Imperial College London

Applications of Advanced Spectroscopic Imaging to Biological Tissues

Cai Li Song

July 1, 2020

Department of Chemical Engineering

Imperial College London

Supervised by Prof. Sergei G. Kazarian

Submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy in Chemical Engineering of Imperial College London and the Diploma of Imperial College London

I would like to dedicate this thesis to my family for all their help and support.

Declaration of Originality

This thesis is a description of the work carried out in the Department of Chemical Engineering at Imperial College London between October 2017 and March 2020 under the supervision of Professor Sergei Kazarian. Except where acknowledged, the material is the original work of the author and no part of it has been submitted for a degree at any other university.

Copyright Statement

Abstract

The objectives of this research were to develop experimental approaches that can be applied to classify different stages of malignancy in routine formalin-fixed and paraffin-embedded tissues and to optimise the imaging approaches using novel implementations. It is hoped that the approach developed in this research may be applied for early cancer diagnostics in clinical settings in the future in order to increase cancer survival rates. Infrared spectroscopic imaging has recently shown to have great potential as a powerful method for the spatial visualization of biological tissues. This spectroscopic technique does not require sample labelling because its chemical specificity allows the differentiation of biocomponents to be achieved based on their chemical structures. Experiments were performed on 3-μm thick prostate and colon tissues that were deposited on 2 mm-calcium fluoride (CaF_2) which were subsequently deparaffinised.

The samples were measured under IR microscopes, in both transmission and attenuated total reflection (ATR) mode. In transmission, thermo-spectroscopic imaging of the prostate samples was first carried out to investigate the potential of thermography to complement the information obtained from IR spectral. Spectroscopic imaging has made the acquisition of chemical map of a sample possible within a short time span since this approach facilitates the simultaneous acquisition of thousands of spatially resolved infrared spectra. Spectral differences in the lipid region (3000 – 2800 cm^{-1}) were identified between cancer and benign regions within prostate tissues. The governing spectral band for classification was anti-symmetric stretching of CH_2 (2921 cm^{-1}) from PCA analysis. Nonetheless, the difference in tissue emissivity at room temperature was minimal, thus the contrast in the thermal image is low for intra-tissue classification. Besides, the thermal camera could only capture IR light between 3333–2000 cm^{-1}.

To record spectral data between 3900 – 900 cm^{-1} (mid-IR), Fourier transform infrared (FTIR) spectroscopic imaging was used to classify the different stages of colon disease. An automated processing framework was developed, that could achieve an overall classification accuracy of 92.7%. The processing steps included unsupervised k-means clustering of lipid bands, followed by Random Forest (RF) classification using the 'fingerprint' region of the data. The implementation of a correcting lens and the effect of the

RMieS-EMSC correction on the tissue spectra were also investigated, which showed that computational RMieS-EMSC correction was more effective at removing spectral artefacts than the correcting lens.

Furthermore, the effect of the fluctuations of surrounding humidity where the experiments were carried out was studied by using various supersaturated salt solutions. Significant peak changes of the phosphate band were observed, most notably the peak shift of the anti-symmetric stretching of phosphate bands from 1230 cm^{-1} to 1238 cm^{-1} was observed. By regulating and controlling humidity at its lowest, the classification accuracy of the colon specimens was improved without having to resort to alteration on the RF machine learning algorithm.

In the ATR mode, additional apertures were introduced to the FTIR microscope, as a novel means of depth profiling the prostate tissue samples by changing the angle of incidence of IR light beam. Despite the successful attempts in capturing the qualitative information on the change of tissue morphology with the depth of penetration (d_p), the spectral data were not suitable for further processing with machine learning as d_p changes with wavelengths. Apart from the apertures, a 'large-area' germanium (Ge) crystal was introduced to enable simultaneous mapping and imaging of the colon tissue samples. Many advantages of this new implementation were observed, which included improvement in signal-to-noise ratio, uniform distribution, and no impression left on the sample. In short, the research done in this thesis set a groundwork for clinical diagnosis and the novel implementations were transferable to studies of other samples.

Acknowledgements

I would like to thank my supervisor Prof. Sergei Kazarian for giving me the chance to undertake this PhD project. I appreciate his guidance and support throughout my PhD. I am especially grateful for the many opportunities given by him to me to present my work to a wide range of people at seminars and conferences, interact with academics and industry professionals, and gain insight into the current state and direction of current research.

In addition, I would like to thank Prof. Robert Goldin of St. Mary's Hospital for providing the samples for my research study; as well as Prof. Amparo Galindo, Ms. Sarah Payne and other staff members of the Chemical Engineering department for their support throughout my education and I gratefully acknowledge the department for funding my PhD study.

I would also like to thank the present and past members of the Vibrational Spectroscopy and Chemical Imaging group at Imperial College London, Dr. James Kimber, Dr. Patrick Wray, Dr. Andrew Ewing, Dr. Ben Turner, Dr. Alessandra Vichi, and Dr. Martha Vardaki. I benefited from their help, training, advice, and kindness over the years. Also, many thanks to my friends and colleagues, Ms. Hannah Tiernan, Mr. Huiqiang Lu, Mr. James Beattie, and Mr. Guan-Lin Liu for bringing laughter to my PhD research life with their good sense of humour. I truly enjoyed the many interesting scientific and non-scientific discussions we had, be it in the research office, in the lab, or over a cup of coffee in the departmental common room. I also extend my thanks to the colleagues in my office for their camaraderie and for bringing in good food for sharing all the time.

Lastly, I owe much gratitude to my parents, grandparents, and siblings for instilling the values of education and learning in me. This work would not have been possible without their care and devotion throughout my life. I would like to thank my sister who is my role model to achieve higher in my life and my brother who gives me strong emotional support at difficult times.

Contents

List of Figures

List of Tables

List of Symbols

Symbol	Description	Unit
E	Energy of radiation	J
h	Planck's constant	$J\ s$
ν	Frequency of light	s^{-1}
$\widetilde{\nu}$	Wavenumber of light	m^{-1}
V_c	Speed of light	ms^{-1}
k_f	Spring constant	$kg\ s^{-2}$
μ	Effective mass of molecules	kg
n_L	Energy level	–
N	Number of atoms in a molecule	–
A	Absorbance of light beam	–
I	Intensity of light beam	–
ε	Molar absorptivity	$L\ mol^{-1}\ m^{-1}$
c	Concentration of sample	$mol\ L^{-1}$
l	Effective path length	m
δ	Bending vibration	–
ν_s	Symmetric stretching vibration	–
ν_{as}	Asymmetric stretching vibration	–
n_i	Refractive index of sample i	–
d	Diameter of a particle	m

λ	Wavelength of light	m
Q	Mie extinction coefficient	–
θ	Angle of incidence	$^\circ$
d_p	Penetration depth	m
d_e	Effective thickness	m
T	Absolute temperature	K
ε_λ	Emissivity of an object	–
τ	Transmissivity of an object	–
R	Reflectivity of an object	–
σ	Stefan Boltzmann constant	$W\ m^{-2}\ K^{-4}$
a	Area of measurement	m^2
α	Significance level	–
ϕ	Crystal diameter	m

List of Acronyms

ANN	Artifical neural network
ATP	Adenosine triphosphate
ATR	Attenuated total reflection
BPH	Benign prostatic hyperplasia
CaP	Prostate cancer
CEA	Carcinoembryonic antigen
DL	Digital level
DNA	Deoxyribonucleic acid
DRE	Digital rectal examination
EM	Electromagnetic
EMSC	Extended multiplicative signal correction
FAP	Familial adenomatous polyposis
FDA	Food and Drug Administration
FFPE	Formalin fixed paraffin embedded
FFT	Fast fourier transform
FOV	Field of view
FPA	Focal plane array
FTIR	Fourier transform infrared
H&E	Hematoxylin and eosin
HCA	Hierarchical cluster analysis

HeNe	Helium-neon
HPLC	High performance liquid chromatography
HSI	Hyperspectral imaging
IR	Infrared
IRE	Internal reflection element
KK	Kramers-Kronig
LDA	Linear discriminant analysis
MRI	Magnetic resonance imaging
NA	Numerical aperture
OPD	Optical path difference
PCA	Principal component analysis
PDMS	Polydimethylsiloxane
PLS-DA	Partial least squares discriminant analysis
PLSR	Partial least square regression
PMMA	Polymethyl methacrylate
PP	Peak-to-peak
PS	Polystyrene
RF	Random forest
RH	Relative humidity
RMieS	*Resonant* Mie scattering
RNA	Ribonucleic acid
ROC	Receiving operator characteristic
SG	Savitzky-Golay
S/N	Signal-to-noise
TME	Tumour microenvironment
TPR	True positive rate
TRUS	Transrectal ultrasound

UV	Ultraviolet
WHO	World Health Organization
ZPD	Zero path difference

Publications

The results presented in this thesis are described in the following publications.

1. Song, C.L. and Kazarian, S. G. The effect of controlled humidity and tissue hydration on colon cancer diagnostic via FTIR spectroscopic imaging. Analytical Chemistry (2020) 92, 9691 - 9698.
2. Song C.L. and Kazarian S. G. Micro ATR-FTIR spectroscopic imaging of colon biopsies with a large area Ge crystal. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy (2020) 228, 117695.
3. Song C.L., Vardaki M.Z., Goldin R.D., and Kazarian S.G. Fourier transform infrared spectroscopic imaging of colon tissues: evaluating the significance of amide I and C-H stretching bands in diagnostic applications with machine learning. Anal Bioanal Chem (2019) 411, 6969 - 6981.
4. Song C.L. and Kazarian S. G. Three-dimensional depth profiling of prostate tissue by micro ATR-FTIR spectroscopic imaging with variable angle of incidence. Analyst (2019) 144, 2954 – 2964.
5. Song C.L., Ryu M., Morikawa J., Kothari A., Kazarian S. G. Thermal effect on dispersive infrared spectroscopic imaging of prostate cancer tissue. Journal of Biophotonics (2018) 11, e201800187.

Other Publications (not included in this thesis)

1. Byrne B., Beattie J. W., Song C.L., and Kazarian S. G. (2020). 'Chapter 1 - ATR-FTIR Spectroscopy and Spectroscopic Imaging of Proteins', in Ozaki, Y. (ed.) Vibrational Spectroscopy in Protein Research. Elsevier, 1-22.
2. Chan K.L.A., Altharawi A., Fale F., Song C. L., Kazarian S. G. et al. Transmission Fourier Transform Infrared Spectroscopic Imaging, Mapping, and Synchrotron Scanning Microscopy with Zinc Sulfide Hemispheres on Living Mammalian Cells at Sub-Cellular Resolution Applied Spectroscopy (2020) 74, 544-552.

3. Yang, H., Song, C., Lim, Y., Chen, W., Heng, J. Selective crystallisation of carbamazepine polymorphs on surfaces with differing properties. CrystEngComm (2017), 19(44), 6573-6578.

Chapter 1

Introduction

The use of infrared (IR) spectroscopy, particularly the Fourier transform infrared (FTIR), for the imaging of biological samples is an expanding field of research over the last decade, largely due to the improved acquisition speed of the data and the increased availability of processing tools and statistical algorithms to analyse a large dataset (Sreedhar, et al., 2015). In contrast to conventional histology using staining methods to examine tissues under visible microscopes, FTIR spectroscopic imaging technique is a label-free method which has the capability to provide biochemical information about the biological samples under examination, thus providing a more objective analysis. It can be employed to identify chemical bonds in molecules based on the specific IR spectra that are acquired from the samples, which contain distinctive molecular fingerprints proved to be very useful in the analysis of the biological tissues.

In this PhD project, prostate cancer and colon cancer were chosen to be studied because they are the most common cancer among males and the third most common cancer in the UK, respectively (Cancer Research, 2014). Prostate cancer is a type of malignant tumor that occurs in the prostate which is located in men below the bladder and near the rectum, whereas colon cancer may develop via polyp pathways in the colon. These cancers have been subjected to intense study and recent work has highlighted the diversity of the cancer lesions. The Prostate and Bowel Cancer Screening Program in the UK are aimed at detecting and accurately categorising these lesions, as well as the early cancers, as the basis for planning further follow up and treatment (Phalguni, et al., 2015).

Currently, the methods commonly used for screening of colon cancer are bowel visualization and measurement of the faecal biomarkers (Phalguni, et al., 2015), and digital rectal examination for prostate cancer, however the distinctions between the different disease stages made by pathologists remain a challenge as there is a significant degree of discordance, even between experts, based on the visual morphology alone. Failure to accurately identify these cancer lesions may potentially result in over or under treatment,

which should be avoided. This makes prostate and colon cancer important diseases to study. Thus, there is a strong need to develop approaches which can identify early stages of the development of cancer in a way that goes beyond the current approach adopted by pathologists in order to increase the survival rates of patients suffering from these cancers. Advances in biopsy analysis including immunohistochemical, molecular markers and new imaging techniques have successfully identified key biomarkers. However, the complexity, the high variability of these approaches and the insufficient consistency of the outcomes have prevented their adoption in routine clinical use. There is, therefore, a raising need to search for a more cost and clinically effective approach for the study of colon and prostate cancer and FTIR spectroscopy was used to demonstrate that in this research.

In order to assess the FTIR spectra rigorously, multivariate analysis is employed. A variety of different multivariate approaches such as partial least squares (PLS), principal component analysis (PCA), and other clustering methods are often used to classify the tissue biopsies as part of the machine-learning strategies. Benchmarking the performance and improvements of FTIR spectroscopic imaging using multivariate approaches will answer fundamental questions such as: 'How does changing the mode of measurement affect the sensitivity and specificity of classification?' or 'How does the performance of the classification change with adjustments in the statistical processing methodology as well as other factors such as the experimental environment?'. It should also be noted that these developed algorithms do not aim to displace histopathologists completely at this stage of research, rather they will be used to augment the existing complementary data identified by histopathological staining to further increase the diagnostic accuracy. In addition, the move towards diagnosis of a higher accuracy means that novel implementations to the current FTIR spectroscopic technique should be explored to yield spectra of better quality.

The end-goal is the development of a robust and reliable protocol that utilise FTIR spectroscopic imaging to study the tissues for cancer diagnostics. It is hoped that the success of the research presented in this thesis will bring this imaging technique closer towards clinical acceptance.

1.1 Objectives

The main aim of this research is to differentiate and classify the different stages of disease for prostate and colon cancerous tissues using FTIR spectroscopic imaging technique.

The specific objectives of this research are outlined as follows:

- Develop protocols and processing frameworks for FTIR spectroscopic imaging of the

prostate and colon tissues

- Optimize the imaging approaches using novel implementations to obtain qualitative and quantitative information of the biopsy specimens studied
- Validate the results obtained from FTIR spectroscopic imaging by comparison with histopathologically stained images
- Apply multivariate analysis and machine learning to achieve an understanding of the imaging parameters on the reliability of diagnostics and classification

Chapter 2

Literature review

2.1 Cancer biology

Cancer is caused by continual abnormal proliferation of cells which grow and divide uncontrollably to form a tumour, due to a defect in the regulatory systems of such cells. The irregular growth of the cancer cells, if not inhibited in time, could result in them invading the normal surrounding tissues and distant organs and to a more serious extent spread throughout the body in a process known as metastasis (Cooper 2000). In fact, metastasis accounts for approximately 90 % of cancer-related deaths. The high cancer mortality is mainly due to the inability to manage the disease once it disseminates through the body, therefore metastasis is often viewed as one of the terminal stages of cancer (Chaffer & Weinberg 2011). Metastasis follows a chain of sequential steps. To put it briefly, the primary step involves the detachment of these cells from its main tumour and their intravasation into the bloodstream as well as the lymphatic system; follows by the effusion of the metastatic cells at distant capillary bed to settle and form into secondary malignant tumour (Seyfried & Huysentruyt 2013). The difference of the cancer cells from normal cells in their life cycles is detailed in Appendices 6.1.

It is well known that cancer is a complicated genetic disease – that is, cancer is a result of an unnatural mutation or damage to the genes. Genes are segments of DNA, and these segments, in turn, are the blueprint to produce proteins.[1] The bases that make up the backbone of DNA, namely Adenine, Thymine, Guanine, and Cytosine, are the codes that are interpreted. Each sequence of three bases, called a codon, codes for a specific amino acid (Griffiths et al. 2000).[2] Two of the main types of genes that play a

[1] By definition, a gene is a sequence of nucleotides in DNA or RNA that encodes the synthesis of either RNA or protein.

[2] Codon is a triplet of adjacent nucleotides in the messenger RNA chain that codes for a specific amino acid in the synthesis of a protein molecule.

role in cancer induction are oncogenes and tumour suppressor genes (Lodish et al. 2000*b*). Often, proto-oncogenes encode proteins that function to stimulate cell division, inhibit cell differentiation, and halt cell death. These processes are crucial for normal human development and for the maintenance of tissues and organs. When a proto-oncogene mutates or replicates in an uncontrolled manner, it becomes a 'bad' gene that can become permanently turned on or activated when it is not supposed to be. An example of an oncogene is the *HER2* gene that is present in greatly increased numbers in breast cancer tumours as well as some lung cancer tumours (Bose et al. 2013, Cappuzzo et al. 2005). Oncogenes are a major molecular target for anti-cancer drug design (Ke & Shen 2017). Tumour suppressor genes (or anti-oncogenes), on the other hand, function to inhibit the development of tumour and cell proliferation through apoptosis induction. They represent the opposite side of cell growth control (Levine et al. 2008). One of the hallmarks of cancer is the ability of malignant cells to evade apoptosis due to a mutation on the tumour suppressor genes. Examples of tumour suppressor genes include *BRCA* genes and the *p53* gene in the breast and ovarian cancer (Lee & Muller 2010).

A suitable analogy here to understand the contribution of oncogenes and tumour suppressor genes to cancer is to think of the cell as a car. The former can be compared to a gas pedal which controls the speed of the car, much like how fast the cell grows and divide. An oncogene is like a gas pedal that is stuck down, hence the cell continues to divide out of control. On the contrary, in this analogy, a tumour suppressor gene behaves like a brake pedal, which stops the car when it is going too fast. When the gene stops functioning properly, the 'molecular switch' for cellular check cannot be turned off, permitting uncontrolled multiplication of the cell, leading to carcinogenesis and subsequently tumour development. In cancer, tumours, once formed, exhibit another dimension of complexity as they are not just masses of 'malignant' cells but a repertoire of other recruited normal cells, including cells of the immune system, the tumour vasculature and lymphatics, as well as fibroblasts, pericytes and sometimes adipocytes, which together create the tumour microenvironment (TME) via their interactions (Balkwill et al. 2012). The malignant cells thrive in this TME. However, not all tumours are cancerous – they could be 'benign', meaning that they do not invade surrounding tissues nor spread around the body.

Underlying the hallmarks of cancer during the multistep development of tumour are genome instability or in other words genetic mutation. Gene damage can happen in several ways. Simplistically, this may involve missense mutation, nonsense mutation, insertion, deletion, duplication, frameshift mutation, or repeat expansion of the DNA base pair (Clancy 2008). The basic types of genetic mutation are acquired (somatic or sporadic) mutation and hereditary (germline) mutation. The former being the most common one (American Society of Clinical Oncology (ASCO) 2018). Hereditary gene mutation is passed on in family whereas acquired mutation is a genetic change that occur at some point in a person's life and will not be passed on to his/her offspring (Griffiths et al. 2000). Acquired

mutations can also be caused by external factors such as exposure to radiation, for instance, UV rays from the sun or chemicals like carcinogens in tobacco (National Cancer Institute (NIH) 2018).

There were an estimated 18 million cancer cases around the world in 2018, of these 9.5 million cases were in men and 8.5 million in women. The most common cancer (excluding non-melanoma skin cancer) globally were the lung and breast cancers, each contributing 12.3 % of the total number of new cases diagnosed in last year, followed closely by colorectal cancer and prostate cancer (CaP) (World Cancer Research Fund (WCRF) International 2018). Cancer has caused serious global economic and public health burden and is one of the most significant challenges need to be addressed in this century (Brown et al. 2001). Cancer remains the leading cause of deaths worldwide despite improvement in the diagnostic and therapeutic strategies, thus there is a need for ongoing research to understand cancer and aid in its diagnostic (Liu et al. 2017).

2.2 Prostate cancer

The prostate (or prostate gland) is a chestnut-size muscular gland that weighs about three-fourth of an ounce (or 20 grams) which constitutes part of the reproductive and urinary system of a man. It is located deep inside the pelvis and is positioned inferiorly to the neck of the bladder and superiorly to the external urethral sphincter (Canadian Cancer Society 2020). The name 'prostate' reaches back to ancient Greece with similar sounding 'prostatēs', which literally means 'someone who stands before someone or something'. The term describes the position of the prostate gland. From the bottom view where the urethra leaves the gland, the prostate 'stands before' the bladder (Josef Marx & Karenberg 2009). The main function of prostate is the secretion of one of the components of semen, i.e. the prostate fluid (Ashford 2020). During ejaculation, the muscles of the prostate gland help propel this seminal fluid into the urethra. This alkaline fluid provides lubrication and nutrition for the sperm. Moreover, the male sex hormone testosterone is transformed to a biologically active form – dihydrotestosterone (DHT) in the prostate (Carson & Rittmaster 2003), which is an androgen famously known for causing male pattern balding, also called androgenic alopecia (English 2018).

The three major conditions that affect the prostate are benign prostatic hyperplasia (BPH), prostatitis, and prostatic cancer. BPH is a common condition in which the prostate increases in size when a man reaches late 40's. The enlargement of prostate could cause uncomfortable urinary symptoms from the physical compression it exerts on the urethra that results in anatomic bladder outlet obstruction. As the name itself suggests, BPH is not linked to cancer and does not increase the risk of developing CaP; nonetheless, the symptoms of the two are very often very similar (Patel & Parsons 2014). Likewise,

prostatitis is non-cancerous; yet unlike BPH, it can affect men of all ages. Prostatitis is associated with the inflammation of the prostate which can be acute or chronic and may or may not be caused by bacterial infection (Krieger et al. 2008). BPH occurs in the transition zone, whereas most CaP begins in the peripheral zone (Canadian Cancer Society 2020). The anatomy of the prostate is described in Appendices 6.2.

2.2.1 Epidemiology of prostate cancer

CaP is the second most frequent cause of cancer deaths in men in most developed countries after lung cancer, and the incidence has increased significantly over recent years. In 2018, 1,276,106 new cases of CaP are recorded worldwide, 28.1 % of which resulted in deaths (Bray et al. 2018). In the United States itself, the American Cancer Society's estimates for CaP mortality is 31,620, out of the 174,650 new cases reported. This high number of cases suggests that around 1 in 9 men is likely to be diagnosed with this type of cancer during his lifetime (American Cancer Society 2020*b*). In the United Kingdom, over the last decade, CaP incidence rates have increased by around a twentieth (4 %). In fact, the incidence rates for CaP are projected to rise at an alarming rate of 12 % in the UK between 2014 and 2035. CaP can be fatal, but in fact, CaP survival is improving and has tripled in the last 40 years in the UK, most likely due to prostate-specific antigen (PSA) testing (Cancer Research UK 2017*b*).

Aetiology and risk factors

The well-established CaP risk factors are advanced age, ethnicity, genetic factors and family history (Rawla 2019), out of which age is the most important risk factors. The cases of CaP diagnosed under the age of 40 is uncommon; however, the odds increase exponentially with age. About two-thirds of all CaPs are diagnosed in men age 65 and older (WebMD 2020). Prevalence of CaP also varies substantially by race. The highest global incidence rate is confirmed in African descent men, with the lowest rates in Asian men (Rebbeck & Haas 2014). Other than the difference in socioeconomic conditions, biological factor such as the genetic predisposition has been proposed by several studies to be associated with the disparity . In one research study, African-American men are found to have the more common chromosome *8q24* variants, which have been associated with increased CaP risk (Okobia et al. 2011). Some other research studies have also demonstrated that African descent men have a high rate of variations in tumour suppressor genes such as *EphB2* (Robbins et al. 2011) and *BCL2* which regulates cell apoptosis (Hatcher et al. 2009). Besides, family history plays an important role. About 5 – 10 % of all CaP diagnosed are hereditary, meaning that a man with close relative suffering from CaP is at a higher risk of developing it too (Memorial Sloan Kettering Cancer 2020). To

date, research has identified several inheritable genes that could contribute to the increased risk of getting CaP, for instance, mutations in the genes *BRCA1* or *BRCA2* (Castro & Eeles 2012). In one recent study, it was found that patients who harbour germline *BRCA2* mutations exhibit the worse clinical outcomes than non-carriers after treatment (Taylor et al. 2019).

Diagnosis and treatment

The first test for CaP is digital rectal examination (DRE). It is a relatively simple check to help detect abnormality in prostate. For DRE, the size and the hardness of the prostate is felt by a physician who has his finger inserted into the rectum. The proximity of the position of the prostate to the rectum allows this check to be carried out quickly. The second test is equally simple: the prostate-specific antigen (PSA) test. It is a test that measures the level of PSA present in blood. PSA is a glycoprotein enzyme secreted by the epithelial cells of the prostate gland, believed to be instrumental in dissolving seminal coagulum as well as the cervical mucus that aids sperm mobility into the uterus (Arneth 2009). The use of the PSA test in conjunction with a digital rectal exam (DRE) to test asymptomatic men for CaP was first approved by FDA in 1994; However, neither tests are 100 % accurate (National Cancer Institute (NIH) 2017). In the past, an elevated PSA level above 4.0 ng mL^{-1} (conventional threshold) would be associated with potential risk of CaP and biopsy often be recommended consequently; but recent study has disapproved the consensus that PSA is a serum marker for cancer detection by showing that some men with PSA levels below 4.0 ng mL^{-1} have CaP and that many men with higher levels do not have CaP. Other prostate disorders can also cause the PSA level to fluctuate, for instance, a man's PSA level often rises if he has prostatitis or a urinary tract infection (Thompson et al. 2004). The positive predictive value of a PSA test only lies in the range of 20 – 30 % (Hayes & Barry 2014). The next step to screen for CaP is often the biopsy, either through transperineal (TP) biopsy or transrectal (TR) biopsy. The difference being the target site: via the rectum for TR and via the perineum (area between the anus and scrotum) for TP. Due to the high risk of complications arising from bacterial infection from the faecal matter introduced to the prostate in the TR biopsy, TP is nowadays commonly performed in most clinics (Xiang et al. 2019). The biopsy is then sent to the lab to be examined by pathologist and a corresponding Gleason score is assigned in the presence of cancer. In addition, there are other diagnostics methods such as transrectal ultrasound (TRUS) or imaging with MRI, just to name a few (Mayo Foundation for Medical Education and Research (MFMER) 2020*b*). Men diagnosed with localized CaP have few primary options: expectant management ('active surveillance'), surgery, and radiotherapy (or brachytherapy) or the less common cryotherapy. The treatment received is most often largely dependent on the severity of the cancer and its growth rate. For metastatic CaP, these might be chemotherapy, or hormone therapy (Litwin & Tan 2017).

2.2.2 Pathological evaluation

Gleason grading system

An important milestone in the history of pathology occurred on Dec. 28th, 2008 with the death of Dr. Donald F. Gleason from heart attack at the age of 88 years, whose work has been the gold standard for staging CaP. An established prognostic indicator of CaP is the Gleason grading system, which was proposed by Gleason in 1966 and further validated in the year 1974 in a research in which 2,900 patients were involved. The score is based on a pathologist's microscopic examination for the 'Gleason' histology or features of prostate tissue that has been chemically stained after a biopsy (Humphrey 2004). A significant attribute of this scoring system is that it is independent on the assessment of nuclear morphology but is based upon architectural changes that can be sufficiently evaluated at low power magnification (Delahunt et al. 2009). The Gleason patterns are detailed in Table 2.1 and Fig. 2.1.

The Gleason score ranges from 1 – 5 and describes how much the cancer from a biopsy looks like healthy tissue (lower score) or abnormal tissue (higher score). Due to the histological variation within the tumours, a 2-digit system, for example 2-1 or 4-3 is introduced; with the primary score based on the predominant pattern or most common cell morphology by area and the secondary score based on the non-dominant pattern with the highest grade. In cases where a single morphology is identified, the secondary score receives the same designation as the primary score, for example 3-3. The final Gleason score is the addition of these two numbers (Gleason & Mellinger 1974). CaP of Gleason patterns 1 and 2 are rarely seen. Gleason pattern 3 is by far the most common and in large majority of cases, it is found in pure form, such that the most common Gleason score is $3 + 3 = 6$. (Chen & Zhou 2016).The World Health Organization (WHO) endorsed the Gleason grading system in the classification of CaP in 2004 and following that, treatment received by the patients is dictated by the grading score assigned (Montironi et al. 2016).

2.3 Colon cancer

The colon, also known as the large intestine, makes up the last part of the digestive system in a human body. The colon functions to absorb water, salts, and nutrients from the materials that have not been digested and remove the leftover waste product, via a process called peristalsis. The colon is made up of c.a. 1.8 meter-long of dense muscle. The first part of the colon, the cecum, connects to the ileum of the small intestine and the rest of the colons are further divided into four parts – the ascending colon that runs up the right side of the abdomen, transverse colon that travels across the abdomen, the descending colon that travels down the left abdomen, and lastly the curving sigmoid colon

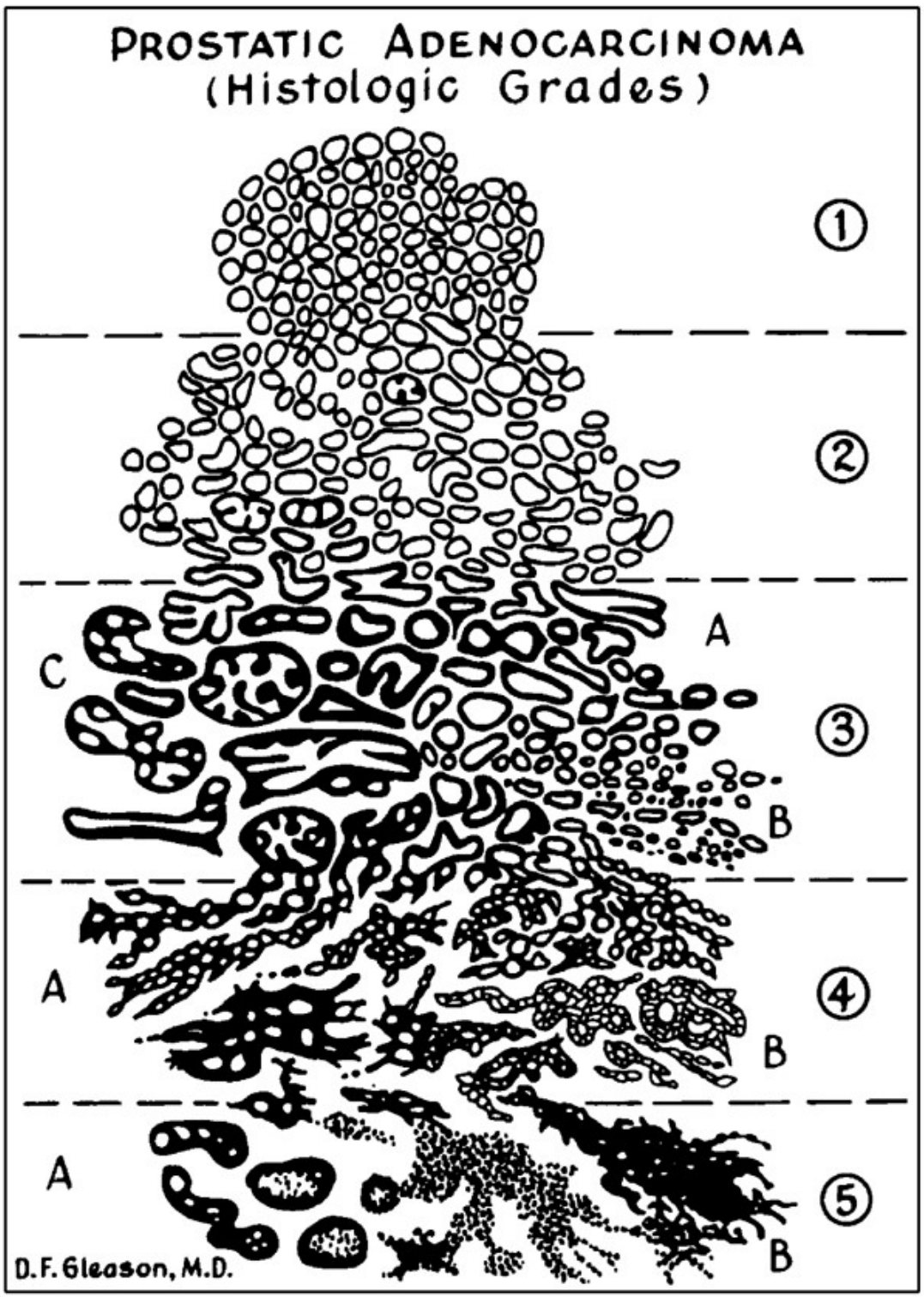


Figure 2.1: The Gleason grading pattern, reproduced from (Humphrey 2004) with the permission from Springer Nature

Table 2.1: Table of features assigned to the different Gleason patterns from the Gleason grading system

Gleason pattern	Corresponding feature(s)
Grade 1	This is a pattern of very well-differentiated growth of small, uniform and closely packed acini, which closely resembles healthy prostate tissue.
Grade 2	The nodular mass of glands is still circumscribed but is less well differentiated. A distinctive characteristic is the increase in variability in gland size and shape compared to grade 1.
Grade 3	This corresponds to a moderately-differentiated carcinoma. Infiltrative tumour with ill-defined edges and irregular extension into stroma can be observed. They are often long or angular.
Grade 4	This is a high grade and poorly differentiated carcinoma growth. The glands look fused together with rare lumen and raggedly infiltrative edges into surrounding tissue. They are difficult to be distinguished from one another.
Grade 5	This comprises of raggedly infiltrative sheet-like growth of anaplastic adenocarcinoma cells and comedo necrosis. The neoplasms have no glandular differentiation.

just before the rectum.

Colon cancer is sometimes referred to as colorectal cancer – a term that combines colon cancer and rectal cancer. It usually begins as polyps in the inside lining of the colon. According to (American Cancer Society 2020*a*), polyps are abnormal growth that are benign. For polyps that have the potential to turn into cancer, they are the adenomatous polyps or pre-cancerous polyps. Hyperplastic polyps, for example, are generally not pre-cancerous; but some in the medical community perceives them as signs of future colon cancer (Bradford 2016). On the other hand, dysplastic polyps can develop into cancer over time. People who suffer from ulcerative colitis or Crohn's disease for many years are at a greater risk of growing this kind of adenomatous polyps. When the cancer starts to develop from the polyps, the malignant cells can start invading the wall of the colon or rectum as time passes on. Typically, this starts in the innermost layer, known as the mucosa, and can grow outward to other layers of the colon. The layers of the colon wall are described in Fig. 2.2.

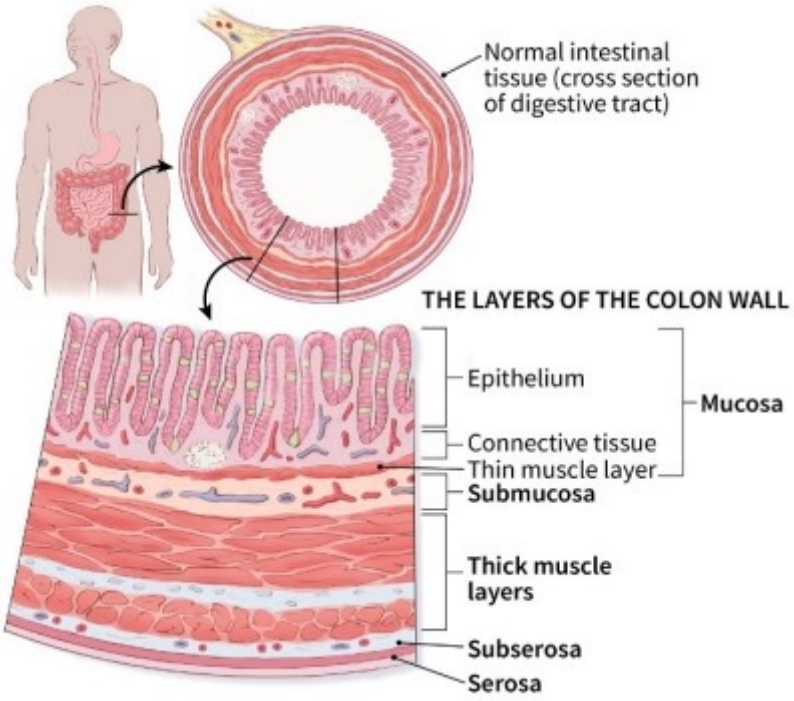


Figure 2.2: The layers of the colon wall. By courtesy of American Cancer Society (ACS), copyright 2020; used with permission

2.3.1 Epidemiology of colon cancer

Colon cancer is the fourth most common cancer in the UK in the year 2016. Between 2014 – 2016, there are approximately 42,000 new cases per year (Cancer Research UK 2017*a*). In the United States, it is the third leading cause of cancer related deaths in men and

women. In the year 2020 alone, it is estimated that colon cancer will cause about 53,200 deaths in the country; however, compared to previous decades, the death rate has been dropping, possibly due to the more frequent screening (colonoscopy) for colon polyps and their subsequent removal before they have the chance to turn into cancer. Besides, the treatment for colon cancer has also improved over the last decade. Despite the drop in the death rate, people younger than the age of 55 have seen an increase of 2 % annually between 2007 and 2016. (American Cancer Society 2020*a*)

Aetiology and risk factors

Compared to other types of cancer, colon cancer has the strongest link to lifestyle-related factors like diet, weight, and exercise. People who are overweight and physically inactive are at a higher risk of getting colon cancer. A diet that is high in red meats or processed food may also raise the risk. Other than these, nearly 1 in 3 people who develop colon cancer has a family history. The most common inherited syndromes linked to colon cancer are Lynch syndrome (hereditary non-polyposis colon cancer, HNPCC). This disorder is caused by a defect in the inherited genes *MLH1* or *MSH2* [3]. The Lynch syndrome made up about 5 % of the people diagnosed with colon cancer, the other 1 % is caused by familial adenomatous polyposis (FAP), mutations in the *APC* gene. Almost all people with this condition suffer from colon cancer if the colon has not been surgically removed by the age of 40 (American Cancer Society 2020*a*).

Diagnosis and treatment

Research is ongoing for early detection of colon cancer to help reduce the mortality rate of this cancer (American Cancer Society 2020*a*). This includes effort to define colon cancer by its sub-categories. In other words, this means classifying based on evaluation on changes that occur to the colon cells, such as the genetic mutations, how the cell looks and behave, the speed of the cells divides, and morphological features of the tumour. It is the ultimate goal to better understand the disease progression and outcome, which can also lead to a higher precision treatment via continuous effort in research in this area. Besides, blood test may also give an insight into the chemicals produced by colon cancer or carcinoembryonic antigen (CEA). Based on the level of CEA present in the blood, medical examiner could understand the prognosis better and provide a way to track how the cancer is responding to the treatment provided. At the very early stage of cancer when the cancer is still contained within the polyps, the polyps are removed during colonoscopy. This is an easy

[3] These genes help repair DNA that has been damaged. People who inherited this condition also have a high risk of developing cancer of the endometrium. The Lynch syndrome has been linked to ovarian, stomach, pancreas, kidney, prostate, breast, and ureters cancer (Centers for Disease Control and Prevention (CDC) 2020).

and quick procedure known as polypectomy. For larger polyps, however, this requires the removal of a small amount of inner lining of the colon close to the polyps (endoscopic mucosal resection) but is still minimally invasive. In situations where the polyps are unable to be removed through colonoscopy, another procedure called the laparoscopic surgery is introduced, which involves partial colectomy, i.e. the removal of part of the colon that contains the cancer. Where surgery is not possible, other treatments are often issued to alleviate the symptoms of colon cancer. These could be chemotherapy, radiation therapy, targeted drug therapy, and immunotherapy (Mayo Foundation for Medical Education and Research (MFMER) 2020*a*).

2.3.2 Pathological evaluation

Number staging system

The most frequently applied staging system is the American Joint Committee (AJCC) TNM system, an abbreviation for **t**umor (T), lymph **n**odes (N), and **m**etastasis (M). The latest grading system is based on the examination of the tissues removed during operation. Once a person's T, N, and M categories have been determined, this information is combined in a process called stage grouping to assign an overall stage. The assignment of the stage is provided in Table 6.1 in Appendices 6.3.

However, the grading systems mentioned above have been subjects of controversy, primarily because they are based upon visual criteria of pattern recognition that are operator-dependent and subject to intra- and inter-observer variability. Thus, there is a need for molecular based techniques to grade tissue samples in a reliable and reproducible manner. In this case, FTIR spectroscopic imaging coupled with statistical pattern recognition techniques could be used in order to demonstrate histopathologic characterization of prostatic tissue and to differentiate benign from malignant prostatic epithelium.

2.4 Fundamentals of IR spectroscopy

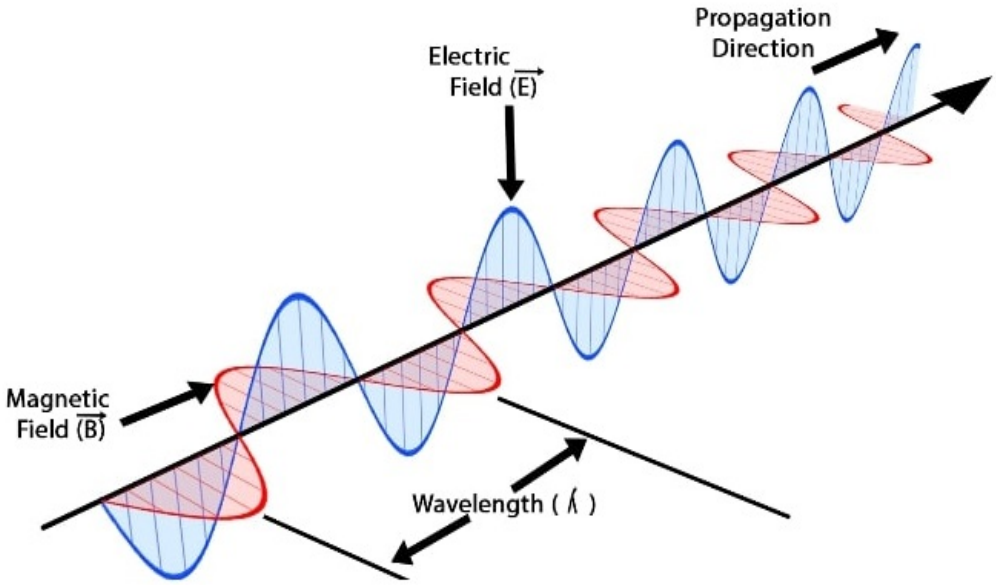
The discovery of infrared (IR) radiation is accredited to Friedrich William Herschel (1738 – 1822), who called IR the 'rays that occasion heat'. In his experiment, he observed that the temperature of the mounted thermometers increased beyond the red portion of the visible light after the sun rays passes through a prism. Shortly after, he discovered that by placing a sample in the path of IR and changing the spectrum that transmitted through the sample, at some point the temperature would drop. His work marked the earliest experiment on the measurement of light absorption in the IR (Rowan-Robinson 2013).

With the formulation of the Maxwell equations in 1861 and the proposal of the electromagnetic light theory in 1864 by Maxwell, the interaction of light with matter could be handled theoretically (Horvath 2009). According to Maxwell classical theory, light is an electromagnetic (EM) radiation, consisting of electric field and magnetic field vector oscillating perpendicularly to each other. These fields are in phase and are being propagated as a sine wave, as shown in Fig.2.3a. The absorption of light is the interaction of the electric field of light with matter. Apart from the classical theory, light can also be explained from the perspective of quantum theory that was explained by Albert Einstein regarding the photoelectric effect of light in the early 20th century. In Einstein's quantum theory, light is composed of particles called photons which have wave-like properties associated with them (Einstein, 1905). A collection of light at different frequency is known as the EM spectrum, illustrated in Fig.2.3b. Infrared (IR) light is defined as a region which encompasses a broad range of wavelength from 700 nm ($\sim$14 000 cm^{-1}) to 1000 μm ($\sim$10 cm^{-1}) (Ball 2003). This IR region lies between the visible light and microwave radiation. IR radiation is further sub-categorized into mid-IR ($\sim$4000 cm^{-1} – $\sim$400 cm^{-1}), far-IR ($\sim$400 cm^{-1} – $\sim$10 cm^{-1}), and near-IR ($\sim$14 000 cm^{-1} – $\sim$4000 cm^{-1}) (Atkins et al. 2019). Mid-IR radiation, which is of particular interest in this thesis, is used to measure infrared spectra as most molecules exhibit fundamental vibrational modes in this region compared to far-IR which corresponds to rotational modes and many other low-frequency vibrations, whilst near-IR corresponds to the overtones and combinations of molecular vibrations (El-Azazy 2019).

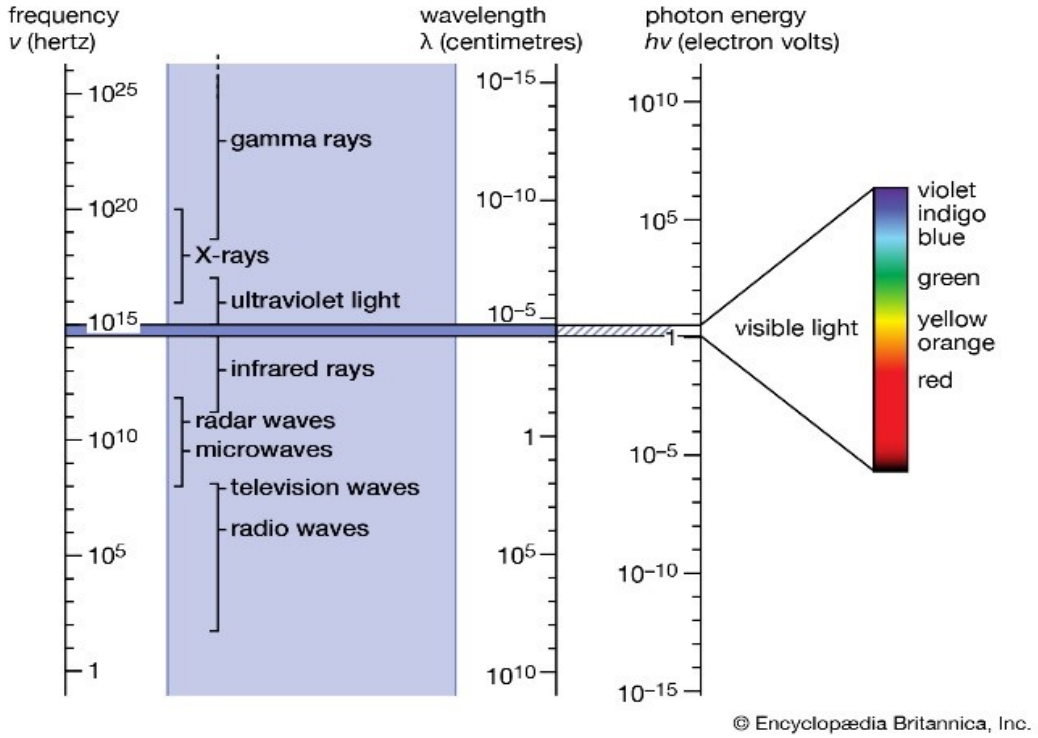
Velocity of propagation of an EM wave in vacuum is constant at $3 \times 10^8\,\mathrm{ms}^{-1}$. This is known as the velocity of light, V_c. If one complete wave travelling a fixed distance each cycle is visualized, it may be observed that the velocity of this wave is related to its wavelength, λ which arise from the periodic modulation of light, such that

$$E = \frac{hV_c}{\lambda} = h\nu = hc\widetilde{\nu} \tag{2.1}$$

where E is the energy of the radiation (or in absorption spectroscopy, this refers to the amount of energy absorbed from the incident radiation to excite a molecule from the ground state to the higher energy excited state); h refers to Planck's constant with a value of 6.626×10^{-34} Js; ν is the frequency of light; and $\widetilde{\nu}$ refers to the wavenumber. In general, the bond of a molecule experiences various types of vibrations and rotations; however, only vibrational motion (and some contributions from rotations) is considered in IR spectroscopy as the resonant frequency that can excite a vibrational motion of a molecules falls in the IR region. At this resonant frequency, the atoms of a molecules vibrate at a greater magnitude, resulting in the stretching of bond lengths and a change in the bond angles. The vibrational degrees of freedom may be described using the models of molecular oscillators. The most approximate description of molecular vibration is fulfilled



(a) The EM wave (by courtesy of Hamamatsu Photonics, copyright 2020; used with permission)



(b) The EM spectrum (by courtesy of Encyclopædia Britannica, Inc., copyright 2020; used with permission)

Figure 2.3: Electromagnetic wave and spectrum

by a classical harmonic oscillator. The vibrating bond of a molecule is analogous to a spring-mass exemplary system – therefore, IR spectrum reveals information about the properties of the system where wavenumber is related to the force constant (k_f) and the effective mass of (μ) the molecules, by solving the Schrödinger equation for the harmonic oscillator using Hooke's law that results in Eq. 2.1, whereby $\widetilde{\nu} = \frac{1}{2\pi c}\sqrt{\frac{k_f}{\mu}}$ and for diatomic molecules, $\mu = \frac{m_1 m_2}{m_1 + m_2}$.

In harmonic approximation, only equidistant vibrational levels are permissible. The transition of energy occurs between adjacent levels i.e. $\Delta n = \pm 1$. The most likely vibrational transition at room temperature is the fundamental transition (Chalmers & Griffiths 2002). According to Boltzmann distribution, the majority of the molecules populate the vibrational ground level at moderate temperature as there is not sufficient thermal energy to excite the molecules. In the case of fundamental transition, approximation with harmonic oscillation works well; however, the true nature of molecular vibrations is more complicated than that. At elevated temperature, the vibrational transitions from excited levels can take place, for instance, $\Delta E_{1\to 2}$ – the IR bands observed in this kind of situation is termed the 'hot bands' which occurs at a lower wavenumber compared to fundamental transition. In actual experiments, the transitions between non-adjacent levels, $\Delta E_{0\to 2}$, are also observed. The anharmonicity of the vibrational model needs to be considered here for a correct representation of these observations.

Take an example of a diatomic molecule. The vibrational potential is as the repulsive force increases when the bond shortens and vice versa when the bond elongates as a molecule reaches the bond dissociation boundary, the binding force decreases. The vibrational potential of an anharmonic oscillator is shown in and may be approximated with Morse potential curve such that $E(q) = D_e(1 - e^{-a(r-r_e)})^2$ and $a = \sqrt{\frac{k_e}{2D_e}}$, where k_e is the force constant at the minimum of potential well and D_e is the dissociation energy, whereas the energy levels of an anharmonic oscillator is governed by the following equation

$$E_n = h\nu_{osc}(n_L + \frac{1}{2}) - xh\nu_{osc}(n_L + \frac{1}{2})^2 \tag{2.2}$$

The anharmonic constant x describes the decrease in spacing towards higher energy levels of an anharmonic oscillator. In Eq. 2.2, the transitions between non-adjacent levels $\Delta n_L = \pm 1, \pm 2, \pm 3...$,etc as described before are allowed. It describes the first overtone ($\Delta E_{0\to 2}$) and second overtone ($\Delta E_{0\to 3}$) (Chalmers & Griffiths, 2006), which are sometimes observed in an IR spectrum.

Not all vibrations are IR active. The gross selection rule for a change in vibrational state brought about by absorption or emission of radiation is that the electric

dipole moment of the molecule must change when the atoms are displaced relative to one another (Atkins et al. 2019) [4]. Homonuclear diatomic molecules such as N_2 and O_2 are IR inactive, since their vibrational motions fail to induce any change in the dipole moment; but heteronuclear diatoms (i.e. CN) do have dipole moments that depend on internuclear distance, so they exhibit vibrational spectra. This explains why N_2 and O_2 (present abundantly at 78 % and 21 % by volume in the atmosphere, respectively) do not constitute as the greenhouse gases and why IR spectrometer can operate at atmospheric condition. Polyatomic molecules undergo more complex vibrations that can be summed or resolved into normal modes of vibration. The normal modes of vibration can be anti-symmetric, symmetric, wagging, twisting, scissoring, and rocking for polyatomic molecules.

The vibrational modes exhibited by a polyatomic molecule can be deduced from the degree of freedom, the number of variables required to describe the motion of a particle completely. For an atom moving in 3-dimensional (3D) space, three coordinates are adequate, so its degree of freedom is three. Its motion is purely translational. If we have a molecule made of N atoms (or ions), the degree of freedom becomes $3N$, because each atom has 3 degrees of freedom. Furthermore, since these atoms are bonded together, all motions are not translational; some become rotational, some others vibration. For non-linear molecules, all rotational motions can be described in terms of rotations around these 3 axes, the rotational degree of freedom is 3 and the remaining $3N - 6$ degrees of freedom (where N is the number of atoms) constitute vibrational motion. For a linear molecule however, rotation around its own axis is not counted because the molecule remains unchanged, in this case, there are only 2 rotational degrees of freedom for any linear molecule leaving $3N - 5$ degrees of freedom for vibration. A good example of IR active molecule to look at here is CO_2, which constitutes 0.04 % (V/V) of the entire atmosphere. CO_2 has 4 vibrational motions [5], namely symmetric stretching ($\sim$1338 cm^{-1}), anti-symmetric stretch ($\sim$2349 cm^{-1}), and out-of-plane and in-plane bending ($\sim$667 cm^{-1}). Out of all the vibrations, symmetric stretching is the only one that does not show up on an IR spectrum as there is no change in the dipole moment. On the other hand, the two bending modes are known to be degenerate since they occur at the same wavenumber, but this degeneracy disappears when CO_2 interact with other molecules, i.e. the carbonyl band in a polyketone sample (Ewing et al. 2015). The presence of CO_2 and water vapour bands in the IR spectrum poses a great challenge for interpretation of the data. This is addressed in Section 2.5.

Band broadening is commonly observed at the high wavenumber end, assigned to the hydrogen bonded vibrations (i.e. NH and OH). The wavenumbers position of NH and OH bands largely depend on the force constant of the bond which is affected when H atom is involved in H-bond interactions with similar molecules or water molecules (such as mois-

[4] The molecule need not have a permanent dipole in order to be infrared active.

[5] There are 3 atoms ($N = 3$) in a CO_2 molecule, and since it is a linear molecule, 4 vibrational modes can be found from $3(3) - 5$.

ture in the environment). These interactions also depend on proximity of other molecules and functional groups which are not static (in crystals, for example), therefore, many molecules with NH and OH groups have a number of slightly different H-bond interactions with basic functional groups (such as carbonyls, N atoms, etc.) or the oxygen atom in water. As a result, there is a broad distribution of the strengths of these interactions. Furthermore, NH and OH groups may also form combination bands with intermolecular stretching and bending mode of H-Bonding, which have corresponding wavenumbers at around 150–100 cm^{-1} and 60–20 cm^{-1}, respectively. $\nu_{OH} + \nu_{intermolecular\ stretching\ or\ bending\ vibrations}$ will result in many combinations bands which also contribute to the broadening of OH and NH bands (Coulson & Robertson 1974).

2.4.1 Interpretation of IR spectrum

During a spectroscopic measurement, a spectrometer is used to measure the decrease of the radiant power (or a decrease in the number of photons) that reached the detector due to the absorption by molecules at a given wavelength. In an ideal model, absorption at a wavelength would result in a sharp line of a decreased radiant power exactly at that wavelength. However, in reality, physical effects descibed by the Heisenberg's uncertainty principle or Doppler shift as well as instrumental factors will lead to line broadening and consequently absorption bands with different profiles that are approximated by Gaussian or Lorentzian spectral profile functions and bandwidth are observed (Grabska et al. 2017). The measured spectrum can further become increased through overlapping of the absorption bands. An IR spectrum that is useful for interpretation is a plot of IR light intensity versus the light property. In conventional practice, this is often the absorbance versus the wavenumber of IR light. Absorbance (A) can be calculated from intensity (I) plot obtained after FFT such that

$$A = log\frac{I_o}{I} \tag{2.3}$$

The value is relative to the reference radiant power emitted from the radiation source of the spectrometer. The inter-conversion between absorbance and transmittance ($\%T$) is relatively straightforward following

$$\%T = 100\frac{I}{I_o} \tag{2.4}$$

Although both conventions are widely adopted, absorbance spectrum is more meaningful and well suited for quantitative analysis as it has a linear relationship on the concentration of the sample measured, related by the Beer-Lambert's Law, which states

that

$$A = \varepsilon cl \tag{2.5}$$

where ε is the molar absorptivity of the sample, c is its concentration, and l is the effective path length of measurement. Based on the Beer-Lambert's Law, quantitative analysis on any sample can be carried out easily. It can also be seen that absorbance is a factor of path length, therefore, the sample thickness needs to be controlled to avoid truncated peaks in the spectrum. This is particularly true for measurement in transmission mode; in attenuated total reflection (ATR), sample thickness is not an issue which is explained in Section 3.1.5. In addition, qualitative analysis on the IR spectrum acquired is also possible. The peak positions of the spectral bands in an IR spectrum correlate with the molecular structure. Extensive spectrum libraries from a great number of infrared spectra measured over the years are now available, and the spectral bands of known molecules derived from these spectra can be used to identify the molecules in an unknown sample.

2.4.2 IR on biological applications

IR spectroscopy has long been applied to study complicated molecules. By 1950, this spectroscopic technique has been used to study complex proteins (Elliott & Ambrose 1950). However, the capability of IR spectroscopy is far beyond just proteins. DNAs and lipids have also been successfully studied since (Theophanides 2012). This makes IR spectroscopy a powerful analytical technique in bioscience.

