pillai's V is a statistical measure, often used in multivariate analysis of variance (MANOVA), 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:

$$V = \text{trace}(BT^{-1}) = \sum_{i=1}^h \frac{\lambda_i}{\lambda_i + 1} \dots\dots\dots (2.12)$$

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 $h = L-1$, L being the number of classes. A high Pillai's V means a high amount of separation between the samples of classes, with the between-group variance being relatively large compared to the total variance.

2.4.4 Multiclass Receiver Operating Characteristic Analysis

For quantifying the algorithms' performance for the different classification tasks, a multiclass receiver-operating characteristic (ROC) analysis [108] is carried out on the classification

results of the diagnostic algorithm using the Hand and Till measure (HTM). HTM is based on multi-class Receiver Operating Characteristic (ROC) approximation [108] and extends the Area Under the Curve (AUC), one of the most popular measures for binary classifiers to multiclass tasks. Given ‘ L ’ 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 $L(L-1)/2$ pairs of classes and given by Hand and Till measure (HTM) [108] as:

$$\text{HTM} = \frac{2}{L(L-1)} \sum_{i < j} \text{AUC}(i, j) \quad \dots\dots\dots (2.13)$$

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. As is the case for two-class, the closer the HTM equals to 1, the more accurate the corresponding diagnostic algorithm is.

2.5 Summary

To summarize, in this chapter, we have presented the development of a portable, NIR Raman system amenable for use in a clinical setting. The development of a novel range-independent background subtraction algorithm capable of rapid and automated retrieval of Raman signatures from the measured raw Raman spectra is detailed next. The algorithm fulfills all the requirements of an faithful background subtraction method that is desired for successful use of Raman spectroscopy for real-time, noninvasive, automated diagnosis of various cancers in a clinical situation. Finally, a brief description of the probability based multiclass diagnostic algorithm which can classify tissue into different pathological classes based on their measured spectral signatures along with various methods of analyses of spectral data are presented. In the next chapter we will describe the results of a clinical *in-vivo* study carried out at Tata Memorial Hospital, Mumbai where the portable Raman system was used for

measurement of *in-vivo* Raman spectra and the diagnostic algorithm along with the statistical methods of data analysis were used for discriminating healthy volunteers from patients with various lesions of oral cavity.

Chapter 3

In-Vivo Raman Spectroscopy for Detection of Oral Neoplasia

3.1 Introduction

Recent studies have demonstrated the applicability of Raman spectroscopy as a promising alternate approach for real-time and non-invasive diagnosis of cancer of various organs [41], [44], [52], [99], [109]–[111]. As far as oral lesions are concerned, the existing reports demonstrating the potential use of the Raman spectroscopy are limited to either *ex-vivo* studies [59]–[65] or *in-vivo* studies targeting only a part or so of the oral cavity (e.g. buccal mucosa or tongue) [45], [49], [65], [112]. There is no published report, thus far, of a full-scale clinical study 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. This is important because availability of a spectral database of the whole oral cavity may facilitate rapid screening of a large population (e.g. in a community setting) for any abnormality in the oral cavity with the help of an appropriately trained diagnostic algorithm. Another limitation of the earlier studies [45], [112] is that the patient population included therein was not a true representative of the tissue types interrogated, since it involved only patients already identified of having malignancy of the buccal mucosa.

We present, in this chapter, the results of a clinical study carried out to evaluate of the efficacy of *in-vivo* Raman spectroscopy for differential detection of oral lesions in the whole of the oral cavity.

3.2 Clinical Protocol for the Studies

The *in-vivo* Raman spectroscopic studies were 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 for the *in-vivo* study. The patients included in the study had no history of malignancy or dysplasia, and were suspected by the examining physician of having either cancer or leukoplakia or submucosal fibrosis on visual examination of the oral cavity. Patients having gone through any prior treatment like surgery, chemotherapy or radio therapy for earlier cancers or with recurrences were excluded from the study.

All the spectral measurements were performed by the participating head and neck surgeon using a protocol which was maintained for all individuals in this study. Briefly, 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 no subject 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.

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. Similarly, no biopsies were allowed from the tissue sites of healthy volunteers. 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. Histopathology was taken as the “gold standard”. All the Raman spectra from across the whole of the oral cavity were categorized in accordance to their histological identities and grouped into oral squamous cell carcinoma (OSCC), oral submucous fibrosis (OSMF), leukoplakia (OLK) or normal oral tissue.

3.3 Study Design and Spectral Measurements

The study involved 28 healthy volunteers with no history of the disease of the oral cavity and 171 patients enrolled for medical examination of the oral cavity at TMH. Informed consent was obtained from each patient as well as the healthy 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 $\sim 6:1$. As far as tobacco habits are concerned, 98% of the patients and 71% of the normal volunteers had habits of either smoking or chewing tobacco, and the remaining candidates had no history of tobacco consumption.

The good quality (signal-to-noise ratio ≥ 10) *in-vivo* Raman spectra were recorded using developed portable Raman spectroscopic system (Fig. 2.1) from a total of 515 tissue sites of 171 patients with an integration time of 5s. Out of these, 94 sites were identified as

OSMF by the examining doctor and from these no biopsies were taken. Of the remaining tissue sites, 316 were histopathologically characterized as OSCC and 105 as OLK. Spectra were also recorded from 287 sites from healthy squamous tissue of 28 normal volunteers. The details of the histopathological distribution of the tissue sites included in the study are summarized in the Table 3.1. Each site was treated separately and classified via the diagnostic algorithm developed.

Table 3.1: *Histopathological distribution of tissue sites included in the clinical pilot study.*

No. of Individuals	No. of Sites	Category
113	316	Oral Squamous Cell Carcinoma (OSCC)
25	94	Oral Submucous Fibrosis (OSMF)
33	105	Oral Leukoplakia (OLK)
28	287	Normal

3.4 Data Pre-processing and Analysis

Prior to Raman spectral measurements from a subject, the wavenumber axis was calibrated with the excitation laser line, acetaminophen, and naphthalene standards. For each measured Raman spectrum, the signal from the CCD was binned along the vertical axis to create a single spectrum per measurement. Prior to any signal processing, the spectrum was truncated to only include the region from about 900 cm^{-1} to 1750 cm^{-1} . A sequence of pre-processing steps was then executed on this binned, truncated spectrum following the procedure described by Motz *et al.* [37]. 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. The next step was to remove the artifacts introduced in the

measured tissue spectrum as a result of laser-induced artifacts generated in the fiber-optic probe. This was done by recording the spectrum of 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 spectrum that results in 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 noise smoothed using a second-order Savitzky–Golay filter and then background subtracted using the range-independent background subtraction algorithm (described in the Section-2.3) to retrieve the weak tissue Raman spectrum. Each background-subtracted tissue Raman spectrum was normalized with respect to its mean spectral intensity across all the Raman bands. The normalized Raman spectra were used for subsequent data analysis.

The probability based direct multiclass classification algorithm developed earlier [101] was employed to analyze the diagnostic content of the spectra measured sequentially from the same set of oral tissue sites with fluorescence and Raman techniques. The algorithm consists of two steps: i) feature extraction (or dimensionality reduction) through nonlinear Maximum Representation and Discrimination Feature (MRDF) [102], and ii) probabilistic classification via Sparse Multinomial Logistic Regression (SMLR) [103]. All analyses were performed using leave-one-individual-out cross validation as described in the Section-2.4.1.

To quantitatively compare the relative performance of the diagnostic algorithms, a multiclass receiver-operating characteristic (ROC) analysis was carried out on the classification results yielded by the corresponding algorithms. The formulation extends the two-class ROC analysis for the multiclass case and computes a generalized metric (HTM) indicative of overall performance measure of a given multiclass diagnostic algorithm and detailed in the Section-2.4.4.

3.5 Results

Fig. 3.1a shows the average normalized Raman spectra for OSCC (n=316), OLK (n=105), OSMF (n=94), and normal squamous tissue sites (n=287) of the oral cavity, with the error bars representing the spectral standard deviations. From the figures it is evident that the variation in the measured spectral intensity is comparable for all the tissue types investigated. The percentage variation (σ/\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 four histopathological categories 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.

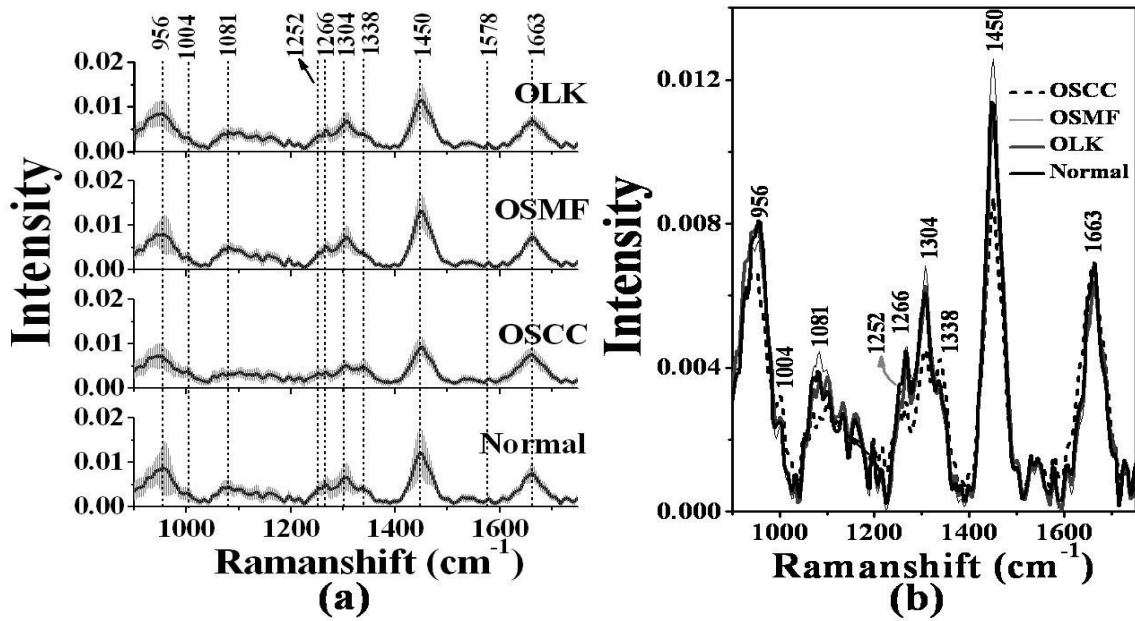


Fig. 3.1: Mean, normalized Raman spectra of OSCC (n=316), OSMF (n=94), OLK (n=105), and normal (n=287) oral tissue sites. The error bars (gray) represent ± 1 standard deviation.

For comparison of spectral differences among the different tissue types, the average Raman spectra are plotted without error bars in Fig. 3.1b. The spectra are similar to those reported in Raman spectroscopic studies of human oral cavity [59]–[67] by others, with major

peaks located at ~ 956 , 1004 , 1081 , 1252 , 1266 , 1304 , 1338 , 1450 , and 1663 cm^{-1} corresponding to various Raman-active biomolecules, notably collagen, elastin, keratin, lipids, and minerals known to be present in oral tissues [45], [47], [49], [59-67], [99], [100], [113]. Although the peak shapes and locations are consistent across all the pathology classes, subtle but significant differences are observed in peak intensities between the different pathology categories indicating biochemical differences inherent in the tissue types of different pathologies.

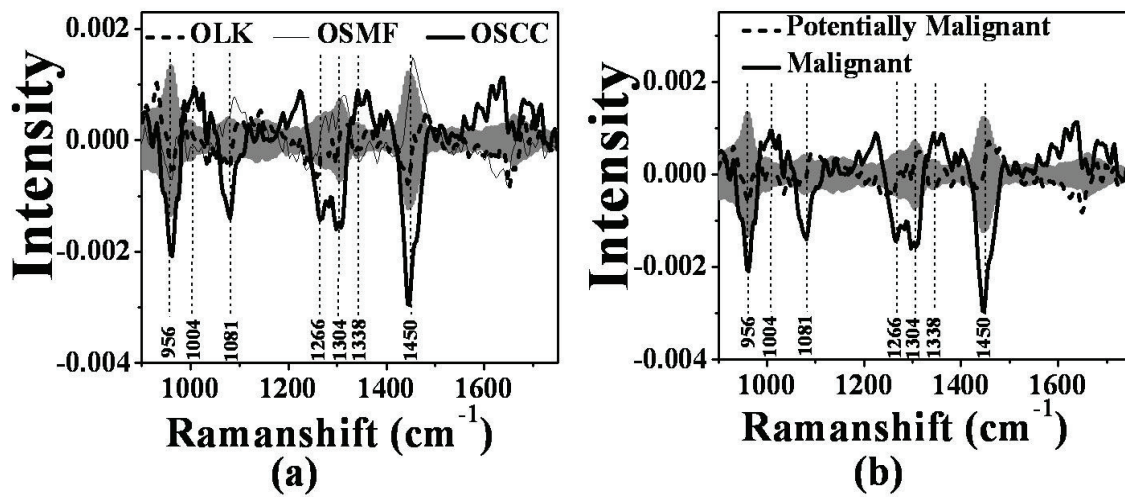


Fig. 3.2: Mean difference spectra showing statistical differences between different pathologies and normal oral tissue spectra. Gray bands indicate the 90% confidence intervals of the difference determined by standard error confidence intervals.

Fig. 3.2 shows the mean difference spectra obtained by subtracting the mean tissue spectrum of each lesion from the mean spectrum of normal oral tissue. The grey bands show the confidence intervals calculated by multiplying SE with a t-value corresponding to 90% confidence intervals and the degrees of freedom equal to number of spectral measurements (corresponding to the pathology minus the number of pathologic categories). A number of significantly different Raman bands are observed for each lesion with respect to normal. The

portion of the difference spectra outside the confidence interval represents the region of statistically significant spectral differences ($p < 0.1$) [99].

Table 3.2 shows the four-class (normal, OSCC, OSMF and OLK) classification results in the form of a confusion matrix displaying comparisons of the pathological diagnosis with that of the MRDF-SMLR based spectroscopic diagnostic algorithm. The classification results were obtained based on leave-one-subject-out cross validation of the MRDF-SMLR algorithm on the entire data set. One can see that the algorithm provided an overall classification accuracy of 86% (690 out of 802). It proved most adept at classifying OSCC tissues with a classification accuracy of 89%, though it fared worse in classifying other tissue types, and errors were spread among the various classes. Normal tissue spectra were correctly classified in 85% of the sites, while OLK and OSMF spectra were classified correctly in 85% and 82% of the sites.

Table 3.2: Confusion matrix displaying classification of the Raman spectra of oral tissue sites into four classes: normal, OSCC, OSMF, and OLK using the MRDF-SMLR based diagnostic algorithm. ‘n’ represents size of the spectral data in the corresponding tissue category.

Pathology Diagnosis	Raman Diagnosis			
	Normal	OSCC	OSMF	OLK
Normal (n=287)	85%	8%	4%	3%
OSCC (n=316)	8%	89%	0%	3%
OSMF (n=94)	14%	0%	85%	1%
OLK (n=105)	13%	5%	0%	82%

Table 3.3 shows the confusion matrix depicting the results of classification with the MRDF-SMLR algorithm validated, in leave-one-subject-out cross-validation fashion, on the whole set of spectra separated into three categories (instead of four), normal, malignant

(OSCC) and potentially malignant where the OSMF and OLK spectra were put together to form the last category. Although the overall discrimination accuracy is seen to be reduced marginally by 3% (from 86 % to 83%), the accuracy with which the OSMF and OLK spectra together (belonging to the potentially malignant category) can be discriminated is found to improve to 88% with 176 out of 199 spectra of this category being classified correctly. However, the algorithm fared worse in classifying the normal squamous tissue spectra where the classification accuracy is found to be only 77%.

Table 3.3: *Confusion matrix displaying results of classification of the Raman spectra of oral tissue sites into three classes: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC) using the MRDF-SMLR based diagnostic algorithm.*

Pathology Diagnosis	Raman Diagnosis		
	Normal	Potentially Malignant (OSMF+OLK)	Malignant (OSCC)
Normal (n=287)	77%	13%	10%
Potentially Malignant (n=199)	9 %	88%	3%
Malignant (n=316)	11%	5 %	84%

Table 3.4 shows the sensitivities and specificities yielded by the algorithm in delineating spectra of the healthy volunteers (i.e. the normal category) from three separate categories: malignant, potentially malignant and a category termed “abnormal” which includes spectra of tissue sites belonging to OSCC, OSMF and OLK all pooled together. One may note that the algorithm could discriminate the spectra of normal from that of all three categories with mean sensitivity and specificity of ~96% and ~97% respectively. Further it is also noted that the performance of the algorithm is relatively superior for binary classification as compared to the 3-class and 4-class classification.

Table 3.4: *The classification results of the Raman spectra of oral tissue sites into two classes using the MRDF-SMLR based diagnostic algorithm. Here the spectra belonging to OSCC are referred to as “Malignant”, those belonging to OSMF and OLK pooled together are referred to as “Potentially Malignant”, and the spectra belonging to OSCC, OSMF and OLK pooled together are referred to as “Abnormal”.*

Pathology Diagnosis	Raman Diagnosis	
	Sensitivity	Specificity
Normal (n=287) vs. Malignant (n=316)	96%	99%
Normal (n=287) vs. Potentially Malignant (n=199)	99%	98%
Normal (n=287) vs. Abnormal (n=515)	94%	94%

In addition to assigning class labels, the diagnostic algorithm also yielded posterior probabilities of the measured tissue sites belonging to each class of oral tissue. Figs. 3.3-3.5 illustrate these posterior probabilities computed by the algorithm for the measured tissue spectra of each tissue class of belonging to that particular class for three different classification modes: 4-class, 3-class and binary. While the open symbols in the figures represent probabilities of correct class-membership, the closed symbols denote the probabilities for the misclassified tissue sites. One may note that while for the two-class case, the probability of belonging to the correct class is always greater than 50%, for classification involving more than two classes the corresponding probability of correct class-membership has fallen below 50%, for a few spectra, despite its being the highest.

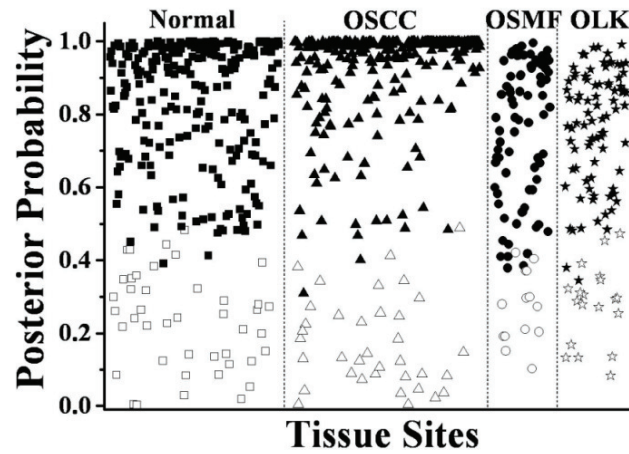


Fig. 3.3: Posterior probabilities for being classified as normal, OSCC, OSMF, and OLK for the Raman spectra of the oral tissue sites interrogated. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.

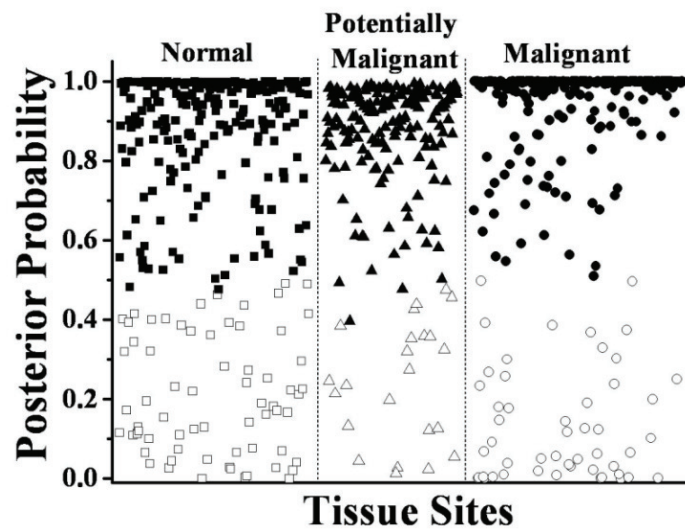


Fig.3.4: Posterior probabilities for being classified as normal, potentially malignant (OSMF & OLK grouped together), and malignant (OSCC) for the Raman spectra of the oral tissue sites interrogated. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.

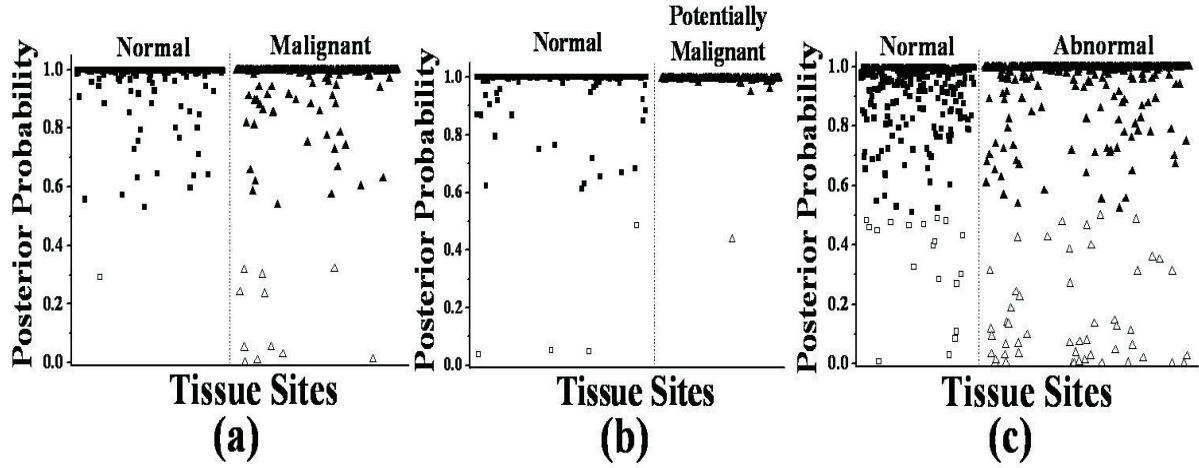


Fig. 3.5: Posterior probabilities for being classified as (a) normal and malignant (OSCC), (b) normal and potentially malignant (OSMF and OLK grouped together), and (c) normal and abnormal (OSCC, OSMF and OLK grouped together) for the Raman spectra of the oral tissue sites interrogated.

The ROC analyses of the classification results provide a quantitative evaluation of the overall performance of the diagnostic algorithm for different classification scenarios. While AUC corresponds to the area under the ROC curve and is a measure of performance of the algorithm for binary classification, HTM is the performance measure for multi-class classification and corresponds to the average of pair-wise area under the ROC curves between $L(L-1)/2$ pairs of classes, L being the number of classes. Table 3.5 lists the AUC value for binary classification and the HTM values obtained for 3- and 4-class classifications. While the estimated HTM values of the algorithm for the 4-class and 3-class classifications are 0.95 and 0.93 respectively, for the binary classifications the AUC value is found to be 0.99 for all three cases. It is important to mention here that the HTM (as well as the AUC) value is a quantitative measure of the gross performance of an algorithm and the HTM (and the AUC) for an ideal diagnostic algorithm will have a value of 1.

