NEAR-INFRARED RAMAN SPECTROSCOPY WITH RECURSIVE PARTITIONING TECHNIQUES FOR PRECANCER AND CANCER DETECTION

TEH SENG KHOON

(B. Eng, National University of Singapore)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DIVISION OF BIOENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2009

To my parents, sister, girlfriend and friends for their
love, support and encouragement

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude towards Dr. Huang Zhiwei, from the Division of Bioengineering, National University of Singapore, who is the supervisor of this research project. I would also like to acknowledge the following collaborators, Assoc. Prof. Teh Ming from Department of Pathology (NUHS (National University Health System) Singapore), Prof. Ho Khek Yu from Department of Medicine (NUHS, Singapore), Assoc. Prof. Yeoh Khay Guan from Department of Medicine (NUHS, Singapore), Assoc. Prof. Jimmy So Bok Yan from Department of Surgery (NUHS, Singapore), and Dr. David Lau Pang Cheng from Department of Otolaryngology (Singapore General Hospital (SGH)), for their invaluable help rendered throughout this entire project for the past 3 years. I would further want to thank all the nurses and colleagues including Amy from the Department of Surgery (NUHS, Singapore), Angela, Nana, Vinnie, and Dr. Zhu Feng who are in the Gastric Clinical Epidemiology Program, the nurses in the Endoscopy Centre from National University Hospital (NUH) and colleagues such as Dr Zheng in the Optical Bioimaging Laboratory who have provided various guidance and assistance during the course of this research work. On top of these, I would like to show earnest appreciation towards my girlfriend (Clarissa), parents, sister, and friends who have inspired me continuously to complete this project. Last but not least, I would also like to acknowledge the following funding agencies for providing financial support to this project, as well as my M.Eng study: Academic Research Fund from Ministry of Education, the Biomedical Research Council, the National Medical Research Council, and the Faculty Research Fund from the National University of Singapore.

Many sincere thanks to you all,

Teh Seng Khoon

NUS, Singapore 2009

PUBLICATIONS (PEER-REVIEWED JOURNALS)

- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Near-infrared Raman spectroscopy for optical diagnosis in the stomach: Identification of *Helicobacter-pylori* infection and intestinal metaplasia", International Journal of Cancer 2009; DOI: 10.1002/ijc.24935.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Near-infrared Raman spectroscopy for early diagnosis and typing of adenocarcinoma in the stomach", British Journal of Surgery 2009; DOI: 10.1002/bjs.6913.
- Z. Huang, **S. K. Teh**, W. Zheng, J. Mo, K. Lin, X. Shao, K. Y. Ho, M. Teh, K. G. Yeoh, "Integrated Raman spectroscopy and trimodal wide-field imaging techniques for real-time *in vivo* tissue Raman measurements at endoscopy", Optics Letters 2009; 34: 758-760.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Near-infrared Raman spectroscopy for gastric precancer diagnosis", Journal of Raman Spectroscopy 2009; 40: 908-914.
- **S. K. Teh**, W. Zheng, D. P. Lau, Z. Huang. "Spectroscopic diagnosis of laryngeal carcinoma using near-infrared Raman spectroscopy and random recursive partitioning ensemble techniques", Analyst 2009; 134: 1232-1239.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang. "Diagnosis of gastric cancer using near-infrared Raman spectroscopy and classification and regression tree techniques", Journal of Biomedical Optics 2008; 13: 034013.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Diagnostic potential of near-infrared Raman spectroscopy in the stomach: differentiating dysplasia from normal tissue", British Journal of Cancer 2008; 98: 457-465.

PUBLICATIONS (CONFERENCES)

- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, S. Manuel, Z. Huang, "Image-guided Raman endoscopic probe for *in vivo* early detection of gastric dysplasia", Best free paper won on the GIHep Singapore 2009, Grand Copthorne Waterfront, Singapore, 20-21 June 2009.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, S. Manuel, Z. Huang, "Image-guided Raman endoscopic probe for *in vivo* early detection of high grade dysplasia", Poster presentation presented on the Digestive Disease Week® 2009, Mccormick place, Chicago, Illinois, 30 May-4 June 2009.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, S. Manuel, Z. Huang, "Early diagnosis and histological typing of gastric adenocarcinoma with near-infrared Raman spectroscopy", Poster presentation presented on the American Association for Cancer Research 2009, Colorado Convention Center, Denver, Colorado, 18-22 April 2009.
- Z. Huang**, S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, "Image-guided near-infrared Raman spectroscopy for *in vivo* detection of gastric dysplasia", Oral presenation presented on the SPIE/BIOS Photonic West 2009, San Jose Convention Center, California, USA, 24-29 January 2009.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Near-infrared Raman spectroscopy to identify and grade gastric adenocarcinoma", Best oral presentation won on the National Health Group Annual Scientific Congress 2008, Suntec Singapore International Convention and Exhibition Centre, Singapore, 7-8 November 2008.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Near-infrared Raman spectroscopy for early diagnosis of *Helicobacter-pylori*-associated chronic gastritis", Poster presentation presented on the Digestive Disease Week® 2008, San Diego Convention Center, San Diego, California, 17-22 May 2008.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, S. Manuel, Z. Huang,

"Detection of *Helicbacter-pylori*-associated chronic gastritis using Raman spectroscopy", Poster presentation presented on the American Association for Cancer Research 2008, San Diego Convention Center, San Diego, California, 12-26 April 2008.

- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Discrimination between normal gastric tissue and intestinal metaplasia by near-infrared Raman spectroscopy", Oral presentation presented on the SPIE/COS Photonics West 2008, San Jose Convention Center, California, USA, 19-24 January 2008.
- **S. K. Teh**, W. Zheng, D. P. Lau, Z. Huang, "Raman spectroscopy for optical diagnosis of laryngeal cancer", Oral presentation presented on the SPIE/COS Photonics West 2008, San Jose Convention Center, California, USA, 19-24 January 2008.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Near-infrared Raman spectroscopy for optical diagnosis of gastric precancer", Poster presentation presented on the SPIE/COS Photonics Asia 2007, Jiuhua Grand Convention and Exhibition Center, Beijing, China, 11-15 November 2007.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Discrimination of gastric cancer using near-infrared Raman spectroscopy and multivariate techniques", Oral presentation presented on the World Congress of Bioengineering 2007, Twin Towers Hotel, Bangkok, Thailand, 9-11 July 2007.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Optical diagnosis of dysplastic lesions in the human stomach using near-infrared Raman spectroscopy and multivariate techniques", Poster presentation presented on the Digestive Disease Week® 2007, Washington DC, United States of America, 19-24 May 2007.
- **S. K. Teh**, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, "Discrimination of malignant tumor from benign tissue in the GI tract using Raman spectroscopy", Poster presentation presented on the Office of Life Sciences conference 2007, Center of Life Sciences, Singapore, 5-6 February 2007.
- Z. Huang**, S. K. Teh**, W. Zheng, J. C. H. Goh, "Raman spectroscopy for

evaluation of structure deformation in stressed bone tissue", Oral presenation presented on the 15th International Conference on Mechanics in Medicine and Biology 2006, Furama Riverfront Hotel, Singapore, 6-8 December 2006.

- Z. Huang**, S. K. Teh**, W. Zheng, Casey K. Chan, "Assessment of degeneration of human articular cartilage using Raman spectroscopy", Oral presenation presented on the Singapore Orthopaedic Association 29th Annual Scientific Meeting 2006, Grand Copthorne Waterfront, Singapore, 8-11 November 2006.

TABLE OF CONTENTS

SUMMARY

Raman spectroscopy is a molecular vibrational spectroscopic technique that is capable of optically probing the biomolecular changes associated with disease transformation. To effectively translate molecular differences captured in Raman spectra between different tissue types into clinically valuable diagnostic information for clinicians, chemometrics would need to be deployed for developing effective diagnostic algorithms for Raman spectroscopic diagnosis of precancer and cancers. However, most of the chemometrices (principal component analysis (PCA)) applied for Raman tissue diagnosis cannot adequately provide the physical meanings of component spectra for tissue classification This dissertation presents the investigation on the diagnostic utility of near infrared (NIR) Raman spectroscopy with recursive partitioning techniques such as classification and regression trees (CART), and random forests to construct clinically interpretable diagnostic algorithm for tissue Raman classification.

A rapid-acquisition dispersive-type NIR Raman system was utilized for tissue Raman spectroscopic measurements at 785 nm laser excitation. A total of 146 tissue samples obtained from 70 patients who underwent endoscopy investigation or surgical operation were used in this study. The histopathogical examinations showed that 94 were gastric tissues (55 normal, 21 dysplastic, and 18 cancerous), and 50 were laryngeal tissues (20 normal, and 30 cancerous).

CART was explored to be used together with NIR Raman spectroscopy for gastric cancer diagnosis. CART achieved a predictive sensitivity and specificity of 88.9% and 92.9%, respectively, for separating cancer from normal. In addition, CART also determined tissue Raman peaks at 875 and 1745 cm^{-1} to be two of the most significant features in the entire Raman spectral range to discriminate gastric cancer from normal tissue. This affirmed the utility of CART to be used for NIR Raman spectroscopy detection of cancer tissues.

To improve diagnostic performance (e.g., stability) of CART, the random ensemble approach (i.e., random forests) was further utilized. Random forests yielded a diagnostic sensitivity of 88.0% and specificity of 91.4% for laryngeal malignancy identification, and also provided variables importance plot that facilitates correlation of significant Raman spectral features with cancer transformation. These confirmed the diagnostic potential of random forests with NIR Raman spectroscopy for detection of malignancy occurring in the internal organs (i.e., larynx).

Comprehensive evaluation of the performance of the empirical approach that utilizes Raman peak intensity ratio, PCA-linear discriminant analysis (LDA), and random forests algorithm was also carried out. Raman peak intensity ratios representing biomolecular signals for collagen, proteins and lipids achieved diagnostic accuracy of approximately 88% for NIR Raman spectroscopic detection of gastric dysplasia from the normal gastric tissues. Further investigation on the use of PCA-LDA achieved obtained a diagnostic accuracy of 93%, while random forests achieved diagnostic accuracy of 90% for gastric

dysplasia detection. Receiver operating characteristics (ROC) curves further confirmed that PCA-LDA and random forests techniques have comparable overall diagnostic accuracy rate which are more superior compared to the empirical approach.

Overall, this dissertation demonstrates that NIR Raman spectroscopy in conjunction with powerful chemometric techniques such as random forests have the potential to generate interpretable clinical Raman information, and to yield high diagnostic accuracy classification results for the rapid diagnosis and detection of precancer and cancer tissues.

LIST OF FIGURES

FIGURES PAGE

LIST OF TABLES

TABLES PAGE

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION AND MOTIVATION

As the majority of cancers (~90%) are epithelial in origin, early detection and localization with immediate removal (e.g., surgery) of malignant tumors is critical towards decreasing the mortality rate of the patients [1]. However, early identification of cancer lesions in the lining of the internal organs such as stomach and larynx can be very challenging through conventional diagnostic method such as the white-light endoscope which heavily relies on the visual examination of gross morphological changes of tissue, leading to a poor diagnostic accuracy [1]. Endoscopic biopsy currently remains the standard approach for most cancer diagnosis, but is invasive and impractical for screening high-risk patients who may have multiple suspicious lesions [2]. Hence, it is highly desirable to develop noninvasive optical diagnostic techniques for direct assessments of biochemical information of suspicious lesion sites during clinical examinations.

Optical spectroscopic methods such as light scattering spectroscopy, fluorescence spectroscopy, and Raman spectroscopy have been comprehensively investigated for cancer and precancer diagnosis and evaluation [1-24]. Raman spectroscopy is a vibrational spectroscopic technique that is capable of probing specific biochemical fingerprints of biological tissues based on inelastic light scattering processes [5]. This technique has shown great promise for detecting molecular alterations associated with diseased transformation [5-12]. With the use of near-infrared (NIR) lasers as excitation light sources, NIR Raman spectroscopy holds significant advantages in that water exhibits very low absorption at the working wavelength range, and tissues exhibit far less autofluorescence than with visible light excitation [12]. Less water absorption makes it easy to detect other tissue components and results in deeper light penetration into the tissue [12]. As a result, NIR Raman spectroscopy has been widely studied for early detection of pre-malignancy and malignancy in a number of organ sites [1, 5, 6, 15], including the stomach [14, 25-28] and larynx [10, 21, 24].

In order to convert molecular differences subtly reflected in Raman spectra between different tissues types into valuable diagnostic information for clinicians, different statistical techniques have been explored in developing effective diagnostic algorithms for Raman spectroscopic of precancer and cancer diagnosis [5, 6, 20, 29, 30]. Due to the complexities of the biological tissues, multivariate statistical techniques (e.g., principal component analysis (PCA)), which are able to take into account of the whole range of Raman spectral features of the tissue, have often been applied to construct high diagnostic accuracy algorithms for different tissue type classification [7-9, 11-13]. However, most of these multivariate statistical techniques (e.g., PCA) could not adequately furnish the clinicians with physical meanings of diagnostic features derived for tissue characterization [29]; thereby, the development of robust algorithms, which not only produce a high predicted diagnostic accuracy, but also provide useful biomolecular diagnostic information from the high dimensional Raman spectral datasets, is highly desirable.

1.2 Specific Aims of the Dissertation

The primary aim of this dissertation was to evaluate the clinical potential of NIR Raman spectroscopy combined with different chemometric algorithms, especially the recursive partitioning techniques for detection of precancer and cancer tissues. Hence, the following specific aims were developed:

1. Assessment on the feasibility of using a rapid fiber-optic NIR Raman spectroscopy system for clinical evaluation of human tissues, and to characterize the Raman properties of internal organ tissues (i.e., gastric and laryngeal tissue).

2. Exploration on the potential of classification and regression trees techniques (CART) for use with NIR Raman spectroscopy in stomach cancer diagnosis.

3. Investigation on the ensemble technique for recursive partitioning algorithms (i.e., random forests) in identification of laryngeal carcinoma from normal laryngeal tissues with the use of NIR Raman spectroscopy.

4. Study of empirical method for gastric precancer detection with NIR Raman spectroscopy.

5. Comprehensive comparison of the potential of empirical method (i.e., intensity ratio) with the multivariate statistical techniques (i.e., PCA and linear discriminant analysis (LDA)) to be used together with NIR Raman spectroscopy for discrimination of gastric dysplasia from normal.

6. Evaluation of random forests technique together with empirical method (i.e., intensity ratio) and the multivariate statistical techniques (i.e., PCA-LDA) for NIR Raman spectroscopic detection of gastric dysplasia.

1.3 ORGANIZATION OF THE DISSERTATION

The study is structured into three main parts. This dissertation begins with providing a detailed background on the Raman instrumentations, the preprocessing method and the types of human tissues samples which have been employed throughout the entire study. The second part of this study is focused on the development of recursive partitioning algorithms from the construction of a single classification tree (i.e., CART), to an ensemble of approximately 1000 classification trees (i.e., random forests) for cancer tissue diagnosis using NIR Raman spectroscopy. The third part is to assess the performance of random forests with respect to two commonly utilized diagnostic algorithms (i.e., intensity ratio and PCA-LDA) for NIR Raman spectroscopy tissue diagnosis. A thorough evaluation of the three different diagnostic algorithms was conducted through the use of precancer tissues to also affirm the diagnostic utility of random forests with NIR Raman spectroscopy for precancer diagnosis.

Specifically, Chapter 2 provides the overview of Raman technique and its development for precancer and cancer diagnosis, extensive review on the application of Raman technology for pre-malignancy and malignancy detection in different organ sites, and the summary of the various diagnostic algorithms which have been utilized to understand and translate Raman molecular signals into clinically useful information. Chapter 3 illustrates the hardware instrumentation, data preprocessing techniques and the type of tissues that have been utilized in this dissertation. Chapter 4 gives the introduction of recursive partitioning technique (i.e., CART) for NIR Raman spectroscopy diagnosis of cancer tissue. In chapter 5, application of the ensemble recursive partitioning algorithms (i.e., random forests) for NIR Raman spectroscopic diagnosis of cancer tissue will be shown. Chapter 6 describes the empirical approach (i.e., intensity ratio) which has been commonly utilized to construct a simple, yet useful diagnostic algorithm for detection of precancer tissues using Raman spectroscopy. Chapter 7 further demonstrates the diagnostic utility of multivariate statistical techniques (i.e., PCA-LDA) in conjunction with Raman spectroscopy for diagnosing precancer tissue. Chapter 8 verifies the diagnostic performance of random forests for precancer tissue in comparison with the empirical and multivariate statistical techniques. The final chapter concludes the work in the dissertation and proposes possible work in the future.

CHAPTER 2

OVERVIEW ON RAMAN SPECTROSCOPY FOR PRECANCER AND CANCER DIAGNOSIS

The discoverer of the Raman effect was Chandrasekhara Venkata Raman who published in *Nature* entitled ‘The color of the sea’, in which he showed that the color of the ocean is due to scattering of light [22]. He continued his investigation on scattering of light and eventually discovered the Raman effect in 1928 [31]. The Raman effect is an inelastic light scattering process whereby a very small proportion of incident photons are scattered (~1 in 10^8) with a corresponding change in frequency. The difference between the incident and scattered frequencies corresponds to the vibrational modes of molecules participating in the interaction. These Raman scattered light can be collected by a spectrometer and displayed as a ‘spectrum’, in which its intensity is displayed as a function of its frequency change.

As most biomolecules are Raman-active scatterers, each with its own spectral fingerprint, and Raman spectra usually exhibit sharp spectral features that are characteristic for specific molecular structures and conformations of tissue, it can provide more specific molecular information about a given tissue or disease state [5]. Therefore, in the past decade, Raman spectroscopy has been comprehensively investigated for precancer and cancer diagnosis and evaluation in humans including in the bladder, brain, breast, cervix, gastrointestinal tract, head and neck, lung, oral, skin, and prostate. Many of these studies have shown that specific spectral features of Raman spectra could be used to correlate with the molecular and structural changes of tissue associated with neoplastic transformation [1, 5-21, 32-34]. In combination with multivariate statistical analysis such as PCA and LDA, NIR Raman spectroscopy has demonstrated promising diagnostic accuracy (~90%) for Raman detection of precancer and cancer tissues in different organ sites (i.e., stomach) [1,2,6-9,12,13,15,16, 21,24].

The present chapter presents an overview on the development of Raman technology for cancer tissue diagnosis, and a review on the different analytical algorithms commonly applied for tissue Raman diagnosis so as to provide comprehensive background knowledge on this project work.

2.1 TECHNOLOGICAL ADVANCEMENT FOR CLINICAL RAMAN SPECTROSCOPY SYSTEM

As Raman scattering (inelastic scattering) is inherently very weak, typically 10^{-9} to 10^{-6} of the intensity of the Rayleigh background (elastic scattering), intense monochromatic excitation and a sensitive detector are critical towards obtaining observable Raman signals [22]. Hence, advancement of Raman spectroscopy for biomedical application only began with the development of lasers and sensitive detector in 1960s [22, 35, 36].

The first laser-based Raman spectroscopy system for biological application arises from the use of visible (VIS) excitation with a photomultiplier or multi-channel optical detector used to detect scattered photons in the frequency range of interest [37].

However, as the techniques for biological application progressed and due to technological advancement, NIR laser excitation gradually became the frequent choice for Raman spectroscopic investigation on biological tissues [22]. This section shall cover the development of NIR Raman technology for biological tissue diagnosis.

2.1.1 Excitation wavelength strategies for biomedical Raman Spectroscopy

The use of different excitation lights such as ultraviolet (UV), VIS and NIR light for Raman spectroscopic studies [22] will generate different light scattering, absorption and emission phenomenon in biological and biomedical systems. In this sub-section, a summary on the investigation of the different laser wavelength for Raman spectroscopy to be used in biomedical application will be presented.

2.1.1.1 Visible (VIS) and near Ultra-violet (UV) excitation

Most biological tissues exhibit significant autofluorescence signals which will severely interfere with weak Raman signals with the use of VIS or near-UV excited Raman spectroscopy. Hence, in order to reduce background autofluorescence signals emitted from biological tissues, the samples had to be photobleached (pre-irradiated) before recording reliable Raman signals [38]. To date, only corneal collagen and lens proteins have been found to produce very little or no autofluorescence signal with VIS excited Raman spectroscopy [22]. As a result, to avoid photobleaching biological tissues which would change the tissue biomolecular conformation and structures, and circumvent the strong autofluorescence signals with the use of VIS or near-UV excited Raman spectroscopy, deep UV (>300nm) and NIR excited Raman spectroscopy, instead, could be utilized for biological application [38, 39].

2.1.1.2 Deep UV resonance Raman spectroscopy

The resonance Raman effect occurs when the excitation laser wavelength is in close proximity with an electronic transition (i.e., absorption band) of the analyte. Thus, by selecting the appropriate excitation wavelength, Raman bands of molecules can be selectively greatly enhanced in the midst of a myriad of overlapping vibrations from various tissue components [22]. On top of the resonance enhancement effect, the scattering cross-section is also increased. These combined effects lead to tremendous increase in Raman intensity, which allow detection of biomolecules in very low concentration. As the penetration of UV light on biological tissue is shallow (<50 μm), it can also effectively target biomolecules on the superficial tissue surface layer, such as the epithelial tissue where most cancerous lesions often originate from. However, there is a potential problem associated with the photomutagenicity on the use of UV light on biological tissues [22].

2.1.1.3 Near-infrared (NIR) excitation Raman spectroscopy

The autofluorescence signal decreases very rapidly at longer excitation wavelengths, and most biological tissues exhibit little or no autofluorescence signals when excited in the NIR spectral range [38]. In addition, NIR light has a relatively small extinction coefficient (absorption coefficient) in biological tissues, and so facilitating a deeper light penetration, in the order of millimeters, which can probe larger tissue volume information [40]. The small absorption coefficient will also not result in photo-degradation of the interrogated biological samples [22]. Furthermore, water is a relatively weak absorber in the NIR. Thus, even though biological cells are usually composed of about 70-95% of water by weight, water will not significantly interfere with NIR Raman spectroscopy for biological application [40]. On top of this, the use of NIR excitation light is compatible to be used with fiber-optic technology, which makes NIR excitation Raman spectroscopy technique highly possible to directly collect remote in situ tissue signals from all parts of human body [23]. As a result, in comparison with UV and VIS excitation Raman spectroscopy, NIR excited Raman spectroscopy provides the most benefits for biological application. Therefore, most of the Raman spectroscopic studies on biomedical application are centered on the use of NIR light.

The earliest form of NIR Raman spectroscopy system (i.e., Fourier-Transform (FT) Raman) primarily uses 1064 nm from a neodymium-doped yttrium aluminium garnet (Nd:YAG) laser as the excitation source, a cooled indium gallium arsenide (InGaAs) detector, and a Michelson interferometer system [41-44]. By working with 1064 nm in the NIR, background autofluorescence is almost entirely eliminated [22]. However, the signal-to-noise ratio (S/N) produced from the NIR FT Raman spectrosocpy is limited by both reduced scattering cross-section at the 1064 nm excitation wavelength, and the intrinsic noise associated with the InGaAs detectors in the spectral range of 1100 – 1350 nm (~Raman shift of 300-200 cm^{-1}) [22]. Hence, long integration time of about 30-60 mins for acquiring high-quality Raman signal from biological tissue is often required for the use of NIR FT Raman 41-43]. Long acquisition time for collection of reliable Raman signal is the main drawback for NIR FT Raman to be employed for biomedical application. On top of this, the throughout advantage of interferometer-based FT Raman spectroscopy is lost due to the incompatibility of the numerical aperture (NA) of the system with the optical fibers which can be used for clinical application [22]. This greatly hinders the development of NIR FT Raman system for remote spectroscopic clinical application.

With technological advancement, a more efficient NIR Raman system, which can provide a high S/N, could be achieved, and so greatly shortened the integration time needed to record a reliable Raman signal. The following subsections (Section 2.1.2 – 2.1.5) will elaborate more in details on the different essential Raman components which are critical towards the development of NIR Raman spectroscopy for biomedical diagnosis.

2.1.2 Charged-coupled device (CCD)

As the noise level of a CCD-based NIR Raman system is signal shot noise limited, while the noise level of an InGaAs-based NIR Raman system is limited by detector noise (e.g. dark current and read-out noise) which is several orders of magnitude larger than the CCD-based NIR Raman system, the CCD-based NIR Raman system could result in a higher S/N [22]. Hence, in order to achieve a better performance, most NIR Raman works progressively focused on the CCD-based NIR Raman system.

There are a variety of different types of CCD such as front illuminated, thinned back-illuminated and front- or back-illuminated deep depletion CCD which are used for different applications [22]. For Raman study, as the Raman signal is very weak, a highly sensitive CCD which can obtain the highest possible photon detection efficiency is the most important criteria. Thus, a thinned back-illuminated CCD detector is often the choice to be used for Raman system as it has higher quantum efficiencies than a front-illuminated CCD. However, in the NIR spectral region, thinned back-illuminated CCD detectors introduce the etalon effect [22]. The newer deep-depletion back-illuminated CCD is specially fabricated and optimized for the NIR light to minimize this elatoning effect. As a result, most current Raman clinical systems employ the use of the deep-depletion back-illuminated CCD detector to maximize quantum efficiency and minimize etalon artifacts.

One important factor to note is that most CCD detectors are only efficient to about 1100 nm wavelength as quantum efficiency drops considerably due to silicon absorption [6]. Furthermore, the high quantum efficiency (QE) of CCD detector, especially at the VIS-excitation range, though enable weak Raman emissions to be detected, it also collect strong fluorescence signals arising from biological tissues which could be beyond the dynamic range of the CCD [22]. This fluorescence signal will also produce shot noise which may interfere with extraction of Raman information. Therefore, due to the limitation of current CCD detector technology, for biological tissue application, the optimal NIR excitation wavelength range is generally between 750 to 850 nm for collection of high quality Raman emission signals within a few seconds [6,22], with most Raman work centered on the use of either 785 [5,6, 9]or 830 nm [30,32].

2.1.3 SPECTROGRAPH

A spectrograph is an important instrument that can separate an incoming light into different frequency on the CCD detector in real-time. For clinical application through using optical fiber and low-power laser excitation to collect incoming tissue scattered light into the spectrograph for CCD collection of spectral data with high spectral resolution of about 8-10 cm^{-1}, careful selection of spectrograph would be necessary [22]. The employment of volume-phase transmissive dispersive grating spectrograph which has its f-number matched with the optical fiber could provide both the high throughout and flat image field at the detector plane required for sensitivity at low laser fluence and spectral resolution at the range of interest [22].

In addition, an important factor which determines the sensitivity of a Raman spectrometer is the usable detection area (i.e., usable slit width x height) [45]. For the majority of Raman application, the larger the sampling area, the most scattered Raman signals can be gathered, which will increase the sensitivity of the Raman system [45]. On a multi-channel detector dispersive-based spectrometer, a given spectral resolution often limits the slit width [45]. Extending the slit height using a straight slit usually causes the image to be curved on the detector due to optical effects [45]. If the optical effects are not corrected, the curved slit image will degrade the peak shape and spectral resolution. 45, 46] One of the ways to correct this image distortion (i.e., aberration) effect is to use a curved entrance slit, opposite to the image curve distortion effect, so that a straight slit image can be achieved [46]. Most details will be provided in Chapter 3 on the corrected image aberration Raman spectroscopy system been utilized in this study.

2.1.4 Fiber-optic probe

Medical applications usually require remote sampling use of optical fibers in which the sizes of the Raman probe and the fiber bundle are strictly limited by anatomic considerations [35]. For instance, in order to endoscopically evaluate stomach mucosa for gastric cancer with Raman spectroscopy, the size of the probe must be small enough (~2 mm in diameter) and long enough (several meters) to be inserted into a narrow-diameter channel [35, 47]. Moreover, the design and material of the Raman probe must be able to undergo regular hospital instrument sterilization procedures [22].

In addition to the physical demands which the Raman probe needs to face, there are also optical characteristic requirements which the Raman probe must possess in order to be clinically applied. For example, as the Raman probe can only probe a small tissue area of interest, it would require the guidance of different wide-field imaging modalities to the suspicious tissue area for evaluation [48]. Hence, the design of the fiber-optic Raman probe must be able to collect high S/N Raman signal in approximately 1s with safe levels of laser exposure for accurate clinical application of the spectral model used for analysis, while also minimizing the light interference from the use of the different wide-field imaging modalities [48]. On top of the external interference, the optical fiber also generated significant intrinsic spectral interference which must be greatly reduced [47]. Fiber fluorescence and absorption interference can be minimized through the use of highpurity low-hydroxyl fibers, and Raman interference from the optical fibers can be removed through installing appropriate high performance filters at the distal tip of the probe [47]. However, the production of such Raman probe which demands high performance criteria, and yet requiring the size of the probe to be small (~2 mm) is of great technical challenge; hence, to date, only a few Raman endoscopic probe have been successfully developed [39, 45, 49, 50]. Note that one particular group has explored an alternative approach for designing the Raman probe such that the fabrication of such Raman probe is very much simpler. They have introduced the exploration of so-called "high-wavenumber" spectral range for tissue Raman diagnosis as this spectral range has minimal interference from the probe, thereby requiring less optical components in the design [51]. Hence, the use of "high-wavenumber" enable the use of a single, unfiltered optical fiber for guiding laser light to the sample and for collecting the back-scattered light to the spectrometer [51].This alternative approach is still an on-going area of research [51, 52] to unravel the potential which the "high-wavenumber" Raman spectroscopy could bring about for tissue diagnosis using Raman spectroscopy.

2.2 AUTOFLUORESCENCE ELIMINATION APPROACHES TO ACHIEVE BACKGROUND-FREE RAMAN SPECTRUM

Biological tissues under NIR excitation wavelength range of between 750 to 850 nm will not only collect weak Raman signals, but will also pick up intrinsic tissue autofluorescence emissions; thereby posing a significance challenge in recovering background-free Raman signals [47]. Through examining the dissimilarity in the inherent optical property of fluorescence and Raman emissions, various techniques have been attempted to recover Raman signals free from its concomitant autofluorescence background signal.

2.2.1 TIME-GATING TECHNIQUES

For instance, the lifetimes of fluorescence and Raman emissions are distinctly different: a fluorescence lifetime is roughly in the order of 10^{-9} to 10^{-7} second, while Raman scattering events are normally around 10^{-11} to 10^{-13}s [53]. Hence, time-gating techniques such as Kerr-gating have been investigated as one of the potential method to separate Raman and fluorescence signals [6, 53]. As this technique aims to reject the fluorescence before it is detected, it can uniquely remove photon shot noise associated tissue autofluorescence signals. However, due to the need to use high excitation fluence pulsed lasers, on top of the high cost and complicated system design required, this technique is unsuitable for clinical application [54].