In the development of diagnostic strategies for early cancer detection, the biochemical changes of a cancer cell or tissue have preceded the morphological changes. This led to the discovery of the use of IR spectroscopy to study disease and early cancer. 'Biospectroscopy' is the general term used to describe the application of different forms of spectroscopy to biological and biomedical studies (Trevisan et al. 2012). In recent years, the rapid development of IR spectroscopy has allowed it to be envisaged as an objective and robust tool to be used in cancer screening and diagnosis. In contrast to conventional histopathological method of staining with dyes, IR spectroscopy is relatively fast and non-destructive to the samples (Bird et al. 2008). Chemical knowledge of the composition of a sample of interest can rapidly be obtained, which contains useful information for the classification of the cancer. In biomedical research with IR spectroscopy, the type of samples used are determined, to a large extent, by the instrumentation and methodology employed to carry out the study (Bel'skaya 2019). They could be biological fluid, tissues or cells; which are measured either in-vivo or ex-vivo.

The applications of spectroscopy to bodily fluids present a less invasive approach

than performing biopsy on patient. An example of bodily fluids extensively study is the human blood serum in the detection of cancer (Bonnier et al. 2014). In this context, human blood serum is rich in chemical information that can be obtained to analyse the health condition of a person as it contains no less than 20,000 proteins of a huge variety with the total concentration of protein estimated at 1 mM. From as little as $1\,\mathrm{ng\,L^{-1}}$ of troponin (or other lower abundance of serum proteins including human growth hormone and prostate specific antigen) to as high as $50\,\mathrm{g\,L^{-1}}$ of serum albumin, the complex composition of the proteins is a promising candidate for cancer diagnostic purposes (Adkins et al. 2002, Pieper et al. 2003). In particular, the low molecular weight fraction of the serum, known as 'peptidome', which is bound to albumin is identified as a potential cancer biomarker that helps to interpret the molecular events taking place in the presence of malignant cells or the TME (Liotta & Petricoin 2006). In the context of cancer detection, it is recognised that for breast cancer, carcinoembryonic antigens, i.e. CA 15-3, HSP90A, and PAI-1 can be considered as serum biomarkers (Bast et al. 2001, Kazarian et al. 2017); for high grade gliomas (or brain tumour) YKL-40 and MMP-9 have been found to be potential serum biomarkers (Hands et al. 2016, Hormigo et al. 2006); for prostate cancer, the prostate specific antigen was found in higher levels in the serum of patients (Catalona et al. 1991, Labrie et al. 1992); for colorectal cancer, carcinoembryonic antigen in blood serum has been reported to be a potential diagnostic biomarker (Locker et al. 2006) in research studies in recent decades. The low concentration the analytes in the serum often leads to spectra of poor signal-to-noise (S/N) ratio in IR spectroscopy, presenting a huge challenge to the ability to identify these components. This is made even worse if aqueous solution is used as the strong absorption of water will mask the spectral bands of analytes of interest (Fabian et al. 2005). The scissoring vibrational mode of the O-H functional group of water exhibits feature at $\sim$1638 $\mathrm{cm^{-1}}$ which superimposes with the amide I band – characteristic of protein rich samples – and also partially overlapping with the amide II band region between 1580 – 1490 $\mathrm{cm^{-1}}$ (Bonnier et al. 2014).

Although the spectral contribution of water could be overcome by the use of ATR-FTIR spectroscopic approach (more so in the multi-reflection set up) due to its shallow penetration depth of the evanescent wave into the samples, water associated peaks could still make it difficult to analyse the convoluted spectral bands obtained in the fingerprint region, and to a more serious extent, this could lead to misinterpretation of the spectra. In addition, it is not uncommon that components of low concentration in the serum may not be detected by ATR-FTIR spectroscopy until it is completely dry in a phenomenon called the 'Vroman effect' – the redistribution of protein in the drying process based on their abundance and the biochemical and electrochemical affinity to the deposition surface (Hirsh et al. 2013). Therefore, to overcome these issues, experimental studies on human serum have predominantly been carried out on air dried droplets deposited on IR transparent substrates such as CaF_2 (Cameron et al. 2018, Ghimire et al. 2017, Butler et al. 2019). Other than blood serum, to date, tears (Nagase et al. 2005, Travo et al.

2014), urine (Paraskevaidi et al. 2018, Yu et al. 2017), saliva (Takamura et al. 2018), and cerebrospinal fluid have also been employed to study with vibrational spectroscopy (Horosh et al. 2016).

Having said that, spectroscopic measurement of air dried biofluid samples present another kind of concern. Coffee-ring formation, cracking and gelation patterns have all been observed in biofluid drops and it has been shown by optical and spectroscopic assessment that this deposition is heterogeneous. Non-uniform spectra have been observed for different patterns due to the inhomogeneous distribution of components within the sample (Cameron et al. 2018). Distortions to IR spectra were found to arise due to this chemical heterogeneity across the drop surface. In the well-known phenomenon known as the 'coffee-ring effect', it was observed that absorption at amide I and II regions around $1647\,\mathrm{cm}^{-1}$ and $1542\,\mathrm{cm}^{-1}$ respectively; as well as lipid band at $1442\,\mathrm{cm}^{-1}$, and nucleic acid at $1078\,\mathrm{cm}^{-1}$ differ across the droplets in an experiment carried out with IR spectroscopy in transmission mode. The absorption values of the bands are higher at the periphery of the ring (Hughes et al. 2014). In the same transmission study, spectral distortions affecting the absorbance ratio and baseline of the amide I and II bands due to light scattering at the cracks which appear in the center of the ring were seen. At the same time, peak shifting in spectral features between 1500 – $1000\,\mathrm{cm}^{-1}$ was also observed at different points of measurement across the 'coffee-ring' (Bonnier et al. 2014). As discussed, the inhomogeneity of the deposits of air-dried biofluids is the main limiting aspect for potential clinical practice, hence consistent and improved protocols are needed to handle biofluids. One of such proposed methodology is sample concentration by centrifugal filtering that helps to improve the quality of the spectra (Bonnier et al. 2014).

The advent of IR microscope with high resolution limit allows the spectral information of cells and tissues to be studied (Naumann 2013). A number of researches involving cancer cell lines and tissues such as ovarian cancer (Li et al. 2018), colon adenocarcinoma (Gao et al. 2015), prostate cancer (Baker et al. 2010), skin cancer (Mostaco-Guidolin et al. 2009), and lung cancer (Lee et al. 2009, Liberty et al. 2015) have been published since. Furthermore, spectral features within a single eukaryotic cell can be successfully determined with this IR microscope, as demonstrated by (Lasch et al. 2002).

2.4.3 IR features of molecular composition of cell

Biological samples contain macromolecules, such as nucleic acids, proteins, lipids and carbohydrates that have characteristic and well-defined IR vibrational modes. These bands can be used as markers for pathologies (Sabbatini et al. 2017). The low wavenumber region, also known as the 'fingerprint region', loosely defines spectral range between 1500 or 1400 – $500\,\mathrm{cm}^{-1}$. This region usually contains a very complicated series of IR absorptions. Within the fingerprint region, it is difficult to recognise individual bonds from this collective of

convoluted bands; however, the fingerprint region is important in the analysis of compound as no two compounds will exhibit the same 'fingerprint' (Smith 2011). The identity of the samples can thus be determined by analysing the fingerprint region. Commonly utilised regions for detection of cancer biomarkers are between 3000 – 2800 cm^{-1} and 1800 – 900 cm^{-1} (Baker et al. 2014) (Fig. 2.4) Some of the most important spectral bands and their corresponding assignment to the biomolecules can be found in literature (Movasaghi et al. 2008) and are listed in Table 2.2.

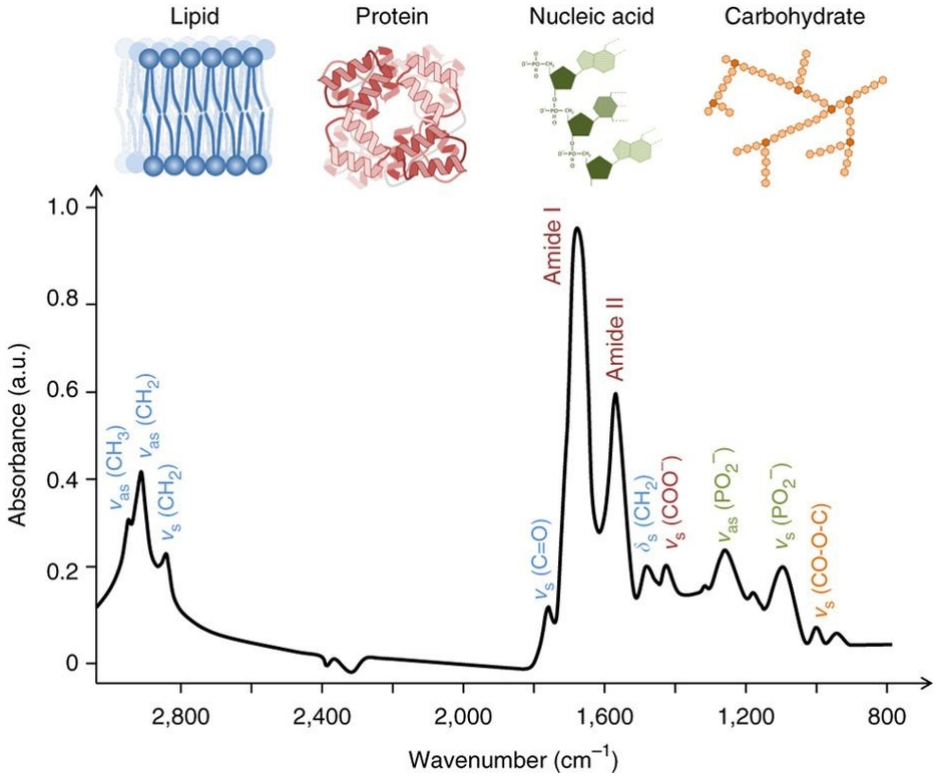


Figure 2.4: An example of IR spectrum obtained with biological sample, reproduced from (Baker et al. 2018) with the permission from Springer Nature

2.4.4 Expected metabolic changes and IR signatures in cancer cells

Common metabolisms expected in cancer cells include aerobic glycolysis, reduced oxidative phosphorylation, and the increased production of biosynthetic intermediates required for cell growth and proliferation (DeBerardinis & Chandel 2016). Most cancer cells demonstrate the 'Warburg effect'. For example, colon cancer cells often exhibit exacerbated glucose (carbohydrate) uptake and glycolysis utilization leading to increased lactate production (Brown et al. 2018). The spectral absorbance in the region between 1200 – 900 cm^{-1} is attributed mainly to glucose (carbohydrates) and glycoprotein (Sahu et al. 2017). In addition, colon and prostate cancer proliferation are also encouraged by lipid intake

Table 2.2: Assignment of the spectral bands of a biological sample

Spectral band (cm^{-1})	Band assignment	Ref.
966 – 964	ν_s C–C and C–O in deoxyribose of DNA backbone	(Gioacchini et al. 2014, Sahu et al. 2004)
998 – 996	ν_s C–O ribose and C–C in uracil ring of RNA	(Sahu et al. 2004)
1020	ν_s in deoxyribose of DNA	(Prabhakar et al. 2012)
1030	Glycogen and collagen	(Movasaghi et al. 2008)
1050	ν_s C–OH (coupled with δ C–O) of polysaccharides or carbohydrates	(Tseng 1997)
1085 – 1080	ν_s PO^{2-} in nucleic acids	(Movasaghi et al. 2008)
1121 – 1120	ν_s ribose C–O in RNA	(Prabhakar et al. 2012)
1241	ν_s PO^{2-} in nucleic acids	(Giorgini et al. 2017)
1255 – 1252	Amide III (δ N-H bending, v C–N stretching, ν C–C stretching)	(Chiriboga et al. 1998)
1400	δ CH_3 in amino acid and fatty acid side chains	(Fung et al. 1996)
1468 – 1460	δ CH_2 in amino acid and fatty acid side chains	(Fabian et al. 1995)
1553	Amide II (ν_s C-N and δ C-N weakly coupled to the Vs C=O)	(Rath et al. 1991)
1650 – 1648	ν_s C=O (and weak δ C-N and δ N-H) in Amide I	(Dovbeshko 2000)
1745 – 1744	ν_s C=O in ester of triglycerides	(Wu et al. 2001)
2859	ν_s CH_3 (predominantly fatty acid chains or lipid, with little contributions from proteins, carbohydrates, and nucleic acids)	(Mignolet et al. 2016)
2926	ν_{as} CH_2 (predominantly lipid)	(Mignolet et al. 2016)
2970 – 2967	ν_{as} CH_2 (predominantly lipid)	(Mignolet et al. 2016)

(Long et al. 2018). The other hallmark for many cancers is the increased cholesterol biosynthesis and uptake (because of the increased absorbance of the bands between 3000 – 2800 cm^{-1}) via the mevalonate pathway (Huang et al. 2020). As for nucleotides, cancer cells have been shown to stimulate de novo nucleotide synthesis to support nucleic acid and protein synthesis along with energy preservation, signaling activity, glycosylation mechanisms, and cytoskeletal function (Villa et al. 2019). An increase in the spectral absorbance of nucleic acids is thus expected in most cancer cells.

Case study 1: Prostate cancer (CaP)

The study of CaP with FTIR spectroscopy can be dated back 1990's, whereby the discrimination between normal prostate tissues, benign prostatic hyperplasia (BPH), and adenocarcinoma tissues is achieved by clustering of the spectral data measured with PCA, based on two regions of the spectra, i.e. 1174 – 1000 cm^{-1} (assigned to strong stretching vibrations of the PO^{-} and C–O groups of the phosphodiester- deoxyribose structure) and 1499 – 1310 cm^{-1} (assigned to weak NH vibrations and CH in-plane deformations of the nucleic acids) (Malins et al. 1997). It is thus concluded that progression of normal prostate to cancer involves structural alterations in DNA that are distinctly different. It was later showed that histologically normal tissues adjacent to tumours also displays abnormalities in the nucleic acid bands, providing clues for the onset of carcinogenesis in prostate (Malins et al. 2005). IR spectroscopy has seen substantial progress in the last decade, following this, its applications on biological samples have also been widely employed (Diem et al. 2004) and this includes the study on prostate cancer. FTIR micro-spectroscopy was applied to study the FFPE tissues samples of benign and cancerous prostate, as well as the prostate cancer cell lines. It was found that the ratio of peak areas of 1030 and 1080 cm^{-1} (corresponding to glycogen and phosphate vibrations respectively) could be utilised for the differentiation of benign from malignant cells (Gazi et al. 2003). Further studies on prostate cancer were also carried out since (Lasch, Mahadevanansen & Diem 2004, Baker et al. 2009, 2008, 2014, Bassan et al. 2009, 2012, 2013, Diem et al. 2004, Gazi et al. 2006, 2004, Harvey et al. 2007). Some of the potential biomarkers identified to distinguish between normal and malignant prostate are amide I, II, and III; as well as protein regions (1400 – 1585 cm^{-1}), DNA/RNA (ν_s PO_2^-) (1080 cm^{-1}), and DNA (ν_{as} PO_2^-) (1230 cm^{-1}) regions (Siqueira et al. 2018).

Case study 2: Colon cancer

IR spectroscopic studies on colon tissues have shown important spectral biomarkers for colon cancer, which differentiate the malignant tissue sections from the healthy ones. In transmission FTIR experimental studies involving non-deparaffinised FFPE human

colon tissues, the spectral bands correspond to the vibrations of the functional groups of deoxyribonucleic rings (1085 cm^{-1}, 1249 cm^{-1}), proteins (1550 cm^{-1}, 1648 cm^{-1}), and water (3250 cm^{-1}) are not observed in tumour compared to healthy tissue. An additional oscillation at 1385 cm^{-1} observed in tumour, which is not present in healthy tissue, was reported to be an important cancer biomarker. This additional spectral feature is likely to arise due to the mutations in DNA during cancer genesis (Depciuch et al. 2017, Kaznowska et al. 2017). Similarly, changes in amide I band and nucleic acid-related vibrations at 1090 cm^{-1} and 1235 cm^{-1} are detected on de-paraffinized cancerous colon tissues (Kallenbach-Thieltges et al. 2013). The weaker absorption bands of cancerous colon over a broad region of infrared, in particular the absorption peaks at 1240 and 1076 cm^{-1} assigned to the anti-symmetric and symmetric stretching of PO_2^- in nucleic acids respectively are, likewise, observed by (Mahadevan-Jansen et al. 2000). This reported decrease in phosphate concentration relative to normal colonic tissue is entirely different from other types of cancer such as breast cancer or skin cancer which typically show a rise in phosphate absorption. The cause of this behaviour is not well understood, although hypotheses were put forward to explain this observation – the rapid cell metabolism in the cancerous colon causes a boost of phosphate consumption by the cells but the low availability of intracellular glucose in adenocarcinoma cells leads to glucose starvation which would consequently cause the degradation of DNA Topoisomerase II a by proteasome, which in turn reduces the DNA synthesis (Kim et al. 1999). Moreover, a shift in the DNA peaks, for example the shifting of 970 cm^{-1} peak in normal to 966 cm^{-1} in cancer, is also an indicator for changes in DNA (Argov et al. 2004). They also suggested the potential use of FTIR spectroscopy for routine monitoring of colon in patients with a continuous history of bowel diseases by predicting the progression of Crohn's disease and inflammatory bowel disorder to cancer, based on parameters such as the ratio of integrated absorbance of tyrosine/phosphate (1512 cm^{-1}/1020 cm^{-1}) and RNA/DNA (996 cm^{-1}/966 cm^{-1}). For lipid bands, significant difference between malignant colon tissues and colitis was recorded in another published work on frozen tissues. C-H stretching bands (2966 cm^{-1}, 2927 cm^{-1}, 2858 cm^{-1}) decrease and even disappear in the spectra of colon cancer; mainly because the development of carcinoma requires increased lipid consumption for the generation of more energy to sustain its proliferation (Li et al. 2012, Rigas et al. 1990). In the same study, the relative absorbance of amide II (1550 cm^{-1}) to amide I (1643 cm^{-1}) band of protein is found to decrease in the spectra of malignant colon tissues. Further deconvolution of the amide I band also suggested that the cellular proteins are to a large extent α-helices with considerable segments of β-sheet; yet the relative amount of β–sheet with respect to that of the α-helical segment is greater in cancer than in normal tissues (Rigas & Wong 1992). When the human colon cell lines are investigated, some of the important spectroscopic features observed in malignant colon tissues are also shared by the colon cancer cell lines, namely: (a) increased hydrogen bonding of the phosphodiester groups of nucleic acids; (b) decreased hydrogen bonding of the C-OH groups of carbohydrates and proteins; (c) a prominent band at 972 cm^{-1}; and (d) a shift of the band normally appearing at 1082 cm^{-1}

to 1086 cm^{-1} (Rigas & Wong 1992). In the high wavenumber spectral region between 3100 — 2800 cm^{-1}, an increase ratio of spectral absorbance of CH_3/CH_2, reflecting the formation of shorter fatty acid chains in colon cancer, was found in a research on de-paraffinized colon tissues (Sahu et al. 2005). Short-chain fatty acids can promote cell migrations in colonic crypts that can lead to abnormal crypt proliferation, a precursor of colon cancer (van der Beek et al. 2017).

2.5 Challenges in IR spectroscopy on biological tissues

Despite the increasing popularity of the field, translation of research studies of cancer with infrared spectroscopy to actual clinical environment have been difficult; there are a few challenges with regards to preparation of samples, IR instrumentation and data processing which need to be addressed before bio-spectroscopy can become a routine process in the clinical settings (Baker et al. 2014).

2.5.1 Tissue de-paraffinization

The preparation of tissues for measurement with IR spectroscopy can have a serious impact on spectral interpretation for their biochemical significance and for imaging studies, the spatial distribution of biomolecules could be affected (Lyng et al. 2010).

For FFPE tissue, the presence of paraffin poses an important issue. Paraffin is known to have significant peaks at 2920 cm^{-1}, 2846 cm^{-1}, 1462 cm^{-1}, 1373 cm^{-1}, and ∼954 cm^{-1}, which may mask solvent-resistant methylene components of native tissue (Sahu et al. 2005, Baker et al. 2014). In addition, paraffin also alters the frequency of the C–O vibration, which could explain the reduction of the band at 1938 cm^{-1}. Formalin fixation of the tissue could also lead to the peak shifting of amide I and II spectral bands by ∼10 cm^{-1} (O Faolain et al. 2005). An investigation on the protein secondary structure after ethanol fixation of rat femur revealed changes in the amide I and II bands near 1650 and 1550 cm^{-1}1, respectively, a result of alteration of the protein conformation of the tissue (Pleshko et al. 1992). Although not all the effects of formalin fixation are well understood yet, the alteration of amide by formalin could be explained as follows: aldehydes in formalin form cross-links (or methylene bridge) between proteins creating a gel, thus retaining the cellular constituents in the tissues. Soluble proteins become fixed to structural proteins. The majority of cross-links are formed between the nitrogen atom of lysine and the nitrogen atom of a peptide linkage. Secondary amide is altered to a tertiary amide from this cross-link (O Faolain et al. 2005).

Currently, there is a lack of consensus regarding the standard protocol for de-

paraffinization of FFPE tissues (Lyng et al. 2010). There are a lot of preparation protocols reported to dewax samples. For example, a study on colon cancer used xylene (or xylol), C_8H_{10}, and alcohol for de-paraffinization of the samples prior to measurement (Sahu et al. 2005). It was shown that after two washes with xylene (10-min with three changes for each wash with mild shaking to hasten the paraffin removal process), there is complete paraffin removal. This is verified by examining the disappearance of spectral bands from paraffin at wavenumbers from 1483 to 1426 cm^{-1}. The samples were then left in 70 % alcohol overnight for ∼12 h and changes in the region 1185 to 900 cm^{-1} were observed; which indicates that alcohol could have removed the xylene and some sugars in the tissues. The processing of tissues could, of course, result in a differential extraction of small, easily removable lipids between normal and cancer tissues in the short time that they are exposed to xylene and alcohol – minor changes , i.e. reduction in absorbance of the spectral bands, were observed in the region 3100 to 2800 cm^{-1} during de-paraffinization; however, this could also be beneficial and a minor factor contributing to the better diagnostics observed in colon tissues compared to cell lines, as demonstrated in the study (Sahu et al. 2005). Nonetheless, the use of spectral peak of paraffin at 1462 cm^{-1} alone as an assessment for the complete paraffin removal in FTIR spectroscopy was questioned by (Faolain et al. 2005) whose work showed that a number of strong signals from vibrations of C–C and CH_2 functional groups of paraffin were still observable using Raman spectroscopy at 1062 cm^{-1}, 1296 cm^{-1}, and 1441 cm^{-1} when the parenchymal tissue from placenta was dewaxed with xylene followed by ethanol. They found that Raman spectroscopy produced more detailed spectra and is intrinsically more sensitive to some paraffin bands than FTIR. In fact, IR spectral peak at 1462 cm^{-1} is not unique to paraffin as it is a bending (scissoring) vibration of methylene (CH_2). Biological molecules themselves will have abundant naturally-occurring methylene bending vibrations of similar strength at this wavenumber (Hughes et al. 2014). Other less toxic xylene substitutes such as Citroclear has also been used to remove the paraffin (Gazi et al., 2006). Among all dewaxing agents tested to date, hexane, C_6H_{14}, was shown to be the most effective dewaxing agent, resulting in almost complete removal of wax (Faolain et al. 2005, Lyng et al. 2010). It was also shown that when using either xylene or hexane as the solvent, paraffin removal can be achieved in a period of ten minutes (Hughes et al. 2014).

The effect of the dewaxing process on human tissue is still a matter of debate and commonly discussed within the biomedical field. It has been suggested that methylene chains of free, unbound tissue lipids are leached from the tissue assumingly in the first tissue processing stages prior to paraffin embedding (Hughes et al. 2014). The situation is complicated by the overlap of paraffin bands with the lipid bands. Although de-paraffinized FFPE tissue sections may display a change in spectral band observed as mentioned before, all of the bands in fresh tissue are present and many are still diagnostically useful (Faolain et al. 2005). Despite the leaching of free lipids, not all information is lost. The solvent-resistant lipids remain present in de-paraffinized FFPE tissues due to them being locked

into the protein–lipid complex matrices, predominantly in the membranes. It is these lipids signals that are subsequently detectable by spectral pathology and have important diagnostic values (Hughes et al. 2014). A recent study comparing the spectra of bone tissue embedded in paraffin and de-paraffinized tissue showed that paraffin has no statistically significant impact on FTIR results (Chaber et al. 2017). Nonetheless, much doubts have been raised as to whether to deparaffinise samples or not. Omitting paraffin de-waxing could have helped eliminate the disturbances caused by variation in de-paraffinization protocols, as well as the tissue damage which could potentially arise due to inappropriate specimen processing and paraffin-embedding technique (Werner et al. 2000).

Routine de-waxing with hexane or xylene is time consuming. Alternate approaches to de-paraffinize tissues were introduced. One of these is modelling the contributions of paraffin and mathematical de-paraffinization of the samples based on the model; an approach which was shown to be highly effective in several studies on colon and skin tissues (de Lima et al. 2017, Ly et al. 2008, Nallala et al. 2012, Wolthuis et al. 2008). Apart from the high efficiency of digital de-waxing, this method also confers index matching thereby potentially reducing scattering artefacts (Nallala et al. 2015). The common approach for digital de-waxing is the Extended Multiplicative Signal Correction (EMSC) algorithm, which was first developed in 1980s for near-IR spectroscopy (Geladi et al. 2016). The same EMSC algorithm is descibed for Mie scattering correction. This algorithm is adapted and modified to neutralize the spectral contributions of paraffin using models consisting of the mean paraffin spectrum and its first ten principal components accounting for the paraffin variability and the target spectrum. Using this approach the variability arising from the paraffin is then neutralized across all the pixels in the dataset and only the chemical variability from cells and tissues is retained (Nallala et al. 2015).

2.5.2 Patient-to-patient variation

Another challenge with disease diagnosis is the interpatient variability, which arise from patients of different genders, ages, ethnic background, lifestyles, medical histories, and physical conditions including hormonal status. These interpatient variabilities are all thought to manifest themselves as disparities in the chemical content of the biological samples sourced from the patients. As a result, it has to be taken into account when designing a diagnostic framework or validating a model for disease classification based on the IR spectra of such samples. By using a large patient cohorts with varying stages and different grades of disease would minimise the interpatient variability (Kallenbach-Thieltges et al. 2013). Appropriate selection of patients and control subjects is of utmost importance to reduce the risk of false positives. Unmatched comparison may lead to biased results and differences between the spectra detected by these diagnostic models could be caused by these confounding factors instead of the disease of interest. The

downside of sourcing from a large patient database is the increased time needed to collect huge amount of IR spectra. This is shown to be made possible by performing feature extraction to identify the salient spectral information, thus reducing the patient variance by targeting the most discriminatory regions during spectral collection to reduce collection times (Hands et al. 2016). That said, it was based on the single spectra of blood serum which are easier to sourced and measured. The story is different for tissue samples, especially with imaging, whereby a significant amount of time is needed to acquire the data of a small sample area. However, the acquisition of a large patient cohort might not be necessary in the study of colon tissues. It has been suggested in previous research studies that the patient-to-patient variations in spectral patterns are actually smaller than those arise from the different stages of disease, or even due to different tissue types. From their findings, it can be inferred that training the spectra from a smaller dataset to achieve reliable results in the study of colon cancer with machine learing is highly possible; but transferring the diagnostic model to another model for a different type of cancer is likely to result in failure of the model to cope (Lasch & Naumann 1997).

2.5.3 Resonant Mie scattering

When electromagnetic radiation encounters an object, in general, it is either scattered, absorbed, or transmitted depending on the morphology (size and shape etc.) of the object, its composition, and the wavelength of the light. In IR transmission spectroscopy study, the amount of light that is transmitted through a sample, representing the loss of light due to absorption by the sample, is calculated in relative to the background spectrum with nil absorbance. However, this only holds true for ideal system with a flat surface. In real situations, a substantial part of the light is lost due to scattering. Samples of a non-flat surface, such as most biomedical samples (single cells and tissue sections), tends to cause light scattering. As a result, the interpretation of the data becomes difficult, since a major part of the incoming light intensity does not reach the detector because of the deviation in light path from scattering. The amount of intensity scattered is strongly dependent on the wavelength of the electromagnetic radiation, the size of the scatterer, and its refractive index. In the field of biomedical optics, the phenomenon called Rayleigh scattering and Mie scattering are commonly encountered. Despite the common use of the terms 'Rayleigh scattering' to refer to scattering by small particles or mass density fluctuations much smaller than the wavelength of light and 'Mie scattering' to refer to light scattering that takes place when the size of the scatterer (or dielectric sphere) is comparable to or larger than the wavelength of light, they are not the correct definitions – Mie scattering is the generic name for scattering by any spheres of any size; whilst Rayleigh scattering is the Rayleigh limit of Mie scattering (Jacques 2013). This scattering process was first described theoretically by Gustav Mie in his famous paper on the subject of light scattering by spherical particles hundred years ago (Mie 1908).

In recent years, the effect of scattering has become a subject of interest as this loss of light is evident in the absorbance spectra of single cells. A broad, undulating scattering background in the IR spectrum of an individual cell onto which the absorption features are superimposed was first observed by (Mohlenhoff et al. 2005), paricularly for nucleus which is denser in terms of biochemical content compared to cytoplasm which is relatively spearse. Depending on the positions of the scattering maxima, the observed spectral intensities and the baseline for absorption spectra are distorted. In this effect they described as 'dispersion artefact' or simply annotated as 'anomalous dispersion' at that time, spectra of isolated single cells also exhibit significant distortion of band shapes, most notably a derivative-like distortion on the high wavenumber side of the amide I band (Bassan et al. 2009). The distortion to the spectra of a single cell or the 'dispersion artefact' is now fully understood to be of the same origin as Mie scattering. The term 'dispersive artefact' circumscribes the fact that absorption resonances lead to fluctuations in the real part of the refractive index, which affect the extinction efficiency and thereby the measured absorbance spectrum (Kohler et al. 2008). The excellent qualitative agreement between the PMMA-sphere model and close packed spheres of the same size in a study by (Bassan et al. 2009) demonstrated the principle that *resonant* Mie scattering (RMieS), i.e. scattering when there is simultaneous absorption, is the main cause of the dispersion artefact. The scattering effect can give rise to the shifting of band positions in both positive and negative direction; hence, this resonant Mie scattering is believed to be different from the reflective dispersion artefact that could be present in trasflection spectra (Bassan et al. 2009) [6].

Spectral distortion which arises from strong Mie scattering in the IR microscopy of cells hampers the chemical interpretation of the absorption peaks, thus it is paramount to come up with a solution to be able to achieve separation of the chemical information (the 'true' absorption properties) of the sample from its physical properties (scattering effect) (Blumel et al. 2016). A set of equations that could approximate the Mie extinction, Q [7], known as the anomalous diffraction approximation of van de Hulst, was formulated (Hartmann 1984), as follows:

$$Q = 2 - \frac{4}{\rho} sin\ \rho + \frac{4}{\rho^2}(1 - cos\ \rho) \tag{2.6}$$

and

[6] The reflective dispersion follows the Kramers–Kronig dispersion but does not exhibit 'opposite' behaviour in the spectral band shift.

[7] The sum of 'scattering' and 'absorption' for light-transmitting biological material is termed 'extinction'. Extinction can therefore be used for both phenomena – 'absorption' and 'scattering'. Mie extinction is then equivalent to Mie scattering.

$$\rho = \frac{2\pi d(n-1)}{\lambda} \tag{2.7}$$

where n and d denote the ratio of real refractive indices of particle to surrounding medium and the diameter of the scattering particle, respectively. However, the approximate Mie scattering equation is a good approximation merely for a non-absorbing dielectric sphere uniformly illuminated with a parallel beam, none of which strictly applies in the case of biological systems. Biomedical samples such as single cells are strongly absorbing materials and IR microscopes use Schwarzschild optics to collect light over a large numerical aperture, hence this equation could not fully describe all the scattering phenomenon observed.

In recent years, several investigations have contributed substantially to the development of the spectral correction strategies for the removal of Mie type scattering on cells. Most notably in 2008, an iterative algorithm based on EMSC, in conjunction with PCA based model, was developed to allow correction of the non-resonant Mie scattering of FTIR absorbance spectra measured for individual lung cancer cells (Kohler et al. 2008). In their work, a simple EMSC model can be used to describe the measured absorbance spectrum $A_{app}(\widetilde{\nu})$ as the reference spectrum $Z_{ref}(\widetilde{\nu})$ multiplied with an effect b, in addition to the deviations from the reference spectrum, denoted as c and $\sum_{i=1}^{A_{opt}} g_i p_i(\widetilde{\nu})$, representing the baseline shift and the sum of multiple scattering arise from heterogenous particles, with a residual term $\varepsilon(\widetilde{\nu})$ to capture the any un-modelled part, as shown in Eq. 2.8.

$$A_{app}(\widetilde{\nu}) = bZ_{ref}(\widetilde{\nu}) + c + \sum_{i=1}^{A_{opt}} g_i p_i(\widetilde{\nu}) + \varepsilon(\widetilde{\nu}) \tag{2.8}$$

where the measured absorbance spectrum can be corrected according to

$$A_{corr}(\widetilde{\nu}) = \frac{A_{app}(\widetilde{\nu}) - c - \sum_{i=1}^{A_{opt}} g_i p_i(\widetilde{\nu})}{b} \tag{2.9}$$

The algorithm was then improved to take into account the fluctuating real refractive index of absorbing materials (Bassan, Kohler, Martens, Lee, Byrne, Dumas, Gazi, Brown, Clarke & Gardner 2010) to correct for resonant Mie scattering. The caveat of noise accumulation in the iterative methods was avoided by using a least square curve-fitting step. Despite failing to account for the complex imaginary part of the refractive index, the EMSC model provides sufficiently good extraction of the pure absorbance spectrum and has been used widely in the field of biomedical spectroscopy (Bassan & Gardner 2010) and is coined the term 'RMieS-EMSC correction' algorithm [8] . The main issue of the

[8]The term 'correction' is often used to describe the RMieS-EMSC algorithm, however, it is important

RMies-EMSC lies in its processing time. Since it requires the iterative calculation of KK transform, the running of this algorithm proves to be very time consuming. Further development was carried out to refine the RMieS-EMSC algorithm. The computation time of the algorithm is shown to improve by a factor of 100 when the KK transform is replaced with a fast Fourier transform (FFT) algorithm [9] (Konevskikh et al. 2016). Recently, a different modelling approach was proposed – extracting a single absorption band of samples of a low refractive index, for instance, biological tissues, has been made possible by modelling the band as Fano resonances [10] (Schofield et al. 2019).

The research work governing Mie scattering mentioned so far are mostly performed on isolated cells. The RMieS-EMSC algorithm was first applied by (Bambery et al. 2012) to cervical tissue and the rat brain tissue to check for its efficiency. The findings showed that the true chemical content of the tissue samples are better presented after correction compared to uncorrected image data, more so for the strongly scattering collagenous fibrils within the connective tissue of the cervical section. (Bassan et al. 2009) mentioned in their study that the significance of resonant Mie scattering is dependent on confluency of cells. Close packed cells in tissue are likely to give rise a low degree of scattering compared with isolated cells and suggested that the difference is actually quite dramatic. Mie scattering mostly affect just the protein bands of the edges of the tissues (Schofield et al. 2019).

to note that the spectra are already correct in that they correctly reflect the underlying chemical and physical (which are both related) structures. A better term would be to 'extract' absorption spectra since the absorption measure is obtained from the recorded spectrum in which absorption and scattering are co-mingled as advised by (Bhargava 2012)

[9] The algorithm of RMieS-EMSC, with the outline of the computational strategies and equations can be found in (Bassan & Gardner 2010).

[10] In physics, a Fano resonance is a type of resonant scattering phenomenon that gives rise to an anti-symmetric line-shape, due to the Interference between a background and a resonant scattering process.

Chapter 3

Materials and methods

3.1 Instrumentation

3.1.1 Development of IR spectrometers

The first commercial IR spectrometer was produced by an English company Adam Hilger Ltd in 1913. The instrument was a single beam spectrometer, which means that background spectrum and sample spectrum need to be recorded one after the other. The complication in operating this type of spectrometer was later eliminated by the production of double-beam spectrometer with chopped IR radiation. When the radiation is chopped, no light arrives at the phototransducer (or detector). The dark current can be measured and subtracted from the overall light measurement this way (Chasteen 2009). The dispersive IR spectrometer was first introduced in the 1940s (El-Azazy 2019).

3.1.2 Dispersive IR spectrometer

Dispersive IR instruments are sometimes called grating or scanning spectrometers (Thermo Nicolet Corp 2002). Rudimentary parts of a dispersive spectrometer include a radiation source, a monochromator, a slit, and a detector. As the name implies, dispersive spectrometers optically disperse the incoming radiation into its spectral (or frequency) components. Common dispersive elements include prisms and gratings (Saptari 2003). These instruments work by the principle of monochromatic tracing and hence, the recording time of a whole spectrum is very long. Classical dispersive instruments also feature other considerable limitations such as tedious external calibration of wavenumbers using an external sample and poor S/N ratio from the detector analysing only a very narrow part of spectral region at a time (Sijbers et al. 1996). An example of the dispersive IR instrument is shown in Fig. 4.1 in section 4.2.

3.1.3 FTIR spectroscopy

During late 1940s and 1950s, there were a large number of publications that used dispersive IR instruments. It was in the 1960s when Fourier transform (FT) instruments came to the scene. The first commercial FTIR spectrometer designed for laboratory use was the FTS-14 marketed by Digilab (now under Agilent Corporation) in 1969 (Griffiths & de Haseth 2007). An FTIR also has an IR source and mirrors, but the similarities to a dispersive instrument end there. The presence of interferometers in FTIR instrument leads to markedly pronounced improvement in the efficiency of IR measurement and subsequently, the widespread use of IR light in a number of scientific research areas (El-Azazy 2019). An example of FTIR spectrometer is shown in Fig. 3.1.

Figure 3.1: Bruker Tensor 27 with a macrochamber for macro spectroscopic measurement. This Tensor FTIR spectrometer contains an MCT detector, swappable with a DGTS detector

Working principles of FTIR spectrometer

The 'heart' of the FTIR spectrometer is an optical device known as interferometer. The design of many interferometers nowadays is based on a two-beam interferometer, known as the Michelson interferometer, named after its inventor in 1891. The term 'interferometer' comes from 'interference meter' and as suggested by the name itself, it is a device capable of splitting a beam of radiation into two and measuring the interference pattern of the beams when they recombine. In physics, the principle of superposition of waves states that when two or more propagating waves are incident upon one another, the resultant displacement of the medium at any point in space or time, is simply the vector sum

of the individual waves. If the waves are in phase, i.e. the crest of a wave meets the crest of another wave of the same frequency at the same point (or trough with trough), then a greater resultant amplitude is obtained—this is 'constructive interference'. On the contrary, if the superimposed waves are out of phase, i.e. the crest of one wave meets the trough of another, then a decrease in the amplitude of the resultant wave is observed—this is known as 'destructive interference' (Stark 2020).

The simplest set-up of a Michelson interferometer is shown in Fig. 3.2. It consists of two mirrors, a fixed (a) and a movable (b) one, placed perpendicularly to each other. The collimated IR beam [1] coming from the light source first passes through a beamsplitter (c), where the rays are partially reflected to the movable mirror (ray B), and at the same time partially transmitted to the fixed mirror (ray A). The optical path difference (OPD) between the two rays introduced by the movable mirror results in the difference in the interference patterns of the light when the beams return to the beamsplitter, where they are partially reflected and transmitted once again. The movement of the movable mirror can be controlled for different purposes, for instance, at constant velocity (for continuous scan) or at a continuous velocity greater than $\sim 0.1\,\mathrm{cm\,s^{-1}}$ (for rapid scan) or be held at equally spaced distance for a fixed period (for step scan). Ultimately, it is the variation of intensity of the beam emerging from the interferometer measured as a function of the path difference collected by a detector (d) which yields the spectral information in an FTIR spectrometer. The interference pattern of the two light beams are recorded in a plot of light intensity versus OPD, known as interferogram (which means 'interference writing').

For monochromatic radiation, understanding the interference pattern is relatively simple – the two light beams reaching the detector will be in phase if their OPD[2] (δ) is exactly an integer multiple of their wavelength (λ), i.e.,

$$\delta = n\lambda, \text{ where } n = 0, 1, 2, 3, ..., \tag{3.1}$$

whilst the OPD is in a Michelson interferometer is always twice the mirror displacement [3] (Δ), given by

$$\delta = 2\Delta \tag{3.2}$$

However, in many cases, a broadband IR source, such as Globar, which gives off light of many different wavelengths (6000 – $50\,\mathrm{cm^{-1}}$) are used in an FTIR spectrometer. They pass through the interferometer together. This results in constructive interference

[1] In an interferometer, a collimating mirror is sometimes placed after the IR light source to collect the light and make its rays parallel. This collimating mirror is not shown in the set up in Fig. 3.2.

[2] The optical path difference (OPD) is also called the retardation.

[3] Light traverses the displaced distance twice on the way to and from the moving mirror.

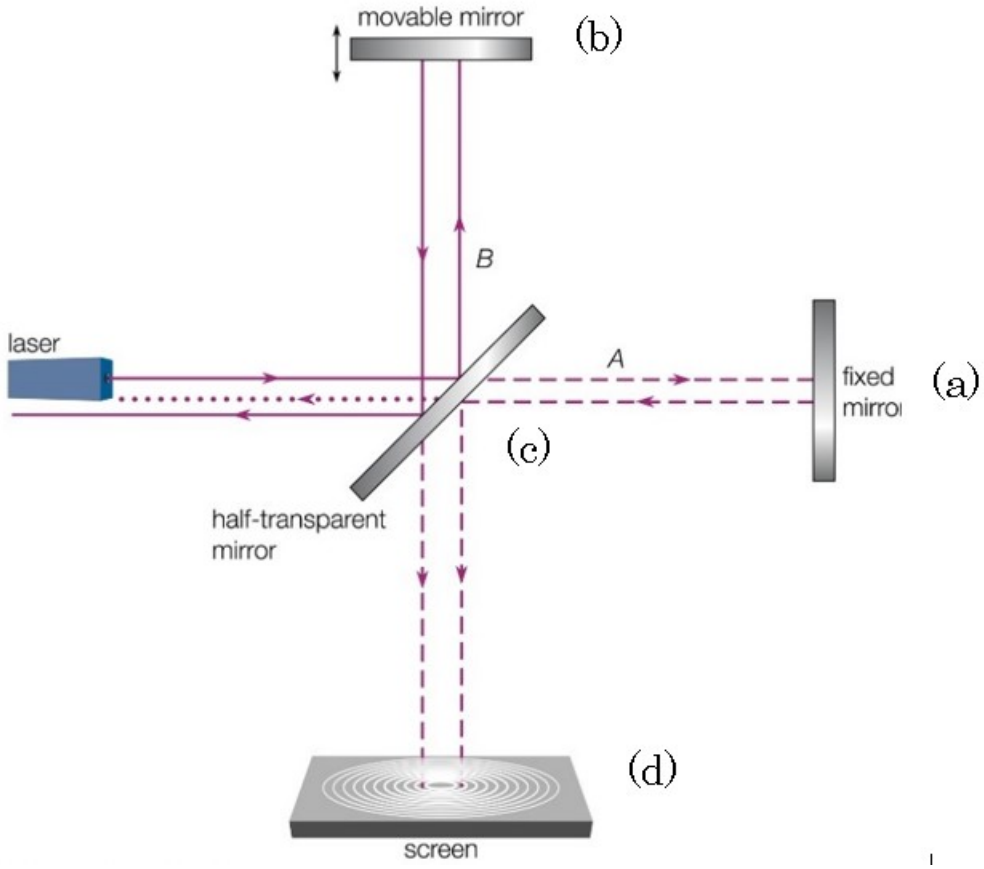


Figure 3.2: An illustration of a Michelson interferometer inside an FTIR spectrometer. By courtesy of Encyclopædia Britannica, Inc., copyright 2020; used with permission (Stark 2020)

takes place only where there is no OPD or zero path difference (ZPD), when n = 0 and all the lights are in phase with one another. The large burst of intensity recorded in interferogram at ZPD is known as the centerburst. The amplitude of the centerburst gives the total amount of IR light reaching the detector. An example of a centerburst is given in Fig. 3.3. On both sides of the interferograms ('wings'), the signal intensity falls rapidly to zero as the optical distance from ZPD increases.

The interferogram is a function of time and it is said to represent signal in the time domain. The time domain is then Fourier transformed to get an IR intensity plot in the frequency domain, as shown in Fig. 3.4.

$$f(t) = \int_{-\infty}^{\infty} F(\widetilde{\nu}) e^{i2\pi\widetilde{\nu}t} d\widetilde{\nu} \;\Leftrightarrow\; F(\widetilde{\nu}) = \int_{-\infty}^{\infty} F(t) e^{-i2\pi\widetilde{\nu}t} dt \tag{3.3}$$

Development of Fast Fourier Transform (FFT) algorithms was a key breakthrough enabling practical implementation of FTIR spectroscopy (Brigham 1988). Nowa-

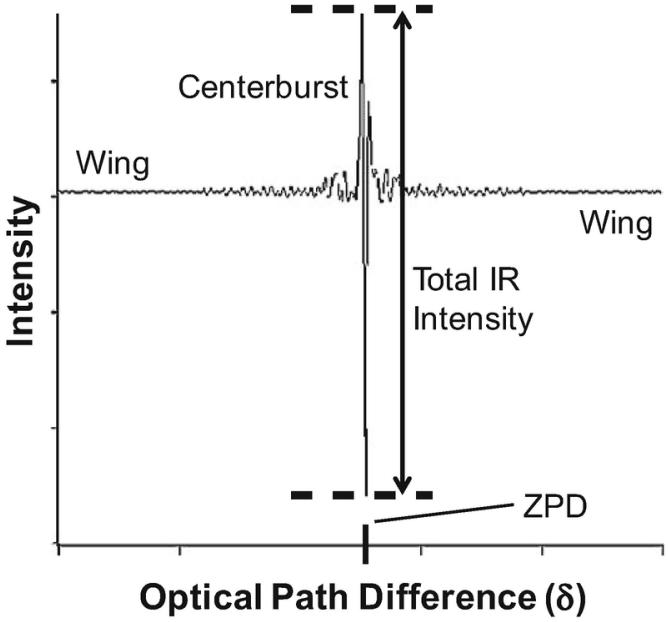


Figure 3.3: An example of the centerburst with no sample being measured, reproduced from (Khan et al. 2018) with the permission from Springer Nature

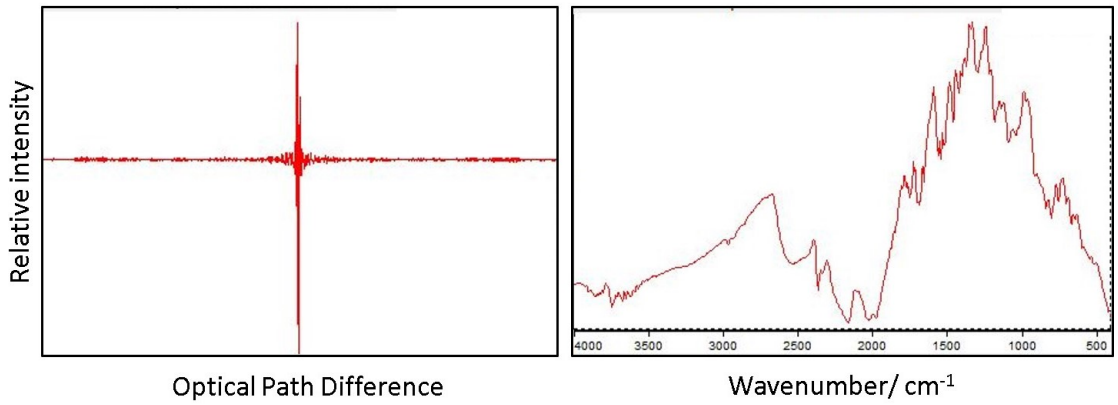


Figure 3.4: Interferogram of air in the sample compartment and its spectrum after Fourier transformation, measured using Alpha system (Bruker Inc.) in ATR-FTIR mode

days, FFT algorithms are commonly used in commercial FTIR spectrometer. The most commonly used FFT is the Cooley-Tukey algorithm. The success of FT spectrometers resulted from technology progress, as high mechanical and optical precision is essential in interference-based devices. Otherwise, a number of detrimental effects may appear as a result of beam divergence, alignment errors, velocity errors, and double modulation, for example (Chalmers & Griffiths 2006). Other residual effects need to be mitigated by numerical processing, such as the apodization procedure necessary to conceal artefacts appearing as a result of a finite path difference (Norton & Beer 1976).

Advantages of FTIR spectroscopy

FTIR spectrometer has practical advantages over dispersive IR instrument in terms of its mechanical simplicity, acquisition speed (Felgett advantage), sensitivity (Jacquinot advantage), and its capability of internal self-calibration (Connes advantage) (Thermo Fisher Scientific Inc 2013). First of all, with the use of interferometer, all wavenumbers of the IR beam can be measured simultaneously, thus speeding up the process of measurement significantly. Measurement with FTIR can be completed in a matter of seconds. On top of that, the fast measurement time allows several co-added scans of a sample to be taken to improve the signal-to-noise (S/N) ratio of the final spectrum obtained – commonly referred to as 'signal averaging'. Typically, repeating the measurement n times increases the S/N ratio by $\sqrt{n}$ times (Fellgett 1949). The amount of light reaching the detector is also higher compared to dispersive spectrometer as the use of monochromator slit in the latter, which restricts the amount of light passing through, is eliminated. Lastly, most modern FTIR spectrometer is calibrated by a laser beam of known wavelength, such as a HeNe laser of 15,800 cm^{-1}. The OPD is monitored by coupling the monochromatic laser beam collinearly with the IR beam into the interferometer. The zero crossings (destructive interferences) of the sinusoidal laser interferogram measured simultaneously at a photodiode identify the OPD in multiples of half the laser wavelength, providing an internal calibration and a signal to digitize the interferogram at equidistant OPDs (Fahmy 2013). This is much more stable and accurate than in dispersive instruments where the scale depends on the mechanical movement of diffraction gratings.

3.1.4 Measurement modes of FTIR spectroscopy

With respect to FTIR spectroscopy, there are three modes of measurements, including transmission, reflectance, and attenuated total reflection (ATR). Transmittance FTIR spectroscopy is commonly employed to obtain high-quality spectra but requires IR transparent substrates. In addition, transmission mode is subjected to total absorption patterns and limited by a certain thickness of the measured samples. On the other hand, FTIR measurement in reflectance mode has the advantage of obtaining FTIR spectra of thick and opaque samples; whilst ATR provides an alternative mean of probing the surface region of a sample (Zhao et al. 2018). ATR-FTIR spectroscopy is discussed in further detail in the following section.

3.1.5 ATR-FTIR sepctroscopy

The operation principle of ATR-FTIR spectroscopy is based on the total internal reflection of light at the interface between a high refractive index element (n_1), called an internal

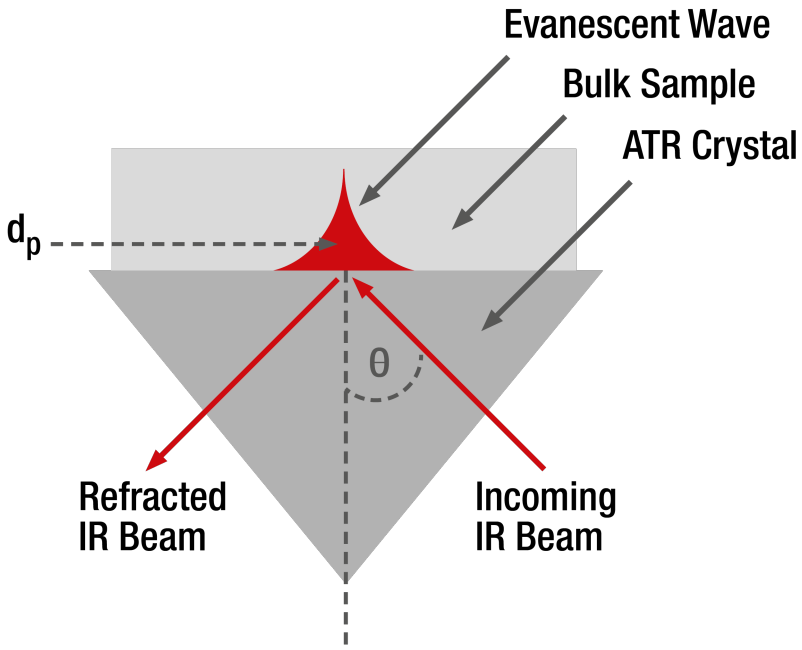


Figure 3.5: Schematic of ATR and the evanescent wave. By courtesy of Anton Paar GmbH, copyright 2020; used with permission

reflection element (IRE) and a sample of a lower refractive index (n_2). This occurs when the angle of incidence of the incoming light beam (θ) is greater than the critical angle (θ_c) – the minimum angle at which light is reflected off the internal surface of the IRE. The schematic of ATR is shown in Fig. 3.5. The critical angle is dependent on the refractive indices of both sample and crystal and can be calculated from Snell's Law [4].

$$\theta_c = sin^{-1}\frac{n_2}{n_1} \tag{3.4}$$

At the point of reflection, a remnant of the electric field that propagates from the interface into the medium when ATR takes place is known as the evanescent wave (Smith 2011). The amplitude of the electric field is largest at the interface and decays exponentially as it moves away from the surface. An important property of the ATR, known as the penetration depth, describes the distance where the amplitude of this electric field decreases to e^{-1} of its maximum value. The expression of the penetration depth [5], d_p, is given by

[4]Snell's law states that the ratio of two refractive indices is equal to the inverse ratio of the angle of incidence and the angle of refraction $\frac{n_1}{n_2} = \frac{sin\theta_r}{sin\theta_i}$

[5]Note that the value of d_p is the same for both s- and p-polarized light. This expression is only applicable for supercritical internal reflection; below the critical angle, d_p is an imaginary value. (For angles of incidence below the critical, internal reflection is subcritical internal reflection, and above the critical angle, it is supercritical internal reflection (Milosevic 2013).)

$$d_p = \frac{\lambda}{2n_1\pi\sqrt{sin^2\theta - {n_{21}}^2}} \tag{3.5}$$

as a function of wavelength (λ), the angle of incidence of the light beam (θ), and the refractive indices of medium to IRE (n_{21}) (Specac Limited 2020). ATR-FTIR spectroscopy requires that the refractive index of the IRE to be real – in other words, this translates into assuming that it is always fully transparent in the mid-IR region. In addition, the IRE needs to have a high refractive index for total internal reflection to take place. Most commonly used IREs for ATR-FTIR are diamond ($n_{Di} = 2.4$), germanium ($n_{Ge} = 4.0$), zinc selenide ($n_{ZnSe} = 2.4$), zinc sulphide ($n_{ZnS} = 2.2$), and silicon ($n_{Si} = 3.45$), just to name a few. The evanescent wave is absorbed by the absorbing medium and consequently, some of the energy from the incident wave is transferred to replenish this loss of energy. This results in a decrease in the energy of the reflected light compared to the incident beam and the internally reflected wave is no longer total – this type of reflection is known as 'attenuated' reflection (Milosevic 2012).

There are several advantages of ATR measurement over transmission, most notably in its efficiency. The ATR technique requires minimal sample preparation and hence little to no destruction to the sample; solid or liquid sample can be loaded directly onto the IRE. Liquid samples by their nature inevitably makes good contact with the IRE crystal, but the solid samples might be more challenging. In particular for imaging in ATR, good contact between the sample and the measuring surface of IRE needs to be achieved in order to ensure a high quality spectrum. The light radiation penetrates only a few micrometres into the sample species. For sample greater than this depth, the ATR technique for the collection of IR spectra is independent of the sample thickness that is crucial to the transmission mode of measurement. All these factors contribute towards highly reproducible results using ATR-FTIR. Spectra of a higher S/N ratio can be obtained if a multi-reflection ATR is used instead of single beam ATR (PIKE Technologies, Inc. 2020). A theoretical increase in sample absorbance that can be achieved with multiple beam reflection is equal to the multiple of the effective thickness and the number of reflections of the IR beam at the sampling surface (Kempfert 2004). Nevertheless, there are fewer spectral libraries for ATR users, which could be inconvenient when an unknown sample is tested (Specac Limited 2019).

ATR-FTIR spectroscopy is particularly suited to study in-situ intermolecular interactions; for instance, the supercritical fluid processing of polymers (Ewing & Kazarian 2018) and the deposition of asphaltenes from crude oil (Shalygin et al. 2019), as well as to study proteins (Glassford et al. 2013), pharmaceutical samples (Kimber et al. 2012), and biomedical specimens (Kazarian & Chan 2006).

ATR-FTIR spectroscopy in many ways resembles the transmission FTIR spec-

troscopic technique, in which the reflected light (or the transmitted light for the latter) has a reduced intensity compared to the incident beam. Likewise, the absorbance spectra of an ATR can be analysed in the same way as in a transmission experiment – they are mutually related. That is to say, sample quantification via the Beer-Lambert law is valid as long as the 'sample thickness' in ATR-FTIR spectroscopic measurements is known such that the absorbance of the spectral band measured in both ATR and transmission modes is the same. This is known as the effective thickness (d_e) and can be obtained from the following correlations[6].