Table 3.5: *The results of ROC analyses for binary, three-class and four-class classifications using the MRDF-SMLR based diagnostic algorithm.*

Binary Classification			Multiclass Classification	
AUC	AUC	AUC	HTM	HTM
Normal vs. OSCC	Normal vs. Potentially Malignant	Normal vs. Abnormal	Normal vs. Malignant vs. Potentially Malignant	Normal vs. OSCC vs. OSMF vs. OPK
0.99	0.99	0.98	0.93	0.95

3.6 Discussions

The primary basis for Raman spectroscopic detection is the array of biochemical changes that take place as tissue undergoes neoplastic transformations. A close examination of the observed differences in the Raman bands (Fig. 3.2) 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 example, the increased intensities of the characteristic protein Raman peaks (1004, and 1338 cm^{-1}) in the OSCC spectra as compared to the spectra of other categories can be correlated to an over-expression of protein, believed to take place in OSCC [59], [60], [62]–[67], [100], [113]. On the other hand, a decreased concentration of the phospholipids and fatty acid as tissue changes from normal to OSCC [114] is the plausible explanation for the reduced intensities of the lipid-specific bands (1081, 1266, 1304, and 1450 cm^{-1}) [47], [60], [113] observed in the Raman spectra of OSCC and likely contributor to the differences in the regions of 1057-1100, 1248-1318, 1430-1474 cm^{-1} (Fig.

3.2). In contrast, the differences in the Raman spectra of OLK and OSMF with respect to normal are seen to be rather subtler (Fig. 3.2).

For the purpose of diagnosis, it is more relevant to explore the significance of the aforementioned spectral differences (observed between the different oral tissue types) towards pathological classification. A critical evaluation of the diagnostic results listed in Tables 3.2-3.4 reveals the following points that are worth noting. In the four-class classification case, the OSCC spectra are classified with the highest accuracy (89%) whereas the OSMF and OLK spectra are classified correctly in 85% and 82% of the sites. This is quite expected given the observation (see Fig. 3.2) that the spectral differences between the OSCC and the normal squamous tissue are more conspicuous as compared to that between normal and OSMF or OLK. In the three-class mode, the classification accuracy of the OSMF and OLK spectra together (belonging to the potentially malignant category) is found to improve to 88 %, though for the normal and the OSCC spectra the accuracies are found to reduce to 77% and 84 % respectively. In general, the overall accuracy for the three-class classification is observed to be poorer as compared to the four-class case (Tables 3.2 and 3.3). This observation is quantitatively supported by the results of the multi-class ROC analysis (listed in Table 3.5) which shows an HTM value of 0.95 for the 4-class as compared to 0.93 for the 3-class classification. The reason for this relatively poorer performance of the algorithm in the three-class mode is most likely due to the fact that compared with the spectral difference between normal mucosa and the individual OSCC or OLK or OSMF, the spectra of normal and pooled OSMF and OLK (potentially malignant lesions) show relatively small differences thereby (Fig. 3.2) resulting in a larger misclassification of normal tissue sites. The situation, however, drastically improves with a substantial increase in the accuracy of discrimination of the normal oral tissue sites (Table 3.4) when the algorithm switches to binary classification mode. One may see that the normal oral tissue sites are consistently classified correctly with

a specificity of over 94% and the improvement in specificity in going from normal vs. abnormal (OSCC, OSML and OLK put together) through normal vs. potentially malignant (OSMF and OLK) to normal vs. malignant (OSCC) classifications is only incremental. The sensitivities in all three cases of binary classification are also found to be significantly improved (as compared to the multi-class cases) with accuracies lying in the range between 94% and 99%. These observations are further supported by the two-class ROC analyses that resulted in AUC values of 0.99. The improvement in classification performance for the binary as compared to multi-class classification is not quite unexpected. This is because in binary classification an algorithm can carve out the appropriate decision boundary for only one of the classes, the other class is simply the complement. In contrast, the multiple class classification is intrinsically harder because here the classification algorithm has to learn to construct a greater number of separation boundaries or relations [115]. Since each of the multiple classes needs to be explicitly predicted, misclassification errors can occur in the construction of any one of the many decision boundaries. Further, the errors increase when there is significant overlap among the class members as in the present case.

It is pertinent to mention here, although our observations, *prima facie*, are seen to be grossly consistent with those recently reported by Singh *et al.* [45] (for example, in both the cases the overall accuracy of leave-one-out-cross-validation is found to be ~83% in classifying the tissue Raman spectra into malignant, pre-malignant and normal), there exist a few differences between the two studies due to which a direct comparison is not possible. For example, in their case, “normal” corresponded to the normal appearing mucosa in the oral cavity of a patient having malignant lesions at the contralateral side and “pre-malignant” referred to patches of lesions scattered in the same visually normal region. In contrast, we had chosen “normal”, as the normal oral mucosa of healthy volunteers with no history of any oral diseases and “potentially malignant” as either oral submucosal fibrosis or leukoplakia in the

oral cavity of patients who did not have any known oral malignancy. The reason underlying our choice is the fact that the normal appearing region surrounding the malignant tumor of a patient might have some subvisual malignant signatures due to the field effect of malignancy and associated biochemical changes [17] and this effect, although, expected to be more pronounced at the advanced stage of the disease, cannot be completely ruled out even at early stages [105]. In the present context, it means that spectral data from many of the uninvolved tissue sites of patients assumed to be normal might not be truly normal due to the field effect of malignancy [17]. Due to the similar reason, there is also a possibility that many of the spectra that were measured from the tissue sites of lesions sitting in the contralateral uninvolved region and were assumed to be premalignant might not be truly premalignant. In contrast, this possibility does not exist for the squamous tissue sites from healthy volunteers who have no history of any disease of oral cavity.

The availability of probability estimates for the MRDF-SMLR algorithm allows one to follow a systematic approach for selecting the appropriate threshold for classification. The availability of this quantitative information during tissue discrimination would allow the clinician to reassess those sites that are classified with higher relative uncertainty. For example, one may note in Fig. 3.5c that most of the normal tissue sites have been classified with a posterior probability of greater than 80%. However, a few normal tissue sites are seen to show a very low posterior probability of being classified as normal. Here, the probabilistic approach can offer an important advantage for optimizing the discrimination goals. Compared to a non-probabilistic classification scheme like PCA-LDA [45], [47], [48], [59], [64], [67] 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 miss any abnormal sites, as may be required for accurate screening of the oral cavity.

3.7 Summary

To summarize, we have presented the results of a pilot study carried out to investigate the clinical applicability of *in-vivo* Raman spectroscopy for discriminating normal from neoplastic lesions of human oral cavity. The *in-vivo* Raman spectra were measured from multiple sites of normal oral mucosa and of lesions belonging to three other histopathological categories, viz. oral squamous cell carcinoma (OSCC), oral submucous fibrosis (OSMF) and leukoplakia (OLK). In order to test the ability of the measured Raman spectra to predict pathological designation of the interrogated tissue sites, the probability based multi-class diagnostic algorithm was applied on the set of Raman spectra corresponding to the different oral tissue sites. With respect to histology as the gold standard, the diagnostic algorithm was found to provide a leave-one-subject-out cross validation accuracy of up to ~89% in classifying the oral tissue spectra into the different tissue categories. When employed for binary classification, the algorithm resulted in a sensitivity and a specificity of 94% each in delineating the normal from all the abnormal oral tissue spectra belonging to OSCC, OSMF and OLK pooled together. The results clearly demonstrate that Raman spectroscopy along with an appropriate diagnostic algorithm has considerable potential to provide real-time, non-invasive diagnosis of malignant and potentially malignant lesions of oral cavity in a clinical situation. While there exists reports to demonstrate the potential of Raman spectroscopy for *in vivo* discrimination of lesions in only a part of the oral cavity (i.e. buccal mucosa or tongue tissue) [45], [49], [65], [112], this is the first a full-scale clinical study 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.

The objective of the present study was a comprehensive evaluation of the efficacy of *in-vivo* Raman spectroscopy for differential detection of oral lesions in the whole of the oral cavity. The intrinsic variability of the tissue Raman spectra that might result due to various

factors like variations in the anatomical locations within the oral cavity being interrogated, influence of tobacco usage, age, gender etc. were not taken into account in the present study. It should be noted that these factors might also affect the algorithm's performance, which in turn might influence the diagnostic efficacy of *in-vivo* Raman spectroscopy has not yet been demonstrated. The result of detailed clinical studies undertaken to address these issues will be presented in the chapters 4 and 5.

Chapter 4

Tobacco Consumption Induced Changes in the Healthy Oral Mucosa and its Effect on Differential Diagnosis of Oral Lesions

4.1 Introduction

The common general objective of all the studies reported thus far [45-49], [59]–[67], [116], [117] 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 present, in this chapter, 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 present the results of our investigation on 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 [4], [5].

4.2 Study Design

In-vivo Raman spectra were measured using an in-house assembled, compact, and portable Raman spectroscopic system described earlier in the Section-2.2.

Table 4.1: *Histopathological distribution of tissue sites included in the study.*

# Patients	# 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)

The *in-vivo* study was conducted at the Tata Memorial Hospital (TMH), Mumbai with the approval of the TMH Ethical Committee. All the spectral measurements were performed by the participating head and neck surgeon (PI) using a standard protocol as described earlier in the Section-3.2 which was maintained for all individuals in this study. 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 4.1. 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.

4.3 Results

Fig. 4.1 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.

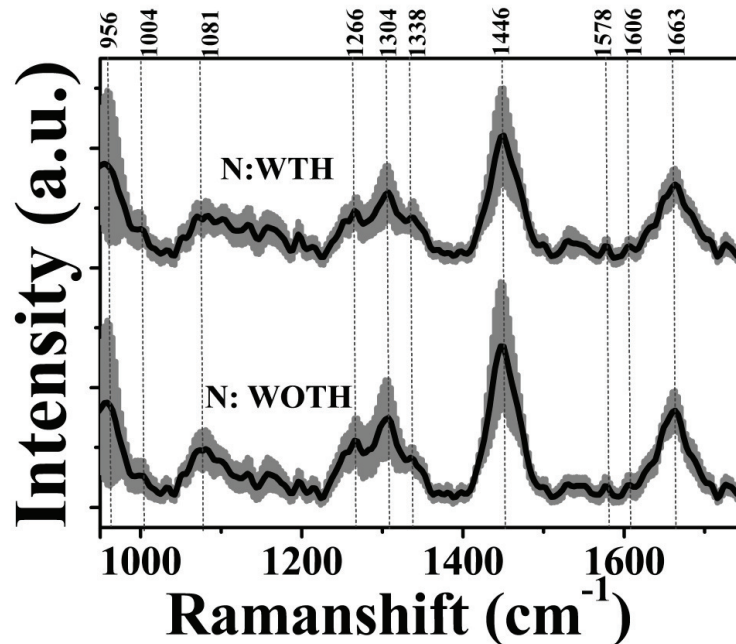


Fig. 4.1: Mean, normalized Raman spectra of the oral mucosa of healthy volunteers with tobacco consumption habit (N:WTH) and without any tobacco consumption habit (N:WOTH). The error bars (gray) represent ± 1 standard deviation.

For illustrating the spectral differences between the tissue types, mean difference spectrum obtained by subtracting one mean spectrum from the other is shown in Fig. 4.2. The grey 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 $1244\text{--}1272\text{ cm}^{-1}$, $1297\text{--}1313\text{ cm}^{-1}$, $1434\text{--}1456\text{ cm}^{-1}$ and $1643\text{--}1672\text{ cm}^{-1}$ are found to be considerably higher in the spectra of tobacco non-user as compared to those of tobacco users indicating changes in the lipid contribution.

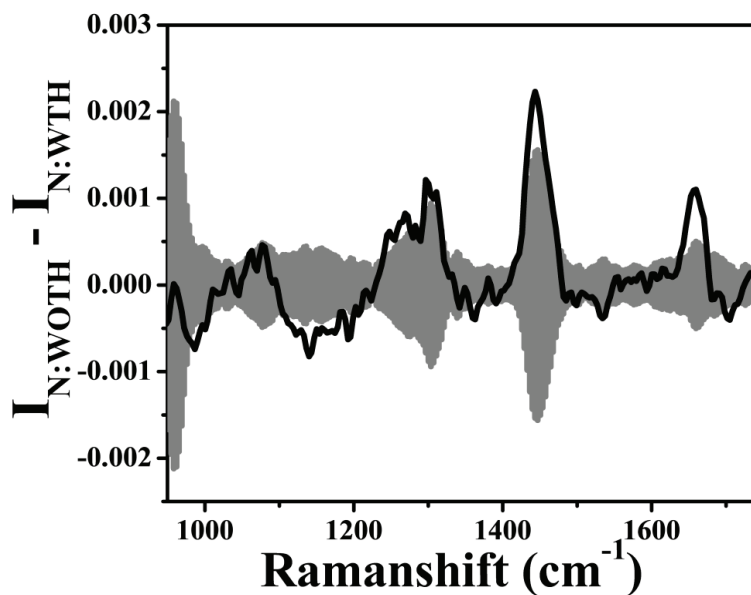


Fig. 4.2: Mean difference spectra showing statistical differences between oral tissue Raman spectra of healthy volunteers without any tobacco consumption habit (N:WOTH) and with tobacco consumption habit (N:WTH). Gray bands indicate 95% confidence intervals of the difference determined by standard error confidence intervals.

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 [101] was applied in binary mode on the spectral data sets. Table 4.2 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.

Table 4.2: 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.

	Raman Classification	
Tissue Category	N:WOTH	N:WTH
N:WOTH (n=83)	95%	5%
N:WTH (n=204)	5%	95%

Fig. 4.3 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.

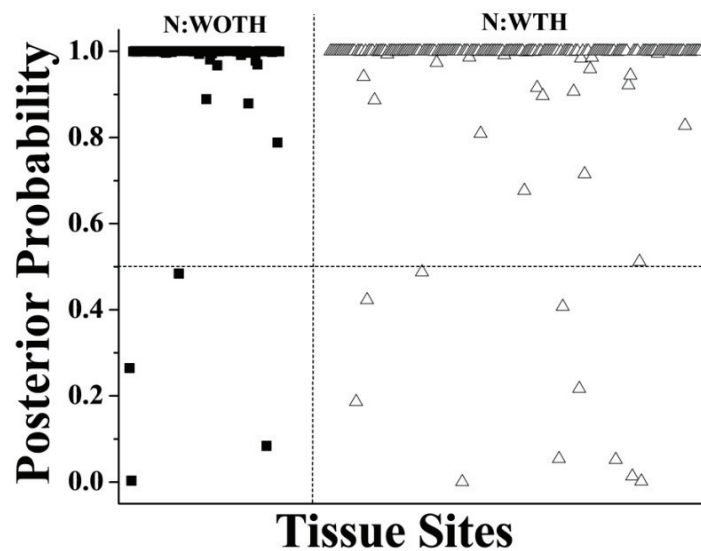


Fig. 4.3: Posterior probabilities of being classified as normal without any tobacco consumption habit (N:WOTH) and normal with tobacco consumption habit (N:WTH).

Fig. 4.4 shows the mean Raman spectra of the different oral lesions investigated. The spectra are 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, [41], [44], [52], [59]–[61], [64], [99], [109–111] 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 1663 cm^{-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.

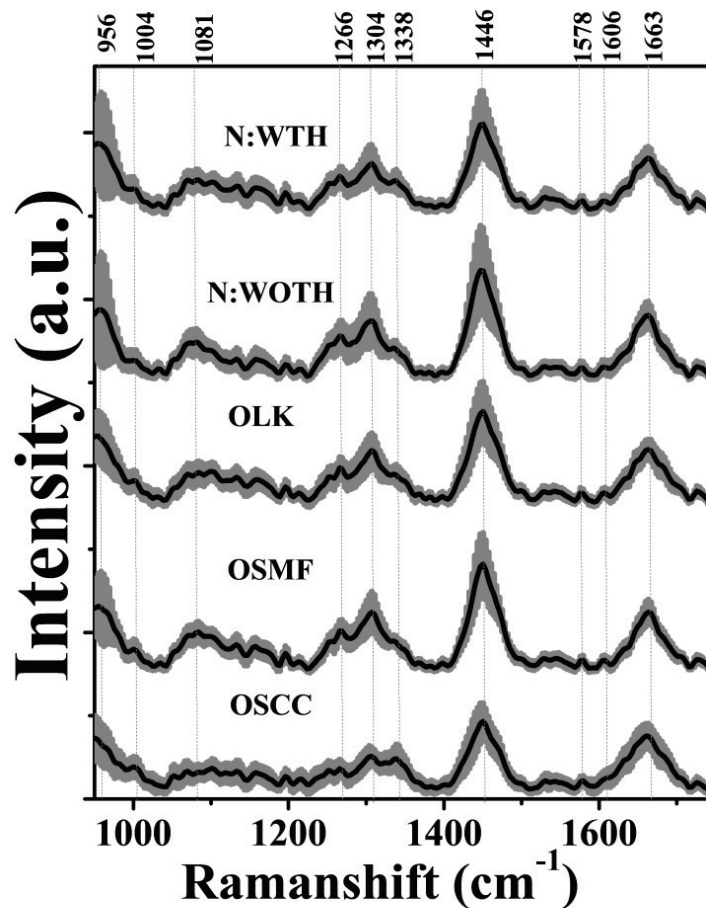


Fig. 4.4: Mean, normalized Raman spectra of OSCC ($n=316$), OSMF ($n=94$), OLK ($n=105$), N:WOTH ($n=83$), and N:WTH ($n=204$). The error bars (gray) represent ± 1 standard deviation.

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 4.3-4.5 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%. Figs. 4.5a, 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 4.3: 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.

Tissue Category	Raman Classification		
	N:WOTH	N:WTH	PM
N:WOTH (n=83)	93%	0%	7%
N:WTH (n=204)	1%	92%	7%
PM (n=199)	9%	4%	87%

Table 4.4: 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.

Tissue Category	Raman Classification			
	N:WOTH	N:WTH	PM	OSCC
N:WOTH (n=83)	86%	5%	3%	6%
N:WTH (n=204)	4%	82%	10%	4%
PM (n=199)	2%	8%	85%	5 %
OSCC (n=316)	2%	4%	5%	89%

Table 4.5: 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 using the MRDF-SMLR based 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)	92%	5%	2%	1%	0%
N:WTH (n=204)	2%	80%	10%	6%	2%
OSCC(n=316)	1%	6%	88%	1%	4%
OSMF (n=94)	0%	10%	2%	86%	2%
OLK(n=105)	1%	5%	5%	3%	86%

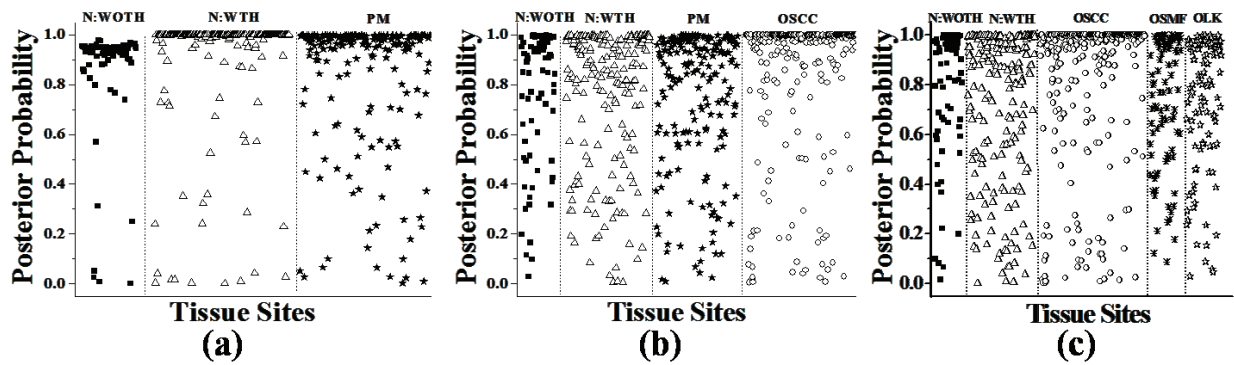


Fig. 4.5: Posterior probabilities for the Raman spectra of the oral tissue sites of being classified as: **(a)** normal without any tobacco consumption habit (N:WOTH), normal with tobacco consumption habit (N:WTH), and potentially malignant (PM; consisting of spectra of OSMF and OLK tissue sites pooled 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.

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 4.6-7 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 4.8 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.

Table 4.6: 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.

Tissue Category	Raman Classification			
	N:ALL	OSCC	OSMF	OLK
N:ALL (n=287)	85%	8%	4%	3%
OSCC (n=316)	8%	89%	0%	3%
OSMF (n=94)	14%	0%	85%	1%
OLK (n=105)	13%	5%	0%	82%

Table 4.7: 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) using the MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.

Tissue Category	Raman Classification			
	N:WOTH	OSCC	OSMF	OLK
N:WOTH (n=83)	96%	3%	1%	0%
OSCC (n=316)	2%	90%	2%	6%
OSMF (n=94)	1%	0%	97%	2%
OLK (n=105)	1%	5%	2%	92%

Figures 4.6a 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 ~58% of the correctly classified tissue sites in each tissue category have a posterior probability >0.80 for Case-2 (spectra of tobacco users excluded from normal), the corresponding fraction is reduced to ~81%, 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 4.8 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.

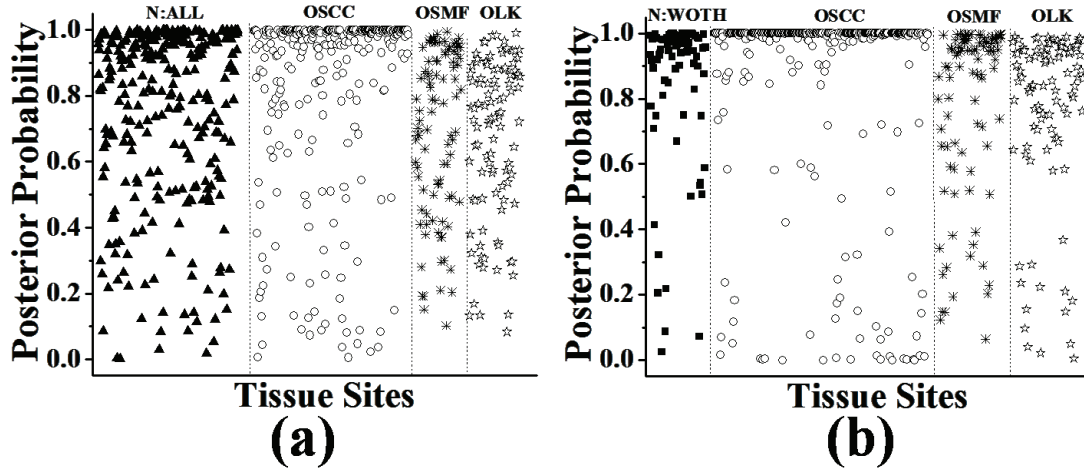


Fig. 4.6: Posterior probabilities for the Raman spectra of the oral tissue sites of being classified as: **(a)** N:ALL (spectra of tobacco users and non-users put together), OSCC, OSMF, and OLK, and **(b)** N:WOTH (spectra of tobacco users excluded), OSCC, OSMF, and OLK.

Table 4.8: 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).

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

4.4 Discussions

There has been no systematic study, thus far, on the use of *in-vivo* 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 tobacco consumption or any other oral diseases. This is required because recent advances in the Raman spectroscopy for exploring the tobacco consumption induced changes in the oral mucosa of healthy volunteers. Further, no reports exist on investigating whether the outcome of a spectroscopic 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 [10], [11], and tobacco, a genotoxic as well as a local irritant, is a key contributor to that [4]–[6]. In fact, it has been observed that more than 75-90% of patients who develop dysplasia or malignancy of the oral cavity has a history of long-term tobacco use [7]. 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 across 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 [118]. 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 [119]. 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 from other potentially malignant 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 4.2), the accuracy is ~80% when it comes to separating these from the OSCC, OSMF and OLK along with the tobacco non-users in the multi-class classification platform (Table 4.5).

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 Fig. 4.1 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 1004, 1213, 1338, 1578, and 1606 cm^{-1} associated with protein 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 [4], [120] 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 [121].