2.2.2 SHIFTED EXCITATION RAMAN DIFFERENCE SPECTROSCOPY

In addition, the shift response of the fluorescence and Raman effect to small excitation wavelength shifts allows another mechanism for removing the fluorescence background [55]: by Kasha's rule which stated that majority of fluorescence is emitted from vibrationally relaxed states, small changes in excitation wavelength will have a minimal effect on the fluorescence emission; while, Raman emission will shift in energy by the amount of excitation shift [55]. Therefore, with two slightly different excitation wavelength (~5-10 cm^{-1}) the fluorescence spectral profile would be invariant, while the entire Raman spectra will be shifted by the corresponding changes in excitation wavelength (~5-10 cm^{-1}). The difference of the two emission spectral profile would eliminate the fluorescence background signal, leaving the difference Raman spectral. This difference Raman spectral is approximately the first derivatives of the original Raman spectrum, and hence reconstruction techniques such as integration or deconvolution would be able to retrieve the Raman spectrum without its concomitant autofluorescence background [6, 55]. This technique, commonly named "shifted excitation Raman difference spectroscopy" [6, 55, 56], is particularly attractive as it could enable the removal of large fluorescence background as well as other sources of random or systemic noise generated by, for example, the detector where the different sensitivity of individual pixels of the detector can produce systemic effects; thereby permitting sensitivity to true photon shot levels [57]. Hence, it has led to a new development of other modified technique such as the so-called "subtracted-shifted Raman spectroscopy", which offers a more simplistic instrumental design by shifting the spectrometer grating (i.e., wavelength) instead of necessitating a tunable laser [57]. However, regardless of shifted excitation Raman difference spectroscopy or subtracted-shifted Raman spectroscopy, this type of technique would be required to, at least, double the integration time in order to obtain the required spectra; which may not be clinically feasible if long integration time is required [54].

2.2.3 Frequency/wavelength-modulated

Since by Kasha's rule which stated that fluorescence emission is independent of slight change in excitation wavelength, while Raman scattering frequency will change according to the amount of change in excitation wavelength, illumination of samples with frequency-modulated excitation light can also be used to remove fluorescence signal from Raman signal. This is done through modulating the excitation light at low frequencies which will achieve time-invariant fluorescence signal [6, 54], and also yield Raman scattering frequencies which will shift with the modulated excitation light. Even though this technique is also an effective way to separate fluorescence and Raman signals, it requires the use of specialized instrumentation for modulation [54].

2.2.4 Digital post processing

Most Raman signals appear as spike-like features residing on top of a broad band NIR autofluorescence background. Hence, mathematical post-processing techniques can be employed to separate the Raman and fluorescence signals [58]. By linearly transform the original data into another feature space such as carrying out Fourier-transformation on the NIR raw data, before introducing different filtering algorithms, can be applied for separation of the NIR Raman and background autofluorescence signal; however, this often lead to artifacts and distortions of the Raman spectrum, especially during a noisy situation [56]. One of the most efficient, easiest and accurate way to subtract the fluorescence signal with minimal distortion to the Raman signal is to fit the spectrum containing both Raman and fluorescence signals to a polynomial of a high enough order to describe the fluorescence line shape but not the higher frequency Raman line shape [6]. However, a single polynomial fitting for fluorescence background removal may not be sufficient to retain Raman signal with minimal fluorescence background interference [59]. Hence, improvement of the polynomial fitting algorithms has been investigated. For instance, an improved algorithm is such that all data points in the polynomial curve have an intensity value higher than the input spectrum (comprising of Raman and background autofluorescence signal) are automatically reassigned to the original intensity. This process will then re-run repeatedly until there is convergence in the number of data points affected by each iteration, as determined by root test for convergence. The processed base-line spectrum is then subtracted from the original spectrum to achieve Raman signal with minimal autofluorescence background interference [59]. This algorithm works reasonable well but still suffers from a few weaknesses such as unable to take into account of noise interference, and divergence of the polynomial-fitting occurring at the endpoint of the selected spectral region of the tissue. Hence, new improved algorithms have been investigated to address these weaknesses for a more robust and unbiased removal of the tissue intrinsic autofluorescence background to achieve reliable Raman signals from biological tissue [60, 61]. Note that this is still an on-going active area of research [62, 63]. Overall, polynomial fitting provides the simplest (involves the simplest hardware configuration, and the simplest computational complexity among other methods) but relatively accurate method to extract Raman emissions with minimal autofluorescence background signal from biological tissues.

Table 2.1 presents the Raman peak features which have been commonly found in the literature for biomedical studies, together with corresponding tentative biochemical assignments after removal of autofluorescence background with polynomial fitting. These multiple Raman biochemical features give rise to a "fingerprint spectrum" which can be very specific for tissue diagnosis, especially for precancer and cancer diagnosis [5]. The next section will provide a comprehensive literature review on Raman technique application for cancer and precancer diagnosis.

Table 2.1 Raman peak features commonly found in the literature for biomedical studies with tentative biochemical assignments [1, 2, 5, 6, 7, 10, 15, 21, 29, 65]

Peak position (cm^{-1})	Protein assignments	Lipid assignments	Others
1745w		ν(C=O)	
1655vs	ν(C=O) amide I (α-helix conformation, collagen)		
1620w			ν (C=C) porphyrin
1585vw	ν(C=C) olefinic		
1558vw	ν(CN) and δ(NH) amide II		ν (C=C) porphyrin
1514			ν (C=C) carotenoid
1445vs	δ(CH_2), δ(CH_3)	δ(CH_2) scissoring	
1379vw		δ(CH_3) symmetric	
1336mw (sh)	δ (CH_2), δ (CH_3), twisting, collagen		
1302vs	δ(CH_2) twisting, wagging, collagen	δ(CH_2) twisting, wagging	
1265s	ν(CN) and δ(NH) amide III (α-helix conformation, collagen)		
1208vw	ν(C-C_6H_5) phenylalanine		
1168vw		ν(C=C), δ(COH)	ν (C-C), carotenoid
1122mw (sh)		$ν_s$(CC) skeletal	
1078ms		ν(CC) skeletal	ν(CC),$ν_s$(PO_2^-) nucleic acids
1030mw (sh)	ν(CC) skeletal, keratin		

1004mw	ν(CC) phenylalanine ring	
973mw (sh)	$\rho(CH_3)$, δ(CCH) olefinic	
935mw	$\rho(CH_3)$ terminal, proline, valine; ν(CC) α-helix keratin	
883mw	$\rho(CH_2)$	
855mw	δ(CCH) phenylalanine, olefinic	polysaccharide

ν, stretching mode; ν_s, symmetric stretch; ν_{as}, asymmetric stretch; δ, bending mode; ρ, rocking mode; v, very; s, strong; m, medium; w, weak; sh, shoulder

2.3 Review on cancer biology

Cancers mainly arise in epithelial tissue (~85%), which are the cells lining the surface of organs. Due to the constant exposure to carcinogens, they are likely to trigger a cascade of carcinogenesis events [65]. Most of these changes begin with early biochemical alterations, and a fraction of these diseased tissues will eventually progress to become malignant tumors [66]. At the cancerous stage, obvious morphological and tissue architecture changes can often be noticed and detected by the clinicians. However, the early lesions (e.g., early cancer and precancer), which can predispose to become invasive neoplasia, are often associated with subtle signs of tissue transformation that are difficult to identify by the clinicians [65]. Nevertheless, these precancerous and early cancer lesions involve molecular transformation which could be advantageously utilized for tissue diagnosis.

It is important to recognize the various patho-histological features of neoplastic tissues which are generally distinct from normal tissues. For instance, malignant tumors are microscopically distinguished from the normal by cellular crowding and disorganization [6], nuclear content, nuclear-to-cytoplasmic ratio, mitotic activity, chromatin distribution, changes in the angiogenesis process and differentiation rate [60]. More specific well-know characteristics of cancerous cells involve the production of a higher quantity of lactate acid for cancerous cells relative to the normal cells, changes in nuclear proteins that regulate cell division and DNA replications, activation of regulatory oncoproteins/oncogenes that results in repeated duplication and amplification of DNA sequences, and increase secretion of proteolytic enzymes such as serine, cysteine, and metalloproteinases [67]. One notes that the morphologic and biochemical changes that accompany neoplasia transformation are many and, most often depend on the specific type and location of the cancer [67]. For instance, alpha fetoprotein level is typically very low for adult human being. However, alpha fetoprotein has been found to increase significantly with the development of liver cancer, but not for other types of cancer such as carcinoma arising in the colon, lung, and pancreas [67]. In another instance, high molecular weight keratin, which is a unique feature of mature epithelium in the cervix, will be replaced by low molecular weight keratin during malignancy transformation; however, cancerous transformation involving all other keratinized tissues (e.g., buccal mucosa cancer) do not share this similar molecular alteration [6]. These confounding factors present a significant challenge for clinicians to detect different type of cancers at different organ sites. Nonetheless, cancerous transformation of tissue at different organ sites will result in changes in nucleic acid, protein, lipid, and carbohydrate conformational structures and compositions. Technology which can tap upon these biomolecular changes would be useful diagnostic tools for early detection of cancers.

2.4 Review on Raman Technique for precancer and cancer diagnosis in different organ sites

This section will briefly provide an up-dated overview of the application of NIR Raman spectroscopic technique in precancer and cancer diagnosis for different organ parts.

2.4.1 Bladder cancer

Several studies have demonstrated the potential of Raman spectroscopy for bladder caner diagnosis [1, 33, 68, 69, 70]. For instance, Crow *et al.* demonstrated the feasibility of differentiation between benign (normal and inflammatory) and malignant tissues with a sensitivity and specificity of 90-95% and 95-98%, respectively [68]. They have also shown that NIR Raman spectroscopy could be used to stage bladder cancer into noninvasive and invasive malignant bladder tissues (i.e., transitional cell carcinoma) with an accuracy of 96% [68]. However, these proof-of-concept studies have been carried out with a free-space Raman system (i.e., Raman microscopy) testing on *ex vivo* bladder tissue samples. Hence, Crow *et al.* further verified that *ex vivo* bladder tumor can be discriminated from the *ex vivo* nontumor tissues with a NIR Raman spectroscopic system equipped with a fiber-optic probe which is compatible with the working channel of a flexible cystoscope [69]; thereby, suggesting the potential of NIR Raman spectroscopy which could be compatible with a conventional cystoscope for bladder cancer tissue diagnosis. Their group further utilized NIR Raman spectroscopy to determine the biochemical basis for different bladder pathologies which include high, moderate and low grades transitional cell carcinoma, carcinoma in situ (CIS), cystitis, and normal urothelium tissues [70]. These Raman studies provided an insightful understanding of the molecular changes in associated with bladder cancer, and also demonstrated the diagnostic utility of NIR Raman spectroscopy for detection of diseases occurring within the bladder.

2.4.2 Brain cancer

Complete removal of malignant tissue, while preserving healthy tissue is a common aim of most oncosurgical procedures, particularly for brain surgery as imprecise targeting of brain tumors may increase the risk of damaging vital brain areas that may result in functional impairment (e.g. speech, movement and etc) [23]. On top of this, tissue samples obtained by stereotactic surgery are relatively small, sampling errors may easily occur which may miss crucial features essential for determining the follow-up procedures of brain cancer such as glioma, which is the most common type of brain cancer [23]. Hence, most of the Raman research in brain cancer area [52, 71, 72] has been aimed at the development of an *in vivo* Raman technique to assist stereotactic surgery for real-time intraoperative optical biopsy guidance, which to date, has not been realized. Recent *ex vivo* Raman works have also demonstrated the potential to detect and grade gliomas [71], and the feasibility of differentiation among glioma, meningioma and normal brain tissues [52]. These *ex vivo* results preliminary concluded that Raman spectra contain tissue molecular information which can be potentially employed for brain cancer detection.

2.4.3 Breast cancer

Breast cancer is the most common cancer, and also the leading cause of cancer death in women worldwide [30]. This fatal disease has been extensively studied by many groups to explore the feasibility of Raman spectroscopy to diagnose breast cancer [1, 30, 32, 34, 41, 73, 74, 75]. For instance, Feld *et al*. have shown the potential of Raman spectroscopy for *ex vivo* tissue identification of infiltrating carcinoma, fibroadenoma, fibrocystic change, and normal breast tissue [32]. With the constructed diagnostic algorithm derived from their *ex vivo* tissue samples, they have also demonstrated the possibility of *in vivo* Raman spectroscopic cancer tissue identification during partial mastectomy breast surgery [30]. Recently, Stone *et al*. has also explored the potential of noninvasive Raman technology such as "spatially-offset Raman spectroscopy" and "transmissive Raman spectroscopy" techniques for noninvasive optical differentiation of different types of microcalcifications occurred during breast cancer [74, 75]. The detection of the different calcifications associated with breast cancer is important as it can be related to the metastasis state of breast cancer; however, to date, no reliable mean for assessing the type of microcalcification has been established [74, 75]. It is expected that Raman technology could also eventually provide a noninvasive diagnostic mean of assessing the tumor stage of malignant breast tissue.

2.4.4 CERVICAL CANCER

A decade of relentless effort by Mahadevan-Jansen *et al.* to pursue early detection of cervical cancer has passed [6, 17, 20, 65, 66, 76]. Their group has demonstrated their cervical precancers can be distinguished from benign tissue *in vitro* [17], as well as showing the possibility of Raman *in vivo* detection of cervical precancers [20] following their successful development of a fiber-optic probe for *in vivo* Raman measurements [77]. They have also constructed an organotypic tissue culture to further their understanding on cervical cancer development [78]. Recently, they have also shown the potential of Raman spectroscopy together with the use of sophisticated nonlinear multivariate statistical algorithms to detect metaplastic cervical tissues [66]. In addition, they have also further proved that hormonal variation will affect Raman readings and have robustly demonstrated that with the stratification of data by menopausal status, low grade dysplasia could be identified from normal with accuracy of nearly 100% [79]. In recent years, few other groups have also contributed to NIR Raman spectroscopic understanding on cervical cancer [80-82]. For instance, Jess *et al*. has demonstrated that Raman spectroscopy could be used to detect cervical cells infected with human papilloma virus (HPV) [80], and Vidyasagar *et al*. has shown that Raman spectroscopy could be used to predict radiotherapy response in cervix cancer [81]. Martinho *et al*. have also pointed out that the inflammatory infiltrates can affect Raman spectroscopic detection of low grade dysplasia [82]. The results of these various studies from different groups gradually bring NIR Raman spectroscopy closer to being employed for cervical precancer detection in a clinical setting.

2.4.5 GASTROINTESTINAL CANCERS

Shim *et al*. is the first to demonstrate that NIR Raman spectroscopy can be used to acquire reliable *in vivo* Raman signals in the entire digestive tract during clinical inspection [83]. To further robustly investigate the potential of Raman spectroscopy for detection of different gastrointestinal lesions related to cancer, various groups have performed laboratory-based Raman testings. For example, Kendall *et al*. has demonstrated sensitivities of 77-100%, and specificities of 92-100% for detection of metaplasia, dysplastic and cancerous lesions occurring in the esophagus with the use of Raman spectroscopy [13]. Boere *et al*. further verified the potential of NIR Raman spectroscopy to detect dysplasia occurring in the Barrett's esophagus by using a rat model [84]. Ling *et al.* is the first to illustrate the potential of Raman spectroscopy for detection of gastric cancer [25], and Teh *et al.* further proved the possibility of NIR Raman spectroscopy to discriminate gastric dysplasia from normal tissues with diagnostic accuracy of 90% [12]. In addition, the feasibility of NIR Raman spectroscopy to detect precancers and cancers in the lower digestive tract (i.e., colon) has also been verified [1, 9]. These studies demonstrated the diagnostic utility of NIR Raman spectroscopy for identification of different lesions associated with gastrointestinal cancers.

2.4.6 HEAD AND NECK CANCER

Raman spectroscopy has been shown to possess the potential to discriminate among normal, dysplastic and malignant changes in the epithelial cells of the larynx with diagnostic sensitivities of 83%, 76%, and 92%, and diagnostic specificities of 94%, 91%, and 90% [24]. Discrimination among normal, papilloma, and malignant laryngeal tissues with specificities of 86%, 94%, and 94%, and sensitivities of 89%, 69%, and 88% using NIR Raman spectroscopy has also been reported [21]. Additionally, successful differentiation of nasopharyngeal cancer from normal tissues (i.e., accuracy of 100%) with the use of NIR Raman spectroscopy has also been demonstrated [85]. Note that these works were *in vitro* studies, and *in vivo* study has yet to be demonstrated. Nevertheless, these investigations have shown that Raman spectroscopy possesses the potential as a tool to be used for improving efficacy in the detection of lesions occurring in the head and neck.

2.4.7 LUNG CANCER

Raman study (excitation wavelength at 647 nm) by Bakker Schut *et al.* investigated the level of carotenoids from lymphocytes of individual with and without cancer [86]. Their work indicated a considerable reduction of carotenoids in lung cancer patients compared with healthy individuals. Subsequent lung Raman investigations were carried out by both Yamazaki *et al.* [87] and Huang *et al.* [5] who have each reported on the feasibility of their in-house developed Raman spectroscopy system for *ex vivo* detection of lung cancer tissues [85]. Jess *et al.* further demonstrated that Raman spectroscopy could also be used for grading different malignant lung cells [88]. To further bring Raman technology closer to a clinical endoscopic application, Short *et al.* has successfully developed a fiber-optic Raman probe *in vivo* lung cancer detection [49]. However, they have reported overwhelming native autofluorescence signal from the bronchial tissues which obscured reliable Raman signals acquisition. Hence, they have explored the so-called "high-wavenumber" spectral region to acquire lung tissue Raman signals and has preliminary demonstrated the feasibility of *in vivo* Raman spectroscopic lung cancer detection. Recently, Magee *et al.* has also developed a new Raman endoscopic probe and NIR Raman system based upon "shifted-subtracted Raman spectroscopy" to tackle the high autofluorescence background signals [89]. Their results indicated that their in-house developed Raman system based upon "shifted-subtracted Raman spectroscopy" technique could be effectively circumvent NIR autofluorescence background and acquire reliable tissue Raman data. They have also preliminary demonstrated the potential of their system to differentiate lung cancer from normal lung tissues. Hence, overall, regardless of "high-wavenumber" Raman spectroscopy or "shifted-subtracted Raman spectroscopic technique" used in the fingerprint spectral region, Raman spectroscopy holds great promises as a useful diagnostic instrument for *in vivo* lung cancer detection.

2.4.8 ORAL CANCER

With a rat model, Bakker Schut *et al*. demonstrated the feasibility of a fiber-optic NIR Raman system for *in vivo* detection of dysplastic tissue occurring at the palate anatomical site with 100% accuracy [8]. Oliveira *et al*. further utilized an animal (i.e., hamster) model to cultivate dysplastic and cancerous oral lesions which were subsequently excised for *ex vivo* tissue Raman testings [90]. Their results suggested Raman spectra of dysplastic and cancerous oral tissues were very similar. Nevertheless, their group confirmed the diagnostic potential of Raman spectroscopy for oral cancer diagnosis. On top of the animal model Raman works, Krishna *et al*. validated the diagnostic potential of Raman spectroscopy using human *ex vivo* oral tissues to differentiate different pathologies associated with cancer [91, 92]. Their results indicated that Raman spectroscopy could be used to distinguish the cancer, inflammatory and normal tissues with 100% accuracy; thereby, providing more evidence that NIR Raman spectroscopy could be utilized for disease detection in the oral cavity.

2.4.9 SKIN CANCER

Skin cancers including squamous cell carcinoma, melanoma and basal cell carcinoma, are among the cancers with the highest incidence worldwide [23, 35, 39]. These neoplastic lesions are often fatal if left undetected and untreated. Excisional biopsy currently remains the standard approach for cancer diagnosis, but is invasive, impractical, and could be unacceptable for screening high-risk patients who may have suspicious lesions localized in cosmetically important parts of the body such as the face [39]. The concept of applying Raman spectroscopy for skin cancer diagnosis is particularly appealing as this optical spectroscopic technique is nondestructive and does not require any sample preparation, enabling acquisition of the spectra directly from the skin [23, 35, 39]. Gniadecka *et al*. provided the first Raman spectroscopic studies of skin cancer, in which they observed spectral differences between malignant and benign lesions such as the basal and squamous cell carcinoma *vs*. lentigo maligna, seborroic keratosis and nevus intrademalis [93]. Their group subsequently specifically demonstrated that melanoma could be discriminated from pigmented nevi, basal cell carcinoma, seborrheic keratoses, and normal skin with a sensitivity of 85% and specificity of 99% through the use of Raman spectroscopy in combination with artificial neural network. [94, 95] However, these studies were conducted through the use of FT-Raman, which could not be applied clinically. Nijssen *et al.* further demonstrated that basal cell carcinoma could be discriminated the perilesional skin with the use of a special fiber-optic Raman probe which explored the use of "high-wavenumber" Raman spectroscopy technique [96]. With the latest development of a portable optical fiber Raman microspectroscopy system which allows 40 μm measurement depth to be performed, Lieber *et al.* verified the potential of NIR Raman spectroscopy for tissue diagnosis of basal cell carcinoma and squamous cell carcinoma from normal and inflammatory skin tissues [97, 98]; they have achieved high diagnostic sensitivity of 91%, specificity of 95% for detecting the skin cancer lesions [97]. It should be also noted that Lieber *et al.* has observed that malignancy occurred at a deeper tissue layer (>40 μm) [97], however, Raman spectroscopy appeared to be able to detect malignancy-associated changes in the morphologically normal tissue surrounding the lesions. As a result, their group has further investigated the malignancy-associated changes for skin tissues with the use of organotypic raft culture, and further determined the possibility of Raman spectroscopy to detect malignancy-associated changes in the histologically-appearing normal tissue surrounding the lesions [65]. These works highlighted that Raman spectroscopy possesses the potential to elucidate biochemical changes closely related with diseased transformation, and could be employed as a unique diagnostic tool for clinical inspection.

2.4.10 Prostate cancer

The prognosis and type of therapeutic intervention for prostate cancer is primarily based on clinical stage, serum prostate-specific antigen, and the Gleason score of the cancer [99]. The Gleason score grading is based on microscopic tumor architecture and the extent of the most prevalence pattern. As the tumor grades are indicator of intrinsic tumor behavior, characterizing the molecular phenotype of grade is of potential clinical importance [99]. As Raman spectroscopy is capable of optically probing the biomolecular changes associated with diseased transformation, it could be a potential instrument for detection and grading of prostate cancer. To date, Crow and Stone *et al*. has been the only group who has explored the possibility of NIR Raman spectroscopy for prostate cancer diagnosis [1, 100]. They have demonstrated the diagnostic efficacy of Raman spectroscopy to detect and grade prostate cancer *in vitro* with an overall accuracy of 89% [100]. Further investigation by their group has showed the feasibility of implementing NIR Raman spectroscopy with a fiber-optic probe for prostate cancer diagnosis [69]. In addition, they have investigated the biochemical basis for the spectral differences in correlation with benign prostatic hyperplasia, prostatitis, and three different grades of malignant prostate cancer [70]. Their results confirmed the molecular information which could be extracted by Raman spectroscopy for diagnosis of different prostate diseases. These studies provided a molecular basis towards clinical implementation of NIR Raman technology for possible noninvasive tissue diagnosis of prostate cancer in living patients.

To summarize this chapter, the potential of NIR Raman spectroscopy for precancer and cancer diagnosis has been demonstrated in many different clinical fields such as the dermatology, gynaecology, gastrointestinal, neurology, respiratory, and urology [1,2,6-9,12,13,15,16, 21,24, 73-100]. Even though Raman technique has progressively moved from *in vitro* tissue validation towards *in vivo* clinical trials due to substantial improvement in the hardware technology, further development of Raman equipment is still necessary, as the current state-of-the art instrumentation is not fully ready to be incorporate in the clinical practice yet [6, 15, 23, 35, 38, 39]. In addition, construction of a robust diagnostic algorithm which can robustly translate Raman spectral information into interpretable clinical information for real-time accurate tissue diagnosis still requires considerable development [23, 35, 39]. The following chapter will introduce on the different analytical techniques which could be implemented for Raman tissue diagnosis.

2.5 ANALYTICAL TECHNIQUES FOR RAMAN CLASSIFICATION

Raman spectroscopy can probe great wealth of biomolecular information ranging from nucleic acids, proteins, lipids and carbohydrates from biological tissues, and present the biochemical information on a Raman spectrum. However, the Raman spectrum usually contains many overlapping bands, and so data interpretation can not be easily made by simple visual inspection for subtle change in tissue pathology [22]. Hence, different statistical techniques would need to be implemented in order to made use of the biomolecular Raman signal for tissue analysis and classification.

The Raman spectrum data usually consists of the results of observations of many different variables (i.e., Raman shift) for different cases (i.e., normal and diseased). Each of these variables could be considered to represent a different dimension. Hence, given n variables, each of the cases may be regarded to be located in a unique position in an n-dimensional hyperspace, which is often very difficult to visualize. Therefore, various statistical algorithms have been explored to reduce this massive dimensional space to an interpretable dimensional space [36].

Very often, tissue Raman spectra can be modeled using linear analysis as Raman spectrum of a mixture of biochemicals is approximately a linear superposition of the mixture's component spectra, and the Raman signal intensity varies with the biochemical component concentration [22]. Thus, most of the dimensional reduction techniques which have been commonly explored for tissue Raman analysis are of the linear analysis methods [39]. The linear dimensional reduction techniques basically comprises of two types: unsupervised or supervised learning algorithms [22, 39]. The unsupervised learning algorithms aim to find intrinsic differences and similarities among the different cases, and group them into clusters. Well known examples of this approach are PCA, and hierarchical cluster analysis (HCA) [36]. In contrast, supervised pattern recognition methods utilize prior information on class memberships to construct a classifier using a portion of samples available (assigned as training set), and the rest will be used as test set to check the classifier performance. Examples of supervised techniques include LDA, and logistic regression (LR) [36]. In recent years, to further improve the diagnostic accuracies for tissue diagnosis using Raman spectroscopy, nonlinear learning algorithms such as support vector machines (SVM) [11], and artificial neural network (ANN) [94, 95] have also been explored. Below provides a short description of the different algorithms commonly applied for tissue diagnosis in Raman spectroscopy.

2.5.1 Principal component analysis (PCA)

Briefly, PCA decomposes the spectroscopic data matrix S into scores T and loading P, according to the relation,

$$S = T.P \tag{2.1}$$

With this equation, PCA transforms a number of correlated variables into a number of uncorrelated variables called principal components (PCs) which describe the greatest variance of the spectral data. It is usually employed as a method for variable or data reduction by retaining the first few principal components. In addition, inspecting the plots generated using scores provides a mean to assess the relationships between samples, since it helps to identify some clusters related to a certain feature and also for detecting potential outliers [39].

2.5.2 Hierarchical cluster analysis (HCA)

In general, hierarchical cluster analysis is the partitioning of a dataset into clusters so that the differences between the data within each cluster are minimized and the differences between clusters are maximized according to a specific distance measure [36]. This is achieved through calculating the symmetric distance matrix (size n x n) between all considered spectra (number n) as a measure of their pairwise similarity/dissimilarity [39]. The algorithm searches for the minimum distance, collects the two most similar/dissimilar spectra into the first cluster, and recalculates spectral distances between all remaining spectra and the first cluster. In the subsequent step, the algorithm performs a new search to cluster more objects (spectra or already formed clusters) together. This searching and formation of clusters algorithm will be repeated n-1 times until all spectra have been merged into a single cluster. The final result will be displayed in a tree-like, two dimensional dendrogram in which one axis refers to the reduction of clusters with increasing number of iterations and the other axis to the respective spectral distances.

2.5.3 Linear discriminant analysis (LDA)

LDA seeks linear combinations of the measurement variables which separate the objects from different classes as much as possible [101]. Factors which determine the separability of classes include the distances (e.g. Euclidean distances) between groups and the compactness of each group. It then follows that the ratio of the between-to-within variability of the transformed training data vectors (i.e. spectra) should be maximized (i.e., S_{max}).

$$S_{max} = \{Variance_{between} / Variance_{within}) \qquad (2.2)$$

In other word, the aim of LDA is to find a discriminant function line that maximizes the variance in the data between groups and minimizes the variance between members of the same group [101]. Unknown object or specimen in our case can then be classified according to its position with respect to the discriminant function line [101].

2.5.4 Logistic regression (LR)

The fundamental assumption of logistic regression is that the probability, p_i, that the *i*th case in a set of spectral data belongs to a particular category, for instance, the data illustrating cancerous tissue will be described by the logistic function, which is of the form,

$$p_i = \{1+ \exp (\alpha + \sum \beta_i x_i)\}^{-1} \qquad (2.3)$$

with x_i the score associated with the *i*th basis spectrum, β_i the corresponding weighting coefficient and α a constant offset [64]. This probability varies from zero to one with a sigmoid linkage function. It is small (i.e., zero) for large values of *x* and approaches to one for small values of *x*. The parameters β and α are often determined through maximum likelihood iterative technique, which aims to maximize the probabilities predicted for obtaining the specific calibration set [102].

2.5.5 Support vector machines (SVM)

SVM algorithm was firstly introduced by Vapnik [103]. It basically classifies two linearly separable set of data that belong to two different groups through the use of a hyperplane [104]. Although there are infinitely many hyperplanes that separate the two classes, SVM classifiers find the hyperplane that maximizes the distances between the two groups by solving a quadratic optimization equation, the Lagrangian dual problem, as shown below [104]:

$$\min_{w,b,\xi} = (1/2) \parallel w \parallel^2 + (C/2) \sum \xi^2$$
$$y_i (<w.x_i>+b) \geq 1 - \xi_i,\ i=1,\ldots..,n. \qquad (2.4)$$

Note that w represents the weights, ξ corresponds to a slack variable, b signifies the bias term, C denotes the penalty cost, x indicates the data set with n number of variables, and y illustrates the binary class information from the support vectors represented by the two integer, {-1, 1}. SVM can be extended to nonlinear classification by implicitly transforming the original data in a nonlinear space using kernel functions [104].

2.5.6 ARTIFICIAL NEURAL NETWORK (ANN)

ANN is another type of non-linear statistical data modeling tool which can be used to develop a classification algorithm [94]. One of the most popular and accurate ANN model is the multilayer perceptions based on the backpropagation algorithm [95] for training neural networks, which can be modeled as,

$$y = f \{b + \sum w_i x_i\} \qquad (2.5)$$

with x_i representing the variables, w_i denoting the weighting coefficient, b corresponding to a constant offset, f denoting a function, and y signifying the output. Backpropagation ANNs are constructed with a layered structure in which each node is connected to all nodes of the preceding and subsequent layer with different weights [95]. The input layer (variables) and the output layer (class membership) are generally connected by a single hidden layer. The input to each node is the linear combination of the outputs of the previous layer using the respective weights for the connections. Each node comprises of a Gaussian or sigmoid activation function. The output of the node is the result of the activation function for the input value, and the outputs will reveal the predicted class membership for the input cases [95].

Overall, these different statistical algorithms have been commonly applied to construct diagnostic models for classification of different diseases with Raman spectroscopy. However, these different techniques could not sufficiently furnish the clinicians with physical meanings of diagnostic features derived for tissue characterization. Hence, the development of robust algorithms which can achieve a high predicted diagnostic accuracy and provide useful biomolecular diagnostic information from the high dimensional Raman spectral datasets will be clinically useful.

2.5.7 RECURSIVE PARTITIONING TECHNIQUES

Classification and Regression Tree (CART) is a form of recursive partitioning statistical technique that can selectively employ variables which are of utmost importance from a large number of input variables in databases for binary discrimination [105]. It is implemented by growing a tree structure with a root node containing all the objects, which are then further divided into nodes by recursive binary splitting. The theory of this technique has been robustly established by Leo Breiman *et al*. in 1984, and was originally meant to be utilized in mass spectroscopy analysis [105]. Following the next two decades, this algorithm has generated great attention in the scientific and clinical field due to the successful applications in numerous applications [106-110]. To further enhance the performance of CART, different variation of recursive partitioning techniques [111, 112] such as the random forests [113] have been successfully developed. Despite the usefulness of recursive partitioning algorithms in many different fields, including in the mass spectroscopy application, these algorithms have yet been reported in detail for Raman spectroscopic biomedical application in the literature.