For s-polarised beam:

$$d_{e\perp} = \frac{\lambda n_{21} cos\theta}{n_1 \pi (1 - {n_{21}}^2)\sqrt{sin^2\theta - {n_{21}}^2}} \tag{3.6}$$

and for p-polarised beam:

$$d_{e\parallel} = \frac{\lambda n_{21} cos\theta (2sin^2\theta - {n_{21}}^2)}{n_1 \pi (1 - {n_{21}}^2)[(1 + {n_{21}}^2)sin^2\theta - {n_{21}}^2]\sqrt{sin^2\theta - {n_{21}}^2}} \tag{3.7}$$

where λ is the vacuum wavelength, θ is the angle of incidence of light beam at the surface of IRE, n_1 is the refractive index of IRE, n_2 is the refractive index of the sample, and n_{21} is the ratio of n_2 to n_1. In brief, the effective thickness for s-polarised light is always less than the effective thickness of the for p-polarised light when all parameters, such as the angle of incidence of IR light, are kept constant. In general, when the efficiency of the instrument in s- and p-plane is similar, the average values of $d_{e\perp}$ and $d_{e\parallel}$ can be taken as the effective path length of non-polarized light in Eq. 3.8 below.

$$d_e = \frac{d_{e\perp} + d_{e\parallel}}{2} \tag{3.8}$$

Macro vs Micro ATR-FTIR spectroscopic imaging

The micro ATR-FTIR spectroscopic approach measures a relatively small area, c.a. 0.01 mm^2, of any sample whilst a commercial macro ATR-FTIR spectroscopic accessory may measure area of about 4 – 20 mm^2. On the flip side, micro ATR-FTIR spectroscopic imaging offers a higher spatial resolution of ∼3 – 6 μm compared to macro ATR-FTIR spectroscopic imaging at ∼13 – 18 μm (Chan & Kazarian 2003). In study where a high spatial resolution is not needed, imaging using macro ATR-FTIR spectroscopy at a lower magnification but increased field of view (FOV) would be more beneficial.

[6]These definitions of effective thickness exist only for angles of incidence above the critical angle and is valid in the low absorption approximation.

Macro ATR-FTIR spectroscopic imaging

A magnification of 1:1 can be found in commercially available macro ATR-FTIR spectroscopic system. Therefore, the FOV is only limited by the size of the imaging detector. For example, a 64 × 64 FPA with a pixel size of 40 μm will generate an image of 2.56 × 3.58 mm^2 (Kazarian & Chan 2010). For simple sample handling, an inverted IRE prism is often used in the macro ATR-FTIR accessory, with two plane mirrors that reflect the light to the prism and direct the light beam to detector. A large sample compartment (Fig. 3.6b) can be used to accommodate the ATR accessories (also known as the 'Golden Gate' for a single reflection diamond ATR accessory with more complex optics). The 'Golden Gate' is shown in Fig. 3.6a. The alignment of the 'Golden Gate' in the sample compartment can be done by adjusting the mirrors or shifting the accessory along the beam path. The optical design of the sample compartment determines the magnification of the system.

(a) 'Golden Gate' accessory (Specac Inc.) (b) Bruker Tensor IMAC compartment

Figure 3.6: The set-up of macro ATR-FTIR spectroscopic system

There is the issue with image distortion when the inverted prism is used. The reason for the image distortion was discussed by (Chan & Kazarian 2003). Imagine the light beam to be cone (or cylindrical)-shaped, when the light enters the IRE prism at 45° with respect to the sampling surface, the area imaged on the surface is elliptical; yet the beam collected after internal reflection is once again in cone shaped. This results in a change in aspect ratio of the image obtained, i.e. the image is compressed in the direction of the beam path. A schematic diagram showing the deformation of the image is presented in Fig. 3.7.

The good news is the distortion of the image can easily be corrected by applying a scale factor to the axes where distortion takes place, that is determined by the angle of

incidence of the light beam. With the example above at 45°, the correction factor is thus $\sqrt{2}$.

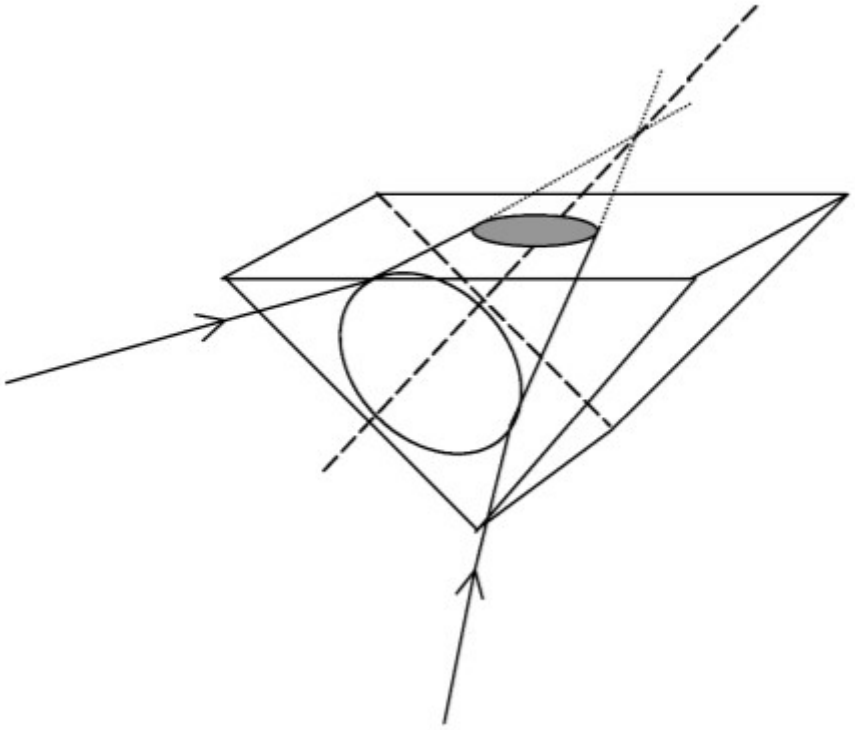


Figure 3.7: Image distortion in the macro-ATR set up, reproduced from (Chan & Kazarian 2003) with the permission from SAGE Publications

Spectral distortions in ATR

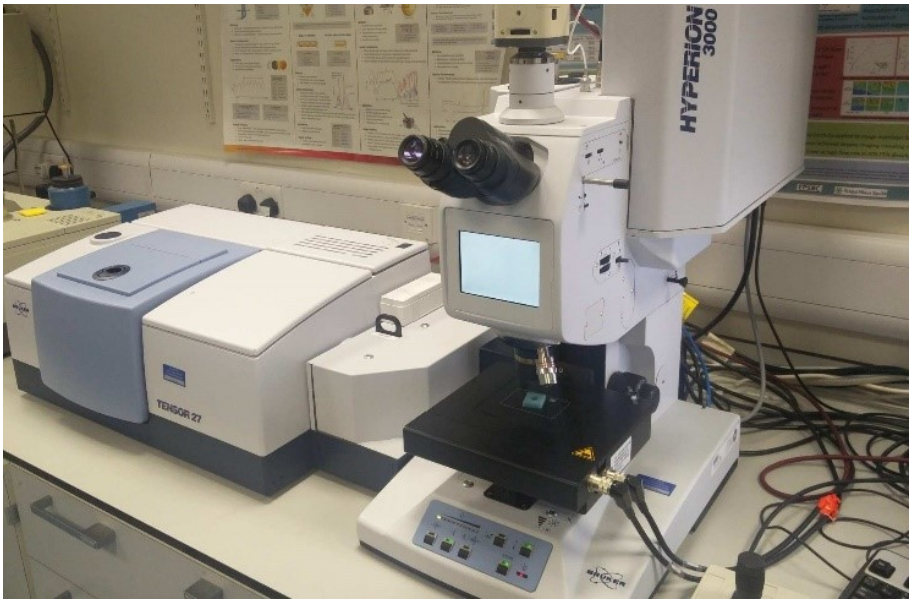
The IR spectrum obtained in ATR measurement mode is somewhat different from that obtained in transmission measurement. The distortion of the relative peak absorbance measured by ATR can be explained in terms of the path length of measurement. Principally in transmission experiment, this refers to the thickness of the sample and is thus constant across the different wavelengths of IR light. This is different for ATR measurement whereby the penetration depth (d_p) is a function strongly dependent on wavelength of light, the angle of incidence, and the refractive indices of both samples and the IRE (Eq. 3.5). Besides, most commonly established band positions and spectrum libraries nowadays are available for transmission spectrum. As a result, the ATR spectrum needs to be corrected from its distortion of the relative intensities of the bands and the shifting of the bands to resemble transmission spectrum if it were to be compared to transmission experiment. The origin of the spectral distortions and the correction algorithms available are described in Appendices section 6.4.

FTIR microscopy

The first IR microscope was built by C.R. Burch in 1947 (Burch 1947) and manufactured by Perkin-Elmer at an early stage. IR microscopes have become very widely used since FTIR was introduced. They are applied to obtain spectra of very small samples or of small parts of larger samples. Today, essentially all commercial FTIR spectrometer manufacturers offer optimized FTIR microscope systems, and FTIR microscopy has become a routine analytical tool. Most of these systems provide some form of hyperspectral imaging, an extremely powerful tool for advanced materials characterization (Vichi et al. 2018), forensic identification (Ewing & Kazarian 2017) and, most importantly, medical diagnostics (Kimber & Kazarian 2017). The FTIR microscopes used in this study are shown in Fig. 3.8a and Fig. 3.8b.

Operation of IR microscopes in transmission mode requires more sample preparation, one of which is the control of the sample thickness. The main requirement is that the sample needs to be sufficiently thin to produce an absorbance value of ideally less than 0.8 in order for Beer-Lambert law to be applied (Chalmers & Griffiths 2006). A general rule of thumb for an optimum sample thickness to ensure good qualitative and quantitative analyses is around 10 μm (Chalmers & Griffiths 2006). Sample preparation could involve flattening the samples, however, the methodology that can be used to control the thickness is highly dependent on the sample's physical properties and whether the alteration on the sample can be done without affecting its physical properties. In the case of biopsy section where the sample integrity must be maintained, the sample is microtomed to the desired thickness. Once the samples are suitably thin, it is mounted on a support that is transparent in IR. Common mounting materials commercially available include CaF_2, NaCl, and KBr substrates (with useful spectral range of $40\,000\,\mathrm{cm}^{-1}$ to $\sim$$600\,\mathrm{cm}^{-1}$, respectively). One of the major issues to be taken into account when conducting a transmission experiment is the shift in the focus of the microscope as a result of the refraction in the substrate (Messerschmidt 1987), which is illustrated in Fig. 3.9. The change in focus is dependent on the refractive index and thickness of the substrate. For example, a 2 mm thick KCl window ($n_{KCl} = 1.468$ at $1700\,\mathrm{cm}^{-1}$) will induce a focal shift of $\sim$ 800 μm (Chalmers & Griffiths 2006). An approach has been taken that include readjustment of the focus with a hemispherical lens (Kimber et al. 2016) which is described later in Section 4.3.

In a FTIR microscope, light from the source is first focused onto the sample using a condenser. Light transmitted by the sample is then collected by a Schwarzchild objective, which houses reflective surfaces with spherical geometries. Most optical microscopes use lenses to focus light and remove chromatic aberrations by the principle of refraction; however, FTIR microscopes employ curved mirror surfaces to condense and collect IR radiation. This is because optical microscope performs over a limited range of wavelength



(a) Hyperion 3000 FTIR microscope (Bruker Inc.)

(b) Cary 620 FTIR microscope (Agilent Inc.)

Figure 3.8: FTIR microscopes used in this thesis

~ 0.5 μm whereas IR microscopes need to be able to operate over a range of ~ 1 mm which is far too large for the chromatic aberration to be corrected via refraction. Most manufacturers provide objective with NAs of 0.6 at 6× and 15× magnification. Objectives at higher magnification (i.e. 32×, NA = 0.65) are also available at the expense of working distance. After the sample, there is an aperture typically made of four independently

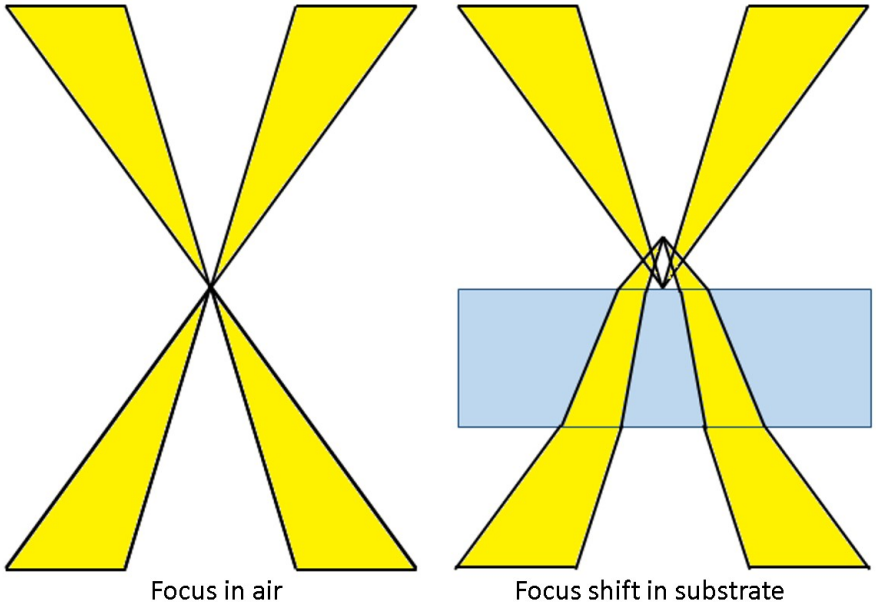


Figure 3.9: Focal shift in transmission mode in FTIR microscopes

adjustable knife blade that can be used to limit the size of the image. Lastly, a magnified image of the illuminated sample is recorded by a detector. The optical path in an IR microscope is shown in Fig. 3.10. In addition, by switching mirrors in the optical train, the microscope can be converted from transmission mode to reflectance mode, and vice versa.

Micro ATR-FTIR spectroscopy

Similar to macro ATR-FTIR spectroscopy, micro ATR-FTIR spectroscopy offers the perks of minimal sample preparation and also a shorter probing length which is ideal for highly absorbing material. In particular, in the microscope configuration, there is an additional advantage of increase in spatial resolution over the transmission experiment (spatial resolution improves with the refractive index of IRE), yet the measurement area is smaller (Lewis & Sommer 1999). Typical contact areas for ATR-FTIR microscope range from 10^{-3} to 10^{-1} mm^2 (Chalmers & Griffiths 2006). The implementation of ATR measurements on a FTIR microscope can be as simple as using a Ge hemisphere in the objective of the microscope, in a set up illustrated in Fig. 3.11. An interesting extension of micro ATR-FTIR spectroscopy is the 'mapping' method with the use of mapping stages. Manufacturer, Bruker for example, has come up with software that allows the stage to be lowered and raised automatically between different sampling locations. However, in this kind of procedure, there is always the risk of sample contamination, i.e. the material from one section of the sample is transferred to the crystal and is subsequently transferred to

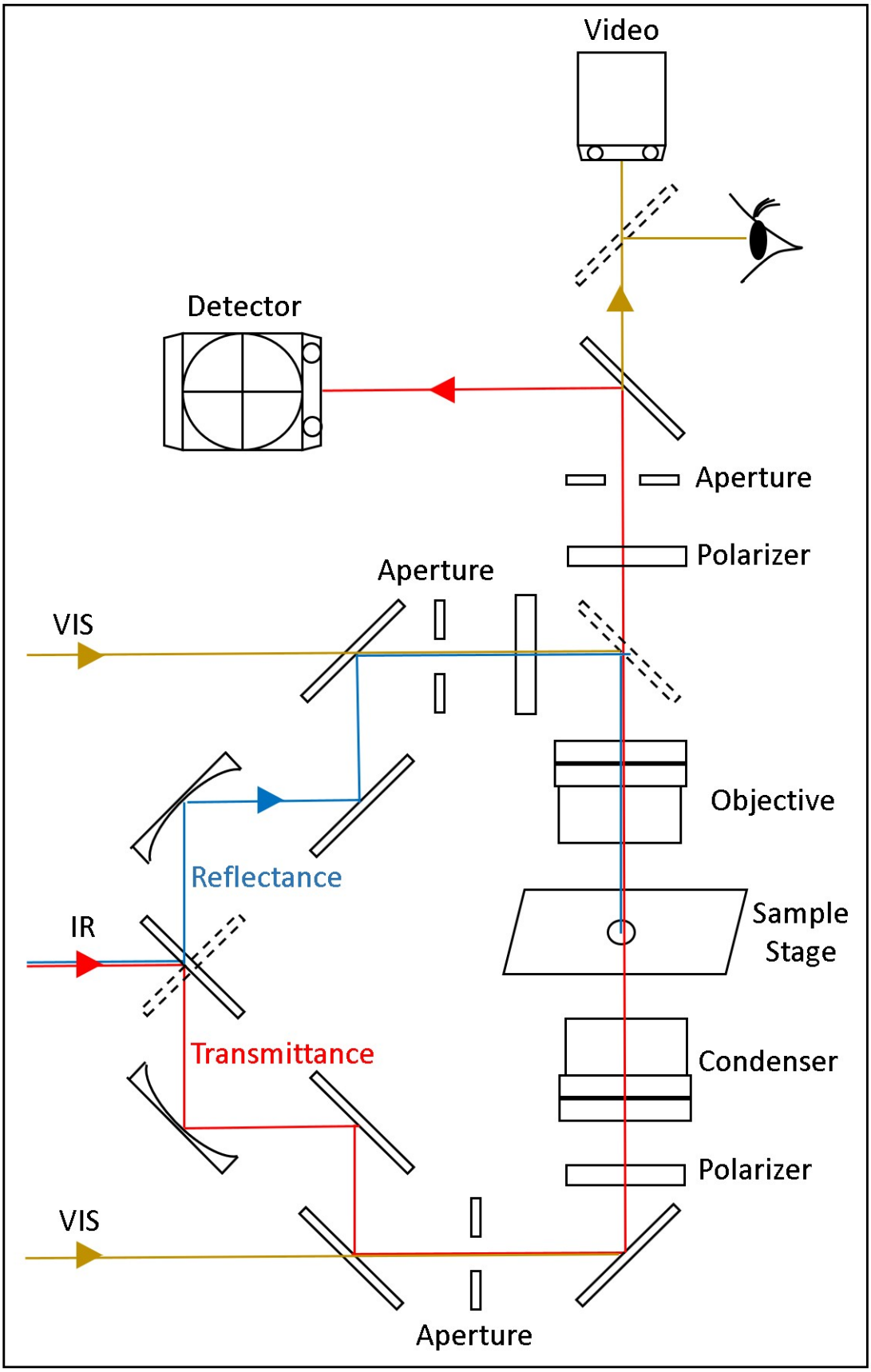


Figure 3.10: Schematic of the beam path inside an FTIR microscope, adapted with permission from Bruker Inc.

the next measurement section, as well as the lack of time between measurements to relax between neighbouring locations (Nakano & Kawata 1994). A new micro ATR-FTIR set up involving a large area Ge crystal devised to map soft materials is introduced in this thesis in Section 4.6.

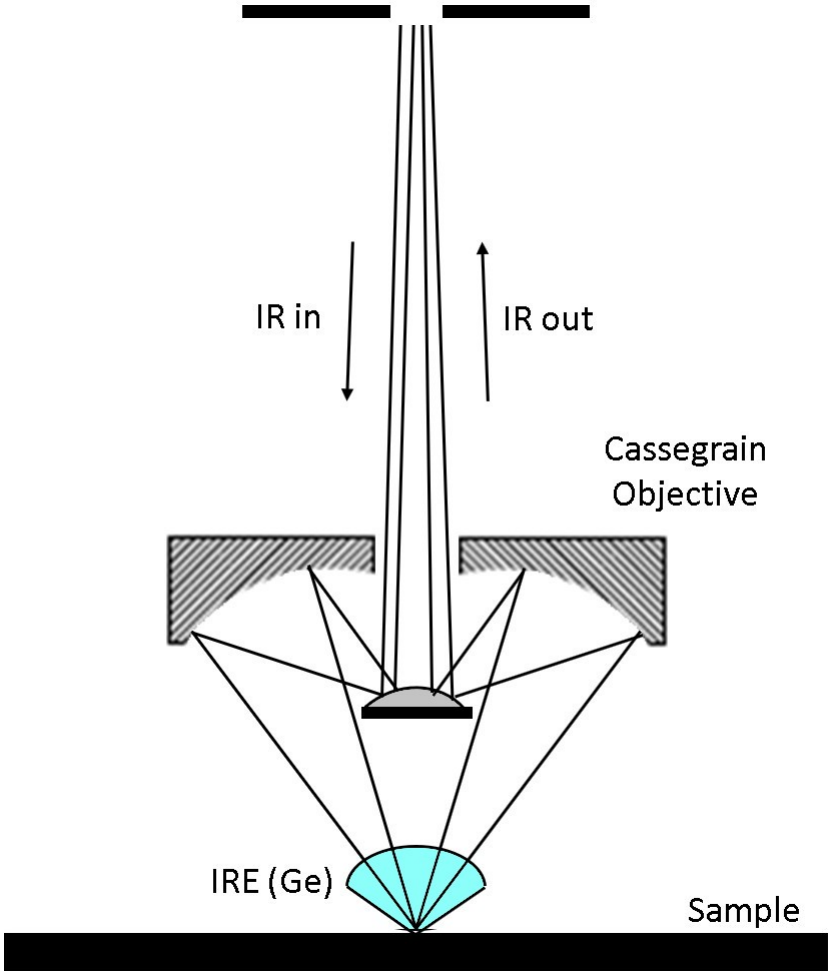


Figure 3.11: Schematic showing the set-up of a germanium IRE in ATR-FTIR microscope

The building blocks common for all FTIR spectrometers are a radiation source and a detector. The source is an inert material that is heated to an elevated temperature for thermal emission of radiation (black-body emission spectrum) with a maximum intensity in the IR region; whilst a focal plane array (FPA) detector is used for hyperspectral imaging to acquire FTIR spectroscopic images (see Appendices 6.5 for types of sources and detectors).

Hyperspectral imaging

The more advanced use is the hyperspectral imaging (HSI) technique where the advantages of optical spectroscopy as an analytical tool are combined with two-dimensional (2D) ob-

ject visualization obtained by optical imaging (Vasefi et al. 2016). Data obtained from HSI systems are 3-dimensional (3D) structures that consist of 2 spatial (x- and y-directions) and 1 spectral dimension (Su & Sun 2018), as illustrated in Fig. 3.12. This data hypercube can be created through various scanning method, such as single point mapping, linear array mapping, and FPA imaging (Lu & Fei 2014). FPA imaging is described in detail here as it is the approach used in this thesis. FPA imaging poses several undeniable advantages such as fast acquisition over the other methods, thanks to its ability to acquire thousands of spectra simultaneously.

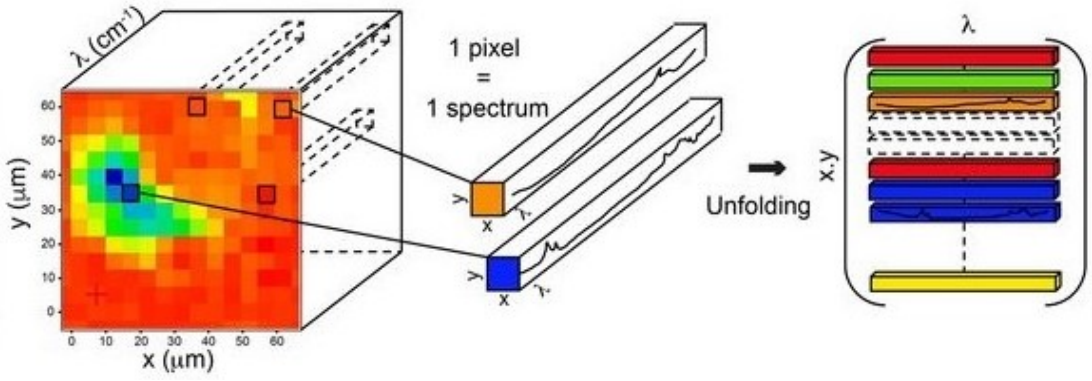


Figure 3.12: Hyperspectral data cube to unfolded matrix acquired from FPA imaging, reproduced from (Pisapia et al. 2018) with the permission from Springer Nature

FPA consists of an array of photon-sensitive pixels with its surface being a photovoltaic panel. Similar to other photovoltaic detectors, each pixel detects photons at certain wavelengths and generate a corresponding voltage proportional to the number of photons. The digitised output voltages are used to construct an image of the sample, also known as the chemical map or chemical image, via the hyperspectral cube of information (Dorling & Baker 2013). This is possible as each pixel collects completely independent IR spectrum from different spatial location of the measured sample. The best commercially available FPA detectors today consists of 128 $\times$ 128 pixels, permitting the simultaneous collection of 16,384 individual spectra at one time (Kazarian & Chan 2013).

3.2 Sample preparation

3.2.1 FFPE tissue

Tissues are first collected from patients through surgical procedure. Formalin-fixed paraffin embedded (FFPE) technique is a way of preserving the tissue specimen obtained that could be used later in examination or experimental research. The first most important step in this process is tissue fixation, whereby a chemical called formaldehyde (also known as formalin) is used to react with the tissue. Vital morphology and proteins' states within

the tissue is preserved in this step. Next, the tissue is dehydrated, commonly with ethanol, isopropanol, or glycol ether dehydrants in tissue processing to remove the free water. Following this, the tissue section is impregnated or embedded with paraffin wax to produce tissue block; this makes it easier for it to be microtomed into thin sections (Feldman & Wolfe 2014).

FFPE vs frozen tissue

Both FFPE and frozen biopsy samples can preserve the specimens adequately well but each of them has their own pros and cons (BioChain Institute, Inc. 2018). FFPE tissue blocks have been archived for decades for later use in 'biobanks' developed by universities and hospitals. This has resulted in FFPE being a plentiful resource of research material as well as being able to offer a vast collection of historical perspective. This is because FFPE samples remain stable after a long time as well as can be stored in a cabinet at room temperature. Furthermore, since FFPE is a well-established method of preserving the tissue specimens, pathologists are accustomed to making diagnoses from them. However, formalin involved in this procedure is toxic. The fixation process can also be time consuming. More importantly, the proteins in FFPE, despite being preserved, is no longer biologically active as they are often denatured in the fixing procedure. Additionally, the nucleic acid (DNA and RNA) is not very well preserved, thus limiting the use of FFPE tissue samples in molecular analysis. Frozen tissue sample proved to be better for molecular genetic analysis (Geneticist Inc. 2018). Proteins are kept in its native state, which makes frozen tissues ideal in immunohistochemistry. Nonetheless, storage of frozen tissues can be costly, and the tissues are more vulnerable in situations where there are power outages and mechanical failures. When compared to FFPE tissues, this method of preserving tissue specimens is relatively newer, hence the collection of frozen tissue in biobanks is smaller and limited for experimental research (Luder Ripoli et al. 2016).

3.2.2 Tissue deposition and de-paraffinisation

The samples were microtomed at 3-μm thickness from a paraffin-embedded specimen block, one of which was mounted onto a 2-mm-thick CaF_2 window (Crystran Ltd., Dorset, UK) with the adjacent section mounted onto a glass slide and stained with haematoxylin and eosin (H&E) by pathologists for viewing under visible microscope. The sections deposited on the CaF_2 window were deparaffinised for IR measurements. The protocol used for deparaffinisation was as follows: the tissues were washed with hexane (HPLC grade) for 5 minutes, followed by 100, 95, 70, and 50 (v/v%) ethanol (HPLC grade), each for 2 minutes before being air dried for 1 hour at ambient condition. The tissue samples were kept in a desiccator at room temperature when no measurements were taken throughout the study.

The samples used in this study were:

- Two prostate biopsy samples (1 control and 1 diseased) from the same patient were provided by Kingston Hospital (London, UK). A Gleason score of 3 + 3 was assigned to the areas of malignancy identified within the H&E stained tissues. (see Fig.4.2)
- Six colon biopsies (2 control and 4 diseased), from different patients, at different disease stages of malignancy (hyperplasia, dysplasia, and cancer), provided by Prof. Robert Goldin at St. Mary's Hospital (Imperial College London, UK). (see Fig. 4.16 H&E and 4.17 H&E)

The tissues were microtomed by pathologists. The disease 'class' or 'category' of the adjacent tissue sections, that were retained for pathological H&E staining, were identified by pathologists.[7] Research ethics code for the polyps is 14/EE/0024 and approved by Imperial College London Research and Ethics Committee.

The prostate tissue samples were already de-paraffinised by postdoctoral researcher, Dr. Martha Vardaki, from the same research group; whilst the de-paraffinisation of colon tissues was carried out by myself with the same protocol used for prostate tissues as outlined above. The absence of spectral peak of paraffin at 1462 cm^{-1} was used as an assessment for the paraffin removal. Although chemical dewaxing does not warrant complete paraffin removal as described earlier in Section 2.5.1, examination of its characteristic spectral bands here made sure its presence was greatly reduced to minimize any potential impacts it has on the interpretation of the spatial heterogeneities of the chemical images. The positive outlook was supported by recent studies that the presence of paraffin has null effect on the absorbance, surface area, full width at half maximum, and peak position observed in the spectrum peaks of the biological samples when the spectral bands of paraffin are minimized accordingly (Depciuch et al. 2016, Chaber et al. 2017).

3.2.3 H&E staining

Staining is a common process for medical examination of cancer in which a dye is applied on tissue specimen to locate the tumorous cells (Alturkistani et al. 2015). The standard stain of anatomical pathology diagnostic is the H&E stained tissue section, but before tissue staining can take place, the FFPE tissue section needs to be de-paraffinized. The basis of H&E is selective staining based on the acidity or basicity of the stains and their ability to form cross-linkages with ionizable radicals within the tissue. Hematoxylin is a positively charged (cationic) basic dye. Eosin is a negatively charged (anionic) acid dye.

[7] Pointers to the disease category were given to general area, not all areas of the samples fully categorised by the histologists. The areas examined were suggested by pathologists and the borderline of the cancer types falls beyond the small section assigned.

Tissue elements that exhibit affinity for the basic dye are termed basophilic or hematoxyphilic; whilst tissue elements that react with the acidic dye are termed acidophilic or eosinophilic. Sequential application of the dyes to histologic section results in cytoplasm and extracellular matrix being stained with varying degree of pink whereas nuclei are stained blue (or purple) (Fischer et al. 2008). Red blood cells are stained intensely red. It is important to note that lipid, partially dissolved out of the cells by the reagents used during tissue processing will lack staining and appear as empty spaces (Chan 2014). Table 3.1 summarises the properties of the most important cellular organelles in the H&E staining method. The interplay of colour shown by the H&E stain gives the internuclear details. Some other advanced stains in the same line as routine H&E are immunohistochemical (immunological labelling with fluorescent or enzymatic stains) and the *in situ* hybridization (Musumeci 2014).

Table 3.1: Nature of the organelles in a living cell

Basophilic	• Nucleus (and parts of cytoplasm that contains RNA) • Rough endoplasmic reticulum (RER) • Ribosome
Eosinophilic	• Mitochondria • Cytoplasm • Most proteins (including cytoplasmic filaments in muscle cells, intracellular membrane, and extracellular fibres)

3.3 Spectral processing

This section discusses the spectral processing to enhance the spectral information and make spectral interpretation easier. Firstly, spectral subtraction is discussed, which is a method applied to extract spectrum of an individual components from a mixture of different compounds. Next, baseline correction is described. Third, spectral derivatives to determine peak positions of spectral bands and deconvolve overlapping peaks are presented. Lastly, the S/N ratio of any spectrum can be improved by smoothing.

3.3.1 Spectral subtraction

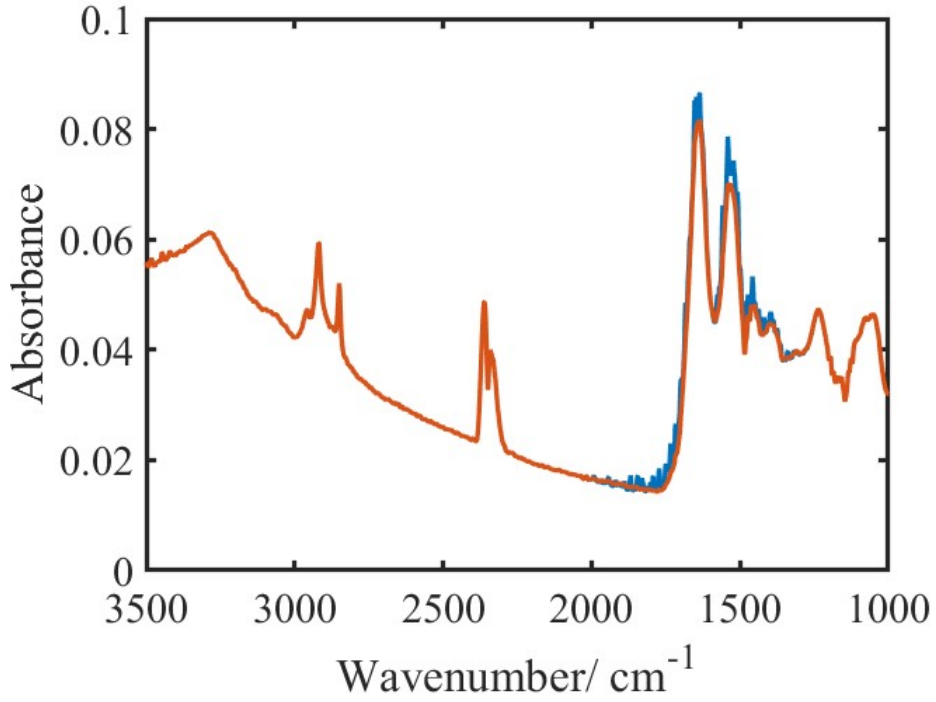
Spectral subtraction is commonly applied to eliminate unwanted contribution, such as that of atmospheric CO_2 and water vapour which will lead to erroneous conclusion if not removed (see discussion under section 2.5). The sample spectra in Fig. 3.13 are from healthy colon tissue, whilst the spectrum used for subtraction are the atmospheric background. For subtraction to work properly, a reference peak needs to be identified. The spectra were checked for water vapour correction based on its second derivative spectra in the $1900 - 1750\,\mathrm{cm}^{-1}$ region. Unfortunately, the absorbance of the spectral bands of sample to reference is not always 1:1, which can make subtraction complicated. However, this situation can be accounted for by introducing scale factor. The mathematical algorithm behind spectral subtraction is relatively straightforward, as follows:

$$\text{Resultant absorbance} = \text{Sample absorbance} - \text{scale factor} \times \text{Reference absorbance} \quad (3.9)$$

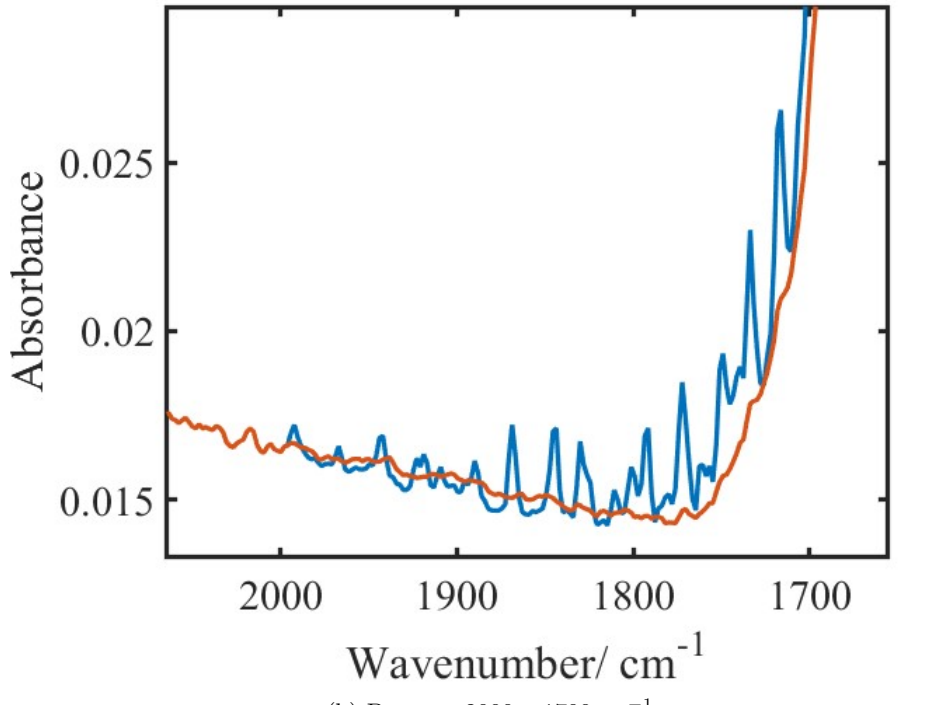
The scale factor is a user adjustable parameter. Ideally, the best scale factor is identified when the resultant absorbance contains no features from reference absorbance or is said to be free from contaminant. Although spectral subtraction is a subjective but legitimate way of simplifying a mixture spectrum, the resultant absorbance after subtraction is noisier than the original spectrum. A good rule of thumb here is to use only original data with a high S/N ratio to yield useful data after subtraction (Smith 2011).

3.3.2 Baseline correction

An ideal IR spectrum should have a flat baseline at zero absorbance (Griffiths & de Haseth 2007), however, this does not happen in most cases. The baseline distortion generally falls into few categories, one of which is the offset of the entire baseline of the spectrum, equivalent to adding a constant value to all the absorbance values in a spectrum. This happens when an equal amount of IR light is reflected or absorbed at all wavenumbers, possibly due to a thick sample. An example of this is the thick KBr pellet spectrum of aspirin (Smith 2011) which has a constant offset of 0.2. Fortunately, a constant baseline offset does not need to be corrected if the integrated absorbance is analysed as it is independent of this baseline shift. The second type of a baseline distortion is slope, as a result of scattering (Coleman 1993) which has been discussed under Section 2.5.3. Specific particulate matter of samples may scatter the infrared beam and the extent of scattering increases with wavenumber, resulting in a sloped spectrum as shown in Fig. 3.14. Sloped baseline problem due to the nature of the sample like this cannot be usually fixed experimentally, hence we turn to correction algorithm to help solve this issue. The



(a) Between 3500 – 1000 cm^{-1}



(b) Between 2000 – 1700 cm^{-1}

Figure 3.13: Example spectrum from healthy colon tissue before and after water vapour subtraction (blue and orange line, respectively)

algorithm used in the work in this thesis is the RMieS-EMSC correction algorithm which incorporates the baseline correction.

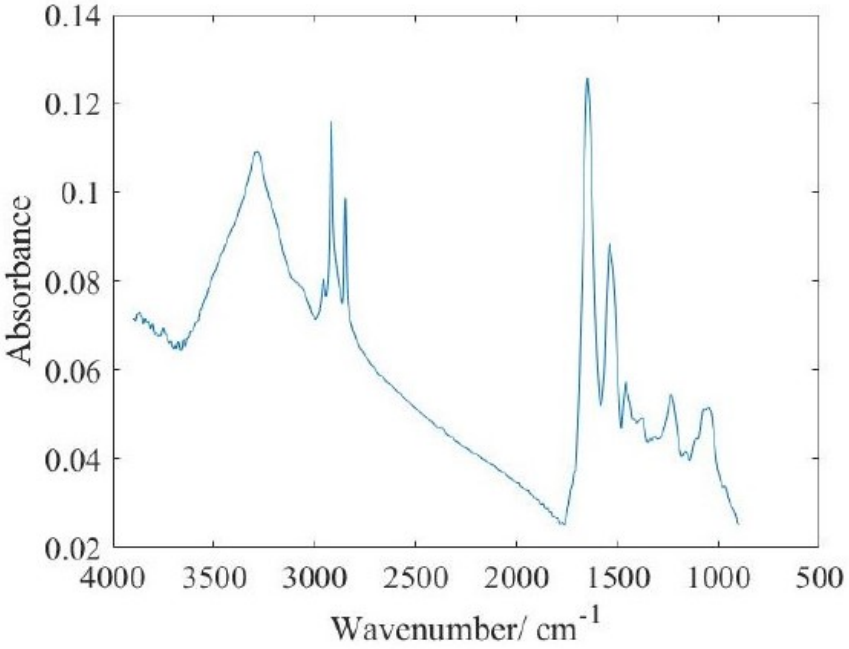


Figure 3.14: Example spectrum showing dispersion artefact arising from resonant Mie scattering effect, reproduced from (Song et al. 2019) with the permission from Springer Nature

3.3.3 Normalization

A normalisation step is essential for comparing variations in absorbance between different spectral bands. Based on Beer-Lambert law, the absorbance of a specific spectral band is dependent on its concentration, thus absorbance of different spectral bands might exhibit different range of absorbance. One simple approach is vector normalisation, which maps the spectra onto a unit sphere in multivariate space. Vector normalization is carried out in the following way: the square root of the sum of the squared absorbances of the spectrum (the 'norm') is first calculated. Then, the spectrum is scaled by dividing each of the spectral absorbances by the 'norm' (Gautam et al. 2015). Other simple approaches include normalisation to constant total and min–max normalisation, which scales each spectrum to between 0 and 1. Without normalisation, the spectral band with lower absorbance in the tissue section will be given less weight in subsequent multivariate analysis, which may mask important variations due to compositional changes that may relate to pathology (Hermes et al. 2018).

3.3.4 Spectral derivatives

From simple calculus, we know that the first derivative of any mathematical function represents its slope. By calculating the first derivatives of an IR spectrum, an absorption band can easily be identified where the slope of the peak is zero (the zero-crossing point) and the two sides of the absorption band have different slopes, i.e. below and above zero respectively. This bipolar function is characteristic of all odd-order derivatives (Owen 1995). The concept of derivatizing spectral data was first introduced in the 1950s and was popularised in the late 1970s, when it becomes practicable to use mathematical methods to generate derivative spectra quickly, easily and reproducibly. This significantly increased the use of the derivative technique with respect to wavelength for qualitative analysis and for quantification, also known as the 'Derivative spectroscopy'.

A huge advantage of taking derivatives of the spectra is that they do not contain offset – their baselines are at zero. Likewise for second derivatives of the spectrum, which measures the change in slope or concavity (Smith 2002). The second derivative has three lobes – a positive value followed by a negative and then a positive value. The peak minimum in second derivatives corresponds to the position of the peak in the spectral band, therefore, second derivative spectra are commonly used in peak picking, peak identification, and library searching. Compared to first derivative, the added advantage of higher derivatives is its ability to deconvolve or separate overlapping peaks without resolving to more complicated processing technique. The simplest way to calculate the derivatives is to apply mathematical techniques to the spectrum with a constant sampling interval. The sampling interval is dependent on the natural bandwidth of the absorbance bands or the bandwidth of the instrument. In FTIR spectroscopy, this interval refers to the spectral resolution, $\Delta\widetilde{\nu}$ i.e. $4\,\mathrm{cm}^{-1}$ or $8\,\mathrm{cm}^{-1}$. The calculation of the first derivative, $\Delta\widetilde{\nu}$, can be carried out for an intermediate wavenumber between two absorbance wavenumbers following the differential equation, such that

$$D_{\widetilde{\nu}+\frac{\widetilde{\Delta\nu}}{2}} = \frac{A_{\widetilde{\nu}+\widetilde{\Delta\nu}} - A_{\widetilde{\nu}}}{\widetilde{\nu}} \tag{3.10}$$

The method is a linear interpolation between adjacent wavenumbers. Similarly, for higher order derivatives, such as the second derivative, absorbance values at three closely-space wavenumbers are used in similar expression as above

$$D_{\widetilde{\nu}} = \frac{A_{\widetilde{\nu}-\widetilde{\Delta\nu}} - 2A_{\widetilde{\nu}} + A_{\widetilde{\nu}+\widetilde{\Delta\nu}}}{\widetilde{\Delta\nu}^2} \tag{3.11}$$

In addition, the derivatives are very useful in obtaining quantitative information of the spectrum. The integrated area of a spectral feature in the derivative spectrum is

proportional to its area under the peak in the zero-order spectrum. In other words, the linear relationship between concentration and absorbance as described in Beer-Lambert Law still applies. The modification to Beer's Law for n^{th} order derivative is

$$\frac{d^n A}{d\widetilde{\nu}^n} = \frac{d^n \varepsilon}{d\widetilde{\nu}^n} cl \tag{3.12}$$

A problem with derivative spectra is that they contain more noise than the spectra from which they are calculated. In order to improve the signal, smoothing algorithm can be applied while calculating the derivatives.

3.3.5 Smoothing

FTIR software packages often come with smoothing functions. This includes the Bruker OPUS software and the Agilent Resolution Pro software which are used in this study. Among others, the best-known and most widely used smoothing algorithm is the Savitsky-Golay (SG) algorithm (Savitzky & Golay 1964). In this algorithm, the first step is to select an appropriate size of smoothing window, i.e. setting the number of points in the windows (5-, 7-, or 9-point being the most common one) (Baker et al. 2014). The amplitude of smoothing is proportional to the number of data points included in the window. Next, a polynomial function ($A_{\widetilde{\nu}} = a_0 + a_1\widetilde{\nu} + ... + a_n\widetilde{\nu}^n$) is fitted to the set of data points in each smoothing window. The higher the degree of polynomial function fitted to the data, the less smoothing is achieved. One can imagine if the polynomial order, n, is less than the number of data points in the window, it is impossible for the polynomial function to fit all the points, hence a smooth approximation of the data is obtained. This is widely used to counteract the degradation of the signal inherited from the derivatization described above in a combined algorithm, in which the coefficients at each wavenumber is multiplied by the factorial of the derivative order, a_1 is the first derivative, $2!a_2$ the second derivative, $3!a_3$ the third derivative, and so on (Owen 1995).

The approach adopted in this study to find the optimum smoothing parameters is to begin the with a small number of data points in the smoothing window and gradually increase the number in small increments. This is terminated when the peaks start to broaden significantly or merge together. In a nutshell, the choice of the processing techniques and their parameters is rather subjective and dependent on the operator's judgement.

3.3.6 Chemometrics

Chemometrics is the use of mathematical and analytical methods to optimize the design of experiments and improve the understanding of chemical information from large and complex datasets (Varmuza & Filzmoser 2009). Naturally, statistical analysis of such data should employ one or more multivariate statistical tools – the simultaneous analysis of dependent variables (outcome) against a plethora of independent variables (predictor). The most common types of multivariate tests for the processing of spectral data include factor analysis such as principal component analysis (PCA), cluster analysis (CA), partial least squares discriminant analysis (PLS-DA), artificial neural network (ANN), and many others (Christou et al. 2018). Experimental data in spectroscopy can be either qualitative or quantitative. The data is considered qualitative when it falls into one of these three categories, i.e. nominal (such as three patients, six tissue samples and so on); dichotomous (male/female); and ordinal (degree of malignancy of tissue samples, such as 1= healthy, 2 = pre-cancerous, 3 = cancer etc.) and quantitative when continuous variables are involved (the wavenumber range of measurements taken). Most often the data is a mix and match of both. The good news is that regardless of the type, all variables can be analysed by the chemometrics approach (Szymanska et al. 2015). The objective of using chemometrics to analyse spectral datasets is to obtain insights into classifying the cancer specimens based on their chemical specificity obtained from FTIR spectra. Due to the nature of each experiment and the extent of available data, it is deemed unsuitable for all aspects of chemometrics to be covered simultaneously in any one research study.

Exploratory analysis

PCA analysis is the most important statistical technique for chemometric analysis of spectral data (Varmuza & Filzmoser 2009). The methods commonly required pre-treatment or pre-processing of the data (Skov et al. 2014), which could be normalization or scaling, as well as anomaly detection and removal (Nunes et al. 2015). These two analytical methods are often employed for exploratory analysis of the data. PCA can be thought of as an unsupervised learning problem – no *a priori* knowledge about the sample is required. The central idea of PCA is to reduce the dimensionality of the data while retaining the largest number of variations possible, achieved by transforming and projecting the data onto new sets of uncorrelated axes, known as the principal components (PCs). The data matrix $\mathbf{X}$ (I $\times$ J) is decomposed by the following equation:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{t}} + \mathbf{E} \tag{3.13}$$

where $\mathbf{T}$ is the (I $\times$ A) scores matrix, $\mathbf{P}$ is the (J $\times$ A) loadings matrix, $\mathbf{E}$ is the (I $\times$ J)

residual matrix, and A is the number of PCs (Granato et al. 2018). The number of PCs is the most important parameters for the meaningful results of PCA. A suitable number of dimensions needed to explain the data variation at a selected significance level can be obtained by subjecting the data to a Bartlett's test of sphericity (The MathWorks, Inc. 2020*a*) or by selecting the number of PCs based on the total variance obtained in a plot of variance against PCs (Gonsales 2018). The latter approach was chosen in this study for its convenience.

Another chemometric analytical method that can be used for exploratory study is the cluster analysis. Among the multiple clustering methods, the non-hierarchical methods such as k-means and k-medians, and hierarchical cluster analysis (HCA) are most frequently used (Baker et al. 2014). In k-means clustering, the data is separated in k different clusters, chosen to be far enough apart from each other spatially in Euclidean distance to produce effective data mining results. Each cluster has a centre, called the centroid, and the data point is clustered into a certain cluster based on how close the features are to the centroid by minimising the objective function shown below in a set of iterations:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(\|x_i - \nu_j\| \right)^2 = 1 \tag{3.14}$$

where $\|x_i - \nu_j\|$ is the Euclidean distance between a point, x_i, and centroid, ν_j, iterated over all k points in the i^{th} cluster for all n clusters. To put it briefly, the algorithm works by first assigning centroids randomly and calculating the distance of each data point to its nearest centroid using Euclidean distance, then finding the new value of centroids by calculating the mean distance of all points belonging to the centroid. Through an iterative process each data points are grouped into a 'cluster' with the minimum distance to the centroid of the cluster and maximum inter-cluster distance. Fuzzy c-means is also used in FTIR spectrocopic study, for instance, on the quantitative estimation of collagen and proteoglycan contents of articular cartilage of rabbits (Kobrina et al. 2012). The use of clustering algorithms has been shown to dramatically increase the information content of FTIR spectroscopic images of colorectal adenocarcinoma tissues as compared to univariate methods of FTIR spectroscopic imaging, i.e. the chemical map obtained from the integrated absorbance of a functional group. It has also been demonstrated that among the cluster imaging methods, HCA proved to be the best in terms of tissue structure differentiation, but HCA is significantly more time-consuming. Therefore, it is justified that HCA cluster analysis for routine analysis of FTIR spectroscopic imaging data recorded by FPA detectors is impracticable (Lasch, Mahadevanansen & Diem 2004). Instead, the task of classifying large datasets can be taken on by supervised classifiers.

Classification methods

PLS-DA is one of the most popular and effective analytical tool that can be utilised on processing a singular covariance matrix where the predictor matrix X (I × J) and a specially constructed response matrix Y (I × K) comprises categorical or 'dummy' variable describing the class membership are fed into the algorithm (Stahle & Wold 1987). PLS-DA will output response values Y predicted for unknown new samples. In the classification methods, the result is a confusion matrix that allows the visualisation of actual and predicted classification. In the past, the results of classification are often presented as type I error, which is the incorrect rejection of a class membership and type II error, which is the wrong acceptance of an object as a member of a class. They are now termed 'sensitivity' and 'specificity' respectively. In attempt to increase both sensitivity and specificity, a large number of samples must be used (Berrueta et al. 2007).

Other techniques that can be used for classification purposes are ANN and Random Forests (RF). The former is a network of connected neurons grouped in layers whereas the latter is an ensemble of decision trees. Since RF is easier to be implemented (lower computational demand and no need for data scaling), RF is successfully applied to a wide variety of high dimensional data, arising from microarrays, time series, and even on spectra. To explain briefly, RF construct many individual decision trees at training and predictions from all trees are pooled to make the final decision. The architecture of a RF is briefly illustrated in the schematic below (Fig. 3.15.).

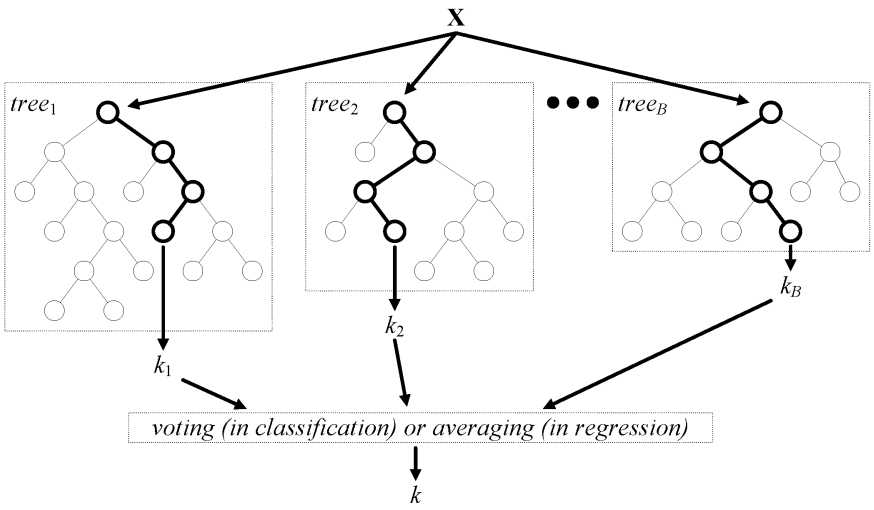


Figure 3.15: The architecture of a RF classifier (Verikas et al. 2016).

As a classifier, RF performs an implicit feature selection, using only a small subset or 'strong variables' for the classification. The outcome of this feature selection can be visualised by the 'Gini importance', which can also be used as a general indicator of feature relevance. This feature importance score provides a relative ranking of the spectral features, and is technically – a by-product in the training of the random forest classifier:

At each node (τ) within the binary trees (T) of the random forest, the optimal split is sought using the Gini impurity ($i(\tau)$) – a computationally efficient approximation to the entropy – measuring how well a potential split is separating the samples of the two classes in this particular node (Menze et al. 2009, Nembrini et al. 2018):

$$i(\tau) = \sum_{j=1}^{J} \widehat{\phi}_j(t)(1 - \widehat{\phi}_j(t)) \tag{3.15}$$

where $\widehat{\phi}_j(t))$ is the class frequency for class j in the node τ. The higher the value of Gini, the more important the feature. By default, Gini impurity is used to represent the feature important for classification with RF.

Chapter 4

Results and Discussion

4.1 Overview

This chapter is divided into several main investigations with the aims to provide insights to the challenges and literature gaps that were discussed in Section 2.5. The motivation behind each investigation is discussed here.

The first section – Section 4.2 presents and discusses the spectroscopic measurements of prostate tissue using dispersive spectrometer in transmission mode [1]. The possibility of identifying the biochemical differences between cancer and benign areas within these prostate tissues based on their spectra were analysed. This is important as despite the popularity of FTIR spectroscopy, dispersive instruments are still being used in some research labs. If the benign and cancer tissues can be successfully differentiated and the potential biomarkers are identified, it would mean that there is a great potential that a new system which relies on only several discrete wavelengths can be developed for the purpose of tissue classification. Such a system is not only cost saving, but also has a high efficiency as the acquisition time would significantly reduce. Apart from that, this experiment was novel as the dispersive IR spectroscopic imaging system was combined with thermography to study the potential of thermographic imaging on the tissues without having to analyse the IR spectra. In other words, the computational effort required to process the spectra can be eliminated as the thermal images alone are sufficient. That said, dispersive spectroscopy has some limitations including the low S/N ratio compared to, discussed in Section 3.1.2. Therefore, the other studies were carried out with FTIR spectroscopy.

The second section – Section 4.3 presents the results of FTIR spectroscopic imag-

[1] This work was a 3-weeks' research carried out in collaboration with Prof. Junko Morikawa at Tokyo Institute of Technology, Japan. The spectral data were saved and kept in the lab of Prof. Morikawa.

ing of colon biopsy tissues in transmission mode in combination with machine learning for the classification of different stages of colon malignancy. The availability of the samples was the main factor colon tissues are used in this study instead, and not the prostate tissue. There were four different grades of colon tissues available for classification which makes it interesting for classification with machine learning, but only a type of prostate tissue (see Section 3.2.2 for more details on the samples) was available for measurements. The processing method in this investigation was not new – i.e. it followed that outlined in previous literature described under Section 3 where the samples were first measured to obtain their FTIR spectra, followed by pre-processing methods, and analysis with both unsupervised and supervised machine learning techniques. However, interesting results and observations were obtained, making this study rather unique. Besides, this was the first time machine learning was used to compare the effect of two different approaches, an optical and a computational one, for the elimination of the *resonant* Mie scattering effect.

The third section – Section 4.4 discusses the effect of fluctuations of surrounding environment on the spectra, as a follow-up studies of the previous part mentioned in the paragraph above. As the presence of CO_2 and water vapour has been accounted for (Section 3.3.1), all that is left for further investigations are the temperature and humidity of the surrounding where the measurements were carried out. Unfortunately, dealing with biological tissues means that changing temperature is not an option as they will undergo degradation at high temperature; therefore, the focus was on humidity at room temperature. The effect of humidity on the hydration of tissue and also on the classification of tissue was investigated.

The fourth section – The previous experiments were all carried out in transmission mode, but the application of spectroscopic imaging on biological tissues could also be extended to measurements in ATR mode. With studies in different modes, their difference and effectiveness in biological applications could be compared. Section 4.5 investigated the possibility of examining embedded components within a prostate tissue specimen using ATR-FTIR spectroscopic imaging by depth profiling. This was achieved by changing the angle of incidence with the use of home designed apertures to cut off the light beam at different angle. It is known that ATR gives higher spatial resolution in the x- and y-direction (Section 3.1.5); but might result in impressions left on the tissues, thus prostate tissues were measured first before deciding if the application of depth profiling should be carried out on colon tissues.

The fifth section – ATR-FTIR imaging on biological tissues has been measured extensively, thus it is necessary to develop an improved method of ATR to make it more efficient. Here, section 4.6 discusses the idea of combining imaging and mapping together to get a bigger picture of the tissues measured. This was carried out by using a new 'large-area' Ge ATR crystal. This idea of implementing a large crystal is to potentially reduce the pressure applied on the soft colon tissues, thus making it possible for them

to be measured for classification purposes without leaving any impressions on them. It is important to make sure the research has not affected the colon tissue samples, for reproducibility reasons.