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 4.2) 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 4.6 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 4.8 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 [101].

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 are required to address these issues.

4.5 Summary

To summarize, we have presented in this chapter the results of a clinical study 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 the probabilistic multi-class diagnostic algorithm employed to discriminate malignant and potentially malignant oral lesions from the healthy oral mucosa. An important finding of the study is 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 can serve as predictor of increased risk of dysplasia or malignancy in an individual. Another notable finding is that exclusion of the spectral data of the oral cavity of the healthy volunteers from the reference normal database provided 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. These 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. In the next chapter we will describe the results of an in-depth analysis of the variability of the tissue Raman spectra due to variations in the anatomical locations and their effect on the diagnostic efficacy of *in-vivo* Raman spectroscopy.

Chapter 5

Anatomical Variability of In-Vivo Raman Spectra of Normal Oral Cavity and its Effect on Oral Tissue Classification

5.1 Introduction

Successful use of *in-vivo* Raman spectroscopy for oral cancer diagnosis requires an appropriate diagnostic algorithm that can best classify the measured spectra from an unknown tissue by using a stored database of spectra of tissues of known histopathologic classification [48]. The major factor that confounds the development of a diagnostic algorithm and thus the outcome of clinical spectroscopic diagnosis, is the large intra-patient as well as inter-patient variability in the intensity and line shape of the measured tissue Raman spectra. While a part of this variability originates from several unavoidable factors like the presence of body fluids, variable contact of the diagnostic probe with the tissue surface etc., the intrinsic anatomical differences over the different tissue sites interrogated in a given organ may also significantly add to this variability. Guze *et al.* [48] were the first to address the issue of anatomical variability of the *in-vivo* Raman spectra of the human oral cavity. They measured Raman

spectra in the higher wave number ($1800\text{-}3000\text{ cm}^{-1}$) region from different anatomical regions of the oral cavity of healthy volunteers and found that the spectra showed significant spectral variability in the observed C-H stretch bands near 3000 cm^{-1} which they correlated to the different degrees of keratinization of the oral mucosa. This was followed by an *in-vivo* study by Huang and co-workers [47] who measured *in-vivo* Raman spectra from the healthy volunteers in the conventional fingerprint region ($800\text{-}1800\text{ cm}^{-1}$). In this spectral region also considerable differences in intensity and line shape of the Raman spectra were observed across different anatomical locations. However, in both of the these studies no attempt was made to investigate whether the observed anatomical variability had any bearing on the performance of a diagnostic algorithm employed for stratifying different oral lesions based on their measured Raman spectra.

In this chapter, we present the results of an *in-vivo* study carried out to characterize the variability of the *in-vivo* Raman spectra measured from the different anatomical sites of healthy volunteers and investigate the effect of inter-anatomical spectral variations on the performance of a diagnostic algorithm that was used to discriminate malignant and potentially malignant oral lesions from the healthy oral mucosa. While an unsupervised classifier based on Fuzzy c-means clustering [122] was used for analyzing the Raman spectra of the different anatomical sites of the normal oral cavity, the diagnostic algorithm was the probability based non-linear MRDF-SMLR algorithm [101] capable of direct multi-class classification of different oral tissue pathologies. It was found that use of an unsupervised classifier based on Fuzzy c-means clustering grouped the normal oral tissue sites into four different anatomical clusters based on their measured Raman spectra. Consideration of these anatomical clusters was found to provide an overall classification accuracy of 95% as against an accuracy of 87% obtained in the case of pooled spectral set without any anatomical clustering.

5.2 Materials and Methods

5.2.1 Study Design

In-vivo Raman spectra were measured using a compact and portable Raman spectroscopic system (Fig. 2.1) as described in the Section-2.2. The *in-vivo* study was conducted at the Tata Memorial Hospital (TMH), Mumbai with the approval of the TMH Ethical Committee. All the spectral measurements were performed by the participating head and neck surgeon using a protocol which was maintained for all the individuals (patient and healthy volunteers) participating in this study. The detailed protocol was already described in the Section-3.2. This study involved 26 healthy volunteers with no history of the disease of the oral cavity and 113 patients enrolled for medical examination of the oral cavity at TMH. The *in-vivo* Raman spectra were recorded from multiple sites of the oral cavity of healthy volunteers as well as patients. On an average, 11 tissue sites were interrogated from a healthy volunteer and 3-4 sites from a patient. The different tissue sites investigated belonged to either of the following anatomical sites: buccal mucosa (BM), dorsal tongue (DT), lateral tongue (LT), ventral tongue (VT), outer lip (OL), lip vermillion border (LVB), hard palate (HP) and soft palate (SP). From each tissue site single spectrum was measured. A quality filter was employed as part of the data collection scheme. The exclusion criterion was based on the SNR of the measured tissue spectra with a cut-off value of 10. Only those spectra which were having $SNR > 10$ were treated as good quality tissue Raman spectra and retained for subsequent analysis. This resulted in a total of 268 spectra from 26 healthy volunteers and 337 spectra from 112 patients. Out of the 337 spectra from patients, 73 were of OSMF from 19 patients, 188 were of OSCC from 69 patients and 76 were of OLK from 25 patients. The details of the histopathological distribution of the tissue sites included in the study are summarized in the Table 5.1. Each site was treated separately and analyzed via the unsupervised and supervised algorithms developed.

Table 5.1: Histopathological distribution of tissue sites included in the *in-vivo* study.

# Individuals	# Sites	Histopathological Diagnosis
69	188	OSCC
19	73	OSMF
25	76	OLK
26	268	Normal

5.2.2 Data Analysis

The sequence of pre-processing steps executed on the measured raw Raman spectra as detailed in the Section-3.4. Following pre-processing steps, various multivariate statistical methods were employed for analyses of 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. A brief description of the methods is as follows.

Unsupervised Classification

In order to characterize the variability in the set of Raman spectra measured from different anatomical locations of the oral cavity of healthy volunteers, an unsupervised classification algorithm based on the principle of Fuzzy c-means (FCM) clustering was developed and applied on the spectral data set. The goal was to reveal the underlying structure of the spectral data and segment the data in groups with similar spectral patterns.

The theory of FCM clustering is detailed elsewhere [122]. In brief, it seeks to partition a set of input data in such a way that the data belonging to the same clusters show a certain degree of closeness or similarity, whereas data belonging to different clusters are as dissimilar as possible. The clusters have “fuzzy” boundaries, in the sense that each data value

belongs to each cluster to some degree or other. The development of the FCM clustering algorithm involved two steps, first, finding the best location for the cluster centers, and next, determining the optimum number of clusters. Since for the given spectral data there is no *a priori* information about the structures in the data, one has to intuitively specify the number of underlying clusters. For a specified number (L) of clusters, the algorithm finds the best location for the centers of the clusters by minimization an objective function (OB) given by:

$$OB_m = \sum_{i=1}^S \sum_{l=1}^L (\mu_{il})^m \|X_i - C_l\|^2 \quad \dots\dots\dots (5.1)$$

where, m is the fuzzification parameter whose value can be any real number >1 , ' S ' is the number of data measured (size of the data), X_i is the i^{th} of D -dimensional input data, C_l is the centre of the l^{th} cluster, μ_{il} is the membership of X_i data in the l^{th} cluster and $\|*\|$ represent the norm expressing the similarity between any measured data point and the cluster center.

Fuzzy partitioning is carried out through an iterative optimization of (Eq. 5.1) with an update of membership (μ_{il}) for each i^{th} data point ($1 \leq i \leq S$) belonging to cluster ' l ' ($1 \leq l \leq L$) given by following equation:

$$\mu_{il} = \frac{1}{\sum_{j=1}^L \left[\frac{\|X_i - C_l\|}{\|X_i - C_j\|} \right]^{2/m-1}}, \quad 1 \leq i \leq S, \quad 1 \leq l \leq L, \quad \sum_{l=1}^L (\mu_{il}) = 1 \quad \dots\dots\dots (5.2)$$

Following this the cluster centre C_l were updated by using equation:

$$C_l = \frac{\sum_{i=1}^S (\mu_{il})^m X_i}{\sum_{i=1}^S (\mu_{il})^m}, \quad 1 \leq l \leq L \quad \dots\dots\dots (5.3)$$

The iterative procedure continues until the cluster centers stabilize. The iteration stops when, $\max_{il} (\|\mu_{il}^k - \mu_{il}^{k+1}\|) < \varepsilon$, where ε is a termination criterion normally chosen to be very small (~ 0.001) and k is number of the iterations. The algorithm incorporates fuzzy set's concept of partial membership and assigns the membership value to the data points for the clusters within a range of 0 to 1 with a restriction that the sum of memberships of a data point in all the clusters must be equal to one. The fuzzification parameter ' m ' determines the degree of fuzziness in the clusters and $m=2$ was chosen in the present algorithm. It is pertinent to mention here that the value of m can be any real number greater than or equal to one ($1 \leq m \leq \infty$). As ' m ' approaches one, the partition becomes hard, while as ' m ' tends to infinity, the partition becomes completely fuzzy. The preferred choice of ' m ' has been shown to be $m=2$ [123].

After having calculated the location of the cluster centers and determined the degree of membership for the data points of belonging to these clusters, the next task of the FCM algorithm is to select the optimum number of clusters (from among the finite set of numbers of clusters specified) which are well separated and compact. The selection of the optimum number of clusters is done by minimizing the ratio of average distance within the cluster to the distance between the cluster centers and the optimum number of clusters is which yields minimum ratio.

Supervised Classification

Supervised classification was used to classify the oral lesions of different pathologic categories based on their measured Raman spectra. A probability based multivariate statistical algorithm [101] was employed for that purpose which has been described previously in the Section-2.4.1. It [101] consists of two steps: 1) extraction of diagnostic features from the spectra using nonlinear maximum representation and discrimination feature

[102] (MRDF) and 2) a probabilistic scheme of classification based on linear sparse multinomial logistic regression [103] (SMLR) for classifying the nonlinear features into corresponding tissue categories.

For quantifying the effect of anatomical variability on the outcome of the supervised classification, a multiclass receiver-operating characteristic (ROC) analysis, described in details in the Section-2.4.4 was carried out on the classification results of the multi-class diagnostic algorithm.

5.3 Results

Fig. 5.1 shows the mean, normalized Raman spectra for different anatomical locations of the oral cavity of healthy volunteers, with the error bars representing one standard deviation. Each spectrum is the average of the spectra from the respective number of tissue sites interrogated in each anatomical location. From the figures it is evident that the variation in the measured spectral intensity is comparable for all the anatomical locations investigated. The percentage variation (σ / \bar{x}) in the spectral intensities from the different measurement sites was observed to lie in the range of ~13%-25% over the respective number of tissue sites included in the anatomical locations. Here, \bar{x} and σ are the mean and (one-) standard deviation of the intensity values of the tallest peak in the Raman spectra of different tissue sites of each category.

The observed spectra are similar to those reported in Raman spectroscopic studies of human oral cavity by others, with major peaks located at ~956, 1004, 1081, 1266, 1304, 1338, 1450, 1578, 1606, 1663 cm^{-1} corresponding to various Raman-active biomolecules, notably collagen, elastin, keratin, lipids, and minerals known to be present in oral tissues [41], [47], [100], [113], [59]–[61], [63]–[67]. Although the peak shapes and locations are consistent across all the anatomical locations, subtle but significant differences are observed

in peak intensities between the different anatomical locations indicating biochemical differences inherent in the tissue types of different anatomies.

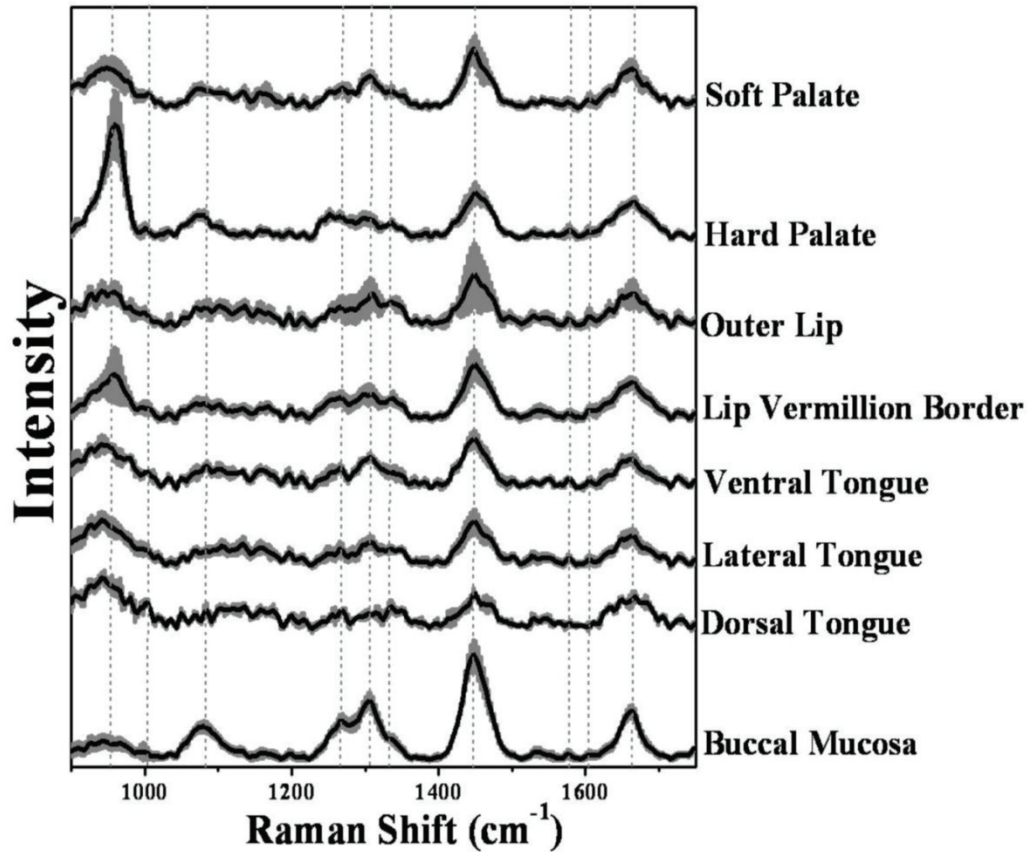


Fig. 5.1: Mean, normalized Raman spectra of the different anatomical sites of oral cavity of healthy volunteers. The error bars (gray) represent ± 1 standard deviation.

For determining the optimum number of clusters, the ratios of average distance within the cluster(s) to the distance between the clusters were calculated for different number of clusters and the optimum number of cluster was selected which yielded the minimum ratio value. Fig. 5.2 shows the plot of the ratio value for varying number of clusters. It is apparent from the figure that the optimum number of clusters corresponds to four.

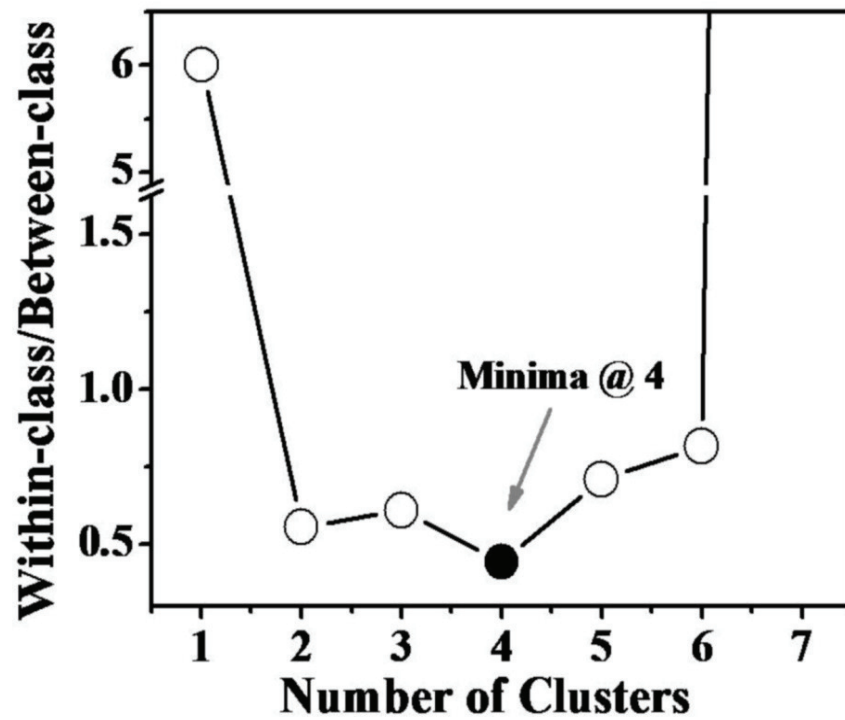


Fig. 5.2: Ratio of within-class to between-class separation as a function of number of clusters. The minimum represents the optimum number of clusters (=4).

A cluster was assigned to a spectrum for which its predicted class-membership of belonging to that cluster was the highest. Table 5.2 lists the number of spectra of different anatomical locations that comprised the four different clusters resulted from Fuzzy c-means cluster analysis. It is evident that the whole set of spectra is primarily made up of four prominent anatomical clusters (AC), namely AC-1 containing spectra of lip, AC-2 containing spectra of buccal mucosa, AC-3 containing spectra of hard palate and AC-4 containing spectra of tongue and soft palate. Table 5.3 lists the results of unsupervised classification in the form of a confusion matrix displaying comparison of predicted membership with the actual membership of the four anatomical clusters.

Table 5.2: The number of spectra of different anatomical locations comprising the four different clusters resulted from Fuzzy c-means cluster analysis.

Anatomy	Findings of Fuzzy c-means cluster analysis			
	Cluster-I	Cluster-II	Cluster-III	Cluster-IV
OL (N=47)	38	4	4	1
LVB (N=24)	19	1	3	1
BM (N=46)	8	33	5	0
HP (N=25)	0	0	23	2
DT (N=26)	0	0	2	24
VT (N=25)	0	0	3	22
LT (N=52)	1	0	11	40
SP (N=23)	2	3	8	10

Table 5.3: Results of unsupervised classification in the form of confusion matrix displaying the comparison of predicted membership with the actual membership of the four anatomical clusters. DT: dorsal tongue, LT: lateral tongue, VT: ventral tongue, SP: soft palate, OL: outer lip, LVB: lip vermillion border, HP: hard palate and BM: buccal mucosa. N: total number of tissue sites belonging in a group.

True Anatomical label	Anatomical label predicted by cluster analysis			
	AC-I	AC-II	AC-III	AC-IV
AC-I (N=71) OL, LVB	80%	7%	10%	3%
AC-II (N=46) BM	17%	72%	11%	0%
AC-III (N=25) HP	0%	0%	92%	8%
AC-IV (N=126) DT, LT, VT, SP	2%	3%	19%	76%

Fig. 5.3 shows the mean, normalized Raman spectra corresponding to the four different anatomical clusters obtained from the cluster analysis. It is apparent that the spectra belonging to the different clusters show significant variability in line shape and intensity. Fig 5.4 shows the mean difference spectra obtained by subtracting pair-wise the mean tissue spectrum of each cluster from that of the other. The grey bands show the confidence intervals estimated by multiplying SE with a t-value corresponding to 95% ($p < 0.05$) confidence intervals and the degrees of freedom equal to the number of spectral measurements (corresponding to the clusters) minus the number of clusters. A number of significantly different Raman bands (characteristics of proteins, lipids and hydroxyapatite) are observed for each cluster with respect to each other which further confirms the statistical significance of the spectral differences between the clusters.

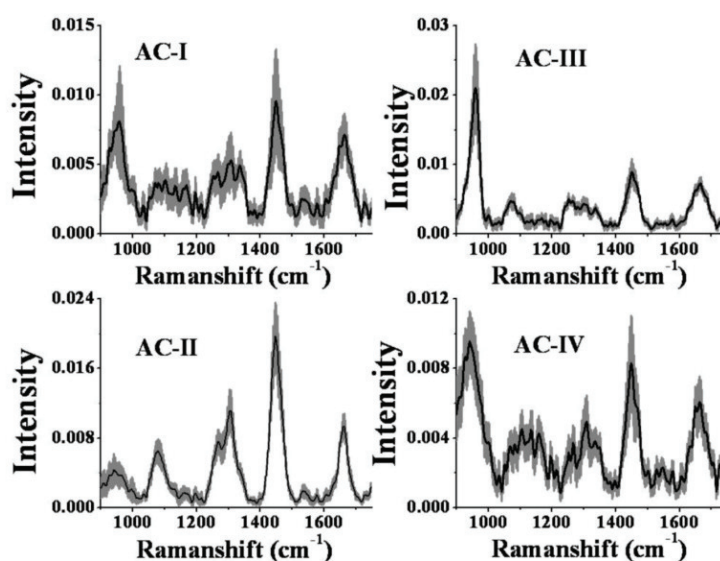


Fig. 5.3: Mean, normalized Raman spectra of the different anatomical sites of normal healthy tissues belonging to the four different anatomical clusters (AC): **(i)** AC-I: outer lip and lip vermillion border; **(ii)** AC-II: buccal; **(iii)** AC-III: hard palate, and **(iv)** AC-IV: dorsal, lateral, ventral tongue and soft palate.

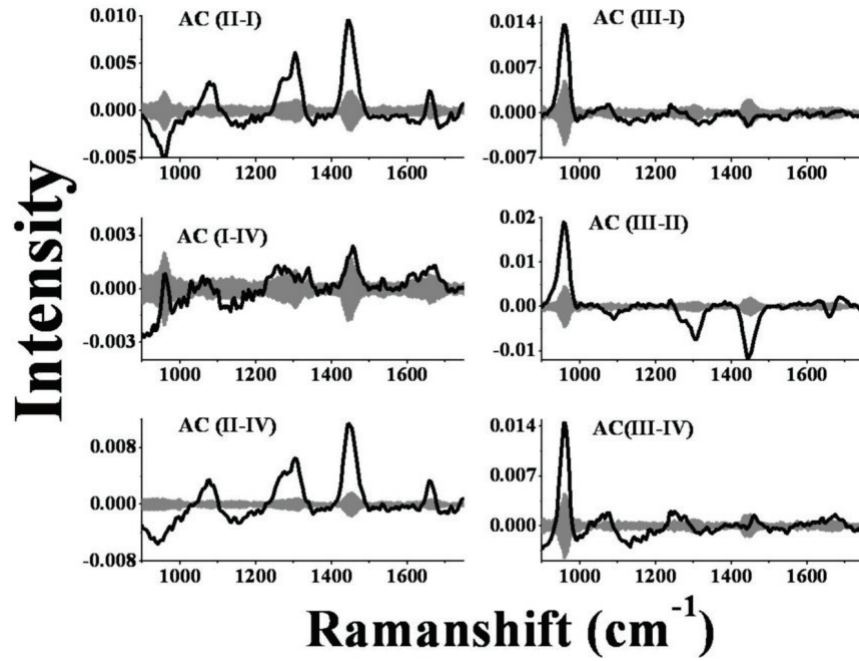


Fig. 5.4: Mean pair-wise difference spectra showing statistical differences between the four anatomical clusters. Gray bands indicate the 95% confidence intervals of the difference determined by standard error confidence intervals.

The whole set of Raman spectra measured from the tissue sites of healthy oral mucosa as well as of lesions, irrespective of their anatomical locations, were grouped into three categories: (1) normal comprising spectra of tissue sites belonging to healthy mucosa, (2) malignant comprising spectra from OSCC tissue sites, and (3) potentially malignant consisting of spectra of OSMF and OLK tissue sites pooled together. Fig 5.5 shows the mean Raman spectra averaged over all the tissue sites investigated irrespective of their anatomy for these different tissue pathologies. 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, 1252, 1338, 1578 and 1606 cm^{-1} Raman bands, believed to be due to proteins [41], [45], [49], [59]–[61], [63]–[67] were found to be higher for

malignant tissues as compare to normal. On the other hand, the lipid-specific Raman peaks at $\sim 1081, 1266, 1304, 1450$, and 1663 cm^{-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.