The following two chapters will present in details on the novel application of CART and random forests for Raman spectroscopic study on cancer diagnosis. The last three chapters will illustrate the commonly utilized analytic algorithms (i.e., empirical method (i.e., intensity ratio) and multivariate statistical techniques (i.e., PCA-LDA)) used for NIR Raman spectroscopy studies on different tissue/cell types for precancer diagnosis, and will also compare the diagnostic efficacy of these conventional Raman analytical algorithms with random forests.

CHAPTER 3

ASSESSMENT ON THE FEASIBILITY FOR USING A RAPID FIBER-OPTIC NIR RAMAN SPECTROSCOPY SYSTEM TO CHARACTERIZE RAMAN PROPERTIES OF HUMAN TISSUE

Despite the great advantages that NIR Raman spectroscopy could offer, there are technical challenges to overcome. For instance, achieving a high signal-to-noise (S/N) ratio while avoiding interference from silica Raman signals in a rapid manner can be difficult for *in vivo* tissue Raman measurements [12]. This is because tissue Raman scattering is inherently very weak, and the fiber-optic probes used to collect *in vivo* signals exhibit strong silica Raman scattering in the fingerprint region. Also, the integration times and irradiance powers for *in vivo* Raman measurements must be limited for practical and safety reasons [12, 47]. Therefore, the primary aim of this chapter was to investigate the use of a rapid in-house developed fiber-optic NIR Raman spectroscopy for characterization Raman properties of human mucosa tissues.

3.1 Raman Instrumentation

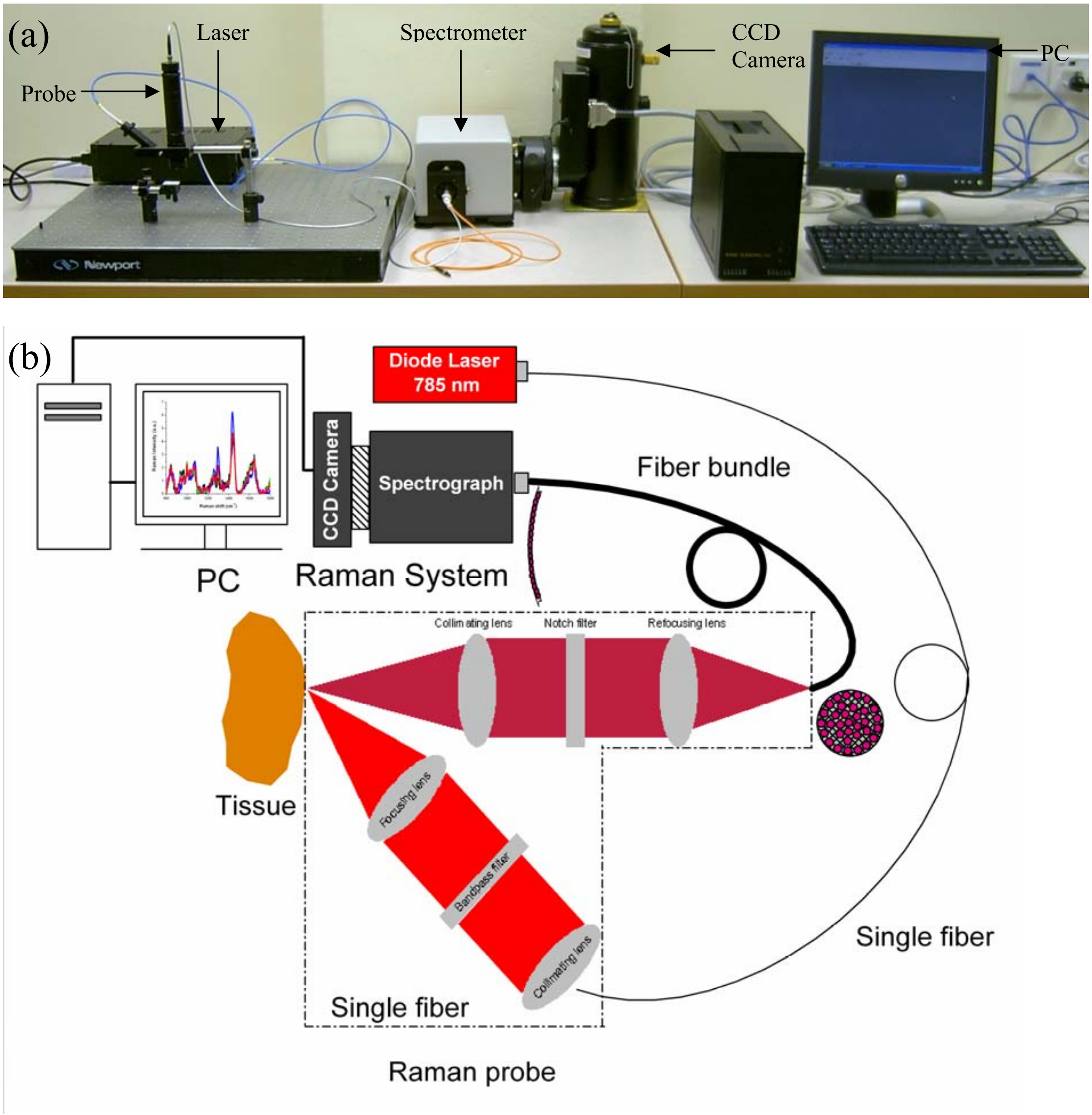

Figure 3.1 (a) Photograph of the in-house developed Raman system used to acquire tissue Raman measurements. (b) Schematic of Raman spectroscopy system used for Raman collection. CCD: charge-coupled device; PC: personal computer.

Figure 1(a) shows the instrument used for tissue Raman spectroscopic studies [12, 46]. Briefly, this system consists of a 785 nm diode laser, a transmissive imaging spectrograph with a Kaiser holographic grating, an NIR-optimized back-illuminated, deep-depletion CCD detector (Princeton Instruments, Trenton, NJ), and an unique fiber-optic Raman probe. The 785 nm laser is coupled to a 200 μm core diameter fiber (NA=0.22) and the fiber is connected to the Raman probe via a subminiature version A (SMA) connector. The Raman probe was designed to maximize the collection of tissue Raman signals while reducing the interference of Rayleigh scattered light, fiber fluorescence and silica Raman signals (Figure 1(b)). One optical arm of the probe consists of a collimating lens, a bandpass filter (785 ± 2.5 nm, Chroma Technology Corp., VT) and a focusing lens to deliver the laser light onto the tissue. The other arm of the probe equipped with collimating and refocusing lenses and a holographic notch plus filter (optical density >6.0 at 785 nm; Kaiser) is used for collecting tissue Raman signals. The holographic notch filer was placed between the two lenses to block the Rayleigh scattered excitation laser light while passing the frequency-shifted tissue Raman signal. The refocusing lens then focused the filtered beam onto the circular end of the fiber bundle (58 x 100 μm core diameter fibers, NA = 0.22). Tissue Raman photons collected by the fiber bundle in the Raman probe are fed into the entrance of the transmissive spectrograph along a parabolic curve, and the holographic grating disperses the incoming light onto the liquid nitrogen-cooled CCD array detector controlled by a personal computer (PC) . The tissue Raman spectra associated with autofluorescence background are displayed on the computer screen in real time and can be saved for further analysis. The system acquired Raman spectra over the wavenumber range of 800-1800 cm^{-1}, and each spectrum was acquired within 5 seconds with light irradiance of 1.56 W/cm^2. The spectral resolution of the system is 4 cm^{-1}.

3.1.1 Unique feature of the in-house developed Raman system

All straight slit usually causes the image to be curved on the detector due to the diffraction effect: To understand this, consider the diffraction equation [45]:

$$m\lambda = d\,(\sin\alpha + \sin\beta) \qquad (3.1)$$

where m is the diffraction order, λ is the wavelength, d is the groove spacing, and α and β are the incident and diffraction angles respectively. This equation can only be applied to the light rays originating from a single point, which is the slit centre. After collimation of light before interacting with the grating groove, the light rays are perpendicular to the grating grooves, and are also parallel to the dispersion plane. Hence, the light rays originating from above or below the slit centre (i.e., a single point) will form a vertical out-of-plan off axis angle δ with the dispersion plane. The grating mimics a mirror with respect to the vertical angle δ and reflects light rays with the same vertical angle. As a result, the dispersion in the diffraction plane would be altered and will be represented by the below modified equation [45]:

$$m\lambda = d\cos\delta\,(\sin\alpha + \sin\beta) \qquad (3.2)$$

As can be observed, it is the factor cos δ causes the curvature in the image plan. Briefly, the shape of the slit image is computed to be in the form of a parabolic curve

$$x' = c'\,y'^2 \qquad (3.3)$$

where x' and y' are the distances in the x and y axis formed by the curved slit image in the image plan, and c' is a constant for a certain wavelength and is determined by the following [45]:

$$\begin{aligned} c' &= (m\lambda) / (2df_2\ \cos\beta) \\ &= (\sin\alpha + \sin\beta) / (2f_2\ \cos\beta) \qquad (3.4) \end{aligned}$$

where f_2 is the focal distance from the focusing lens to the image plan. Note that the diffraction angle β's dependence on δ is removed in equation (3.4) to illustrate the on-axis condition where $\delta = 0$. As a result, from the equation, it is obvious that the image on a straight slit in the shape of a parabolic curve. Four observations can be derived from this equation [45].

1. The curvature becomes more prominent as the diffraction angle increase.
2. When $m = 0$, and $\beta = - \alpha$, which will represent specular reflection, the curvature disappears.
3. For opposite diffractions orders, their shapes are symmetric around the zeroth-order reflection.
4. The curvature becomes more prominent as the focal length f_2 decreases. Hence, for compact spectrometers which usually utilize short focal distance lenses, the curvature is very much obvious. However, short lenses are necessary because of the high numerical aperture it provides (or low f/number).

Therefore, in order to obtain a straight slit image, a curve entrance opposite to the curved slit image can be employed. Using symmetry, the shape of the entrance slit is also a parabolic curve and is described by [45]:

$$x = cy^2 \qquad (3.5)$$

$$c = (\sin \alpha + \sin \beta) / (2f_1 \cos \beta) \qquad (3.6)$$

where x and y are the distances in the x and y axis in the entrance slit, and where f_1 is the focal distance from the entrance slit to the collimating lens. Note that the shape of the parabolic curve at the entrance slit needs to be accounted for the magnification lens used in the spectrometer.

For ease of manufacturing, an arc to approximate the parabolic curve can be carried out. Hence, solving the below equation will yield the radius of the arc (R) [45, 46].

$$R = 1/2c \qquad (3.7)$$

Since the ideal radius of the arc is dependent on wavelength, only one wavelength can be fully corrected, and residual curvature will exist for all other wavelengths. For the Raman system employed in this dissertation, spectrograph image aberration was corrected by a parabolic-line fiber array, permitting complete CCD vertical binning, thereby yielding a 3.3-16 fold improvement in S/N ratio for the use of 785 nm excitation rays [12, 46]. To date, this is the unique design to correct for image aberration effect for Raman spectroscopic study in the literature. The greatly enhanced S/N ratio allows rapid acquisition of weak Raman signals from biological tissues within 1 sec [46].

3.2 Data preprocessing

Due to the gradual change in quantum efficiency of the CCD detector across the NIR spectral range which affects the instrument sensitivity with increasing wavelength, etaloning in the CCD, and variation in the individual CCD pixel sensitivity, this can result in a rapid oscillation in the sensitivity of a Raman instrument with changing wavelength [1]. Hence, in order to extract reliable data independent of these inherent effects from the Raman system, all wavelength-calibrated spectra were corrected for the wavelength-dependence (system response) of the system using a standard lamp (RS-10, EG&G Gamma Scientific, San Diego, CA). Figure 3 shows an example of a tissue raw spectrum before and after correcting for the system response.

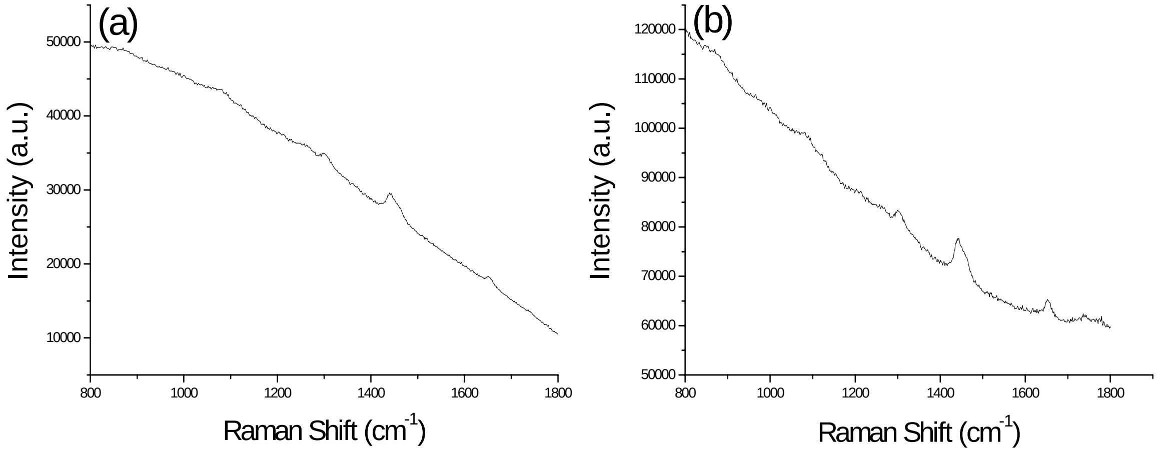

Figure 3.2 Example of a tissue raw spectrum (a) before and (b) after correcting for the system response.

The raw spectra acquired from gastric tissue in the 800-1800 cm^{-1} range represented a combination of prominent tissue autofluorescence, weak tissue Raman scattering signals, and noise [12]. Thus, the raw spectra were preprocessed by a first-order Savitsky-Golay filter (window width of 3 pixels, which corresponded to the system spectral resolution) to reduce noise (Figure 3(a)) [114]. A fifth-order polynomial was found to be optimal for fitting the broad autofluorescence background in the noise-smoothed spectrum (Figure 3(b)), and this polynomial was then subtracted from the raw spectrum to yield the tissue Raman spectrum alone (Figure 3(c)) [12].

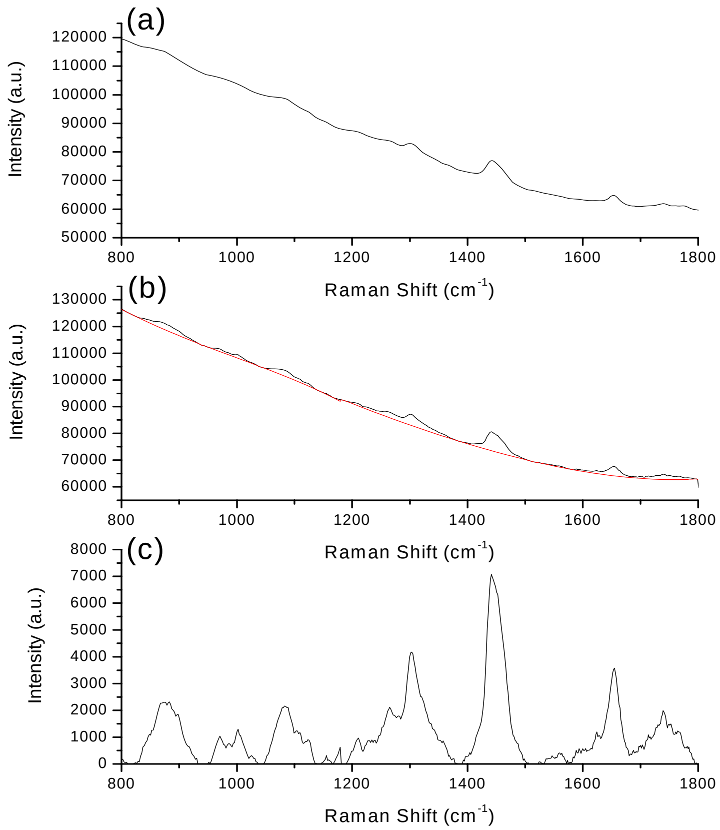

Figure 3.3 Example of a tissue raw spectrum (a) after noise removal via Savitsky-Golay filter, (b) followed by fitting the autofluorescence background with a 5th order polynomial, and (c) this polynomial was then subtracted from the raw spectrum to yield the tissue Raman spectrum alone. Note: tissue raw spectrum and tissue Raman spectrum, black; 5th order polynomial autofluorescence background, red.

In order to achieve tissue Raman data independent of Raman spectroscopic measurement conditions such as excitation/detection geometries, excitation light power fluctuations, probe-tissue positioning variations, different tissue sample size, and etc, each of background-subtracted Raman spectra was also normalized to the integrated area under the curve from 800-1800 cm^{-1} to enable a better comparison of the spectral shapes and relative peak intensities among the different tissue samples [5, 16].

3.3 EX VIVO TISSUE SAMPLES

In this study, a total of 146 gastric tissue samples were collected from 70 patients who underwent resection or endoscopic biopsies with clinically suspicious lesions. All patients preoperatively signed an informed consent permitting the investigative use of the tissues, and this study was approved by the Ethics Committee of the National Healthcare Group (NHG) of Singapore. After biopsies or surgical resections, tissue samples were immediately sent to the laboratory for Raman measurements. After spectral measurements, the tissue samples were fixed in 10% formalin solution and then submitted back to the hospital for histopathologic examination. The tissue specimens comprised five histological groups: normal gastric tissue, dysplastic gastric tissue, cancerous gastric tissue, normal laryngeal tissue, and cancerous laryngeal tissue. Table 3.1 lists the number of samples associated with each of the histological type of gastric and laryngeal tissues. Note that the tissue samples were typically approximately 3 × 3 ×2 mm in size, and the incident laser light with a beam size of 1 mm was focused on the tissue mucosal surface to mimic the *in vivo* clinical measurements.

Table 3.1 Type and number of human tissues collected.

Type of tissue	Histology type	Number of samples
Gastric	Normal	55
	Dysplasia	21
	Cancer	18
Laryngeal	Normal	20
	Cancer	30

3.4 RAMAN MEASUREMENTS

To assess intra-sample variability, multiple Raman measurements (n=5) on each of the tissue were made at different locations of the same samples. Figure 3.4 shows an example of the mean normalized Raman spectrum ± 1 standard deviation (SD) measured from a normal gastric tissue, illustrating spectral intensities variation of 30% about the mean for normal tissue. Overall, the relative Raman peak heights, shapes and positions showed little intra-sample variability for all gastric and laryngeal tissues, indicating the relative homogeneity of tissue samples used in this study [12].

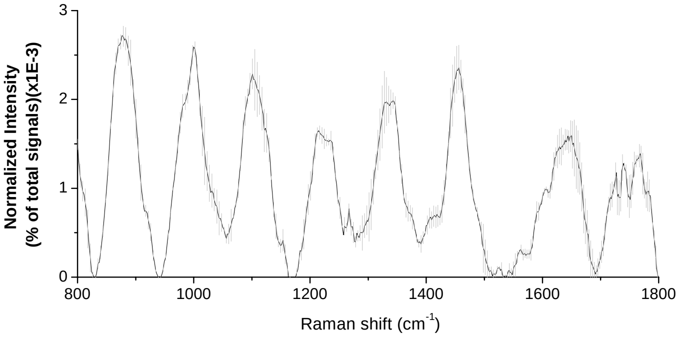

Figure 3.4 Mean normalized gastric Raman spectra (solid line) ± 1 standard deviation (SD) (gray area) obtained from a normal by multiple measurements (n=5) at various locations for each sample. Each spectrum was normalized to the integrated area under the curve to correct for variations in absolute spectral intensity. All spectra were acquired in 5 seconds with 785 nm excitation and corrected for spectral response of the system.

Figure. 3.5 shows the mean normalized Raman spectra ± 1 SD of the different types of gastric and laryngeal tissue. Table 3.2 further lists the tentative biochemical assignments for the 8 major Raman vibrational bands consistently observed in all different tissue types [10, 12]. As can be observed in Figure 3.5, the Raman spectral pattern between the different tissue types (i.e. normal gastric, dysplasia gastric, cancerous gastric, normal laryngeal, cancerous laryngeal) could be very similar, and there are significant inter-sample variability which may obscure the inter-pathology variability. Hence, it is highly desirable to develop robust diagnostic approaches to extract all possible diagnostic information contained in tissue Raman spectra for well correlation with the different tissue types.

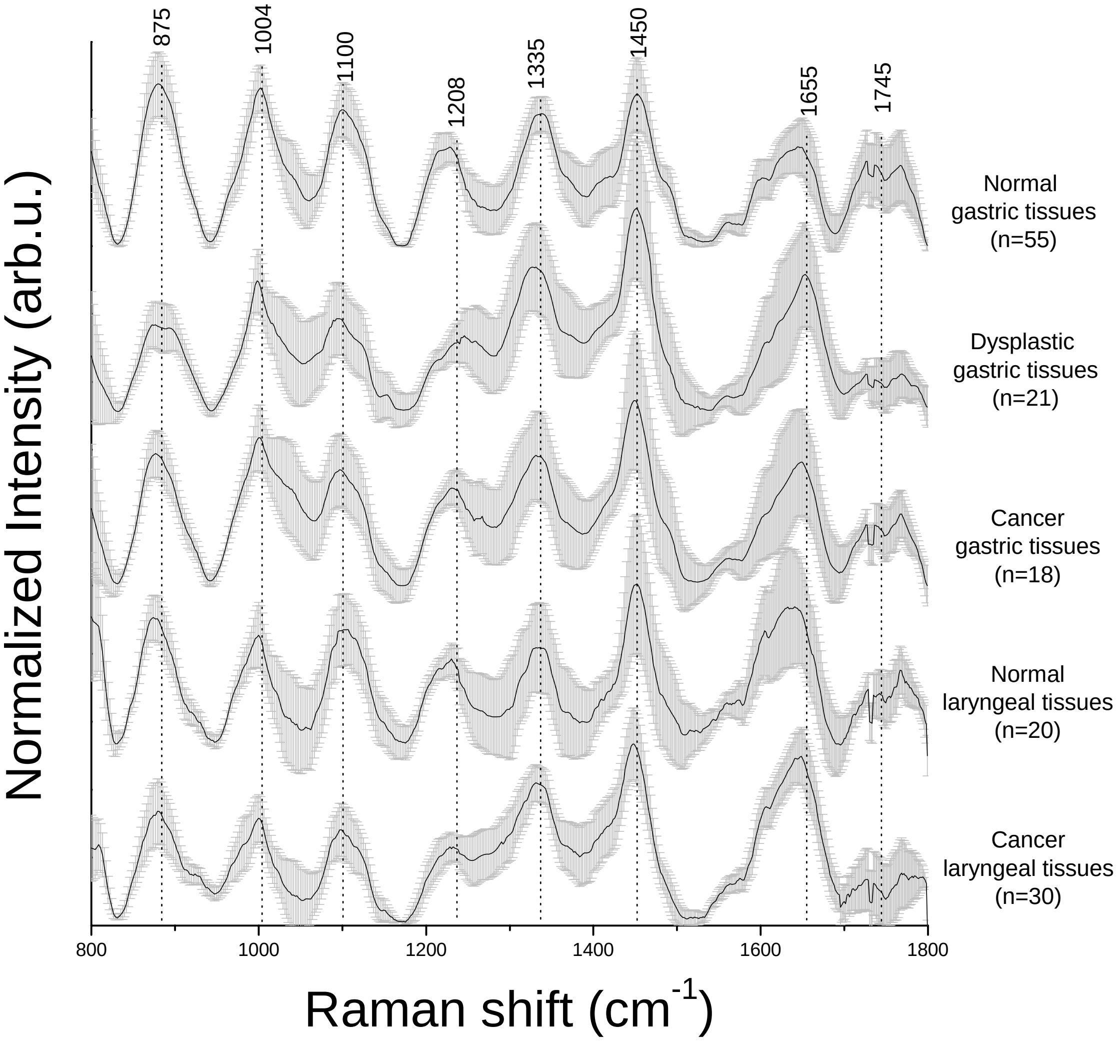

Figure 3.5 Mean Raman spectra ± 1 SD of normal gastric tissues (n=55), dysplastic gastric tissues (n=21), cancerous gastric tissues (n=18), normal laryngeal tissues (n=20), and cancerous laryngeal gastric tissues.

Table 3.2 Tentative assignments of the major Raman peaks identified in gastric and laryngeal tissues [10, 12].

Raman wavenumber (cm^{-1})	Tentative biochemical assignment
875	*v* (C-C) of hydroxyproline
1004	v_s(C-C) ring breathing of phenylalanine
1100	*v* (C-C) of phospholipids
1208	*v* (C-C_6H_5) of tryptophan and phenylalanine
1335	CH_3CH_2 wagging mode of nucleic acids
1450	δ (CH_2) of proteins
1655	*v* (C=O) of amide I, α-helix of proteins
1745	*v* (C=O) of phospholipids

Note: *v*, stretching mode; δ, bending mode

In conclusion, an in-house developed NIR Raman system (Section 3.1) coupled with effective pre-processing techniques (Section 3.2) could be efficiency used to acquire reliable tissue Raman spectrum in the stomach and larynx.

CHAPTER 4

NOVEL DIAGNOSTIC ALGORITHM FOR RAMAN TISSUE CLASSIFICATION: RECURSIVE PARTITIONING TECHNIQUE – CLASSIFICATION AND REGRESSION TREES (CART) FOR GASTRIC CANCER DIAGNOSIS

Despite a falling incidence rate of gastric cancer, it is still the fourth most common malignancy and also the second leading cause of cancer deaths in humans, accounting for 600 000 deaths worldwide [115, 116]. If the tumor is detected early and treated before it has invaded the gastric wall, the survival rate of the patient will increase tremendously [116]. However, early identification and demarcation of such lesions in the stomach can be very difficult to detect by the conventional diagnostic method-white-light endoscope which heavily relies on the visual observation of gross morphological changes of tissue, leading to a poor diagnostic accuracy. Excisional biopsy currently remains the standard approach for cancer diagnosis, but is invasive and impractical for screening high-risk patients who may have multiple suspicious lesions.

Raman spectroscopy which makes use of inelastic light scattering process to capture "fingerprints" of specific molecular structures and conformations of a given tissue or disease state, has shown to be a promising optical diagnostic technique for identifying malignant tissues in various organ [5, 13-15, 25-28]. In order to convert molecular differences subtly reflected in Raman spectra between different tissues types into valuable diagnostic information for clinicians, multivariate statistical techniques have been successfully deployed in developing effective diagnostic algorithms for Raman spectroscopic diagnosis of cancers [1,2,6-9]. Due to the complexities of the biological tissues, PCA, which is able to take into account of the whole range of Raman spectral features of the tissue, has often been applied to simplify the computational complexities for the development of effective classifier algorithms (e.g., LDA, LR) without compromising diagnostic accuracy [1, 7, 13, 30, 34]. However, PCA does not necessary provide the physical meanings of component spectra for tissue classification [29]. Very recently, the classification and regression tree (CART) technique, which bases on the recursive partitioning for generating discriminatory algorithms and possesses potential to uncover interactions among prognostic factors in complex dataset, has received extensive attention in biomedical fields, such as proteomics, genomics and mass spectroscopy [105-110]. For instance, John *et al.* applied both neural network and CART on liver cancer proteomes and found that both algorithms produced equally good predictive ability [106]. Garzotto *et al.* employed CART to identify prostate cancer from normal tissue with the sensitivity of 96.6% [107]. Zhang *et al.* also made use of CART on mass spectral urine profiles to achieve the sensitivity of 93.3% and specificity of 87.0% for separating transitional cell carcinoma of from normal bladder tissue [108]. Despite these successful applications, to date, CART technique has yet been applied to Raman spectroscopy for elucidation of Raman spectra in tissue diagnosis. In this chapter, we explore the feasibility of applying the CART technique to develop effective diagnostic algorithms for differentiation of near-infrared (NIR) Raman spectra between normal and cancer tissue, and to further understanding of molecular changes reflected in Raman spectra of tissue associated with the onset of malignancy in the stomach [14].

4.1 THEORY OF CLASSIFICATION AND REGRESSION TREES

Classification and Regression Tree (CART) is a statistical technique that can selectively employ variables which are of utmost importance from a large number of input variables in databases for binary discrimination [105]. It is implemented by growing a tree structure with a root node containing all the objects, which are then further divided into nodes by recursive binary splitting. The split which gives the best reduction in impurity between the mother group (t_p) and the daughter groups (t_l and t_r) at different nodes of the tree is sequentially selected in the construction of CART tree. The maximization of change of impurity function[20], Δ $i(t)$, at each node, is defined as:

$$\Delta\, i(t) = \underset{x_j \le x_j^R,\, j=1,\ldots..,M}{\arg\max}\ [i(t_p) - P_l i(t_l) - P_r i(t_r)], \qquad (4.1)$$

where x_j represents different variables for different values of j from a total of M variables; x_j^R represents the best splitting value of x_j; $i(t_p)$, $i(t_l)$ and $i(t_r)$ are the impurity functions belonging to the parent node t_p, left child node t_l, and right child node t_r of the parent node, respectively. P_l and P_r are the probabilities of achieving left and right nodes, respectively. CART will search through all possible values of variables for the best splitter at the maximal Δ $i(t)$ ($x_j < x_j^R$). In this study, Gini index is used to determine the impurity, $i(t)$, at each node which forms the criterion for splitting [105].

$$\text{Gini} = 1 - \sum_{j=1}^{c} \left(\frac{n_j}{n} \right)^2, \qquad (4.2)$$

where c is the number of different classes, n is the total number of objects and n_j is the number of objects from class *j* present in the node. Generally, a tree is firstly grown to its maximal size until the terminal nodes are sufficiently small. However, the maximum size tree that is usually overfitted with noise could not generalize well for future dataset. Henceforth, the tree is usually gradually shrunk by pruning away terminal nodes that lead to the smallest decrease in accuracy [67]. For each subtree T, a complexity-misclassification cost function, $R_\alpha(T)$, is generated:

$$R_\alpha(T) = R(T) + \alpha \, |T|, \tag{4.3}$$

where R(T) is the resubstitution misclassification error of T; |T| and α represent the number of terminal nodes and the cost of complexity per terminal node, respectively. During each successive pruning process which resulted in a smaller subtree (the subtree $T' \leq T$ that minimizes $R_\alpha(T')$) with a smaller number of terminal nodes, α will gradually increase. As a result, searching for an optimal tree size (defined with respect to expected performance on the cross-validated dataset) is equivalent to finding the correct α so that the information in the learning dataset is best fit rather than overfit or underfit [105].