4.2 Thermo-dispersive IR spectroscopic measurements on prostate tissue

Thermal radiation, or thermal emission, is an EM radiation generated by all matters with a temperature greater than absolute zero (Landau & Lifshitz 2013). The IR radiation emitted is detectable with an infrared camera, where information related to the emitter's material properties (Brugel 1965) and temperature distribution (Christensen et al. 2001) can be obtained. The basis of thermal emission spectroscopy is that different types of compounds will take on different temperatures when exposed to the same amount of light. Thermal emission researches are broadly conducted at high temperatures (Dyachenko et al. 2016, Yeng et al. 2012), such that the signal dominates over the thermal background emitted by components of the measurement instrument, the Johnson-Nyquist noise in the detector, or the room that houses the experiment. Recently, there has been a growing interest in measuring thermal emitters at room temperature (Liu & Padilla 2017). However, the measurement can be challenging due to the large thermal background and the low S/N ratio compared to heated samples (Xiao et al. 2019). The near-room temperature emissivity measurement ensures that Kirchoff's Law, as originally conceived to be applicable to situations where the sample is isothermal and is in thermal equilibrium as the background to which it radiates, can be applied (Salisbury et al. 1993). The Kirchoff's Law states that at thermal equilibrium, the power radiated by an object must be equal to the power absorbed (Riedl 2001). This is the definition of a blackbody. In reality, most radiation sources are not blackbody. The radiant emittance of a real object is less than that of a blackbody, as some of the energy is reflected or transmitted. The ratio of the two radiant emittance values is the 'emissivity', a value measured in a typical thermal emission experiment.

4.2.1 Set-up of the thermo-dispersive IR spectrometer

Prostate tissue samples were prepared, as described in Section 3.2. The instrument was housed in the laboratory of Prof. J. Morikawa (Tokyo Institute of Technology, Japan). The components and optical path of the instrument are as follows. IR light is emitted by a ceramic heater, which passes through a mechanical chopper. The chopper can be switched between 'open' and 'close' position, allowing the alternate acquisition of IR and thermal images. In "open" position, light beam is allowed to reach the sample; while it is cut off in 'close' position. The period between these two measurements was set at 5 ms. To prevent a high order of diffraction, the light beam is passed through an order sorting filter, after which a grating monochromator (CT-10; JASCO Corporation, Tokyo, Japan) is in place to produce monochromatic light of various discrete wavelengths from 3 to 5 μm (3333 –2000 cm^{-1}) at a spectral resolution of 2 nm or 1.33 cm^{-1} (1000 data points).

The frequency of the chopper is controlled by a DC power supply to match that of the grating motion. The emitted beam is then condensed with a parabolic mirror and focused onto the sample in the transmission mode. The parabolic mirror has a 3× magnification. The objective lens of IR camera is designed considering the wavelength of the spectral band and the angle of the view of the sensor, with a numerical aperture at 0.7. The anti-reflective coating on the achromatic lens is composed of materials suited for mid-IR. The transmitted light is recorded with an IR high-performance camera (Phoenix, Indigo Systems Corporation, California, USA). It consists of a closed-cycle Stirling cryogenic cooled indium antimonide (InSb)-based FPA detector [2], which operates at a temperature of 77 K; 128 × 128 pixels are used to compose the images. The pixel size is 30 × 30 μm^2, giving a total field of view of 512 × 512 μm^2 under 7.5× magnifications. The input trigger of the camera is highly synchronized with the signal to the chopper to achieve high simultaneity in measurement with the in-house written LabView software (Tokyo Institute of Technology, Tokyo, Japan). A schematic of the optical setup is shown in Fig. 4.1.

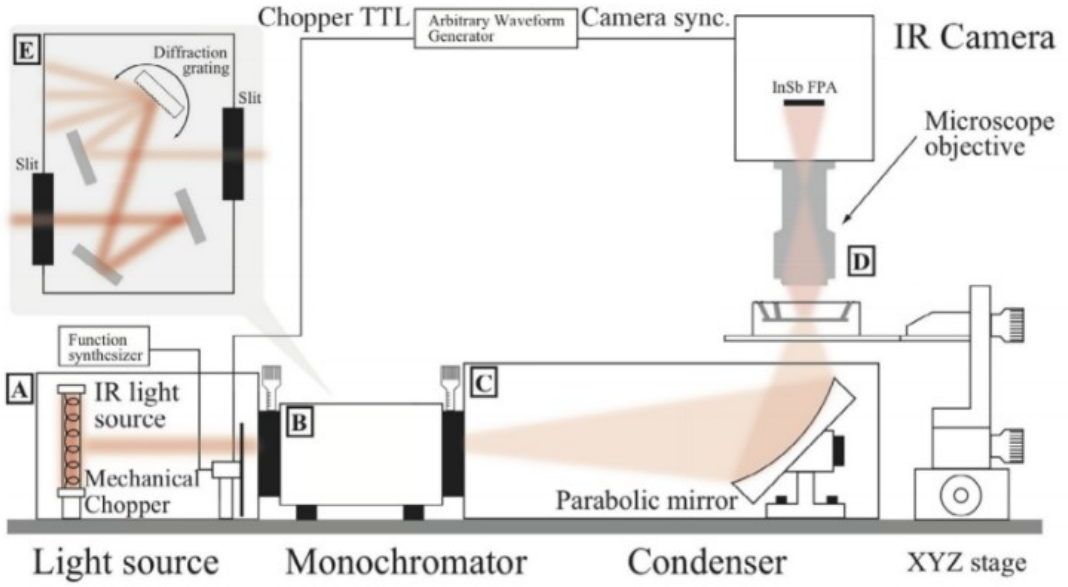


Figure 4.1: Set-up of the dispersive IR spectrometer, adapted from (Ryu et al. 2017) with the permission from Elsevier

4.2.2 Data processing

The data obtained from the system (S) involve contribution from both transmitted IR (I) and thermal emission expressed in the digital level of the IR camera (E), which is related to thermal energy or temperature after calibration with a reference. Eq. 4.1 relates S, I, and E, where x and y are the spatial coordinates (m), λ is the wavelength (m), and T is the absolute temperature (K). All data are multidimensional arrays (hypercubes).

[2] Photovoltaic detector of p-n junctions that delivers high sensitivity in the atmospheric window between 3 – 5 μm.

$$S(x, y, \lambda, T) = I(x, y, \lambda, T) + E(x, y) \tag{4.1}$$

E is further defined as

$$E(x, y) = \varepsilon_\lambda \sigma a (T^4 - {T_0}^4) \tag{4.2}$$

where ε_λ is the thermal emissivity (< 1 for a grey body); σ is the Stefan-Boltzmann constant ($\mathrm{Wm^{-2}K^{-4}}$); and a is the area of measurement ($\mathrm{m^2}$). Rearranging Eq. 4.1, IR only (I) image is obtained by subtracting E image from S image. The sample spectra were divided against background scans which were obtained from a region of no tissue sample. Absorbance as a function of wavenumber of the incident IR light was then calculated from the logarithmic transmittance ratio. Chemical images made up of a total of 16,384 (128×128) pixels were constructed from the integrated absorbance of the spectral band of 2940 to 2900 $\mathrm{cm^{-1}}$. Further spectral data processing was performed using MATLAB (Mathworks Inc., Natick, Massachusetts).

The IR imaging data were subjected to a spectral quality test to eliminate spectra from areas with poor S/N ratio or without any tissue. The signal is the peak absorbance at 2921 $\mathrm{cm^{-1}}$, while the noise is the SD of data between 2810 and 2760 $\mathrm{cm^{-1}}$. Spectra which pass the quality test set at a threshold of 5 % of the maximum S/N ratio were then differentiated twice to deconvolute overlapping bands for comparison of the spectra between cancer and healthy areas of the tissue. SG smoothing algorithm with 13 smoothing points was implemented on the second derivatives. To determine the natural grouping of the spectra, the data were subjected to successional k-means clustering. Clusters of maximum inter-cluster variance but minimum intra-cluster variance were formed from this unsupervised clustering technique. The obtained pseudo-colour cluster images were compared directly with the IR and H&E stained images taken from the same sample. Randomly chosen data from each cluster were then compared with PCA, whereby the first PC explains most of the data variance, followed by second independent PC accounting for most residual variance and so on. A plot of the projection of the spectra on the PCs, represented by scores, on a 2-dimensional space was used to validate the difference between each cluster. Reproducibility of the results was attained by repeating the experiment on another biopsy tissue of the same kind and same stage of malignancy from the same patient. The two samples used were denoted as 'Sample 1' and 'Sample 2'.

4.2.3 Thermal effect on IR image

In this study, the micro-scale areas of cancer cells and benign stroma in the biopsy tissues of prostate were identified from IR spectroscopic and thermal images, by comparing them to histopathological H&E stained images of adjacent sections, described in Section 3.2.3. Fig. 4.2 shows the H&E stained prostate cancer tissues of two different samples (Sample 1 and Sample 2). Both tissues are assigned a Gleason score of 3 + 3, the least aggressive tumour that can be identified on prostate biopsy. In H&E staining, nuclei are stained purple due to its basophilic nature, while the cytoplasm is stained pink (Chan 2014). The areas where a high concentration of nuclei are found were identified to be cancerous by pathologists.

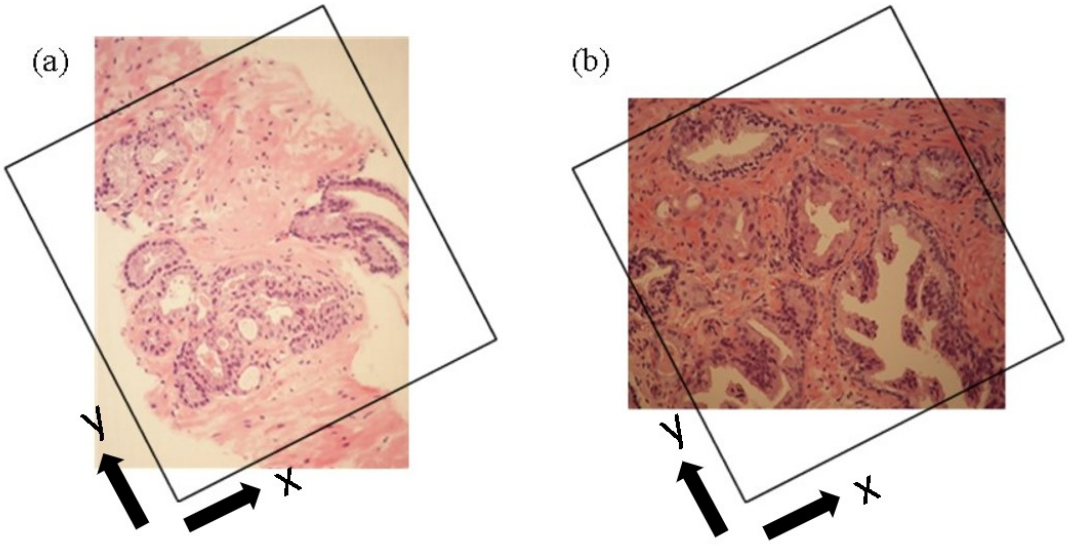


Figure 4.2: Photomicrographs of the H&E stained prostate tissue sections under 20× magnifications: (a) Sample 1 and (b) Sample 2. The squares represent the IR spectroscopic sampling areas (512 × 512 μm^2) with the orientation of the tissues when measurements were taken in our experiment

The chemical images displaying the integrated absorbance of the lipids or fatty acids band (2940 – 2900 cm^{-1}), attributed to the anti-symmetric stretching mode (ν_{as}) of the CH_2 group, are shown in Fig. 4.3. A comparison of the IR images before and after correction of the thermal effect is also depicted.

From Fig. 4.3, the images are different before and after removal of thermal contribution, particularly within the sections outlined in red boxes. A larger area of low concentration of fatty acids in the tissue (pixels with dark blue colour) is seen on images on the left. When the IR chemical images, after subtraction of thermal effect, were compared with Fig. 4.2, the coverage of the dark blue areas coincides with the regions identified as cancer. On the contrary, when comparing images before thermal subtraction to H&E stain, the same regions are unable to be classified from the surrounding areas due to the

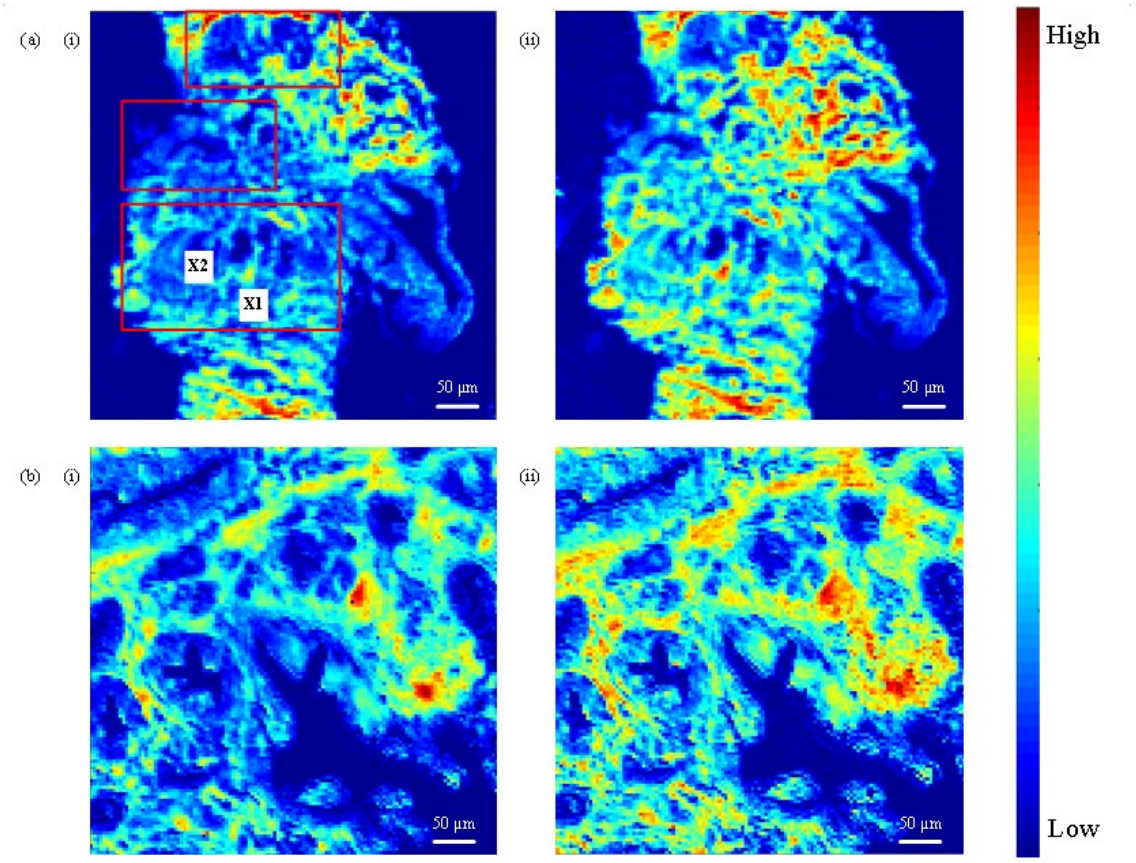


Figure 4.3: IR images of (a) Sample 1 and (b) Sample 2, showing the distribution of the integrated absorbance of the stretching of the C-H band under 7.5× magnifications (i) after removal of thermal effect and (ii) before removal of thermal effect. Areas in the boxes of Sample 1 (numbered as box 1 to 3 from top to bottom respectively) were further examined under 10× magnifications and analysed through unsupervised clustering technique to identify different regions of the tissue. The colorbar on the right indicates the colour scale from low to high absorbance

lower contrast between the cancerous and benign cells from thermal noise. The observation was quantified by examining the mean spectra of several pixels of low intensity (marked as 'X2' in Fig. 4.3); and neighbouring pixels of high intensity ('X1'). The spectra are presented in Fig. 4.4 below. The average ratio of the corresponding integrated absorbance of A1/A2 at 2940 – 2900 cm^{-1} for these two areas is 4.5; whilst the value is lower for A3/A4 (for spectra measured before removal of thermal effect) for the same two areas at 2.0, thereby confirming that thermal effect contributes to the reduced contrast of the image in identifying different regions within a tissue, however, it is also noted that the S/N ratio is higher for the spectra obtained from the same pixels prior to subtraction of thermal radiation, therefore correction for thermal contribution does not help to improve spectral quality. Other noise reduction methods, such as increasing the number of co-added scans for any single spectrum, need to be implemented to improve the S/N ratio.

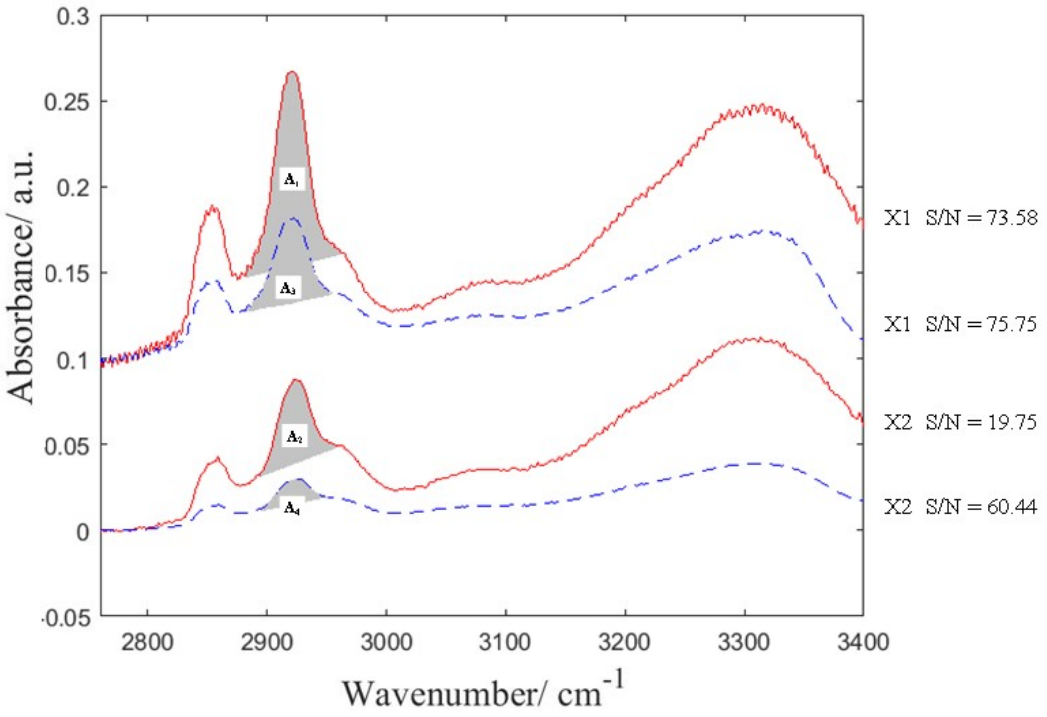


Figure 4.4: The average spectra of areas from 9 pixels marked with 'X1' and 'X2' in Fig. 4.3, representing areas of high and low integrated absorbance of the spectral band assigned to ν_{as} of CH_2 respectively. The spectra before subtraction of thermal contribution are shown in dotted blue lines while the spectra after correction are shown in solid red lines

4.2.4 Distribution of thermal signal intensity in tissue samples

Thermal DL signal is dependent on temperature and emissivity of the biopsy tissues. Emissivity of an object, in turn, is correlated to its absorption of incident radiation, also known as its absorptivity according to Kirchoff's law of thermal radiation (Planck 2013). The value of emissivity is unique to every object. In our research, the integrated thermal and IR spectroscopic measurements allow the absorption to be estimated from the integrated absorbance of the IR spectra at the wavelength specific to the sample. Images showing the distribution of thermal signal intensity taken on the prostate biopsy samples are shown in Fig. 4.5.

Fig. 4.5 shows the distribution of thermal DL signal, which was calculated by first obtaining the average value of all combined pixels of the background, which refers to CaF_2 window only without the presence of tissue. The value of each single pixel is then corrected for the non-uniformity in thermal image. Areas of high absorbance at 2940 – 2900 cm^{-1} (benign stroma in the tissue) also shows high thermal signal as it has higher emissivity compared to cancer cells. The thermal DL signal can be translated to the temperature of a certain area of the sample if the relationships between concentration, temperature, and infrared emisssion are determined beforehand. This requires calibration for concentration

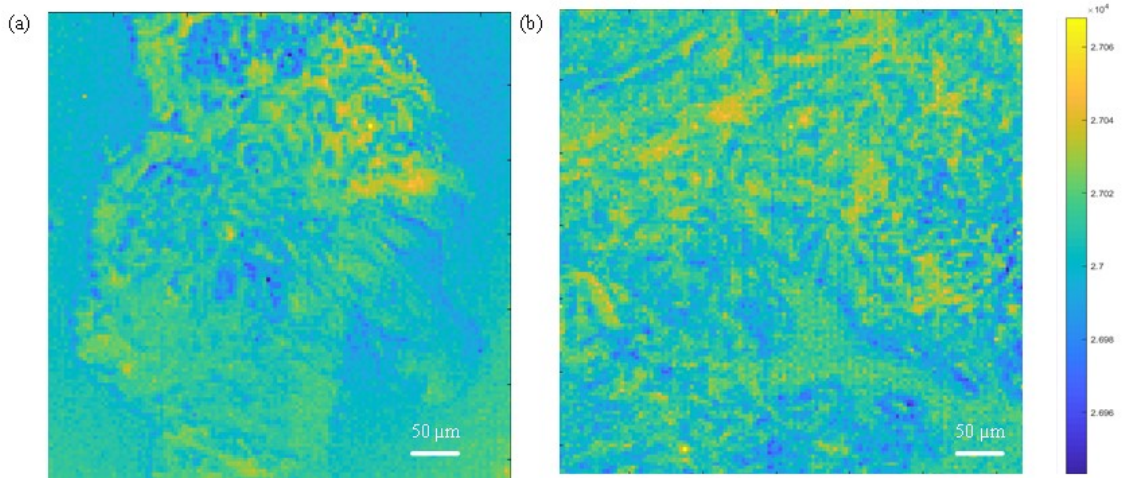


Figure 4.5: Thermal images of (a) Sample 1 and (b) Sample 2 obtained for the same area of prostate tissues as IR spectroscopic images. Cancer is shown in dark shade of blue, while stroma is highlighted in yellow. The colour bar on the right indicates the thermal signal intensity

versus DL signal and temperature versus DL signal on the sample examined, as discussed by (Ryu et al. 2017). Nonetheless, subjecting biological specimens to heating can result in irreversible cellular and structural changes (Rossmanna & Haemmerich 2014), thus the calibration is avoided in this case. Instead, the temperature distribution of sample of the same region with similar concentration of fatty acids was interpreted, which translates to constant emissivity, thereby limiting the thermal signal to a function of temperature only. A uniform DL signal is obtained within areas of tissues with same emissivity. Based on the results, it can be inferred that the thermal variation of studied samples mainly comes from differences in IR absoprtion of tissue structures., thus fortify the findings that thermal radiation has a substantial effect on dispersive IR imaging measurements of a heterogeneous samples of various emissivities.

It is also noted that thermal signal is the lowest at the interface between tissue structures, i.e. at the edge between tissue and non-tissue and also between cancer and benign areas. The increased diffuse thermal reflectance at these interfaces could be the reason behind this phenomenon although the real reason is unclear. Unlike absorptivity which positively correlates to emissivity, reflectance has an opposite effect on thermal emission spectra according to the formula $\varepsilon = 1 - R - \tau$, where ε is emissivity, τ is transmissivity, and R is reflectivity of an object. As depicted in Fig. 4.5, the 'thermal sink' effectively outlines various regions of interest (The 'thermal sink' here refers to area of low thermal signal) and can be explored for future inspection of cancerous areas.

The main visual difference between the IR spectroscopic and thermal images is the diminished contrast of the latter. This is due to the heterogeneity in tissue components that introduces multiple degrees of reflectance, which effectively suppresses the contrasts

in the image. The same phenomenon is observed by (Byrnes et al. 2007) when studying thermal emission of glass.

4.2.5 Analysis of IR spectroscopic imaging data

Sample 1

Unsupervised k-means clustering was applied to sort the data that exhibit similar features. It uses an iterative algorithm to randomly assigned a center for the clusters until a distinct boundary between different clusters is obtained. Each spectrum is assigned a class membership through this method. The number of clusters set in this experiment was 4 to match the regions wanted to be identified (prostate glands, cancer and its surrounding tissues, and benign stroma), with the maximum number of iterations set at 100. The process continued until maximum inter-cluster but minimum intra-cluster 'euclidean' distances were achieved. Fig. 4.6 depicts the pseudo-colour images assembled in 2D spatial coordinates where each cluster is represented by a colour. The clusters clearly defined the areas of cancer and surrounding cells in the tissue, and when compared to H&E images side-by-side for the assessment of regions, they correspond to each other. Each region is labelled 1 – 4 for comparison of the spectral information.

Spectral data in each cluster were extracted, and the mean and their corresponding smoothed second derivatives with SG algorithm are presented in Fig. 4.7. Comparison with second derivative is more precise as the baseline shift is eliminated. The peaks of the second derivatives are assigned to the vibrational modes of the functional groups: symmetric stretching of methylene group, ν_s CH_2 ($2852\,cm^{-1}$); anti-symmetric stretching of methyl group, ν_s CH_3 ($2885\,cm^{-1}$); anti-symmetric stretching of methylene group, ν_{as} CH_2 ($2921\,cm^{-1}$); and anti-symmetric stretching of methyl group, ν_{as} CH_3 ($2967\,cm^{-1}$). The phospholipids, lipids, triglycerides, and proteins absorb within the studied IR region of $3000 - 2800\,cm^{-1}$ (Bogomolny et al. 2007).

From Fig. 4.7(a), prostate glands, cancer, and benign stroma show an increasing order of absorbance at the $2852\,cm^{-1}$ and $2921\,cm^{-1}$ bands. This observation was supported by the hypothesis from (Liu 2006) that fatty acid oxidation becomes the dominant bioenergetics pathway in prostate cancer, in contrast to the common glycolysis pathway. An increase in the utilization of fatty acids as an important energy source to provide adenosine triphosphate (ATP) is crucial to sustain the energy requirement for the enhanced proliferation of cancer cells and their high metabolism. Their finding consolidates the observation in Fig. 4.3, such that cancerous lesions have a lower concentration of lipid compared to surrounding benign tissues. Differentiation of the cancerous regions employing the fatty acid spectral bands has already been investigated for breast cancer (Baker et al. 2008), but employing region between $3000 - 2800\,cm^{-1}$ only to identify cancerous le-

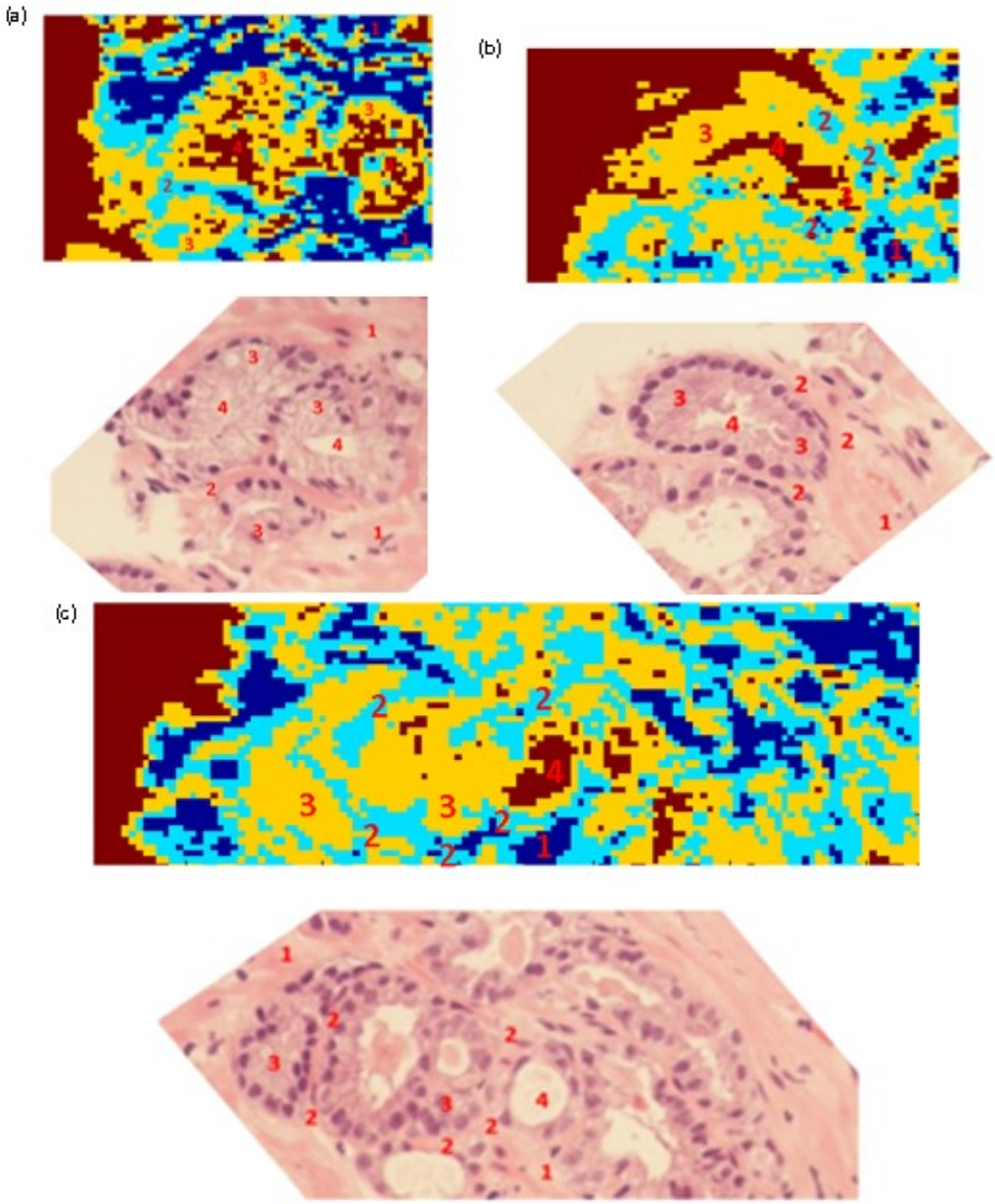


Figure 4.6: Top: Pseudo-colour cluster image of areas of cancer lesions of (a) Box 1, (b) Box 2, and (c) Box 3 of Patient 1 respectively, following the convention used in Fig. 4.3(a)(i), constructed from the second derivative spectra. Each cluster is assigned to a colour (Brown = region 4; Light blue = region 2; Dark blue = region 1; and Yellow = region 3). Bottom: H&E stained image to differentiate the cluster assigned: region 1 = benign stroma at a distance from the cancer; region 2 = stroma in between cancer; region 3 = cancer; and region 4 = prostate glands. The H&E stain images were cut out from the boxes in Fig. 4.2 and re-oriented for easier comparison, resulting in the irregularities in their shapes

sions in biopsy, without utilizing the spectral bands of phosphates and carbohydrate from

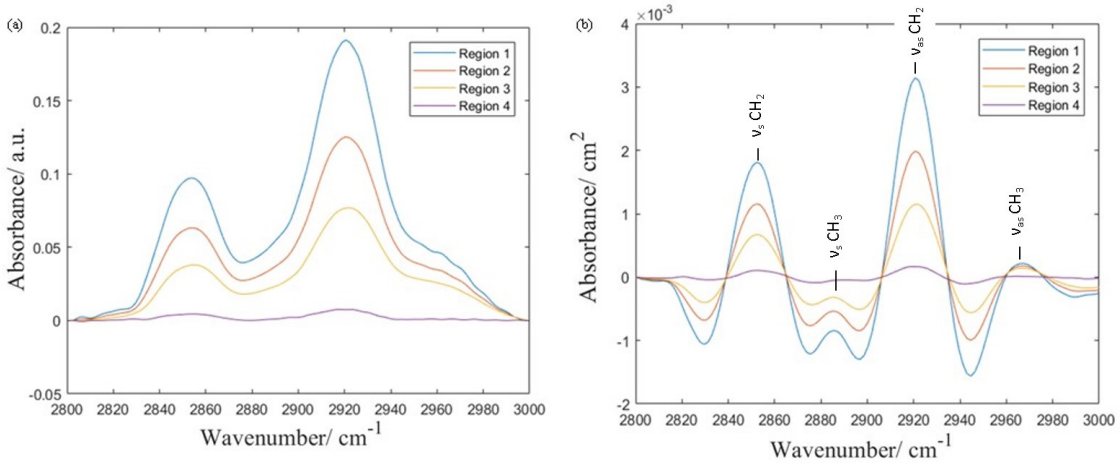


Figure 4.7: (a) Average spectra of all the spectra coming from each of the 4 clusters identified from unsupervised clustering in the range of 3000 – 2800 cm^{-1} of Sample 1. (b) Second derivatives of the mean spectra of each cluster. Two more unresolved peaks at 2885 cm^{-1} and 2967 cm^{-1} were obtained from the second derivatives

IR spectroscopic imaging is uncommon. As can be seen in Fig. 4.7(b), the number of overlapping peaks in the spectral region of stretching C-H vibrational modes of fatty acids is fewer than in the fingerprint region of 1800 – 900 cm^{-1}. Therefore, direct comparison of the results without having to resolve the convoluted peaks can be made, saving the computational power required to analyse the data.

The ratios were calculated for the integrated absorbance of both symmetric and anti-symmetric stretching modes of CH_2 functional group. The average ratio of absorbance of the corresponding bands of CH_2 group (ν_s CH_2: ν_{as} CH_2) is found to decrease for tissue areas closer to cancerous region at 0.57, 0.36, and 0.21 for regions 1, 2, and 3 respectively. The peak absorbance were calculated by taking the integrated absorbance of the second derivative spectra between the nearest minimum points of the specific spectral peak. The ratios of these bands qualitatively monitor the conformational change of chain and packing of fatty acids (Mendelsohn et al. 2006). This ratio is useful as an indicator of cancer cells. The values have been shown to increase for malignant tissues of oesophagus, breast, and skin; whilst a decrease in the ratio has been recorded for leukaemia and invasive cervical cancer cells (van den Driesche et al. 2011). In this study, a decrease in the CH_2 ratio is seen closer to the cancerous section of the prostate tissues. The reduced CH_2 ratio in prostate cancer could be due to an increase in cholesterol in these cells, as suggested by (van den Driesche et al. 2011). This is consistent with the fact that cholesterol, which is already present in large quantity in normal prostate tissues, further increases during progression to prostate cancer (Krycer & Brown 2013). The absorbance at 2852 cm^{-1} (ν_s CH_2) is assigned to lipid or fatty acid, thus a change in CH_2 stretch ratio can be associated with changes in lipoprotein of membrane that governs the cell permeability to transport metabolites for rapid cell growth. Apart from that, ν_{as} CH_2: ν_{as} CH_3 absorption ratio

provides information on the composition of fatty acids. It is found to be higher in benign stroma than cancer at 31.1 and 11.3 respectively, thus it is postulated that lipids of shorter chain could be found in prostate cancer. The postulation is supported by a study on colon where short-chain fatty acids have been reported to induce cell migration in colon polyps that leads to abnormal crypt growth (Sahu et al. 2005). In addition, a reduction in the cytoplasmic to nucleic size in cancer could possibly contribute to the ratio change. No frequency shift of the CH_2 and CH_3 spectral bands is observed in measured spectra.

PCA score plots (Fig. 4.8) were used to explain the difference in spectra of the clusters formed from unsupervised clustering. Twenty randomly chosen data points from each cluster were used to generate the figure. Two PCs that make up 98.71 % of the total variance are identified, namely PC1 (95.04 %) and PC2 (3.67 %). In these plots, the scores are oriented primarily in the direction of PC1 (between cluster 2 and 3, and 3 and 4), followed by PC2. PC1 corresponds to variation along ν_{as} CH_2 whereas PC2 corresponds to ν_s CH_2. The other wavelengths have little relevance in understanding the total variability of the system. As the data points from region 2 and 3, as well as region 3 and 4, do not overlap with one another in the PCA score plots, it is inferred that cancer, represented by region 3 can significantly be distinguished from the surrounding tissues and prostate glands by examination of the CH_2 vibrational modes. The results were analysed with student's t-test and showed that there is a significant difference ($P < 0.001$) among them. This is not the case for region 1 and 2 since the data points slightly overlap with each other. This is understandable as region 1 and 2 both stand for stroma, thus they share similarity in their chemical composition with the only difference between them is the distance from cancer, resulting in only a slight disparity between the concentration of lipid.

Sample 2

The mean integrated absorbance at $2967\,\mathrm{cm}^{-1}$ is found to be ~ 0.40 for region 4, ~ 1.50 for region 3, ~ 2.85 for region 4, and ~ 3.75 for region 1. The values for absorbance were used as cut-off threshold on an independent set of spectra obtained from Sample 2, another tissue of the same type, degree of malignancy, and thickness as Sample 1, presented in Table 4.1. Fig. 4.9 shows that the threshold introduced are validated as the areas of the regions can be defined from false colour image after implementing the lower and upper limits for classification of the region. It is shown in later study, however, that machine learning could be more useful at data classification when a larger range of spectral data is involved.

Apart from that, the time taken for measurement within a short range of wavenumber is greatly reduced. In our experiment the acquisition time was reduced to within 5 minutes, significantly more efficient compared to FTIR spectroscopic imaging where mea-

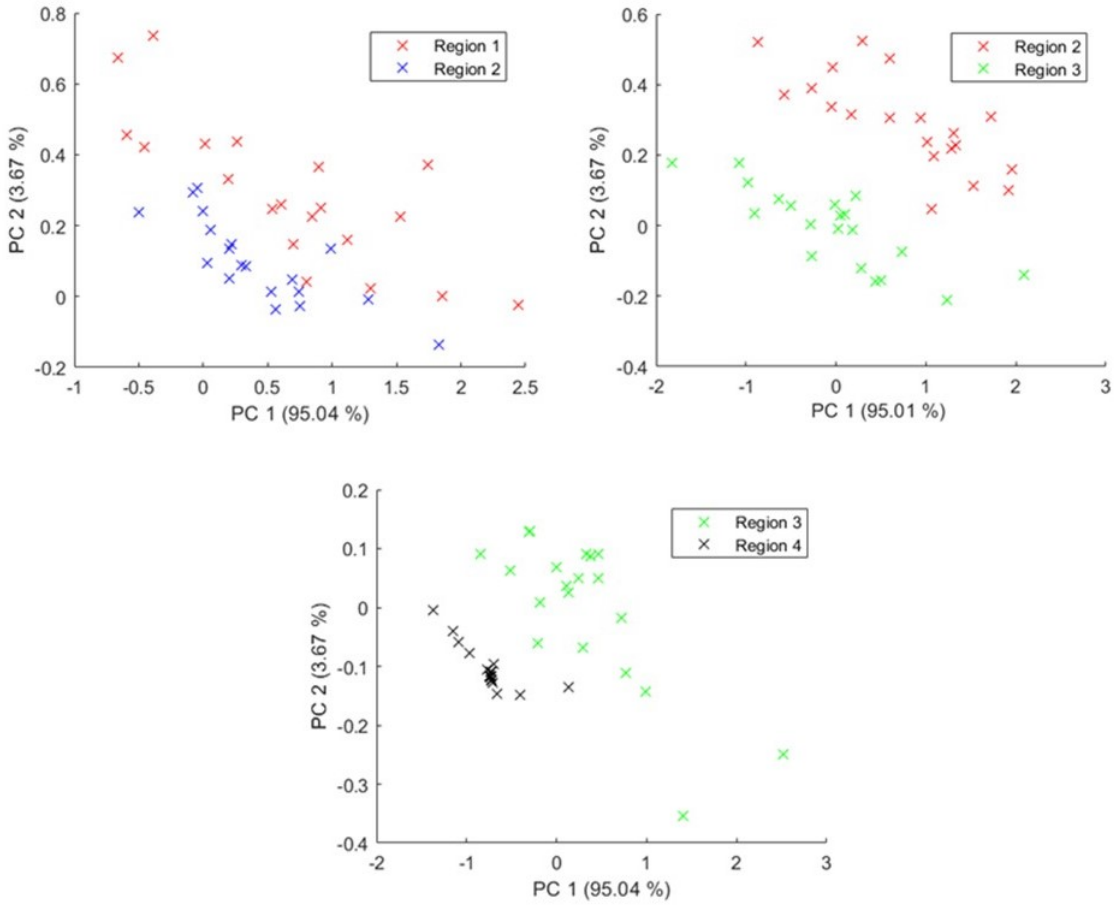


Figure 4.8: Score plot of randomly selected data from the clusters along the first two principal components (PCs)

surements on a similar area of measurement can take up 2 hours for spectral ranging from 3900 – 900 cm^{-1} at spectral resolution of 4 cm^{-1} with 512 co-added scans (Bruker Tensor 27 with Hyperion 3000 Microscope, Bruker Optics, Ettlingen, Germany). The image of Sample 2 obtained from FTIR spectroscopic measurements under 15× magnifications and dispersive IR spectroscopic imaging instrument under 10× magnification is compared in Fig. 4.10 which shows that image on the left is more pixelated. The image resolution that is achieved in FTIR spectroscopic imaging is higher due to a smaller projected pixel size [3] of 2.7 μm, than the dispersive IR instrument used in our study at 3.0 μm.

[3]The projected pixel size is not the same as actual spatial resolution, which can be estimated by measuring the IR absorbance across a sharp interface of two substances of similar refractive indices, as demonstrated in the following sections.

Table 4.1: The lower and upper limits of each clusters used for validation of the workflow and results on the other independent spectral data obtained from Sample 1

Assigned cluster	Corresponding area	Integrated absorbance
Region 1	Benign stroma	<0.8
Region 2	Cancer surrounding benign stroma	>0.8 and <2.2
Region 3	Cancer	>2.2 and <3.5
Region 4	Prostate gland	>3.5

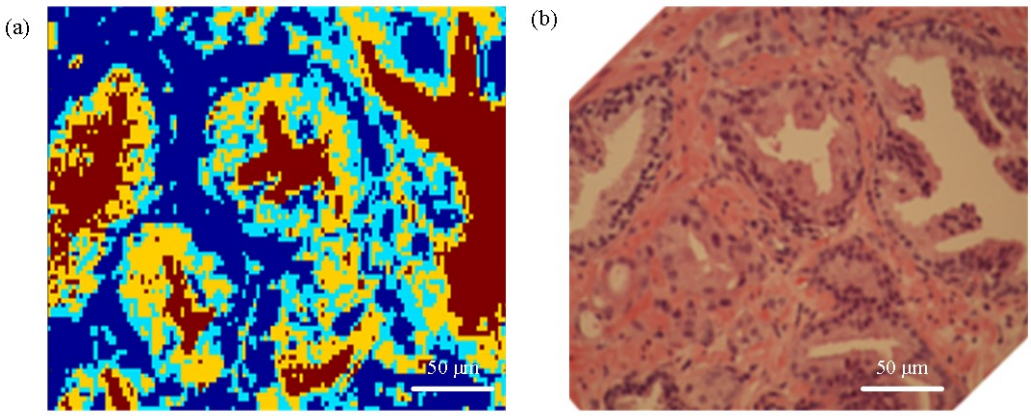


Figure 4.9: (a) False colour image generated from the cut-off thresholds set from Sample 1 to validate the reproducibility of the results on Sample 2 and (b) the corresponding H&E image of the area under focus

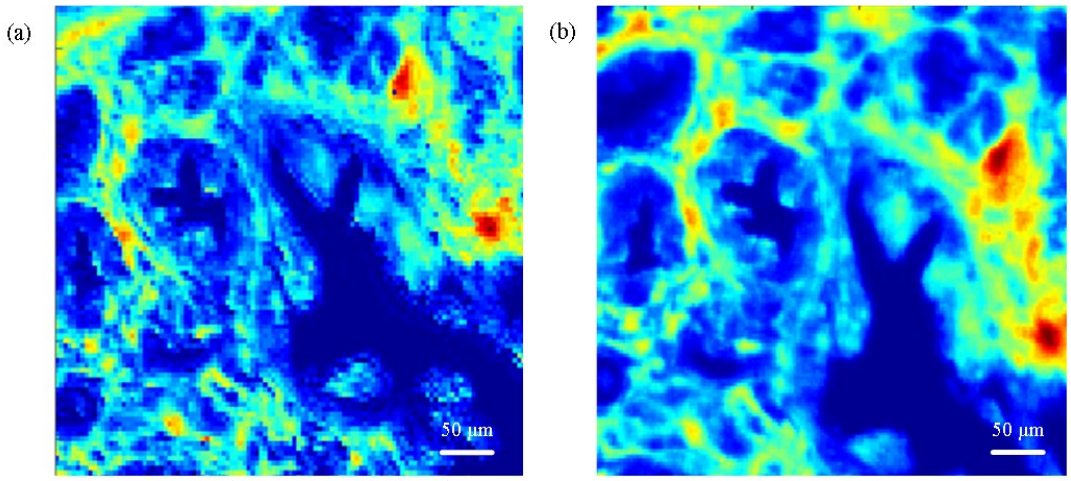


Figure 4.10: IR images (470 × 470 μm^2) showing distribution of integrated absorbance of peak at 2920 cm^{-1} of Sample 2 taken with (a) dispersive IR and (b) FTIR spectroscopic imaging instrument

4.2.6 Summary

A system that combines dispersive IR micro-spectroscopic imaging and thermography has been developed to study the effect of thermal radiation on the IR absorption spectra of prostate biopsy samples. The system allows the distribution of thermal signal intensity as a function of emissivity to be interpreted from the integrated absorbance obtained by spectroscopic imaging. Biochemical differences between cancer and benign areas within the specimens were identified in the spectra. Side-by-side comparison of H&E stained adjacent tissue sections with infrared images constructed before and after removal of thermal effect showed that the latter strongly support differentiation of regions within tissues. A systemic methodology was implemented to process the data, firstly by k-means clustering on the second derivative spectra, followed by PCA analysis. Four distinct regions within the tissue samples were successfully classified based on the anti-symmetric stretching mode of methylene functional group. Separation between data in clusters occured when projecting spectra on a PCA score plot on a plane made by first two PCs. The significance of the disparity was verified with statistical test.

4.3 FTIR imaging of colon tissue in transmission mode

FTIR spectra contain a wealth of information about the sample. As such, in analysis of spectra of biological systems, multivariate statistics and machine learning algorithms are frequently applied to extract the important information. The two main strategies in chemometrics used to analyse FTIR spectral data are unsupervised learning and supervised learning. The variety of the methods are detailed by (Goodacre 2003). The aim of this study is to utilise the well established machine learning approach, random forest (RF) in this case, to examine the FTIR spectral ranges obtained in transmission mode that contain the most important spectral biomarkers to distinguish between colon specimens of various degree of malignancy.

4.3.1 Experimental set-up

Colon tissue samples were prepared, as described in Section 3.2. The experiments were carried out in transmission mode at 15× magnification (NA = 0.4), with a Hyperion 3000 FTIR microscope coupled to Tensor 27 FTIR spectrometer (Bruker Corp.). A liquid nitrogen cooled 64 × 64-pixel FPA, which has FOV of 170 × 170 μm^2, was used for simultaneous acquisition of FTIR spectral dataset. As imaging was combined with mapping, 3 × 3 individual images were stitched into one, resulting in a total measured area of 510 × 510 μm^2 for each tissue. The spectral images from 8 sample areas were acquired (2 control/healthy and 6 diseased sample areas). A new background was recorded before measuring each individual image. All measurements were taken in the mid-IR range from 3900 cm^{-1} to 900 cm^{-1}, at 4 cm^{-1} spectral resolution and with 521 co-added scans. An additional CaF_2 lens, which has been shown to significantly reduce Mie scattering (Chan & Kazarian 2013), was also employed for imaging of the exact same tissue areas. The design and set-up of the lens for combining imaging and mapping were described in Fig. 4.11 by (Kimber et al. 2016). To put it briefly, the added lens is kept in focus with an external holder whilst the stage is shifted in x- and y-direction for different areas to be measured. The additional lens implemented to correct for the chromatic aberration in infrared measurement is referred to as 'correcting lens' (Kimber et al. 2016). The assessment of the performance of the correcting lens has never been examined with machine learning prior to the work discussed here.

4.3.2 Data processing framework for the classification of disease

The spectral data were processed with Matlab R2019b (The MathWorks, Inc.). The spectral data in the range of 1800 – 1000 cm^{-1} and 3000 – 2800 cm^{-1} were used for further analysis. The region between 2800 – 1800 cm^{-1} contains no important spectral

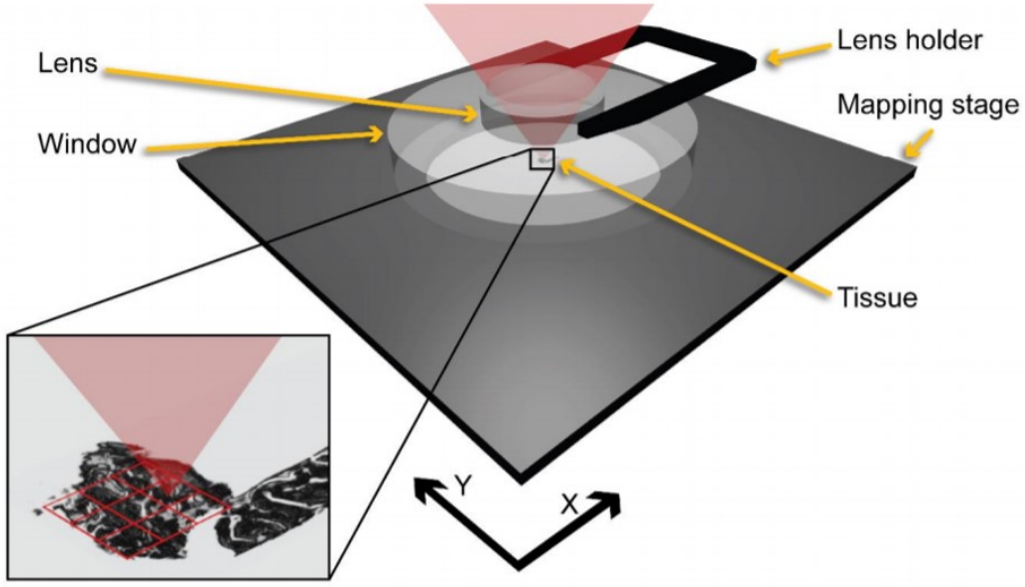


Figure 4.11: Illustration showing the set-up of the imaging and mapping approach using the correcting lens. The lens is fixed in line with the objective with an external lens holder while the substrate and sample are moved underneath to allow mapping to take place (shown here is a sample of Barrett's oesophageal adenocarcinoma). Reproduced from (Kimber et al. 2016) with the permission from Royal Society of Chemistry

information whilst region $> 3000\,\mathrm{cm}^{-1}$ is sensitive to water content within the tissues. Second derivatives of the obtained spectra were calculated with SG 9-point smoothing, which were then vector normalised. The spectra, second derivatives, and normalized second derivative data were then separately subjected to unsupervised machine learning, in this instance, the k-means clustering algorithm (tested for 2 to 6 clusters, each with 5 replicates and infinite iteration until the solution converges to a local minimum). A total of 2000 individual sample spectra were recorded and used for machine learning. Training and test models were created, each made up of 500 random spectra sampled from each cluster without replacement, for tissue at the same disease stage. In other words, the model consists of 2000 spectral data- 500 for healthy (H); 500 for hyperplastic polyps (HY); 500 for dysplastic polyps (D); and 500 for cancer sections (C), which are identified by H&E staining. The models were from different individuals ensuring that the inter-patient variability is included in the study. Employing machine learning to study imaging data has been demonstrated in previous works (Goodacre 2003, Berisha et al. 2019).

The training model, after undergoing data dimensionality reduction with PCA, was subsequently supplied to random forest (RF) classifier to generate a prediction model on the test model. RF operates by constructing multiple decision trees for classification on the data, gets prediction from each tree and thus outputs the class mode by means of voting. In this study, bootstrapping as well as a 5-fold cross validation of the dataset is implemented (Breiman 2001). Among various supervised machine learning classifiers,

Table 4.2: Parameters of RMies-EMSC algorithm used in Matlab to correct for Mie scattering effect

Number of iterations	10
Number of PCs used	8
Lower range for scattering particle diameter /μm	2
Upper range for scattering particle diameter /μm	8
Lower range for average refractive index	1.1
Upper range for average refractive index	1.5
Reference spectrum	Matrigel

RF is preferable since it is faster and insensitive to over-fitting (Chen et al. 2015) The prediction accuracy of the RF model is presented in the form of a confusion matrix. Inter-model predictability was carried out with independent training and test set. The size of training to test models were varied from 1:1 to 1:6. The analytical procedure was repeated with the measurement data obtained from the added correcting lens, as well as for no-lens data but corrected with RMies algorithm (provided by Peter Gardner's Lab, University of Manchester) (Bassan, Kohler, Martens, Lee, Byrne, Dumas, Gazi, Brown, Clarke & Gardner 2010, Kohler et al. 2008, Martens & Stark 1991). The parameters of the RMies algorithm are given in Table 4.2. Several machine learning parameters, namely the number of clusters, the spectral range for supervised and unsupervised classification, the size of training and test models, and the variance of retained PCA have been tried and tested, to optimize the prediction model. The important features are selected from the Gini index – a score of the feature importance that is derived from the training of the RF classifier, which technically correlates to the optimal 'Gini impurity' split at each nodes within the binary trees (Menze et al. 2009). Based on the selected features or spectral range, a flowsheet depicting all the different pathways to re-training machine leaning in categorizing the different stages of the colon cancer is shown in Fig. 4.12. The prediction accuracy of the test models was used as the criteria to cross-check the spectral range highlighted by the machine learning as the 'key biomarker' that can be utilized to understand different degree of malignancy of colon. Fig. 4.12 shows all the pathways that were tested with unsupervised and supervised approach via training and re-testing of the spectral data.

4.3.3 Physical and computational correction of Mie scattering effect

Mie scattering effect was significantly reduced at the edges of the tissues when correcting lens was added, shown in Fig. 4.13, where increase in absorbance of the amide I band and reduction in the sharp derivative-like distortion to the spectra at $\sim$1710 cm^{-1} were observed. Correction with the added lens did not significantly shift the peak position of

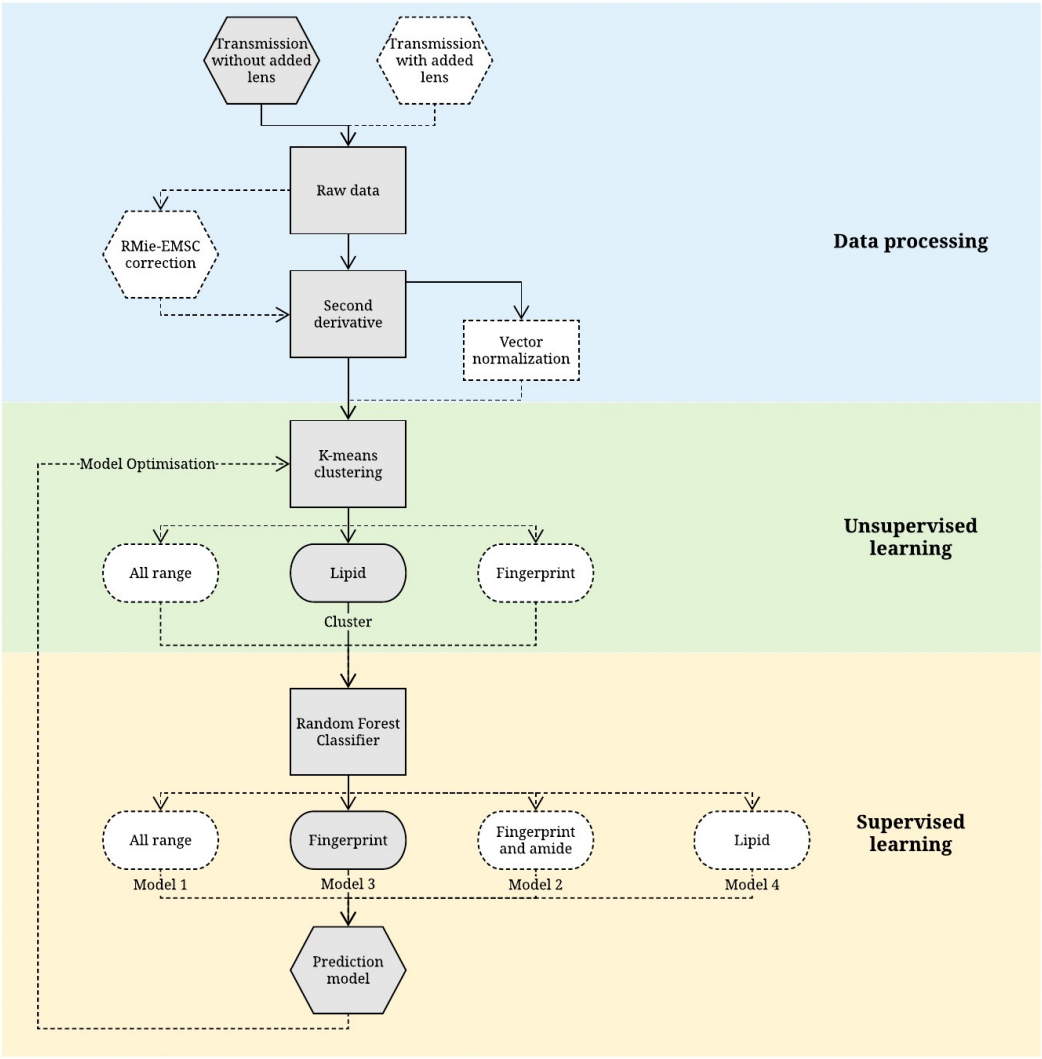


Figure 4.12: Schematic overview of the data processing and machine learning steps explored in this study. The best pathway leading to the optimised result is highlighted in grey

the amide I band (1652 cm^{-1}), most likely because the chromatic aberration was only partially corrected with a single lens (Fig. 4.15(a)). With the added lens, which acted like an immersion objective as reported by Kimber et al. (Kimber et al. 2016), the image has $\sim$ 40% increase in magnification (total area of 360 $\times$ 360 μm^2 compared to 510 $\times$ 510 μm^2 for image without lens) and the image is flipped due to the arrangement of the tissue during measurement whereby the tissue was placed facing downwards with the correcting lens added on top of it. Computation correction with RMies algorithm was more efficient at recovering a flat baseline of the spectra (Fig. 4.14) compared to correction with the added lens but was more time consuming and without the added benefit of increasing the magnification (spatial resolution). The peak position of amide I band was corrected to where the peak is supposed to be at 1654 cm^{-1} with the computational method.