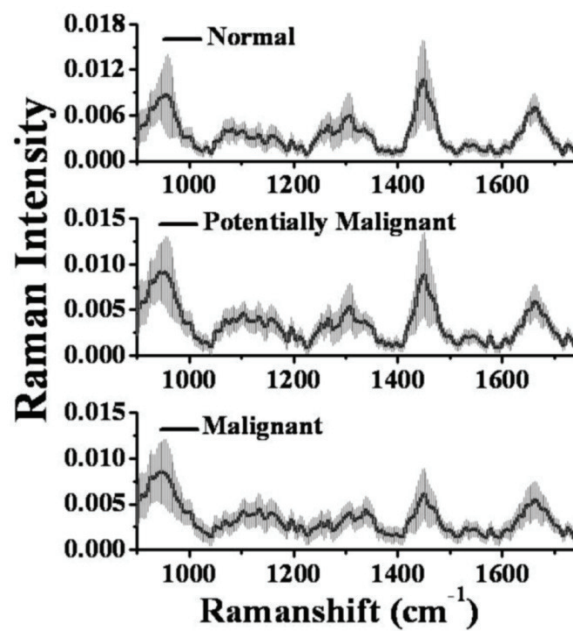


Fig. 5.5: Mean, normalized Raman spectra of malignant, potentially malignant and normal oral tissue sites. The error bars (gray) represent ± 1 standard deviation.

In order to investigate the effect of the observed inter-anatomical spectral variations on the outcome of supervised classification, the MRDF-SMLR algorithm [101] was applied on the set of normal, potentially malignant and malignant tissue spectra corresponding to each anatomical cluster as well as the whole set of spectra without any anatomical clustering. The common objective was to classify the measured tissue Raman spectra into three pathologic categories: normal, potentially malignant and malignant. Table 5.4 shows the confusion matrices listing classification results of different tissue pathologies for AC-I, AC-

II, AC-III and AC-IV. 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 classification results reveals that the supervised diagnostic algorithm, when applied on the set of spectra stratified into the anatomical clusters, correctly discriminated normal, malignant and potentially malignant tissue sites with an overall accuracy of 94% (94 / 100) for AC-I, 94% (201 / 214) for AC-II, 93% (42/45) for AC-III and 96% (237 /246) for AC-IV respectively.

Table 5.4: Confusion matrices displaying results of classification of the Raman spectra of oral tissue sites, into three classes: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC) using the MRDF-SMLR based diagnostic algorithm for different anatomical clusters.

Cluster	Pathology Diagnosis	Raman Diagnosis			Overall Classification Accuracy
		Normal	Malignant	Potentially Malignant	
AC-I	Normal (N=71)	66	0	5	94%
	Malignant (N=9)	0	9	0	
	Potentially Malignant (N=20)	1	0	19	
AC-II	Normal (N=46)	45	0	1	94%
	Malignant (N=82)	0	81	1	
	Potentially Malignant (N=86)	3	8	75	
AC-III	Normal (N=25)	23	1	1	93%
	Malignant (N=9)	0	9	0	
	Potentially Malignant (N=11)	1	0	10	
AC-IV	Normal (N=126)	117	4	5	96%
	Malignant (N=88)	0	88	0	
	Potentially Malignant (N=32)	0	0	32	

Table 5.5: Confusion matrix displaying results of anatomy matched overall classification of the Raman spectra of oral tissue sites into three classes: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC) using the MRDF-SMLR based diagnostic algorithm. Each element of the matrix was obtained by computing the sum of the corresponding elements of the matrices listed in Table-5.4 and dividing the sum by the total sum of the elements over the corresponding rows multiplied by 100.

Pathology Diagnosis	Raman diagnosis		
	Normal	Malignant	Potentially Malignant
Normal (N=268)	94%	2%	4%
Malignant (N=188)	0%	99%	1%
Potentially Malignant (N=149)	3%	6%	91%

Table 5.6: Confusion matrix displaying results of classification of the Raman spectra of oral tissue sites into three classes: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC) using the MRDF-SMLR based diagnostic algorithm without considering anatomical clustering.

Pathology Diagnosis	Raman diagnosis		
	Normal	Malignant	Potentially Malignant
Normal (N=268)	86%	5%	9%
Malignant (N=188)	7%	88%	5%
Potentially Malignant (N=149)	7 %	7%	86%

Table 5.5 shows the overall classification results for the anatomy matched spectral data in a single confusion matrix. Each element of the matrix was obtained by computing the

sum of the corresponding elements of the matrices listed in Table 5.4 and dividing the sum by the total sum of the elements over the corresponding rows multiplied by 100. For comparison's sake, the classification results obtained for the whole set of spectra without any anatomical clustering is shown in Table 5.6. The overall classification accuracy is found to be higher (95%) in anatomy-matched classification (Tables 5.5) as compared to that (87%) when inter-anatomical variability was not considered during classification (Table 5.6).

Table 5.7 shows the Pillai's V values obtained for the set of Raman spectra belonging to different pathology classes before and after anatomical clustering. One can see higher values of Pillai's V for the anatomy matched spectral set indicating a larger separation between the malignant, potentially malignant and normal oral tissue spectra in this case.

Table 5.7: Pillai's V and HTM values for the set of Raman spectra belonging to different pathology classes before and after anatomical clustering. While Pillai's V is a quantitative measure of the separation between different pathology classes, HTM is a measure of the performance of a diagnostic algorithm.

Measure	Considering anatomical clustering				Pooled data (Without considering anatomical clustering)
	AC-I	AC-II	AC-III	AC-IV	
Pillai's V	1.42	1.18	1.70	1.77	0.72
HTM	0.99	0.99	0.97	0.99	0.96

In addition to assigning class labels, the diagnostic algorithm also yielded posterior probabilities of the measured tissue sites belonging to an 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. Figs. 5.6a and b illustrate these posterior probabilities computed by the algorithm for the measured tissue spectra of each

tissue class of belonging to that particular class for two different cases: Case-1 when anatomical clustering was considered and Case-2 when anatomical clustering was not considered. While the closed symbols in the figures represent probabilities of correct class-membership, the open symbols denote the probabilities for the misclassified tissue sites. It is apparent from the figures that while more than 95% of the correctly classified tissue sites in each tissue category have a posterior probability >0.80 for Case-1 (anatomy-matched classification), the corresponding fraction is reduced to $\sim 85\%$, when anatomical clustering was not considered.

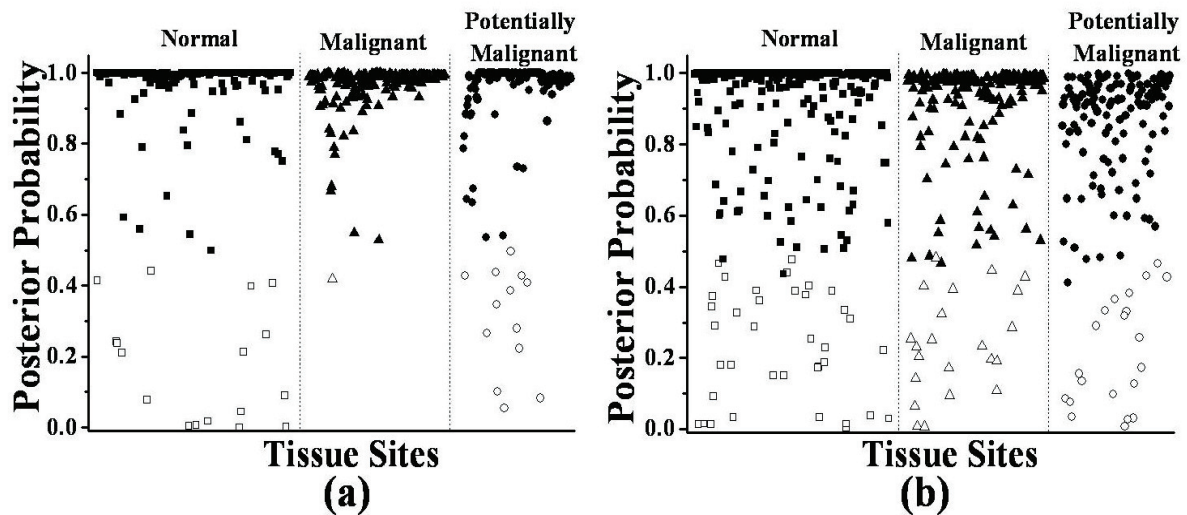


Fig. 5.6: Posterior probabilities computed by the MRDF-SMLR algorithm for the measured tissue spectra of each tissue class of belonging to that particular class **(a)** when anatomical clustering was considered and **(b)** when anatomical clustering was not considered. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.

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 5.7 lists the HTM values obtained for the two cases. While the estimated HTM value

for Case-1 is seen to be 0.99, for Case-2 the corresponding value is seen to be 0.96. 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.

5.4 Discussions

The results show significant variability in line shape, peak position and relative intensity of the measured Raman spectra across the different anatomical locations of the oral cavity. Intrinsic heterogeneity in the biochemical make-up of the oral cavity due to biologically and functionally different anatomical regions [124] seems to be the plausible origin of the observed variability in the measured Raman spectra from the different tissue sites. For example, the high intensities of the lipid-specific peaks at around 1081, 1266, 1304, 1450 and 1663 cm^{-1} [47], [100] in the Raman spectra of the tissue sites from buccal mucosa (Fig. 5.1) corroborate the presence of significant amount of lipids in the buccal mucosa [47]. The other non-keratinized stratified squamous epithelia such as lip which also is known to contain lipid is evident from the appearance of fairly intense Raman peak around 1450 cm^{-1} . In contrast, the spectra of the sites belonging to keratinized tissue like tongue show larger intensity of the 1338 cm^{-1} Raman peak characteristic of proteins and a relatively broader amide-I peak around 1663 cm^{-1} indicating dominant contributions of keratin in tongue. On the other hand, the hard palate shows a dominant peak around 956 cm^{-1} associated with bone minerals characteristics of phosphate stretching vibration of hydroxyapatite [125]. The spectra also show fairly intense Raman peak around 1338 cm^{-1} characteristics of keratin [47] revealing keratinization of hard palate.

An important goal of the present study was to characterize the oral mucosa of the healthy volunteers based on their measured Raman spectra. The algorithm developed using the

principle of Fuzzy c-means clustering was used for this purpose. The method of Fuzzy clustering was chosen because the conventional hard clustering methods (like k-means clustering etc.) assign each data point to a unique cluster thereby resulting in a very crisp segmentation [126]. However, in real life situations, like in the present case where there is a considerable overlap in the inherent structure of spectral data over different anatomical locations, this kind of crisp segmentation is expected to result in unsuitable clusters that may not properly reflect the structure of the data set [126]. One of the major advantages of Fuzzy clustering is that it is a soft segmentation method which allows the data elements to belong to more than one clusters and also expresses the confidence in the assignments of data points to these different clusters [126]. Thus, each cluster is a fuzzy set of all the data patterns. In fact, it represents the similarity a data point shares with each cluster with a function (termed the membership function) whose values (called memberships) are between zero and one. Memberships close to unity (i.e. larger membership values) indicate higher confidence in the assignment of the data to the cluster signifying a high degree of similarity between the data and the cluster while memberships close to zero imply less similarity. Further, another important advantage of Fuzzy clustering is that it is more robust in terms of reducing the local minima of the objective function (it seeks to minimize) as compared to the conventional hard clustering techniques, and thus, it can avoid the possibility of undesired clustering results originating from the undesired local minima of the objective function [126].

The significance of the spectral differences observed across the different anatomical locations of the oral cavity is grossly reflected in the findings of the cluster analysis. It is seen that the whole set of spectra of healthy volunteers can be categorized into four anatomical clusters: (1) OL, and LVB classified into AC-I with an accuracy of 80%; (2) BM into AC-II with an accuracy of 72%; (3) HP into AC-III with an accuracy of 92% and (4) DT, LT, VT and SP into AC-IV with an accuracy of 76%. A close examination of the observed spectra

(Fig. 5.3) of the four clusters and their pair-wise difference spectra (Fig. 5.4) help clarify the basis of this categorization. It is evident from the figures that the four clusters can largely be separated based on the Raman features related to either protein or lipid or hydroxyapatite. For example, the difference spectra for AC-III and the rest of the anatomical clusters reveal that the main differences appear in the spectral region containing the Raman peak characteristics of hydroxyapatite [47], [125]. Also statistically significant ($p < 0.05$) differences between AC-III and AC-II were observed in the spectral region containing lipid related signatures. The differences between AC-II and AC-I, AC-I and AC-IV, and AC-II and AC-IV, were again related to differences at in the spectral region containing both protein and lipid related signatures [47], [100], [113], [125]. A detailed analysis of the classification results reveals that most the AC-IV spectra that have been misclassified (13 out of 30) belong to soft palate. This is not quite unexpected given the differences in biochemical make-up of soft palate and tongue as evident from Fig. 5.1. Out of the thirteen (total 23) soft palate spectra that have been misclassified, eight fall in AC-III (Table 5.2) probably because of their close proximity with the hard palate. Further, most of the AC-II spectra that have been misclassified into AC-I (8 of total 13 misclassified spectra) are from normal volunteers having habits of either smoking or chewing tobacco for more than 10 years and have less intense lipid peak as compare to the correctly classified spectra of AC-II.

While the findings of the cluster analysis reveal that the Raman spectra of the healthy squamous tissues of the oral cavity can be grouped into four separate clusters, it is more relevant to find the significance of these inter-anatomical spectral variations towards pathological classification. One can see from the classification results listed in the confusion matrices (Tables 5.5 and 5.6) that the anatomy-matched classification provides an improved overall classification accuracy of 95% as against an accuracy of 87% when anatomical clustering was not taken into account during classification. These observations are further

supported by the multi-class ROC analyses that resulted in an HTM value 0.99 for anatomy-matched classification and 0.96 for the case when clustering was taken into consideration. The reason for this improvement in the classification accuracy can be understood from Table 5.7 which lists the values of the Pillai's V before and after anatomical clustering. It is seen that Pillai's V values corresponding to all the clusters are larger than its value obtained for the pooled spectral data (without any clustering). 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 [107].

5.5 Summary

To summarize, we present the results of a clinical study carried out to characterize the variability of the *in-vivo* Raman spectra of the different anatomical sites of the oral cavity of healthy volunteers and investigate the effect of inter-anatomical spectral variations on the performance of the probabilistic multi-class diagnostic algorithm employed to discriminate malignant and potentially malignant oral lesions from the healthy oral mucosa. An unsupervised cluster analysis using Fuzzy c-means clustering algorithm was conducted for quantifying the underlying structure of the normal oral tissue spectra. The algorithm was found to segment the normal oral tissue sites, based on similarity of spectral patterns, into four major anatomical clusters (AC): (1) outer lip, and lip vermillion border into AC-I with an accuracy of 80%; (2) buccal mucosa into AC-II with an accuracy of 72%; (3) hard palate into AC-III with an accuracy of 92%; (4) dorsal, lateral and ventral tongue and soft palate into AC-IV with an accuracy of 76%. Further, a probabilistic multi-class diagnostic algorithm, developed for supervised classification, was used to classify the whole set of measured tissue Raman spectra into three categories: normal, potentially malignant and malignant. The results showed that the diagnostic algorithm, when applied on the pooled set

of spectra from all the clusters, correctly discriminated normal, malignant and potentially malignant tissue sites with 86%, 88%, and 86% accuracy respectively, which amounts to an overall accuracy of 87%. However, when the anatomy-matched data sets were considered, the overall classification accuracy was found to improve to 95% with the algorithm correctly discriminating the corresponding tissue sites with 94%, 99%, and 91% accuracy respectively. The results of the present study demonstrate that the lesser the inter-anatomic variability in the spectra included in the database, the better the accuracy of classification yielded by the algorithm. In terms of clinical application, this signifies that one should incorporate anatomy-matched spectral data in the training set to have an improved performance of the algorithm (i.e. more accurate diagnosis) for the test set.

Chapter 6

A Comparison of Fluorescence and Raman Spectroscopy for Clinical Diagnosis of Oral Neoplasia

6.1 Introduction

The results presented in chapters 3-5 have established with evidence that Raman spectroscopy has significant potential in providing noninvasive and accurate diagnosis of oral lesions in a clinical situation [42], [117]. It is also known that it has the intrinsic ability to detect biochemical information in a tissue in greater detail than fluorescence spectroscopy [44], [52]. However, given currently available technologies, fluorescence is a significantly stronger candidate than Raman for practical applications because of the simplicity and considerably lower cost involved in its instrumentation. In fact, owing to the same reason, fluorescence spectroscopy was one of the first to be developed as a diagnostic tool for oral cancer detection [21], [53]. Though each of these techniques has been validated separately in clinical settings for oral cancer detection [22], [50], [51], [112], a comprehensive, side-by-side evaluation of the relative efficacies of the two methods has not been addressed in the

literature. In this chapter we will describe the results of an *in-vivo* clinical study carried out to evaluate and compare the relative capabilities of fluorescence and Raman spectroscopy for *in-vivo* discrimination of various oral tissue pathologies in a clinical setting. Such a comparative evaluation is required because it will help understand which approach is best suited for differential diagnosis of different abnormal lesions of oral cavity.

6.2 Materials and Methods

6.2.1 Clinical Spectroscopy Systems

A nitrogen laser based, compact and portable fluorescence spectroscopic system (Fig 6.1a) developed earlier [29] by us was used to measure *in-vivo* autofluorescence spectra of human oral cavity. Light delivery to and collection from tissue was achieved with a fiber-optic probe consisting of seven 400 micron core diameter fibers (0.22 NA) arranged in a six-around-one configuration (Avantes Inc. Broomfield, CO 80021).

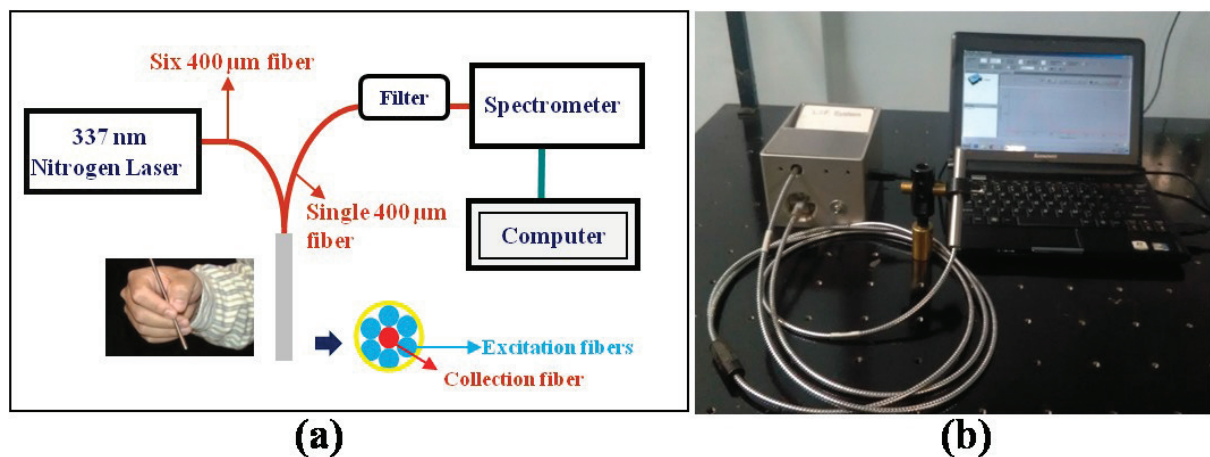


Fig. 6.1: (a) Schematic, and (b) setup of N_2 laser based fluorescence spectroscopy system (Ref. [29]).

The schematic of experimental setup for fluorescence measurement is shown in Fig. 6.1b. The six peripheral fibers deliver excitation laser light to the tissue surface while the

central fiber collects emitted autofluorescence light from the tissue surface. The fluorescence emission collected by the fiber-optic probe is then dispersed and detected with a chip-based spectrometer (model number USB-4000, Ocean Optics, Dunedin, FL). The nitrogen laser had 337 nm emission, 10-Hz repetition rate, 5-nanosecond pulse width, and average pulse energy of 50 ± 5 μ J at the tissue surface. An integration time of 500 ms was used for each spectral measurement. The overall spectral resolution of the system was ~ 5 nm.

In-vivo Raman spectra were measured with the portable Raman spectroscopy system described in the Section-2.2.

6.2.2 Spectroscopic Measurements

The study was conducted at Tata Memorial Hospital Mumbai with the approval of Institutional Ethical Committee. All patients participating in this study pre-operatively signed an informed consent, permitting the investigative *in-vivo* fluorescence and Raman spectroscopic acquisition of oral tissues. The spectral measurements were performed by the participating head and neck surgeon using a protocol as described in the Section-3.2. This was maintained for all individuals in this study.

From each of the investigate tissue site, fluorescence and Raman spectra were measured sequentially. *In-vivo* optical spectra recorded from a total of 800 sites were included in the study. Of these, 300 sites were from normal squamous tissue of 30 healthy volunteers and 500 sites were from abnormal oral cavity tissue of 150 patients. Out of these abnormal tissue sites, 90 (from a total 25 patients) were identified as OSMF by the examining doctor and from these no biopsies were taken. Of the remaining tissue sites, 310 (from a total 95 patients) were histopathologically characterized as OSCC and 100 (from a total of 30 patients) as OLK. The healthy volunteers had no clinically observable lesions of the oral

cavity and also had no history of malignancy. Each site was treated separately and classified via the diagnostic algorithm detailed elsewhere [101] and briefed in Section-2.4.1.

6.2.3 Data Processing and Analysis

The *in-vivo* autofluorescence spectra were recorded in the 375-700 nm spectral range and process as described in details elsewhere [29]. In brief, prior to fluorescence measurement from tissue site of a subject, a background spectrum was acquired and subtracted from all subsequently acquired spectra from that subject. Spectra were further binned along the wavelength axis in 2 nm intervals and filtered with a first-order, eleven point Savitzky-Golay smoothing filter for noise removal. The resultant spectra were corrected for the system spectral response by using a NIST traceable calibration lamp (LS-1, Ocean optics, Inc., Dunedin, FL). From each site, three spectra were recorded and were averaged to yield a single spectrum per site. Following data processing, a method of normalization was adopted to remove the absolute intensity information from the spectra that might be affected by many unavoidable experimental factors. In the case of fluorescence, the spectrum from each site of a subject was normalized with respect to the integrated intensity from that site.

In the case of Raman, the sequences of pre-processing steps as detailed in the Section-3.4 were executed on the measured raw Raman spectrum. Each background subtracted Raman spectrum was normalized to its mean spectral intensity across all the Raman bands. The normalized fluorescence and Raman spectra were used for subsequent data analysis as describe previously in the Section-2.4.1.

6.3 Results and Discussions

The average fluorescence and Raman spectra are shown in Figs. 6.2 (a) and (b) respectively for OLK (n=100), OSMF (n=90), OSCC (n=310), and normal squamous (n=300) tissues,

with the error bars representing plus and minus one standard deviation. The measured fluorescence and Raman spectra showed a variety of spectral differences between different tissue types. These differences are the manifestation of the different biochemical changes (in the concentration or conformation of different biomolecules present in oral tissue) associated with disease transformation.