Although only one variable would be selected as the best splitter at any node in a CART tree, there would always be a second best variable which may perform nearly as good as the best splitter. These second best variable(s) could be masked by the best splitter(s) and would not appear in the final CART tree. As such, to avoid masking the importance of any variables used in CART, the relative importance of each input variable is assessed based on its importance over all possible nodes and splits by "variable ranking method" [67]. Using the variable ranking method, the importance of a variable X_m is defined as:

$$M\ (X_{m}) = \sum_{t \in T} \Delta I(\tilde{s}_m, t)\ , \qquad (4.4)$$

with $\Delta I\ (\tilde{s}_m, t) =$ max $\Delta I_{C_1}\ (s_m, t)$, which equals the maximal decrease in node impurity for the division of a parent node t into daughter nodes C_1 and C_2 guided by a surrogate split $\tilde{s}_m$. A surrogate split is defined by a surrogate variable. This variable is the second best variable, following the selected variable and giving the second best reduction in impurity of the mother group into the daughter groups. This maximal decrease in node impurity is summed for all the nodes of the optimal subtree, T, to obtain the importance of a variable.

In this study, a 10-fold cross-validation was chosen to select the optimal tree size [14]. The learning dataset is randomly divided into 10 subsets. One of the subsets is used as independent testing dataset, while the other 9 subsets are combined and used as training dataset. The tree growing and pruning procedure is repeated 10 times, each time with a different subset as testing dataset. For each tree size, the resubstitution and cross-validation error are calculated, averaged over all subsets. The misclassification cost obtained for each subtrees on the cross-validation subset is matched with the subtrees of the complete model learning dataset using the α values. The optimal sized tree is proposed to be the tree within one standard error (SE) of the complexity-misclassification rate for the tree with the minimum complexity-misclassification rate [105].

4.2 Development of CART diagnostic algorithm for Raman gastric cancer detection

4.2.1 tissue Raman dataset

Figure 4.1 shows the mean Raman spectra of normal (n=115) and cancer (n=61) gastric tissue in the model learning dataset. 9 prominent Raman peaks were observed in both normal and cancer tissue at the following locations with its respective tentative biochemical bond assignment: ~875 cm^{-1}, C-C stretching modes of proteins; ~1004 cm^{-1}, C-C_6H_5 symmetric ring breathing of phenylalanine; ~1100 cm^{-1}, C-C stretching of phospholipids; ~1230 cm^{-1}, C-N stretching and N-H bending modes of amide III of proteins; ~1265 cm^{-1}, C-N stretching and N-H bending modes of amide III of proteins; ~1335 cm^{-1}, CH_3CH_2 twisting of proteins and nucleic acids; ~1450 cm^{-1}, CH_2 bending of proteins and lipids; ~1655 cm^{-1}, C=O stretching of amide I of proteins; ~1745 cm^{-1}, C=O stretching of phospholipids [14]. The gastric cancer tissues showed higher intensities at 1265, 1450, and 1655 cm^{-1}, while lower at 875, 1004, 1100 and 1745 cm^{-1}, compared to normal tissues. Besides the intensity differences, there are also significant differences in Raman spectral shapes between normal and cancer tissues, which includes broadening of Raman bandwidths at around 1450 and 1655 cm^{-1}, and red shifting of Raman peak at 1655 cm^{-1} for gastric cancer tissues, confirming the potential role of Raman spectroscopy for gastric cancer diagnosis.

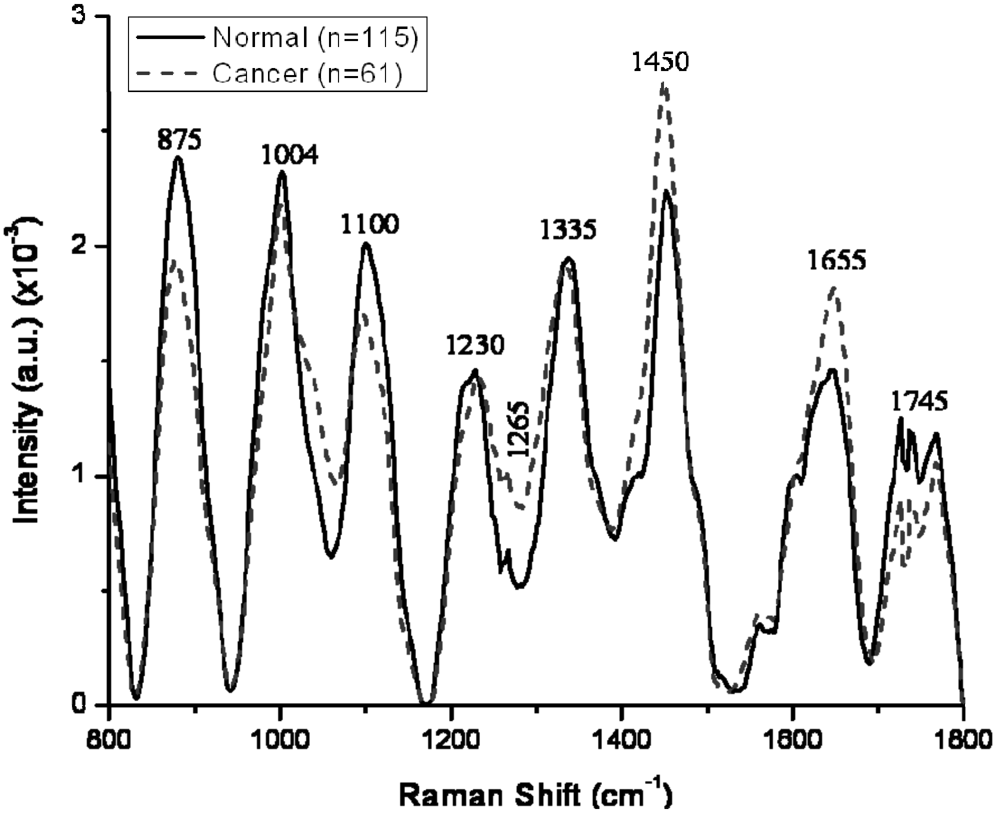

Figure 4.1 Mean Raman spectra of gastric tissues from (a) normal (n=115) and (b) cancer (n=61) in learning Raman dataset.

Table 4.1 lists the mean and SD values of 7 prominent Raman peaks that are diagnostically significant (unpaired two-sided Student's *t*-test, $p < 0.05$) for tissue classification, demonstrating significant overlappings of intensities between gastric normal and cancer tissue. Using each of these individual Raman intensities for diagnosis, the overall sensitivity and specificity varied by 60% – 82% and 60% – 75%, respectively. To further improve tissue classification, CART was subsequently employed to correlate all the diagnostically significant Raman peak intensities with tissue pathologies.

Table 4.1 Statistical characteristics of diagnostically significant Raman peaks (unpaired two-sided Student's *t*-test, p<0.05; 80% of total Raman dataset).

Raman peak (cm^{-1})	**Normal (mean±SD)**	**Cancer (mean±SD)**	**Sensitivity (%)**	**Specificity (%)**	**P-value**
875	0.011 (0.002)	0.009 (0.003)	70.5 (43/61)	74.8 (86/115)	0.0000011
1004	0.011 (0.002)	0.010 (0.002)	62.3 (38/61)	66.1 (76/115)	0.0075823
1100	0.011 (0.002)	0.008 (0.002)	68.9 (42/61)	73.9 (85/115)	0.0000184
1265*	0.003 (0.002)	0.005 (0.002)	65.6 (40/61)	65.2 (75/115)	0.0000002
1450*	0.011 (0.003)	0.013 (0.003)	82.0 (50/61)	60.0 (69/115)	0.0000016
1655*	0.007 (0.002)	0.008 (0.002)	70.5 (43/61)	61.7 (71/115)	0.0000019
1745	0.005 (0.003)	0.004 (0.003)	60.7 (37/61)	61.7 (71/115)	0.0000061

Note:

SD: standard deviation

The symbol * denotes a particular Raman peak intensity with cancer tissue being higher than normal tissue

4.2.2 CART APPLICATION TO THE TISSUE RAMAN DATASET

Figure 4.2 (a, b) shows the relationship of complexity with respect to the misclassification cost and the number of terminal nodes for both cross-validated and resubstitution error after 10-fold cross-validation of the model learning dataset. The misclassification cost for the resubstitution error increases monotonically as the complexity increases with a corresponding decrease in terminal nodes. On the other hand, the misclassification cost for the cross-validated error increases at a slower rate compared to the resubstitution error. A local minimum misclassification cost of 0.1875 is found at complexity of 0.00568 for the cross-validated error. Consequently, the optimal sized tree, defined according to the cross-validated dataset, was chosen to be at complexity of 0.00852 with 13 terminal nodes that is within one SE of the complexity-misclassification cost of the local minimum complexity-misclassification cost.

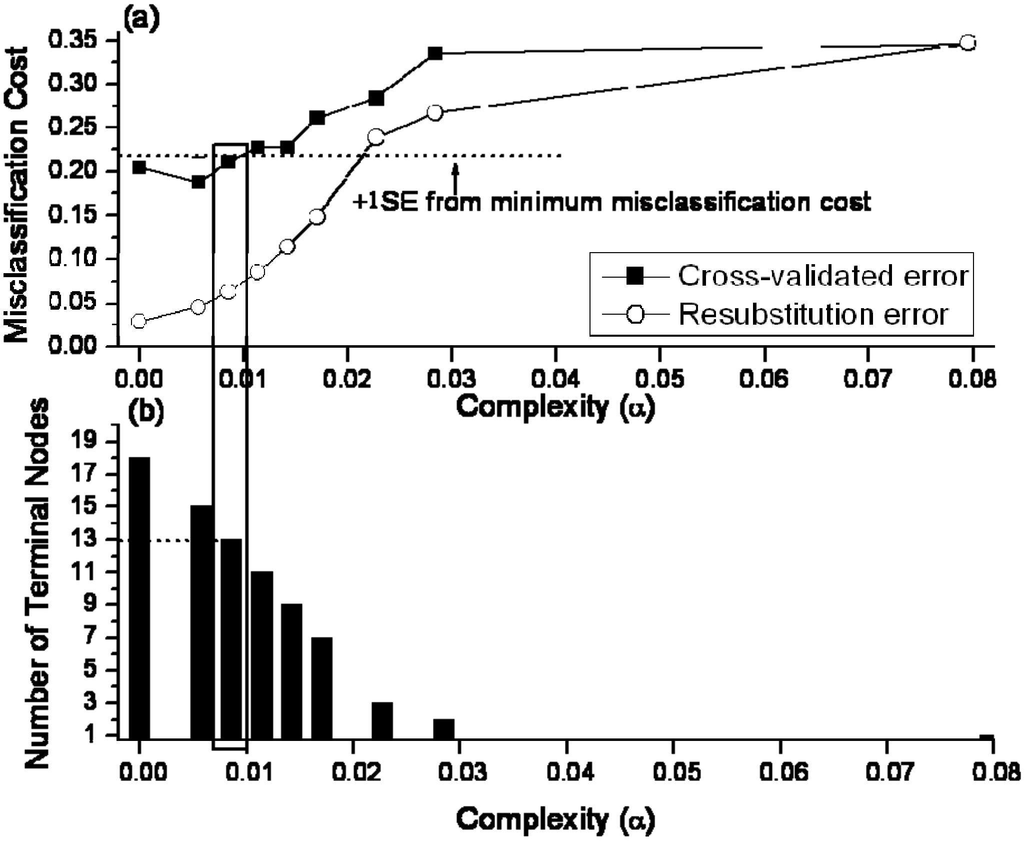

Figure 4.2 Dependence of complexity,α, on (a) misclassification cost nodes for cross-validated error after 10-fold cross-validation, and resubstitution error, and on (b) number of terminal nodes for resubstitution error of the CART model learning dataset. The optimal sized tree was chosen to be at complexity of 0.00852 with 13 terminal nodes within one SE of the complexity-misclassification cost of the local minimum complexity-misclassification cost.

Figure 4.3 displays the CART analysis procedure in a classification model based on the model learning dataset (80% of total dataset). With the CART model, 6 diagnostically significant Raman peaks at 875, 1100, 1265, 1450, 1655, and 1745 cm^{-1} are inter-linked differently to build the following 13 subgroups (designated as either normal or cancer in the terminal subgroups): normal - Group 1, Group 3, Group 6, Group 7, Group 9, Group 11, Group 13; cancer - Group 2, Group 4, Group 5, Group 8, Group 10, Group 12. All these 6 significant Raman peak intensities are combined differently to build the 7 normal and 6 cancer subgroups for best tissue classification.

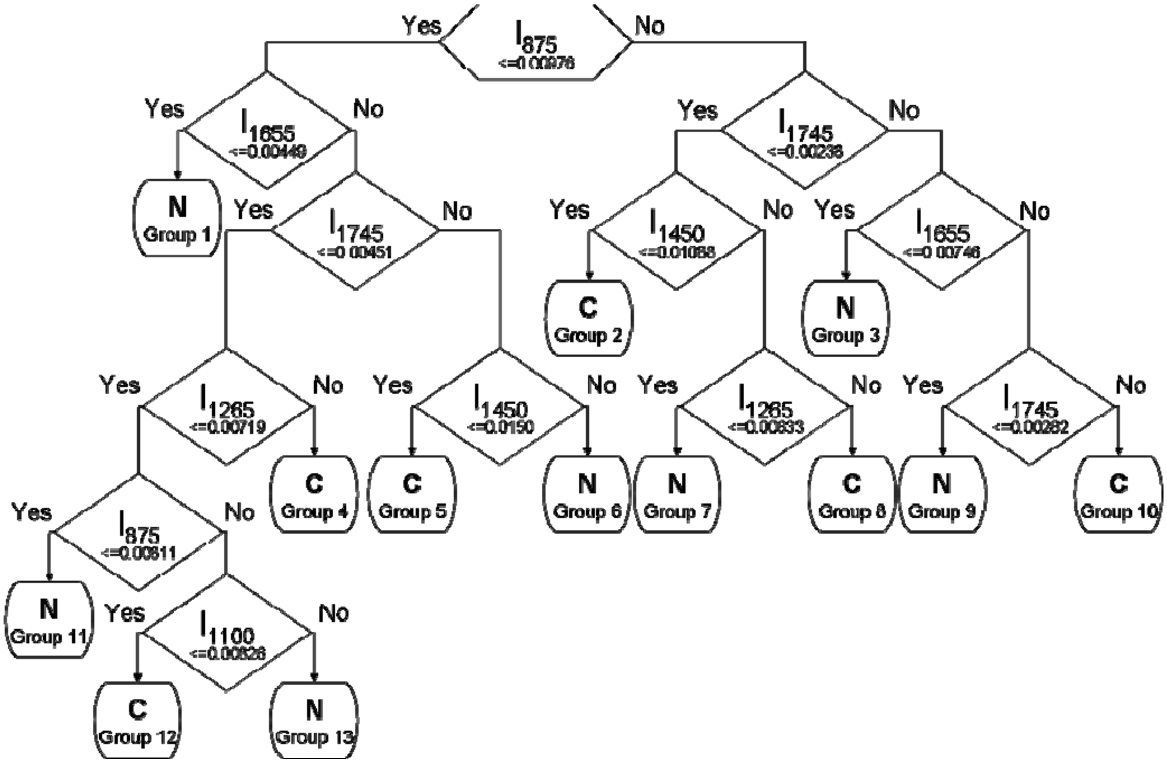

Figure 4.3 The optimal classification tree generated by CART method after 10-fold cross-validation of the model learning dataset by utilizing 6 significant Raman peaks (875, 1100, 1265, 1450, 1655, and 1745 cm^{-1}). The binary classification tree composed of 12 classifiers and 13 terminal subgroups. The decision making process involves the evaluation of if-then rules of each node from top to bottom, which eventually reaches a terminal node with designated class outcome, i.e., normal (N) or cancer (C).

Table 4.2 tabulates the variable ranking of the Raman peaks and the total number of appearing times of different intensity features to generate the CART-based diagnostic model (Fig. 4.3). According to the variable ranking method, the most and least important Raman peaks are found to be located at 875 and 1004 cm^{-1}, respectively. By assessing the final CART-based diagnostic model, the Raman peaks which appeared the most and least number of times are located at 1745 and 1004 cm^{-1}, respectively. For Raman peaks located at 1655, 1265, 1100 and 1450 cm^{-1}, in the order of descending variable rankings, appeared 2, 2, 1 and 2 times, respectively, in the final CART model. As a result, Raman peaks at 875$_{,}$ 1100$_{,}$ 1265$_{,}$ 1450, 1655$_{,}$ and 1745 cm^{-1} are found to be most constructive towards building the final CART-based diagnostic model, with Raman peaks at 875 and 1745 cm^{-1} as the most important variables for tissue classification.

Table 4.2 The variable rankings of all the input Raman peak intensity features (n=7) computed by the CART algorithm, with the corresponding total number of times of the respective feature appearing in the final CART-based diagnostic model.

Raman peak (cm^{-1})	**Number of times appearing in the final CART model**	**# Variable Ranking (Importance)**
875	2	1
1004	0	7
1100	1	5
1265	2	4
1450	2	6
1655	2	3
1745	3	2

Note:

The symbol # denotes a particular Raman peak intensity with variable rankings (1 as the most importance and ranking 7 as the least importance).

4.2.3 Evaluation of the CART algorithm with prospective study

To evaluate the performance of the CART-based diagnostic algorithms for predicting the prospective cases (generalization), a randomly selected validation dataset (20% of total dataset) was used in which 6 important Raman peaks (875, 1100, 1265, 1450, 1655, and 1745 cm^{-1}) were utilized as an input in the final CART-based diagnostic model. Table 4.3 summarizes the classification results of the 2 pathologic groups (normal *vs.* cancer) for both the model learning dataset (after 10-fold cross-validation) (80% of total dataset), and the validation dataset (20% of total dataset). The sensitivity of 90.2% and specificity of 95.7% can be obtained for the model learning dataset, while a predictive sensitivity and specificity of 88.9% and 92.9% can be achieved for the independent validation dataset. The results show that CART-based diagnostic algorithms that utilize the most diagnostically important peaks of Raman spectra are powerful and robust for accurately predicting the tissue types in the prospective new cases.

Table 4.3 Classification results of Raman prediction of the 2 pathological groups with the model learning dataset (80% of total dataset) using the 10-fold cross-validation method, and the validation dataset (20% of total dataset) using a CART-based diagnostic algorithm.

Pathology and Classification		**Raman prediction**		**Total**
		Normal	**Cancer**	
Learning model	**Normal**	110	5	115
	Cancer	6	55	61
(after 10-fold cross validation)	**Sensitivity (%)**	90.2		
	Specificity (%)	95.7		
Testing model	**Normal**	26	2	28
	Cancer	2	16	18
	Sensitivity (%)	88.9		
	Specificity (%)	92.9		

Applying CART technique for classification of tissue Raman spectra, the high predictive diagnostic sensitivity and specificity can be achieved for the independent validation dataset (Table 4.3). Further model validation analysis via 5-fold cross validation reveals that the cross validated result is almost similar to the independent validation dataset classification results and the PCA-LDA approach (PCA-LDA is slightly higher compared to CART) (data not shown) [105]. These further reinforced that the CART-based diagnostic algorithms generated are robust and powerful for tissue diagnosis and characterization by Raman spectroscopy. Besides the ability for tissue classification, CART diagnostic model could also provide a novel way for better understanding of the relationship between the disease-related biochemical changes of Raman spectra and tissue pathologies. We use the CART technique to evaluate how different Raman molecular information are correlated with tissue types by analyzing how the different Raman peaks are inter-linked to optimally form different subgroups for tissue classification. For instance, in Figure 4.3, it is found that Group 5 (cancer subgroup) has the highest probability of incidence, followed by Group 3 which is a normal subgroup for both the model learning and validation datasets. In Group 5, CART indicates that cancer gastric tissues are associated with a relative increase in Raman peak intensities at 1655 and 1745 cm^{-1}, while a decrease at 875 and 1450 cm^{-1}. These results are in fact in agreement with the reports on the decrease of Raman intensity ratio at 1655 cm^{-1} to 1455 cm^{-1} associated with malignancies in the cervix and lung [5, 20]. CART also notes that the Raman peaks at 875 and 1745 cm^{-1} representing collagen and phospholipids, respectively, appear to be significantly correlated with the Raman peaks at 1450 and 1655 cm^{-1} for identifying the cancer subgroup (Group 5). Conversely, the Raman peaks at 875, 1655 and 1745 cm^{-1} utilized to construct cancer subgroup (Group 5) could be employed for identifying the normal subgroup (Group 3). CART also indicates that compared to cancer tissue, normal gastric tissues tend to be related with high collagen contents (Raman peak at 875 cm^{-1}) in extracellular matrix, high lipid contents (Raman peak at 1745 cm^{-1}) present in both extracellular matrix and cytoplasm, and a lower histones content (Raman peak at 1655 cm^{-1}) in the nucleus. Further investigation on other subgroups shows that heterogeneous molecular changes may occur in tissue, enabling cancer subgroups to be distinguished from normal. For example, the collagen content typically decreases with malignancy, but there are quite a number of other cancer subgroups (Group 2, Group 8 and Group 10) to be associated with increase in collagen content. These subgroups are accompanied by either less phospholipids or higher histones content which could enable them to be distinguished from normal tissue. The above CART-Raman analysis results indicate that most biochemical/biomolecular information from tissue and cells are essential for tissue discrimination, and the CART-based diagnostic model is able to partition different subgroups based on different compositions of Raman molecular information for separating gastric cancer from normal. Therefore, the CART-Raman algorithm may provide new insights into the biochemical/biomolecular changes associated with malignant transformation [12].

In summary, the CART technique was first introduced and implemented to develop effective diagnostic algorithms for classification of Raman spectra between normal and cancer gastric tissues. This work shows that NIR Raman spectroscopy in combination with powerful CART algorithms has potential to provide an effective and accurate diagnostic means for cancer diagnosis in the gastric. Further studies to investigate the use of a more powerful diagnostic algorithm which is primarily based on the recursive partitioning technique has also been carried out and will be elaborated more in details in the next chapter.

CHAPTER 5

IMPROVED RECURSIVE PARTITIONING TECHNIQUE FOR RAMAN TISSUE DIAGNOSIS: AN ENSEMBLE APPROACH – RANDOM FORESTS FOR IDENTIFICATION OF LARYNGEAL MALIGNANCY

Laryngeal squamous cell carcinoma is one of the most common malignancies in the head and neck and also the sixth most common malignancies worldwide [117, 118]. In Southeast Asia, the rates of incidence and mortality due to laryngeal cancer are notably higher than many parts of the world [119]. Early identification and accurate demarcation of such malignant lesions coupled with effective therapy (particularly for partial laryngectomy, transoral excision, photodynamic therapy and radiotherapy) are crucial to improving the survival rates of the patients [118, 120]. However, identification of early malignancy in the larynx can be very challenging even for the very experienced otolaryngologists with the aid of conventional diagnostic techniques such as white-light endoscope (e.g. microlaryngoscopy, transnasal esophagoscopy); since the white-light endoscopy heavily relies on the visual observation of gross morphological changes of pathologic tissues, leading to a poor diagnostic accuracy. This limitation poses a great challenge in an ear-nose-throat (ENT) clinic and, therefore, there is of significant clinical values for the development of new endoscopic diagnostic techniques to complement white-light endoscopy for improving diagnostic performance of malignancy detection in the head and neck.

In the past decade, optical spectroscopic methods, such as light scattering spectroscopy, fluorescence spectroscopy, and Raman spectroscopy, have been comprehensively investigated for cancer and precancer diagnosis and evaluation [1-10, 21, 24]. Particularly, NIR Raman spectroscopy has received great interest for optical diagnosis of malignant tissues in various organs [1, 5, 7], including the larynx [10, 21, 24]. In order to convert molecular differences subtly reflected in Raman spectra between different tissues types into valuable diagnostic information for clinicians, different algorithms ranging from empirical approaches (e.g. intensity ratio algorithms) to sophisticated multivariate statistical techniques (e.g. PCA, LDA, SVM, CART, PCA-ANN) have been actively explored for classification of Raman spectra for precancer and cancer diagnosis [1, 11, 14, 15, 62]. For instance, Lau *et al.* employed a fiber-optic rapid NIR Raman spectroscopy together with PCA-LDA techniques to identify laryngeal papillomatosis and squamous cell carcinoma with accuracies of 91 and 87%, respectively [21].

In recent years, the ensemble-based statistical techniques [121], such as random forests algorithm [113] have gained increased attention due to their capability of effectively combining multiple classifiers for data classification in biomedical areas [122-129]. For example, Jiang *et al.* introduced random forests to distinguish real microRNA precursors from pseudo ones with a sensitivity of 98.2% and specificity of 95.1% [122]. Wu *et al.* demonstrated that random forests algorithm could yield the highest diagnostic accuracy among other diagnostic algorithms such as LDA, quadratic discriminant analysis (QDA), k-nearest neighbor classifier, bagging and boosting classification trees and SVM, for discriminating ovarian cancer from normal serum samples using mass spectroscopy [123]. Despite of these successful investigations, the utility of ensemble-based algorithms on Raman spectroscopy for biomedical diagnosis has yet been reported in the literature. In this chapter, we extend our previous study on the use of CART for Raman tissue diagnosis, to further explore the feasibility of developing effective diagnostic algorithms based on an ensemble of classification trees-random forests technique for differentiation of NIR Raman spectra between normal and cancer tissue in the larynx [10].

5.1 RANDOM FORESTS THEORY

Figure 5.1 illustrates the procedure for generating the random forests algorithm for classification of tissue Raman spectra. The random forests is fundamentally an ensemble of unpruned classification trees [113, 125]. Different trees are grown via different bootstrap sampling of the original dataset. In each of these growing trees, the random forests algorithm selects a fixed-size random subset of all variables to search for the best split, defined by the Gini Criterion, at every node encountered. The final diagnostic random forests is then constructed by combining multiple classification trees through the use of majority voting technique [113].

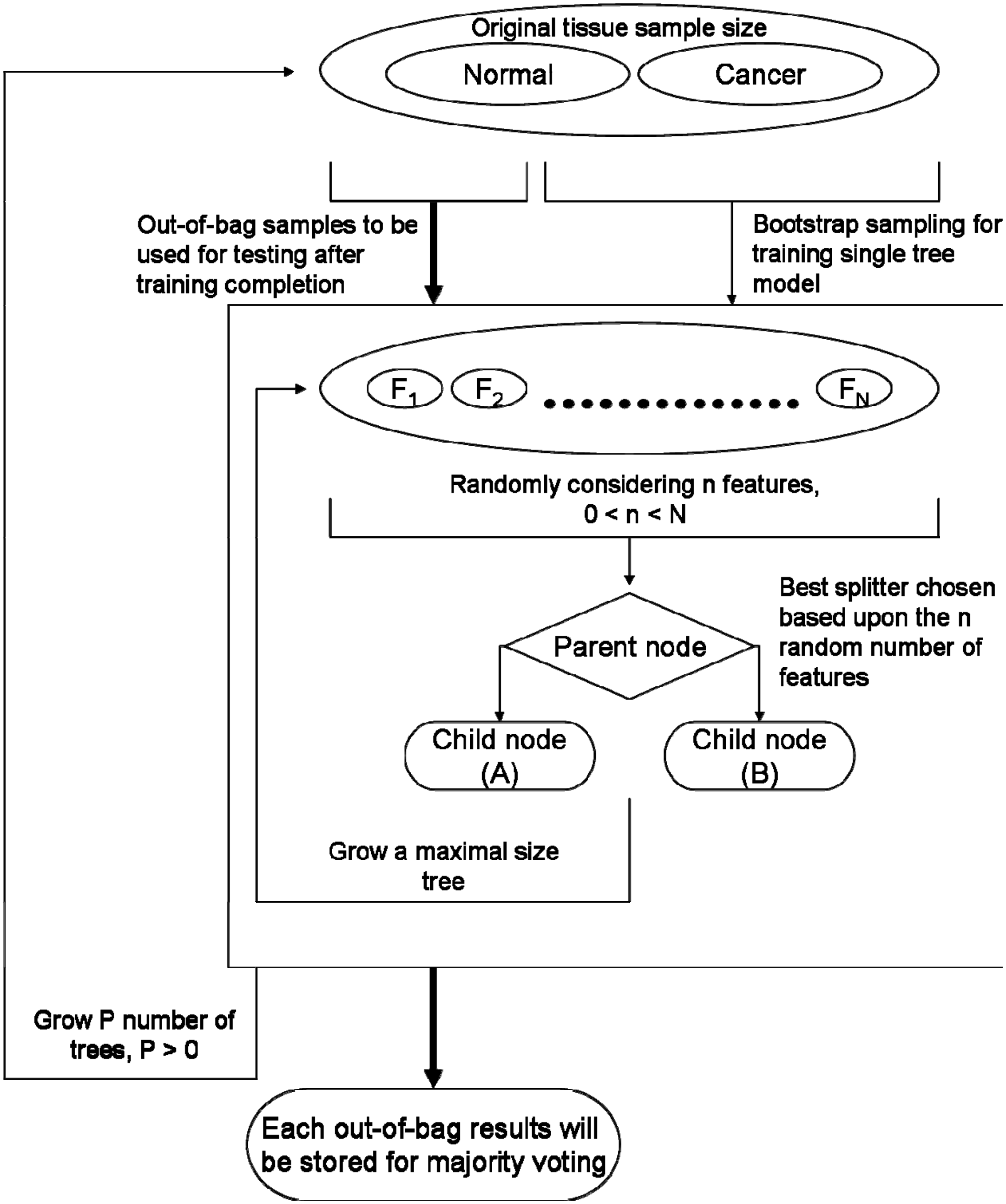
Original tissue sample size
Normal
Cancer
Out-of-bag samples to be used for testing after training completion
Bootstrap sampling for training single tree model
F1
F2
FN
Randomly considering n features, 0 < n < N
Best splitter chosen based upon the n random number of features
Parent node
Child node (A)
Child node (B)
Grow a maximal size tree
Grow P number of trees, P > 0
Each out-of-bag results will be stored for majority voting

Figure 5.1 Illustration of procedures for generating the random forests algorithm for tissue classification.

Since random forest employed bootstrap sampling, performance of this ensemble classifier could be assessed with the prediction error for the objects left out in the bootstrap procedures. This is also known as the out-of-bag estimation (internal cross-validation) [113]. Specifically, after each process of growing a maximal tree with a particular bootstrap dataset, the remaining data which were not used in the tree construction would constitute the "out-of-bag" dataset and be used to assess the respective tree prediction performance. The final predicted class of the data is calculated by majority vote of all the "out-of-bag" predictions on the dataset. Hence, the estimate of the error rate (ER) for the random forests classification algorithm is computed:

$$\mathrm{ER} \approx \mathrm{ER}^{\text{out-of-bag}} = \mathrm{N}^{-1} \sum_{i=1}^{N} I(Y^{out-of-bag}(X_i) \neq Y_i) \tag{5.1}$$

where N is the size of forest (i.e., number of trees). $I(Y^{out-of-bag}(X_i) \neq Y_i)$ represents the misclassification indicator function of the ensemble prediction in which $Y^{\textit{out-of-bag}}$ is based on each training dataset and X_i is on the respective out-of-bag dataset Y_i. Besides the determination of data classification (either true positive or true negative), the voting outcome can also generate probability of the identification being true positive (cancer) or true negative (normal) for each data [124].

In this study, the number of variables to be randomly selected at each node in a tree for classification is approximately derived from the square root of the total number of available variables; the optimal size of the forests (i.e., the number of trees and the number of bootstrap sampling) would correspond to the smallest forest size which the out-of-bag error estimation stabilizes at a constant value with enlargement of forest size

[125]. Based on the predicted probability for each data from the optimal forest size constructed from the training dataset, ROC curve was also generated by successively changing the thresholds to determine the probability of prediction being cancer or normal for all tissue Raman data.