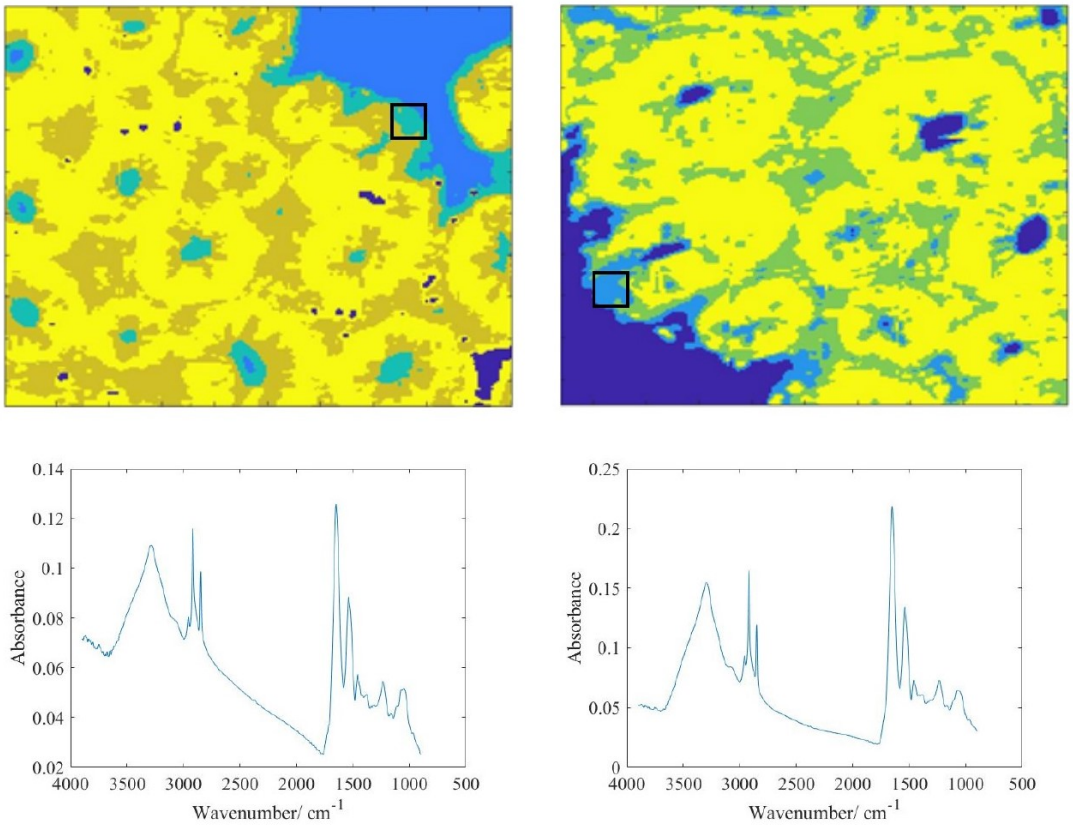


Figure 4.13: Top: false colour k-means cluster images of healthy colon tissue without the lens (left) and with the lens (right) obtained by mapping from nine stitched images. Each of the chemical images has a size of 510 $\times$ 510 μm^2. Cluster represented in light blue shade (box) indicates the edges of the tissue. Bottom: the average measured spectra from the areas representing the edges of the tissue

4.3.4 Analysis of FTIR spectroscopic images

Eight different tissue regions, which comprise of 4 training and 4 test models (as described in data processing procedure), were measured and analyzed. Chemical images showing

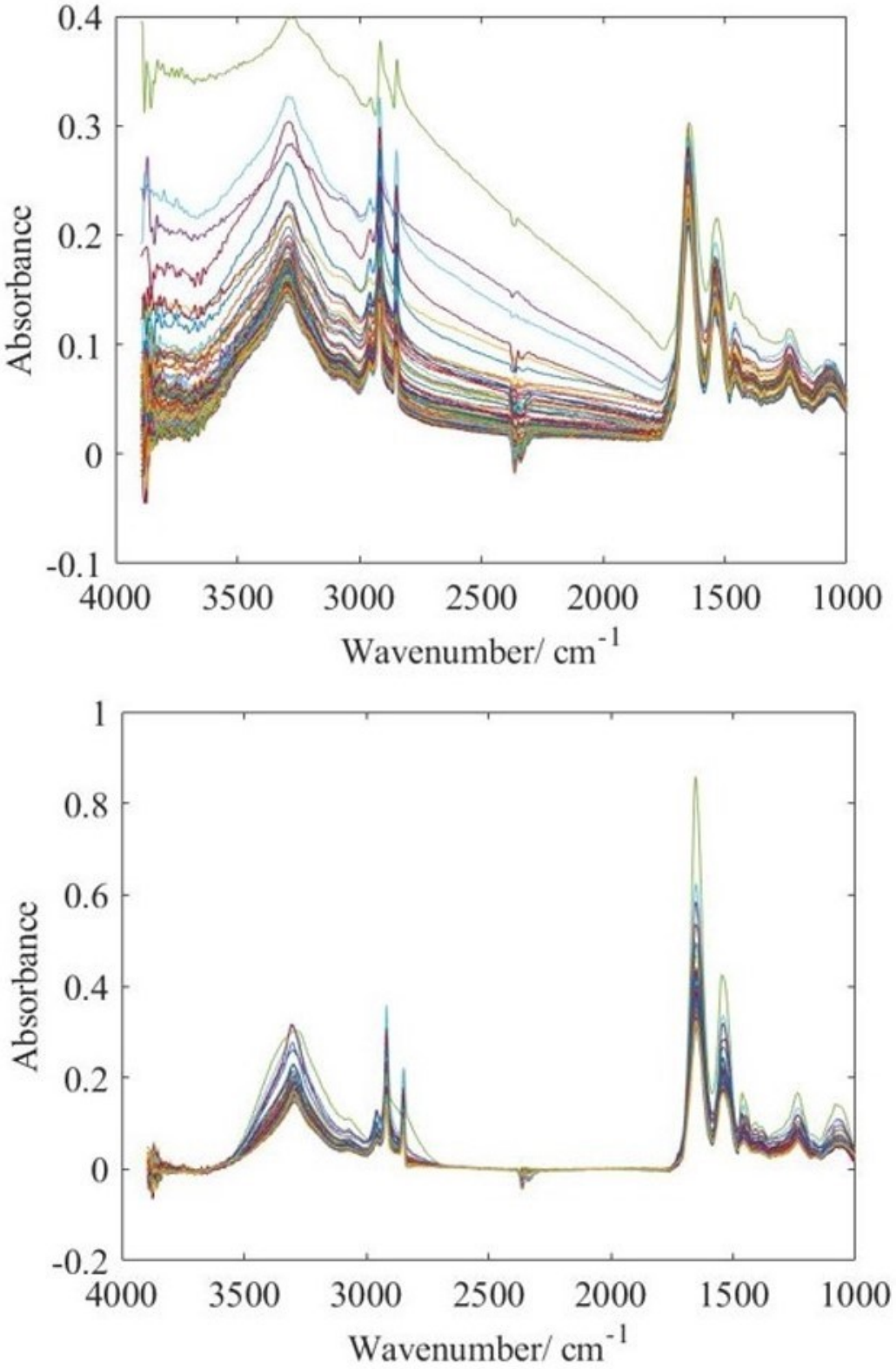


Figure 4.14: The raw spectra of 100 random pixels before and after computational RMieS-EMSC correction, shown on the left and on the right respectively. Resonant Mie scattering effect can be seen in the figure on the top prior to correction

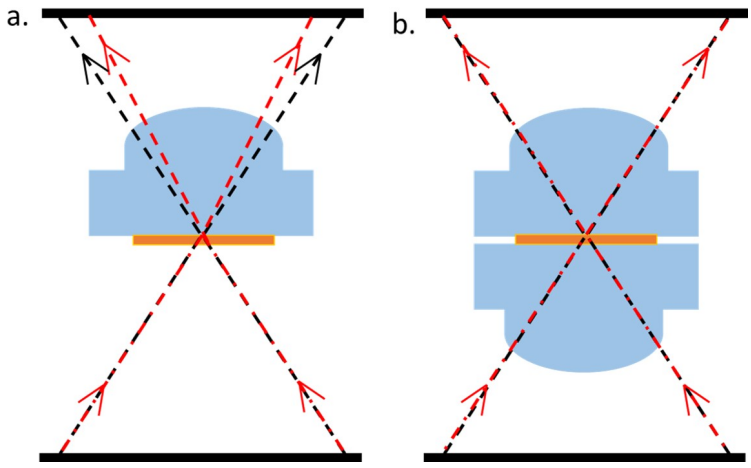


Figure 4.15: Schematic of (a) a single correcting lens and (b) two correcting lens on both sides of the sample

the distribution of integrated absorbance, estimated with trapezoidal rule of integration, at the spectral bands of 1272 – 1184 cm^{-1}, 1712 – 1589 cm^{-1} and 2944 – 2880 cm^{-1}, which are assigned to asymmetric phosphate stretching of nucleic acid, amide I, and CH stretching of lipid respectively (Movasaghi et al. 2008) are represented in Fig. 4.16 and 4.17, alongside the H&E stain images, which were used by pathologists to assign the stage of tissue malignancy.

As can be seen in the images in Fig. 4.16 and 4.17, the integrated absorbance of nucleic acid band at 1271 – 1184 cm^{-1} is lowest for healthy colon biopsy; likewise for amide I band. The opposite is observed for the lipid spectral band within 2944 – 2880 cm^{-1}, whereby the lowest integrated absorbance is achieved in cancer tissues. This is in agreement with the high nucleic acid-to-cytoplasmic volume ratio observed in colon cancer tissues (Li et al. 2014) as well as the loss of normal glandular architecture. The inner lining or mucosa of healthy colon is lined with columnar epithelium and large number of goblet cells, where numerous secretory vesicles containing mucus (glycoprotein) are present, in addition to the secreted mucin in the intestinal epithelial surface layer. Mucus is a complex biochemical layer made up of carbohydrates, antimicrobial peptides, immunoglobulins, electrolytes, and lipids (Bansil & Turner 2018). For diseased tissue however, the goblet cells are not differentiated well to perform its function, instead they become highly metastasizing cells with high metabolic rate, which might progress to cancer (an aggregation of undifferentiated cells).

The difference between different stages of cancer is also highlighted in the mean average spectrum obtained after taking their second derivatives. The evaluation of the variation is not very straightforward, thus the need for machine learning to perform the task of classification of colon disease. The machine learning algorithm only required the input of 'features', which is the wavenumber and 'label', the stage of disease. The sec-

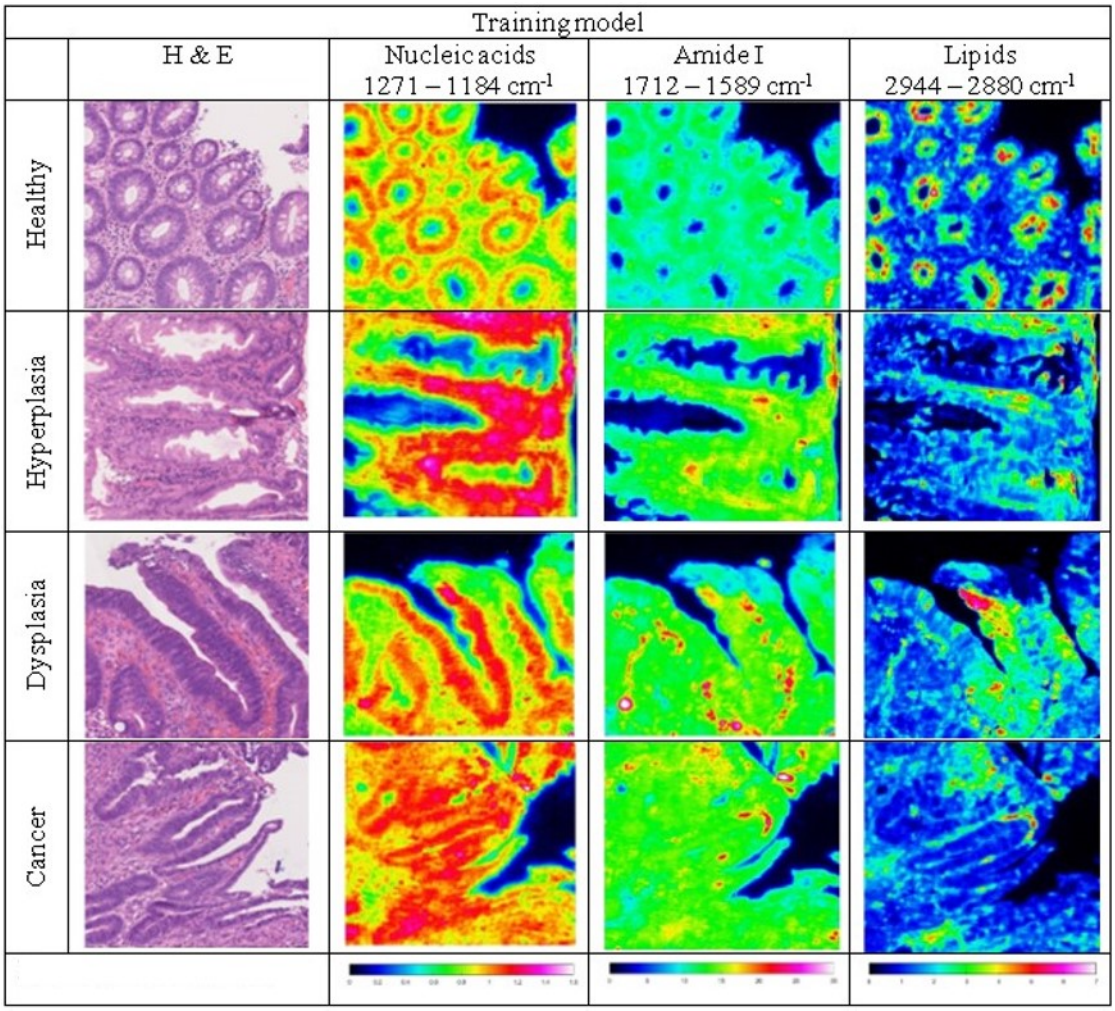


Figure 4.16: FTIR spectroscopic images of the colon biopsy used in the training models, depicting the distribution of different components by evaluating the integrated absorbance at various spectral ranges, which are labeled at the top of each column. The first column gives the H&E stained images identified by the pathologist. Each image has a size of 510 $\times$ 510 μm^2

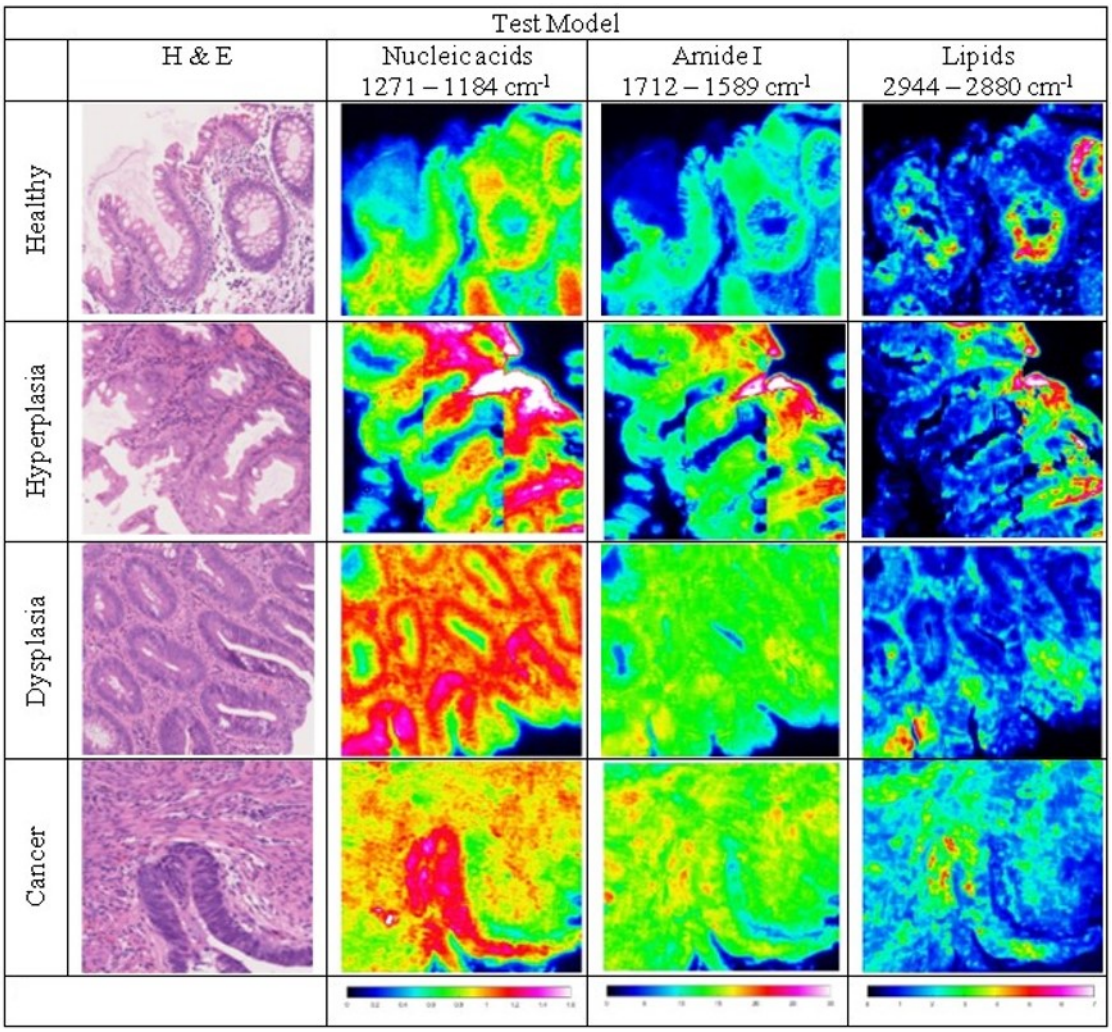


Figure 4.17: FTIR spectroscopic images of the colon biopsy used in the test models

ond derivative spectral bands and their corresponding band assignment are, nonetheless, provided in Fig. 4.18 and table 4.3 to demonstrate the potential variation that might be picked up by the machine learning classification model. The differences were picked up by the machine learning algorithm by recognising the peak shift and the intensity of the trough of the second derivative data.

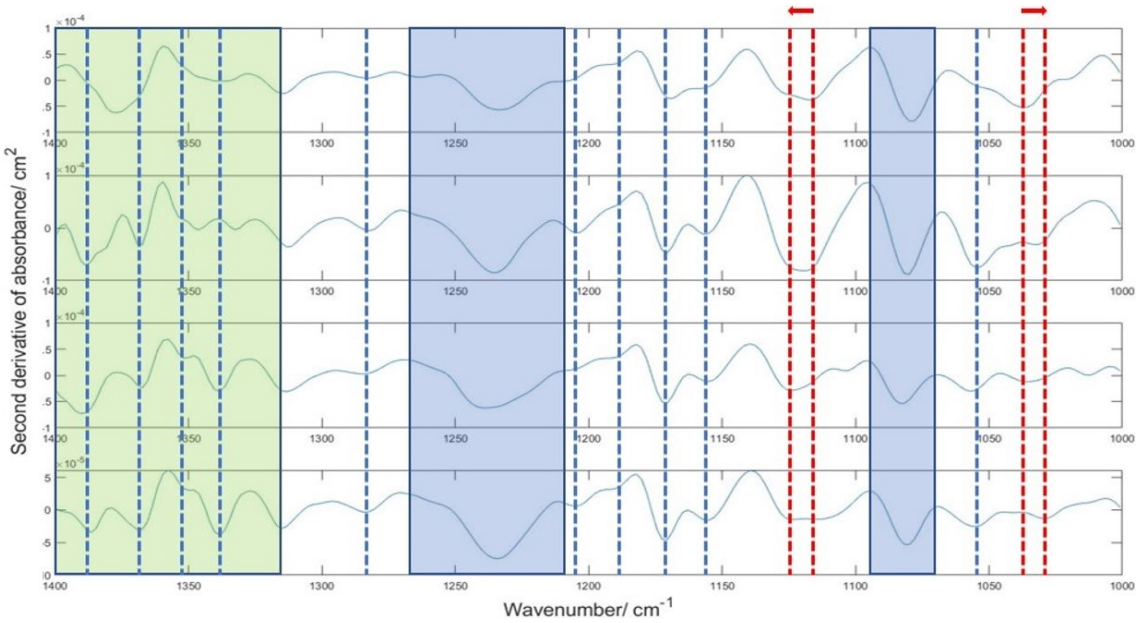


Figure 4.18: The second derivative spectra of colon biopsy tissue (from top to bottom: healthy, hyperplasia, dysplasia, and cancer) within the spectral range of 1400 – 1000 cm^{-1} by taking average of all pixels of high lipid absorbance (high lipid cluster classified via k-means clustering technique after water vapour subtraction). The red dotted lines show the shift in spectral band as colon cancer progresses, whilst the blue dotted lines denote the peak where only slight change is detected in the intensity of the trough is observed. The blue regions show the spectral ranges where significant changes in intensity are observed. On the other hand, the green region denotes the spectral range susceptible to minor interference of the water vapour peaks. The second derivative spectra in this region is compared before and after water vapour subtraction in Fig. 4.19. The details of the spectral observation are tabulated in Table 4.3

4.3.5 Unsupervised learning

K-means clustering were used for intra-tissue classification, by maximizing inter-distance variance between data within a tissue. It is important to recognize here that the optimum parameters for clustering in this study, after assessing the outcome of the supervised predictive model by comparing different spectral ranges and the number of clusters (results not shown), were based on the second derivative of the spectra between 3000 – 2800 cm^{-1} (introduced as the 'lipid region' henceforth, although strictly speaking the spectral band within this region is not limited to lipid, it is assigned to the C-H stretching of methyl and methylene groups) (Movasaghi et al. 2008). The three clusters identified were considered

Table 4.3: Differences in second derivative spectra with the increase in progression of colon cancer (from healthy to hyperplasia, followed by dysplasia and lastly cancer). Band assignment is taken from (Movasaghi et al. 2008)

Wavenumber/cm^{-1}	Band assignment	Observation in peak intensity	Observation in peak shift
1037	C-C, CH_2OH, C-O stretching coupled with C-O bending	–	Shift of band to lower wavenumber at 1030 cm^{-1}
1050	C-O stretching coupled with C-O bending of the C-OH of carbohydrates;Glycogen	Insignificant in healthy tissue, significant in diseased tissues	–
1080	Symmetric phosphate PO_2^- stretching; Collagen; Phosphodiester groups of nucleic acids	Decrease	–
1117	C-O stretching vibration of C-OH group of ribose (RNA)	–	Shift of band to higher wavenumber at 1124 cm^{-1}
1155	C-O stretching vibration	Insignificant in healthy tissue, significant in diseased tissues	–
1171	CO-O-C asymmetric stretching	Insignificant in healthy tissue, significant in diseased tissues	–
1190	Deoxyribose	–	–
1205	Amide III; Collagen	–	–
1235	Composed of amide III as well as phosphate vibration of nucleic acids	Increase	–
1263	PO_2^- asymmetric (phosphate I)	Significant only in healthy tissue	–
1282	Amide III; Collagen	–	–
1315	Amide III	Increase	–
1338, 1352, 1367, 1386	CH_2 wagging; Stretching C-O, deformation C-H, deformation N-H	–	–

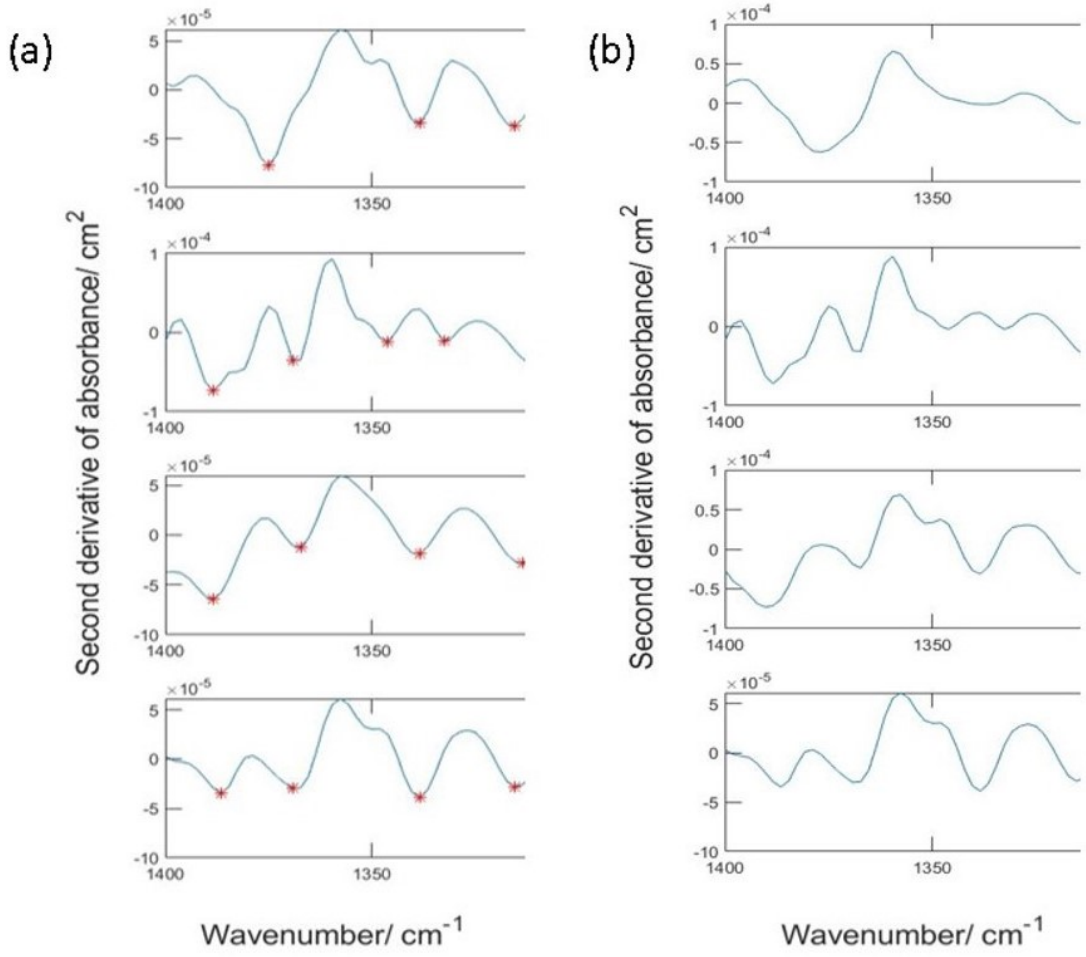


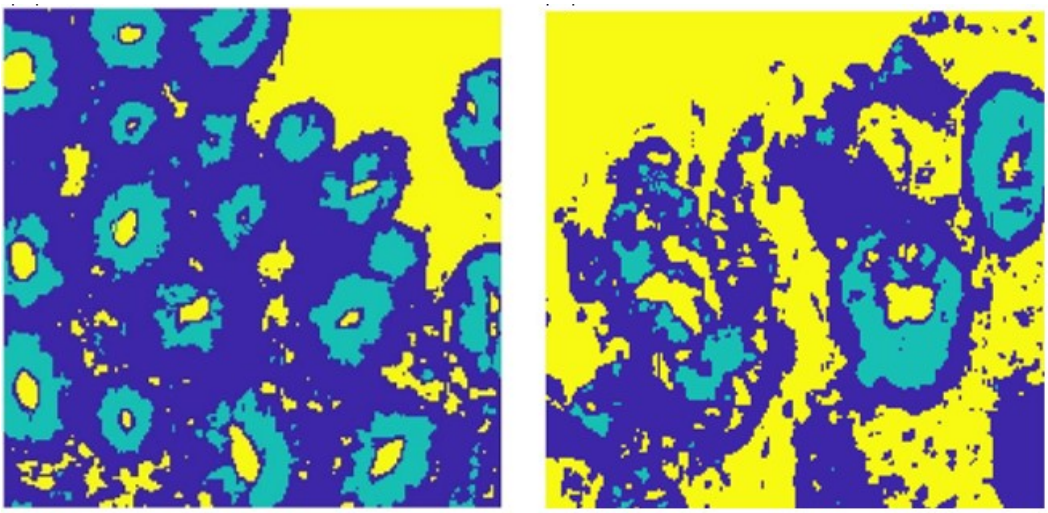
Figure 4.19: Second derivative spectra in the range of 1400 – 1325 cm^{-1} (a) before and (b) after water vapour subtraction. The contribution of water vapour was very little in this study, so elimination of water vapour via water subtraction method did not necessarily improve the performance of the RF predictive model

sufficient in this study following these reasonings: first of all, the various tissue morphology categorized by the clusters were fed into supervised machine learning independently as a way of phasing out unnecessary regions of the tissue since not all morphology or clusters showed essentially distinct spectra between different stages of colon disease, the highest performance was obtained with the spectra from high lipid region, which can be easily classified with just 3 clusters. Secondly, the higher the number of unsupervised clusters implemented, the higher the degree of similarity of the spectral data within each cluster, the lower the tolerance for dissimilarity of the test datasets, in other words, overfitting of data was introduced. In addition, since k-means is an unsupervised imaging approach, higher number of clusters has a tendency to cluster data that are close to each other which should have been treated as one (the sum of squared distances between each cluster decreases exponentially with increasing number of clusters). Besides, the main objective of this study is to assess the importance of lipid and amide bands in the prediction ability of

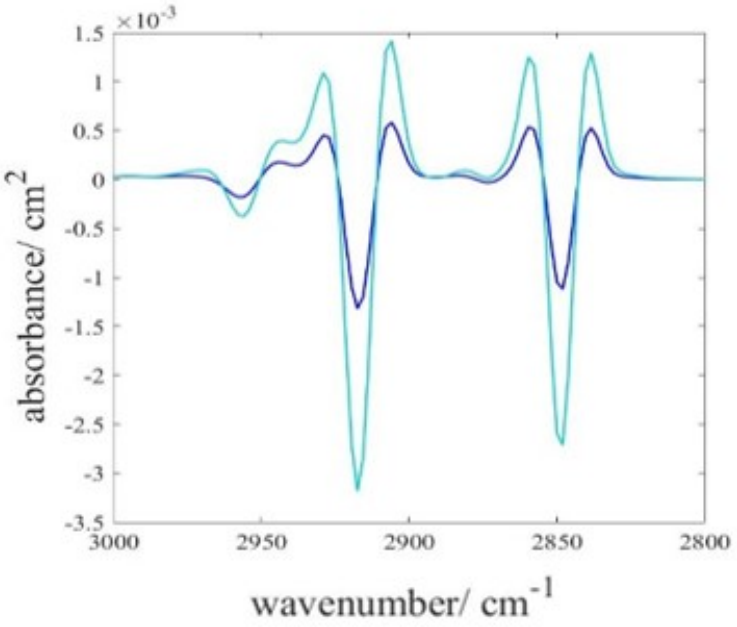
the RF machine learning, the least number of clusters which can output a good predictive performance is desirable, in this case, three clusters for the intra-tissue differentiation could warrant a prediction outcome greater than 90 % accuracy. Most importantly, higher category of classification (> 3 groups) was not needed as the lipid region is of secondary importance, as discussed before. Although the higher number of clusters were useful at exploration of the various histopathological architecture of colon adenocarcinoma, as analyzed by (Lasch, Haensch, Naumann & Diem 2004) using similar multivariate imaging approaches, higher number of clusters was not explored here. Classification of the tissue morphology was shown in Fig. 4.59 in Section 4.6. The false-color images generated from k-means clustering and their corresponding second derivatives are shown in Fig. 4.20.

From the mean second derivative spectra by averaging all the pixels within the same cluster, the tissue regions were effectively classified into low and high lipid absorbance regions (cluster 1 and cluster 2 respectively), which were fed into the supervised learning algorithm separately. This reinforced the previous findings in Section 4.2. that spectral bands of lipid are useful biomarkers for intra-tissue classification, despite the lower Gini importance index. The lipid spectral region contains a wealth of information. (Bassan et al. 2014) has also demonstrated that the high wavenumber spectral range (O-H, N-H, and C-H stretches occurring at ca. $3800 - 2500\,\mathrm{cm}^{-1}$) was useful for the generation of false color classification image of breast tissue microarrays on glass substrate. They were free from interference from the spectral bands of water vapor and Mie scattering, with the only possible variation coming from the de-paraffinization process on the formalin fixed tissues. This variation was controlled and minimized by strictly adhering to the de-paraffinization protocol.

It is possible that the cancerous tissues are more susceptible to change during solvent-based removal of material, that was carried out prior to the paraffin embedding process. The FFPE process required fixation of fresh tissue in formalin for 6 to 24 hours, followed by multiple washes in ethanol/water with increasing ethanol concentration until water has been removed. Xylene, or possibly isopropanol, was then used to remove the ethanol, taking with it much of the fats within the natural tissue. Finally, the tissue was soaked in molten paraffin, usually at 60 °C. Precautions were taken to conduct the de-waxing process in a closely controlled manner, so that each of the three samples were treated in the same way; however, the manner in which the FFPE was first conducted is out of our control, including the amount of fats and other materials that might have been removed in that process. That said, surprisingly similar observations were made on prostate cancer tissues that were supplied by different pathologists but de-waxed with the same procedure (Song et al. 2018), that this wavenumber region ($3000 - 2800\,\mathrm{cm}^{-1}$) is different between normal and cancer samples. Thus, the explanation that tissues of different malignancy retains various amount of fats after de-paraffinisation essentially still offers a different kind of 'key biomarker' for cancer differentiation in FTIR imaging study.



(a)



(b)

Figure 4.20: Representative color-coded k-means clustered images of healthy colon biopsy sections of (a) test and (b) training model. Cluster represented in light blue is for areas dominated by goblet cells (denoted as cluster 2), dark blue for basal membrane (denoted as cluster 1) and yellow for areas without tissue. (c) Average second derivative spectra of the corresponding clusters in the high wavenumber spectral region (3000 – 2800 cm^{-1}), following the color code in k-means cluster

4.3.6 Supervised machine learning

RF classifier was shown to be an efficient supervised machine learning technique for the classification of spectral data in previous studies (Smith et al. 2016, Balbekova et al. 2018, Goodacre 2003). In this study, second derivative data (for measurements with and without

correcting lens) from various spectral ranges were used to train the algorithm – Model 1: between 1800 – 1000 cm^{-1} and 3000 – 2800 cm^{-1} (all range), Model 2: 1800 – 1000 cm^{-1} only (fingerprint region with amide bands), Model 3: 1500 – 1000 cm^{-1} only (fingerprint region), and Model 4: 3000 – 2800 cm^{-1} only (lipid region). To clarify, re-training and re-testing of the RF models was still required after Gini selection to subjectively assess the prediction performance, hence the results are organized in the way shown in the workflow (Fig. 4.12).

The overall prediction accuracy for each model is shown in Fig. 4.21. A typical fingerprint region of infrared measurement is loosely defined to be between ~1600 cm^{-1} or 1500 cm^{-1} to 500 cm^{-1} (Baker et al. 2014, Stuart 2004). To avoid confusion, here, the fingerprint region is referring to spectral range within 1500 – 1000 cm^{-1} inclusive. At this part of the analysis, no computational correction of resonant Mie scattering was applied.

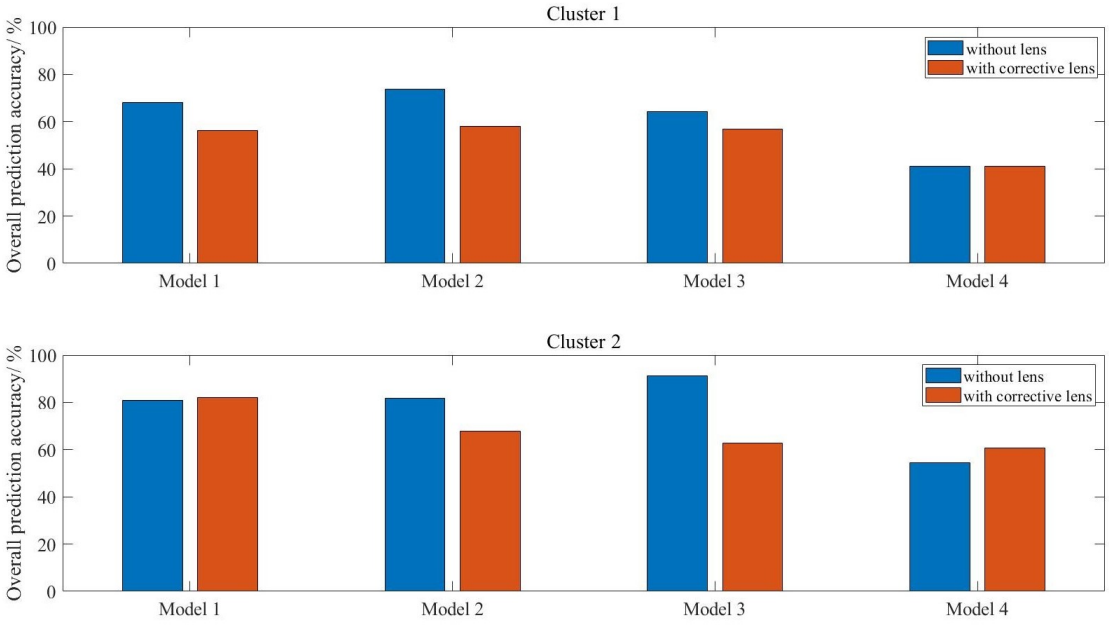


Figure 4.21: The bar chart shows the overall prediction accuracy in percentage of various models for measurement with and without correcting lens (and without computational correction for Mie scattering effect) for cluster 1 of low lipid absorbance and cluster 2 of high lipid absorbance

Fig. 4.21 shows that overall prediction accuracy is higher for data in cluster 2, region of higher lipid absorbance, than cluster 1. A comparison of the performance of measurements with and without correcting lens can be achieved by analyzing cluster 2, which reveals that apart from model 1 and model 4, the measurements with correcting lens, despite its ability to minimize Mie scattering at the edges of the tissues, generally underperform compared to measurements without the added lens. The lowest accuracy of cluster 2 prediction is obtained from model 3 with added lens. This is because while the added lens approach removes the scattering effect and thus improves the quality of amide

I band, the spectra collected in the range of 1100 – 1000 cm^{-1} suffer from enhanced noise, which is not an issue with computational approach. This happens because the additional stacking of lens on top of the CaF_2 substrate (from the way the correcting lens is set up) reduces the throughput of light. Due to the lower photon counts that pass through the sample and the fact that CaF_2 has a cut-off at ∼900 cm^{-1} in transmission, the spectral quality in the low wavenumber region deteriorates significantly compared to the set-up without correcting lens.

Model 2 (with lens) gives a slightly better performance when amide bands are factored into consideration as added lens is shown to improve the absorbance of the spectral band of amide I. Model 4 which considers the data exclusively from the lipid region is undeterred by the noise introduced by the extra lens configuration and model 1 which takes into consideration all the spectral regions shows similar performance with and without additional lens, for reasons discussed above. Instead of CaF_2, a pseudo-hemispherical ZnS lens with infrared cut-off at ∼700 cm^{-1} was suggested to improve the spectral quality (Kimber et al. 2016). However, Mie scattering correction does not play a significant role in optimizing the performance of the supervised learning, reinforced by the finding that the highest prediction accuracy of 92.7 % can be achieved with model 3 (fingerprint region). In other words, the second derivative spectral data within 1500 – 1000 cm^{-1} from cluster of high lipid absorbance region alone, is sufficient to achieve effective discrimination of all the different grades of colon cancer as the fingerprint region is least affected by Mie scattering. Therefore, removal of Mie scattering effect is not necessary as the amide spectral range (1700 – 1500 cm^{-1}) does not need to be included in data analysis at all, as demonstrated here.

On the other hand, model 4 gives the lowest prediction accuracy, this infers that lipid spectral region (or high wavenumber region) alone is not reliable for supervised training of the classification model in the study of colon biopsy. Nevertheless, the possibility of classifying between normal and cancer state of a biopsy, without classification of the stages of disease, based solely on the lipid region is not ruled out. For example, the study by (Pilling et al. 2017) for different type of cancer (breast cancer) and using different substrates showed that relying on high wavenumber spectral range alone allowed for rapid discrimination between normal epithelium, malignant epithelium, normal stroma, and cancer associated stroma of breast biopsies with classification accuracy as high as 95 %. However, the categorization of the different stages of breast cancer was not shown in their study.

The breakdown of the true positive rates (or true negative rates for non-healthy tissue) of each cancer grade (measurement without additional lens) is shown for all models of cluster 2 in Fig. 4.22. From these results, it is apparent that healthy and malignant tissues are easily distinguished from other stages of the disease, whereas dysplastic tissue is often misclassified as hyperplasia, if the correct spectral range is not implemented.

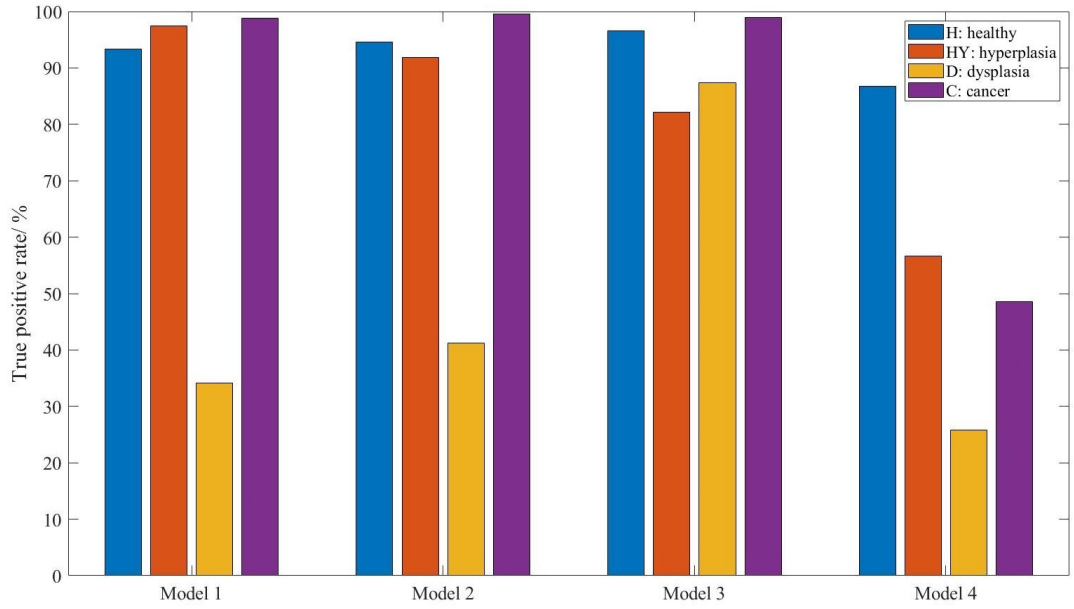


Figure 4.22: The bar chart shows the prediction accuracy of different stages of colon disease within each model of cluster 2 (high lipid absorbance area) for measurement without lens

Hyperplastic and dysplastic tissues rely heavily on differences within 1500 – 1000 cm^{-1}, possibly from the change in concentration of the nucleic acid and carbohydrates in the tissues (Baker et al. 2014), and can be classified at a high accuracy when only the fingerprint region is used. Hyperplasia and dysplasia exhibit very similar spectral pattern above 1500 cm^{-1}, hence they are best differentiated from each other when the amide and lipid bands, which have higher absorbance than the nucleic acid bands, were eliminated from the training dataset (Model 3). The results from supervised learning give a significant insight into assessing the spectral biomarkers of colon cancer.

It is important to note that a high intra-model prediction (prediction within the training model without test model) does not warrant a high inter-model prediction (prediction with the test model). In this case, inter-model prediction was employed as a better and more reliable guide to verify the efficiency of the machine learning and should be carried out where possible. The stability of the training model was confirmed by decreasing the ratio of the size of training to test models from 1:1 to 1:6, the error in the prediction accuracy is a mere ± 2.0 %. The optimum variance of PCA for the training of the model is found to be 99 %, ca. 20 % of the second derivative data within the fingerprint region contains useful information for data classification (Fig. 4.23 for results tested with variance of PCA retained ranging from 87 % to 100 %). The final best results of the prediction model summarized in a confusion matrix plot (The MathWorks, Inc. 2020*a*) is presented in Fig. 4.24.

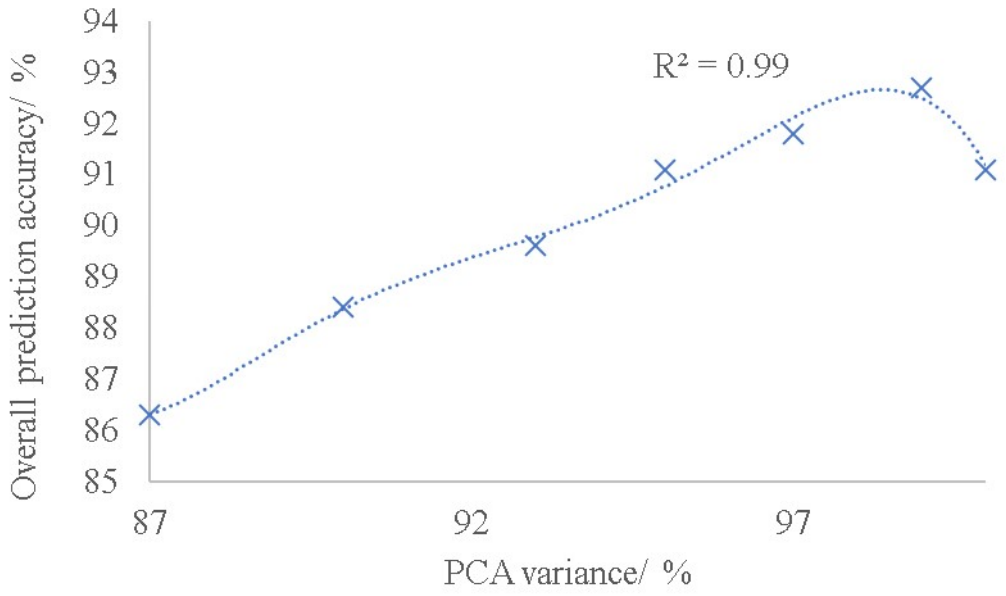


Figure 4.23: A plot of overall prediction accuracy of RF classifier of the same range (within fingerprint region only) for PCA with variance ranging from 87 % to 100 %

The findings are reinforced by comparing the prediction outcome with that obtained from spectral data after correction with RMieS algorithm (and without the correcting lens). The performance of the fingerprint region with amide bands after RMieS correction shows significant improvement in the prediction accuracy compared to the second derivative data before correction (from 81 % to 91 % prediction accuracy), despite being slightly lower than that of the fingerprint region alone, due to the correction of the amide I band (92 %). Correction with the RMies algorithm is computational while correction with the added lens is a practical optical approach, thus as expected the RMies algorithm provides a more precise solution which indeed yields a better overall prediction accuracy. The confusion matrices for both cases are provided in Fig. 4.25. In this study, both correcting lens and RMies correction are shown to be useful at correcting the scattering effect on amide I band but might not be necessary if classification of the stages of the colon adenocarcinoma via machine learning technique is the main objective as the training model without any correction for Mie scattering is sufficient to yield accuracy comparable to that after correction.

4.3.7 Spectral biomarkers from RF classifier

The choice of spectral wavenumbers for classification, as mentioned earlier, is based on the Gini index (Fig. 4.26). The results show that the fingerprint region ($< 1500\,\mathrm{cm}^{-1}$) contained most of the important features. This was followed by the lipid region (3000

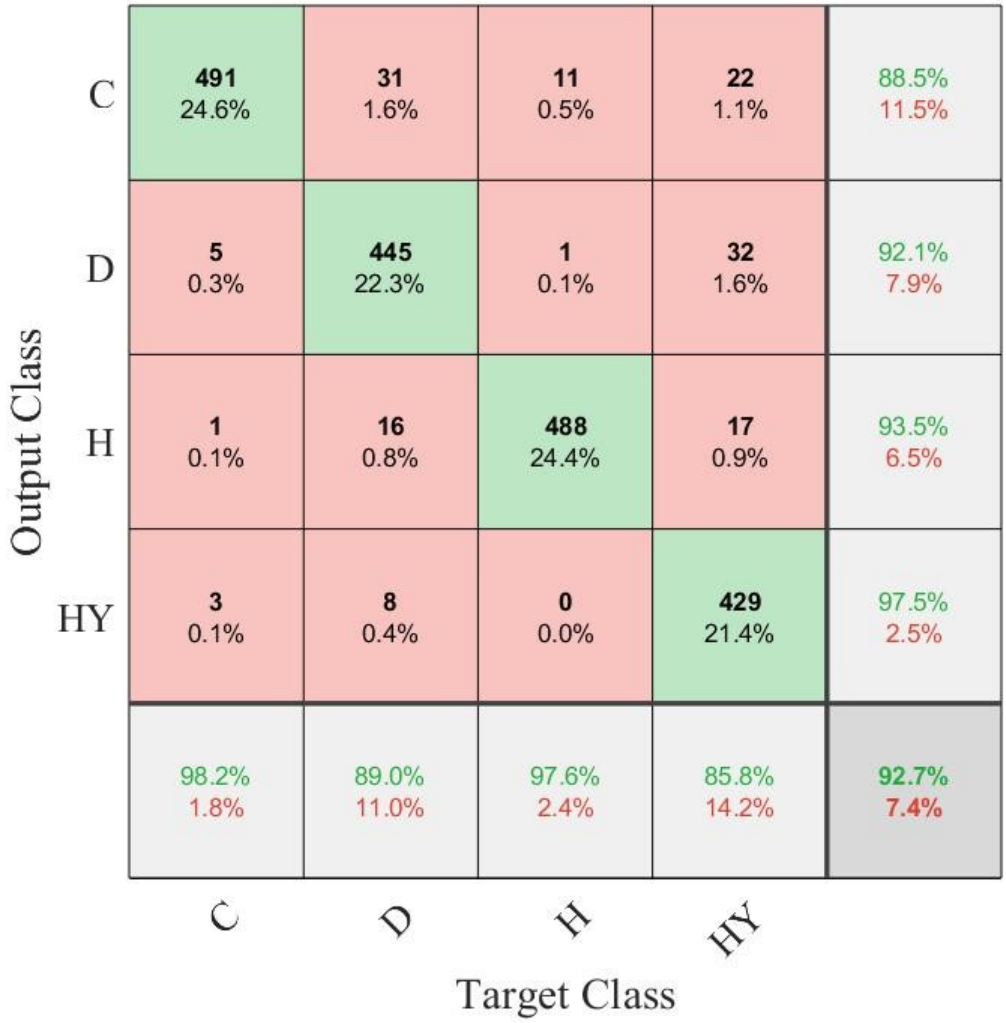
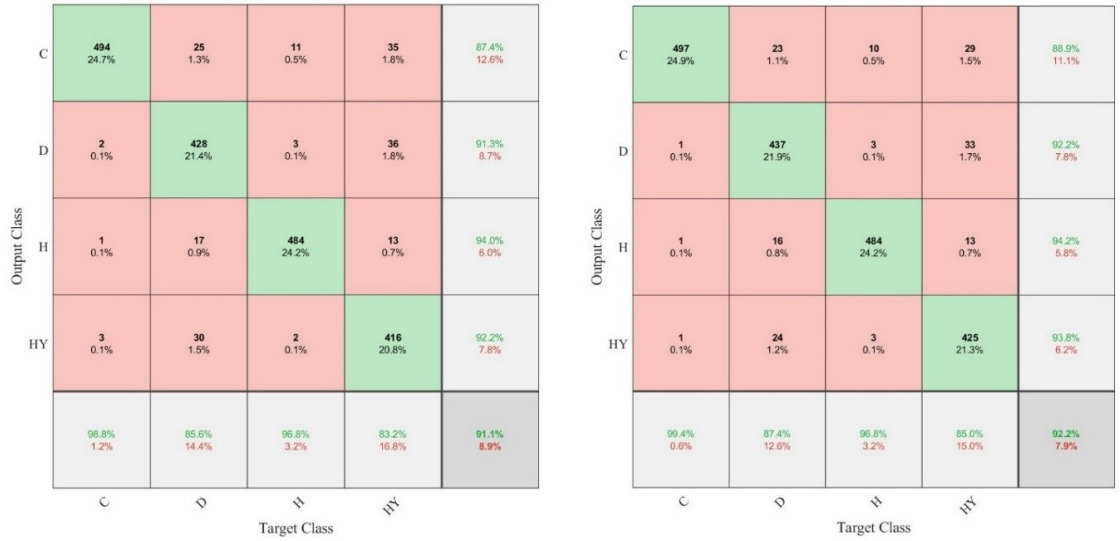


Figure 4.24: The confusion matrix plot shows the best result that can be obtained from fingerprint region of the spectral data with model 3 (C – Cancer; D – Dysplasia; H – Healthy; and HY – Hyperplasia). The rows show the predicted class and the columns represent the true class. The diagonal cells correspond to correctly classified observations, whilst the off-diagonal cells correspond to observations that are incorrectly classified. Both the number of observations and the percentage of the total number of observations are shown in each cell. The column on the right of the plot shows the percentages of all the examples predicted to belong to each class that are correctly and incorrectly classified. The row at the bottom of the plot shows the percentages of all the examples belonging to each class that are correctly and incorrectly classified. Overall accuracy of the prediction of the classifier model is given in the cell in the bottom right of the plot



(a) Fingerprint and amide regions of spectral data

(b) Fingerprint region of spectral data

Figure 4.25: The confusion matrix plots of the prediction outcome trained after correction with RMieS algorithm

– 2800 cm^{-1}) of secondary importance. Surprisingly, the best prediction accuracy was obtained when unsupervised training was applied on the spectral range of secondary importance, whilst the most important features were used for supervised training. A similar machine learning study was performed by (Kuepper et al. 2018) on colon cancer, however the spectral range used for the training was inclusive of the amide I band (Lasch, Haensch, Naumann & Diem 2004). It should be realized that the amide I region showed no significant importance here in this study, based on the Gini values.

4.3.8 Summary

FTIR spectroscopic imaging of colon biopsy tissues in transmission combined with machine learning for the classification of different stages of colon malignancy was carried out in this study. Two different approaches, an optical and a computational one, were applied for the elimination of the scattering background during the measurements and compared with the results of the machine learning model without correction for the scattering. Several different data processing pathways were implemented in order to obtain a high accuracy of the prediction model. This study demonstrates, for the first time, that C-H stretching and amide I bands are of little to no significance in the classification of the colon malignancy, based on the Gini importance values. The best prediction outcome is found when supervised RF classification was carried out in the fingerprint region of the spectral data between 1500 – 1000 cm^{-1}. An overall prediction accuracy higher than 90% is achieved through the RF. The results also show that dysplastic and hyperplastic tissues were well distinguished. This leads to the insight that the important differences between hyperplas-

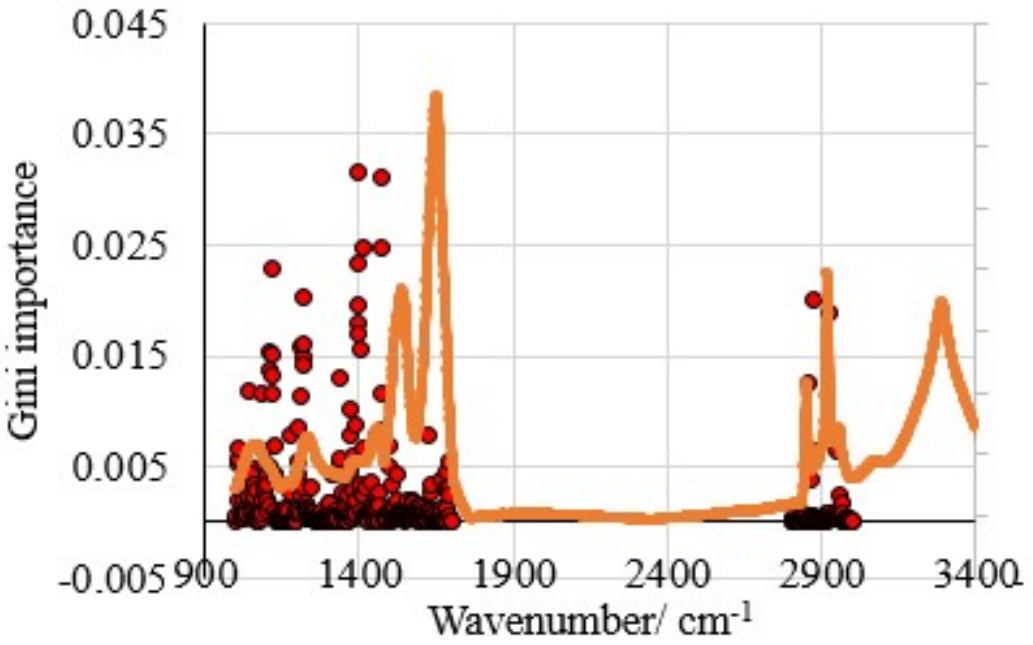


Figure 4.26: Plot of Gini importance values obtained from RF prediction model against wavenumber of colon biopsy tissue, overlaid on the average FTIR spectrum of healthy colon tissues for clarification purpose

tic and dysplastic colon tissues lie within the fingerprint region of FTIR spectra. Besides, the computational RMieS-EMSC correction performed better than optical correction, but the findings show that the disease states of colon biopsies can be distinguished effectively without elimination of Mie scattering effect, if the procedure for data processing shown here is followed.