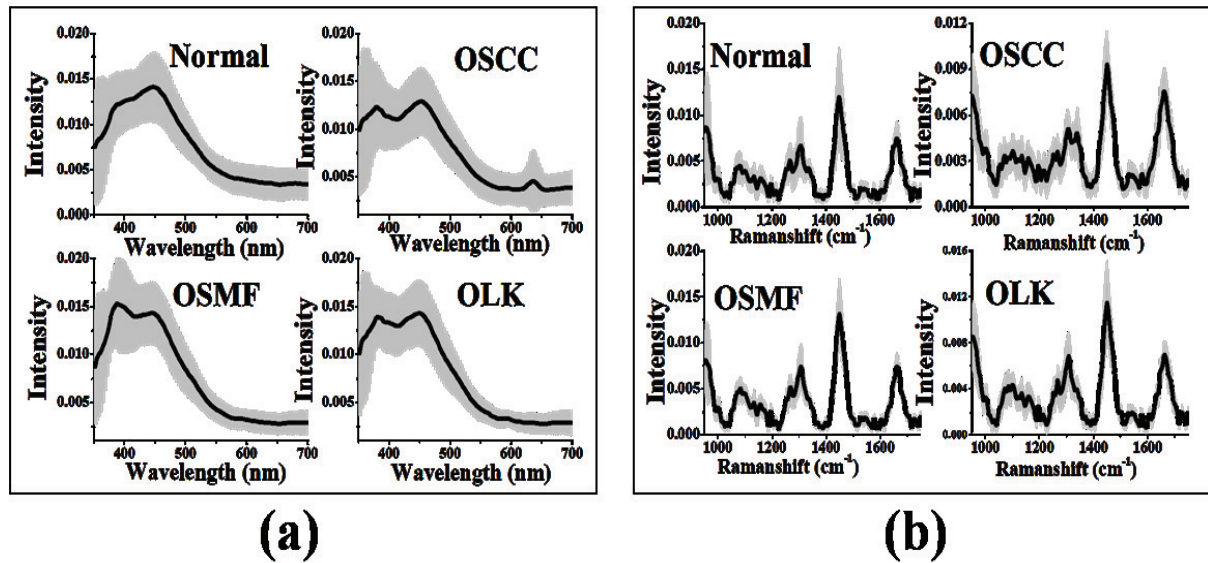


Fig. 6.2: Mean (a) auto-fluorescence spectra, and (b) Raman spectra of human oral tissues belonging to different pathology categories. Spectra are plotted with one standard deviation to represent inter-patient variability.

In the case of fluorescence, the spectral signatures of the structural proteins such as collagen and elastin are characterized by the intensity of 390 nm band [53]. The intensity of this spectral band was found to be the most intense in OSMF tissues (Fig. 6.2a). The differences in concentration and oxidation state of coenzymes such as NADH and FAD contribute to the changes in the intensity of the broad 460 nm band [22], [27], [53] which was found stronger in the fluorescence spectra from normal squamous tissues (Fig. 6.2a). The OSCC lesions were found to have overall decrease in fluorescence intensity as compared to the normal oral

tissue. Further, the fluorescence spectra of OSCC lesions were seen to have a peak around 635 nm, believed to be due to endogenous porphyrins [127]. These changes are consistent with those reported in other studies and result from the structural and metabolic changes associated with cancer [22], [26], [27], [53].

In the case of Raman, the intensity of Raman bands at 1004 cm^{-1} , 1213 cm^{-1} , 1338 cm^{-1} , 1578 cm^{-1} and 1606 cm^{-1} assigned to proteins or DNA were found to be higher for oral lesion as compared to normal tissue indicating the protein over expression [59]–[63], [65]–[67], [114]. On the other hand, the lipid-specific Raman peaks at ~ 1081 , 1266 , 1304 , 1450 , and 1663 cm^{-1} were found to be stronger in the spectra of normal oral tissues. The differences between Raman spectra of normal and malignant (OSCC) are found to be more pronounced as compared to that between normal and potentially malignant tissues (OSMF and OLK).

It can also be seen from Fig. 6.2 that although there are differences in the average autofluorescence and Raman spectra among the oral tissue pathologies, these differences are fairly subtle and swamped by the large inter-patient variability. In order to extract and quantify the diagnostic information that is otherwise hidden in the apparently overlapping set of respective spectral data, each spectral data set was analyzed with the help of the non-linear MRDF-SMLR based spectroscopic diagnostic algorithms. Table 6.1 shows the results of simultaneous multi-class classification performances of the fluorescence and Raman spectroscopic techniques. The results were obtained based on leave-one-individual-out cross validation of the entire data set. The algorithm with fluorescence spectra as input achieved an overall classification accuracy of 77% (618 correctly classified tissue sites out of 800). Based on the fluorescence measurements, the OSCC, OSMF, OLK and normal tissue could be correctly classified with accuracies of 75%, 88%, 78%, and 76%, respectively. In contrast, when Raman spectra were used as the input to the algorithm, the overall classification accuracy was found to improve to 86% with the algorithm providing classification accuracies

of 84%, 96%, 94%, and 83% for OSCC, OSMF, OLK, and normal oral tissues, respectively. While 690 out of a total 800 sites (i.e. ~86%) tissue site could be correctly classified using Raman spectroscopic measurements, with fluorescence spectroscopic measurements the figure was 618 out of 800 (i.e. ~77%).

Table 6.1: *Confusion matrices displaying the results of simultaneous multi-class classification of the fluorescence and Raman spectra of oral tissue sites into four classes: normal, OSCC, OSMF, and OLK using the non-linear MRDF-SMLR based diagnostic algorithm in leave-one-subject-out cross validation mode.*

	Fluorescence				Raman			
Pathology Diagnosis	Normal	OSCC	OSMF	OLK	Normal	OSCC	OSMF	OLK
Normal (N=300)	76%	19%	3%	2%	83%	15%	1%	1%
OSCC (N=310)	21%	75%	3%	1%	14%	84%	1%	1%
OSMF (N=90)	5%	1%	88%	6%	1%	0%	96%	3%
OLK (N=100)	15%	1%	6%	78%	4%	1%	1%	94%

In addition to assigning class labels, the diagnostic algorithms also yielded posterior probabilities of class membership for the measured tissue sites. The posterior probabilities of the predicted results are shown in Fig. 6.3 for the two techniques. The spread of the predicted probability of the classification results were analyzed using multiclass ROC analysis for quantitative comparison of the relative diagnostic performances of the two techniques. The Raman spectroscopy was found to provide better class prediction (area under ROC curve=0.97) as compared to fluorescence spectroscopy (area under ROC curve=0.90).

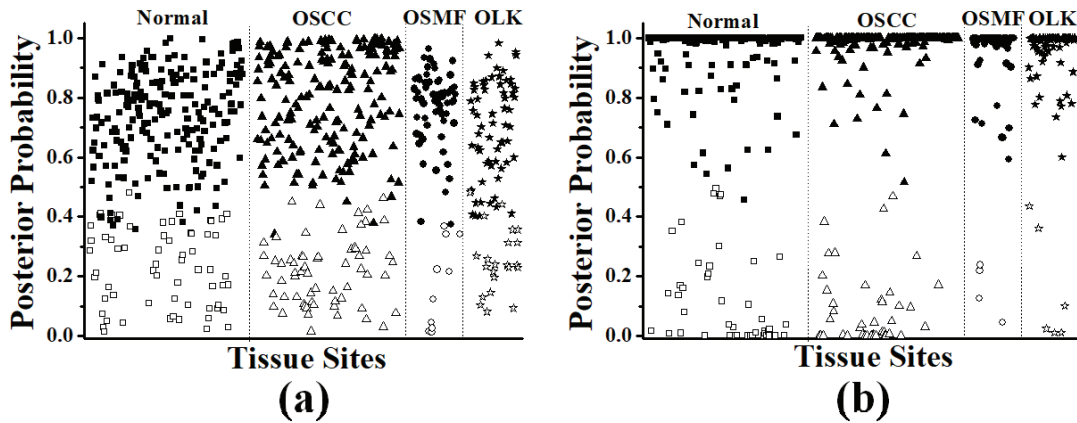


Fig. 6.3: Posterior probabilities for being simultaneous multiclass classification as normal, OSCC, OSMF and OLK for the (a) fluorescence, and (b) Raman spectra of the human oral cavity. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.

Further, the spectra from OSCC, OSMF and OLK lesions were grouped together to form a separate category called abnormal and the performance of the diagnostic algorithm was checked in delineating the normal tissue from all the abnormal oral lesions. Table 6.2 shows the findings of binary classification (normal verses abnormal) performance of the Raman and the fluorescence spectroscopic techniques. In binary classification, the sensitivity and specificity provided by Raman spectroscopy were found to be 95% and 93%, respectively, as compared to 91% sensitivity and 90% specificity provided by fluorescence spectroscopy. The overall classification accuracy for Raman in delineating normal from abnormal tissue was found to be 94% as compared to the fluorescence spectroscopy which provided an overall performance of 90% for the same task. Thus with respect to the current cost of the Raman system which is ten times more as compared to that of the fluorescence system, the increase in the binary classification accuracy is only marginal. The area under the curve (AUC) for ROC analysis also showed the marginal improvement in the posterior probability predictions with Raman data (Fig. 6.4). The AUC increases from 0.95 for fluorescence to 0.96 for Raman

spectroscopy. This clearly indicates that fluorescence spectroscopy can very well serve as a tool for screening of oral cancer with limited economy resources. However when it comes to more detailed classification of different pathologies, Raman spectroscopy outperformed fluorescence spectroscopy by a relatively larger margin ($\sim 10\%$).

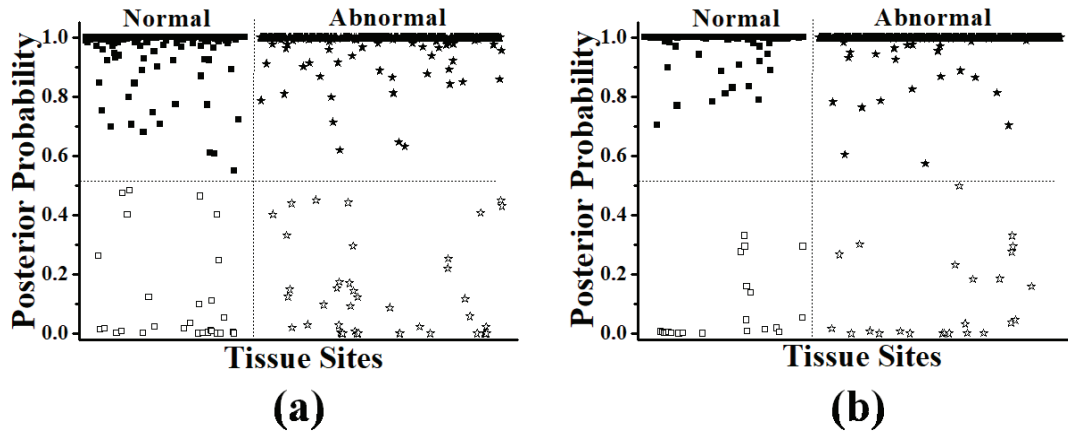


Fig. 6.4: Posterior probabilities for being classified as normal and abnormal (OSCC, OSMF, and OLK) for the (a) fluorescence, and (b) Raman spectra of the human oral cavity. The closed symbols represent probabilities of correct class-membership and the open symbols represent the probabilities for the misclassified tissue sites.

Table 6.2: Confusion matrices displaying the results of binary classification of the fluorescence and Raman spectra of oral cavity into normal and abnormal tissue using the non-linear MRDF-SMLR based diagnostic algorithm. Here the spectra belonging to OSCC, OSMF and OLK pooled together are referred to as “Abnormal”.

	Fluorescence		Raman	
Pathology Diagnosis	Normal	Abnormal (OSCC+OSMF+OLK)	Normal	Abnormal (OSCC+OSMF+OLK)
Normal (N=300)	90%	10%	93%	7%
Abnormal (N=500)	9%	91%	5%	95%

6.4 Summary

To summarize, we have presented the results of a comparative evaluation of the relative capabilities of *in-vivo* fluorescence and Raman spectroscopy for discriminating the different histopathologic categories of human oral tissues in a clinical setting. A probability based multivariate statistical algorithm utilizing nonlinear maximum representation and discrimination feature for feature extraction and sparse multinomial logistic regression for classification was used for direct multi-class classification of different oral tissue pathologies based on their measured spectra. For both fluorescence and Raman spectroscopy, the diagnostic algorithm was found to provide accuracy greater than 90% in delineating the normal from all the abnormal oral tissue spectra (belonging to OSCC, OSMF and OLK pooled together). However, Raman spectroscopy was found to outperform fluorescence spectroscopy in simultaneous multi-class classification of different oral tissue pathologies by a margin of over 9% with an overall improvement in accuracy from 77 % to 86%. The results of this study indicate that if the objective is only to delineate abnormal from normal oral mucosa, as may be required for routine screening procedures, the fluorescence approach can serve as a method of choice with a cost effective system. However, when it comes to more accurate tissue classification, as may be required for clinical diagnosis of the pathological state of a lesion, Raman spectroscopy out performs fluorescence spectroscopy.

Chapter 7

Optical Spectroscopy and Imaging Based Point-of-Care Diagnostic Devices for Automated Screening of Oral Neoplasia

7.1 Introduction

Although the results of the studies presented thus far have established with evidence that optical spectroscopy is a promising tool for noninvasive and accurate diagnosis of various oral lesions, the diagnostic results obtained in these studies were based on offline analyses of the spectral data acquired with the diagnostic systems. However, for achieving the final goal of routinely using optical spectroscopy for the management of oral cancer in a clinical setting, it is imperative to deliver diagnostic results in real time [37], [39], [40], [44], [52]. This requires development of a stand-alone optical spectroscopic device capable of providing non-invasive and accurate diagnostic feedback in real-time. In order to meet this requirement it is essential to equip the device with an appropriate software that will not only provide the

necessary interface for the hardware control of the device and automation of data acquisition, but also have the ability to deliver online diagnosis of the interrogated tissue sites based on an instant analysis of the spectra being acquired thereof. To be able to make reliable diagnostic predictions about the acquired optical spectra from the tissue sites under interrogation, the software will need to be integrated with a robust diagnostic algorithm appropriately trained on a statistically significant number of previously measured (using the device) optical spectra from various pathologically certified oral lesions as well as healthy oral tissues.

Thus the overall development involves a number of important tasks: (1) design and development of a compact, portable and preferably low-cost point-of-care diagnostic system, (2) development of an easy-to-use software interface for hardware control of the device for automated measurement of spectral data and its documentation, and (3) development and integration of a robust diagnostic algorithm and database of spectra (on which the algorithm will be trained) of representative tissue pathologies with the data acquisition software to be able to deliver online diagnosis of the tissue site based on the instant analysis of the spectra.

We have seen in the previous chapter that although Raman spectroscopy allows for the most accurate tissue classification as may be useful for clinical diagnosis of the pathological state of a lesion, fluorescence spectroscopy holds promise for use in the applications such as routine or mass screening of population having high risk of oral cancer occurrences. On the basis of this observation, two point-of-care devices were developed for screening of human oral cavity abnormalities, one based on combined fluorescence and diffuse reflectance spectroscopy, and the other based on fluorescence imaging. The results of the clinical evaluation of efficacies of these devices in detecting lesions of oral cavity in patients with oral neoplasia at Raman Foundation Cancer Hospital and Research Centre, Indore are also detailed in the chapter.

7.2 The Optical Spectroscopy based Point-of-Care Diagnostic Device

The optical spectroscopy based point-of-care diagnostic device, called OncoDiagnoScope (Fig 7.1), is a light emitting diode (LED) based compact and portable diagnostic system intended for use in a clinical environment for oral cancer detection. A touch-screen enabled graphic user interface (GUI) software developed and installed in a tablet computer integrated with the device provides the necessary interface for the hardware control of the device and automation of data acquisition and analyses. The changes in the optical properties of oral mucosa taking place during carcinogenesis can be measured with this USB powered device using a pencil-sized stainless steel fiber optic probe. The fiber optic probe is brought in contact with the suspected tissue of the oral cavity of a patient and light is shone upon it. The light coming out of the tissue is captured and fed to the tablet computer where it is analyzed by a smart diagnostic algorithm which can instantly determine whether the tissue is cancerous or not. Two types of light (fluorescence and reflectance) returned from the oral tissue, following illumination by light from two LEDs, are used to determine the pathology status without disturbing or destroying the tissue. The entire investigation procedure per subject using this device is less than 15 minutes as compared to several hours required by the conventional procedure of biopsy followed by histopathology. Using this device, oral lesions can be accurately separated in a non-invasive manner from healthy oral tissues based on their natural characteristics in response to light.

7.2.1 Device Hardware

The system consists of a 365 nm UV LED (for inducing fluorescence), a broadband white LED (for exciting diffuse-reflectance) and a chip-based miniaturized fiber-optic spectrometer (USB-4000, Ocean Optics Inc.) all accommodated in a rectangular acrylic box (~27x16x4 cm) fitted with a tablet computer on its top. The box is SMA connected to three optical

fibers running through the three legs of a custom-designed, trifurcated fiber-optic diagnostic probe (Applied Optical Technologies P Ltd, Thane, India) made of stainless steel. An in-house (designed and) developed miniaturized electronic data acquisition card is mounted inside the acrylic box and connected to the USB port of the tablet computer from where it draws power and produces electrical signal for driving two constant current sources to power up the LEDs. The powering up of the LEDs is done sequentially and in synchronization with the spectrometer such that at a given time only one of the LEDs is on and the corresponding spectrum is recorded.

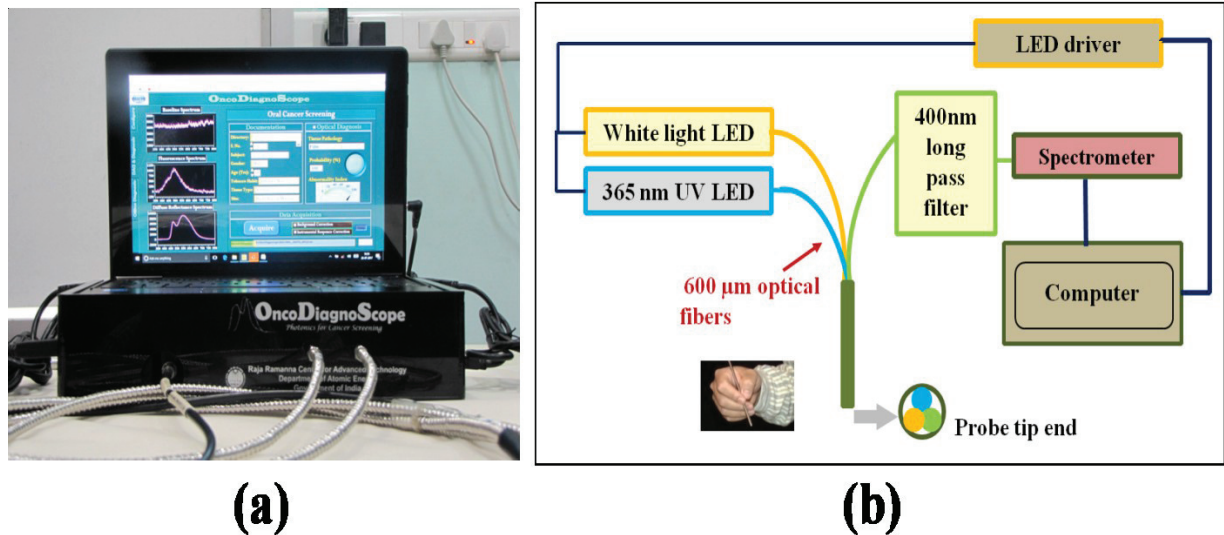


Fig. 7.1: OncoDiagnoScope (a) developed system, and (b) schematic.

Light delivery to and collection from the target tissue is achieved with the tri-furcated fiber-optic probe. The three legs of the probe, each comprising a fused silica fiber (0.22 NA) of core diameter of 600μm, merge to form a common fiber bundle enclosed in an SS tube of 5 mm outer diameter and 80 mm length. Two of the fibers sequentially deliver (fluorescence and diffuse reflectance) excitation light to the tissue surface and the third fiber collects fluorescence or diffuse reflectance from the surface area directly illuminated by the excitation

light. While the white LED light is directly coupled to the SMA connected end of one of the excitation fibers, the light from the 365 nm LED is spectrally cleaned by passing through a narrow band-pass filter before coupling to the SMA connected end of the other excitation fiber. A custom-made miniaturized 400 nm long-pass filter placed at the tip of the SMA connected end of the collection fiber blocks the back-scattered 365 nm light from entering the spectrometer. The overall spectral resolution of the system is ~ 5 nm. The spectra can be recorded in the wavelength range of 350-800 nm.

7.2.2 Device Software

In order to control the hardware of the device for automated acquisition of spectral data and also for subsequent analyses of the acquired spectra to get real-time diagnostic feedback about the interrogated tissue, the important next step is to have an appropriate software enabling interfacing with the device. A graphic user interface (GUI) software was developed in-house and installed in the tablet computer for that purpose. The software has a modular structure with three major applications modules intended to serve three specific domains of functionality. The modules can be accessed through three different tabs positioned one below the other on the left side of the startup (or home) page which is displayed on starting the software (Figs. 7.2-7.4).

The first module, which can be accessed by clicking on the tab titled “Configuration”, consists of a set of user interface controls which recognize the hardware components (LED driver and spectrometer) of the device and is intended for configuring these components. Using this module, which is menu driven, one can set and adjust parameters related to data acquisition (like integration time, number of spectra to be averaged and the window of the boxcar smoothing etc.), and can also view continuous streaming of fluorescence of diffuse reflectance spectra from a test sample quite often required for checking the optimum light

coupling between fiber and LEDs, and optimization of the integration time for achieving the desired signal strength. The module can also be used to re-calibrate the wavelength axis of the spectrometer, if required, by editing the unique wavelength calibration coefficients stored onto an EEPROM memory chip (which the GUI has access to read) of the given spectrometer. The un-checking of the ‘Detect System’ (in Fig. 7.2) allows the running of the software without an error even if the spectroscopy system is not electronically connected to the laptop. This also helps the user for offline review of the recorded data using the same software even without connecting the system. Any error arises in the normal operation of the system can be viewed in this module.

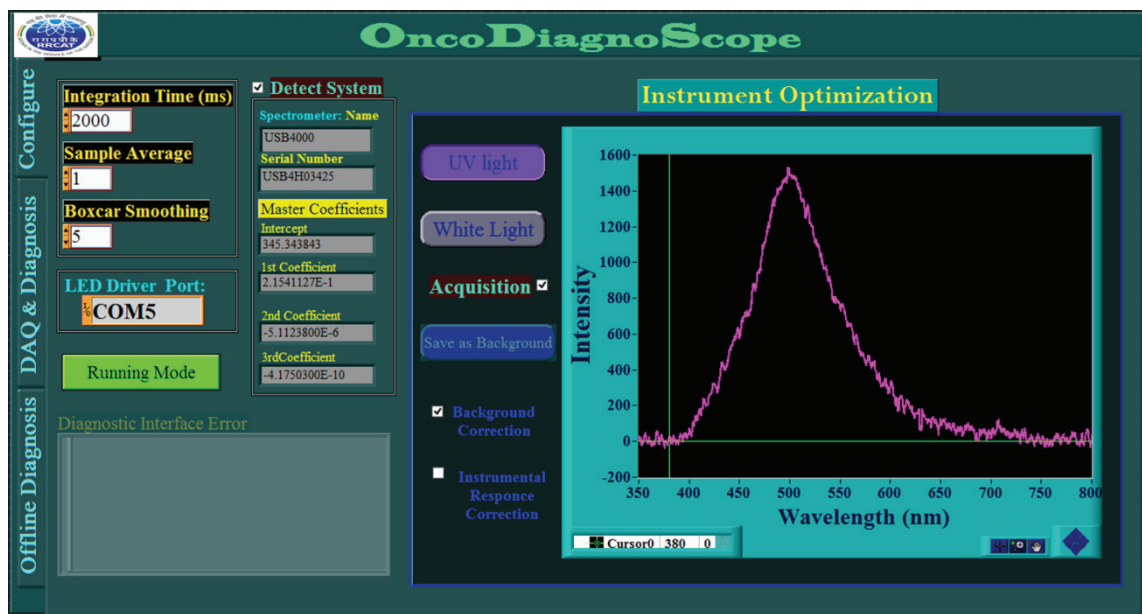


Fig. 7.2 The graphic user interface of the OncoDiagnoScope indented to be used for configuration of the system.