When a variable that substantially contributes to prediction performance is replaced with random noise, the performance of the prediction will be noticeably degraded [113, 125]. Conversely, if a variable is irrelevant to the prediction performance, replacing it with random noise only has trifle effect on the performance [125]. To assess the importance of different variables, the permutation accuracy importance measure is utilized in the random forest algorithm as shown in Figure 5.1. For each tree, the prediction accuracy on the out-of-bag dataset is first recorded. After which, the values of a specific variable is randomly permutated in all the trees, and the new predictions based on the same out-of-bag dataset will also be recorded. Consequently, the permutation accuracy importance measure of a specific variable will be equivalent to the difference between the error rate for the new and original results based on the same out-of-bag dataset of a specific predictor variable, normalized by the standard error [113]. Henceforth, the larger the difference, the more important the predictor variable contributes towards the construction of the random forest [113, 125]. In this work, the variable importance algorithm defines a variable importance value of 1 belonging to the most important variable, whereas a variable importance of 0 belonging to the least important variable.

5.2 Evaluation of Random forests diagnostic algorithm for Raman laryngeal cancer diagnosis

5.2.1 Laryngeal tissue Raman dataset

Figure 5.2 shows the mean normalized Raman spectra of normal (n=70) and cancer (squamous cell carcinoma) (n=117) laryngeal tissue. Prominent Raman peaks are observed in both normal and cancerous laryngeal tissue, which are located at around 875 cm^{-1} (C-C stretching of hydroxyproline), 935 cm^{-1} (C-C stretching mode of proline and valine in α-helix conformation), 1004 cm^{-1} (C-C symmetric stretch ring breathing of phenylalanine), 1100 cm^{-1} (C-C stretching of phospholipids), 1208 cm^{-1} ($C-C_6H_5$ stretching mode of tryptophan and phenylalanine), 1265 cm^{-1} (C-N stretching and N-H bending modes of amide III of proteins), 1335 cm^{-1}, (CH_3CH_2 twisting of proteins and nucleic acids), 1450 cm^{-1} (CH_2 bending of proteins and lipids), 1552 cm^{-1} (C=C stretching mode of tryptophan), 1582 cm^{-1} (C=C bending mode of phenylalanine), 1601 cm^{-1} (C=C in-plane bending mode of phenylalanine), 1655 cm^{-1} (C=O stretching of amide I of proteins), and 1745 cm^{-1} (C=O stretching of phospholipids)[6,7,9-10,13-15,34], respectively [10]. The intensity differences between the two tissue types are remarkable. For example, cancer tissue show higher intensities at 935, 1265, 1335, 1450, and 1655 cm^{-1}, while lower at 875, 1004, 1100, 1208, 1560, 1582, 1601 and 1745 cm^{-1}, compared with normal tissues. This suggests that there is an increase or decrease in the percentage of a certain type of biomolecules relative to the total Raman-active constituents in cancer tissue. There are also obvious changes of Raman peak positions and bandwidths in the ranges of 1100-1190 cm^{-1}, 1200-1500 cm^{-1}, and 1500-1800 cm^{-1} which are related to the

C-C stretching of phospholipids, the amide III and amide I of proteins, CH_3CH_2 twisting of proteins/nucleic acids, and C=C stretching of phospholipids, respectively, for malignant laryngeal tissue. The differences in Raman spectra between normal and cancer tissue demonstrate the utility of Raman spectroscopy for laryngeal cancer diagnosis and detection.

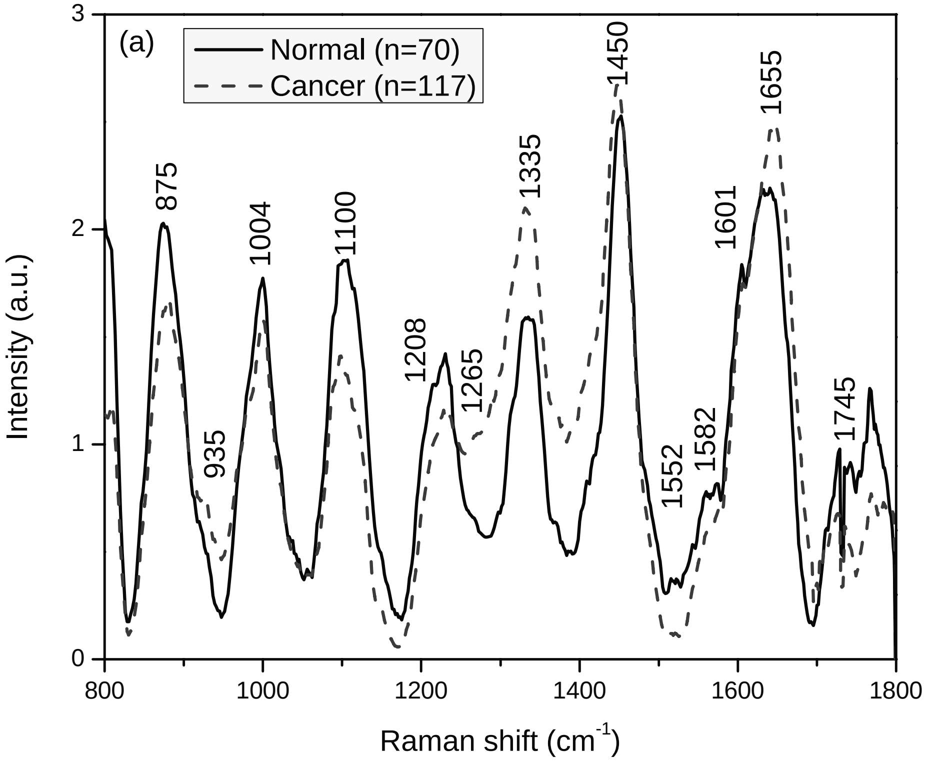

Figure 5.2 Comparison of the mean normalized Raman spectra of normal (n=70) and cancer (n=117) laryngeal tissue.

5.2.2 Employment of random forests to the tissue Raman dataset

We employ the ensemble of recursive partitioning algorithms approach (i.e., random forests) by incorporating the entire Raman spectra (each Raman spectrum ranging from 800-1800 cm^{-1} with a set of 544 intensities) to determine the most diagnostically significant Raman features for improving tissue analysis and classification. In this study, the number of variables tested for each split was set to 23 (√544) and the error rate of different forest size from 1-1500 trees was investigated. Fig. 5.3 (a) shows the relationship of error rate with respect to the different number of trees. The optimal forest size was chosen to be at 973 when the error rate stabilizes to about 0.107 after the forest size is more than 972 trees. To evaluate the performance of this optimal diagnostic algorithm, ROC curve (Fig. 5.3 (b)) is also generated by constructing different threshold levels from the probability associated with each data after the majority voting. The integration areas under the ROC curve is 0.964 for the forest size of 973 trees, proving the efficacy of this random forest algorithm derived for laryngeal cancer diagnosis. The results show that random forests-based diagnostic algorithm is robust for laryngeal cancer diagnosis.

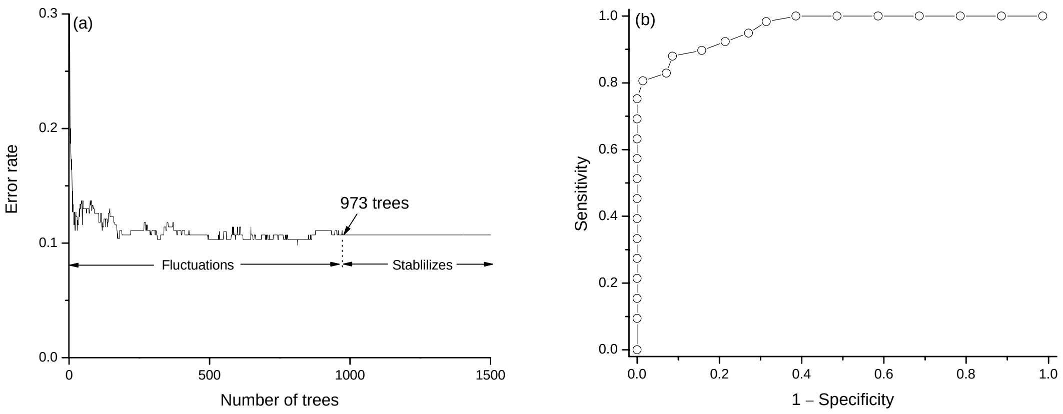

Figure 5.3 (a) Different error rates belonging to different sizes of the random forests (i.e., different number of trees) after the voting process on all the tissue Raman spectra. Due to the "strong law of large number", the error rate stabilizes to 0.107 when the forest has more than 972 trees, highlighting that the random forests algorithm does not overfit. Note that each of the individual trees is grown to the maximal size and left unpruned. (b) ROC curve of tissue classification belonging to the final optimal random forests tree size of 973 with an AUC of 0.964, illustrating the diagnostic ability of Raman spectroscopy and random forests algorithm to identify cancer from normal laryngeal tissue.

Figure 5.4 shows the variables importance plot based upon the construction of 973 trees associated with cancer transformation. The permutation accuracy importance measure reveals that Raman intensities at 820, 850, 870, 938, 1004, 1085, 1104, 1123, 1172, 1208, 1240, 1314, 1335, 1370, 1421, 1490, 1552, 1576, 1601, 1672 and 1745 cm^{-1} are among the most important variables in the laryngeal tissue Raman spectra for distinguishing cancer from normal (at 95% confidence interval). Table 5.1 lists the tentative biochemical representations assigned to the significant Raman intensities. Raman intensities at these Raman wavenumber positions of the original dataset were subsequently subjected to unpaired two-sided Student's *t*-test for investigating the statistical significances ($p<0.05$) between normal and cancerous tissue [10]. Raman peak intensities at 938, 1314, 1335, 1370, 1421, and 1672 cm^{-1} were significantly higher for cancerous laryngeal tissue as compared to normal, while Raman intensities at 820, 850, 870, 1004, 1085, 1104, 1123, 1172, 1208, 1240, 1490, 1552, and 1745 cm^{-1} were significantly lower for cancerous laryngeal tissue. Conversely, Raman intensities at 1576 and 1601 cm^{-1} were found not to be statistically significant ($p=0.19$ and $p=0.29$) for cancer diagnosis. The results indicate that most prominent Raman peaks, particularly in the spectral range of 820-1745 cm^{-1} representing proteins, lipids and nucleic acids are the most important variables for tissue grouping. Hence, random forests-based permutation accuracy importance measure algorithm provides a novel way to identify variables which are important towards creating the final diagnostic model for tissue diagnosis and characterization.

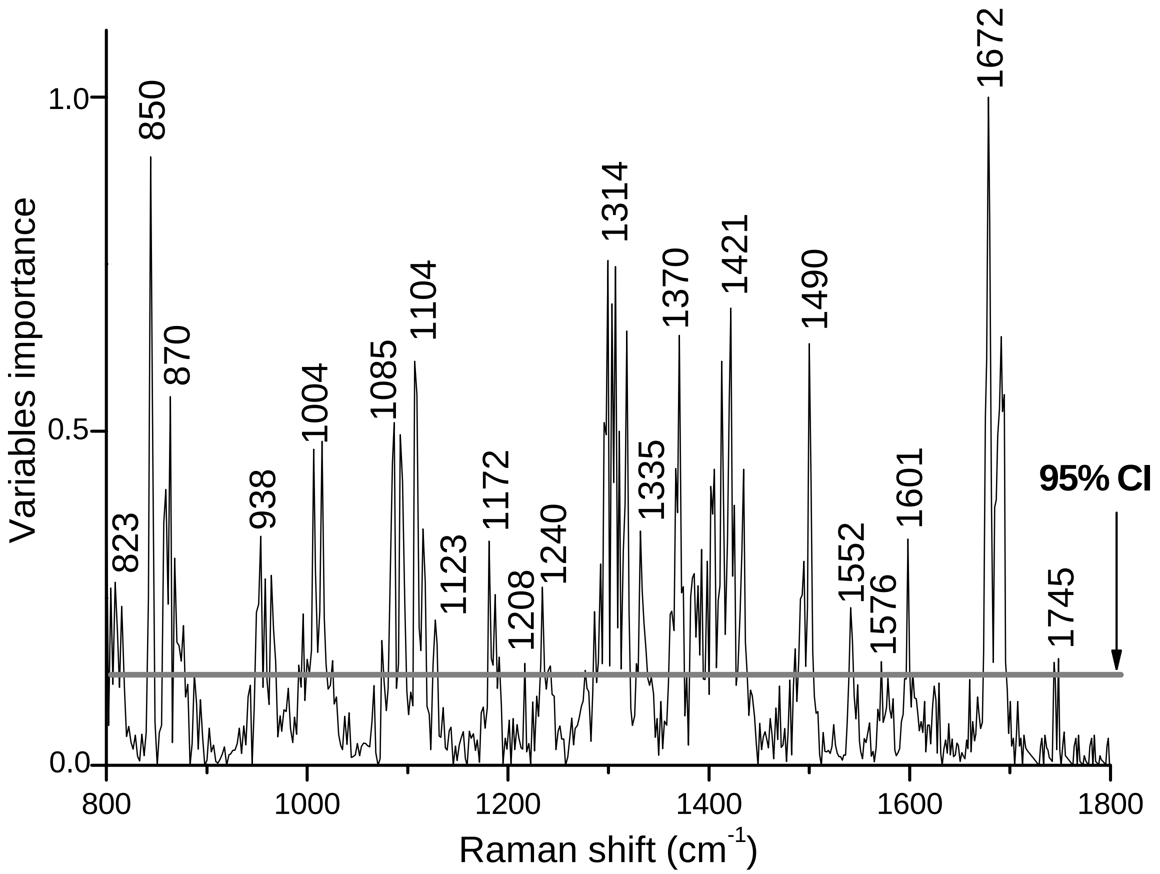

Figure 5.4 Variables importance plot for the Raman spectral region 800-1800 cm^{-1} generated from random forests size of 973 trees which was used for discrimination of cancer from normal laryngeal tissue. The variable importance algorithm defines the most important variable as 1, whereas the least important variable as 0. Major Raman spectral features above the bold grey line (95% confidence interval, 13.7) are identified and listed in Table 5.1.

Table 5.1 Tentative assignments of the Raman peaks identified in laryngeal tissue (Fig. 5.4, variables importance plot), mean intensity changes (increase +/decrease −) of cancer with respect to normal, and p-values of unpaired two-sided Student's *t*-test on Raman peak intensities of normal and cancer laryngeal tissue.

Raman wavenumber (cm^{-1})	**Tentative assignment**	**Mean change (Cancer-Normal)**	*p*-value
820	Out-of-plane, ring breathing of tyrosine	−	3E-2
850	δ (CCH) ring breathing mode of tyrosine, Polysaccharide	−	3E-3
870	*v* (C-C) of hydroxyproline	−	8E-5
938	*v* (C-C) in α conformation of proline and valine	+	4E-7
1004	v_s(C-C) symmetric ring breathing of phenylalanine	−	4E-2
1085	*v* (C-N) of proteins (lipids mode to lesser degree)	−	2E-4
1104	*v* (C-C) of phospholipids (in gauche conformation)	−	3E-8
1123	*v* (C-N) of proteins	−	9E-11
1172	δ (C-H) of tyrosine	−	1E-6
1208	*v* ($C\text{-}C_6H_5$) of tryptophan and phenylalanine	−	8E-8
1240	*v* (C-N), δ (N-H) amide III, α-helix of proteins	−	7E-3
1314	CH_3CH_2 twisting mode of proteins	+	3E-14
1335	CH_3CH_2 wagging mode of proteins and nucleic acids (DNA-purine bases)	+	6E-11
1370	Adenine, thymine, guanine	+	2E-10
1421	δ (CH_2), δ (CH_3) of proteins, δ (CH_2) scissoring of phospholipids, deoxyriboses	+	3E-2
1490	*v* (C-N) in-plane vibration of guanine (Adenine to lesser degree)	−	2E-8
1552	*v* (C=C) of tryptophan	−	6E-3
1576	δ (C=C) of phenylalanine	−	6E-2
1601	δ (C=C) (in-plane) of phenylalanine	−	3E-1
1672	*v* (C=O) of amide I in β-helix conformation of proteins, C=C lipid stretch	+	4E-7
1745	*v* (C=O) of phospholipids	−	9E-8

Note: *v*, stretching mode; v_s, symmetric stretching mode; δ, bending mode.

Figure 5.5 displays the probabilistic classification results based on the random forests technique together with the leave-one sample (i.e., all spectra associated with the sample)-out, cross validation method. The random forests diagnostic algorithm yields the diagnostic sensitivity of 88.0%, and specificity of 91.4% for separating cancer from normal laryngeal tissues. The results show that Raman spectroscopic technique combined with the random forests diagnostic algorithm is robust and powerful for cancer diagnosis.

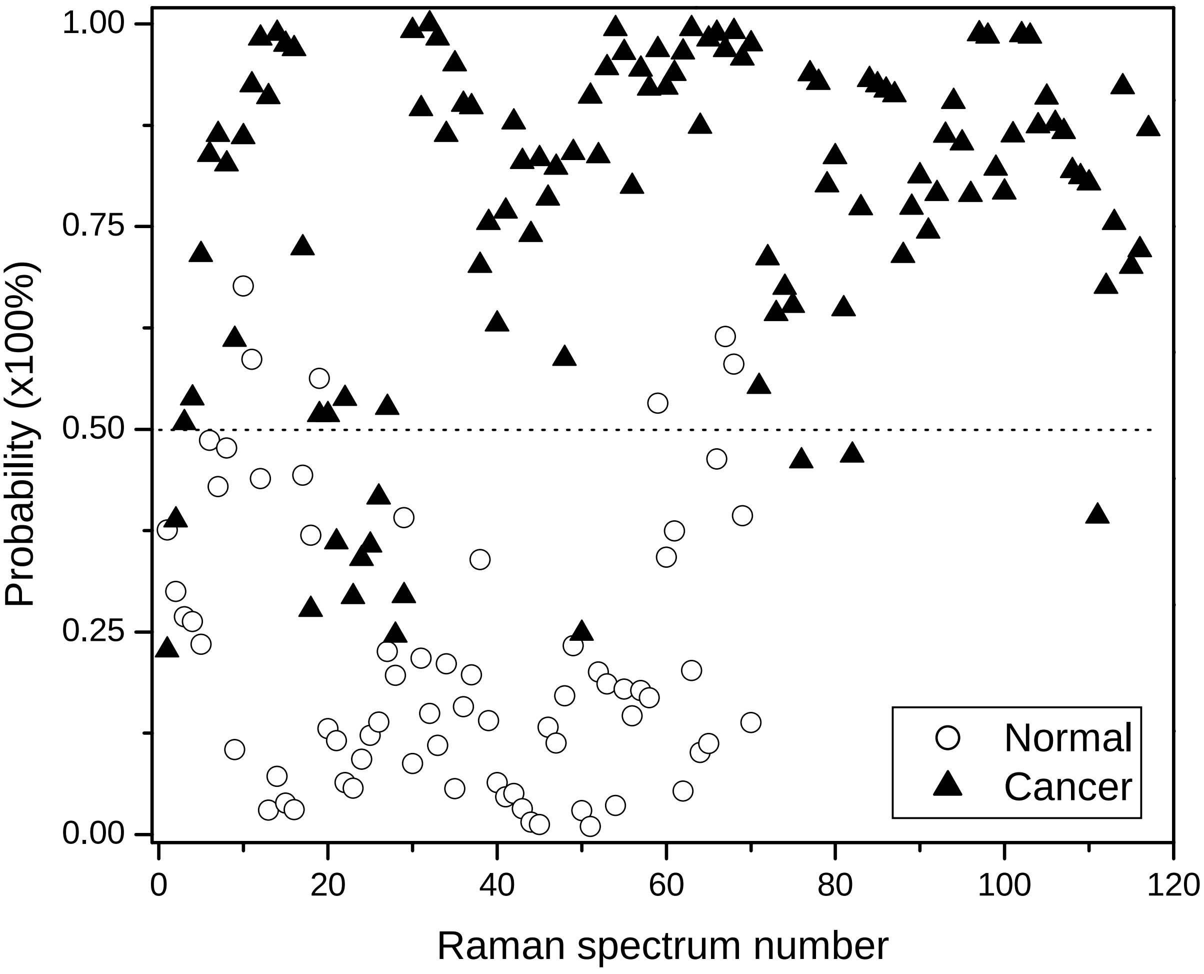

Figure 5.5 Scatter plot of the generated probabilistic scores belonging to the normal and cancer categories using the random forests technique together with leave-one sample-out, cross validation method. The separate line yields a diagnostic sensitivity of 88.0%

(103/117) and specificity of 91.4% (64/70) for differentiation between normal and cancer laryngeal tissue.

Raman spectroscopy holds great promise for molecular diagnosis of different tissue types in biomedicine [1,5, 7, 10]. To harvest the great wealth of biomolecular information of complex tissue contained in Raman spectra for reliable diagnostic applications, different multivariate statistical algorithms have been extensively investigated [1, 6, 11, 14, 15, 62]. In this study, we developed a robust random forests algorithm using the ensemble of classification trees for Raman spectroscopic diagnosis of laryngeal cancer. The result shows that the developed random forests algorithm based on the tissue Raman dataset yields a predictive diagnostic sensitivity of 88.0% (103/117), specificity of 91.4% (64/70), and overall accuracy of 89.3% (167/187) for laryngeal cancer detection. Since the original motivation for the development of random forests algorithm was to enhance the classification accuracy of single classification tree (i.e., CART technique), for comparison purposes, we also apply the CART technique combined with the leave-one sample-out, cross-validation method on the same tissue Raman dataset. We found that a predictive sensitivity of 82.9% (97/117), specificity of 81.4% (57/70) and accuracy of 82.4% (154/187) can be obtained for using CART algorithm on laryngeal cancer diagnosis. Clearly, the random forests diagnostic algorithm can give a fairly higher level of diagnostic accuracy compared to the CART model, indicating that an ensemble-based algorithm is more powerful than a single classifier technique. This is probably due to the fact that effective ensemble-based algorithms generally possess the properties of low biasness, low variance, and low correlations of classifiers (diversity) for data classification [130, 131]. Also, the combination of unpruned trees inherently preserves low biasness and variance, and the limited number of predictors generated at each node in a tree also ensures that correlations among the resultant classification trees are very small [125], leading to the improved classification performance of the ensemble-based technique (random forests) for cancer identification as compared to the CART algorithm. Furthermore, through the employment of the majority voting procedure for the ensemble of multiple tree classifiers, the random forests diagnostic model could also generate probability of acquired spectra belonging to cancer or normal group (Fig. 5.5). This facilitates the clinicians to effectively assess the accuracy of the predicted tissue types or pathologies associated with NIR Raman spectroscopy technique. On top of these, the random forests algorithm possesses an inherent characteristic that the error rate can converge to an asymptotic value due to the "strong law of large number" thereby enabling this diagnostic algorithm to be resistant towards overfitting for tissue classification (Figure 5.3(a)) [132]. These further confirm that the ensemble-based random forests diagnostic technique together with NIR Raman spectroscopy is robust and powerful for laryngeal tissue diagnosis and characterization.

Currently, most of the diagnostic algorithms (e.g. PCA-LDA, SVM, logistic regression,) employed in Raman spectroscopy diagnosis of diseased tissue could not adequately furnish the clinicians with physical meanings of diagnostic features derived for tissue characterization [1, 6, 11, 14, 15, 62]. Hence, the development of robust algorithms which not only produce a high predicted diagnostic accuracy, but also provide useful biomolecular diagnostic information from the high dimensional Raman spectral datasets is highly desirable. As the random forests algorithm provides the permutation accuracy importance measure [133] to uncover the disease-related biochemical changes of Raman spectra that can correlate with tissue pathologies, we have constructed the robust random forests diagnostic model for distinguishing laryngeal cancer [10]. We found that the ensemble-based classification trees diagnostic algorithm could be used to select distinctive spectral regions that are optimal for tissue differentiation. We have identified twenty-one significant Raman features related to particular biochemical and biomolecular changes (e.g. proteins, lipids, nucleic acids, and carbohydrates) that are associated with cancer transformation in the larynx (Fig. 5.4 and Table 5.1). For instance, the Raman intensity at 1672 cm^{-1} has been found to be one of the most important Raman features in the construction of random forests-based diagnostic model for discriminating malignancy from normal laryngeal tissue (Fig. 5). The finding is consistent with Huang *et al.*'s report on the observation of a shoulder band at 1668 cm^{-1} in malignant lung tissue [5]. This suggests that laryngeal cancerous transformation could be related with an increase in the relative amounts of proteins in the β-pleated sheet or random coil conformation, and lipids [5, 10]. In addition, the Raman intensities at 820, 850 and 1172 cm^{-1} that are presumably ascribed to tyrosine [10] are also found to be significantly lower for cancer, indicating the decrease of tyrosine with laryngeal malignancy. These findings are in agreement with Verschuur *et al.*'s study [134] in which the activities of protein tyrosine kinases were found to be lower for head and neck squamous cell carcinoma. On top of this, Raman intensities at 938, 1004, 1208, 1552, 1576 and 1601 cm^{-1} that are representative of different amino acids such as proline, valine, phenylalanine and tryptophan, and Raman intensities at 1085, 1123, 1240, 1314 cm^{-1} for proteins in general are also found to play pivotal roles towards tissue classification. This may signify significant changes of different proteomic activities (e.g. enzymatic, hormones and etc) with malignancy which could be an indication of increase of mitotic activities in cancerous cells [10]. Furthermore, Raman intensity at 875 cm^{-1} has been found to decrease significantly with malignancy, signifying a reduce in the percentage of collagen contents relative to the total Raman-active components in the stroma layer of cancer tissue. This observation coincides with the report in literature that cancerous cells proliferate, invade into underlying layer, and express as a class of metalloproteases leading to a decrease in the amount of collagen level [93]. Besides, the thickening of the epithelium associated with cancerous progression may attenuate the excitation laser power and also obscure the collagen Raman emission from the deep collagen basal membrane, thereby resulting in an overall decrease of Raman intensity at 875 cm^{-1} [135]. We also observed that the Raman intensities at 1335, 1370, 1421, 1490 cm^{-1} that mainly represent nucleic acids are generally higher for cancer laryngeal tissue, whereas Raman intensities at 1104 and 1745 cm^{-1} for phospholipids are lower as compared to normal tissue. These results are in consistent with the findings of the increase of nucleic acids to lipids ratio in cancerous tissues in literature [10, 136]. Since the random forests has taken into account of the massive multiple interactions among different Raman intensities (544 variables), the random forests algorithm is able to capture diagnostically significant features contained in the Raman spectra. Hence, the random forests algorithm is robust and powerful for distinguishing the origins of biochemical/biomolecular changes of Raman spectra for tissue carcinogenesis analysis.

To conclude for this chapter, significant differences in Raman spectra are found between normal and cancer tissue in this study, demonstrating the utility of Raman spectroscopy for laryngeal cancer detection. We have also employed the random forests technique for the first time to develop the random recursive partitioning ensemble algorithms to realize effective classification of Raman spectra between normal and cancer laryngeal tissue [10].

CHAPTER 6

EMPIRICAL STATISTICAL ANALYSIS FOR GASTRIC PRECANCER DIAGNOSIS

Gastric cancer is among one of the most common malignancies worldwide and it continues to be one of the worst 5-year survival rate statistics among other malignancies [137, 138]. Increasing the survival rate of patients with gastric cancer is important and recent statistics have shown that if diagnosis occurs at an early stage for gastric cancer, the 5-year survival rate of the patient is expected to be higher than 90% [102, 103]. Hence, early diagnosis of gastric cancer represents the most important measure to decreasing disease-associated mortality [139]. Additionally, the identification of gastric precancer (i.e., dysplasia) would offer the best prognosis as it is essential for planning optimal therapy, particularly for photodynamic therapy, endoscopic submucosal dissection (ESD) or endoscopic mucosal resection (EMR), as compare to surgery and chemotherapy [140]. However, early detection of gastric dysplasia renders a great challenge to the endoscopists as these flat lesions are usually lack of obvious gross morphological changes to be visualized under conventional white-light endoscopy.

NIR Raman spectroscopy has received much interest for optical diagnosis of diseased tissue in a number of organs [1-5-10], including malignant tumor in the stomach [12, 14, 16, 25-28]. To correlate the Raman spectral changes with different pathologic conditions in a straight forward way, the nonparametric diagnostic algorithms based on prominent tissue Raman bands intensities and/or intensity ratios have been practiced for effective distinction of gastric cancer from normal mucosa tissue. Diagnostic algorithms derived from intensity ratios is advantageous as it could provide tissue diagnosis with being inherently independent of Raman spectroscopic measurement conditions such as excitation/detection geometries, excitation light power fluctuations, probe-tissue positioning variations, etc [141]. With the potential of Raman intensity ratio diagnostic algorithms to also amplify the molecular distinction between normal and cancer tissues for better tissue classification, Hu *et al.* utilized the Raman peak intensities ratios of I_{1587}/I_{1156} and I_{1156}/I_{1660} to successfully discriminate malignant tissue from normal stomach tissue with a diagnostic accuracy of nearly 100% [28]. Very recently, Kawabata *et al*. reported that simply using a Raman band intensity at 1644 cm^{-1} for proteins of NIR Raman spectra could yield a diagnostic accuracy of 70% for 123 neoplastic and 128 non-neoplastic gastric tissue samples, which was close to the diagnostic accuracy level produced by the multivariate statistical techniques (e.g., PCA, LDA) [27]. These work highlighted the high efficacy of Raman spectroscopy associated with prominent Raman peak intensities or intensity ratios for gastric cancer detection. In this chapter, we explore the potential of NIR Raman spectroscopy (at 785 nm excitation) in conjunction with pairwise combinations of different Raman peak intensity ratios as diagnostic algorithms for identifying dysplasia from normal gastric mucosa tissue [16].

6.1 COMPARISON OF SPECTRAL DIFFERENCES BETWEEN NORMAL AND DYSPLASIA GASTRIC TISSUES

Figure 6.1(a) shows the mean normalized NIR Raman spectra of normal (n=44) and dysplasia (n=21) gastric tissues. A comparison of Raman spectra of dysplasia tissue with respect to the normal reveals remarkable differences in spectral shapes and intensities (Figure 6.1(b)), with dysplasia tissue showing lower intensities at 875, 1004, 1100, 1208 and 1745 cm^{-1}, while being higher at 1265, 1335, 1450 and 1655 cm^{-1}, respectively. There are also obvious changes of Raman peak positions and bandwidths in the spectral ranges of 1200-1500 cm^{-1}, and 1600-1700 cm^{-1} which are related to the amide III and amide I of proteins, CH_3CH_2 twisting of proteins/nucleic acids, and CH_2 bending mode of proteins and lipids for dysplasia. Hence, the significant differences in Raman spectra between normal and dysplasia tissue confirm a potential role of NIR Raman spectroscopy for precancer diagnosis in the stomach.