4.4 Controlled humidity study on the classification of colon disease

Despite the increasing popularity of biomedical research with FTIR spectroscopy, translation of research studies of cancer with infrared spectroscopy to actual clinical environment have been difficult; there are a few challenges with regards to preparation of samples, such as FTIR instrumentation and data processing, as well as the ethics involved, which need to be addressed before bio-spectroscopy can become a routine process in the clinical settings (Baker et al. 2014). In particular, spectral performance could be affected by the humidity of the environment where the study is carried out. Humidity fluctuations may occur in practice because of varying weather conditions (Oliver et al. 2016). When applying analytical protocols to FTIR spectroscopic data obtained from dried versus hydrated materials, it is important to take into consideration the possibility that the bio-components within the tissue samples could potentially exhibit distinct FTIR characteristics in both the hydrated and dried forms.

Regulation of surrounding air humidity for experimental works is not uncommon. Experiments in the laboratory, such as the study of moisture sorption isotherm, hygrometer calibration, material conditioning to the study of hydration reaction such as the erosion of metals, require a precise control of humidity in its environment. FTIR spectroscopic imaging under controlled humidity has been reported previously both in transmission (Chan et al. 2004, Chan & Kazarian 2004, 2006) and macro ATR imaging modes (Chan & Kazarian 2007*b*). In the simplest and most convenient way, the control of humidity can be achieved by using salt solutions, either saturated or unsaturated, in a small sealed container to regulate the relative humidity (RH) (Rockland 1960). At a given temperature, the saturated salt solutions fixed at a defined concentration in a sealed environment with restricted flow of air will reach the desired equilibrium vapor pressure. The solution remains saturated in the presence of modest sources or sink conditions when excess solute is introduced, and the saturation is easily determined in the case where the solute is a solid in the pure phase (Greenspan 1977). The RH value is specific to the salt and the temperature; hence, this method can be used over a broad range of relative humidity, by simply changing the salts to obtain the desired RH.

The objective of this research was to find out the effect of hydration on the mid FTIR spectra of colon biopsy samples and the subsequent diagnostic performance by regulating the air humidity in a controlled environment. The added diagnostic values of humidity control to fixed colon biopsy tissues were the first to be explored in this area.

4.4.1 Regulation of humidity with saturated salt solutions

The instrumentation and sample preparation were as described in Section 4.3.1, without involving the use of additional correctiing lens. The visible images of the tissue sections measured in this experiment are shown in Fig. 4.27.

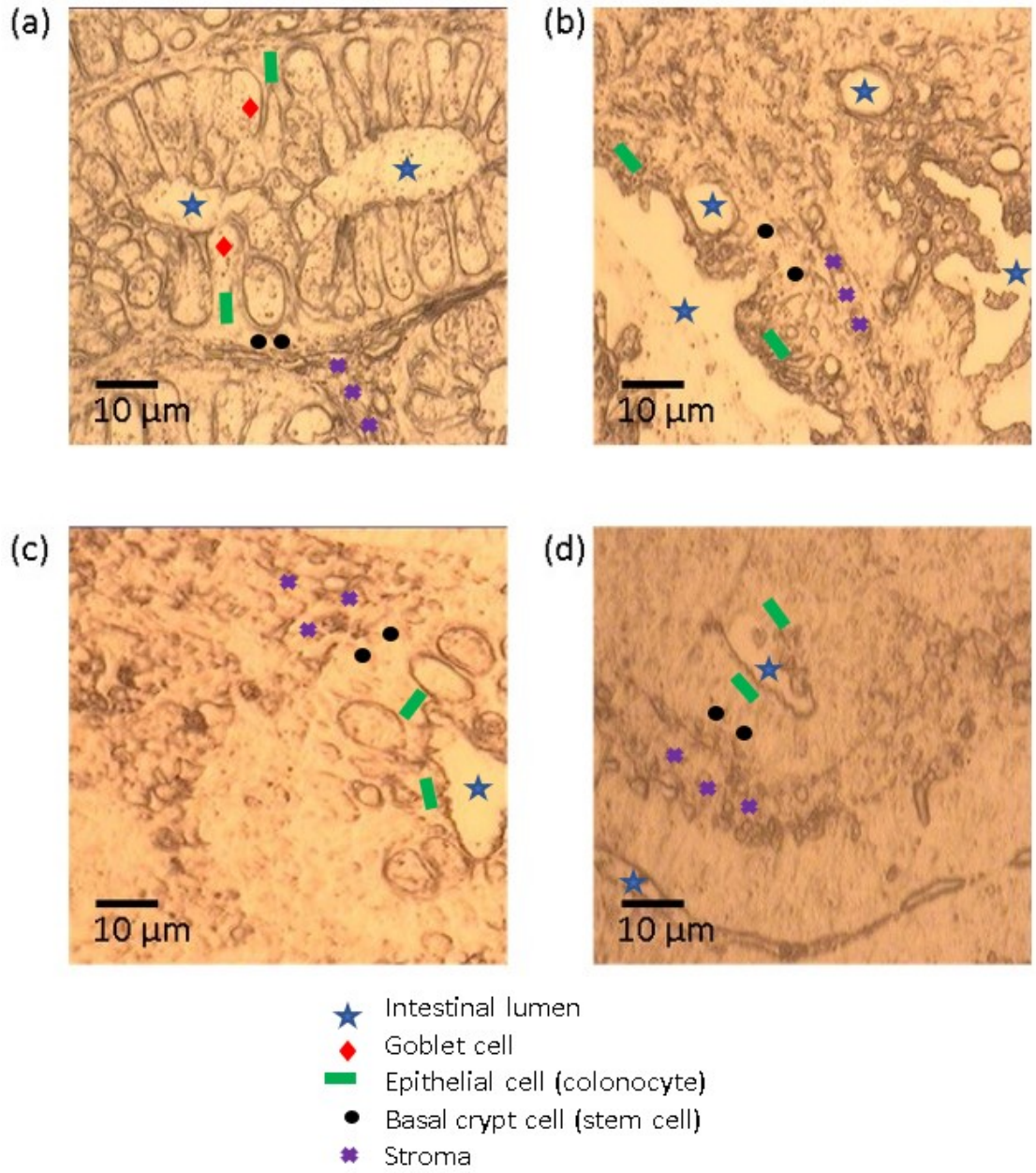


Figure 4.27: Visible (unstained) images of (a) healthy colon tissue to various degree of malignancy from (b) hyperplasia to (c) dysplasia and (d) cancer measured with visible camera under microscope at 15× magnification; each has an area of 70 × 70 μm^2

The humidity was regulated and controlled by saturated salt solutions. Five saturated salt solutions, sodium hydroxide (NaOH), magnesium chloride ($MgCl_2$), sodium bromide (NaBr), sodium chloride (NaCl), and potassium chloride (KCl) which yield RH

values of 16 %RH, 35 %RH, 58 %RH, 77 %RH, and 88 %RH at 20 °C respectively (O'Brien 1948), were prepared prior to measurement. In the preparation of each solutions, excess solute was added to a beaker of water heated on a hot plate up to 70 °C and magnetically stirred at 200 RPM until no more solute could be dissolved. The salt solution was then removed from the heat source and set aside to cool down to room temperature. The relative humidity of the environment where the measurement took place were maintained at specific levels by keeping the dish of salt solutions in a plastic box. A hole was made in the container to house the Cassegrain objective. A hygrometer was also placed in the plastic box to monitor and record the RH. A closed environment was ensured by sealing the container with sealing tape around the openings. The values of RH varied slightly (2 – 3 %) between experiments due to the variation in room temperature where the microscope was housed. A trial experiment was performed prior to the actual measurements to identify the time span needed for the humidity level to reach equilibrium. The equilibrium was achieved after 5 h. A measurement was taken when the equilibrium was initially achieved and another was taken after 24 h, and both spectra were identical. Therefore, it was deemed suitable that the samples were kept in the humidity box at the specific RH overnight before taking the actual measurements to ensure the equilibrium has been reached.

Between each measurement, the controlled humidity box was removed to prevent contamination and due to the sensitive nature of the Schwarzschild objective to moisture, it was not exposed to the high humidity (greater than atmospheric humidity which is approximately 45 %RH at 20 °C) for long periods. The measurements were taken as soon as the intended RH value was achieved. An illustration of the set-up of the controlled humidity box is shown in Fig. 4.28. A new background was recorded before measuring each individual image, all at the desired level of RH. The steps for data processing were the same as outlined in Section 4.3.2.

4.4.2 Chemical images as a function of humidity

The FTIR spectral data obtained were analysed in the form of chemical images also known as chemical maps showing the distribution of a specific component, whereby the values of each pixel are obtained by taking the integral area of the absorbance of a spectral band. For the colon tissues many absorption bands are found in the fingerprint (1500 – 1000 cm^{-1}) and amide (1800 – 1500 cm^{-1}) spectral region and at high wavenumber (3900 – 2800 cm^{-1}). The absorbance and positions of spectral bands are highly reliant on the structure and chemistry of a bio-component. The most notable vibrations of the bio-components captured in the FTIR spectrum are those of PO_2^- of nucleic acids, proteins (amide I), and CH_2 of lipids. The chemical images, based on their corresponding bands, showing distribution of these components are presented in Fig. 4.29.

The vibrational modes of the liquid phase of water manifest themselves as spectral

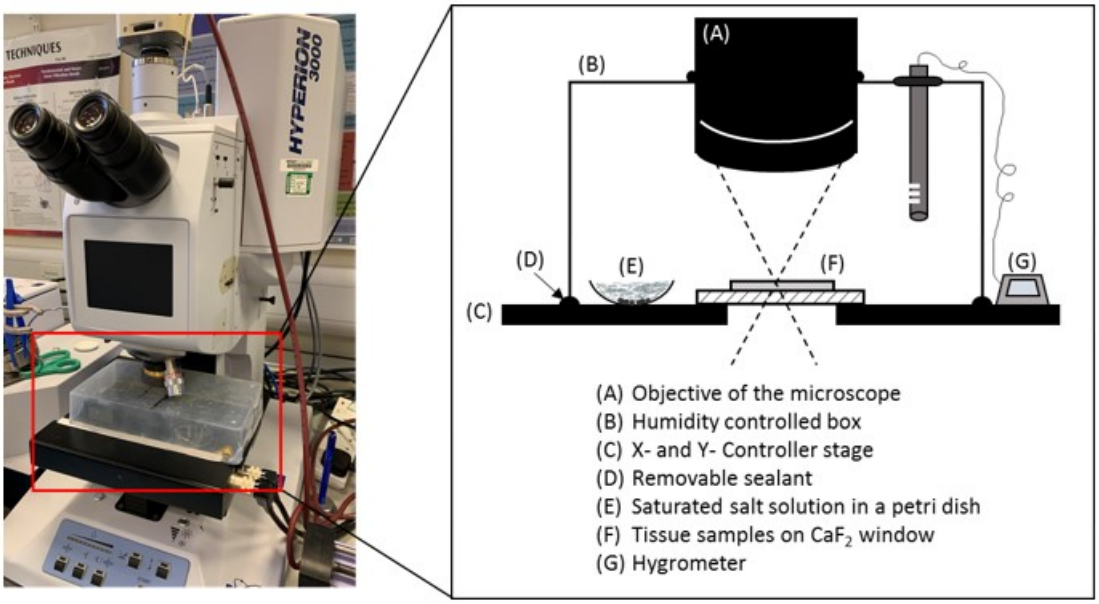


Figure 4.28: Picture showing the actual set-up of the controlled humidity box on the instrument in lab and the 2D schematic illustration of the set-up

bands at ∼1643 cm^{-1}, ∼2127 cm^{-1}, ∼3404 cm^{-1}, ∼3450 cm^{-1} and ∼3600 cm^{-1}, which correspond to the bending vibration (ν_b), the combination of its libration and bending vibration, overtone of the bending vibrations (2 ν_b), its symmetric stretching (ν_s) and anti-symmetric stretching (ν_{as}) respectively (Venyaminov & Prendergast 1997). The broad band corresponding to the stretching modes of water (3700–3100 cm^{-1}) has been used in the past to monitor sorption of water at different RH (Chan & Kazarian 2004). Likewise, in this study, the spectral band of water at ∼1643 cm^{-1} overlaps with the amide I band of colon tissue; thus the high wavenumber region between 3539 – 3332 cm^{-1} is utilised to analyse the hydration of tissue. From Fig. 4.29 column (v), it can be seen that at low humidity of 16%, the tissues are in a de-hydrated state where the water spectral peaks at this high wavenumber band were absent, although the healthy tissue retains an insignificantly low amount of water. As the humidity of the surrounding (the environment in the box) increases, it is expected that the water content of the tissue increases as well. A thin, healthy colon biopsy sections of 3 μm has the capability to absorb (maximally) around 368 % of the water from the air (based on the values of the absorbance of a corresponding spectral band) in comparison to its de-hydrated state, depicted in Fig. 4.30. From the graph, healthy colon tissue seems to exhibit a tendency to be hydrated at a greater degree, followed by hyperplastic and dysplastic tissue, and finally cancerous section, which absorb around 207 %, 164 %, and 76 % of water from its surrounding (based on the integrated absorbance of the OH band between 3539 – 3332 cm^{-1}) at 82 – 88 %RH respectively. It is hypothesised that this may be due to the high nuclear-cytoplasmic ratio of the cancerous sample that obstructs the diffusion of water vapour into the sample, or the presence of colonic crypt that encourages water retention in the healthy tissue.

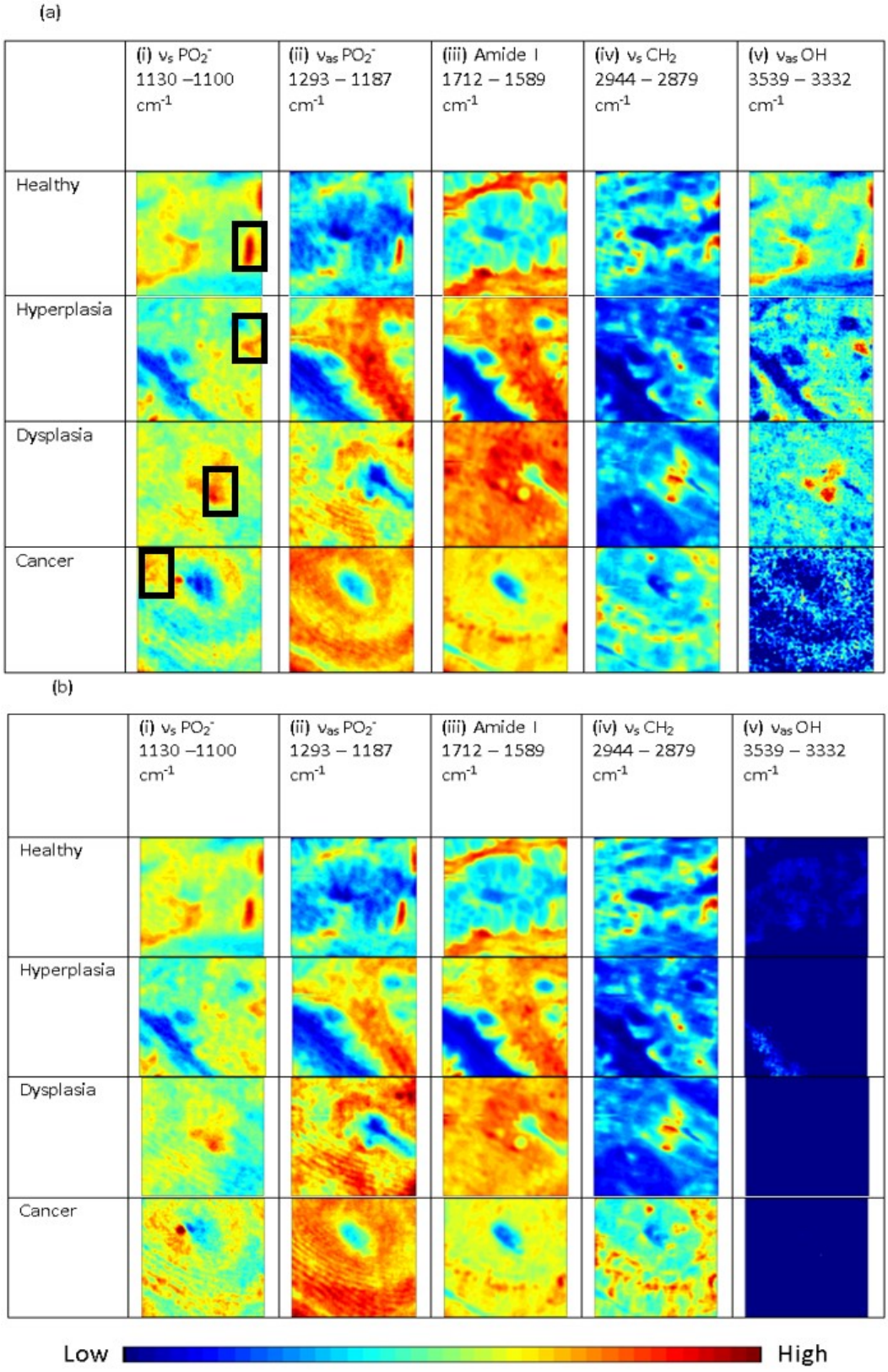


Figure 4.29: Chemical images generated for five different spectral bands across tissues of different grades of malignancy – (a) at high humidity of 88 %RH and (b) at low humidity of 16 %RH. Each image has a size of 70×70 μm^2. The intensity of the images are presented in jet colormap. Spectra are extracted from the box areas in column (i) of subfigure (a) for further chemometric analysis with PCA and RF

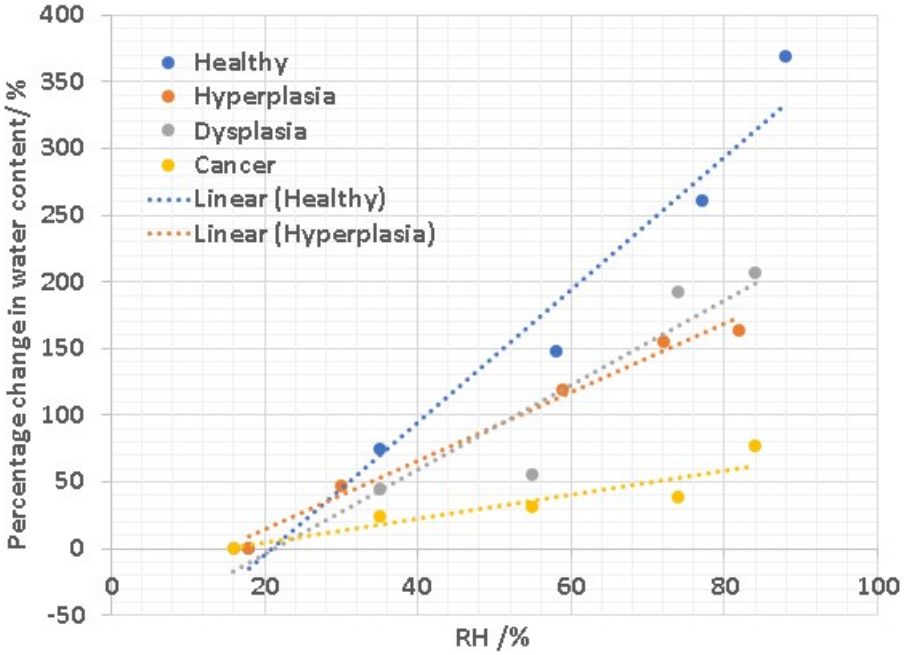


Figure 4.30: Plot of percentage change in the amount of water in the tissues across the different levels of humidity, based on the analysis of the water spectral band between 3539 – 3332 cm^{-1}

Moreover, at the high humidity of 88 %RH, the chemical images obtained from the analysis of the water band (Fig. 4.29 column (v)) show that the hydration of the tissue was in fact non-uniform throughout the tissue section. Water was absorbed mostly in the areas of the high absorption of spectral band between 1143 – 1100 cm^{-1} with slight coincidence with phosphate PO_2^- (1293 – 1187 cm^{-1}) and methyl (and methylene CH_2) rich areas. Between 1143 – 1100 cm^{-1}, the spectral bands correspond to ν_s PO_2^- and ν CO^- of ribonucleic acid RNA (and a small contribution from polysaccharides) (Movasaghi et al. 2008). Thus, it is expected that nucleic acids, especially RNA, follow a different behaviour in the IR spectra acquired at high humidity due to their interaction with the water absorbed into the tissues.

Furthermore, it can be seen from the chemical images in Fig. 4.29 column (iii) that the amide I (1712 – 1589 cm^{-1}) rich region is found in the colonocytes and lamina propria mucosa. However, the absorbance values of amide I of the same tissue cross-section vary across humidity level. They are higher at high RH compared to low RH. The overlapping of water and amide I band results in an inaccurate representation of the amide I present in the tissues if the amount of water is not subtracted. It has been reported in Section 4.3 that this amide I band is not the most important spectral biomarker in the classification of colon cancer (Song et al. 2019). While the main objective of this study is to compare the effect of humidity on the spectra, the tissues were not subjected to the iterative RMieS-EMSC correction algorithm (Bassan, Kohler, Martens, Lee, Jackson,

Lockyer, Dumas, Brown, Clarke & Gardner 2010). This is also justified by previous findings on the same samples that unlike cells, colon tissues are less sensitive to Mie scattering due to its closely packed nature (Song et al. 2019). On the other hand, analysis of the spectral band at 2879 – 2844 cm^{-1} (Fig. 4.29 column (iv)), assigned to ν_s CH_2 of fatty acids, lipids, phospholipids, and cholesterol, reveals that they are abundant in goblet cells, the region closest to the central lumen of the crypt in the healthy tissue but this characteristic is lost in the other tissues. This is because following the progression of colon cancer, the presence of the specialised goblet cells that line the mucosal surfaces is not seen. Moreover, the high abundance region of this lipid (or phospholipid) components are complementary to the amide I rich region – where one is high in absorbance, the other is low.

4.4.3 Analysis of the spectra

The second derivative of the spectral bands between 1300 – 1100 cm^{-1}, also known as the phosphodiester region, that has been identified previously to coincide with the areas of hydration within the tissue from analysis of the chemical images are shown in Fig. 4.31. Second derivative spectra have the advantage that they are free from the baseline shift, as well as enhancing resolution by de-convoluting overlapping peaks from multiple components (Gutierrez 1992). As the surrounding humidity changes, it is apparent that the second derivative spectra vary, reflecting a change in the conformation or arrangement of biomolecular components of the tissue.

A critical finding from this experiment is the consistent peak shift, perceptibly from 1230 cm^{-1} to 1238 cm^{-1} across all tissues, regardless of their malignancy states, as the humidity gradually rises. This peak correlates with the ν_{as} $PO_2{}^-$ and is often associated with the nucleic acid DNA. The dynamics of en masse DNA conformation have shown to be dependent on several factors, one of which is the water content (Franklin and Gosling 1953). In a humidity study on pure DNA by (Falk et al. 1963), it was hypothesised that the $PO_2{}^-$ group of DNA was the first to become hydrated, followed by the interaction of water with the C-O-P and the C-O-C groups of phosphodiester linkage at humidity and finally the hydration of the bases. The wavenumber and absorbance changes of the $PO_2{}^-$ spectral bands were not completely unsubstantiated. A conformational change in the DNA has been shown to manifest in the blue shift of the ν_{as} $PO_2{}^-$ peak from 1227 cm^{-1} to 1236 cm^{-1} which according to the findings is a diagnostic of a transition from hydrated B-DNA to dehydrated A-DNA (Whelan et al. 2013*a*). In normal physiological condition, A-DNA is rarely present – it starts appearing when the humidity of the environment is below ∼75% RH. A-DNA and B-DNA are both right-handed double helices made up of deoxyribonucleotides. By comparison, the structure of A-DNA is generally broader than B-DNA. A-DNA comprises 11 base pairs per turn with a length of 2.86 nm; whilst the

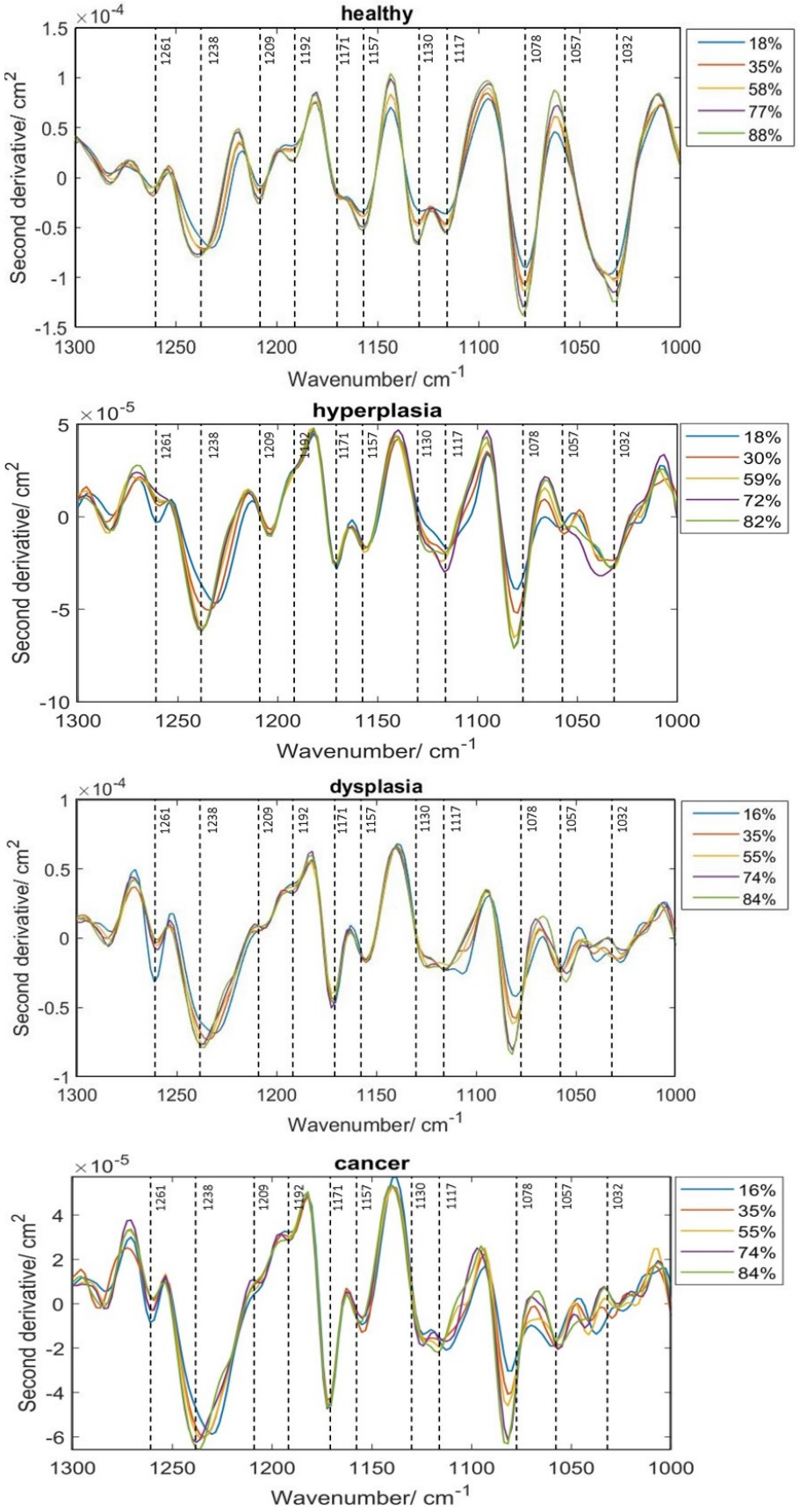


Figure 4.31: The average second derivative spectra in the phosphodiester region between 1300 – 1000 cm^{-1} across the RHs

latter consists of 10 base pairs per turn with a length of 3.4 nm (Berg et al. 2002). It should be noted, the shifting of the peak at ∼1227 cm^{-1} to a higher wavenumber, as observed in (Whelan et al. 2013*a*), is not recorded in this study. A plausible explanation here, keeping in mind that the samples measured are de-paraffinised tissues, is that the DNA that has been through the FFPE process is incapable of undergoing the conformational change. Therefore, this study discriminates from those involving live cells and pure, isolated DNA (Kondepati et al. 2008). In fact, the presence of the band at 1238 cm^{-1} reaffirms the adoption of a more disordered A-DNA conformation in the de-paraffinised FFPE colon tissues (Wood 2016). Hence, the small shift of this band by 8 cm^{-1} could potentially arise from the lack of hydrogen bonding of the phosphate group in the A-form of DNA.

Furthermore, the region below 1030 cm^{-1} suffers from poor S/N ratio. Thus, the changes of the spectral bands annotated to another left-handed double helix DNA – the Z-DNA (1018 – 1014 cm^{-1}) (Wood 2016) are hardly observed in this study due to its weak signal in the transmission FTIR spectroscopy. Additionally, at a high level of humidity, the spectra of healthy tissue shows a deconvolution of a broad spectral band between 1144 – 1095 cm^{-1} to two distinct peaks at 1117 cm^{-1} and 1130 cm^{-1}, that corresponds to CO stretching vibration of C-O-P of ribose (RNA). Specifically, vibrations of the skeletal structure around the $2'$-OH group of the ribose residue (Zucchiatti et al. 2016), and of disaccharides (or carbohydrates) respectively, which are not successfully resolved in a dry environment, are recorded when the humidity is high. This behaviour is not seen in other diseased colon biopsy specimens. Our results on the lack of dependence of RNA on hydration compared to DNA are identical to previous findings by (Pevsner & Diem 2003). This is because RNA exists only in the A-form and it is energetically unfavourable for any transformation of the conformation to take place due to the attachment of the hydroxyl group to the carbon of the ribose ring. In the non-healthy tissues, these two spectral bands convolute into one broad band. Apart from the band shift and deconvolution that are discussed above, an alteration to the vibrations of the phosphate functional groups of the nucleic acid components is congruous with the observation that the intensity of the second derivative peak at 1078 cm^{-1} (ν_s PO_2^-) amplifies with RH; whilst the peak at 1261 cm^{-1} (ν_{as} PO_2^-) behaves contrary to the former trend. Another band that increases in intensity includes the C-O stretching vibration at ∼1045 cm^{-1} (of the healthy tissue). An identical remark on the enhancement of absorbance at 1087 cm^{-1} and 1050 cm^{-1} was also noted in the re-hydration spectral of mammalian lymphocytes and rodent fibroblasts (Tan & Chen 2006, Whelan et al. 2013*b*). The same general pattern displayed here upon re-hydration indicates that the response appears omnipresent not just across eukaryotic cells, but also tissues.

Despite the slight overlapping between water sorption areas in the tissues and the C-H rich areas (2944 – 2879 cm^{-1}), there is no difference in the spectra across different levels of humidity within the high wavenumber spectral region (Fig. 4.32); thereby

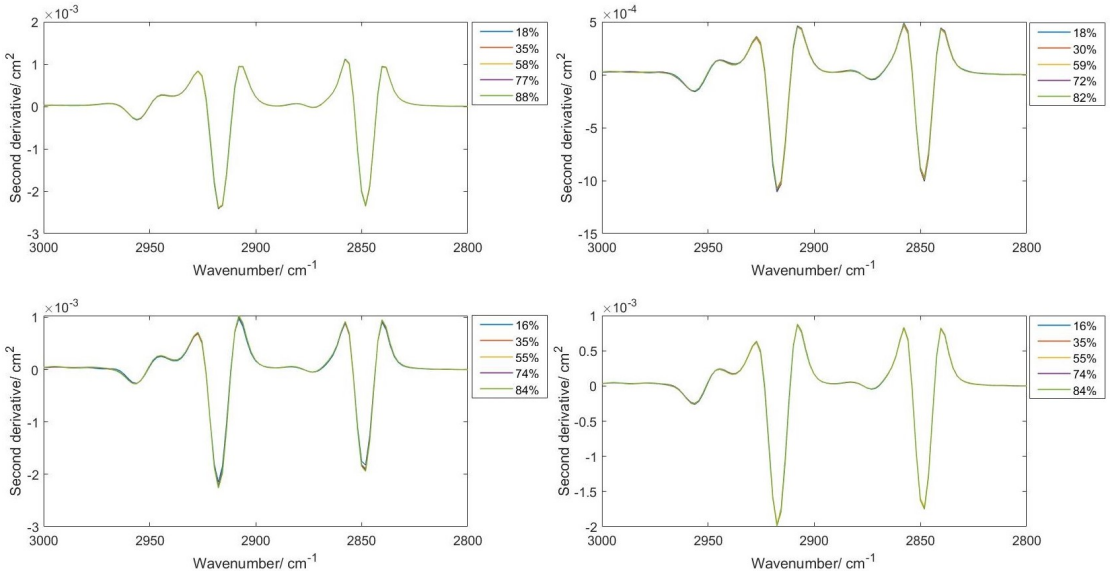


Figure 4.32: The average second derivative spectra in the lipid region between 3000 – 2800 cm^{-1} across the RHs

affirming the lack of interaction between water with the methyl (or methylene) groups of predominantly lipids. The spectral bands of lipid molecules remain relatively constant, independent of the tissue hydration; whereas spectral region beyond 3300 cm^{-1} is sensitive to the change in the hydration of the tissue as demonstrated above in the chemical images (Fig. 4.29), rendering this region ($>$ 3300 cm^{-1}) unsuitable for further chemometric analysis for the classification of colon tissues.

The second derivative spectra were also compared across tissues. First of all, the most evident differences between healthy and diseased tissues are the appearance and disappearance of several peaks with the progression of colon malignancy. Furthermore, the peak at 1192 cm^{-1} disappears beyond the healthy state of colon tissue (which also shows a decrease in its intensity when the humidity is low in healthy tissue); whilst the peak at 1209 cm^{-1} is not observed beyond the hyperplastic state. These peaks and their vibrational modes are summarised in table 4.4. Secondly, there is a discernible progressive increase in ratio of peak intensity of the second derivative spectra at 1171 cm^{-1} to 1157 cm^{-1} $\left(\frac{I_{1171cm^{-1}}}{I_{1157cm^{-1}}}\right)$, attributed to the ν C-O of proteins, such as collagen and carbohydrates. Our findings in this study are consistent with the results reported previously (Rigas et al. 1990). In the spectra of human colon tissues, these bands are essentially from the ν C-O mode of cell proteins. The presence of spectral band at 1173 cm^{-1} – 1164 cm^{-1} was previously recorded to change in both shape and peak absorbance in malignant human colorectal cell lines. Consequently, the peak position was observed to shift from 1164.2 cm^{-1} in normal tissues to 1173.1 cm^{-1} in malignant tissues (Rigas et al. 1990), similar to our findings here.

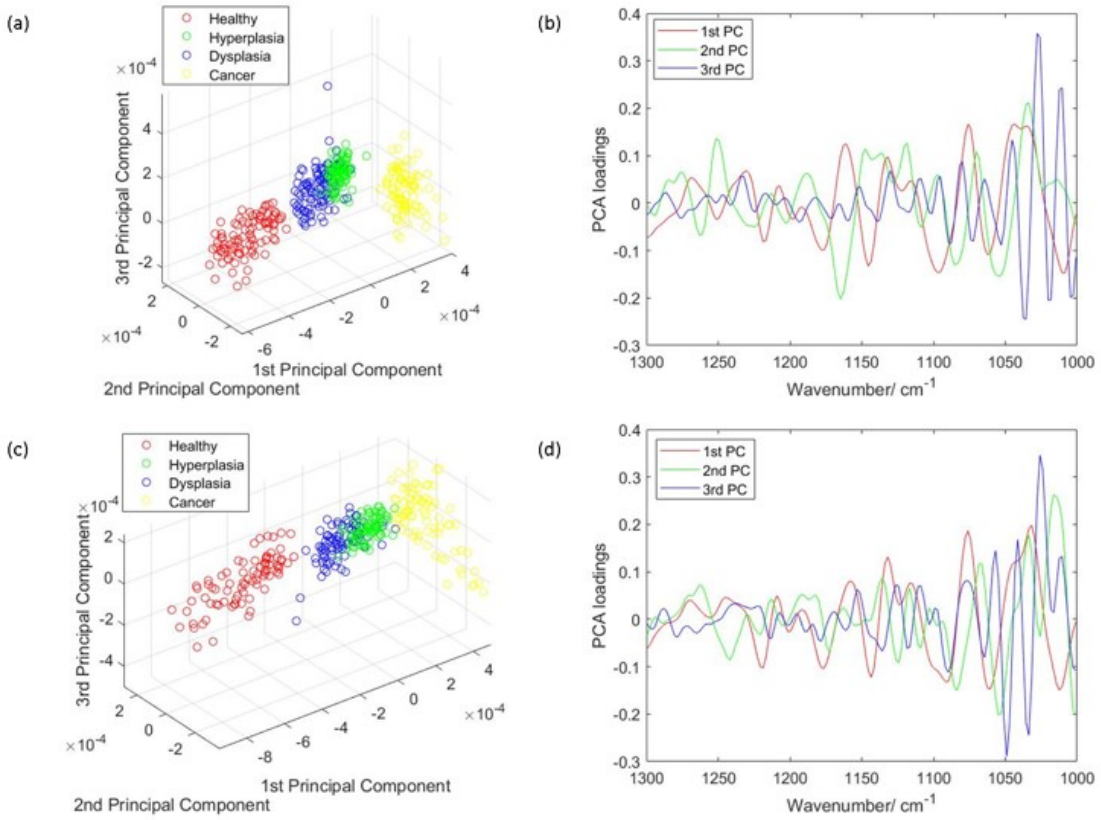


Figure 4.33: PCA score plots (a, c) and loading plots (b, d) at the lowest humidty at 16 %RH and the highest humidity at 88 %RH, respectively

4.4.4 Classification of colon disease at various relative humidty

The second derivative spectra were subjected to dimensionality reduction with PCA. The spectra were extracted from a small area of the tissues where the absorbance at 1143 –1100 cm^{-1} is found to be highest (indicated by black box in Fig. 4.27(a)). PCs that accounts for 90 % of the total variance of the data were retained for subsequent processing with RF classification. Only the two PCA scores at the lowest and highest humidity are presented here as they correspond to the first two models with the best classifications. The data were projected along the first three PCs that corresponds to 45.39 %, 6.45 %, and 3.96 % of the variance respectively, plotted in Fig. 4.33 (a, c).

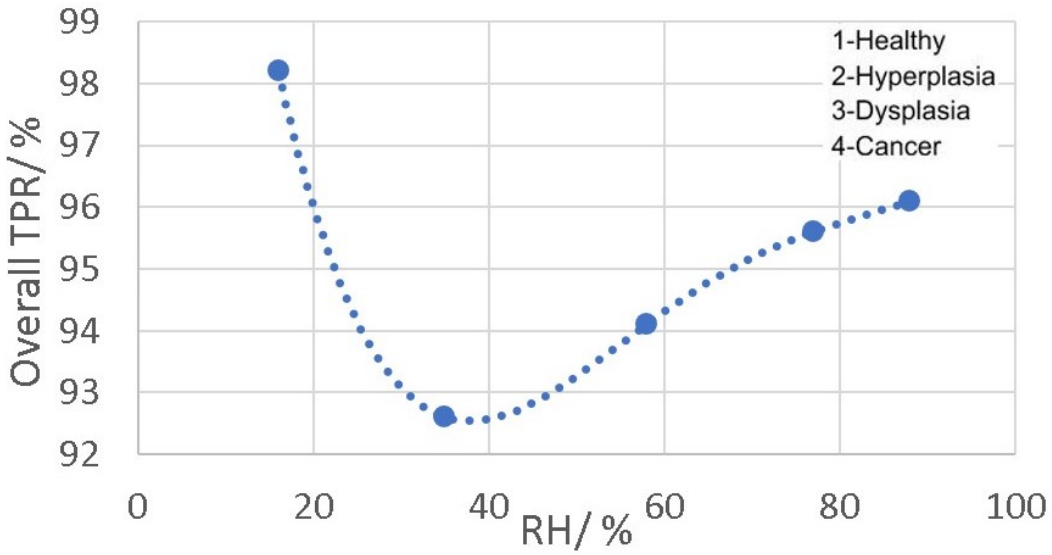
Clusters of data from healthy and cancer samples are well discriminated on the plot when the humidity is at its lowest. At the same time, the data from hyperplastic and dysplastic samples share a low degree of overlap. Also, from the PCA score plot, the cluster of cancer data is observed to shift closer to dysplasia as tissues are hydrated – in other words, we surmise that the classification of healthy and hyperplastic tissues is insusceptible to the change in surrounding humidity, and cancer tissues are easily misidentified when tissue hydration occurs. This finding is also elucidated in the confusion matrices (Fig.

Table 4.4: List of spectral peaks present, marked by '✓' (or absent, denoted by '×') in the healthy colon tissues and those of different malignancy states

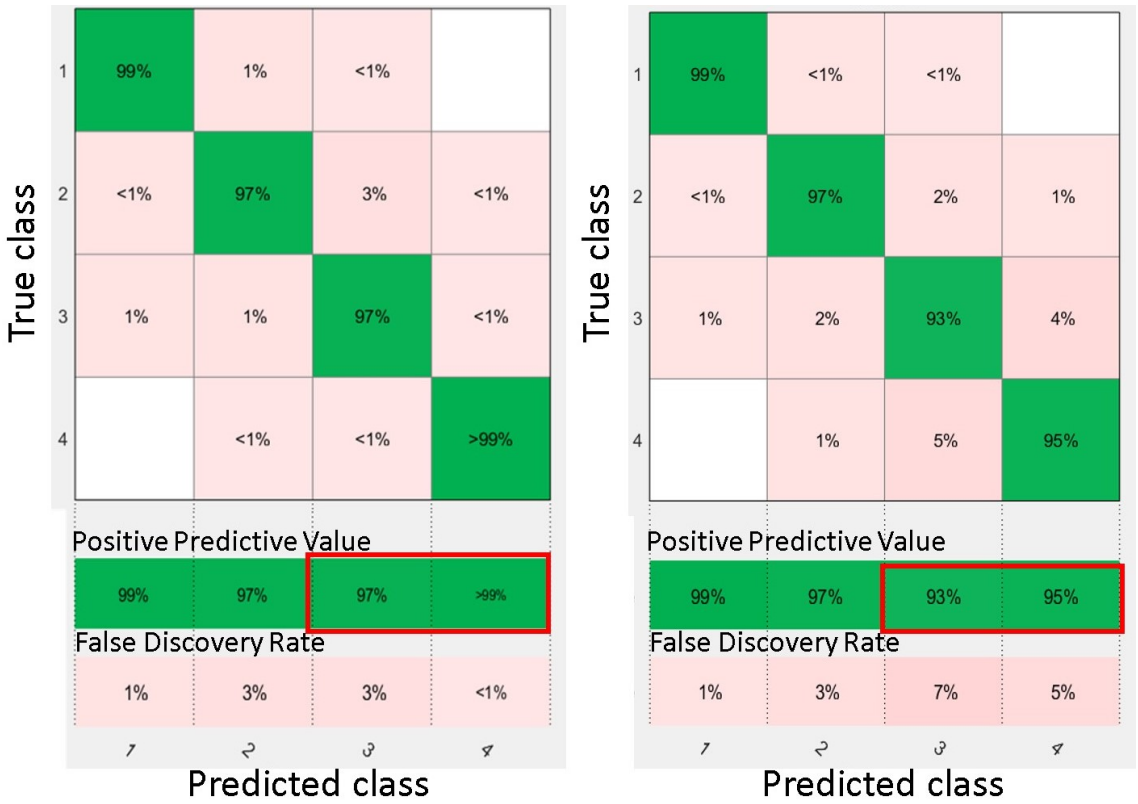
Spectral peak/ cm^{-1}	Band assignment	Healthy	Hyperplasia	Dysplasia	Cancer
1032	ν CC and CO of DNA, collagen, and glycogen	✓	✓	×	×
1057	ν CO of deoxyribose	×	✓	✓	✓
1192	ν_{as} PO_2^- of nucleic acids and collagen	✓	×	×	×
1209	ν_{as} PO_2^- of nucleic acids	✓	✓	×	×

4.34) of the supervised machine learning classifier where the true positive rate (TPR) of dysplasia and cancer notably decreases at the highest humidity. The spectral differences governing the categorisation of samples were previously recognised in table 4.4. Here, important spectral features are identified from the PCA loading plot (Fig. 4.33 (b, d)). A close inspection of the first two PCs shows that most of the prominent 'spectral biomarkers' are identified to be the spectral peaks that have been pointed out before at 1032 cm^{-1}, 1057 cm^{-1}, 1076 cm^{-1}, 1078 cm^{-1}, 1117 cm^{-1}, and also the ratio of the peak intensities between 1182 - 1140 cm^{-1}. The spectral peaks that make up most of the variance in these PCs do not vary much with tissue hydration.

Chemometric analysis via RF classification model utilising just the phosphodiester region (1300 – 1000 cm^{-1}) is effective at tissue classification. Similar approach was employed here to investigate the effect of tissue hydration on the model performance. The confusion matrices and the overall TPR (or test sensitivity) is depicted in Fig. 4.34. Remarkably, the model performs the best at the lowest humidity at 16 %RH, followed by a slightly lower score of TPR at the highest humidity at 88 %RH. Between these two extremes, TPR increases as RH increases. A plausible explanation for the good performance is the better deconvolution of overlapping peaks when tissue is hydrated. On the contrary, when the tissue is de-hydrated, some peak intensities, like the band at 1261 cm^{-1} significantly increases, which could possibly improve the distinction between healthy and malignant samples. The changes of the FTIR spectra has been discussed in the previous section when looking at the second derivative spectra; the combination of all these factors give the highest TPR value at the lowest humidity. Hence, it infers that for this



(a)



(b)

Figure 4.34: (a) The overall true positive rate of the RF classifier based on the spectral region between 1300 – 1000 cm^{-1} at different RH levels, with (b) the corresponding confusion matrices at the lowest and highest humidity, shown on the left and right respectively. The major differences can be seen in the classification of the 'Dysplasia' and 'Cancer', as highlighted with the red boxes

experimental study concerning disease classification with biopsy specimens using FTIR microscopy, the experiment should ideally be run at the lowest RH level possible. Despite the good performance of the supervised classification model at high humidity (only ∼2% lower than that of the best performing model), high hydration is not encouraged as the condensing salt and moisture might pit the surface of the objective of the microscopes and other components that are vulnerable to destruction when subjected to constant exposure to high water vapor content. Thus, the high RH controlled environment is best avoided when possible. At typical ambient condition of 20 °C with humidity at 40 – 45 %RH, the sensitivity can be estimated to be ∼92.5 %, in accordance to the results presented in Section 4.3. This gives prominence to the control of humidity in experiments especially when interpreting the changes that occur in tissues of varying degree of malignancy based on the phosphodiester bands since a small change in RH could lead to substantial changes in the absorbance of the important DNA bands.

4.4.5 Summary

FTIR spectroscopy has been demonstrated as a useful approach in monitoring the hydration of the tissue. It is shown here that the relative humidity conditions had significant impact on the spectra of the colon tissues. FTIR imaging provides appreciation that water is not uniformly absorbed into the tissue, and the degree of water intake might rely on several parameters, such as the tissue types and disease stages. It is observed that water is mostly absorbed into the areas of the tissues with a high absorbance in the phosphate and lipid regions. Based on the observed dependency of spectral bands between 1300 – 1000 cm^{-1} on hydration, insight into the vibrational modes of the phosphate group of nucleic acids was revealed. This led to the reasoning that DNA appears in its A-form in abundance in de-paraffinised fixed tissues. Dehydration of DNA molecules resulted in the transition of DNA to its A-form (Zuccheri & Samorì 2002) (its crystal form confirmed by circular dichroism (Hall et al. 2014)). This premise poses another issue to the already challenging clinical translation of ex-vivo FTIR spectroscopic measurements of biopsy tissues because most DNA in its natural form exists in the B-form. Further analysis with supervised machine learning RF models, coupled with PCA dimensionality reduction technique, strengthens our initial findings and supports that supervised machine learning model with the phosphodiester bands are adequate for the effective categorisation of diseased colons from healthy ones. Furthermore, the diagnostic performance of the FTIR spectroscopy varies with the level of hydration of the samples. It is evident that the diagnostic accuracy shows significant enhancement at both extremes – hydrated and de-hydrated states. This discovery presents an experimental approach to improve the sensitivity of FTIR spectroscopy for diagnostics by controlling the surrounding humidity, without recourse to altering the machine learning model outlined in Section 4.3, sample size, or other statistical parameters to achieve better diagnostic performance.

4.5 Depth profiling of prostate tissues by micro ATR-FTIR imaging

The approach of acquiring 3-dimensional (3D) imaging information by using variable angle macro ATR-FTIR accessory for depth profiling has been demonstrated in the past (Chan & Kazarian 2007*a*), but similar application with micro ATR-FTIR imaging in both qualitative and quantitative studies is very limited. The only paper describing variable angle micro ATR-FTIR spectroscopic imaging was that by (Wrobel et al. 2015); however, that work was on the investigation of polymer laminate samples, i.e. PS, PMMA, and PDMS polymer films only, to demonstrate the approach. Importantly, that results gave an insight on the potential to extend this work to study the structure and morphology of tissue which may be employed for disease identification. This section focuses on the use of variable angle micro ATR-FTIR spectroscopic imaging technique to study the variability of the chemical content of prostate tissues across various probing depths. It should be realised that the methodology used in this study is not the same as constructing a 3D model from a stack of 2-dimensional (2D) slices of images of microtomed thin layers of samples in transmission mode, as it was done by varying the angle of incidence alone and multiple microtoming was not required.

4.5.1 Experimental set-up and design of the apertures

Prostate tissue samples were prepared, as described in Section 3.2. The experiments were carried out with a Cary 620 FTIR microscope coupled to Agilent Varian 670 Spectrometer (Agilent Technologies, Inc.). The microscope was equipped with a slide-on ATR accessory attached to the 15×-cassegrain objective of the microscope, which enabled different apertures to be introduced for the study of controlled angles of incidence. The IRE for ATR in this study was a Ge crystal with a refractive index of 4. A liquid nitrogen cooled 64 × 64-pixel FPA detector, which has a FOV of 70 × 70 μm^2, was used for simultaneous acquisition of the IR spectral data. All measurements were taken in the mid IR range from 3900 cm^{-1} to 900 cm^{-1}, at 8 cm^{-1} spectral resolution and with 256 coadded scans. The data obtained were processed and constructed into 2D- and 3D- chemical models with Matlab R2019b (The MathWorks, Inc. 2020*b*). To warrant the accuracy of the image obtained at different probing depths, a 0.01 g precision digital scale was placed beneath the sample to ensure uniform contact pressure was applied throughout the study at 5.00 ± 0.05 g [4]. Reproducible pressure with the samples was identified to be crucial for ATR-FTIR spectroscopic experiments (Lu et al. 2017).

[4] In this experiment, consistent pressure on tissues was maintained so that the probing depth can truly represent the tissue sections measured without over-compressing the samples on one measurement compared to another. The actual pressure value used was not important.

Laser-cut full circle apertures, with circular slit width kept consistent at 3 mm, were used to introduce multiple depths of penetration into the sample probed in this study. The light beam is circular and hence a circular slit within the aperture enables the maximum amount of light to pass through at the selected angle and be collected by the detector. The apertures were created with varying distance from the centre. The set-up of the aperture is shown in Fig. 4.35 below. The material used as the aperture was black opaque acrylic with a thickness of 2 mm. Seven apertures (A1 – 7) were created in addition to measurement without the aperture (A0), covering various incident angles of the IR beam, as summarized in Table 4.5.

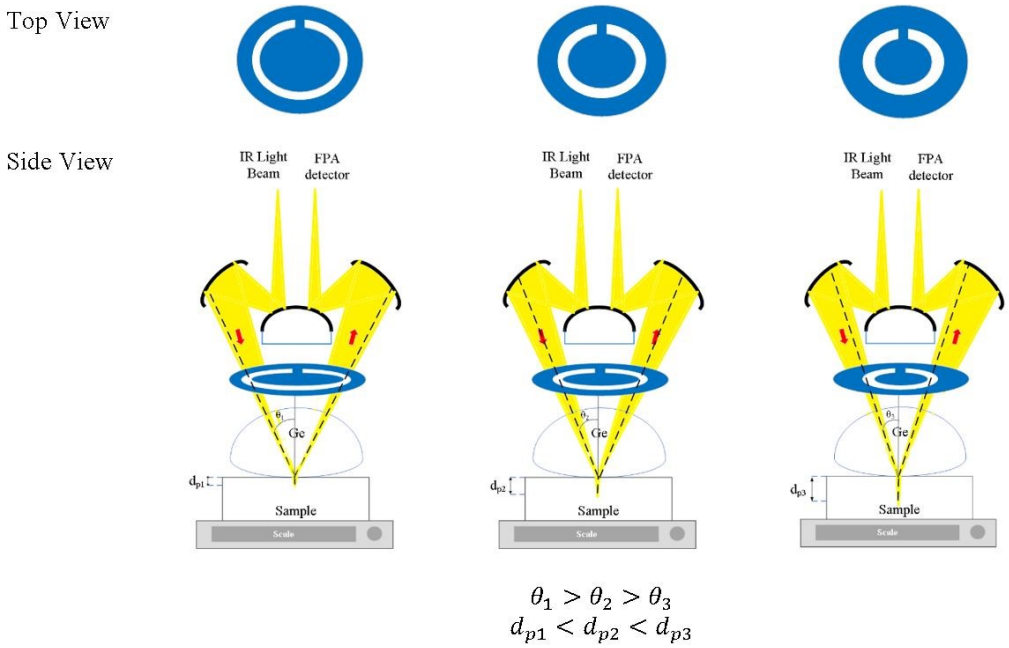


Figure 4.35: Schematic depiction of the insertion of full-circle apertures into the slide-on Ge ATR accessory for probing sample at different penetration depths (d_p) by setting various angles of incidence (θ) of the IR light beam. The illustration is not drawn to scale

4.5.2 Calibration of the angles of incidence from the measured effective thickness

The concept of effective thickness, d_e, was applied to estimate the angles of incidence of the IR light beam with apertures. This quantity is dependent on the polarization of the light beam, angle of incidence and the refractive indices of IRE and the sample. The d_e represents the thickness of a sample that would result in the same absorbance in a hypothetical transmission experiment, given in Eq. 3.6 and Eq. 3.7 in Section 3.1.5 (Harrick & Carlson 1971, Harrick & du Pre 1966). It enables the straightforward application of Beer-Lambert's Law on the spectral data obtained by ATR measurements, which states that the absorbance of a spectral band is directly proportional to the d_e. In these equations, n_1 is the refractive index of the IRE, i.e. Ge ($n_1 = 4.00$), n_2 is the

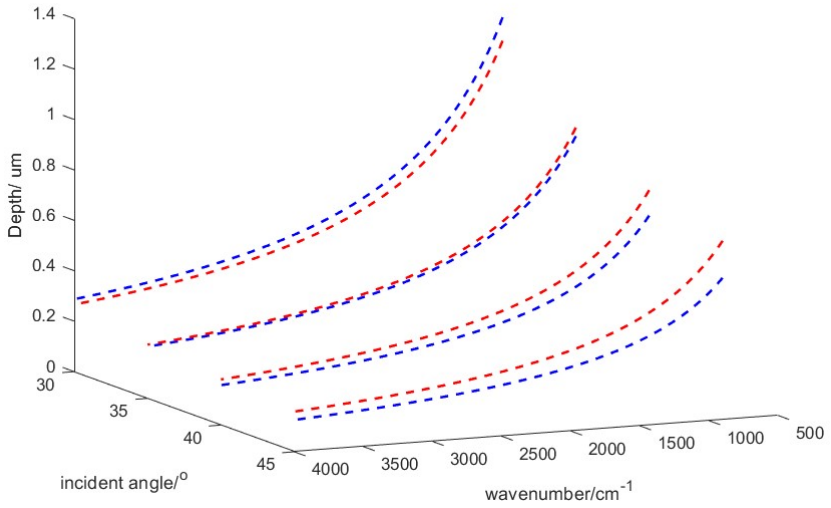
Table 4.5: Summary of the estimated incident angles, aspect ratios of the chemical images, and the range of effective thicknesses measured with different apertures for the prostate tissue samples

Aperture	Estimated $\theta/^\circ$	Estimated d_e @900 cm^{-1}/ μm ($n_2 = 1.45$)	Estimated d_e @4000 cm^{-1}/ μm ($n_2 = 1.45$)
A1	30.7	1.494	0.336
A2	32.0	1.340	0.301
A3	33.0	1.242	0.279
A4	34.6	1.112	0.250
A5	36.5	0.975	0.222
A0	38.3	0.893	0.201
A6	41.3	0.766	0.173
A7	41.8	0.744	0.172

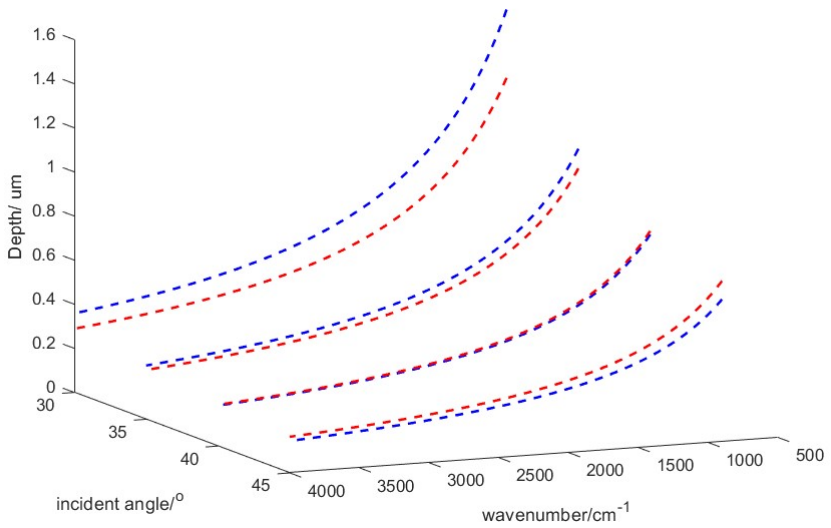
refractive index of the sample, i.e. water ($n_2 = 1.33$) and prostate tissue ($n_2 \approx 1.45$).