The second module (Fig. 7.3) that can be accessed through the tab titled “Data acquisition (DAQ) & Diagnosis” is the major module of the GUI intended to be operated by the users. The page on the computer screen which opens with this module is the home page that shows up every time one starts the GUI. On the basis of the various tasks to be executed within its

scope, the module has three functional segments: data acquisition, documentation and diagnosis with separate tab designated for each of the tasks. Each segment executes a set of processes as a batch to complete the intended task.

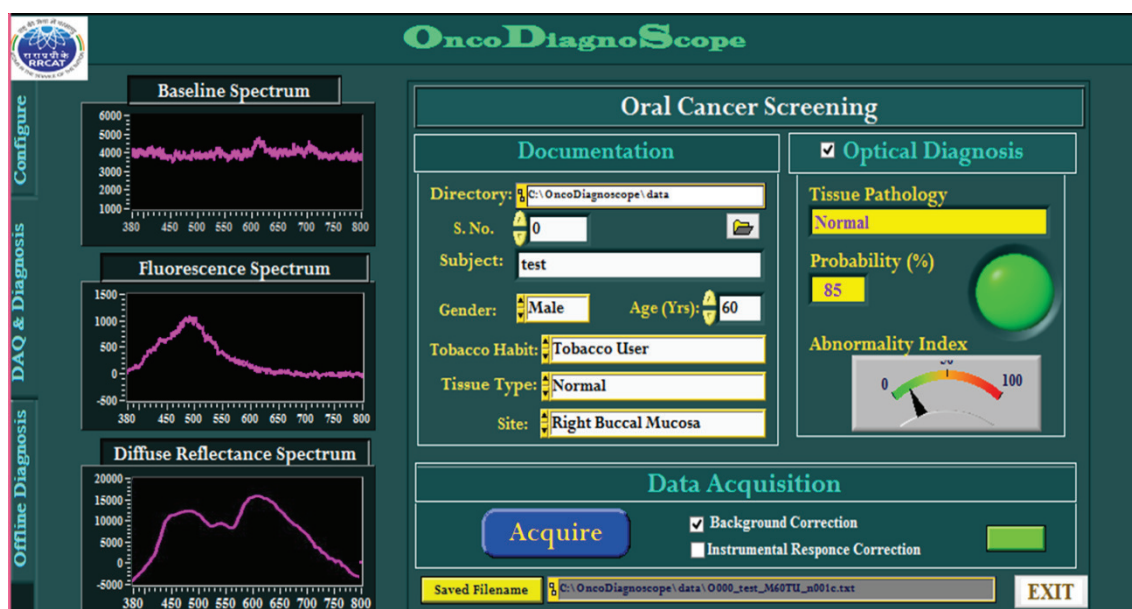


Fig. 7.3 The graphic user interface of the OncoDiagnoScope system for the real-time display of tissue optical spectra, documentation and diagnosis. The total time for data acquisition and analysis was under 10 s.

The data acquisition segment allows automated, sequential acquisition of fluorescence and diffuse-reflectance spectra by controlling the spectrometer, its synchronization with the electrical signal generated by the LED driver card to power-up the LEDs and switching of the respective excitation LEDs. It is designed such that once the “acquisition” command is given, it sequentially performs three tasks: (1) record baseline spectrum (background measurement), with both the excitation LEDs off, (2) measure fluorescence spectrum with 365 nm LED on, and finally, (3) measure diffuse reflectance spectrum with white LED on. It also allows making corrections for the instrumental response profile, if the calibration file is loaded. Once, the acquisition of spectral data is completed, the software prompts a message on the

computer screen as to whether save or discard the recorded data. If all the steps (like touching the probe with the tissue surface, blocking the exposure of overhead or ambient light etc.) are properly followed during the course of spectral measurement, and the lineshape and intensity of the measured spectra look similar to what is expected from an oral tissue, the user can go ahead pressing the “save” tab following which the data get automatically saved in the predefined directory. Otherwise, the user can decide not to save the data. There is also a quality filter (details given in the Section-7.3) integrated in the software for dealing with any improper measurement. If any user, by mistake, saves a data which do not look proper and hence should not have been saved, the quality filter declares it as an outlier.

Before proceeding to acquire spectral data, one necessary step is to register various information related to the details of the patient to be examined, and also the anatomical location and the histological identity of the investigated sites as determined by the examining physician based on his visual assessment. Following documentation of these details, the spectra are acquired using the command “acquisition” as described in the previous paragraph. The acquired spectral data are saved in a predefined directory using regular text editors such as notepad with all the documented information appropriately coded in the file name. For example, the coded file name contains information of the patient age and gender, tobacco consumption history, anatomical location of the spectral measurement site as well as the report on doctor's visual impression or histopathological finding of the investigated tissue sites. For entering information on the type of the suspected lesion and its anatomical location in the oral cavity one can use the tabs ‘*Tissue type*’ and ‘*Measurement Site*’ (Fig. 7.3) respectively, which have drop-down lists. For example, the list under “*Tissue type*’ includes normal, squamous cell carcinoma, sub-mucous fibrosis, leukoplakia, ethroplakia, melenoplakia, verucous cell carcinoma and etc. Similarly, the different measurement sites under ‘*Measurement Site*’ include buccal mucosa (BM), dorsal of tongue (DT), lateral

tongue (RT), ventral tongue (VT), floor of mouth (FM), upper lip (UL), lower lip (LL), vermilion border of lip (VB), hard palate (HP), soft palate (SP), gingivo buccal sulcus (GBS), alveolus (ALV), retromolar trigone (RMT) etc. The coded file name starts with the entry made in the ‘Subject Identification’ column (Fig. 7.3), which is generally the pseudo identification name of the patient to hide the patient identity. Immediately following this is an underscore “_” after which three descriptors indicating the gender (M for Male, F for female, and O for other), age (numeric value), and tobacco habit of the patient (TU for Tobacco user, NTU for Tobacco Non-user, and O for other) appear consecutively. This is followed by an alphabetic character that denotes the tissue type (like the letter ‘c’ denotes OSCC) separated by another underscore “_”. The alphabetic character is followed by numeric characters denoting a particular anatomy. For example, a filename, ‘test_M60TU_c01.txt’ indicates a patient with pseudo identification name ‘test’, gender, ‘male’, age ‘60 years’, known to consume tobacco and measurement taken from his right buccal mucosa (numeric code 01 corresponds to BM) with the doctor’s visual impression of the interrogated tissue site as OSCC. If multiple spectra are taken from the same site and the anatomical location, the file name is further suffixed by the alphabetic characters a,b,c,...etc.

The last segment, titled “diagnosis” is designed for analyses of the acquired spectral data (fluorescence and diffuse-reflectance) and prediction of diagnosis. It delivers online probabilistic diagnosis of the tissue site based on an instant analysis of the spectra being acquired thereof. It essentially incorporates a probability-based, diagnostic algorithm (the details of which will be described in the later section of the present chapter) developed using the framework of sparse multinomial logistic regression (SMLR) for classification of tissue types. This segment also displays the outcome of diagnosis like the predicted pathologic state of the interrogated tissue site as well as the probability of prediction as yielded by the diagnostic algorithm following analyses of spectral data. Depending on the probabilistic

output, red, green or orange flash is displayed with red and green implying confirmed (probability > cutoff decided by the physician) abnormal and normal respectively, while the orange indicating doubtful. The doubtful cases are the ones for which the posterior probability is less than the cut-off decided by the examining doctor and may, probably, require re-assessment. For the measured spectra which do not, by any standard, match any of the spectra (either in terms of spectral profile or intensity) included in the database, an outlier message is displayed. However, a necessary pre-requisite for the exercise of prediction of tissue pathology by the algorithm for a set of measured spectra is that the algorithm gets appropriately trained on a database of spectra (of statistically significant size) previously measured from oral tissue sites of known histopathology. The creation of spectral data base and the task of development and optimization of the diagnostic algorithm are covered in the later sections (7.2.4 and 7.2.5) of the present chapter.

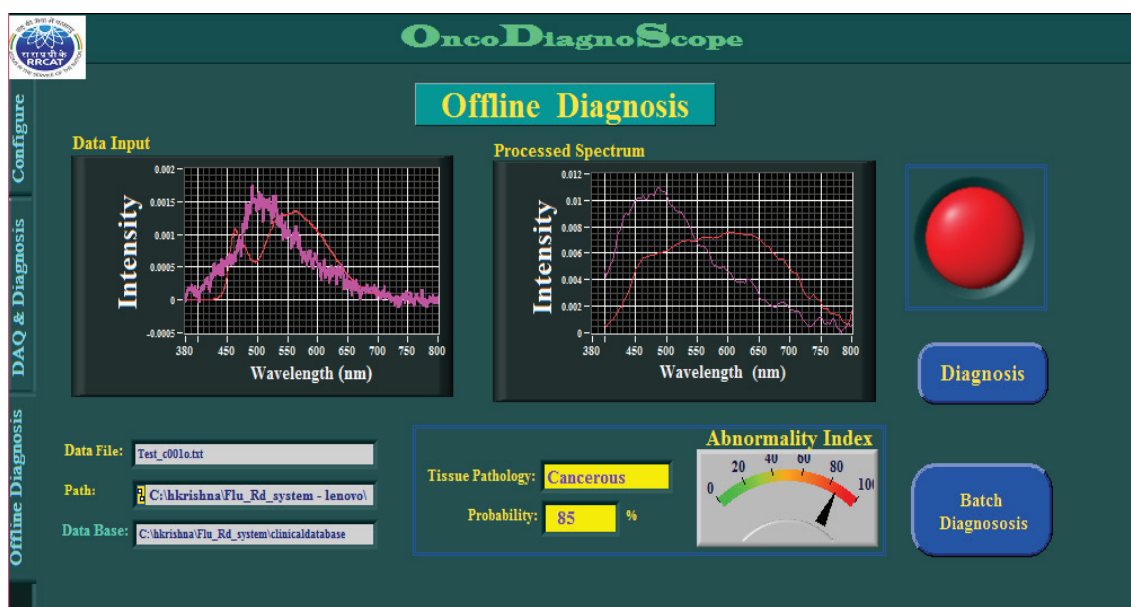


Fig. 7.4 The graphic user interface of the OncoDiagnoScope indented to be used for the offline review and analysis of the saved data.

The module accessed through the tab “Offline diagnosis” (Fig. 7.4) allows offline reviewing of the measured spectral data (like viewing of the measured raw spectra and processed spectra following spectral response correction) and also having information about their histological identity as determined by the diagnostic algorithm. It also enables batch processing of all the spectral data present in a folder to generate an excel file containing the results of optical spectroscopic diagnosis along with the visual/ clinical diagnosis by the examining doctor. This excel file lists the subject identity, observation site, clinical observation, spectral finding with the probability, and abnormality in different column for each spectral data file present in the processed folder.

Overall, the program structure for each of the modules in the GUI software is designed in such a way that once all the parameters of the acquisition are fixed (using “Configure” by developers), the targeted users can easily operate it in online (using the module “Data acquisition (DAQ) & Diagnosis”) or offline (using “Offline diagnosis” module) mode without concerning much about the parameters involved with the hardware of the device. The pages (corresponding to different modules) of the GUI has a dark colored background to minimize the light interference from the peripheral tablet computer with the spectral signal.

7.2.3 Methods of Data Analysis for Prediction of Diagnosis

In this section, a brief description of the methods employed for analyses of the spectral data measured using the module “Data acquisition (DAQ) & Diagnosis” is presented.

Diagnostic Algorithm:

Analyses of the acquired spectral data and prediction of diagnosis by the GUI software requires a diagnostic algorithm which can classify measured set of fluorescence and diffuse

reflectance spectra into appropriate pathologic categories. A probability-based, diagnostic algorithm using the framework of non-linear sparse multinomial logistic regression (NL-SMLR) [103] was developed for that purpose.

In the NL-SMLR approach [103], the concatenated F and Rd input data are first subjected to non-linear transform utilizing the modified radial basis function (RBF) defined as:

$$K(X_{test}, X_{train}) = e^{-\frac{\|X_{test} - X_{train}\|^2}{d * KP}} \quad \dots\dots\dots 7.1$$

Where, X_{train} and X_{test} are the training and test (validation) input data respectively and contain one spectral measurement per row. The parameter $d * KP (= \sigma)$ is the width of the Gaussian with KP being the kernel parameter that needs to be optimized during the learning process of the classifier and d being the dimension of the measured data. The kernel operation transforms the input test data X_{test} with size $N_{test} \times d$ to its nonlinear projections X'_{test} ($= K(X_{test}, X_{train})$) of size $N_{test} \times N_{train}$. These transformed nonlinear projections are used as input to the SMLR for subsequent classification.

In order for the SMLR algorithm to deliver diagnostic feedback about a measured set of optical spectra from an interrogated tissue site, the algorithm needs to be first trained on a library of previously measured spectra of known tissue pathologies. The training of algorithm involves building a model to separate a set of labeled input data into its constituent classes[101]. To separate the data into its constituent classes, the SMLR algorithm computes the posterior probabilities using a multinomial logistic regression model following Bayes' rule [101]. The training requires estimation of a total L (number of classes) set of sparsity promoting model weights $w = [w^{(1)^T}, \dots, w^{(L)^T}]^T$ based on maximum a posteriori (MAP) framework using the nonlinear RBF kernel projections of the training data to itself [X'_{train}

($= K(X_{train}, X_{train})$) and their true class labels as the input data. Using a Bayesian maximum a posteriori (MAP) framework [101], [103] one can explicitly compute:

$$\hat{w}_{MAP} = \arg \max_w [l(w) + \log p(w)] \quad \dots\dots\dots 7.2$$

where, $p(w)$ is a prior distribution on w , $l(w)$ is the log-likelihood function given by:

$$l(w) = \sum_{j=1}^{N_{train}} \log P(y^{(i)} | x, w) = \sum_{j=1}^{N_{train}} \left\{ \sum_{i=1}^L y_j^{(i)} w^{(i)^T} x_j - \log \left[\sum_{i=1}^L \exp(w^{(i)^T} x_j) \right] \right\} \quad \dots\dots\dots 7.3$$

Here $x \in X'_{train}$ is an input data instance of interest from training data set with class label $y = [y^{(1)}, y^{(2)}, \dots, y^{(L)}]^T$ such that $y^{(k)} = 1$, if x corresponds to an example belonging to the class k ($k \in \{1, \dots, L\}$) and $y^{(k)} = 0$ otherwise.

If many of the training set data points are irrelevant in making the decision boundaries for the class prediction, the vector w is said to be sparse and many of its entries are exactly zero. To promote sparsity in w , a Laplacian prior on w is used which means that $p(w) \propto \exp(-\lambda \|w\|_1)$, where $\|w\|_1 = \sum_l |w_l|$ denotes the l_1 -norm and λ is a tunable regularization parameter that needs to be optimized during the learning phase of the algorithm. For computing \hat{w}_{MAP} , the algorithm exploits a bound optimization framework to perform sequential iterative optimization of a concave lower bound which is described in details elsewhere [103].

Once the set of weights is calculated, then the probability that a given sample x belongs to class i can be estimated as:

$$P(y^{(i)} | x, w) = \frac{\exp(w^{(i)^T} x)}{\sum_{i=1}^L \exp(w^{(i)^T} x)} \quad \dots\dots\dots 7.4$$

Where, x is the any row vector of the nonlinear projections $X'_{test} (= K(X_{test}, X_{train}))$ of the input test data point.

Quality Filter:

A quality filter developed and integrated in the GUI software performs outlier detection during data acquisition in an automated way. The outlier detection scheme is based on the principle of principal component analysis (PCA) coupled with the framework of Hotelling's T^2 and Q-residuals [40], [128]. The Hotelling's T^2 measures the variation of the observed data within the PCA model and Q-residuals measure the mismatch between the PCA model and the observed data. In brief, the PCA reduces the dimension of the spectral data by decomposing them into set of linearly uncorrelated variables called principal components (PCs), such that the spectral variations of the data set in transformed space are maximized. PCA model of the data matrix X can be defined as:

$$X = TP^T + E \quad \dots\dots\dots 7.5$$

where T , P and E represent PC scores (data projection in the reduced space) and loading vectors (new orthogonal axis in reduced space), and the residuals respectively. Accordingly, Hotelling's T^2 is the measure of variance captured by the PCA model (sample to model distance) [40], [128] and given as:

$$T_{ik}^2 = t_{ik}(\lambda_k^{-1})t_{ik}^T \quad \dots\dots\dots 7.6$$

where, λ_k^{-1} is the diagonal matrix of the inverse of the k largest normalized eigenvalues λ_i of the co-variance matrix of data X arranged in a descending order, and t_{ik} are the PC scores for i^{th} sample spectrum using k PC components. T^2 gives an indication of extreme values of

variance within the PCA model and statistical thresholds for T^2 can be calculated using the F-distribution as follows:

$$T_{\alpha}^2 = \frac{k(m-1)}{m-k} F_{\alpha}(k, m-k) \quad \dots\dots\dots 7.7$$

where, T_{α}^2 is the threshold value with a significance level of confidence, α ; m is the number of samples used to build PCA model, $F_{\alpha}(k, m-k)$ is the upper 100 α % critical point of the F-distribution with k and $(m - k)$ degrees of freedom.

On the other hand, the Q statistic gives the measurement of variance, which is not captured by the PCA model [40], [128]. It gives an indication of how well the newly observed data conforms to the PCA model. It is defined by:

$$Q_{ik} = \sum (x_i - t_{ik} P_k^T)^2 \quad \dots\dots\dots 7.8$$

Where, x_i is the sample spectrum, Q_{ik} is the sum of squared reconstruction error for i^{th} sample spectrum using k PC scores and loading. The statistical thresholds for Q statistic can be calculated as follows [40], [128]:

$$Q_{\alpha} = \theta_1 \left(\frac{h_0 c_{\alpha} \sqrt{2\theta_2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right)^{\frac{1}{h_0}} \quad \dots\dots\dots 7.9$$

Where, $\theta_1 = \sum_{i=k+1}^n \lambda_i$, $\theta_2 = \sum_{i=k+1}^n \lambda_i^2$, $\theta_3 = \sum_{i=k+1}^n \lambda_i^3$, $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$ and c_{α} is the normal deviate corresponding to the $(1-\alpha)$ percentile.

The Hotelling's T^2 and Q-residuals are the two independent parameters providing quantitative information about the model fit. For both the statistical measures, a normalized 95% confidence interval was used to determine the thresholds which served as the eligibility criteria for deciding whether the measured data qualify as a valid spectral data to be used as input to the diagnostic algorithm.

7.2.4 Preparing the Device as a Stand-Alone Tool for Real-Time Diagnosis in a Clinical Setting

In order to make the device ready for real time prediction of the pathology of the interrogated tissue site based on its measured fluorescence and diffuse reflectance spectral signatures, the device was installed at Raman Foundation Cancer Hospital and Research Centre, Indore to carry out detailed *in-vivo* clinical studies on healthy volunteers and patients with various oral lesions. The objective was two-fold: first, to generate the data-base of spectral data from healthy volunteers with clinically normal oral cavity and patients with clinically confirmed or pathologically certified oral lesions, and second, to evaluate the efficacy of the NL-SMLR algorithm in making accurate predictions of the pathology sites of the interrogated tissues based on their measured spectral data with histology or visual examination (following the guidelines of the Ethical Committee) as the gold standard of diagnosis.

Study Design

The study was conducted with the approval of Institutional Ethical Committee. All patients participating in this study pre-operatively signed an informed consent, permitting the investigative *in-vivo* fluorescence and diffuse reflectance spectroscopic acquisition of oral tissues. The final eligibility of each patient was determined by the participating head and neck surgeon based on the medical condition of the patient such that patient care was not

compromised. Patients having gone through any prior treatment like surgery, chemotherapy or radiotherapy for earlier cancers or with recurrences were excluded from the study. The spectral measurements were performed by the participating head and neck surgeon using a protocol as described in chapter 3 which was maintained for all individuals in this study.

A total of 1421 tissue sites from 229 individuals were interrogated to record the spectral data. The spectral measurements included 413 sites from 129 patients with OSCC lesion, 138 sites from 16 patients with OSMF lesion, 87 sites from 27 patients with OLK lesion, and a total 783 sites from 58 healthy volunteers with no disease of the oral cavity. Biopsies were taken subsequent to acquisition of spectra from the oral cavity sites suspected of being malignant. However, as per the terms of the approval from the Ethical Committee of the hospital, no biopsies were available from the investigated sites for the patients with oral submucous fibrosis (OSMF) and the diagnosis of this condition was based on clinical findings only. Similarly, no biopsies were allowed from the tissue sites of healthy volunteers. 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.

Pre-processing of Spectral Data

The *in-vivo* autofluorescence (F) and diffuse reflectance (Rd) spectra were recorded in the 400-700 nm spectral range. The background spectrum acquired with the both the light sources off was subtracted from all subsequently acquired F and Rd spectra from that measurement. Spectra were further binned along the wavelength axis in 2 nm intervals and filtered with a first-order, eleven point Savitzky-Golay smoothing filter for noise removal. The fluorescence spectra were corrected for the system spectral response by using a NIST traceable calibration lamp (LS-1, Ocean optics, Inc., Dunedin, FL) and the Rd spectra were corrected for the spectral profile of source (white LED). The area-normalized fluorescence

and diffuse reflectance spectra of each tissue site were concatenated to form a single normalized optical spectrum per site and used as input to the diagnostic algorithm for further analysis.

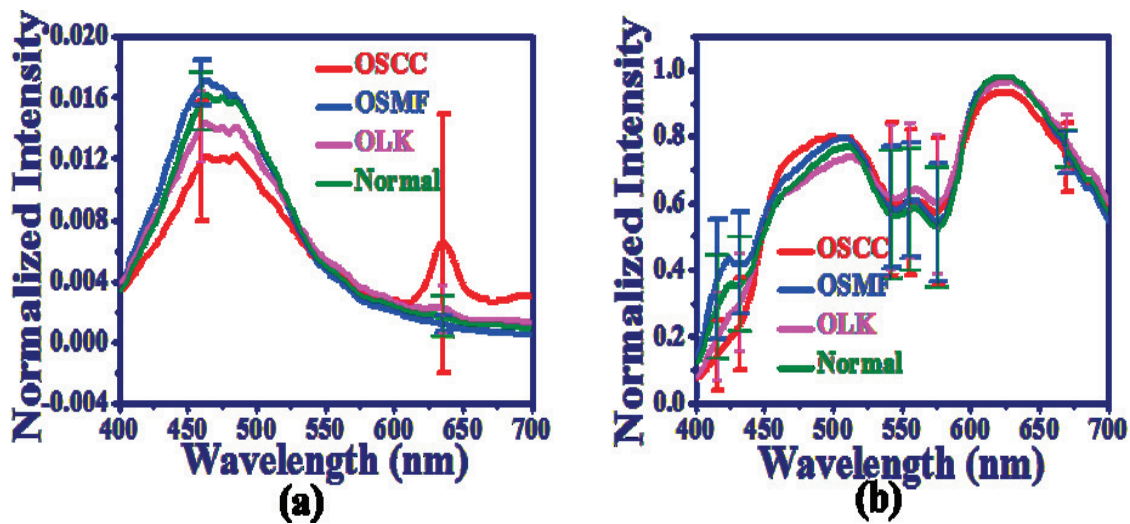


Fig. 7.5 Mean, normalized (a) fluorescence and (b) diffuse reflectance spectra of OSCC, OSMF, OLK, and normal oral tissue sites. The error bars represent ± 1 standard deviation at that wavelength.