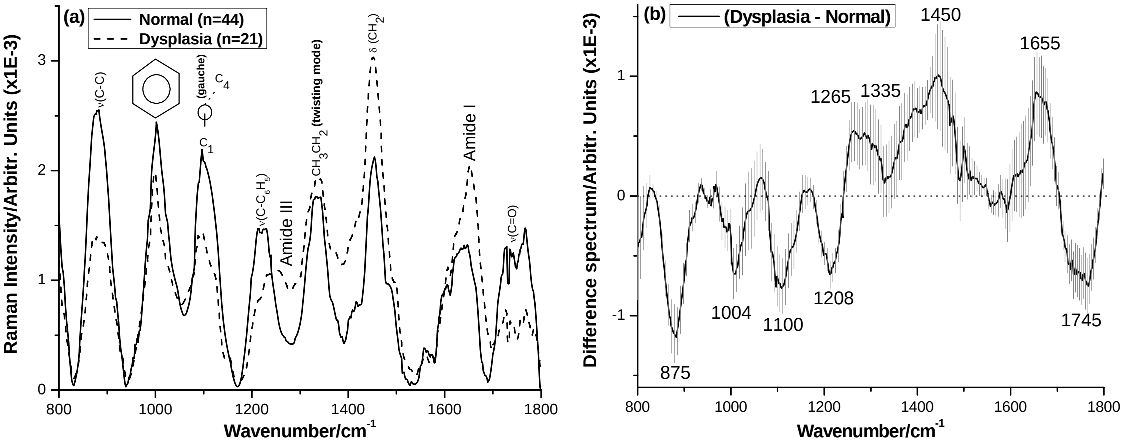

Figure 6.1 (a) The mean normalized NIR Raman spectra from normal (n=44) and dysplasia (n=21) gastric mucosa tissue samples; (b) Difference spectrum ± 1.96 SD calculated from the mean Raman spectra between normal and dysplasia tissue (i.e., the mean normalized Raman spectrum of dysplasia tissue minus the mean normalized Raman spectrum of normal tissue). Solid and dotted lines represent the mean spectra, and shaded areas indicate the variance within 95% confidence interval of the mean difference of the respective spectra.

This result shows that there are distinctive spectral differences between dysplasia and normal gastric tissue [16]. For example, Raman peak intensity at 875 cm^{-1}

(hydroxyproline of collagen) was found to be much reduced in dysplastic tissue. This was probably due to the cytoplasmic mucin depletion and the elevated concentration of metalloproteinase which cleaved collagen in the stroma layer in gastric dysplasia. In addition, the thickening of the epithelium associated with dysplastic progression may attenuate the excitation laser power and also obscure the collagen Raman emission from the deep collagen basal membrane, thereby resulting in an overall decrease of Raman intensity at 875 cm^{-1} from dysplasia tissue [16]. Raman bands for essential amino acids at 1004 and 1208 cm^{-1} (phenylalanine and tryptophan) show lower percentage signals for dysplasia compared to the normal, indicating a decrease in the percentage of phenylalanine and tryptophan relative to the total Raman-active constituents in the dysplasia. This is consistent with Alimova *et al.*'s report on a decrease in the tryptophan phosphorescence signal associated with malignancies in breast tissues [142], and also in agreement with Worthington *et al*.'s study that established a link between essential amino acids deficiency (e.g., phenylalanine) and carcinogenesis [143]. Similarly, in accord with Huang *et al*. Raman studies on lung cancer diagnosis [5], Raman peaks at around 1100 and 1745 cm^{-1} also show lower intensities for gastric dysplasia, indicating the decrease of Raman signals related to different vibrational modes from the hydrophobic chains of phospholipids which make up the membranes of the cell and numerous organelles inside the cytoplasm of the cell. On the other hand, the Raman bands at 1265 and 1655 cm^{-1} for histones (α-helical proteins surrounding the DNA) are significantly higher in dysplasia than normal, suggesting that hyperchromatism may take place in the nucleus with neoplastic transformation in the gastric [144]. Furthermore, the intensity of Raman peak at 1450 cm^{-1} (CH_2 proteins/lipids) and the bandwidths of Raman peaks at 1335 cm^{-1}

(CH_3CH_2 twisting of proteins and nucleic acids) and at 1655 cm^{-1} (amide I band of proteins) have also been found to change (e.g., intensity increase; bandwidth widening and peak shift, as shown in Figure 6.1) with dysplastic transformation, and these biomolecular changes could be related to the increase of mitotic activity occurring in the nucleus [145]. As a result, the distinctive differences in Raman spectra between normal and dysplasia gastric tissue reinforce that Raman spectroscopy can be used to reveal cellular and molecular changes associated with dysplastic transformation.

6.2 Raman intensity ratio

The nonparametric analysis based on intensity ratios of prominent Raman bands which were identified from the difference spectrum in Figure 6.1(b) is explored for tissue diagnosis in a straightforward way. Figure 6.2 shows box charts of the 6 significant Raman peak intensity ratios of I_{875}/I_{1450}, I_{1004}/I_{1450}, I_{1100}/I_{1450}, I_{1208}/I_{1450}, I_{1745}/I_{1450} and I_{1208}/I_{1655} (unpaired two-sided Student's *t*-test, $p < 0.0001$) correlated with their histopathologic findings. Each ratio belonging to normal tissue is significantly higher than that of dysplasia tissue, and the corresponding separation lines (i.e., diagnostic algorithms), as shown in Figure 6.2(a-f), classify dysplasia from normal with a sensitivity of 76.2%, 81.0%, 95.2%, 81.0%, 95.2%, and 76.2%, and a specificity of 90.9%, 90.9%, 77.3%, 88.6%, 75.0%, and 84.1%, respectively. These results indicate that different ratios of Raman peak intensities give different levels of accuracy for tissue classification.

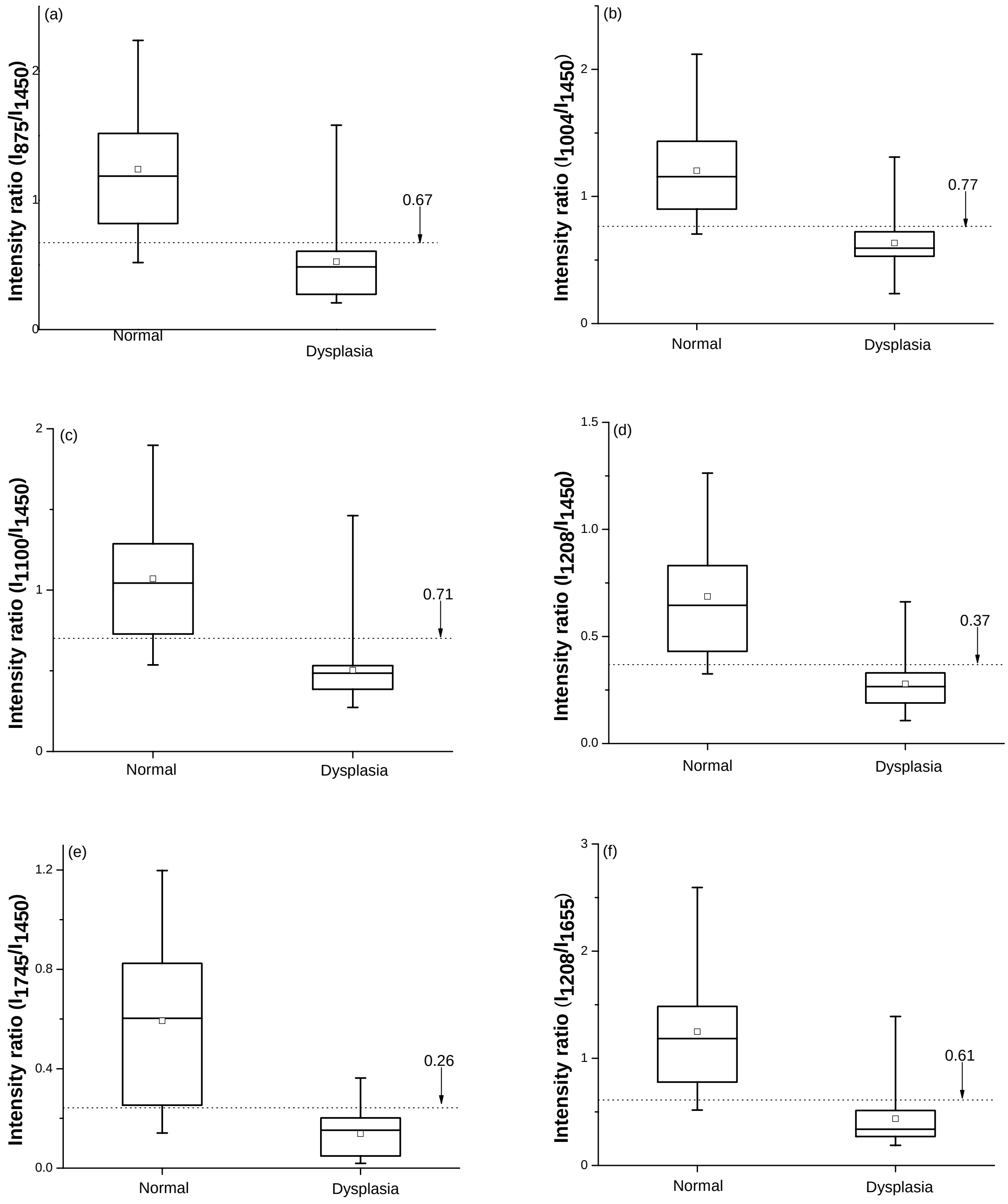

Figure 6.2 Box charts of the 6 significant Raman peak intensity ratios which can differentiate dysplasia from normal gastric mucosa tissue (unpaired Student's *t*-test, $p<0.0001$): (a) I_{875}/I_{1450}; (b) I_{1004}/I_{1450}; (c) I_{1100}/I_{1450}; (d) I_{1208}/I_{1450}; (e) I_{1745}/I_{1450}, and (f) I_{1208}/I_{1655}. The dotted lines ($I_{875}/I_{1450} = 0.67$; $I_{1004}/I_{1450} = 0.77$; $I_{1100}/I_{1450} = 0.71$; $I_{1208}/I_{1450} =$

0.37; I_{1745}/I_{1450} = 0.26; I_{1208}/I_{1655} = 0.61) as diagnostic threshold algorithms classify dysplasia from normal with sensitivity of 76.2% (16/21), 81.0% (17/21), 95.2% (20/21), 81.0% (17/21), 95.2% (20/21), and 76.2% (16/21); specificity of 90.9% (40/44), 90.9% (40/44), 77.3% (34/44), 88.6% (39/44), 75.0% (33/44), and 84.1% (37/44), respectively.

Given the potential of Raman intensity ratios approach to amplifying the molecular distinction between different pathological groups as well as the ability of achieving the biomolecular diagnosis independent of Raman measurement conditions such as excitation light power fluctuation or probe positioning variation [141], we have comprehensively investigated different prominent Raman intensity ratios as nonparametric diagnostic algorithms for classifying dysplasia from normal gastric tissue. As a result, 6 diagnostically significant ratios were found to be able to enhance the molecular differences between normal and dysplasia tissues. For instance, the intensity ratio of Raman peak intensity at 875 cm^{-1} (hydroxyproline of collagen) to the peak at 1450 cm^{-1} (CH_2 mode of proteins and lipids) yielded a diagnostic sensitivity of 76.2% and specificity of 90.9% for separating dysplasia from normal tissue. Further investigation also shows that other intensity ratios such as the Raman peak intensity bands for essential amino acids at 1004 cm^{-1} and at 1208 cm^{-1} to the Raman peak intensity at 1450 cm^{-1} also provide good differentiation between normal and dysplasia tissue. The significant differences of these intensity ratios between normal and dysplasia tissue may reflect the relative changes in the concentration of potential biological markers ranging from cell surface antigens, cytoplasmic proteins and mucin, collagen in the extracellular matrix, enzymes, and hormones associated with dysplastic changes [16]. Hence, the Raman intensity ratios could be utilized for dysplasia identification in the gastric.

6.3 Optimal Raman intensity ratio diagnostic algorithm

To further improve the discriminative ability of the empirical approach but yet allowing interpretations of diagnostic results in a straightforward manner, we also extensively explored the possible pairwise combinations of different Raman intensity ratios for tissue diagnosis and classification (Table 6.1). Table 6.1 shows the results of predicted diagnostic sensitivity, specificity and accuracy using 15 pairwise combinations of the significant Raman peak intensity ratios for tissue classification. The combination of I_{1208}/I_{1655} and I_{875}/I_{1450} is one of the optimal diagnostic algorithms for discriminating dysplasia from normal gastric tissue. Figure 6.3(a) presents the scatter plot of combining the intensity ratios of I_{1208}/I_{1655} and I_{875}/I_{1450} for different pathologic types. The linear discrimination line ($I_{1208}/I_{1655} = -0.81\ I_{875}/I_{1450} + 1.17$) generated from the logistic regression analysis together with the leave-one sample-out, cross-validation method gives a diagnostic sensitivity of 90.5% and specificity of 90.9% for separating dysplastic tissue from normal gastric tissues. To further evaluate the performance of this optimal diagnostic algorithm, receiver operating characteristic (ROC) curve (Figure 6.3(b)) is also produced from the scatter plot in Figures 6.3(a) at different threshold levels. The integration areas under the ROC curve is 0.96 for the combined Raman peak intensity ratios, proving the robustness of this nonparametric algorithm derived for gastric precancer diagnosis.

Table 6.1 Results of predicted sensitivity, specificity and accuracy for discrimination of gastric dysplasia from gastric normal tissue using the pairwise combinations of Raman peak intensity ratios.

Diagnostic Pairwise Combinations			After leave-one sample-out cross validation (Predicted)		
			Sensitivity	**Specificity**	**Accuracy**
$I_{875/1450}$	*v.s.*	$I_{1004/1450}$	80.9% (17/21)	90.9% (40/44)	87.7%(57/65)
		$I_{1100/1450}$	85.7% (18/21)	88.6% (39/44)	87.7%(57/65)
		$I_{1208/1450}$	76.1% (16/21)	88.6% (39/44)	84.6% (55/65)
		$I_{1745/1450}$	85.7% (18/21)	84.1% (37/44)	84.6% (55/65)
		$^{*}I_{1208/1655}$	90.5% (19/21)	90.9% (40/44)	90.8% (59/65)
$I_{1004/1450}$	*v.s.*	$I_{1100/1450}$	76.1% (16/21)	90.9% (40/44)	86.2% (56/65)
		$I_{1208/1450}$	76.1% (16/21)	86.4% (38/44)	83.1% (54/65)
		$I_{1745/1450}$	76.1% (16/21)	86.4% (38/44)	83.1% (54/65)
		$I_{1208/1655}$	80.9% (17/21)	88.6% (39/44)	86.2% (56/65)
$I_{1100/1450}$	*v.s.*	$I_{1208/1450}$	80.9% (17/21)	88.6% (39/44)	86.2% (56/65)
		$I_{1745/1450}$	76.1% (16/21)	84.1% (37/44)	81.5% (53/65)
		$I_{1208/1655}$	80.9% (17/21)	90.9% (40/44)	87.7%(57/65)
$I_{1208/1450}$	*v.s.*	$I_{1745/1450}$	76.1% (16/21)	86.4% (38/44)	83.1% (54/65)
		$I_{1208/1655}$	80.9% (17/21)	90.9% (40/44)	87.7%(57/65)
$I_{1745/1450}$	*v.s.*	$I_{1208/1655}$	76.1% (16/21)	84.1% (37/44)	81.5% (53/65)

Note:

The symbol * denotes the pairwise combination of Raman peak intensity ratios with the highest sensitivity, specificity and accuracy values (lowest estimated generalization error) after the leave-one sample-out, cross validation.

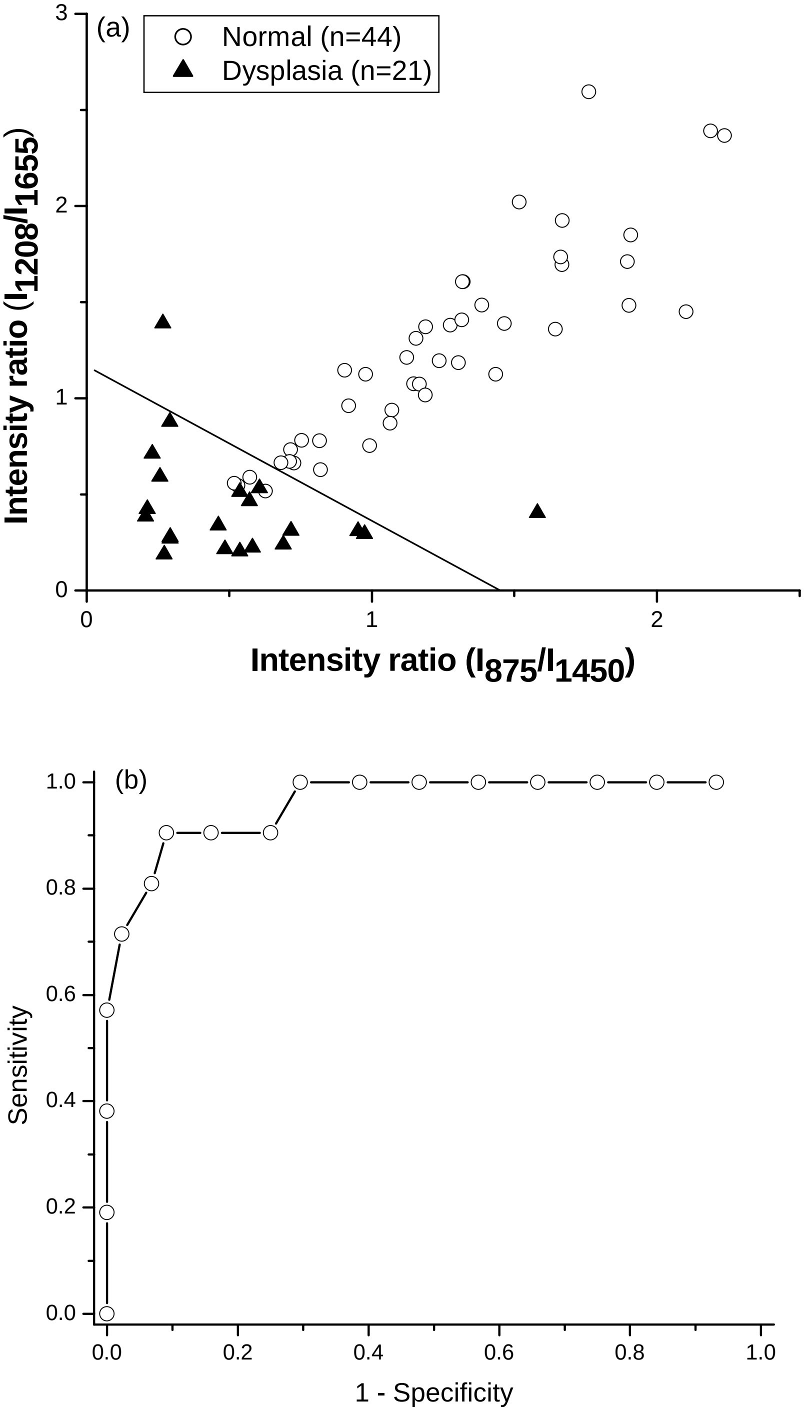

Figure 6.3 (a) Two-dimensional scatter plot showing the distribution of normal and dysplastic gastric mucosa tissues after combining both Raman peak intensity ratios of I_{1208}/I_{1655} and I_{875}/I_{1450} as a discriminating algorithm. A linear diagnostic decision algorithm ($I_{1208}/I_{1655} = -0.81\ I_{875}/I_{1450} + 1.17$) yields a sensitivity of 90.5% (19/21) and a specificity of 90.9% (40/44) for separating dysplasia from normal tissue. (b) Receiver

operating characteristic (ROC) curve with an area under curve (AUC) of 0.96 illustrates the ability of Raman spectroscopy to identify dysplasia from normal gastric tissues.

The logistic regression analysis coupled with leave-one sample-out, cross validation, as well as ROC analysis for unbiased evaluation on different 2-dimensional diagnostic models indicate that the combined intensity ratios of I_{1208}/I_{1655} and I_{875}/I_{1450} can serve as one of the most effective ratio diagnostic algorithms for dysplasia identification. Even though the simplistic empirical analysis here could only utilize up to four prominent Raman peaks each time for tissue classification, these algorithms still contained a mixture of different biomolecular diagnostic information ranging from the nucleus and cytoplasm within the cell, to the extracellular matrix outside the cell. For example, the nonparametric algorithm (I_{1208}/I_{1655} and I_{875}/I_{1450}) contained the diagnostic information that are specifically extracted from collagen present in the extracellular matrix, essential amino acids and CH_2 of proteins/lipids found mostly in tissues and cells, and also histones which are present in the nucleus [5, 6]. Hence, the Raman intensity ratios derived in this work could potentially be used as effective diagnostic algorithms for classifying dysplasia from normal gastric tissue. One notes that the intensity ratio analysis employs the selection of a group of diagnostic Raman intensity features directly derived from the tissue Raman spectra to construct the classification model for tissue diagnosis. Other powerful multivariate feature selection algorithms (e.g. random forests) that take into consideration of the interactions among different spectral features could also be applied to choose the best subset of Raman features for effective tissue classification. Alternatively, multivariate statistical techniques such as PCA-LDA, which fully utilize diagnostically significant features contained in the entire Raman spectrum and transform them into multiple independent linear combinations of the original spectral features, could be used for further improving diagnostic efficacy for gastric precancer diagnosis and classification. Note that a detailed comparison of the PCA-LDA with the Raman intensity ratio algorithm will be discussed in the following chapter [12].

In conclusion, this study shows that there are significant differences in Raman spectra between normal and dysplastic gastric tissue, demonstrating the utility of NIR Raman spectroscopy for differentiating dysplasia from normal tissue in the stomach. Furthermore, with the use of Raman intensity ratios as diagnostic algorithms, NIR Raman spectroscopy could be a clinically useful tool for rapid, noninvasive, *in vivo* diagnosis and detection of gastric precancer at the molecular level [16].

CHAPTER 7

COMPARISON OF PERFORMANCE FOR MULTIVARIATE STATISTICAL ANALYSIS AND EMPIRICAL STATISTICAL ANALYSIS FOR GASTRIC DYSPLASIA DIAGNOSIS

The current gold standard for clinical diagnosis of gastric dysplasia is through histological observation by the pathologist, on the extent of cytological and architectural abnormalities of the histologically prepared tissue samples [145]. These abnormalities involve much molecular alterations which could also be tapped upon for diagnosis, most importantly during routine endoscopic inspection [145]. Hence, Raman spectroscopy that is capable of providing rich biochemical and biomolecular information about tissue may be the promising diagnostic tool to be used for molecular discrimination of gastric dysplasia. However, as gastric dysplasia belongs to part of a widely accepted multi-step, continuum progression cascade from normal gastric tissue to adenocarcinoma [146], it implies gastric dysplasia's vague molecular distinction that may render characterization and discrimination tougher for Raman spectral analysis. As shown in previous chapter (i.e., Figure 6.1), the Raman spectral pattern between normal and dysplastic gastric tissues could be very similar, it is highly desirable to develop robust diagnostic approaches to extract all possible diagnostic information contained in tissue Raman spectra for well correlation with tissue changes associated with neoplastic transformation. Consequently, with a larger sample size, both empirical and statistical techniques were examined in detail in this chapter for a more robust evaluation to attain the likelihood of good clinical discriminators of Raman spectra for separation between normal and dysplastic gastric tissues [12].

7.1 ANALYTICAL APPROACHES

7.1.1 EMPIRICAL APPROACH: INTENSITY RATIO

Nonparametric diagnostic algorithms based on peak intensities, spectral bandwidths, and/or peak ratios have been widely employed in literature to correlate the variations of tissue spectra with tissue pathology in a simple and straightforward fashion [5, 20]. In this chapter, the empirical diagnostic algorithm based on the ratio of the Raman peak intensity at 875 cm^{-1} for hydroxyproline to the peak intensity at 1450 cm^{-1} for CH_2 proteins/lipids was selected for tissue classification. Figure 7.1 shows an example of the scatter plot of the ratio of Raman intensity at 875 cm^{-1} to that at 1450 cm^{-1} grouped according to tissue pathologic types. The mean value (mean± SD) of this ratio for normal tissues (1.13 ± 0.46, n=55) is significantly different from the mean value for dysplastic tissues (0.52 ± 0.33, n=21) (unpaired two-sided Student's *t*-test, p <0.00001). The decision line ($I_{875}/I_{1450} = 0.717$) discriminates dysplasia tissue from normal gastric tissue with a sensitivity of 85.7% and a specificity of 80.0%.

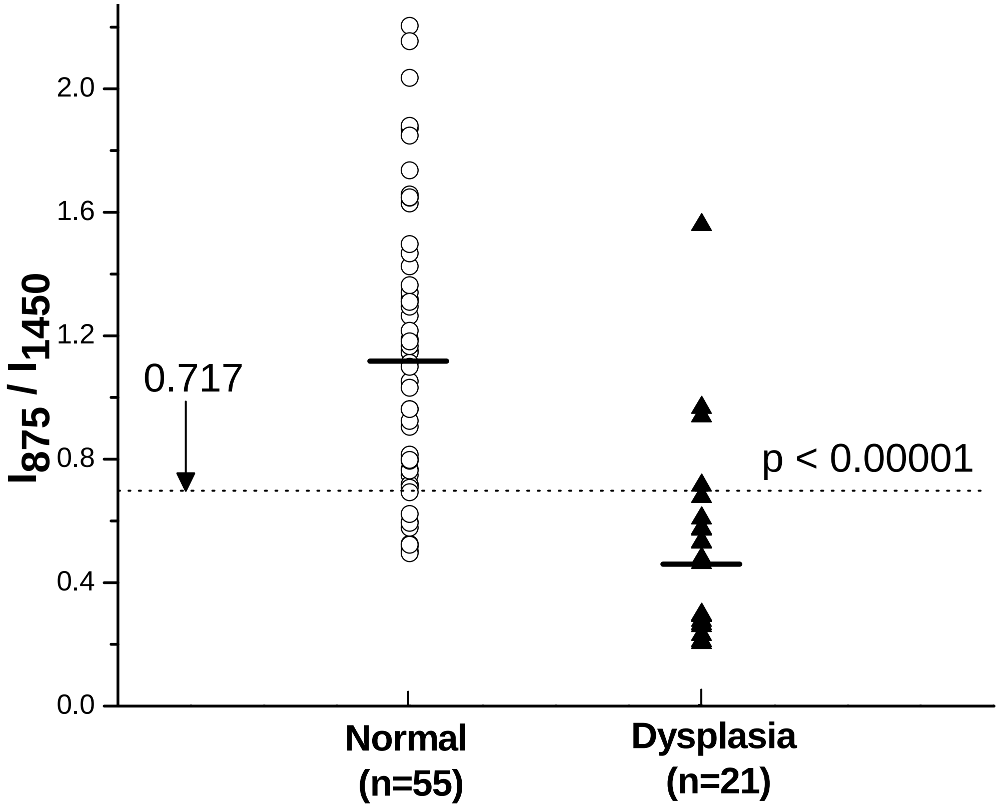

Figure 7.1 Scatter plot of the intensity ratio of Raman signals at 875 cm^{-1} and 1450 cm^{-1}, as measured for each sample and classified according to the histological results. The mean intensity (1.13 ± 0.46,) of normal tissue is significantly different from the mean value (0.52 ± 0.33) of dysplasia tissue (unpaired Student's *t*-test, p<0.00001). The decision line (I_{875}/I_{1450} = 0.717) separates dysplasia tissue from normal tissue with a sensitivity of 85.7% (18/21) and specificity of 80.0% (44/55).

7.1.2 MULTIVARIATE ANALYSIS: PCA

The high dimension of Raman spectral space (each Raman spectrum ranging from 800-1800 cm^{-1} with a set of 544 intensities) will result in computational complexity and inefficiency in optimization and implementation of the LDA algorithms. As such, PCA was firstly performed on tissue Raman dataset to reduce the dimension of Raman spectral space while retaining the most diagnostically significant information for tissue classification. To eliminate the influence of inter and/or intra-subject spectral variability on PCA, the entire spectra were standardized so that the mean of the spectra was zero and the standard deviation of all the spectral intensities was one. Mean centering ensures that the principal components (PCs) form an orthogonal basis [12]. The standardized Raman data sets were assembled into data matrices with wavenumber columns and individual case rows. Thus, PCA was performed on the standardized spectral data matrices to generate PCs comprising a reduced number of orthogonal variables that accounted for most of the total variance in original spectra. Each loading vector is related to the original spectrum by a variable called the PC score, which represents the weight of that particular component against the basis spectrum. PC scores reflect the differences between different classes. Unpaired Student's *t*-test was used to identify the most diagnostically significant PCs ($p<0.05$) and showed that there were four PCs (PC1, PC2, PC4, and PC5) that were diagnostically significant ($p<0.05$) for discriminating dysplasia tissue from normal tissue [12]. Figure 7.2 displays the four significant PCs scores calculated from PCA on the Raman spectra. The first PC accounts for the largest variance (e.g. 42.6% of the total variance), whereas the successive PCs describe the spectral features that contribute progressively smaller variances. Some PC features (Figure 7.2 (a) – (d)), such as peaks, troughs, spectral shapes, are similar to those of tissue Raman spectra.

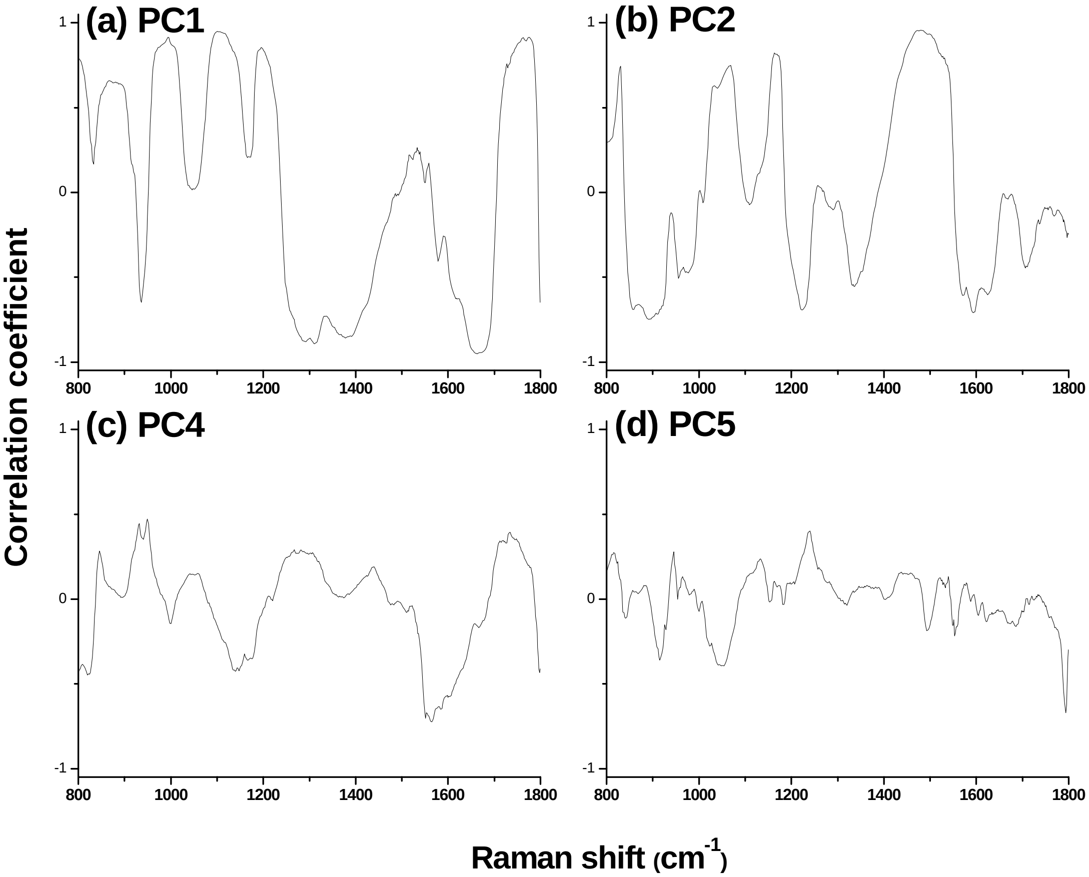

Figure 7.2 The first four diagnostically significant principal components (PCs) accounting for about 78.5% of the total variance calculated from Raman spectra (PC1 – 42.6%, PC2 – 25.4%, PC4 – 7.9%, and PC5 – 2.6%), revealing the diagnostically significant spectral features for tissue classification.