Although Eq. 3.6 – Eq. 3.8 become the standard for calculation of d_e in ATR measurement, they fail to account for the effect of anomalous dispersion on n_2, that is prevalent for strongly absorbing samples. Because of this assumption that n_2 remains constant, Harrick's equations have limitations on its applicability for non-absorbing or weakly absorbing bands only (Averett et al. 2008). Further methods to improve the accuracy of Harrick's equations were suggested and the inclusion of optical constant method to the calculation was found to be the best for strongly absorbing media in their study. Nonetheless, the above less cumbersome formalism was used for the approximate determination of incident angle in this experiment, as pointed out by (Fringelli 2000), when the spectral band of bending mode of water was investigated, the deviation between actual and approximate calculation fell within the experimental error, i.e. less than 3 %.

On the other hand, the penetration depth, d_p, as stated in Eq. 3.5, is essentially different from the d_e by definition. However, the value of d_p approaches that of d_e when $\frac{n_{21}cos\theta}{(1-{n_{21}}^2)}\left(1+\frac{2sin^2\theta-{n_{21}}^2}{(1+{n_{21}}^2)sin^2\theta-{n_{21}}^2}\right)$ is close to 1, independent of the wavelength of the illumination beam. Both quantities of depths are in the order of micrometer in micro ATR-FTIR measurement. d_e and d_p from samples of different refractive indices were compared. For water, these values converge at $\sim 40^\circ$; whilst for tissue specimen, this occurs at $\sim 35^\circ$, thus in our study where the angle falls between 35°– 40°, both Eq. 3.8 and Eq. 3.5 are similarly applicable. The angle where d_e and d_p is the same decreases when n_2 increases, shown in Fig. 4.36 – Fig. 4.37.



(a) Tissue sample (n_2=1.45)



(b) Liquid water sample (n_2=1.33)

Figure 4.36: Plot of depth as a function of angle of incidence and wavenumber of different samples, blue line is for d_p while red line is for d_e. With the same IRE (Ge), both values differ from each other and the difference increases with decreasing wavenumber, but the values converge at an angle depending on the refractive index of the sample investigated

4.5.3 Calibration of the angles of incidence for apertures using water as a sample

The IR absorption of liquid water has been thoroughly investigated through theories and experiments to date. The correlation between its absorption and structures, as well as its dynamics, was well studied (Robertson & Williams 1971, Downing & Williams 1975, Pinkley et al. 1977). Due to the extensive library of information available on the IR

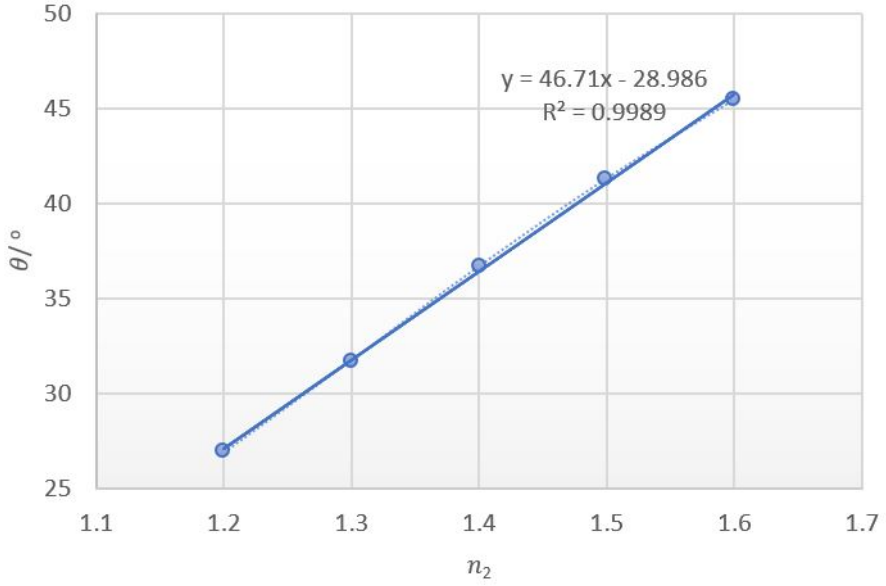


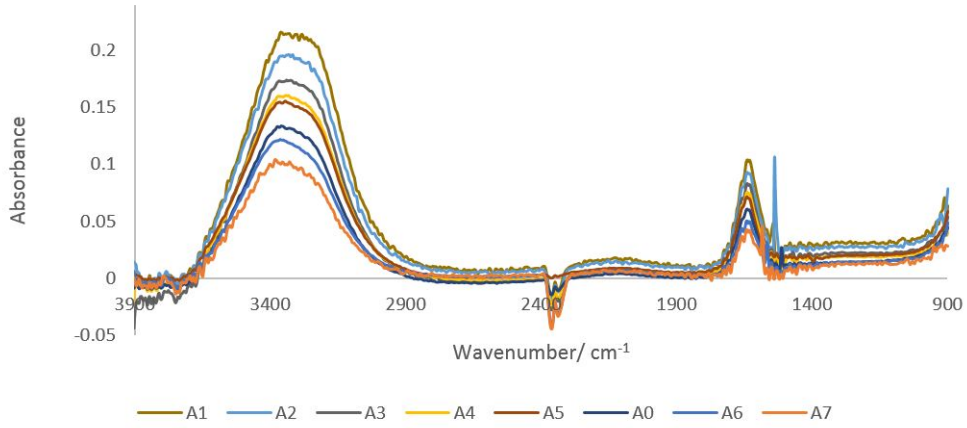
Figure 4.37: Plot of the angle of incidence where d_p is equal to d_e at various refractive indices of sample (n_2). A linear correlation can be inferred from the figure; as n_2 increases, the angle increases as well

spectrum of water, it was chosen as a reliable calibrating sample in this study. The incident angles were estimated from the peak absorbance of water from baseline at spectral band of $1643\,\mathrm{cm}^{-1}$ (assigned to the vibrational bending mode of water), $A_{water(1643)}$, with known $\varepsilon_{water} = 21.8\,\mathrm{M}^{-1}\mathrm{cm}^{-1}$ (Venyaminov & Prendergast 1997) and $c_{water} = 55.5$ M. From the Beer-Lambert's Law, d_e was calculated from $\frac{A_{water(1643)}}{\varepsilon_{water} c_{water}}$, which was then substituted into Eq. 3.6 - Eq. 3.8 for calculation of θ. The spectra of water obtained were shown in Fig. 4.38.

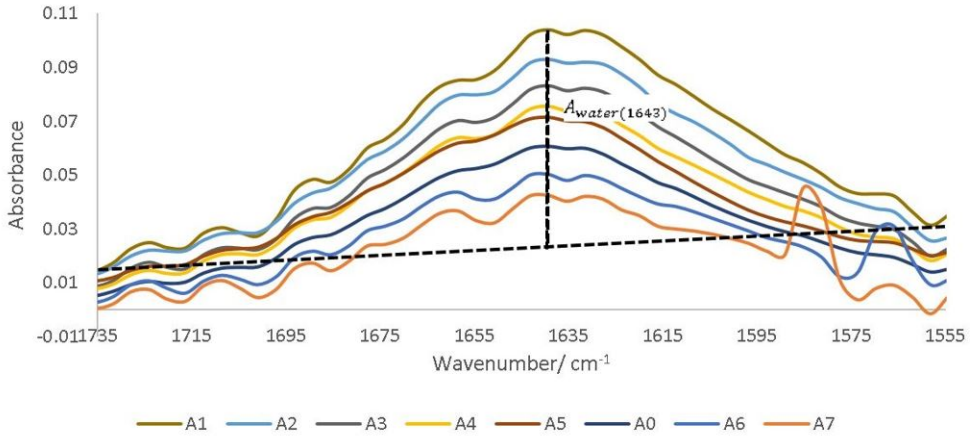
The angles of incidence (θ) were found to be 30.7°, 32.0°, 33.0°, 34.6°, 36.5°, 41.3°, and 41.8° for A1 – 7 respectively, while A0 was 38.3°. The latter value is consistent with the manufacturer's specification of the 15× objective used, given NA = 0.62 (in transmission) and the half angle of the light cone, $\theta_{objective} = sin^{-1}\frac{NA}{n_{air}} = 38.3°$. Aperture A7 has the largest radius, hence allowing light to reach the sample at the greatest angle, probing sample at its outermost (shallowest) layer closest to the surface. The radius of the aperture decreases from A7 to A1, measuring the sample at an increasing d_e.

The range of penetration depths covered between $900\,\mathrm{cm}^{-1}$ and $4000\,\mathrm{cm}^{-1}$ with various apertures created is shown in Fig. 4.39, with the assumption that the tissue had a uniform refractive index of 1.45 (taking the mean of the refractive indices of a biological tissue, which were typically reported to lie between 1.35 – 1.55 (Bolin et al. 1989)), albeit observations in previous studies that inhomogeneity in optical properties of diseased and non-diseased tissue prevails (Svensson et al. 2007, Wang et al. 2011)[5].

[5]The refractive index distribution reveals cellular and subcellular structures in transparent tissue slices,



(a) Spectral range between 3900 – 900 cm^{-1}



(b) Zoom in spectral range between 1735 – 1555 cm^{-1}(Peak absorbance at 1643 cm^{-1} for bending mode of water)

Figure 4.38: Mean micro ATR-FTIR absorption spectra of liquid water from all pixels within the FOV measured with apertures at various angles of incidence between (a) 3900 – 900 cm^{-1} and (b) 1735 – 1555 cm^{-1}. Absorbance at 1662 cm^{-1}, 1677 cm^{-1}, 1689 cm^{-1}, 1708 cm^{-1}, and 1724 cm^{-1} are attributed to the presence of spectral bands of water vapour (Ingle & Crouch 1988)

4.5.4 Micro ATR-FTIR spectroscopic images

Measurements of the prostate biopsy with variable angle of incidence have been performed with the Ge crystal in contact with the sample at a constant pressure. The areas of measurement are outlined in the visible images obtained under a 15× light microscope (Fig. 4.41). The distributions of the integrated absorbance at 1700 – 1600 cm^{-1}, assigned to amide I band (mostly attributed to the C=O stretching mode), were constructed as chem-

for example, the refractive index of sites of calcifications in breast cancer biopsies are different from the stroma (Zhuo et al. 2011), however, the refractive index was assumed to be constant in this study as the identification of the refractive index itself is a cancer biomarker, which can be used for cancer diagnosis.

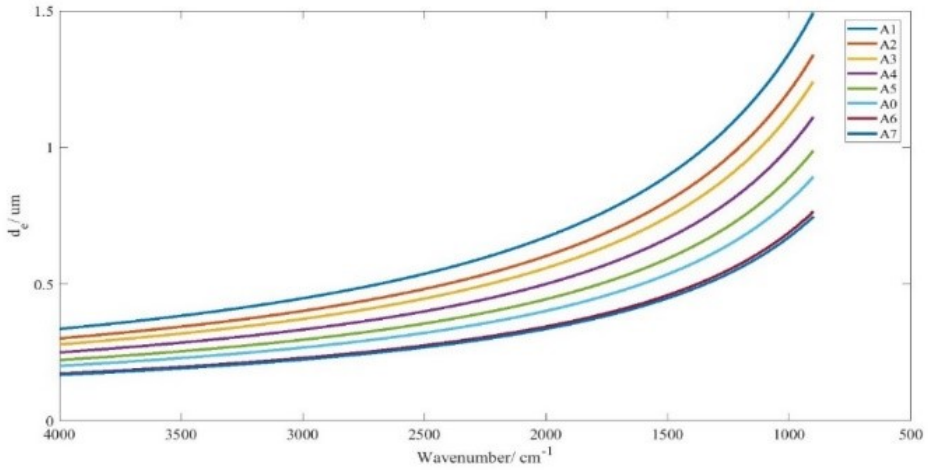


Figure 4.39: Effective thickness (d_e) values of non-polarized light calculated for all the apertures in the wavenumber range of 4000 – 900 cm^{-1} for tissues with refractive index approximated at 1.45

ical images and are shown in Fig. 4.42a (healthy tissue) and Fig. 4.42b (cancerous tissue) respectively. The area of strong amide I absorbance was associated with the cytoplasm of the cells which contains a high amount of proteins. At greater penetration depth, the absorbance of amide I band increases, creating an image of good contrast between tissue and lumen (areas of weak absorbance). This observation was expected as the absorbance is dependent on the sample thickness, which increases with a greater probing depth. The mean absorbance spectra of 3 × 3 binned tissue areas of strong amide I absorbance with each aperture are plotted in Fig. 4.40. The variation in the morphology of the tissue along the z-axis has also been successfully captured in the images obtained from different d_e. In Fig. 4.42b, a lumen that was embedded within the tissue appeared at the centre of the image for A6 and A0, which were not captured in either A7 nor A5; whilst in Fig. 4.42a, cells that are deep within the tissue layers were detected by depth-profiling the tissue block. These areas were circled in the figures for clarity.

The spectral band at 1235 cm^{-1} is attributed to the vibrational mode of nucleic acids, i.e. DNA and RNA ($\nu_{as}PO_2^-$) (Movasaghi et al. 2008). The distributions of the integrated absorbance at this wavenumber were constructed in Fig. 4.43a and Fig. 4.43b after normalisation of the amide I band across various angles, for healthy and cancerous tissue respectively. Chemical images collected with aperture A1 and A7 suffered from poor resolution. These apertures allowed incoming infrared beam at the edges, which carried fewer photons than the beam at the centre. The fewer photons that reached the detector resulted in spectra of low S/N ratio.

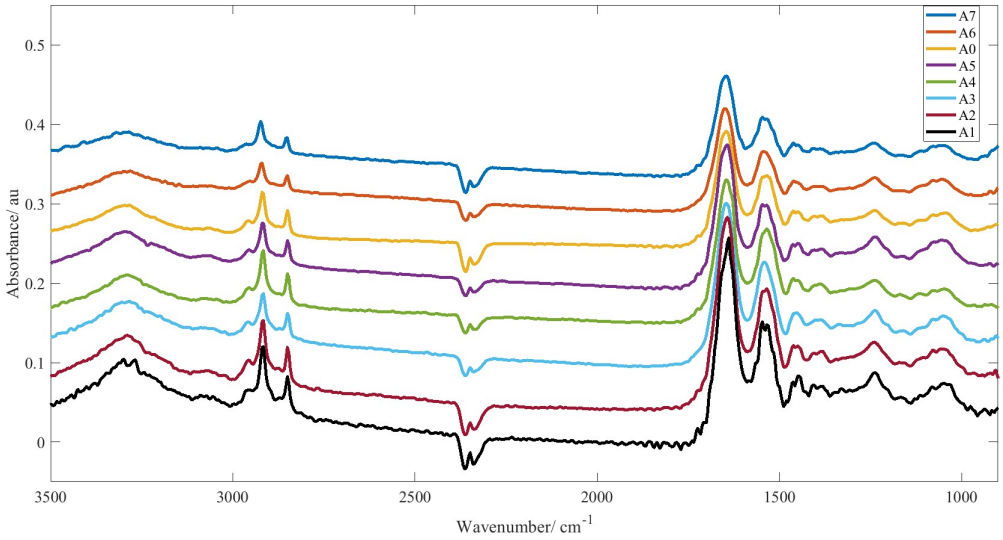


Figure 4.40: Average raw spectra obtained from a 3 × 3 binned tissue area (centre coordinates: x = 50; y = 40). For clarity and easier comparison, all spectra were plotted on the same scale with an offset of 0.05 from one another

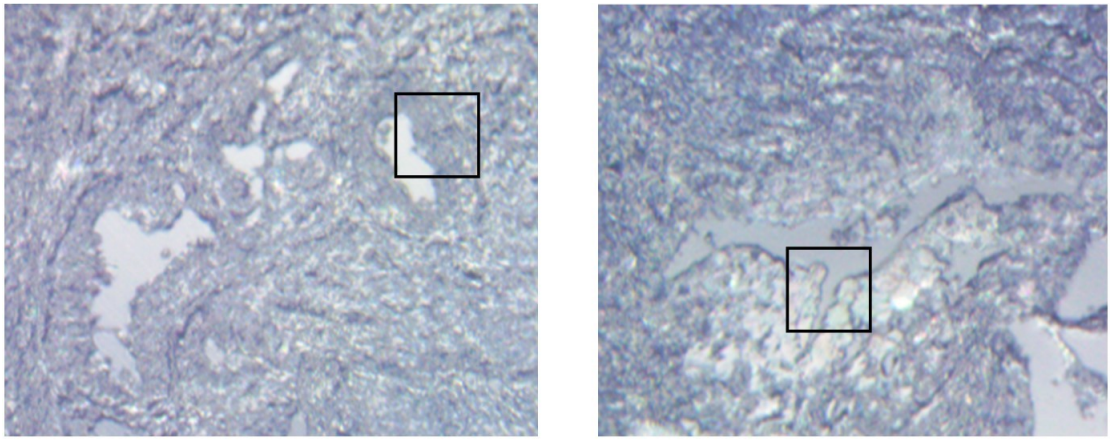
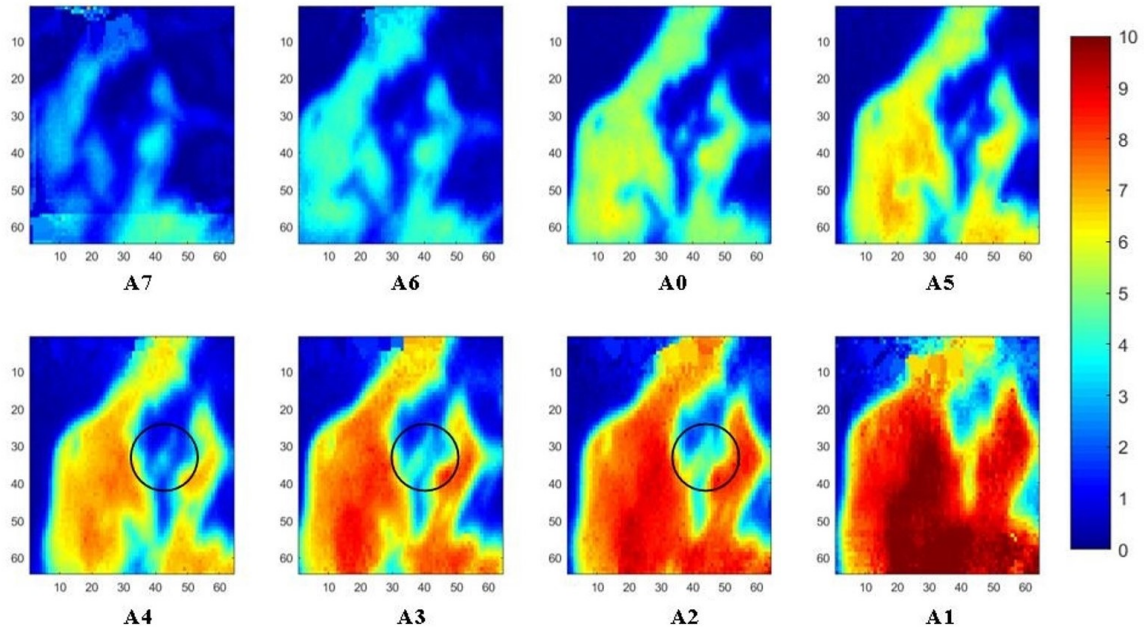


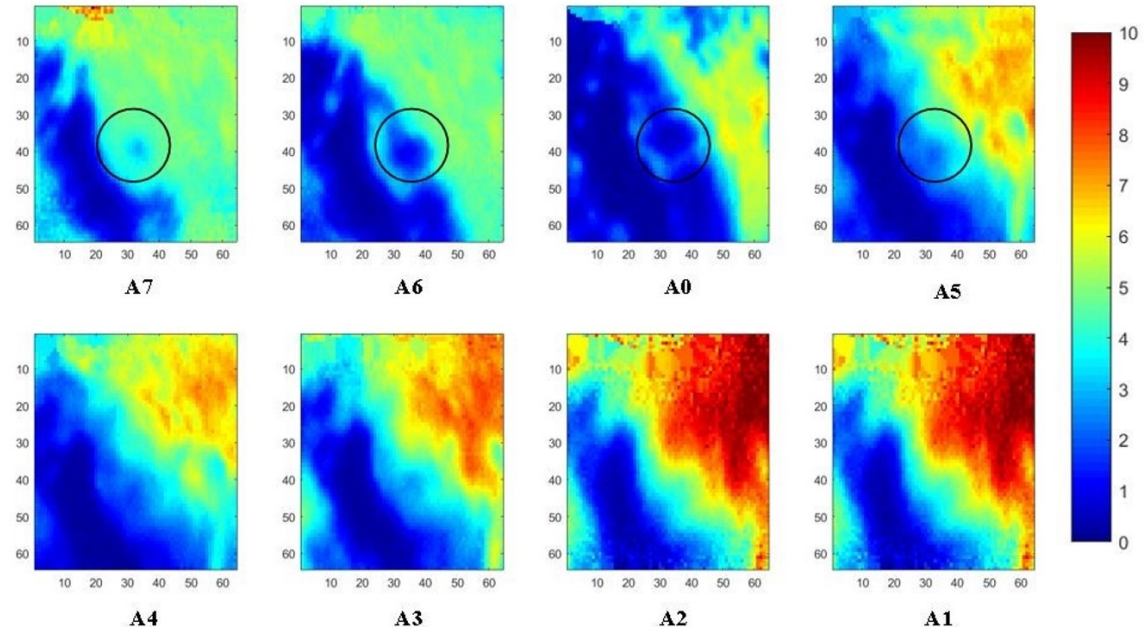
Figure 4.41: Visible images of the surface of cancerous and healthy prostate tissues taken under 15× light microscope in reflection mode, shown on left and right respectively. Areas inside the square boxes were imaged with ATR-FTIR microscope. The visible images have a size of 530 × 530 μm^2 and the box areas are 70 × 70 μm^2

4.5.5 Non-uniformity in spatial resolution

In theory, the spatial resolution, as stated by Rayleigh as the minimum distance between two adjacent points that is just resolved ($2r$), is limited by the wavelength (λ) and the numerical aperture (NA) of the system, given as $2r = \frac{1.22\lambda}{NA}$ (Rayleigh 2009). In ATR measurement, the light passes through Ge IRE with a refractive index of 4, thus increasing the NA of the system by 4 times. In practice, the actual spatial resolution was found to be $\sim 2\lambda$, i.e. between 5 – 20 μm for mid-IR (Bailey et al. 2016), lower than the value obtained from Rayleigh criterion. Measurements of the prostate tissues at variable angles



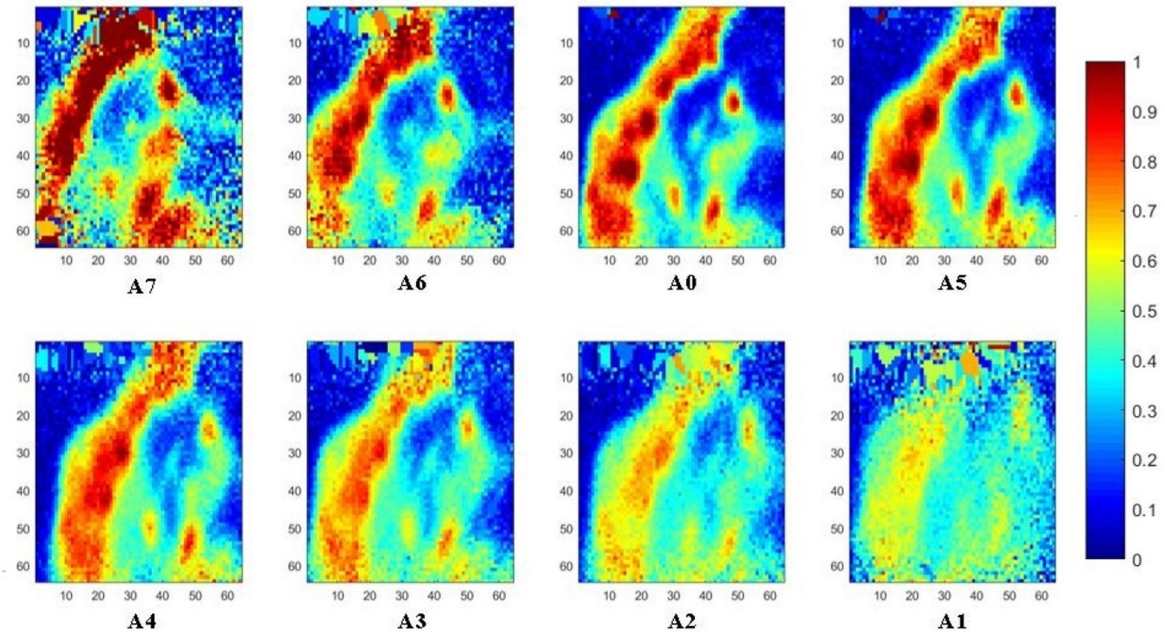
(a) Healthy prostate tissue



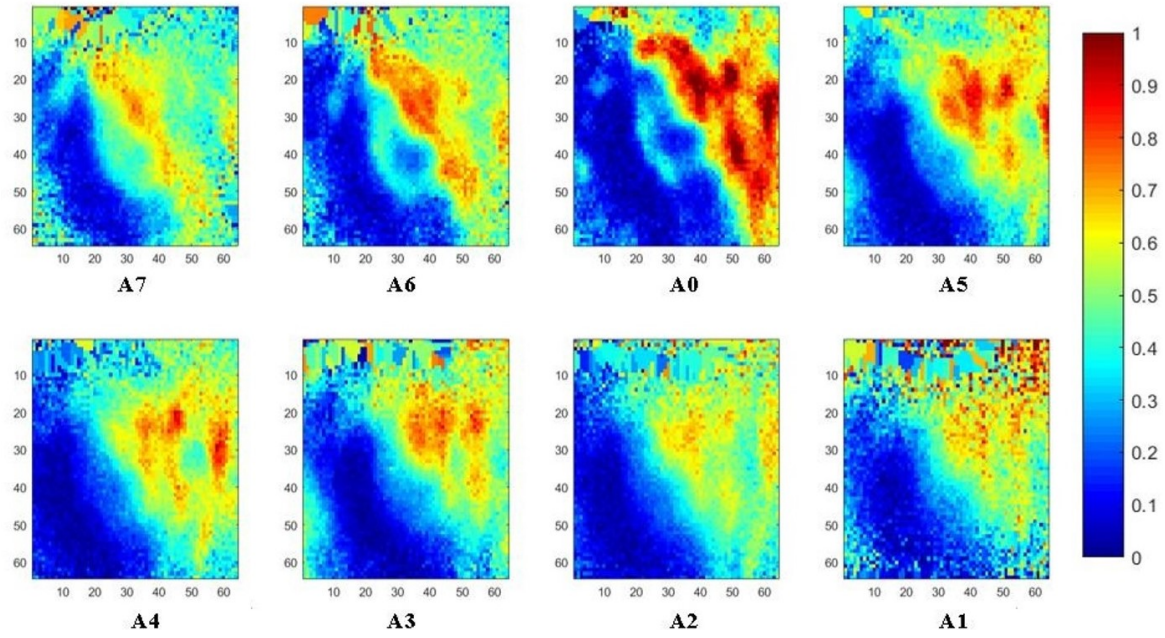
(b) Cancerous prostate tissue

Figure 4.42: Chemical images obtained with micro ATR-FTIR spectroscopic imaging. These images are based on the distribution of the integrated absorbance of the spectral band of amide I between 1700 and 1600 cm^{-1}. d_e = 0.41 μm (A7), 0.42 μm (A6), 0.49 μm (A0), 0.54 μm (A5), 0.61 μm (A4), 0.68 μm (A3), 0.73 μm (A2), and 0.81 μm (A1). Circled regions show the embedded component. The size of each image is 70 × 70 $\mu\mathrm{m}^2$

of incidence resulted in the non-uniformity of the spatial resolution of the 2D chemical images obtained in x- and y-direction, as expected in ATR-FTIR imaging study. A similar observation was noted for macro ATR-FTIR imaging with diamond (Chan & Kazarian 2003) and ZnSe reflection accessory (Nakamura et al. 2000) from the change in the aspect



(a) Healthy prostate tissue

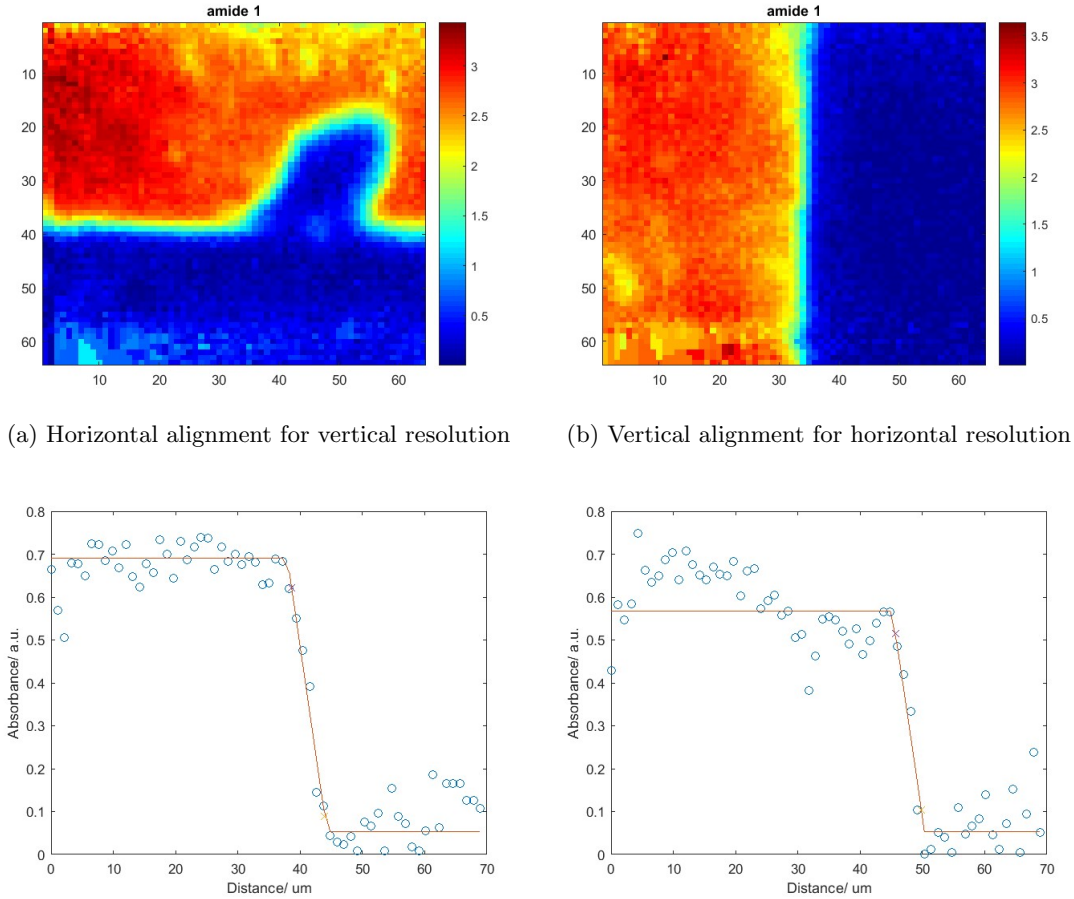


(b) Cancerous prostate tissue

Figure 4.43: Chemical images obtained with micro ATR-FTIR imaging. These images are based on the distribution of the integrated absorbance of the spectral band of $\nu_{as}PO_2^-$ between 1268 and 1200 cm^{-1}.d_e = 0.54 μm (A7), 0.56 μm (A6), 0.65 μm (A0), 0.72 μm (A5), 0.81 μm (A4), 0.90 μm (A3), 0.97 μm (A2), and 1.09 μm (A1). The size of each image is 70 $\times$ 70 $\mu\mathrm{m}^2$

ratio of the imaging areas. The actual vertical and horizontal spatial resolutions of the images in this work were estimated to be 7.3 μm and 6.2 μm respectively with the edge response method by measuring along a line across a sharp polyurethane/PMMA interface

at 1600 cm^{-1} (6.25 μm in wavelength), shown in Fig. 4.44 and Table 4.6, compared to 3.07 μm obtained when calculated with the Rayleigh criterion at the same wavelength. The observed disparity in vertical and horizontal spatial resolution arises from the non-uniform illumination of the objective, since only half of the objective is used to illuminate the sample while the other half allows light to be collected in the setting of the FTIR microscope (Kazarian & Chan 2010).



(a) Horizontal alignment for vertical resolution

(b) Vertical alignment for horizontal resolution

(c) Vertical distance (top to bottom)

(d) Horizontal distance (left to right)

Figure 4.44: (a,b) Micro ATR-FTIR spectroscopic images of a polyurethane/PMMA interface constructed at 1600 cm^{-1}; (c,d) Plot of absorbance vs distance along the dotted line. Drawn in red is the line of best fit

4.5.6 Differentiation of benign from cancer prostate tissue

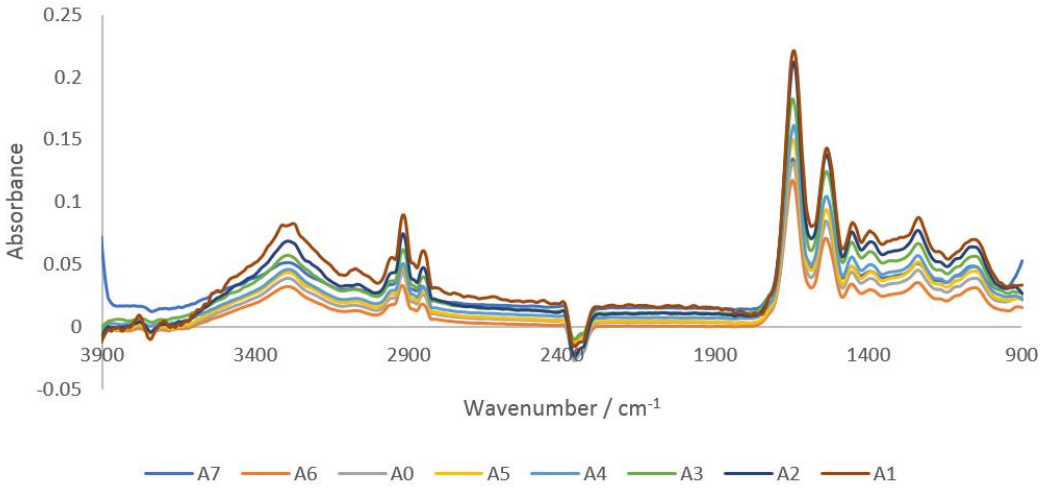
FTIR spectra of healthy and cancerous tissues were recorded for each angle of incidence with A0 to A7. The spectra were subjected to a S/N quality test whereby only spectra with good S/N ratio were retained. This pre-processing step eliminated the non-tissue spectra from interfering with the final spectra analysed, after taking the mean of all pixels representing the tissue only. The signal is the absorbance at 1700 – 1600 cm^{-1} (amide I)

Table 4.6: Calculation showing the distance or resolution. The values are different for vertical and horizontal resolution and is dependent on angle of incidence (38.3°, which can be approximated by taking the inverse cosine of the ratio of resolution in both directions. The resolution is calculated by taking the distance between 95 % and 5 % of the maximum absorbance

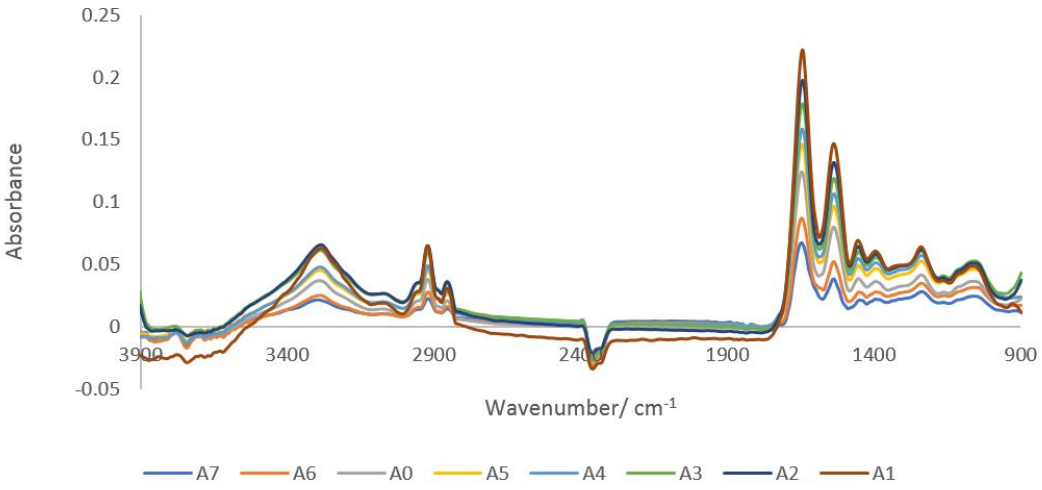
	Vertical resolution	Horizontal resolution
Maximum absorbance of line of best fit @ 1600 cm^{-1}	0.6882	0.558
Minimum absorbance of line of best fit @ 1600 cm^{-1}	0.2226	0.0530
Distance at 95 % of maximum value (μm)	38.62	44.66
Distance at 5 % of maximum value (μm)	45.94	50.84
Difference in distance (μm)	7.32	6.18

while the noise being the standard deviation of data between 2700 cm^{-1} and 2600 cm^{-1}. The spectra were then differentiated twice, to correct for baseline shift for the accurate comparison of the spectra. As noise was inevitably enhanced when the second derivatives were taken, SG smoothing algorithm with 13 smoothing points was implemented to obtain the smoothed second derivative spectra. The raw spectra and derived data for both tissue samples from the imaging datasets are presented in Fig. 4.45 – Fig. 4.46. The complexity of the spectra presented is typical of a biological sample (German et al. 2006). From the FTIR spectra, vibrational modes and the corresponding biological components were assigned to the main spectral bands for the interpretation of the results from comparing diseased and non-diseased tissue. In general, the fingerprint region $\sim$ 1340 – 950 cm^{-1} contains a series of complicated absorption of multiple spectral bands, mostly assigned to the functional groups in nucleic acids; $\sim$ 1700 – 1500 cm^{-1} to amides or proteins; whereas vibrational modes of lipid molecules (or alkyl chains) manifest themselves in the spectral bands that lie in the region of $\sim$ 3000 – 2800 cm^{-1}. The spectral interpretation was taken from reference (Movasaghi et al. 2008).

The analysis protocol performed for the differentiation of the datasets of different samples are similar to previous work on cell identification (Gaigneaux & Goormaghtigh 2013). To assess the significance of the statistical difference between the second derivatives of the spectra of the two tissues, student's t-test was performed at each wavenumber from 1800 – 950 cm^{-1}. Prior to the t-test analysis, the data were found to be normally distributed using the Jarque-Bera normality test at 5% significance level (Matlab R2017b) (Ghasemi & Zahediasl 2012). Paired t-test was suitable for the data of tissues of the same origin, i.e. from the same patient where the only difference is the diseased state. Fig.



(a) Cancerous prostate tissue

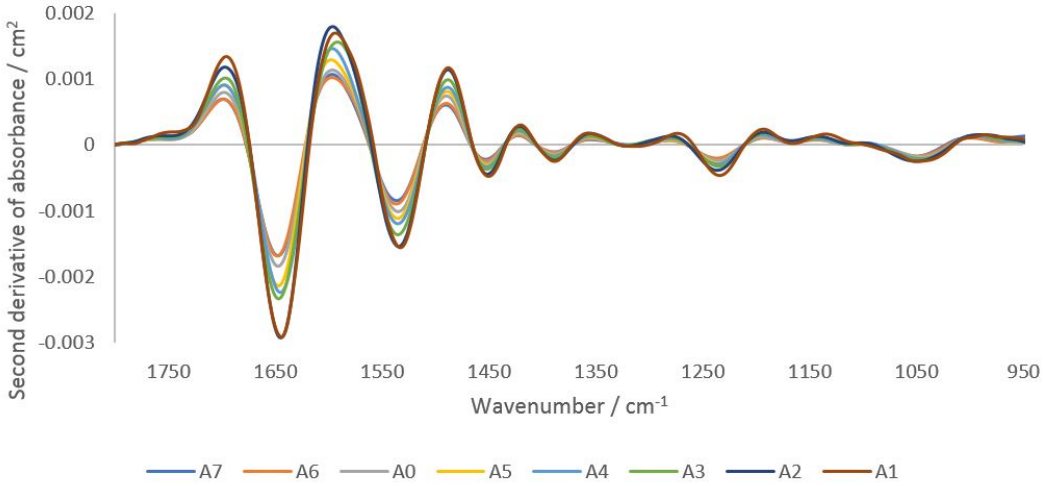


(b) Healthy prostate tissue

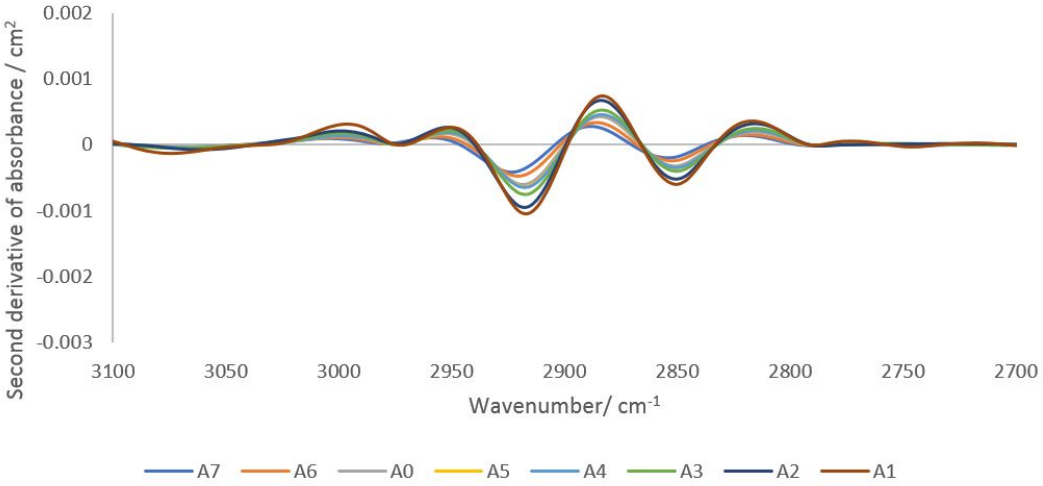
Figure 4.45: Mean ATR-FTIR spectra measured with different apertures in the 3900 – 900 cm^{-1} region

4.47 depicts the results returned by the statistical test. The wavenumbers where the null hypothesis was rejected at 1% significance level ($\alpha < 0.01$) were highlighted in red.

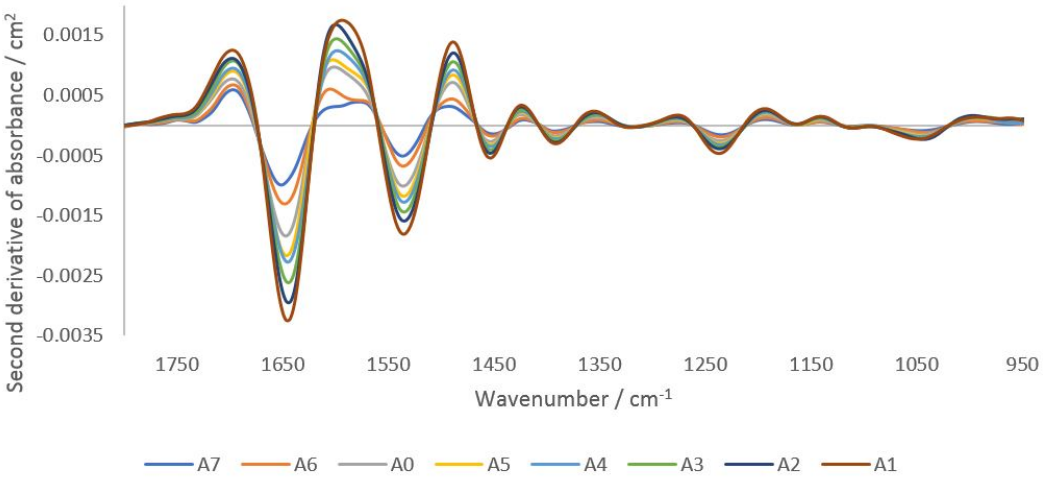
From the results, major differences are found to be in the fingerprint region (1500 – 1000 cm^{-1}) where spectral biomarkers for the identification of prostate cancer have been identified (Baker et al. 2008). The components at different probing depths vary from one another, thus the difference between each layer also varies accordingly. For instance, between 35° (A4) – 42° (A7) good discrimination of the spectral bands at 1087 – 995 cm^{-1} between the tissues was observed; however, decreasing the angle to between 31° (A1) – 33° (A3), the difference becomes significantly important. A mixture of overlapping bands of phospholipids, glycogen, and nucleic acids contributed to this difference (Movasaghi et al. 2008). On the contrary, at 1235 cm^{-1} and 1238 cm^{-1}, the significance of the dissimilarity



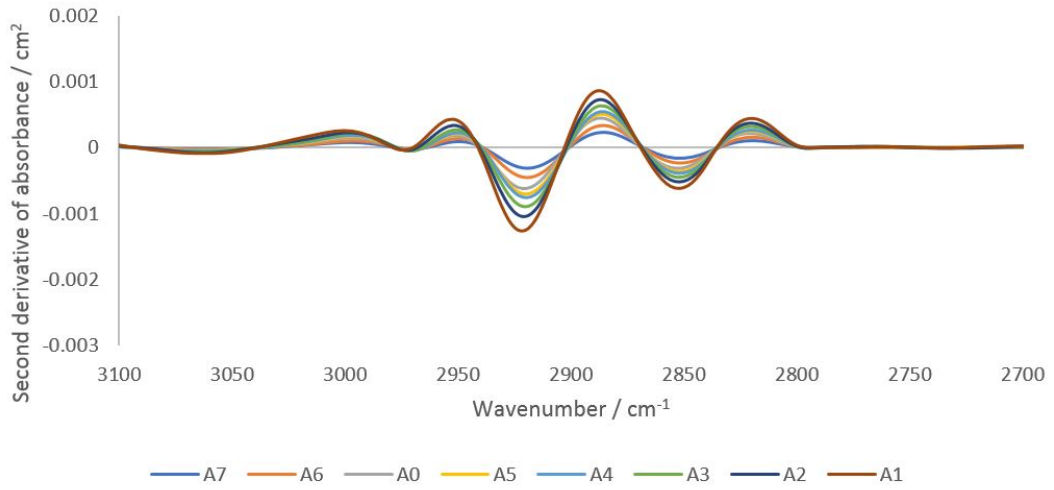
(a) Cancerous prostate tissue between 1800 – 950 cm^{-1}



(b) Cancerous prostate tissue between 3100 – 2700 cm^{-1}



(c) Healthy prostate tissue between 1800 – 950 cm^{-1}



(d) Healthy prostate tissue between 3100 – 2700 cm^{-1}

Figure 4.46: Mean second derivative spectra measured with different apertures

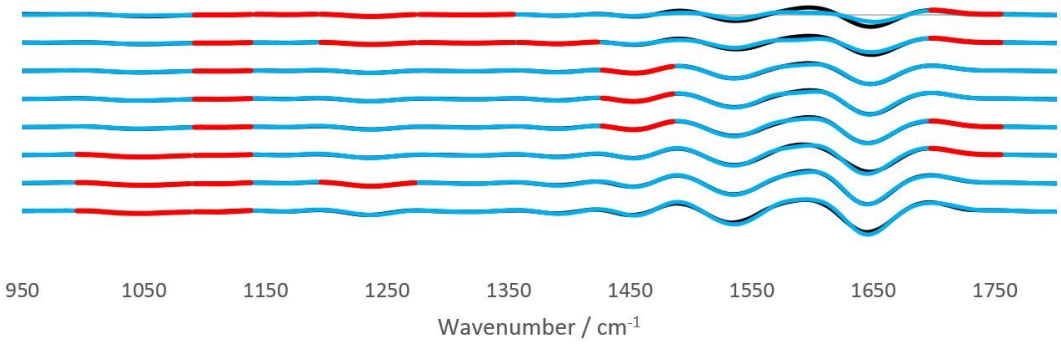


Figure 4.47: Significant differences from paired t-test analysis ($\alpha = 0.01$) highlighted in red on the second derivative spectra obtained at various angles of incidence, from 41.8° (top) to 30.7° (bottom). At each angle, the spectra were plotted on top of each other. The spectra of healthy and cancer tissues were given in black and blue lines respectively

diminishes as the penetration depth increases with decreasing angle. Both wavenumbers are assigned to amide III and nucleic acids band, which are present at great quantities at the surface of the tissue because of the way the tissue was microtomed and measured. Probing the shallow layers at the surface of the tissue via the insertion of apertures has provided the opportunity to measure the distinction between samples in the fingerprint region, as compared to the data obtained without aperture (A0), where the deviation in the amount of nucleic acids was not picked up at the set significance level mentioned before. Apart from that, the disparity in other components was also observed. Lipid band at 1736 cm^{-1} was found to vary significantly between the cancer and healthy tissues at intermittent thickness. Lipids such as fatty acids and sterols are stored in droplets in cells, which are particularly widespread in cytosol (Thiam et al. 2013), hence the difference presumably originated from the randomness of the amount of these particles found at

each layer. As for the spectral band at 1450 cm^{-1} (assigned to small carbon particles), the difference was picked up in the middle layers. Like lipid droplets, they are randomly distributed across the tissue. This spectral band has not been identified as potentially useful spectral biomarker for the detection of prostate cancer. No statistically significant difference was observed for the amide I and II bands for the whole range of penetration depths. In short, probing the shallow layers at the surface of the tissues was more efficient at distinguishing between the two tissues based on the results of the *t*-test.

Although the student's *t*-test was useful in providing a preliminary picture of the statistically significant bands and provided an insight into the effect of probing at different penetration depths, the variables (i.e. spectral bands) were in fact not completely independent of one another. PCA was used on top of the *t*-test to illustrate an overall discrepancy between the tissue specimens in the fingerprint and amide regions (between 1800 – 900 cm^{-1}). The PCA score plot of each angle is shown in Fig. 4.49. The first 2 PCs explained $\sim$ 96 % of the total variance of the data. When the plot was projected on a 2D space along PC1 and PC2, the data spread out along PC1. The variance explained by PC1 for different apertures was as follows: $PC1_{A7}$ = 95 %, $PC1_{A6}$ = 92 %, $PC1_{A0}$ = 91 %, $PC1_{A5}$ = 91 %, $PC1_{A4}$ = 97 %, $PC1_{A3}$ = 91 %, $PC1_{A2}$ = 91 %, $PC1_{A1}$ = 88 %. Between 30° (A1) and 35° (A4), this PC was aligned along the spectral band between 1080 – 1053 cm^{-1} with a minor contribution from the spectral band at 1235 cm^{-1}. The former was attributed to the symmetric vibration (ν_s) of $PO_2{}^-$, while the latter to the asymmetric vibration (ν_{as}) of the same molecule. Beyond 35°, the trend was reversed, 1235 cm^{-1} became the dominating spectral band responsible for the variance observed in the first PC, as depicted in Fig. 4.48 when the percentage variance of the contribution to the PC was plotted at each wavenumber investigated. As the probing depth increased, the distribution of data that discriminated cancer from benign tissue shifted from ν_{as} $PO_2{}^-$ to ν_s $PO_2{}^-$. This observation was consistent with the *t*-test analysis.

The effect of the penetration depth on the discrimination of spectra obtained from cancerous tissue from that of healthy prostate tissue in micro ATR-FTIR measurements were studied. The best separation is obtained when probing at the surface of the tissue where ν_{as} $PO_2{}^-$ is the dominating factor in PC1. At 41.8° (A7), the two clusters of data, plotted in different colours, were distinguishable from each other. A reasonably good separation was obtained at 30° (A1) with slight overlapping of the data; however, superposition is mainly observed when measurements were taken with other apertures. Following PCA, the classification of the data of the first 2 PCs was further evaluated by fitting them to linear discriminant analysis (LDA) classifier in Matlab – a supervised classification model with a 10-fold cross validation (The MathWorks, Inc. 2020*b*). The predictive performance of the model was assessed through classification loss, which is the mean squared error between the predicted value and the true response, in this case the diseased and non-diseased tissue. The cross-validation loss for each aperture is reported

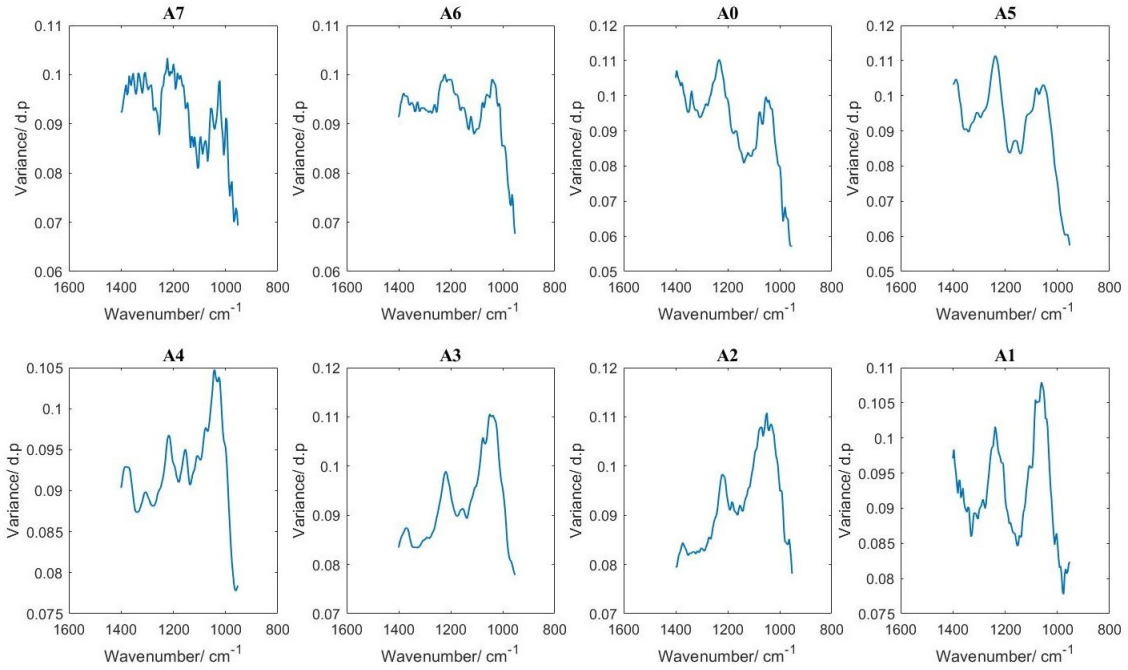


Figure 4.48: Percentage variance at each wavenumber from 1500 – 850 cm^{-1} along which PC1 is aligned. Higher variance indicates a dominating wavenumber for the PC. In the top four plots, 1235 cm^{-1} has the highest variance, whereas for the bottom four, 1062 cm^{-1} is the dominating band

in Fig. 4.49. The lower the value of the loss, the more accurate the LDA model, in other words, the better the separation of the data. The results showed that A7 had the lowest loss at 0.1430, followed by A1 at 0.2318 and A6 at 0.2749. The results from LDA corresponded well to the visual observation of the degree of separation of the PCA. Therefore, it is postulated that $\nu_{as}\ PO_2{}^-$ is a better spectral biomarker for distinguishing cancer tissue than $\nu_s\ PO_2{}^-$.

The findings here, that probing tissues at a greater angle of incidence offered a more reliable differentiation between benign and cancerous tissue compared to probing the thicker layers of the sample, were significant. At the same time, there were some limitations on this study such as the limited number of samples available and the observations for different types of tissue may not be the same. The results may also differ based on the way the tissue is microtomed or the configuration of the components within the tissue. However, this work demonstrated the feasibility of the assessment of the tissue across various effective thicknesses, hence offering a higher possibility of recognising the embedded component or detecting spectral differences which might not be significant when probing at just one depth.

4.5.7 Summary

This section demonstrates a non-destructive, label-free approach of examining heterogeneous biological samples in the z-direction using micro ATR-FTIR imaging, in a qualitative and semi-quantitative studies. The depth of penetration and effective thickness in ATR-FTIR spectroscopic imaging are dependent on the wavelength and the angle of incidence of incoming light beam. It was demonstrated here, for the first time, that variable angle micro ATR-FTIR spectroscopic imaging, which was created via the insertion of circular apertures, was useful at examining the embedded components within a prostate tissue specimen. This was done by constructing a 3D model from the stacks of 2D chemical images obtained, each of which represented the spatial distribution of a chosen spectral band assigned to the component of interest at a different probing depth. Micro ATR-FTIR imaging was also shown to have the ability to resolve subcellular components of cells such as organelles. For differentiation of diseased and non-diseased tissue, statistical tests were employed to analyse the spectral datasets obtained. When the second derivative of the spectral datasets were subjected to t-test analysis, the spectral differences between both samples in the fingerprint region were shown to be more significant at shallow depth of penetration; with the greatest variance at the spectral band of 1235 cm^{-1} (ν_{as} $PO_2{}^-$), depicted by plotting the scores of PCA on its first two PCs.