The mean spectra (Fig. 7.5) showed a variety of differences between the different tissue types consistent with the earlier spectral observations associated with the diseased transformation [26], [27], [32]–[34], [53]. For the fluorescence spectra (Fig. 7.5a), the most prominent differences were seen in the 460 nm and 635 nm spectral bands. While the 460 nm spectral band was the most intense in healthy squamous tissues, the intensity of the 635 nm band was the highest in the spectra from malignant lesions. The diffuse reflectance spectra (Fig. 7.5b), showed dips around 400–450, 540–550, and 570–580 nm characteristics of blood absorption and the extent of these dips were observed to be different implying significant variations in the concentration of hemoglobin in the different tissue types.

Evaluation of the Performance of the Diagnostic Algorithm

The entire set of the pre-processed spectral data corresponding to the healthy volunteers (a total 783 spectrum) and the patients with OSCC lesions (a total 413 spectrum) were randomly split into two sets: training set and validation set such that validation set has 100 spectra per tissue pathology. The purpose of the training set was to optimize the parameters (KP and λ) of the algorithm and the purpose of validation set was to prospectively test its accuracy in an unbiased manner. For a given classification task, NL-SMLR algorithm computed the posterior probabilities of the different interrogated tissue sites of belonging to various tissue classes for both the training and validation set data. Subsequently, the class was assigned to a given spectral data whose posterior probability of belonging to that particular class was the highest. The spectra of training data set were first used to build supervised learning models with varying values of the algorithm parameters and the performance of the models were estimated in classifying the training set data. The model which led to minimum classification errors in training data set is considered to be optimum model. This model is then tested on the validation data set to obtain the unbiased estimated of the diagnostic performance of the system. The predictive accuracy of the diagnostic algorithm was calculated with respect to histopathology as the gold standard of reference.

The random assignment of spectra into the two sets (training and validation) was repeated ten times such that each time the composition of the training and the validation sets was unique and different from each other. The performance of diagnostic algorithm was optimized using each of the ten training data sets formed by random splitting of spectra, and validated each time on the corresponding independent validation data set. The average sensitivity and specificity provided by the algorithm for the validation set in discriminating OSCC lesion from normal oral tissue sites were found to be ~89% [95% confidence interval

(CI): 87% to 90%] and 98% [95% CI: 97% to 99%] respectively, with the overall accuracy of discrimination being ~93% [95% CI: 93% to 94%].

7.2.5 Point-of-Care Device as a Stand-Alone Tool for Real-Time Diagnosis in a Clinical Setting

In order to convert the spectroscopic device into a point-of-care device capable of providing online diagnostic feedback on the interrogated tissue sites of the oral cavity, it is required to integrate the GUI software with a proper database along with the parameters of the diagnostic algorithm optimally trained on this database. To begin with, the device was readied as a stand-alone device for real-time diagnosis of the cancerous lesions of oral cavity. In order for that, the various parameters of the NL-SMLR algorithm were optimized using the set of spectral data recorded earlier from ~129 patients with OSCC lesions and ~58 healthy volunteers with normal oral mucosa. The receiver operating characteristic (ROC) analysis of the classification results was performed on each of the randomly split training and validation data sets in order to select the set of training data yielding best discrimination between OSCC and normal oral tissues. The set of training data that led to the maximum area under the ROC curve was considered as the optimal set of training data. This set of training data (corresponding to best classification) and the SMLR model weights (w) corresponding to the optimal parameters (KP and λ) obtained following training of the algorithm on this set of training data are the only requisites for carrying out (RBF-) kernel transformation and probability calculation for any test data [34], [117], [121], [129].

Apart from selecting the optimum parameters for accurate diagnosis, it is also important to determine whether the measured spectral data qualify as a valid test data. For doing that, the GUI software uses the quality filter (described in the Section-7.2.3), developed using the principles of principal component analysis combined with Hotelling's T^2 and Q-

residuals statistics [40], [128], that performs automatic outlier detection during data acquisition. The Hotelling's T^2 measures the variation of the observed data within the PCA model and Q-residuals measure the mismatch between the PCA model and the measured data. The values of these statistical measures were calculated using the selected training dataset yielding best classification (of the validation set data). For both the statistical measures, a normalized 95% confidence interval was used to determine the thresholds which served as the eligibility criteria for deciding whether the measured data qualify as a valid spectral data to be used as input to the diagnostic algorithm.

In fact, for being able to deliver accurate diagnostic predictions for the data measured by the device, the GUI software needs a number of things to be integrated therein. These include (i) PCA loading vectors and threshold values of the T^2 and Q-residuals statistics corresponding to the optimum training data to be used for the outlier detection, and (ii) an optimum set of training data to be used for the kernel transformation of the input spectral data, and (iii) corresponding optimum SMLR weights w to be used for probability calculation. Thus, once a test measurement is performed, the T^2 and Q statistic for the measurement is calculated using the PCA loading vectors and compared with the stored threshold values for these measures. If the calculated values of both the statistical measures were found to be less than the threshold values, the measured spectra are considered to be qualified as a valid spectral data. Subsequent to checking the eligibility as the valid spectral data, the quality-passed spectra were transformed to its projection in the kernel space and these projected data were immediately fed to the diagnostic algorithm for on-line prediction of tissue pathology. The algorithm calculated the posterior probability of belonging to each class for a given data. The class having the maximum probability of belonging is assigned to that data. Depending on the probabilistic output, red, green or orange flash is displayed with red and green implying confirmed (probability > cutoff decided by the physician) cancerous

lesion and normal tissue respectively, while the orange indicating doubtful. The whole process of spectral acquisition and online data processing requires less than ten seconds to get the diagnostic result.

The performance of the point-of-care device as a stand-alone system was further validated on previously unseen spectra of ~ 60 individuals. The system was found to provide an overall classification accuracy of $\sim 83\%$ (with a sensitivity of 72% and specificity of 89%) for discriminating cancerous lesion from healthy oral tissue sites in those individuals.

It is pertinent to mention here that due to limited size of the spectral data for the potentially malignant lesions (OSMF and OLK put together), the ability of device has been demonstrated only for discriminating cancerous lesions from healthy oral tissues. However, once the acquisition of *in-vivo* spectra from more number of patients with various potentially malignant lesions of oral cavity is done, the device will be further optimized with augmented spectral data of potentially malignant oral lesions.

7.3 Fluorescence Imaging based Point-of-Care Device for Improved Visual Assessment of Oral Cavity Lesions

Patients diagnosed with oral cancer at an early stage have a substantially greater chance of successful treatment and less treatment-associated morbidity than those diagnosed at a late stage. Currently, the most widely used screening test for oral cancer is visual inspection of the oral cavity under white light illumination. However, often it turns out to be difficult to satisfactorily detect changes in oral mucosa, associated with early cancers or pre-cancerous alterations that generally precede invasive cancers, using conventional oral examination under white light illumination. In recent years, fluorescence imaging has been suggested to be an effective alternative method for precisely locating oral lesions [21]. The auto-fluorescence from tissue is attributed to many endogenous fluorophores present in the tissue [22], [27],

[53], and the transformation from normal to neoplastic lesions is accompanied by changes in the concentration or optical properties of these fluorophores [53], [56] thereby making the native tissue fluorescence as a sensitive marker for monitoring neoplastic transformation [53], [56]. The U.S. Food and Drug Administration has already approved the use of autofluorescence imaging device for early detection of oral neoplasia, which is commercially marketed as the VELscope® (LED Dental, Inc., White Rock, BC, Canada). The device uses a blue or violet light to illuminate oral tissue and long pass and notch filters to enable direct visualization of auto-fluorescence from the oral cavity [25], [57]. However, this tool has a limited utility in developing countries like India due to prohibitive cost of the equipment and its maintenance apart from the cost of trained technicians and laboratory infrastructure. With the advent of intense UV LEDs and improvement in the camera technology, now it is possible to make a low cost fluorescence imaging devices and explore the possibility of using UV light induced fluorescence imaging for identification of oral lesions.

This section describes the development and use of a low-cost, handy fluorescence imaging tool for real-time, non-contact and *in-situ* imaging of fluorescence from human oral cavity intended for improved visual assessment of the oral cavity. Using this instrument, regions of oral lesions can be better identified against the healthy oral tissues based on their natural characteristics in response to light. A Graphic User Interface (GUI) software developed and integrated with this imaging tool enables automated acquisition and processing of tissue fluorescence images for highlighting the difference in the fluorescence characteristics of the different oral tissue types.

7.3.1 Device Hardware and Software

The device (Fig. 7.6a) uses a circular array of four UV LEDs emitting light at 365 nm for inducing fluorescence in native tissues. The light from these LEDs is shone onto the tissue

surface of the oral cavity to be examined and the fluorescence emitted from the oral cavity, different for normal and abnormal oral tissues, is detected by a CCD camera to generate two dimensional fluorescence spectral images. A miniaturized long-pass filter with cut off wavelength of 400 nm is fitted at the tip of the camera to avoid the excitation light. All the components of the system, including the LEDs, the camera and the optical filter are accommodated in a custom made cylindrical Perspex-case of diameter 40 mm and length 120 mm (designed and fabricated in-house). The field of view of the imaging system is ~ 3 cm in diameter at a distance ~ 5 cm from the distal end of the system.



Fig. 7.6: (a) Vision Enhanced Module (VEM) and (b) Graphic User Interface (GUI) software developed for VEM.

A modular GUI software was developed and integrated with this imaging device through a laptop computer for automated acquisition and processing of tissue fluorescence images (Fig. 7.6b). The camera as well as the LEDs are powered up by the USB ports of this laptop computer. The software allows displaying fluorescence images of tissue online in the video mode as well as taking a snapshot of the displayed image at any instant in the still

mode on the computer screen. It also has the ability to process the grabbed images offline using different color composites (RGB component) for highlighting the difference in fluorescence characteristics of the different tissue types. The GUI software was also integrated with a pixel connectivity based algorithm [130] developed to demarcate the boundaries of spectrally different region present in the acquired image. The connectivity algorithm, when clicked on any pixel of an image, directly determines whether an adjacent pixel belongs to the same object or a different object and grows the object boundaries until the maximum allowed deviation between intensity of adjacent pixels is violated. Further, the original as well as the processed images can be saved in a single multipage file having '.tiff' image format for offline viewing.

7.3.2 Clinical Validation of the Device

Study Design

In order to evaluate the performance of the device, the device was installed at Raman Foundation Cancer Hospital and Research Centre, Indore to carry out *in-vivo* clinical studies on healthy volunteers and patients with various oral lesions. With the approval of Institutional Ethical Committee a study was conducted at the hospital as well as several cancer screening camps in and around Indore to record the fluorescence images from healthy volunteers and patients with various lesions of the oral cavity. The protocol for the study is already detailed in the Section-3.2 as well as in the Section-7.2.4 of the present chapter. All patients participating in this study signed an informed consent, permitting the investigative fluorescence image acquisition. Prior to recording the fluorescence image, each individual was asked to wear the UV protection goggles (LG4, Thorlabs) to protect their eyes from any accidental exposure to excitation light. For recording the fluorescence images, the individuals were asked to gently open the mouth ensuring that none of the individuals complained about

pain because of mouth opening. The overhead room lights in the OPD room were turned off temporarily during image acquisition to minimize the contribution of the ambient light in the acquired images.

A total of 576 (64 normal and 512 abnormal) *in-vivo* fluorescence images were acquired from the oral cavity of ~196 patients having different types of lesions and ~22 healthy volunteers with no history of the disease of the oral cavity. The different oral lesions include OSCC, OSMF, and OLK. Few of the patients have multiple types of the lesions. Biopsies were taken subsequent to acquisition of images from the oral cavity sites suspected of being malignant. However, as per the terms of the approval from the Ethical Committee of the hospital, no biopsies were available from the investigated sites for the patients with oral sub-mucous fibrosis, and the diagnosis of this condition was based on clinical findings only.

Evaluation of Performance

The recorded fluorescence images showed significant differences between the different lesion types. For example, the Fig. 7.7 shows a few representative images recorded from various tissue pathologies. The images acquired from the oral cavity of healthy volunteers (Fig. 7.7A), in general showed uniform green fluorescence (19 out of 22 individuals) with the exception of a few (3 out of 22 individuals) having uniform red fluorescence (Fig. 7.7B), especially when the images were recorded from the region of the dorsal tongue. This increase in red fluorescence intensity was a frequently observed characteristic found in the images recorded from OSCC lesions. However, the difference is that in this case (i.e. for the OSCC lesions) the red fluorescence was found to be scattered across the suspected region irrespective of its anatomical location. While the fluorescence images of 46% (66 out of 142) patients with OSCC lesions showed this large increase in red fluorescence (Fig. 7.7C). Fluorescence images from rest of the patients with OSCC lesions were found to have an

overall decrease in fluorescent intensity (Fig. 7.7D) as compared to the normal individuals. The leukoplakik lesions, similar to the OSCC ones, also showed reduction in the overall fluorescence (Fig. 7.7E). However, here the reduction in the blue and the red fluorescence was found to be larger as compared to that in the green fluorescence, making these lesions appear as feeble green spots in the images (18 out of 33 patients). All the sub-mucosal fibrosis lesions showed enhanced blue fluorescence (Fig. 7.7F) as compared to the contralateral normal tissue.

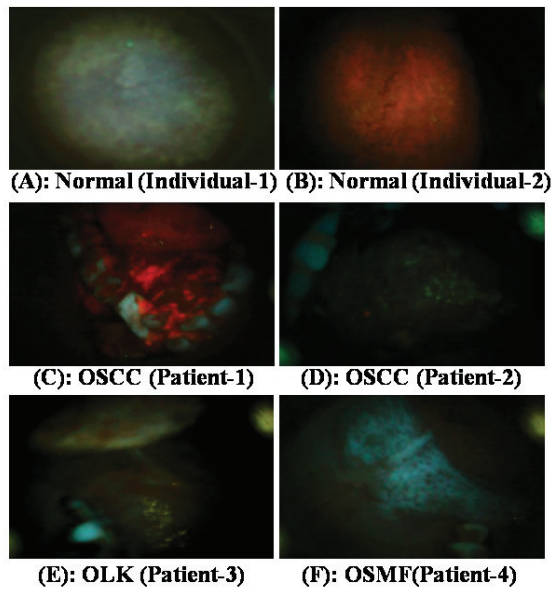


Fig: 7.7: Representative fluorescence images of different oral tissue pathologies as acquired using Vision Enhanced Module (VEM).

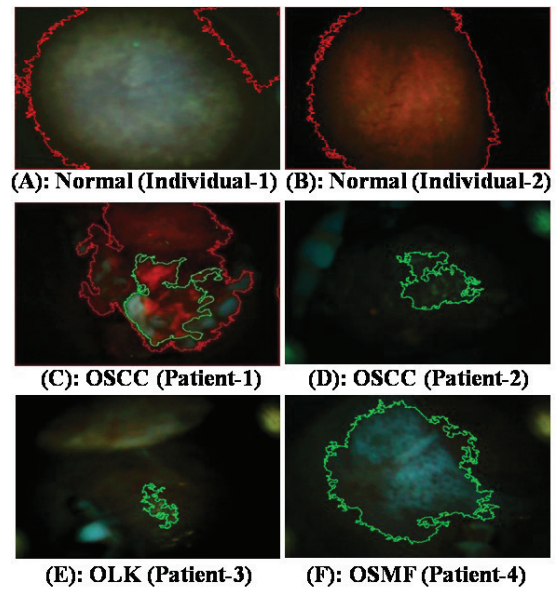


Fig: 7.8: Demarcated area of the suspected tissue sites present in the corresponding fluorescence images of the Figs. 7.7.

The images were further analyzed using the pixel connectivity algorithm [130] to demarcate the boundaries of the region having similar spectral characteristics. Figs 7.8A-F shows the demarcated area of the suspected tissue sites present in the corresponding fluorescence images of the figs. 7.7A-F.

7.4 Summary

In summary, the chapter presents development of two point-of-care devices for screening of human oral cavity cancer. While one of the devices is a tablet computer based, compact and portable instrument which works in point-spectroscopy mode (by measuring fluorescence and diffuse reflectance sequentially from various sites on the surface of a target tissue, following illumination of these sites with light of appropriate wavelengths), the other one is a hand-held tool that works in imaging mode (by detecting fluorescence emitted from a comparatively larger area of a tissue surface following its illumination). However, both the devices are meant to achieve a common goal that is to find whether any potential lesion is present in the oral cavity of a subject being examined. Thus, simultaneous use of the two devices might lead to better management of oral cancer in a real-time diagnostic setting. This is because while the hand-held imaging tool will allow better visual identification of regions of oral lesions (often time missed by visual examination under white light illumination) against the healthy oral tissues, the point-of-care spectroscopy tool can be used to quantitatively ascertain the pathology of these oral lesions.

Chapter 8

Summary and Future Perspectives

To summarize, we have presented in the thesis, the findings of our investigation on the evaluation of clinical applicability of optical spectroscopy and imaging for non-invasive diagnosis of human oral cavity neoplasia. Towards this objective, we first developed a portable Raman system capable of recording *in-situ* Raman signal from tissue in clinically acceptable data acquisition time. An important requirement for the use of Raman spectroscopy for tissue diagnostic applications is an appropriate algorithm that can faithfully retrieve weak tissue Raman signals from the broad, orders of magnitude stronger, background believed to be arising from fluorescence and/or the scattering tail of the laser line. The most widely used method for background subtraction is iterative polynomial fitting to mimic the background. However, in this method, the lineshape and intensity of the extracted Raman spectra are dependent on the wavenumber range selected for the fitting. An important outcome of the present work is the development of a background subtraction algorithm which overcomes this drawback and gives range independent Raman spectra. The developed algorithm performs the iterative smoothening of the measured spectrum such 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.

The portable Raman system was installed at Tata Memorial Hospital (TMH), Mumbai and used to carry out *in-vivo* studies with the approval of TMH ethical committee. The *in-vivo* Raman spectra were recorded from oral cavity of ~ 200 individuals. The different tissue sites investigated belonged to either of the four histopathologic categories: 1) squamous cell carcinoma (OSCC), 2) sub-mucosal fibrosis (OSMF), 3) leukoplakia (OLK) and 4) normal squamous tissue. The Raman spectra of the oral tissues were characterized by several characteristic peaks belonging to proteins, DNA, lipids, and bone minerals and were found to have significant difference over different tissue types. In general, the Raman bands assigned to proteins and DNA were found to be higher for OSCC lesions as compared to normal tissues whereas it is otherway for lipid related Raman peaks. Compared with the difference between normal mucosa and malignant (OSCC), normal and potentially malignant (OSMF and OLK) were found to have relatively small differences. All the measured *in-vivo* tissue spectra were analyzed for differential diagnosis of oral tissues using a probability based multi-class diagnostic algorithm. The algorithm uses the theory of nonlinear maximum representation and discrimination feature (NLMRDF) for feature extraction, and the theory of sparse multinomial logistic regression (SMLR) for classification [101]. The best classification accuracy of 85%, 89%, 85% and 82% were obtained in classifying the oral tissue spectra into four different pathology classes- normal, OSCC, OSMF and OLK respectively, on the basis of leave-one-individual-out cross-validation, with an overall accuracy of 86%.

The another interesting finding of the thesis work is that the Raman spectra measured from the healthy volunteers with and without tobacco habits have a number of statistically significant spectral differences between the different Raman bands. Based on these differences, the spectra of two groups were discriminated with an accuracy of 95% by NLMRDF-SMLR based supervised classification algorithm in the leave-one-subject-out

cross validation mode. Further, it was found out that exclusion of the spectral data of the healthy volunteers with the tobacco consumption habits from the reference normal database considerably improves the overall classification accuracy (92 % as against 86%) of the algorithm in separating the oral lesions from the normal oral tissues.

This thesis also addresses the variability of the measured Raman spectra recorded from the different anatomical sites of the oral cavity of healthy volunteers. An unsupervised Fuzzy means clustering analysis of normal tissue Raman spectra recorded from oral cavity of healthy volunteers reveals the presence of four major anatomical groups. Further, we have examined the effect of inter-anatomical variations on the performance of NLMRDF-SMLR based supervised diagnostic algorithm used to classify the whole set of measured tissue Raman spectra into three categories: normal, potentially malignant (OSMF and OLK pooled together) and malignant (OSCC). The findings showed that the diagnostic algorithm, when applied on the pooled set of spectra from all the clusters, correctly discriminated normal, malignant and potentially malignant tissue sites with 86%, 88%, and 86% accuracy respectively, which amounted to an overall accuracy of 87%. However, when the anatomy-matched data sets were considered, the overall classification accuracy was found to improve to 95% with the algorithm correctly discriminating the corresponding tissue sites with 94%, 99%, and 91% accuracy respectively. So in this work we have established that for classification of oral pathologies the anatomical variability is significant and should be considered as an important parameter for development of Raman spectroscopy based diagnostic tools.

In this thesis work we have also carried out a side-by-side comparison of the relative performances of fluorescence and Raman spectroscopy for *in-vivo* discrimination of various oral tissue pathologies in a clinical setting. The results showed that while the binary (normal vs. abnormal) classification accuracy for the fluorescence based system was comparable to

that for Raman, the multi-class (OSCC, OSMF, OLK or normal tissue) classification accuracy using the Raman system was significantly better (86 % as against 77%), because of higher molecular specificity of Raman. Based on these results and with the availability of high brightness LEDs, an LED based, USB powered, portable diagnostic system (named OncoDiagnoScope) capable of providing automated diagnostic feedback has been developed. The system can record both fluorescence and diffuse reflectance spectra sequentially in one go from a given tissue site. Studies involving ~200 subjects with this system provided an average sensitivity and specificity of ~89% and 98% and an overall accuracy of ~93% in independent training and validation mode in discriminating malignant (OSCC) from normal oral tissue sites.

We have also developed a handy fluorescence imaging tool, called Vision Enhancement Module (VEM), for real-time, non-contact and *in-situ* imaging of fluorescence from human oral cavity intended for improved visual assessment of the oral cavity. Using this device, the extent of oral lesions can be better identified against the healthy oral tissues based on their natural characteristics in response to light. The system was validated at hospitals on subjects enrolled for routine examination of their oral cavity and was found to identify the lesions in the oral cavity with improved efficacy.

Future Perspective

While the findings presented in the thesis have demonstrated with evidence that techniques based on optical spectroscopy and imaging show every indication of being a promising tool in the detection of oral cancer in a clinical setting, they have also pointed towards the possibilities of further work to be done in this important field. For example, in Chapter-7 we discussed that for achieving the final goal of routinely using optical spectroscopy for the management of oral cancer in a clinical setting, it is imperative to deliver diagnostic results in

real time. We have made a stand-alone combined fluorescence and diffuse reflectance spectroscopy based device capable of providing non-invasive and accurate diagnostic feedback in real-time as a first step towards that goal. Since early detection of oral cancer is critical in patient care and the *in-vivo* studies carried out using this device thus far included high risk population of patients most of whom were at an advanced stage of cancer, one urgent future need will be to conduct independent validation via multi-centre clinical trials on a population having all kinds of oral lesions. Although the inherent intra-patient as well as inter-patient variability in the spectra is an issue that limits the performance of the diagnostic algorithms and thus implementation of clinical trials, the development of new automated, robust, cutting-edge diagnostic algorithms based on state-of-the-art pattern recognition methods can be developed to address this issue.