Figure 7.3 shows the correlations between the diagnostically significant PC scores for normal and dysplastic gastric tissue, illustrating the utility of PC scores for classification of Raman spectra between different tissue types. Normal and dysplasia tissues can be largely clustered into two separate groups based on different combinations of significant PCs, and the corresponding separation lines (i.e., diagnostic algorithms) in Figure 7.3 (a-f) classify dysplasia from normal tissue with the sensitivity of 90.5% (19/21), 76.2% (16/21), 71.4% (15/21), 81.0% (17/21), 71.4% (15/21), and 71.4% (15/21); specificity of

90.9% (50/55), 80.0% (44/55), 83.6% (46/55), 80.0% (44/55), 72.7% (40/55), and 72.7% (40/55), respectively. These results show that selection of different combinations of significant PCs will give different levels of accuracy for tissue classification.

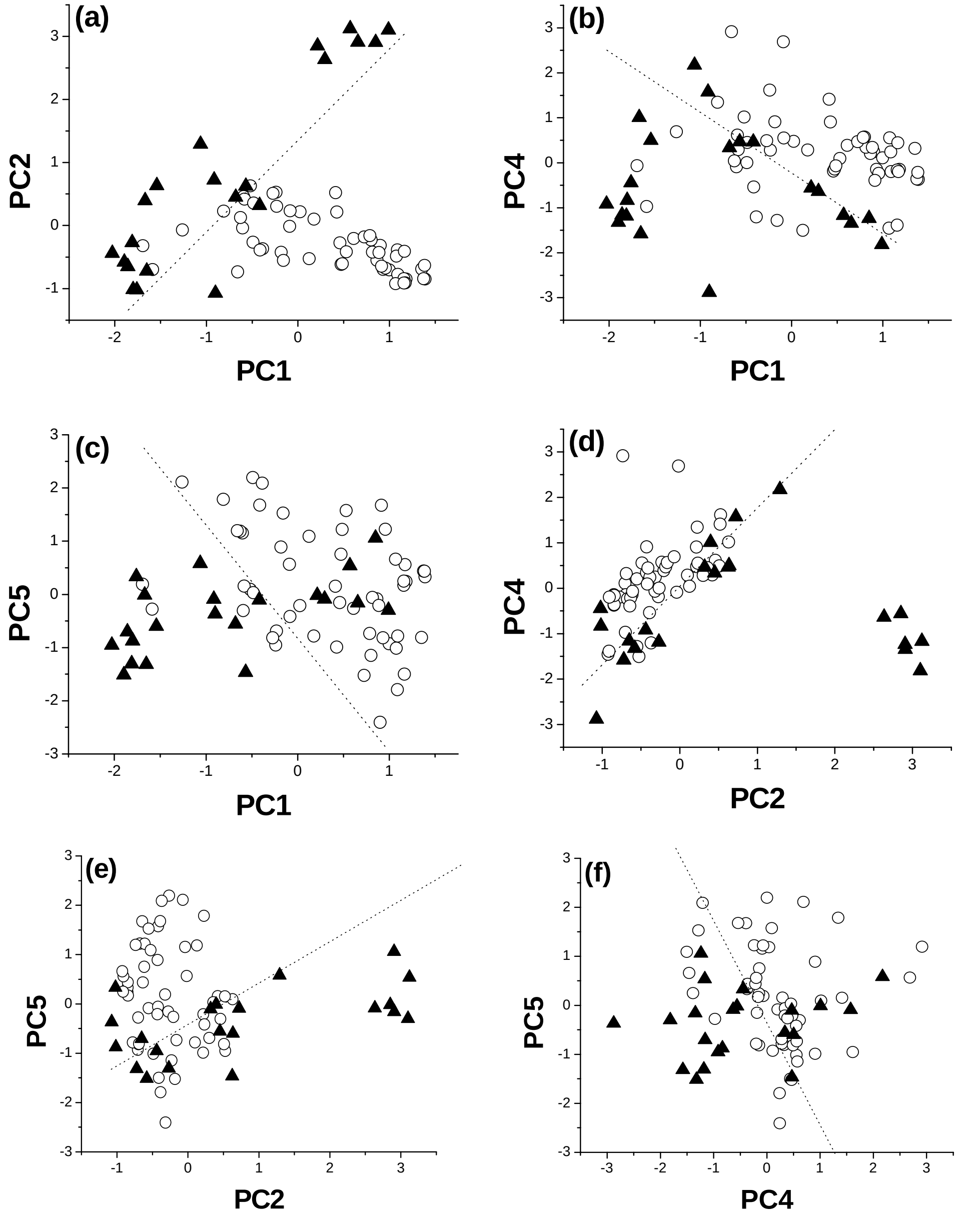

Figure 7.3 Scatter plots of the diagnostically significantly PC scores for normal and dysplastic gastric tissue derived from Raman spectra, (a) PC1 *vs*. PC2; (b) PC1 *vs*. PC4; (c) PC1 *vs*. PC5; (d) PC2 *vs*. PC4; (e) PC2 *vs*. PC5; (f) PC4 *vs*. PC5. The dotted lines

(PC2= 1.46 PC1 + 1.34; PC4= -1.32 PC1 + 0.94; PC5= -2.16 PC1 – 0.89; PC4= 1.74 PC2 + 0.12; PC5= 0. 84 PC2 – 0.381; PC5= -2.05 PC4 – 0.29) as diagnostic algorithms classify dysplasia from normal with sensitivity of 90.5% (19/21), 76.2% (16/21), 71.4% (15/21), 81.0% (17/21), 71.4% (15/21), and 71.4% (15/21); specificity of 90.9% (50/55), 80.0% (44/55), 83.6% (46/55), 80.0% (44/55), 72.7% (40/55), and 72.7% (40/55), respectively. Circle (○): normal; Triangle (▲): dysplasia.

7.1.2 Multivariate analysis: LDA

To further improve tissue diagnosis, all the four diagnostically significant PCs were loaded into the LDA model for generating effective diagnostic algorithms for tissue classification. LDA determines the discriminant function that maximizes the variances in the data between groups while minimizing the variances between members of the same group [12]. The performance of the diagnostic algorithms rendered by the LDA models for correctly predicting the tissue groups (i.e., normal *vs*. dysplasia) was estimated in an unbiased manner using the leave-one sample-out, cross validation method on all model spectra. In this method, one sample (i.e., one spectrum) was held out from the data set and the entire algorithm including PCA and LDA was redeveloped using the remaining tissue spectra. The algorithm was then used to classify the withheld spectrum. This process was repeated until all withheld spectra were classified.

Figure 7.4 shows the classification results based on PCA-LDA technique together with leave-one spectrum-out, cross-validation method. The PCA-LDA diagnostic algorithms yielded the diagnostic sensitivity of 95.2% and specificity 90.9% for separating dysplasia from normal gastric tissues.

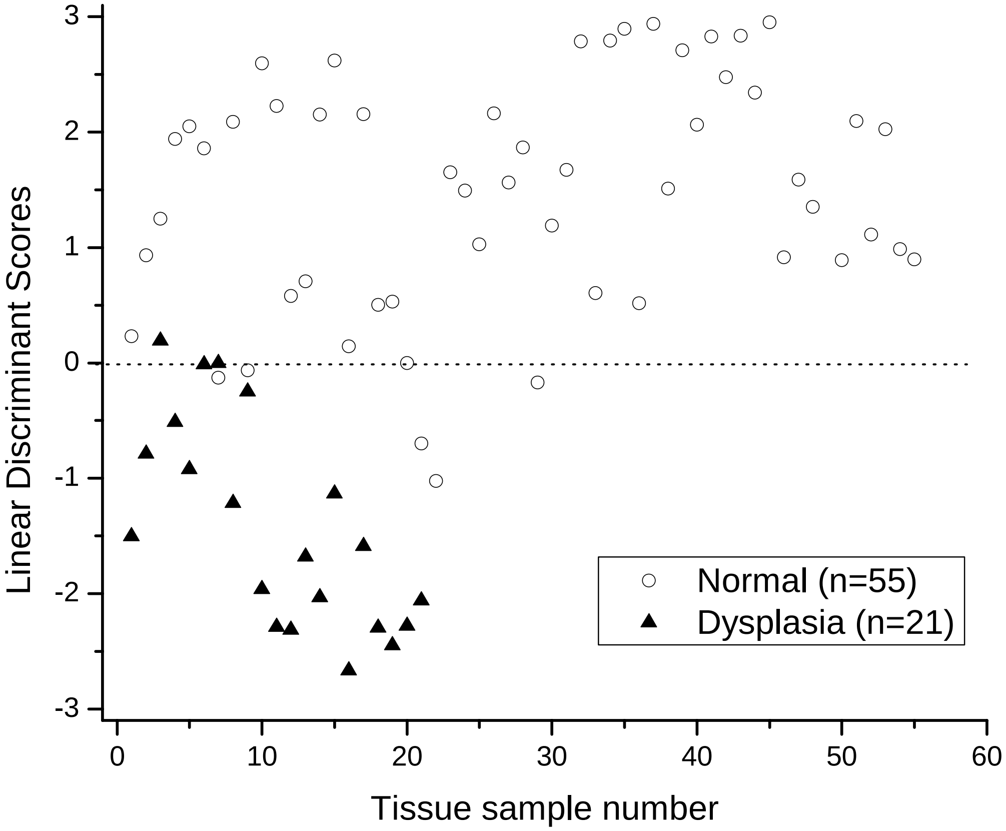

Figure 7.4 Scatter plot of the linear discriminant scores of belonging to the normal and dysplasia categories using the PCA-LDA technique together with leave-one spectrum-out, cross-validation method. The separate line yields a diagnostic sensitivity of 95.2% (20/21) and specificity of 90.9% (50/55) for differentiation between normal and dysplasia tissue.

7.1.3 Comparison of performance for different analytic techniques: ROC

To evaluate and compare the performance of the PCA-LDA-based diagnostic algorithms derived from all the significant PCs of tissue Raman dataset against the empirical approach-based diagnostic algorithm derived from the intensity ratio of I_{875}/I_{1450}, the ROC curves (Figure 7.5) were generated from the scatter plots in Figures 7.2 and 7.4 at different threshold levels, displaying the discrimination results using both diagnostic algorithms. A comparative evaluation of the ROC curves indicates that PCA-LDA-based diagnostic algorithm gives more effective diagnostic capability for detection of gastric dysplasia from normal gastric tissues, as illustrated by the improvement in the diagnostic sensitivity and specificity. The integration areas under the ROC curves are 0.98 and 0.88 for PCA-LDA-based diagnostic algorithms and the nonparametric intensity ratio algorithm, respectively. These results demonstrate that PCA-LDA-based diagnostic algorithms that utilized the entire spectral features of Raman spectra yield a better diagnostics accuracy than the empirical approach.

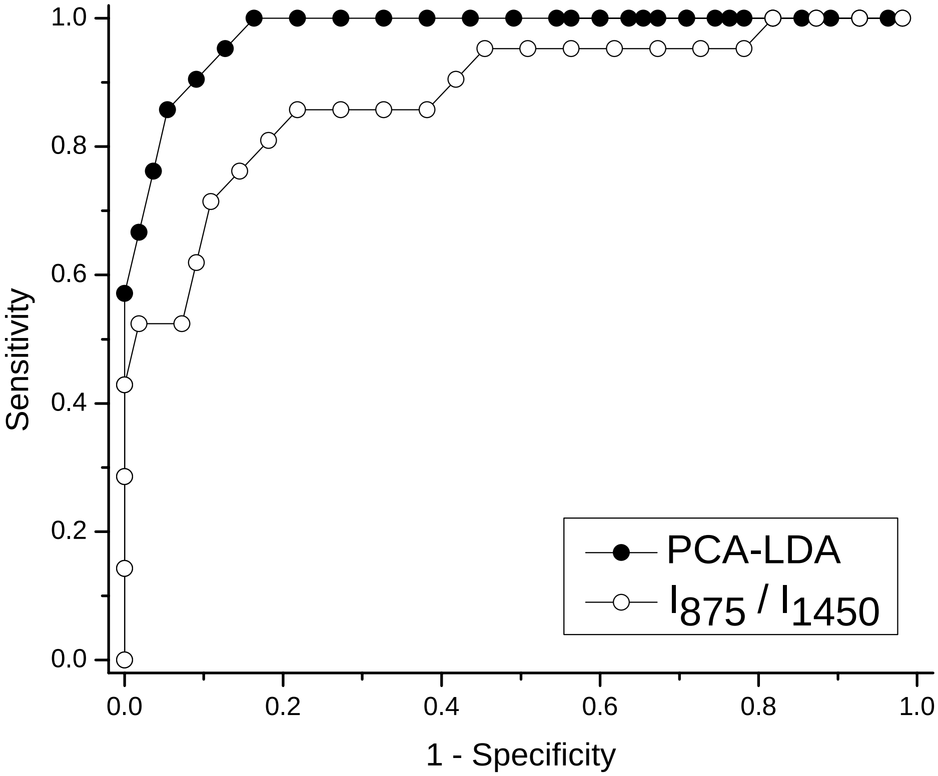

Figure 7.5 Comparison of ROC curves of discrimination results for Raman spectra utilizing the PCA-LDA-based spectral classification with leave-one spectrum-out, cross-validation method and the empirical approach using Raman intensity ratio of I_{875}/I_{1450}. The integration areas under the ROC curves are 0.98 and 0.88 for PCA-LDA-based diagnostic algorithm and intensity ratio algorithm, respectively, demonstrating the efficacy of PCA-LDA algorithms for tissue classification.

The simplistic empirical analysis above only employs a limited number of Raman peaks for tissue diagnosis, most of the information contained in the Raman spectra has not been used for spectral analysis [12]. Since biological tissue is complex, it is likely that there are many biochemical species influencing diseases concurrently. Therefore, a multivariate statistical analysis (e.g., PCA and LDA) that utilizes the entire spectrum to determine the most diagnostically significant spectral features may improve the diagnostic efficiency of Raman technique for tissue analysis and classification. As such, PCA-LDA together with cross-validation technique was applied in this work to the NIR Raman spectra acquired for dysplasia tissue identification. The unpaired, two-sided Student's *t*-test identified that only a few PCs (PC1, PC2, PC4 and PC5) contained the most diagnostically significant information ($p<0.05$) for tissue classification. We note that one of the most statistically significant PCs (e.g., PC5) only describes small amount (2.6%) of the total variance. This indicates that some PCs with small variances can still contain the useful diagnostic information for revealing molecular changes with dysplastic transformation. However, since the noise present in weak tissue Raman signals may affect the determination of significant PCs with smaller variances for tissue diagnosis, caution should be taken when acquiring the weak tissue Raman signals [147]. Hence, the rapid fiber-optic Raman system with a high signal-to-noise ratio (3.3-16 folds improvement) [46] was employed to obtain high quality Raman tissue spectra, and an appropriate data preprocessing was also introduced for further reducing the noise interference in PC analysis. The consistency in identifying similar significant PC scores from run to run during the leave-one spectrum-out, cross-validation testing suggested that the diagnostic algorithms developed were robust for Raman spectral analysis in this study. To develop effective diagnostic algorithms for tissue classification, all the four diagnostically significant PCs were utilized in the LDA model. The diagnostic sensitivity and specificity of 95.2% and 90.9% for identifying dysplasia from normal gastric tissue can be achieved using the PCA-LDA model, which had almost a 10% improvement in diagnostic accuracy compared to the empirical method. ROC analysis (Figure 7.5) further confirms that PCA-LDA-based diagnostic algorithms employing the entire spectral features of Raman spectra are more robust and powerful in distinguishing dysplasia from normal tissue.

In conclusion, this work proved that multivariate statistical technique provided a higher accuracy performance compared to the intensity ratio method for NIR Raman spectroscopy detection of gastric precancer [12]. Therefore, NIR Raman spectroscopy in conjunction with multivariate statistical technique has potential for rapid diagnosis of dysplasia in the stomach based on the optical evaluation of spectral features of biomolecules.

RANDOM FORESTS DEMONSTRATION FOR GASTRIC PRECANCER DETECTION

PCA is primarily for data reduction rather than for identification of biochemical or biomolecular components of tissue [12]. It is usually difficult to interpret the physical meanings of the component spectra [12]. However, with more powerful diagnostic algorithms (e.g., random forests) [10], distinctive spectral regions that are optimal for tissue differentiation may be identified and related to particular biochemical and biomolecular changes (e.g., proteins, lipids, nucleic acid, carbohydrates) associated with neoplastic transformation. This chapter will present the investigation of the efficacy of random forests technique for Raman spectroscopic gastric dysplasia detection.

8.1 RESULTS OF THE EMPLOYMENT OF RANDOM FOREST ALGORITHM FOR GASTRIC DYSPLASIA DETECTION

In this study, the number of variables tested for each split was set to 23 (√544) and the error rate of different forest size from 1-1000 trees was investigated. Fig. 8.1 (a) demonstrated that the diagnostic algorithm stabilized after 284 trees were employed. Hence, we have selected to construct an ensemble of 285 trees to construct an optimal diagnostic algorithm. To illustrate the performance of this optimal diagnostic algorithm, ROC curve (Fig. 8.1 (b)) was generated by constructing different threshold levels from the probability associated with each data after the majority voting. The integration areas under the ROC curve is 0.950 for the forest size of 285 trees, proving the efficacy of this random forest algorithm derived for gastric dysplasia cancer diagnosis. The results show that random forests-based diagnostic algorithm is robust for gastric dysplasia diagnosis.

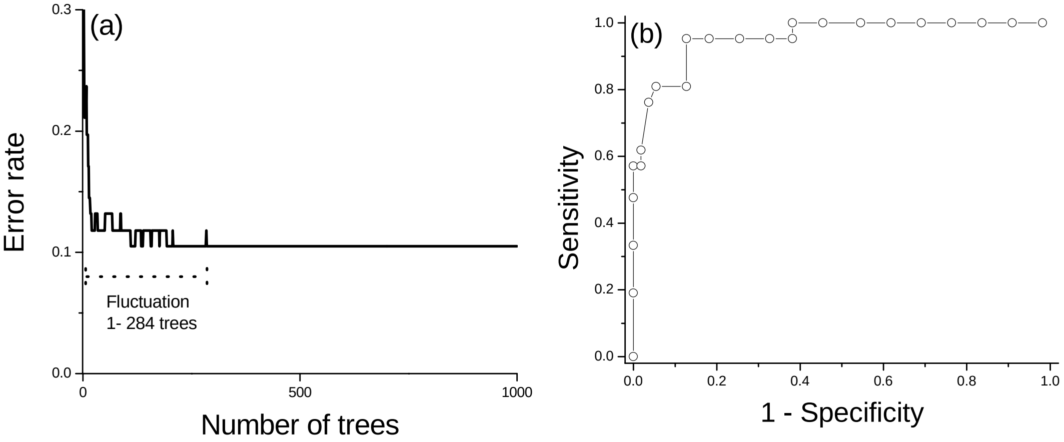

Figure 8.1 (a) Different error rates belonging to different sizes of the random forests (i.e., different number of trees) after the voting process on all the tissue Raman spectra. Stabilization of forests occurred at 0.105 after more than 284 trees, illustrating that the random forests algorithm does not overfit. (b) ROC curve of tissue classification belonging to the final optimal random forests tree size of 1000 with an AUC of 0.950, illustrating the diagnostic ability of Raman spectroscopy and random forests algorithm to identify gastric dysplasia from normal gastric tissue.

Figure 8.2 (a) displays the probabilistic classification results based on the random forests technique together with the leave-one sample (i.e., all spectra associated with the sample)-out, cross validation method. The random forests diagnostic algorithm yields the diagnostic sensitivity of 81.0%, and specificity of 92.7% for separating dysplasia from normal gastric tissues. In addition, Figure 8.2 (b) shows the variables importance plot based upon the construction of 1000 trees associated with gastric dysplasia transformation. The permutation accuracy importance measure notably reveals that Raman intensities at 875 (C-C stretching of hydroxyproline), 980 (*v* (C-C) in α conformation of proline and valine), 1218 (C-C_6H_5 stretching mode of tryptophan and phenylalanine), 1302 (CH_3CH_2 twisting mode of proteins), and 1420 cm^{-1} (δ (CH_2), δ (CH_3) of proteins, δ (CH_2) scissoring of phospholipids, deoxyriboses) are among the most important variables in the gastric tissue Raman spectra for distinguishing the dysplasia from normal. These results show that Raman spectroscopic technique combined with the random forests diagnostic algorithm is robust and powerful for gastric dysplasia diagnosis, and could provide interpretable Raman biomolecular information for the constructed diagnostic algorithm.

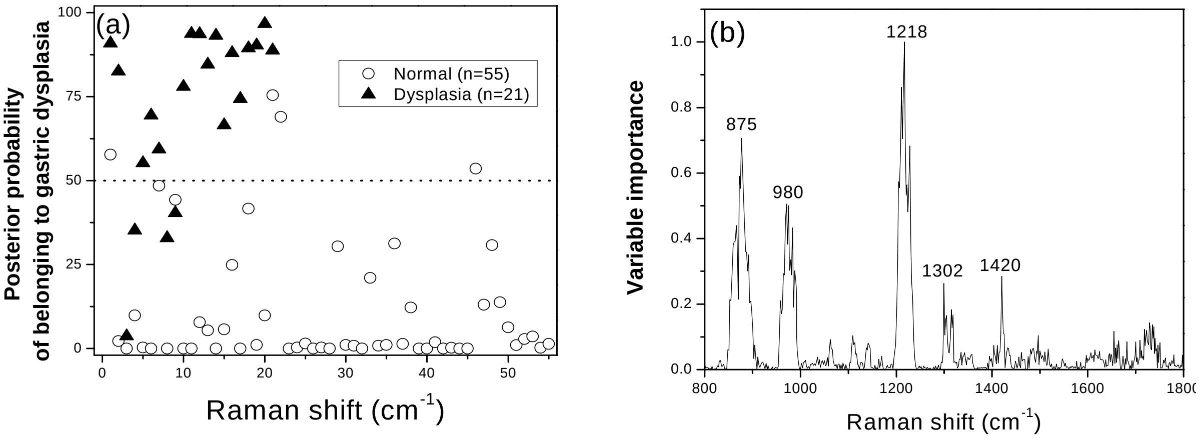

Figure 8.2 (a) Scatter plot of the generated probabilistic scores belonging to the normal and dysplasia categories using the random forests technique together with leave-one sample-out, cross validation method. The separate line yields a diagnostic sensitivity of 81.0% (17/21) and specificity of 92.7% (51/55) for differentiation between normal and dysplastic gastric tissue. (b) Variables importance plot for the Raman spectral region 800-1800 cm^{-1} generated from random forests size of 1000 trees which was used for discrimination of dysplasia from normal gastric tissue. The variable importance algorithm defines the most important variable as 1, whereas the least important variable as 0. Notable peaks are identified.

8.2 COMPARISON OF PERFORMANCE AMONG INTENSITY RATIO, PCA-LDA, RANDOM FORESTS ANALYTIC ALGORITHMS FOR GASTRIC PRECANCER DETECTION

To further evaluate and compare the performance of the different analytical algorithms constructed for gastric dysplasia detection, the ROC curves (Figure 8.3) from Figures 7.5 and 8.1 (b) were plotted together, displaying the discrimination results for the three different diagnostic algorithms. A comparative evaluation of the ROC curves (i.e., ROC-AUC) indicates that the PCA-LDA-based and random forests-based diagnostic algorithm can significantly give a more effective diagnostic capability as compared to the intensity ratio method for detection of gastric dysplasia from normal gastric tissues, as shown by the improvement in the diagnostic sensitivity and specificity. In addition, the ROC curve of PCA-LDA-based and random forests-based diagnostic algorithms appeared to be very similar, illustrating comparable diagnostic accuracy for detection of gastric dysplasia. The ROC-AUC for PCA-LDA-based and random forests-based diagnostic algorithm was 0.98 and 0.95, respectively, confirmed that the two diagnostic algorithms achieved almost equivalent overall diagnostic accuracy for separating gastric dysplasia from normal gastric tissues. This result demonstrated that random forests-based diagnostic algorithms that utilized the entire spectral features of Raman spectra can yield comparable diagnostics accuracy as compared to PCA-LDA-based analytic algorithm, and a better diagnostic accuracy than the empirical approach for discriminating gastric dysplasia from normal tissues.

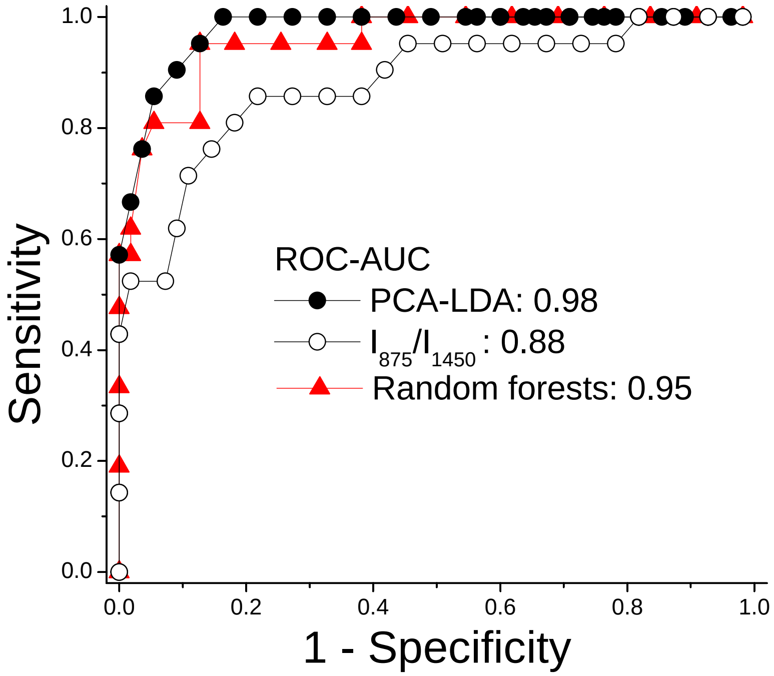

Figure 8.3 Comparison of ROC curves of discrimination results for Raman spectra utilizing the Raman intensity ratio of I_{875}/I_{1450}, PCA-LDA and the random forests algorithm. The integration areas under the ROC curves are 0.88, 0.98, and 0.95 for intensity ratio algorithm, PCA-LDA-based, and random forests-based diagnostic algorithm and intensity ratio algorithm, respectively, demonstrating the efficacy of PCA-LDA algorithms for tissue classification.

Intensity ratio approach can only employ a limited number of Raman peaks for tissue diagnosis, most of the biomolecular information contained in the Raman spectra has not been used for spectral analysis. Since biological tissue is complex, it is likely that there are many biochemical species influencing diseases concurrently [12]. Therefore, sophisticated chemometrics techhnique such as PCA-LDA and random forests that utilizes the entire spectrum to determine the most diagnostically significant spectral features can improve the diagnostic efficiency of Raman technique for tissue analysis and classification. In this study, both techniques have been demonstrated to possess comparable diagnostic utility for gastric precancer diagnosis. Random forests technique provides a unique advantage over the conventional PCA-LDA algorithm but furnishing distinctive spectral regions that are optimal for tissue differentiation can be identified and related to particular biochemical and biomolecular changes (e.g., proteins, nucleic acid, carbohydrates) associated with pre-neoplastic transformation. Note that only 5 prominent features are revealed to be of significant for precancerous lesion detection; where up to 20 Raman features are found to be of diagnostic significant for diagnosing cancer (Chapter 4). The less Raman features revealed in the variable importance plot for precancerous lesions as compared to cancerous tissues indicated that there are less biomolecular changes involving the precancerous lesions as compared to the cancerous lesions. This is in agreement with histopathological signs of carcinogenesis transformation, whereby increase biomolecular activity will occur with the degree of neoplastic transformation [60]. Therefore, random forests could be a useful approach to identify the origins of biochemical/biomolecular changes of Raman spectra for tissue carcinogenesis analysis. This study confirmed the feasibility of NIR Raman spectroscopy in conjunction with random forests for providing diagnostic information necessary for distinguishing precancer from normal tissue.

CHAPTER 9

CONCLUSION AND FUTURE RESEARCH

In this dissertation, we have demonstrated the feasibility of using an in-house developed NIR Raman system for human tissue characterization [12]. In order to facilitate identification of the origins of biochemical/biomolecular changes of Raman spectra for diseased tissue analysis, as well as providing high accuracy diagnostic algorithms for tissue classification, we have introduced the utilization of recursive partitioning technique with NIR Raman spectroscopy for detection of gastric cancer from normal gastric tissues [14]. To enhance the performance (i.e., stability and accuracy) of recursive partitioning technique, ensemble technique was also successfully deployed to be used with NIR Raman spectroscopy for cancer diagnosis [10]. We further assessed the diagnostic performance of random forests in comparison with the use of simplistic empirical method which employs Raman peak intensity ratios and multivariate statistical techniques (i.e., PCA-LDA) for gastric dysplasia diagnosis [12, 16]. ROC curves confirmed that PCA-LDA and random forests techniques have comparable overall diagnostic accuracy rate which are more superior compared to the empirical approach for detection of gastric dysplasia from normal gastric tissues. Overall, this dissertation demonstrated the potential of NIR Raman spectroscopy with sophisticated chemometrics algorithms, particularly the random forests, to construct clinically interpretable diagnostic algorithm which can also yield high diagnostic accuracy for rapid diagnosis of precancer and cancer tissues based on the optical evaluation of spectral features of biomolecules.

In a clinical situation, there is often a need to determine the type of tissues into more than two groups such as various grades of tumor [11]. However, in this dissertation, we have only evaluated binary classification such as precancer *vs*. normal and cancer *vs*. normal. Further studies to investigate the possibility of simultaneously classifying Raman tissue spectra into more than two classes (e.g., normal *vs*. dysplasia *vs*. cancer) with different diagnostic algorithms [148, 149], especially the random forests are warranted. In addition, this study has only assessed the potential of Raman spectroscopy for dysplasia and cancer diagnosis. Evaluation on the feasibility of Raman spectroscopy for detection of diseases accompanying the entire carcinogenesis cascade before the onset of dysplastic changes should also be carried out [149]. Due to the limitation of size of the Raman probe used in the project, this dissertation has only focused on the evaluation of *ex vivo* tissues samples. With the development of miniaturized Raman probes for the collection of tissue Raman signals in a few seconds via endoscope, *in vivo* tissue Raman evaluation of the feasibility for detection of different lesions associated with cancer in the gastric and larynx ought to be carried out [50]. It is expected that NIR Raman endoscopic spectroscopy could be a clinically promising tool for the rapid, noninvasive, *in vivo* diagnosis and detection of gastric and laryngeal lesions at the molecular level in clinical endoscopy.

BIBLIOGRAPHY

1. Stone N, Kendall C, Smith J, Crow P, Barr H. Raman spectroscopy for identification of epithelial cancers. Faraday Discuss. 2004; 126: 141-57.

2. Huang Z, Lui H, McLean DI, Korbelik M, Zeng H. Raman spectroscopy in combination with background near-infrared autofluorescence enhances the *in vivo* assessment of malignant tissues. Photochem Photobiol 2005; 81:1219-26.

3. Huang Z, Zheng W, Xie S, Chen R, Zeng H, McLean DI, Lui H. Laser-induced autofluorescence microscopy of normal and tumor human colonic tissue. Int J Oncol 2004; 24: 59-64.

4. Georgakoudi I, Jacobson BC, Van Dam J, Backman V, Wallace MB, Müller MG, Zhang Q, Badizadegan K, Sun D, Thomas GA, Perelman LT, Feld MS. Fluorescence, reflectance, and light-scattering spectroscopy for evaluating dysplasia in patients with Barrett's esophagus. Gastroenterology 2001; 120:1620-1629.

5. Huang Z, McWilliams A, Lui H, McLean DI, Lam S, Zeng H. Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. Int J Cancer 2003; 107:1047-1052.

6. Mahadevan-Jansen A, Richards-Kortum R. Raman spectroscopy for the detection of cancers and precancers J Biomed Opt 1996; 1:31-70.

7. Stone N, Kendall C, Sheperd N, Crow P; Barr H. Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers. J Raman Spectrosc 2002; 33:564-573.

8. Bakker Schut TC, Witjes MJ, Sterenborg HJ, Speelman OC, Roodenburg JL, Marple ET, Bruining HA, Puppels GJ. *In vivo* detection of dysplastic tissue by Raman spectroscopy. Anal Chem 2000; 72:6010-6018.

9. Molckovsky A, Song LM, Shim MG, Marcon NE, Wilson BC. Diagnostic potential of near-infrared Raman spectroscopy in the colon: differentiating adenomatous from hyperplastic polyps. Gastrointest Endosc 2003; 57:396-402.

10. Teh SK, Zheng W, Lau PD, Huang Z. Spectroscopic diagnosis of laryngeal carcinoma using near-infrared Raman spectroscopy and random recursive partitioning ensemble techniques. Analyst 2009; 134: 1323-9.

11. Widjaja E, Zheng W, Huang Z. Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines. Int J Oncol. 2008; 32: 653-62.

12. Teh SK, Zheng W, Ho KY, Teh M, Yeoh KG, Huang Z. Diagnostic potential of near-infrared Raman spectroscopy in the stomach: differentiating dysplasia from normal tissue. Br J Cancer. 2008; 98: 457-65.

13. Kendall C, Stone N, Shepherd N, Geboes K, Warren B, Bennett R, Barr H. Raman spectroscopy, a potential tool for the objective identification and classification of neoplasia in Barrett's oesophagus. J Pathol. 2003; 200:602-09.

14. Teh SK, Zheng W, Ho KY, Teh M, Yeoh KG. Huang Z. Diagnosis of gastric cancer using near-infrared Raman spectroscopy and classification and regression tree techniques. J Biomed Opt. 2008; 13: 034013.

15. Manoharan R, Baraga JJ, Feld MS, Rava RP. Quantitative histochemical analysis of human artery using Raman spectroscopy. J Photochem Photobiol B. 1992; 16: 211-33.

16. Teh SK, Zheng W, Ho KY, Teh M, Yeoh KG, Huang Z. Near-infrared Raman spectroscopy for gastric precancer diagnosis. J Raman Spectrosc 2009; 40: 908-914.

17. Mahadevan-Jansen A, Mitchell MF, Ramanujam N, Malpica A, Thomsen S, Utzinger U, Richards-Kortum R. Near-infrared Raman spectroscopy for *in vitro* detection of cervical precancers. Photochem Photobiol 1998: 68; 123-32.

18. Mourant JR, Short KW, Carpenter S, Kunapareddy N, Coburn L, Powers TM, Freyer JP. Biochemical differences in tumorigenic and nontumorigenic cells measured by Raman and infrared spectroscopy. J Biomed Opt 2005: 10; 031106

19. Shetty G, Kendall C, Shepherd N, Stone N, Barr H. Raman spectroscopy: elucidation of biochemical changes in carcinogenesis of oesophagus. B Jr Cancer 2006: 94; 1460-64.

20. Utzinger U, Heintzelman DL, Mahadevan-Jansen A, Malpica A, Follen M, Richards–Kortum R. Near-infrared Raman spectroscopy for *in vivo* detection of cervical precancers. Appl Spectrosc 2001: 55; 955-59

21. Lau DP, Huang Z, Lui H, Anderson DW, Berean K, Morrison MD, Shen L, Zeng H. Raman spectroscopy for optical diagnosis in the larynx – preliminary findings. Lasers Surg Med 2005: 37;192-200

22. Hanlon EB, Manoharan R, Koo TW, Shafer KE, Motz JT, Fitzmaurice M, Kramer JR, Itzkan I, Dasari RR, Feld MS. Prospects for *in vivo* Raman spectroscopy. Phys Med Biol. 2000; 45: 1-59.

23. Nijssen A, Koljenović S, Bakker Schut TC, Caspers PJ, Puppels GJ. Towards oncological application of Raman spectroscopy. J Biophotonics. 2009; 2: 29-36.

24. Stone N, Stavroulaki P, Kendall C, Birchall M, Barr H. Raman spectroscopy for early detection of laryngeal malignancy: preliminary results. Laryngoscope. 2000; 110: 1756-63.

25. Ling XF, Weng SF, Li WH, Zhi X, Hammaker M, Fateley WG, Wang F, Zhou XS, Soloway RD, Ferraro JR, Wu JG. Investigation of normal and malignant tissue samples from the human stomach using Fourier transform Raman spectroscopy. Appl Spectrosc 2002*:* 56: 570-73

26. Kumar KK, Anand A, Chowdary MVP, Keerthi, Kurien J, Krishna CM, Mathew S. Discrimination of normal and malignant stomach mucosal tissues by Raman spectroscopy: A pilot study. Vib Spectrosc *2007:*44; 382-87

27. Kawabata T, Mizuno T, Okazaki S, Hiramatsu M, Setoguchi T, Kikuchi H, Yamamoto M, Hiramatsu Y, Kondo K, Baba M, Ohta M, Kamiya K, Tanaka T, Suzuki S, Konno H. Optical diagnosis of gastric cancer using near-infrared multichannel Raman spectroscopy with a 1064-nm excitation wavelength. J Gastroenterol. 2008; 43: 283-90.

28. Hu Y, Shen A, Jiang T, Ai Y, Hu J. Classification of normal and malignant human gastric mucosa tissue with confocal Raman microspectroscopy and wavelet analysis. Spectrochim Acta A Mol Biomol Spectrosc. 2008; 69: 378-82.

29. Zheng F, Qin Y, Chen K. Sensitivity map of laser tweezers Raman spectroscopy for single-cell analysis of colorectal cancer. J Biomed Opt. 2007; 12: 034002.

30. Haka AS, Volynskaya Z, Gardecki JA, Nazemi J, Lyons J, Hicks D, Fitzmaurice M, Dasari RR, Crowe JP, Feld MS. *In vivo* margin assessment during partial mastectomy breast surgery using Raman spectroscopy. Cancer Res. 2006; 66: 3317-22.

31. Raman C, Kirishnan K. A new type of secondary radiation. Nature 1928; 12: 501-502.

32. Haka AS, Shafer-Peltier KE, Fitzmaurice M, Crowe J, Dasari RR, Feld MS. Diagnosing breast cancer by using Raman spectroscopy. Proc Natl Acad Sci U S A. 2005; 102: 12371-6.

33. de Jong BW, Schut TC, Maquelin K, van der Kwast T, Bangma CH, Kok DJ, Puppels GJ. Discrimination between nontumor bladder tissue and tumor by Raman spectroscopy. Anal Chem. 2006; 78: 7761-9.

34. Haka AS, Shafer-Peltier KE, Fitzmaurice M, Crowe J, Dasari RR, Feld MS. Identifying microcalcifications in benign and malignant breast lesions by probing differences in their chemical composition using Raman spectroscopy. Cancer Res. 2002; 62: 5375-80.

35. Kendall C, Isabelle M, Bazant-Hegemark F, Hutchings J, Orr L, Babrah J, Baker R, Stone N. Vibrational spectroscopy: a clinical tool for cancer diagnostics. Analyst. 2009; 134: 1029-45.

36. Ellis DI, Goodacre R. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. Analyst. 2006; 131: 875-85.

37. Lord RC, Yu NT. Laser-excited Raman spectroscopy of biomolecules. I. Native lysozyme and its constituent amino acids. J Mol Biol. 1970; 50: 509-24.

38. Choo-Smith LP, Edwards HG, Endtz HP, Kros JM, Heule F, Barr H, Robinson JS Jr, Bruining HA, Puppels GJ. Medical applications of Raman spectroscopy: from proof of principle to clinical implementation. Biopolymers. 2002; 67: 1-9.

39. Krafft C, Steiner G, Beleites C, Salzer R. Disease recognition by infrared and Raman spectroscopy. J Biophotonics. 2009; 2: 13-28.

40. Manoharan R, Shafer K, Perelman L, Wu J, Chen K, Deinum G, Fitzmaurice M, Myles J, Crowe J, Dasari RR, Feld MS. Raman spectroscopy and fluorescence photon migration for breast cancer diagnosis and imaging. Photochem Photobiol. 1998; 67: 15-22.

41. Alfano RR, Liu CH, Sha WL, Zhu D, Akins L, Cleary J, Prudente R, Cellmer E. Human breast tissues studied by IR Fourier transform Raman spectroscopy. Lasers Life Sci 1991; 4: 23–8.

42. Liu CH, Da BB, Glassman WL Sha, Tang GC, Yoo KM, Zhu HR, Akins DL, Lubicz SS, Cleary J, Prudente R, Celmer E, Caron A, Alfano RR. Raman, fluorescence and time-resolved light-scattering as optical diagnostic techniques to separate diseased and normal biomedical media. J. Photochem. Photobiol. B: Biol. 1992; 16: 187-209.

43. Schrader B, Keller S, Loechte T, Fendel S, Moore DS, Simon A, Sawatzki J. NIR FT Raman spectroscopy in medical diagnosis. J Mol Struct 1995; 348: 293–6.

44. Nie S, Bergbauer KJ, Ho JJ, Kuck JFR Jr,. Yu YT. Application of near-infrared Fourier transform Raman spectroscopy in biology and medicine. Spectroscopy 1990; 5: 24–32.

45. Zhao J. Image curvature correction and cosmic removal for high-throughput dispersive Raman spectroscopy. Appl Spectrosc. 2003; 57: 1368-75.

46. Huang Z, Zeng H, Hamzavi I, McLean DI, Lui H. Rapid near-infrared Raman spectroscopy system for real-time *in vivo* skin measurements. Opt Lett 2001; 26:1782-84.

47. Motz JT, Hunter M, Galindo LH, Gardecki JA, Kramer JR, Dasari RR, Feld MS. Optical fiber probe for biomedical Raman spectroscopy. Appl Opt. 2004; 43: 542-54.

48. Wilson BC. Detection and treatment of dysplasia in Barrett's esophagus: a pivotal challenge in translating biophotonics from bench to bedside. J Biomed Opt. 2007; 12: 051401.

49. Short MA, Lam S, McWilliams A, Zhao J, Lui H, Zeng H. Development and preliminary results of an endoscopic Raman probe for potential *in vivo* diagnosis of lung cancers. Opt Lett. 2008; 33: 711-3.

50. Huang Z, Teh SK, Zheng W, Mo J, Lin K, Shao X, Ho KY, Teh M, Yeoh KG. Integrated Raman spectroscopy and trimodal wide-field imaging techniques for real-time *in vivo* tissue Raman measurements at endoscopy. Opt Lett. 2009; 34: 758-60.

51. Santos LF, Wolthuis R, Koljenović S, Almeida RM, Puppels GJ. Fiber-optic probes for *in vivo* Raman spectroscopy in the high-wavenumber region. Anal Chem. 2005; 77: 6747-52.

52. Koljenović S, Bakker Schut TC, Wolthuis R, de Jong B, Santos L, Caspers PJ, Kros JM, Puppels GJ. Tissue characterization using high wave number Raman spectroscopy. J Biomed Opt. 2005; 10: 031116.

53. Matousek P, Towrie M, Stanley A, Parker W. Efficient rejection of fluorescence from Raman spectra using picosecond Kerr gating. Appl Spectrosc. 1999; 53: 1485-9.

54. Shim MG, Wilson. Development of an *in vivo* Raman spectroscopic system for diagnostic applications. J Raman Spectrosc 1997; 28: 131-42.

55. Matousek P, Towrie M, Stanley A, Parker W. Simple reconstruction algorithm for shifted excitation Raman difference spectroscopy. Appl Spectrosc. 2005; 59: 848-51.

56. McCain ST, Willett RM, Brady DJ. Multi-excitation Raman spectroscopy technique for fluorescence rejection. Opt Express. 2008; 16: 10975-91.

57. Bell SEJ, Bourguignon SO, Dennis Andrew. Analysis of luminescent samples using subtracted shifted Raman spectroscopy. Analyst 1998; 123: 1729-34.

58. Mosier-Boss PA, Lieberman SH, Newbery R. Shifted-spectra, edge detection, and FFT filtering techniques. Appl Spectrosc. 1995; 49: 630-8.

59. Lieber CA, Mahadevan-Jansen A. Automated method for subtraction of fluorescence from biological Raman spectra. Appl Spectrosc. 2003; 57: 1363-7.

60. Cao A, Pandya AK, Serhatkulu GK, Weber RE, Houbei Dai, Thakur JS, Naik VM, Naik R, Auner GW, Rabah R, Freeman DC. A robust method for automated background subtraction of tissue fluorescence. J Raman Spectrosc 2007; 38: 1199-205.

61. Zhao J, Lui H, McLean DI, Zeng H. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. Appl Spectrosc. 2007; 61: 1225-32.

62. Beier BD, Berger AJ. Method for automated background subtraction from Raman spectra containing known contaminants. Analyst. 2009; 134: 1198-202.

63. Afseth NK, Segtnan VH, Wold JP. Raman spectra of biological samples: A study of preprocessing methods. Appl Spectrosc. 2006 Dec; 60: 1358-67.

64. Dollish FR, Fateley WG, Bentley FF. Characteristic Raman frequencies of organic compounds. Wiley, New York (1974).

65. Keller MD, Kanter EM, Lieber CA, Majumder SK, Hutchings J, Ellis DL, Beaven RB, Stone N, Mahadevan-Jansen A. Detecting temporal and spatial effects of epithelial cancers with Raman spectroscopy. Dis Markers. 2008; 25: 323-37.

66. Kanter EM, Majumder SK, Vargis E, Robichaux-Viehoever A, Kanter GJ, Shappell H, Jones H 3rd, Mahadevan-Jansen A. Multi-class discrimination of cervical precancers using Raman spectroscopy. J. Raman Spectrosc. 2009; 40: 205–11.

67. Robbins SL, Cotran RS, Kumar V. Pathologic Basis of Disease. Saunders WB: Philadelphia (1994).

68. Crow P, Uff JS, Farmer JA, Wright MP, Stone N. The use of Raman spectroscopy to identify and characterize transitional cell carcinoma in vitro. BJU Int. 2004; 93: 1232-6.

69. Crow P, Molckovsky A, Stone N, Uff J, Wilson B, WongKeeSong LM. Assessment of fiberoptic near-infrared raman spectroscopy for diagnosis of bladder and prostate cancer. Urology. 2005; 65: 1126-30.

70. Stone N, Hart Prieto MC, Crow P, Uff J, Ritchie AW. The use of Raman spectroscopy to provide an estimation of the gross biochemistry associated with urological pathologies. Anal Bioanal Chem. 2007; 387: 1657-68.

71. Koljenović S, Choo-Smith LP, Bakker Schut TC, Kros JM, van den Berge HJ, Puppels GJ. Discriminating vital tumor from necrotic tissue in human glioblastoma tissue samples by Raman spectroscopy. Lab Invest. 2002; 82: 1265-77.

72. Banerjee HN, Zhang L. Deciphering the finger prints of brain cancer astrocytoma in comparison to astrocytes by using near infrared Raman spectroscopy. Mol Cell Biochem. 2007; 295: 237-40.

73. Frank CJ, Redd DC, Gansler TS, McCreery RL. Characterization of human breast biopsy specimens with near-IR Raman spectroscopy. Anal Chem. 1994; 66: 319-26.

74. Stone N, Matousek P. Advanced transmission Raman spectroscopy: a promising tool for breast disease diagnosis. Cancer Res. 2008; 68: 4424-30.

75. Stone N, Baker R, Rogers K, Parker AW, Matousek P. Subsurface probing of calcifications with spatially offset Raman spectroscopy (SORS): future possibilities for the diagnosis of breast cancer. Analyst. 2007; 132: 899-905.

76. Robichaux-Viehoever A, Kanter E, Shappell H, Billheimer D, 3rd. Jones H, Mahadevan-Jansen A. Characterization of Raman spectra measured *in vivo* for the detection of cervical dysplasia. Appl Spectrosc. 2007; 61: 986-93.

77. Mahadevan-Jansen A, Mitchell MF, Ramanujam N, Utzinger U, Richards-Kortum R. Development of a fiber optic probe to measure NIR Raman spectra of cervical tissue *in vivo.* Photochem Photobiol. 1998; 68: 427-31.

78. Viehoever AR, Anderson D, Jansen D, Mahadevan-Jansen A. Organotypic raft cultures as an effective *in vitro* tool for understanding Raman spectral analysis of tissue. Photochem Photobiol. 2003; 78: 517-24.

79. Kanter EM, Majumder S, Kanter GJ, Woeste EM, Mahadevan-Jansen A. Effect of hormonal variation on Raman spectra for cervical disease detection. Am J Obstet Gynecol. 2009; 200: 512.e1-5.

80. Jess PR, Smith DD, Mazilu M, Dholakia K, Riches AC, Herrington CS. Early detection of cervical neoplasia by Raman spectroscopy. Int J Cancer. 2007; 121: 2723-8.

81. Vidyasagar MS, Maheedhar K, Vadhiraja BM, Fernendes DJ, Kartha VB, Krishna CM. Prediction of radiotherapy response in cervix cancer by Raman spectroscopy: a pilot study. Biopolymers. 2008; 89: 530-7.

82. Martinho Hda S, Monteiro da Silva CM, Yassoyama MC, Andrade Pde O, Bitar RA, Santo AM, Arisawa EA, Martin AA. Role of cervicitis in the Raman-based optical diagnosis of cervical intraepithelial neoplasia. J Biomed Opt. 2008; 13: 054029.

83. Shim MG, Song LM, Marcon NE, Wilson BC. *In vivo* near-infrared Raman spectroscopy: demonstration of feasibility during clinical gastrointestinal endoscopy. Photochem Photobiol. 2000; 72: 146-50.

84. Boere IA, Bakker Schut TC, Boogert JVD, Bruin RWFD, Puppels GJ. Use of fibre optic probes for detection of Barrett's epithelium in the rat oesophagus by Raman spectroscopy. Vib. Spectrosc. 2003; 32; 47-55.

85. Lau DP, Huang Z, Lui H, Man CS, Berean K, Morrison MD, Zeng H. Raman spectroscopy for optical diagnosis in normal and cancerous tissue of the nasopharynx-preliminary findings. Lasers Surg Med. 2003; 32 :210-4.

86. Bakker Schut TC, Puppels GJ, Kraan YM, Greve J, van der Maas LL, Figdor CG. Intracellular carotenoid levels measured by Raman microspectroscopy: comparison of lymphocytes from lung cancer patients and healthy individuals. Int J Cancer. 1997; 74: 20-5.

87. Yamazaki H, Kaminaka S, Kohda E, Mukai M, Hamaguchi HO. The diagnosis of lung cancer using 1064-nm excited near-infrared multichannel Raman spectroscopy. Radiat Med. 2003; 21: 1-6.

88. Jess PR, Mazilu M, Dholakia K, Riches AC, Herrington CS. Optical detection and grading of lung neoplasia by Raman microspectroscopy. Int J Cancer. 2009; 124: 376-80.

89. Magee ND, Villaumie JS, Marple ET, Ennis M, Elborn JS, McGarvey JJ. *Ex vivo* diagnosis of lung cancer using a Raman miniprobe. J Phys Chem B. 2009; 113: 8137-41.

90. Oliveira AP, Bitar RA, Silveira L, Zângaro RA, Martin AA. Near-infrared Raman spectroscopy for oral carcinoma diagnosis. Photomed Laser Surg. 2006; 24: 348-53.

91. Krishna CM, Sockalingum GD, Kurien J, Rao L, Venteo L, Pluot M, Manfait M, Kartha VB. Micro-Raman spectroscopy for optical pathology of oral squamous cell carcinoma. Appl Spectrosc. 2004; 58: 1128-35.

92. Malini R, Venkatakrishna K, Kurien J, Pai KM, Rao L, Kartha VB, Krishna CM. Discrimination of normal, inflammatory, premalignant, and malignant oral tissue: a Raman spectroscopy study. Biopolymers. 2006; 81: 179-93.

93. Gniadecka M, Wulf HC, Nielsen OF, Christensen DH, Hercogova J. Distinctive molecular abnormalities in benign and malignant skin lesions: studies by Raman spectroscopy. Photochem Photobiol. 1997; 66: 418-23.

94. Gniadecka M, Philipsen PA, Sigurdsson S, Wessel S, Nielsen OF, Christensen DH, Hercogova J, Rossen K, Thomsen HK, Gniadecki R, Hansen LK, Wulf HC. Melanoma diagnosis by Raman spectroscopy and neural networks: structure alterations in proteins and lipids in intact cancer tissue. J Invest Dermatol. 2004; 122: 443-9.

95. Sigurdsson S, Philipsen PA, Hansen LK, Larsen J, Gniadecka M, Wulf HC. Detection of skin cancer by classification of Raman spectra. IEEE Trans Biomed Eng. 2004; 51: 1784-93.

96. Nijssen A, Maquelin K, Santos LF, Caspers PJ, Bakker Schut TC, den Hollander JC, Neumann MH, Puppels GJ. Discriminating basal cell carcinoma from perilesional skin using high wave-number Raman spectroscopy. J Biomed Opt. 2007; 12: 034004.

97. Lieber CA, Majumder SK, Ellis DL, Billheimer DD, Mahadevan-Jansen A. *In vivo* nonmelanoma skin cancer diagnosis using Raman microspectroscopy. Lasers Surg Med. 2008; 40: 461-7.

98. Lieber CA, Majumder SK, Billheimer D, Ellis DL, Mahadevan-Jansen A. Raman microspectroscopy for skin cancer detection *in vitro*. J Biomed Opt. 2008; 13: 024013.

99. True L, Coleman I, Hawley S, Huang CY, Gifford D, Coleman R, Beer TM, Gelmann E, Datta M, Mostaghel E, Knudsen B, Lange P, Vessella R, Lin D, Hood L, Nelson PS. A molecular correlate to the Gleason grading system for prostate adenocarcinoma. Proc Natl Acad Sci U S A. 2006; 103: 10991-6.

100. Crow P, Stone N, Kendall CA, Uff JS, Farmer JA, Barr H, Wright MP. The use of Raman spectroscopy to identify and grade prostatic adenocarcinoma in vitro. Br J Cancer. 2003; 89: 106-8.

101. Dillion RW, Goldstein M. Multivariate analysis: methods and applications. John Wiley and Sons: New York (1984).

102. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: Wiley (1989).

103. Vapnik VN. Statistical Learning Theory. Wiley: New York (1998).

104. Cortes C, Vapnik VN. Support vector networks. Mach. Learn. 1995; 20: 273-97.

105. Briman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. *CA*: Wadsworth, Belmont (1984).

106. Luk JM, Lam BY, Lee NP, Ho DW, Sham PC, Chen L, Peng J, Leng X, Day PJ, Fan ST. Artificial neural networks and decision tree model analysis of liver cancer proteomes. Biochem Biophys Res Commun. 2007; 361: 68-73.

107. Garzotto M, Beer TM, Hudson RG, Peters L, Hsieh YC, Barrera E, Klein T, Mori M. Improved detection of prostate cancer using classification and regression tree analysis. J Clin Oncol. 2005; 23: 4322-9.

108. Zhang YF, Wu DL, Guan M, Liu WW, Wu Z, Chen YM, Zhang WZ, Lu Y. Tree analysis of mass spectral urine profiles discriminates transitional cell carcinoma of the bladder from noncancer patient. Clin Biochem. 2004; 37: 772-9.

109. Hess KR, Abbruzzese MC, Lenzi R, Raber MN, Abbruzzese JL. Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. Clin Cancer Res. 1999; 5: 3403-10.

110. Zlobec I, Steele R, Nigam N, Compton CC. A predictive model of rectal tumor response to preoperative radiotherapy using classification and regression tree methods. Clin Cancer Res. 2005; 11: 5440-3.

111. Dettling M, Bühlmann P. Boosting for tumor classification with gene expression data. Bioinformatics. 2003; 19: 1061-9.

112. Amaratunga D, Cabrera J, Lee YS. Enriched random forests. Bioinformatics. 2008; 24: 2010-4.

113. Breiman L. Mach. Learning. 2001; 45: 5-32.

114. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least-squares procedures. Anal Chem 1964; 36:1627-39.

115. Anya NM, Robert S, Ralph C, Fatima C, Johan AO. Early onset gastric cancer: on the road to unraveling gastric carcinogenesis. Curr Mol Med. 2007; 7: 15-28.

116. Khay GY. How do we improve outcomes for gastric cancer? J Gastroen Hepatol. 2007; 22: 970-2.

117. Lee LA, Cheng AJ, Fang TJ, Huang CG, Liao CT, Chang JT, Li HY. High incidence of malignant transformation of laryngeal papilloma in Taiwan. Laryngoscope 2008; 118: 50-5.

118. Genden EM, Ferlito A, Silver CE, Jacobson AS, Werner JA, Suárez C, Leemans CR, Bradley PJ, Rinaldo A. Evolution of the management of laryngeal cancer. Oral Oncol 2007; 43: 431-9.

119. Cao WF, Zhang LY, Liu MB, Tang PZ, Liu ZH, Sun BC. Prognostic significance of stomatin-like protein 2 overexpression in laryngeal squamous cell carcinoma: clinical, histologic, and immunohistochemistry analyses with tissue microarray. Human Pathology 2007; 38: 747-52.

120. Marioni G, Marchese-Ragona R, Cartei G, Marchese F, Staffieri A. Current opinion in diagnosis and treatment of laryngeal carcinoma. Cancer Treat Rev 2006; 32: 504-15.

121. Kuncheva LI. Combining Pattern Classifiers. Methods and Algorithm. John Wiley and Sons (2004).

122. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 2007; 35: 339-44.

123. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 2003; 19: 1636-43.

124. Zhang QY, Aires-de-Sousa J. Random forest prediction of mutagenicity from empirical physicochemical descriptors. J Chem Inf Model 2007; 47: 1-8.

125. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003; 43: 1947-58.

126. Ulintz PJ, Zhu J, Qin ZS, Andrews PC. Improved classification of mass spectrometry database search results using newer machine learning approaches. Mol Cell Proteomics 2006; 5: 497-509.

127. Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins 2006; 63: 490-500.

128. Diaz-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics 2007; 8: 328.

129. Latino DA, Aires-de-Sousa J. Linking Databases of Chemical Reactions to NMR Data: an Exploration of ^{1}H NMR-Based Reaction Classification. Anal Chem. 2007; 79: 854-62.

130. Webb GI, Zheng Z. Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. IEEE T. Knowl. Data. EN. 2004; 16: 980-91.

131. Ge G, Wong GW. Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. BMC Bioinformatics. 2008; 9: 275.

132. Sheridan RP, Korzekwa KR, Torres RA, Walker MJ. Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9.J Med Chem. 2007; 50: 3173-84.

133. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. 2007; 8: 25.

134. Verschuur HP, Rijksen G, van Oirschot BA, Schipper-Kester GP, Slootweg JP, Staal GE, Hordijk GJ. Protein tyrosine (de-)phosphorylation in head and neck squamous cell carcinoma. Eur Arch Otorhinolaryngol 1994; 251: 12-6.

135. Badizadegan K, Backman V, Boone CW, Crum CP, Dasari RR, Georgakoudi I, Keefe K, Munger K, Shapshav SM, Sheetse EE, Feld MS. Spectroscopic diagnosis and imaging of invisible pre-cancer. Faraday Discuss 2004: 126: 265-79.

136. Short KW, Carpenter S, Freyer JP, Mourant JR. Raman spectroscopy detects biochemical changes due to proliferation in mammalian cell cultures. Biophys J. 2005; 88: 4274-488.

137. Clark CJ, Thirlby RC, Picozzi V Jr, Schembre DB, Cummings FP, Lin E. Current problems in surgery: gastric cancer. Curr Probl Surg. 2006; 43: 566-670.

138. Axon A. Symptoms and diagnosis of gastric cancer at early curable stage. Best Pract Res Clin Gastroenterol. 2006; 20: 697-708.

139. Teh M, Tan KB, Seet BL, Yeoh KG. Study of p53 immunostaining in the gastric epithelium of cagA-positive and cagA-negative Helicobacter pylori gastritis. Cancer. 2002; 95: 499-505.

140. Lauwers GY, Srivastava A. Gastric preneoplastic lesions and epithelial dysplasia. Gastroenterol Clin North Am. 2007; 36: 813-29.

141. Zheng W, Lau W, Cheng C, Soo KC, Olivo M. Optimal excitation-emission wavelengths for autofluorescence diagnosis of bladder tumors. Int J Cancer. 2003; 104: 477-81.

142. Alimova A, Katz A, Sriramoju V, Budansky Y, Bykov AA, Zeylikovich R, Alfano RR. Hybrid phosphorescence and fluorescence native spectroscopy for breast cancer detection. J Biomed Opt. 2007: 12; 014004.

143. Worthington BS, Syrotuck JA, Ahmed SI. Effects of essential amino acid deficiencies on syngeneic tumor immunity and carcinogenesis in mice. J Nutr. 1978; 108: 1402-11.

144. Thomas GJ Jr, Prescott B. Secondary structure of histones and DNA in chromatin. Science. 1977; 197: 385-8.

145. Lauwers GY, Riddell RH. Gastric epithelial dysplasia. Gut. 1999; 45: 784-90.

146. Correa P. A human model of gastric carcinogenesis. Cancer Res. 1988; 48: 3554-60.

147. Slobodan Š. Eigenvalues and principal component loadings or heavily overlapped vibrational spectra. Spectrochim Acta A 2001; 57:323-36.

148. Teh SK, Zheng W, Ho KY, Teh M, Yeoh KG, Huang Z. Near-infrared Raman spectroscopy for early diagnosis and typing of adenocarcinoma in the stomach. Br J Surg 2009; DOI: 10.1002/bjs.6913.

149. Teh SK, Zheng W, Ho KY, Teh M, Yeoh KG, Huang Z. Near-infrared Raman spectroscopy for optical diagnosis in the stomach: Identification of *Helicobacter-pylori* infection and intestinal metaplasia. Int J Cancer 2009; DOI: 10.1002/ijc.24935.