In addition to the use of the optical spectroscopic device for *in- vivo* diagnosis of oral cancer, one further possibility could be to employ the same setup for *ex-vivo* diagnosis of oral cancer using body fluids like urine. We have already carried out preliminary studies to evaluate the applicability of the approach on samples of urine collected from a limited number of patients and healthy volunteers [131]. The results of the studies were encouraging and provide enough motivation to pursue further studies to establish the potential of the technique for detecting oral cancer.

Another important point worth considering is that although Raman has the intrinsic ability to detect molecular information in a tissue in greater detail than fluorescence or diffuse reflectance and as a consequence is capable of providing improved multi-class class classification performance (as seen in Chapter-6), given presently used instruments, combined fluorescence and diffuse reflectance is a significantly stronger candidate than Raman for making point-of-care devices. Thus one future challenge would be to develop a low-cost, compact and portable Raman system which can be converted to a stand-alone

cancer diagnostic device. Another possibility could be to combine all three optical spectroscopic techniques in a single instrument which can be used for screening (using the advantage of tissue fluorescence and diffuse reflectance) of the suspected lesions of the oral cavity as well as determining the exact pathological status (using the advantage of tissue Raman) of the lesions screened.

The point spectroscopy based systems developed and used by us for the clinical *in-vivo* studies have provided very good sensitivity and specificity (~90%) in delineating different lesions of the oral cavity. However, they produce single-point diagnostic measurements on the tissue surface and cannot provide the spatial information. Spatial scanning of the probe over the tissue surface is required for obtaining both spatial and spectral information from a larger tissue area. This is not only tedious and time-consuming but also not feasible in a real-time diagnostic setting. In order to obtain both spatial and spectral information from the tissue surface being interrogated it is required to combine spectroscopy with imaging. A compact and portable liquid-crystal tunable filter (LCTF) based spectral imaging system can be developed for that. This system will provide a complete spectroscopic map of every pixel in the two-dimensional tissue image obtained with the system. We have already initiated work in this direction and developed a preliminary setup for hyperspectral imaging of biological tissue [132].

Another important concern is that the oral tissues are known to have sub-surface layers, with different layers having different biochemical and morphological make-up. As tissue transforms from normal to neoplastic, these sub-surface tissue layers are known to undergo various morphological and biochemical changes quite different from that of the surface epithelium. A major limitation of the conventional optical spectroscopy based systems that we have developed is that they collect light only from the surface of the oral tissue. Thus, it does not necessarily contain complete information of the tissue state, since the

light emission at a given point on the tissue surface is volume integrated over these sub-surface depths. Obtaining the depth-wise spectra information is important because it may facilitate a more detailed analysis of the biochemical (and morphological) state of a given tissue thereby leading to an improved diagnostic feedback. One future objective will be to either develop a depth-sensitive optical spectroscopy probe which could be integrated with the presently developed systems or develop a completely new depth-sensitive optical spectroscopic system which is clinically amenable. This would not only further the diagnosis of the disease but also would enable improved understanding of the underlying biochemistry associated with the progression of cancer. Early detection of oral cancer is critical in patient care. Optical spectroscopy is a tool which has the ability to provide the information necessary to make a difference towards this process.

References

- 1 P. N. Notani, *Curr. Sci.* **2001**, 81, 465.
- 2 T. Rastogi, S. Devesa, P. Mangtani, A. Mathew, N. Cooper, R. Kao, R. Sinha, *Int. J. Epidemiol.* **2008**, 37, 147.
- 3 M. K. Nair, R. Sankaranarayanan, *Cancer Causes Control* **1991**, 2, 263.
- 4 A. S. K. Sham, L. K. Cheung, L. J. Jin, E. F. Corbet, *Hong Kong Med. J.* **2003**, 9, 271.
- 5 P. Boffetta, S. Hecht, N. Gray, P. Gupta, K. Straif, *Lancet Oncol.* **2008**, 9, 667.
- 6 A. Khanna, D. Gautam, P. Mukherjee, *Toxicol. Int.* **2012**, 19, 322.
- 7 R. Doll, R. Peto, K. Wheatley, R. Gray, I. Sutherland, *BMJ* **1994**, 309, 901.
- 8 Ministry of health and Family Government of India, Global Adult Tobacco Survey GATS India, **2010**.
- 9 WHO Report, GLOBOCAN 2012: Estimated cancer Incidence, mortality and prevalence worldwide in 2012, **2012**.
- 10 J. B. Epstein, L. Zhang, M. Rosin, *J. Can. Dent. Assoc.* **2002**, 68, 617.
- 11 P. Garg, F. Karjodkar, *Int. J. Prev. Med.* **2012**, 3, 737.
- 12 P. J. Ford, C. S. Farah, *J. Cancer Policy* **2013**, 1, e2.
- 13 P. E. Petersen, *Oral Oncol.* **2009**, 45, 454.
- 14 M. W. Lingen, J. R. Kalmar, T. Karrison, P. M. Speight, *Oral Oncol.* **2008**, 44, 10.
- 15 C. Carreras-Torras, C. Gay-Escoda, *Med. Oral Patol. Oral Cir. Bucal* **2015**, 20, e305.
- 16 S. Hasan, S. Elongovan, *Int. J. Pharm. Pharm. Sci.* **2015**, 7, 29.
- 17 C. K. Brookner, U. Utzinger, G. Staerke, R. Richards-Kortum, M. F. Mitchell, *Lasers Surg. Med.* **1999**, 24, 29.
- 18 N. Ramanujam, in *Encyclopedia of Analytical Chemistry*, John Wiley & Sons Ltd, Chichester, Chichester, UK, **2000**, pp. 20–56.

- 19 J. K. Dhingra, D. F. Perrault, K. McMillan, E. E. Rebeiz, S. Kabani, R. Manoharan, I. Itzkan, M. S. Feld, S. M. Shapshay, *Arch. Otolaryngol. Head. Neck Surg.* **1996**, *122*, 1181.
- 20 S. K. Majumder, H. Krishna, M. Sidramesh, P. Chaturvedi, P. K. Gupta, *Proceedings of SPIE - The International Society for Optical Engineering*, **2010**, 817306.
- 21 P. Lane, M. Follen, C. MacAulay, *Gend. Med.* **2012**, *9*, S25.
- 22 D. C. G. De Veld, M. J. H. Witjes, H. J. C. M. Sterenborg, J. L. N. Roodenburg, *Oral Oncol.* **2005**, *41*, 117.
- 23 T. D. Wang, J. Van Dam, *Clin. Gastroenterol. Hepatol.* **2004**, *2*, 744.
- 24 M. P. Rosin, C. F. Poh, M. Guillard, P. Michele Williams, L. Zhang, C. Macaulay, *Ann. N. Y. Acad. Sci.* **2007**, *1098*, 167.
- 25 C. F. Poh, L. Zhang, D. W. Anderson, J. S. Durham, P. M. Williams, R. W. Priddy, K. W. Berean, S. Ng, O. L. Tseng, C. MacAulay, M. P. Rosin, *Clin. Cancer Res.* **2006**, *12*, 6716.
- 26 I. Pavlova, M. Williams, A. El-Naggar, R. Richards-Kortum, A. Gillenwater, *Clin. Cancer Res.* **2008**, *14*, 2396.
- 27 A. Gillenwater, R. Jacob, R. Ganeshappa, B. Kemp, A. K. El-Naggar, J. L. Palmer, G. Clayman, M. F. Mitchell, R. Richards-Kortum, *Arch. Otolaryngol. Neck Surg.* **1998**, *124*, 1251.
- 28 S. K. Majumder, A. Uppal, P. K. Gupta, *Lasers Life Sci.* **1999**, *8*, 211.
- 29 S. Majumder, H. Krishna, S. Muttagi, P. Gupta, P. Chaturvedi, *J. Cancer Res. Ther.* **2010**, *6*, 497.
- 30 S. K. Majumder, A. Gupta, S. Gupta, N. Ghosh, P. K. Gupta, *J. Photochem. Photobiol. B Biol.* **2006**, *85*, 109.
- 31 S. K. Majumder, S. K. Mohanty, N. Ghosh, *Curr. Sci.* **2000**, *79*, 1089.

- 32 D. C. G. de Veld, M. Skurichina, M. J. H. Witjes, R. P. W. Duin, H. J. C. M. Sterenborg, J. L. N. Roodenburg, *Lasers Surg. Med.* **2005**, *36*, 356.
- 33 R. A. Schwarz, W. Gao, D. Daye, M. D. Williams, R. Richards-Kortum, A. M. Gillenwater, *Appl. Opt.* **2008**, *47*, 825.
- 34 R. Mallia, S. S. Thomas, A. Mathews, R. Kumar, P. Sebastian, J. Madhavan, N. Subhash, *J. Biomed. Opt.* **2008**, *13*, 41306.
- 35 T. Upile, W. Jerjes, H. J. C. M. Sterenborg, A. K. El-Naggar, A. Sandison et al., *Head Neck Oncol.* **2009**, *1*, 25.
- 36 A. Amelink, O. P. Kaspers, H. J. C. M. Sterenborg, J. E. van der Wal, J. L. N. Roodenburg, M. J. H. Witjes, *Oral Oncol.* **2008**, *44*, 65.
- 37 J. T. Motz, S. J. Gandhi, O. R. Scepanovic, A. S. Haka, J. R. Kramer, R. R. Dasari, M. S. Feld, *J. Biomed. Opt.* **2005**, *10*, 31113.
- 38 M. B. Fenn, P. Xanthopoulos, G. Pyrgiotakis, S. R. Grobmyer, P. M. Pardalos, L. L. Hench, *Adv. Opt. Technol.* **2011**, *2011*, 1.
- 39 J. Zhao, H. Lui, D. I., H. Zeng, in *New Developments in Biomedical Engineering*, ed. by Domenico Campolo, InTech, Rijeka, **2010**, Chap. 24.
- 40 S. Duraipandian, *J. Biomed. Opt.* **2012**, *17*, 081418.
- 41 P. Chen, A. Shen, X. Zhou, J. Hu, *Anal. Methods* **2011**, *3*, 1257.
- 42 H. Krishna, S. K. Majumder, P. Chaturvedi, P. K. Gupta, *Biomed. Spectrosc. Imaging* **2013**, *2*, 199.
- 43 M. D. Keller, E. M. Kanter, A. Mahadevan-jansen, *Spectroscopy* **2006**, *21*, 33.
- 44 A. Mahadevan-Jansen, in *Biomedical Photonics Handbook*, 3rd ed., ed. by Tuan Vo-Dinh, CRC Press, Washington D.C., **2003**, Chap. 30, pp. 1–27.
- 45 S. P. Singh, A. Deshmukh, P. Chaturvedi, C. Murali Krishna, *J. Biomed. Opt.* **2012**, *17*, 1050021.

- 46 A. Malik, A. Sahu, S. P. Singh, A. Deshmukh, P. Chaturvedi, D. Nair, S. Nair, C. Murali Krishna, *Head Neck* **2017**, *39*, 2216.
- 47 M. S. Bergholt, W. Zheng, Z. Huang, *J. Raman Spectrosc.* **2012**, *43*, 255.
- 48 K. Guze, M. Short, S. Sonis, N. Karimbux, J. Chan, H. Zeng, *J. Biomed. Opt.* **2009**, *14*, 14016.
- 49 A. Sahu, A. Deshmukh, A. D. Ghanate, S. P. Singh, P. Chaturvedi, C. M. Krishna, *Technol. Cancer Res. Treat.* **2012**, *11*, 529.
- 50 M. J. H. Witjes, *Head Neck Oncol.* **2009**, *1*, 08.
- 51 C. Wu, J. Gleysteen, N. T. Teraphongphom, Y. Li, E. Rosenthal, *Int. J. Oral Sci.* **2018**, *10*, 10.
- 52 N. Stone, C. A. Kendall, in *Emerging Raman Applications and Techniques in Biomedical and Pharmaceutical Fields*, ed. by Pavel Matousek, Michael D. Morris, Springer, Berlin, Heidelberg, **2010**, pp. 315–346.
- 53 N. Ramanujam, *Neoplasia* **2000**, *2*, 89.
- 54 Tuan Vo-Dinh, *Biomedical Photonics Handbook*, CRC Press, Washington D.C., **2003**.
- 55 I. J. Bigio, S. G. Bown, *Cancer Biol. Ther.* **2004**, *3*, 259.
- 56 A. Uppal, P. K. Gupta, *Biotechnol. Appl. Biochem.* **2003**, *37*, 45.
- 57 L. L. Patton, J. B. Epstein, A. R. Kerr, *J. Am. Dent. Assoc.* **2008**, *139*, 896.
- 58 S. A. McGee, *PhD Thesis*, Harvard-MIT Division of Health Sciences and Technology **2008**.
- 59 R. Malini, K. Venkatakrishna, J. Kurien, K. M. Pai, L. Rao, V. B. Kartha, C. M. Krishna, *Biopolymers* **2006**, *81*, 179.
- 60 Y. Li, Z.-N. Wen, L.-J. Li, M.-L. Li, N. Gao, Y.-Z. Guo, *J. Raman Spectrosc.* **2010**, *41*, 142.
- 61 S. N. Sunder, N. N. Rao, V. B. Kartha, G. Ullas, J. Kurien, *J. Orofac. Sci.* **2011**, *3*, 15.

-
- 62 L. Su, Y. F. Sun, Y. Chen, P. Chen, A. G. Shen, X. H. Wang, J. Jia, Y. F. Zhao, X. D. Zhou, J. M. Hu, *Laser Phys.* **2012**, 22, 311.
- 63 K. Guze, M. Short, H. Zeng, M. Lerman, S. Sonis, *J. Raman Spectrosc.* **2011**, 42, 1232.
- 64 A. Deshmukh, *J. Biomed. Opt.* **2011**, 16, 127004.
- 65 S. Devpura, J. S. Thakur, S. Sethi, V. M. Naik, R. Naik, *J. Raman Spectrosc.* **2012**, 43, 490.
- 66 T. C. Bakker Schut, M. J. H. Witjes, H. J. C. M. Sterenborg, O. C. Speelman, J. L. N. Roodenburg, E. T. Marple, H. A. Bruining, G. J. Puppels, *Anal. Chem.* **2000**, 72, 6010.
- 67 A. P. Oliveira, R. A. Bitar, L. Silveira, R. A. Zângaro, A. A. Martin, *Photomed. Laser Surg.* **2006**, 24, 348.
- 68 A. Mahadevan-Jansen, R. Richards-Kortum, *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. “Magnificent Milestones and Emerging Opportunities in Medical Engineering” (Cat. No.97CH36136), **1997**, 2722–2728.
- 69 C. A. Lieber, A. Mahadevan-Jansen, *Appl. Spectrosc.* **2003**, 57, 1363.
- 70 P. A. Mosier-Boss, S. H. Lieberman, R. Newbery, *Appl. Spectrosc.* **1995**, 49, 630.
- 71 A. P. Shreve, N. J. Cherepy, R. A. Mathies, *Appl. Spectrosc.* **1992**, 46, 707.
- 72 J. Zhao, M. M. Carrabba, F. S. Allen, *Appl. Spectrosc.* **2002**, 56, 834.
- 73 S. T. McCain, R. M. Willett, D. J. Brady, *Opt. Express* **2008**, 16, 10975.
- 74 A. C. De Luca, M. Mazilu, A. Riches, C. S. Herrington, K. Dholakia, *Anal. Chem.* **2010**, 82, 738.
- 75 M. Mazilu, A. C. De Luca, A. Riches, C. S. Herrington, K. Dholakia, *Opt. Express* **2010**, 18, 11382.
- 76 M. D. Morris, P. Matousek, M. Towrie, A. W. Parker, A. E. Goodship, E. R. C.

- Draper, *J. Biomed. Opt.* **2005**, *10*, 14014.
- 77 E. V Efremov, J. B. Buijs, C. Gooijer, F. Ariese, *Appl. Spectrosc.* **2007**, *61*, 571.
- 78 J. V Sinfield, O. Colic, D. Fagerman, C. Monwuba, *Appl. Spectrosc.* **2010**, *64*, 201.
- 79 D. V. Martyshkin, R. C. Ahuja, A. Kudriavtsev, S. B. Mirov, *Rev. Sci. Instrum.* **2004**, *75*, 630.
- 80 P. Matousek, M. Towrie, A. W. Parker, *J. Raman Spectrosc.* **2002**, *33*, 238.
- 81 A. Jirasek, G. Schulze, M. M. L. Yu, M. W. Blades, R. F. B. Turner, *Appl. Spectrosc.* **2004**, *58*, 1488.
- 82 J. A. Westerhuis, S. de Jong, A. K. Smilde, *Chemom. Intell. Lab. Syst.* **2001**, *56*, 13.
- 83 P. Geladi, D. MacDougall, H. Martens, *Appl. Spectrosc.* **1985**, *39*, 491.
- 84 H.-W. Tan, S. D. Brown, *J. Chemom.* **2002**, *16*, 228.
- 85 D. Zhang, D. Ben-Amotz, *Appl. Spectrosc.* **2000**, *54*, 1379.
- 86 Z.-M. Zhang, S. Chen, Y.-Z. Liang, Z.-X. Liu, Q.-M. Zhang, L.-X. Ding, F. Ye, H. Zhou, *J. Raman Spectrosc.* **2009**, *41*, 659.
- 87 N. N. Brandt, O. O. Brovko, A. Y. Chikishev, O. D. Paraschuk, *Appl. Spectrosc.* **2006**, *60*, 288.
- 88 R. N. Hager, R. C. Anderson, *J. Opt. Soc. Am.* **1970**, *60*, 1444.
- 89 D. G. Cameron, D. J. Moffatt, *Appl. Spectrosc.* **1987**, *41*, 539.
- 90 A. Mahadevan-Jansen, M. F. Mitchell, N. Ramanujam, A. Malpica, S. Thomsen, U. Utzinger, R. Richards-Kortum, *Photochem. Photobiol.* **1998**, *68*, 123.
- 91 V. Mazet, C. Carteret, D. Brie, J. Idier, B. Humbert, *Chemom. Intell. Lab. Syst.* **2005**, *76*, 121.
- 92 G. Schulze, A. Jirasek, M. M. L. Yu, A. Lim, R. F. B. Turner, M. W. Blades, *Appl. Spectrosc.* **2005**, *59*, 545.
- 93 T. J. Vickers, R. E. Wambles, C. K. Mann, *Appl. Spectrosc.* **2001**, *55*, 389.

-
- 94 J. Zhao, H. Lui, D. I. McLean, H. Zeng, *Appl. Spectrosc.* **2007**, *61*, 1225.
- 95 M. N. Leger, A. G. Ryder, *Appl. Spectrosc.* **2006**, *60*, 182.
- 96 A. Savitzky, M. J. E. Golay, *Anal. Chem.* **1964**, *36*, 1627.
- 97 H. G. Schulze, R. B. Foist, A. Ivanov, R. F. B. Turner, *Appl. Spectrosc.* **2008**, *62*, 1160.
- 98 M. A. Lifshits, *Gaussian Random Functions*, Springer Netherlands, Dordrecht, **1995**.
- 99 C. A. Lieber, S. K. Majumder, D. L. Ellis, D. D. Billheimer, A. Mahadevan-Jansen, *Lasers Surg. Med.* **2008**, *40*, 461.
- 100 J. De Gelder, K. De Gussem, P. Vandenabeele, L. Moens, *J. Raman Spectrosc.* **2007**, *38*, 1133.
- 101 S. K. Majumder, S. Gebhart, M. D. Johnson, R. Thompson, W.-C. Lin, A. Mahadevan-Jansen, *Appl. Spectrosc.* **2007**, *61*, 548.
- 102 A. Talukdar, *PhD Thesis*, Carnegie Mellon University, Pennsylvania, **1999**.
- 103 B. Krishnapuram, L. Carin, M. A. T. Figueiredo, A. J. Hartemink, *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 957.
- 104 I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, **2002**.
- 105 S. K. Majumder, N. Ghosh, P. K. Gupta, *Lasers Surg. Med.* **2005**, *36*, 323.
- 106 S. K. Majumder, N. Ghosh, P. K. Gupta, *J. Biomed. Opt.* **2005**, *10*, 24034.
- 107 J. Cohen, *Statistical power analysis for the behavioral sciences*, Lawrence Erlbaum Associates, New Jersey, **1988**.
- 108 D. J. Hand, R. J. Till, *Mach. Learn.* **2001**, *45*, 171.
- 109 A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, J. Lyons, D. Hicks, M. Fitzmaurice, R. R. Dasari, J. P. Crowe, M. S. Feld, *Cancer Res.* **2006**, *66*, 3317.
- 110 S. K. Teh, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, *Br. J. Cancer* **2008**, *98*, 457.

- 111 M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Y. So, Z. Huang, *Biosens. Bioelectron.* **2011**, *26*, 4104.
- 112 S. P. Singh, A. Deshmukh, P. Chaturvedi, C. M. Krishna, Biomedical Vibrational Spectroscopy V: Advances in Research and Industry, 82190K.
- 113 F. M. Lyng, E. . Faoláin, J. Conroy, A. D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan, H. J. Byrne, *Exp. Mol. Pathol.* **2007**, *82*, 121.
- 114 K. Kolanjiappan, C. . Ramachandran, S. Manoharan, *Clin. Biochem.* **2003**, *36*, 61.
- 115 R. Rifkin, S. Mukherjee, P. Tamayo, S. Ramaswamy, C.-H. Yeang, M. Angelo, M. Reich, T. Poggio, E. S. Lander, T. R. Golub, J. P. Mesirov, *SIAM Rev.* **2003**, *45*, 706.
- 116 A. Sahu, S. Tawde, V. Pai, P. Gera, P. Chaturvedi, S. Nair, C. M. Krishna, *Anal. Methods* **2015**, *7*, 7548.
- 117 H. Krishna, S. K. Majumder, P. Chaturvedi, M. Sidramesh, P. K. Gupta, *J. Biophotonics* **2014**, *7*, 690.
- 118 M. A. Wall, J. Johnson, P. Jacob, N. L. Benowitz, *Am. J. Public Health* **1988**.
- 119 W. K. Al-Delaimy, *Tob. Control* **2002**, *11*, 176.
- 120 D. A. Tipton, M. K. Dabbous, *J. Periodontol.* **1995**, *66*, 1056.
- 121 G. Taybos, *Am. J. Med. Sci.* **2003**, *326*, 179.
- 122 R. Babuska, in *Fuzzy Systems in Medicine*, ed. by P. S. Szczepaniak, P. J. G. Lisboa, J. Kacprzyk, Physica-Verlag, Heidelberg New York, **2000**, pp. 139–173.
- 123 N. R. Pal, J. C. Bezdek, *IEEE Trans. Fuzzy Syst.* **1995**, *3*, 370.
- 124 H. Gray, W. H. Lewis, *Anatomy of the Human Body*, Lea & Febiger, Philadelphia, **1918**.
- 125 R. Smith, I. Rehman, *J. Mater. Sci. Mater. Med.* **1994**, *5*, 775.
- 126 F. Klawonn, *Mathw. soft Comput.* **2004**, *11*, 125.
- 127 F. N. Ghadially, *J. Pathol. Bacteriol.* **1960**, *80*, 345.

- 128 D. Antory, *Mech. Syst. Signal Process.* **2007**, *21*, 795.
- 129 R. Sankaranarayanan, K. Ramadas, G. Thomas, R. Muwonge, S. Thara, B. Mathew, B. Rajan, *Lancet* **2005**, *365*, 1927.
- 130 W. K. Pratt, *Processing Digital Image Processing*, Wiley & Sons, Hoboken, NJ, USA, **2001**.
- 131 H. Krishna, S. K. Majumder, *Proceedings of DAE-BRNS National Laser Symposium-26*, **2017**, BARC Mumbai, Dec 20-23.
- 132 H. Krishna, S. K. Majumder, *Proceedings of DAE-BRNS National Laser Symposium-25*, **2016**, KIT University, Bhuneshwar, Dec 20-23.