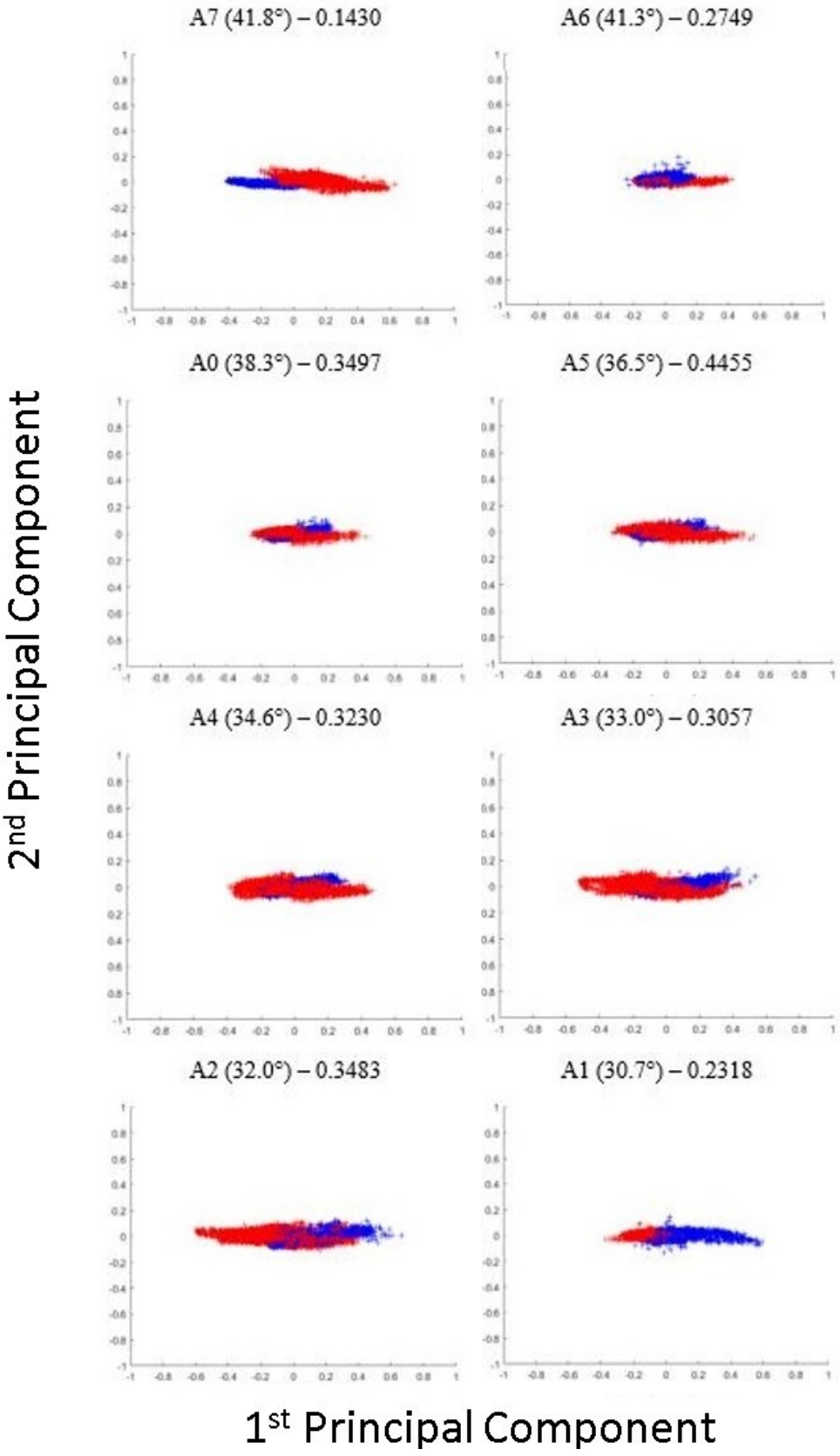


Figure 4.49: PCA score plot of second derivative spectra (red = healthy tissue; blue = cancerous tissue) in the spectral range $1400 - 950\,\mathrm{cm}^{-1}$ at different angles from $\sim 30° - 42°$, projected on the 1^{st} and 2^{nd} principal components. The incident angle and corresponding classification loss of LDA classifier is given on top of each score plot for each aperture

4.6 Mapping of colon tissues in micro ATR-FTIR imaging mode

Although ATR is today one of the most widely used FTIR sampling tool, there are several challenges associated with imaging using commercially available ATR-FTIR microscope. First and foremost, the imaging area is limited at approximately 70 × 70 μm^2 with a 64 × 64-pixel FPA (at 15× magnification, Cary 620, Agilent Inc.). To be able to study a large area with micro-ATR thus requires the combination with mapping which proves to be difficult as a consistent pressure is required during these measurements to ensure perfect stitching of the images altogether, as well as the possibility of leaving impression or causing damage to the sample due to the multiple contacts made in the process of mapping. This is particularly undesirable in the imaging of soft biopsy tissues. On top of that, the spectra obtained tend to suffer from poor S/N ratio which is often compensated by increasing the number of scans and subsequently, the acquisition time increases, thereby reducing the efficiency of the diagnostic process. A demonstration of imaging with mapping of large areas achieved in macro ATR-FTIR spectroscopic imaging using a large Ge inverted prism to obtain the chemical images of human fingerprints was reported (Chan & Kazarian 2008); however macro-imaging has a different set-up compared to micro ATR-FTIR imaging approach discussed here. The spatial resolution of macro-imaging compared to micro- is also worse at ca. 10 μm.

Recently, the use of a large IRE crystal in combination with the FTIR microscope has been shown to improve the FTIR spectroscopic images acquired in ATR mode in polymers and pharmaceuticals (Patterson & Havrilla 2006, Patterson et al. 2007). However, the applicability of the methodology on the softer biological specimens was not studied until lately where cell and brain tissues using a large crystal set-up was investigated (Vongsvivut et al. 2019). Despite the success in achieving chemical images of a high quality, their work required the use of synchrotron IR source, thus there are still areas for continued development and the investigation of the potential of the large crystal in ATR imaging. In this study, as part of the effort to improve and overcome the limitation of the micro ATR-FTIR systems in the market, a novel configuration to improve micro ATR-FTIR imaging was devised and the suitability of the new hemispherical large radius Ge crystal, coupled with a conventional bench top FTIR microscope of globar source, to study a large area of a soft biological material by mapping with FPA imaging was explored for the first time.

4.6.1 Experimental set-up of 'large area' Ge crystal

Colon tissue samples were prepared, as described in Section 3.2. In this experiment, the specimens were subjected to ATR-FTIR measurement with a Cary 620 FTIR microscope

coupled to Varian 670 spectrometer (Agilent Technologies, Inc.). Simultaneous acquisition of the FTIR spectra was carried out with a liquid nitrogen cooled FPA detectors consisting of 64 × 64 pixels. The FOV of a single FTIR spectroscopic image is 70 × 70 μm^2 at 15× magnification (NA = 0.62). The average angle of incidence of the focused IR beam coming from a globar light source within the spectrometer is 38.3° (Song & Kazarian 2019*b*). Spectra were recorded in the range of 3900 – 900 cm^{-1} with spectral resolution of 4 cm^{-1} and 64 co-added scans. The depth of penetration varies from $\sim$ 0.2 μm at the high wavenumber (4000 cm^{-1}) to $\sim$ 0.9 μm at the low wavenumber side (900 cm^{-1}) with the refractive index of the tissue estimated at 1.45. The acquisition parameters were the same as the one described in Section 4.5.1.

A novel approach to the imaging and mapping of a large area of samples using a newly designed hemispherical Ge crystal (PIKE Technologies) was employed. The Ge crystal, with a refractive index of 4, acted as a single reflection ATR accessory. The bottom surface of the crystal which was in contact with the specimen has a diameter of 2 mm; however, the useful area for measurement was limited to $\sim$ 420 × 420 μm^2, due to the refraction of light when the centre of the crystal is not vertically aligned with the objective, as in the case of mapping (Fig. 4.56). The Ge crystal was mounted on the microscope stage and was brought into focus. A specially created microscope stage for the adaptation of the crystal was used to allow the sample to be placed between the stage and crystal before the stage was raised, sandwiching the biopsy sample between the crystal and the stage, while maintaining the position of the crystal. The procedures were controlled and monitored using the Resolutions Pro software (Agilent Technologies, Inc.) which provides a real time FPA view. The set-up of the newly designed Ge (Fig. 4.50 (d)) has the added advantage which allows mapping of a large area of the sample to be carried out without the need for new contact made in-between measurements. The geometry of this crystal with a flatter surface was designed to provide good optical contact with the sample to be analysed without damaging the surface of the soft biopsy tissues. Unlike micro ATR-FTIR measurements where the crystal was fixed on the objective (Fig. 4.50 (a)), the stage and the Ge crystal were moved together as one for subsequent measurements of new areas in mapping mode. A single background spectrum was obtained at the beginning before the mapping measurements on any one sample were carried out.

4.6.2 Data Processing

The spectral data obtained were processed with Matlab R2019b (The MathWorks, Inc.). The selection of the processing steps was subjective (Lasch 2012), the data processing workflow in this study was described as follows. The non-tissue containing pixels were first eliminated by subjecting the FTIR imaging data to a S/N ratio test. The signal was the peak absorbance at 2921 cm^{-1}, while the noise was the standard deviation of data between

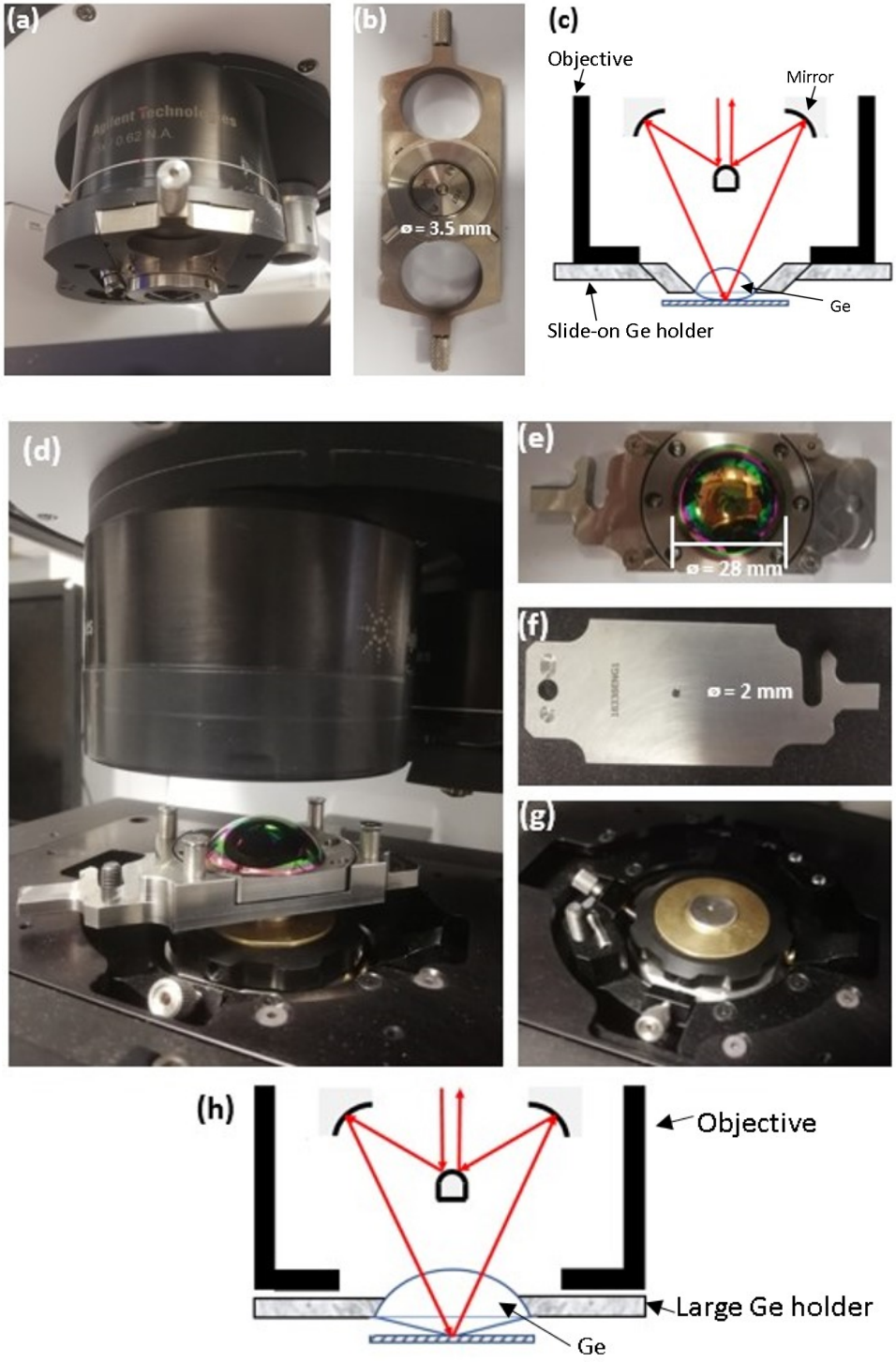


Figure 4.50: Schematic of the set-up of micro ATR-FTIR imaging system compared in this study. (a) small Ge crystal attached to the objective of the FTIR microscope; (b) slide-on small Ge accessory; (c) illustration of the IR beam path within the objective and the small crystal; (d) 'large area' Ge attached to the specially designed stage suitable for mapping; (e) the top view of the Ge crystal to allow more light to be focused on the sample; (f) the bottom view of the crystal which is in contact with the sample; (g) the specially design stage where the 'large area' crystal is screwed into place – the stage is raised by rotating the knob, and moved vertically and horizontally by turning the screws on both sides; and (h) IR beam path within the objective and the large area crystal

2810 cm^{-1} and 2760 cm^{-1}. The spectra which passed the quality test set at a threshold of 5% of the maximum S/N ratio were retained. The spectra were checked for water vapour correction based on its second derivative spectra in the 1900 – 1750 cm^{-1} region, following the procedures described by (Bruun et al. 2006) before they were subjected to SG second derivative smoothing (5-point third-order polynomial smoothing and derivative filter) (Savitzky & Golay 1964). New background measurement was impossible to obtain between each sampling areas, therefore, computational subtraction of the background was necessary. The distribution of the absorbance of a spectral band of interest was identified by analysis of the images, constructed by calculating the integrated absorbance within the spectral ranges of the spectra following the trapezoidal rule. After that, the spectral windows in the wavenumber range of 1700 – 1000 cm^{-1} and 3000 – 2800 cm^{-1} containing important information of a biological sample were selected for further analysis (Baker et al. 2018). PCA and PLS models with 5-fold cross validation were then applied on the vector-normalized second derivative dataset for the classification based on the state of malignancy of the biopsies. The results were compared with H&E images analysed by pathologists and the comparison was also made across measurements in ATR and transmission modes.

4.6.3 Enhanced performance of the ATR-FTIR spectroscopic mapping approach

Prior to actual sampling of the biopsy tissues, the added improvement of the large area crystal was investigated in comparison to the conventional slide-on ATR crystal. The parameter selected to assess the performance of the ATR microscope measurement is the sensitivity, as measured by its S/N ratio, using water (liquid) sample at 20 °C as a standard. For relative comparison of measurements with different set ups, the calculation of S/N ratio is simplified – the signal is taken as 100% with no sample in place and the noise is calculated peak-to-peak (PP) from 2000 – 1900 cm^{-1} of the water spectrum. This spectral region is free from the absorption bands of liquid water, which is found between $\sim$ 1700 – 1500 cm^{-1} (ν_s OH) and $\sim$ 3700 – 2800 cm^{-1} (ν_{as} OH) in an ATR spectrum (Marechal 2011). The average 1/PP noise ratio of varying number of co-added scans at 8, 16, 32, 64, 128, and 512 were plotted in Fig. 4.51. The ratio increases with the number of scans since the IR signal (S) is additive but the noise (N) is stochastic (Robinson et al. 2005); however, it reaches a plateau beyond 64 scans for measurement at a spectral resolution of 4 cm^{-1}. The 'large area' Ge gives spectra of higher quality than the slide-on Ge throughout (Song & Kazarian 2019*b*). With the other parameters such as the integration time kept constant, the S/N ratio is approximately 2× higher for large area crystal, likely because more photons are focused at the sample, consistent with the higher IR beam intensity recorded by the detector (see Fig. 4.52 for a plot of light intensity versus integration time of FPA). The increase in size of the large area Ge also warrants a more uniform distribution of signal intensity within the FOV of the FPA detector (Fig.

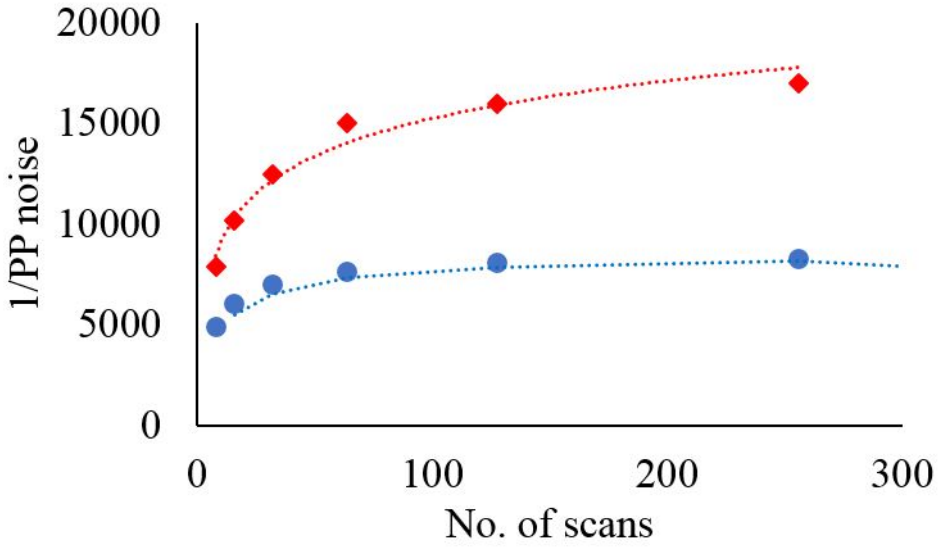


Figure 4.51: Signal-to-noise (S/N) ratio of both small slide-on and 'large area' Ge crystal as a function of the number of scans, represented by the blue and red dotted line respectively

4.53). Moreover, the new ATR design is especially suited for mapping of a large area, up to a total area of $\sim 420 \times 420$ μm^2. The limitations of the small Ge set-up in mapping and the impression left on the colon biopsies which were not observed with 'large area' Ge are illustrated in Fig. 4.54.

4.6.4 Spatial resolution as a function of mapping distance from centre of beam

Spatial resolution of the images taken with the 'large area' hemispherical Ge crystal was assessed along an interface between polyurethane and PMMA, following the method in reference (Kazarian & Chan 2010). The spatial resolution was approximated to be the distance between 95% and 5% of the absorbance of the spectral band at 1600 cm^{-1} (Sommer et al. 2001), measured along a straight line. Using an Agilent Cary 620 microscope, the vertical and horizontal resolution (i.e. resolution in y- and x-direction respectively) is slightly different; the former is worse than the latter at $\sim$ 24% difference (Song & Kazarian 2019*b*). The better horizontal resolution was investigated here and the average value is presented in Fig. 4.55. At the centre position where the crystal is vertically aligned with the infrared beam, the horizontal spatial resolution of the large crystal is ~ 6.55 μm, similar to slide-on small germanium crystal (Song & Kazarian 2019*b*). This new set-up improves the signal intensity but the spatial resolution has not improved. This is expected because the spatial resolution of optical imaging instrument is independent of light intensity; it is limited by the diffraction of light. However, when mapping with the large area is carried out, the spatial resolution is found to suffer with mapping distance from the centre

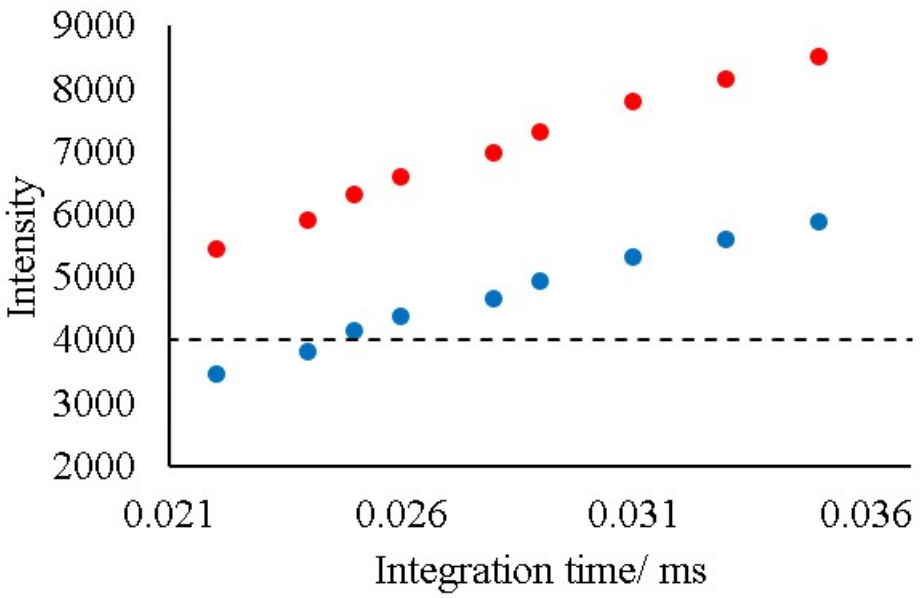


Figure 4.52: A plot of intensity of IR light against the integration time of the FPA detector (red - large area crystal; blue - small slide-on crystal). At each integration time, the intensity of large area crystal is almost twice that of small crystal. The dotted line shows the least intensity required to obtain spectra of acceptable S/N ratio

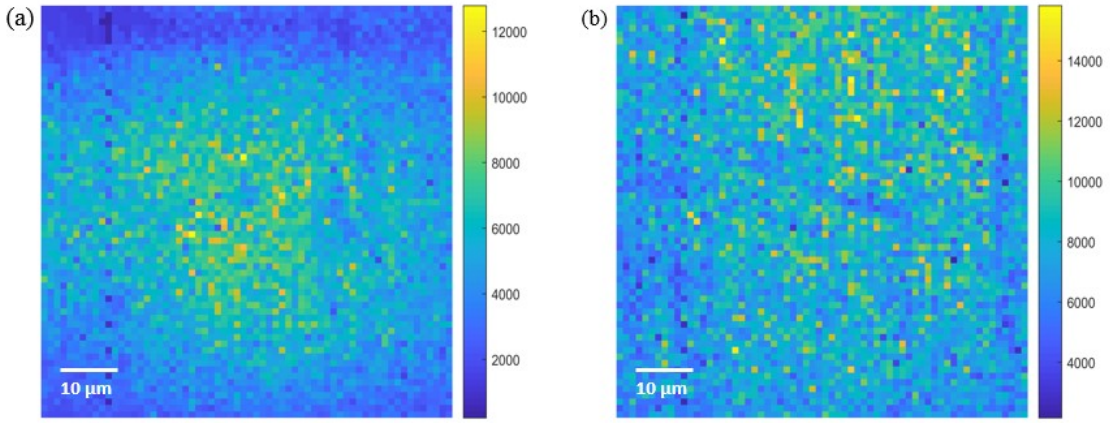


Figure 4.53: Distribution of signal intensity within the FOV of the FPA with a size of 70 $\times$ 70 μm^2 of (a) small slide-on and (b) 'large area' Ge crystal. The former has a good signal limit of up to 60 $\times$ 60 μm^2 whereas the latter is only limited by the size of the FPA. As indicated by the color scalebar, set-up (a) has a lower maximum intensity compared to set-up (b)

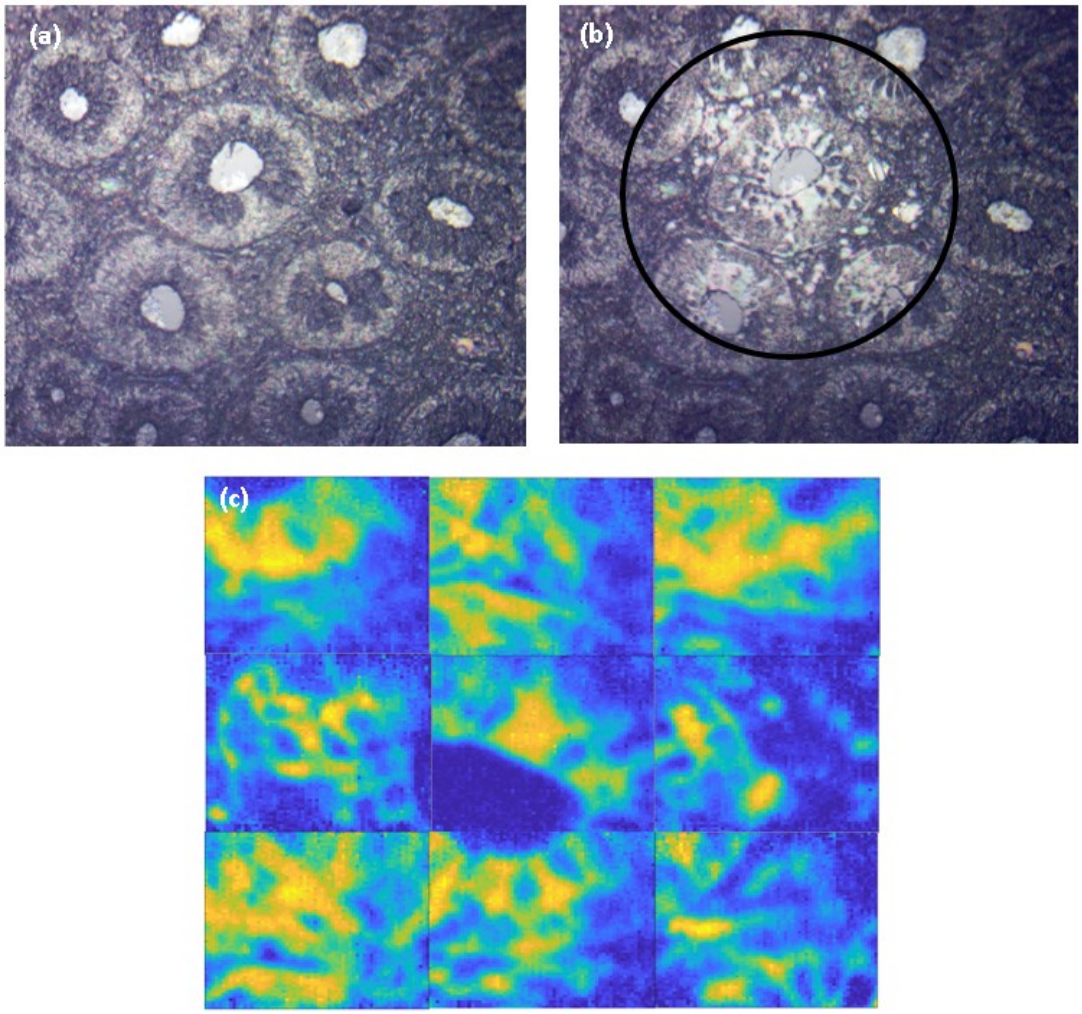


Figure 4.54: (a) The healthy colon biopsy taken under visible light in reflection mode before measurement was taken and (b) the visible image taken after measurement with small Ge crystal. Impression made during contact with the tissue (in circle) can be seen clearly close to the centre of the image where measurement was taken. (c) The mismatch in images upon stitching with small Ge set-up due to the inconsistent contact pressure applied as multiple contact needs to be made during mapping

of beam. For the first mapping distance of 70 μm from the centre position (the crystal is left-shifted), the resolution changes to $\sim$ 6.91 μm, and subsequently lowered to $\sim$ 8.02 μm,$\sim$ 8.75 μm, $\sim$ 9.64 μm and so on (Fig. 4.55(c)). Therefore, the imaging area with mapping is ideally not larger than 420 $\times$ 420 μm^2 as the advantage of imaging in ATR mode to obtain image of higher resolution is forfeited when spatial resolution reaches a value greater than ca. 9.64 μm – measurement in transmission mode is reported to have a spatial resolution in the range of 8.5 μm to 11 μm (using FTIR microscopes in high magnification mode and with added lens) (Chan & Kazarian 2013, Kimber et al. 2016). When the stage is moved from its centre position (in the case of mapping), chromatic aberration caused by dispersion occurs. The refractive index of the lens varies with the wavelength of light. This phenomenon causes difficulty in focusing at the sample. The refraction of light beam at the surface of the Ge crystal during mapping is illustrated in Fig. 4.56. As the spatial resolution is inversely related to the sine of angle of incidence and the angle of reflection of infrared light beam, the change in the angles might contribute to a worse spatial resolution recorded.

4.6.5 Micro ATR-FTIR spectroscopic images from mapping

The spectroscopic images obtained through the new mapping setup are shown in Fig. 4.57 which has a total size of up to $\sim$ 420 $\times$ 350 μm^2 (demonstrated in the figure is an example of the distribution of the integrated absorbance of amide II band between $\sim$1600 – 1500 cm^{-1}), by stitching together images from mapping by moving the stage of the microscope in both vertical and horizontal directions. The perfect stitching of images shows that there is no impression left by the Ge crystal during mapping with large area crystal where the contact between crystal and sample needs to be made once only as the stage and crystal move as one.

4.6.6 Unsupervised classification with *k*-means clustering

ATR-FTIR spectroscopic images were constructed based on the distribution of the integrated absorbance of at 1145 – 985 cm^{-1} (C-O ribose and C-C), 1182 – 1144 cm^{-1} (C-O stretching, glycogen, and collagen), 1581 – 1479 cm^{-1} (Amide II), and 2860 – 2844 cm^{-1} (C-H of mostly lipids), (Movasaghi et al. 2008) shown in Fig. 4.58 (only a small section of 70 $\times$ 70 μm^2 of the whole image is included here for demonstration purposes) . It can be seen from the figure that the distribution of amide II overlaps strongly with that of ribose and glycogen, whilst the lipid regions are complementary to amide distribution. Among all the tissues of different level of malignancy, healthy tissue gives the best separation of the distribution of the components. The possible explanation for this is that the malignant cells are more closely packed together, and the undifferentiated mass of cells made it

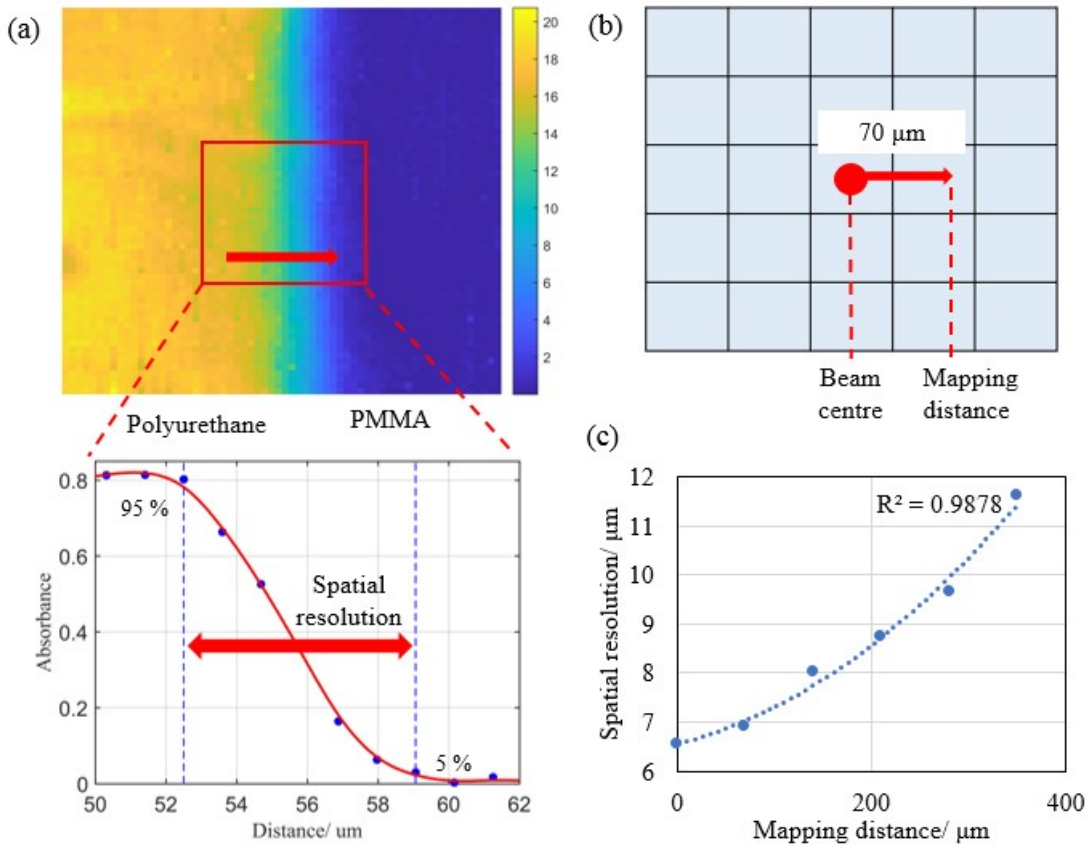


Figure 4.55: ATR-FTIR image of the integrated absorbance at the band of 1600 cm^{-1} along the red arrow across a vertically aligned sharp polymer interface. The spatial resolution is estimated by taking the distance between 95% and 5% of the maximum absorbance. The plot depicted is for measurement with IR beam directly above the crystal at its centre position. (b) Illustration describing the mapping distance of 70 μm from one image to another. (c) A plot of spatial resolution versus mapping distance showing the image resolution becomes worse as the distance from the centre increases (Ge moves left)

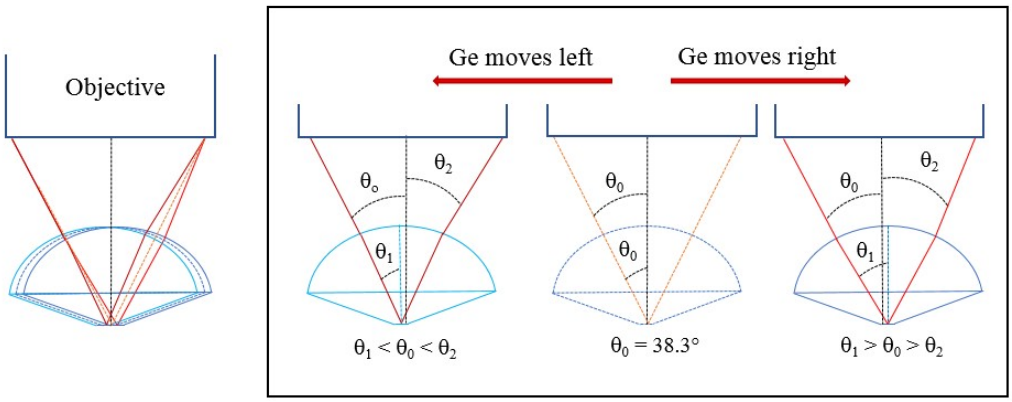
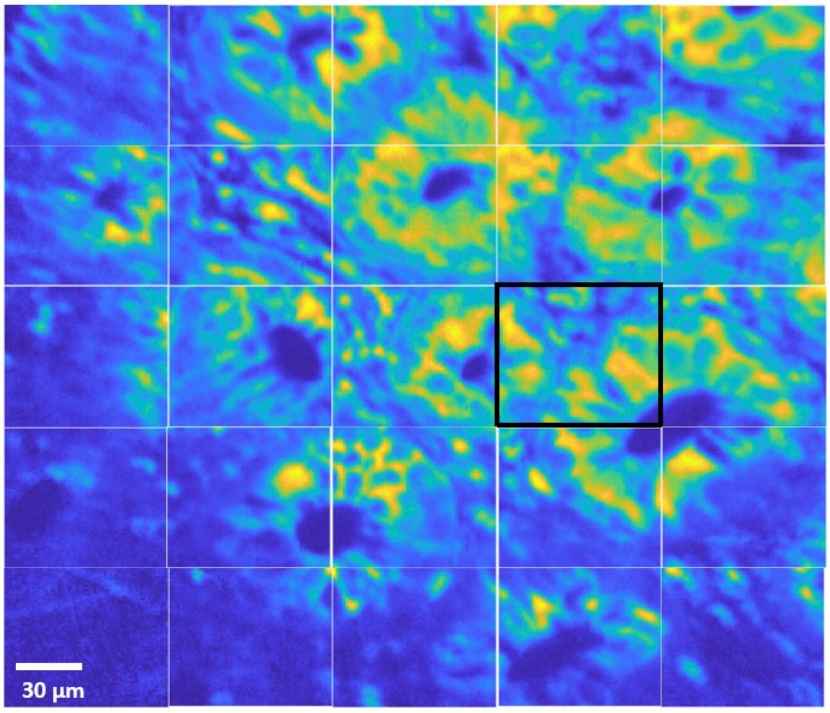


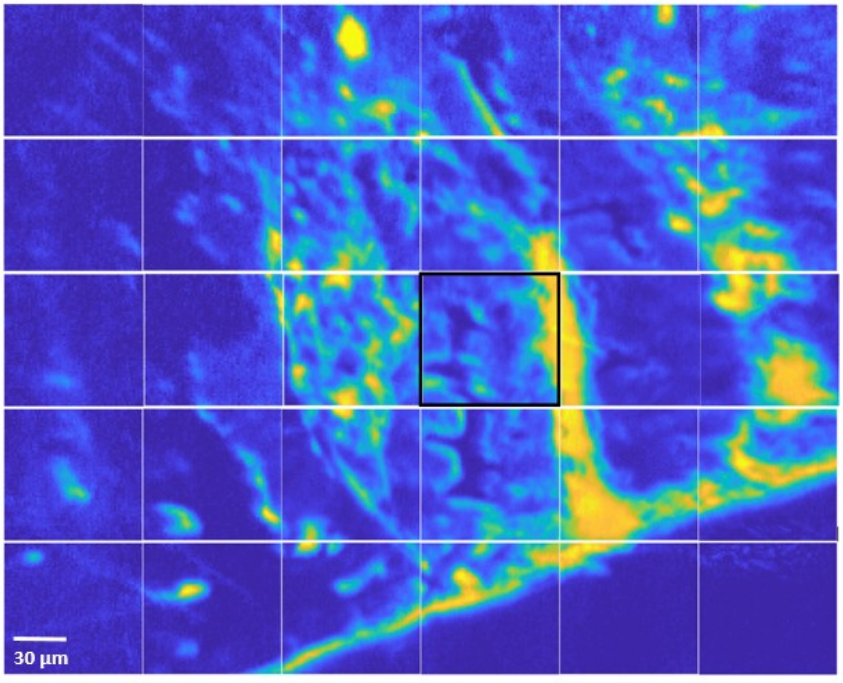
Figure 4.56: Schematic showing the refraction of light beam at the surface of the crystal when the stage was moved from the centre position in the mapping process. A range of angles of incidence and reflection are expected when the incident light beam does not enter the crystal at a perpendicular angle, resulting in the change of spatial resolution as a function of mapping distance from the centre position

difficult for the region to be separated from each other due to the resolution limit of the spectroscopic system employed here. A close examination at the chemical images of the healthy tissues reveal that the distribution of absorbance of the spectral bands at 1144 – 985 cm^{-1} coincides with that of nucleus within the epithelial cells when compared to the H&E images (shown on the far left column of Fig. 4.58). On the other hand, integrated absorbance at 1182 – 1144 cm^{-1} which is suspected to arise from glycogen has a high absorbance in the region around the nucleus, although this is not well differentiated for hyperplastic, dysplastic, and cancer tissues. Spectroscopic images of all the amide II band have also shown that the crypt of the colon tissues (which surrounds the lumen) has a higher absorbance compared to that of stroma and stromal cells, likewise for the other spectral bands.

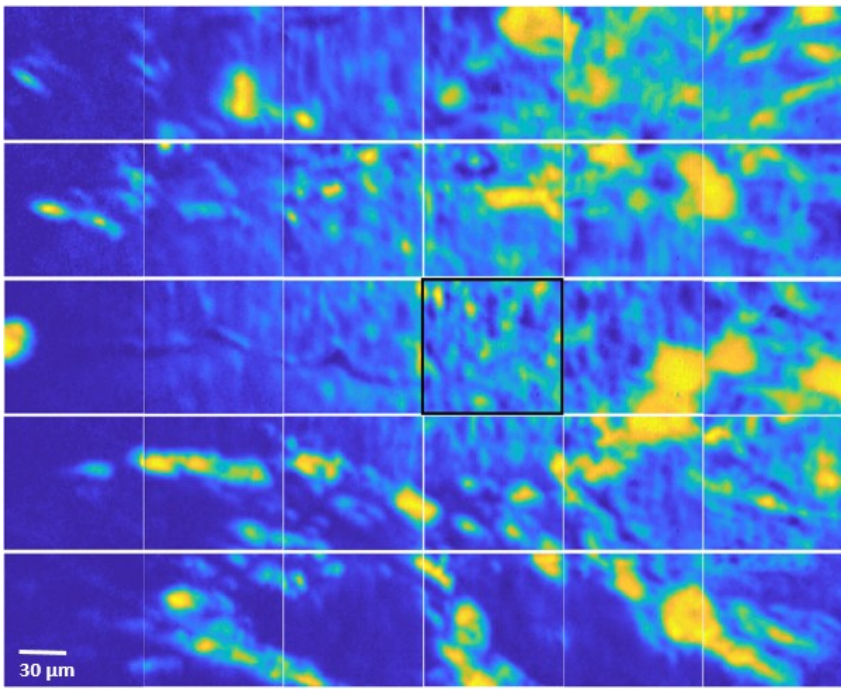
Unsupervised classification of the tissue samples was carried out with k-Means clustering algorithm. Images generated from k-means contain useful information that separates the region within the sample. Five clusters are identified to be the optimum clusters for characterisation of the colon tissue morphology. As can be seen in Fig. 4.58, the region of stroma (blue) and crypt (green) are well differentiated between different clusters, as well as the epithelial cells (also known as colonocytes), goblet cells, and lumen or intestinal gland (dark blue) which can be found in crypt. By comparing the k-means images to other chemical images, amide-rich region and lipid-rich region within the crypt are found to be complementary to each other. Classification of the spectral data into respective groups following k-means algorithm allows the extraction of useful data, for instance, spectral data from stroma region is excluded from analysis due to their overall poorer absorbance in the acquisition range compared to the crypt region.



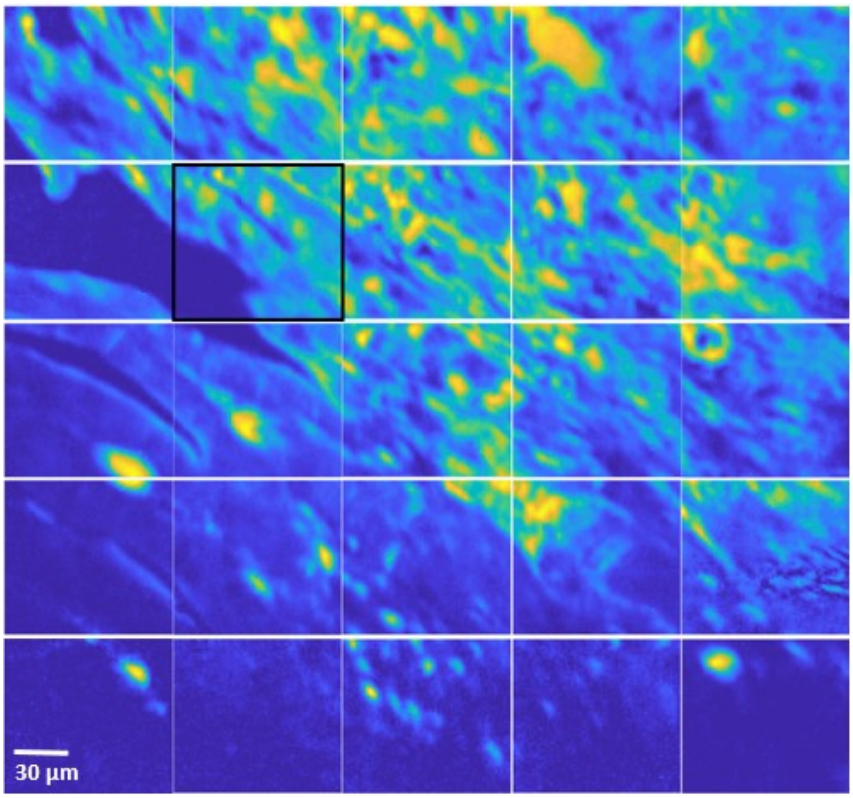
(a) Healthy colon tissue



(b) Hyperplastic colon tissue



(c) Dysplastic colon tissue



(d) Colon cancer

Figure 4.57: Micro ATR-FTIR spectroscopic images constructed from the distribution of the integrated abosrbance of amide II band by mapping with the 'large area' Ge crystal. Perfect stitching of the chemical images without further processing was easily obtained.

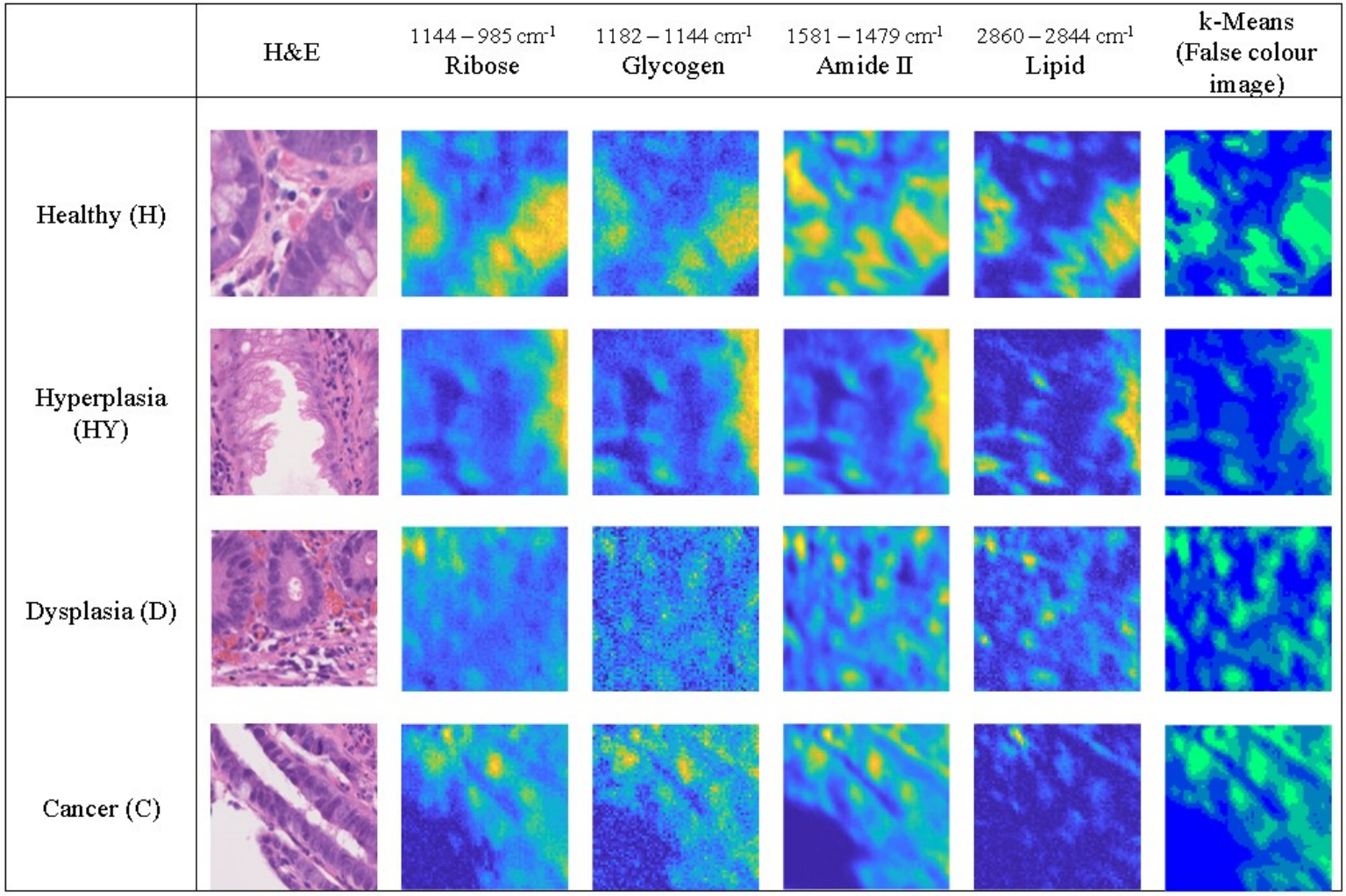


Figure 4.58: Comparison of the ATR-FTIR spectroscopic images generated from the distribution of the integrated absorbance of several different bands in images areas with H&E images for the identification of the biological components within the tissue, as well as the k-means images constructed from the second derivative spectra of each tissues. Each image has a size of 70 $\times$ 70 μm^2. The scalebar is omitted as all the images are rescaled to values between 0 and 1 for easier comparison between images of different components

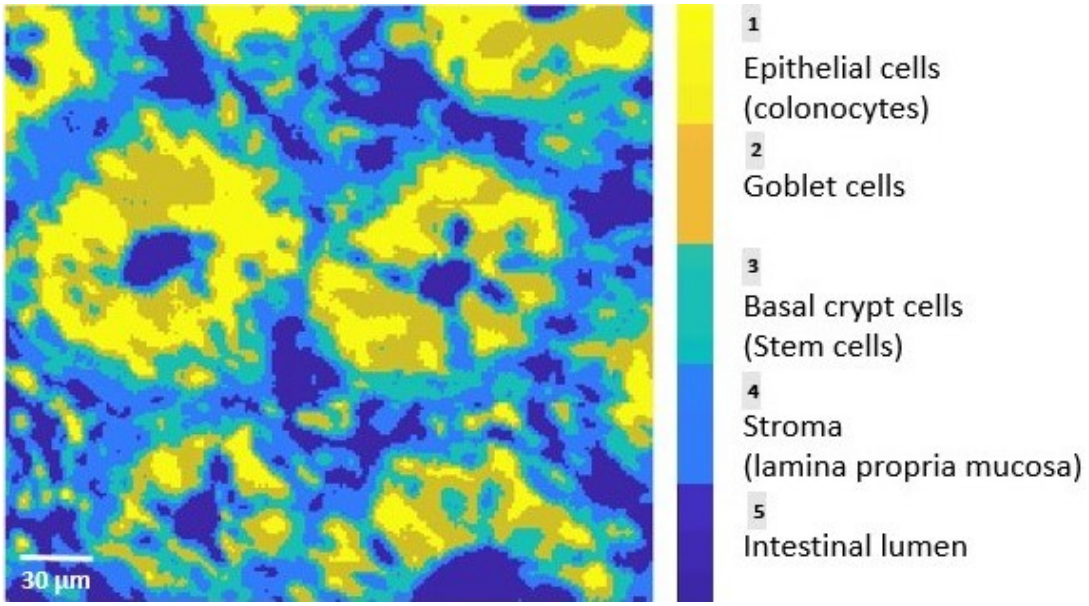
The corresponding morphology of each k-means cluster in comparison to H&E stained images for healthy section is shown in Fig. 4.59 (For demonstration purpose of the various morphologies of colon biopsy cross-section, only healthy tissue with distinct morphology is shown here). The morphology of similar tissue samples was discussed in Section 4.5. The pure spectra corresponding to each cluster identified are given in Fig. 4.60.

4.6.7 Significance of large area ATR mapping

Silhouette plots, shown in Fig. 4.61, were used to get an idea of how well-separated the resulting clusters (The MathWorks, Inc. 2020*c*). The silhouette plots were used to compare the performance of k-means clustering on small imaging area with the 'large area' Ge set-up (70×70 μm^2 without mapping) versus large area with mapping to highlight the advantage of mapping. The documentation for silhouette analysis on k-means clustering can be found in (scikit-learn developers 2019).

Mapping was not used to generate large images in the silhouette plots in Fig. 4.61a , it can be seen that most points in lumen and stroma have a positive silhouette value, indicating that those clusters are most likely correctly classified. However, all clusters have values less than 0.6 – the data points are not well separated from neighbouring clusters. This is even worse for clusters representing the crypt of colon, including the basal, epithelial, and goblet cells where negative silhouette values are observed for all tissues. When mapping was used to produce large image datasets in Fig. 4.61a, the performance of k-means clustering improves significantly. Although negative silhouette values are still observed for clusters representing crypt, the proportion of data with such values is greatly reduced. More importantly, all five clusters now exhibit values greater than $> 0.6 - 0.8$, indicating a good separation between different neighbouring clusters. To summarise this observation, we infer that measurement of a larger area helps with the classification of tissue morphology using unsupervised classification techniques, which is logical as more datasets will be less skewed by the presence of outliers. Furthermore, in the instance of colon tissues, the data that corresponds to each cluster (or morphology) are more equally distributed when a larger area is taken. K-means is sensitive to the size of the clusters since it aims to minimize the intra-cluster sum of squares, the k-means algorithm allocates more 'weight' to larger clusters. This would result in the 'centroid' of the cluster being wrongly assigned (Robinson 2020). As was already mentioned, one of the highlights of the large area crystal is its suitability for consistent large area mapping which is more advantageous compared to small Ge crystal.

(a) H&E stained image



(b) False colour k-means image

Figure 4.59: H&E stained and false colour k-means images of healthy colon tissue, with the anatomical structure labelled accordingly

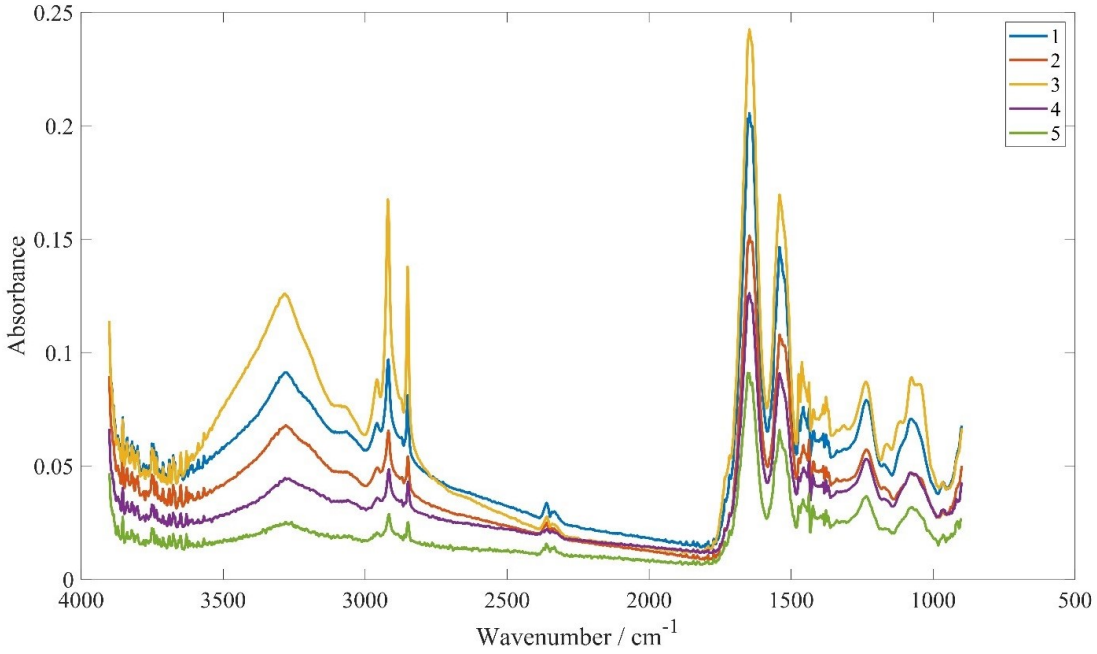
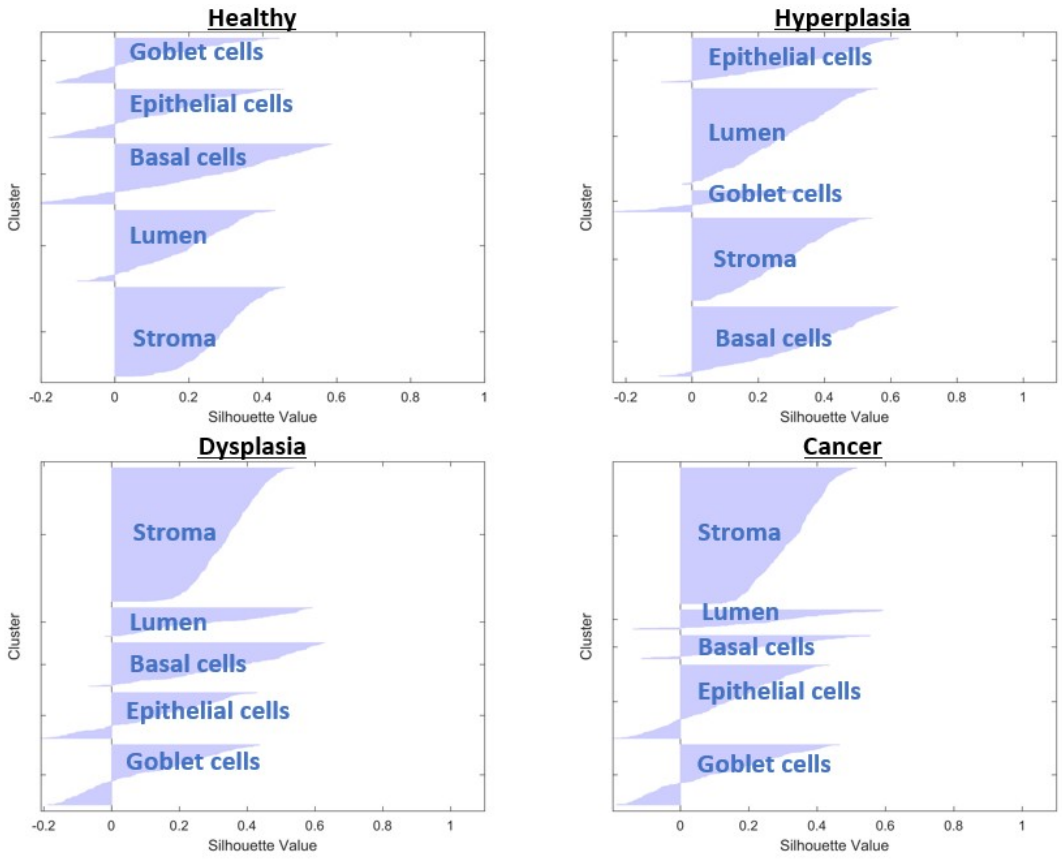
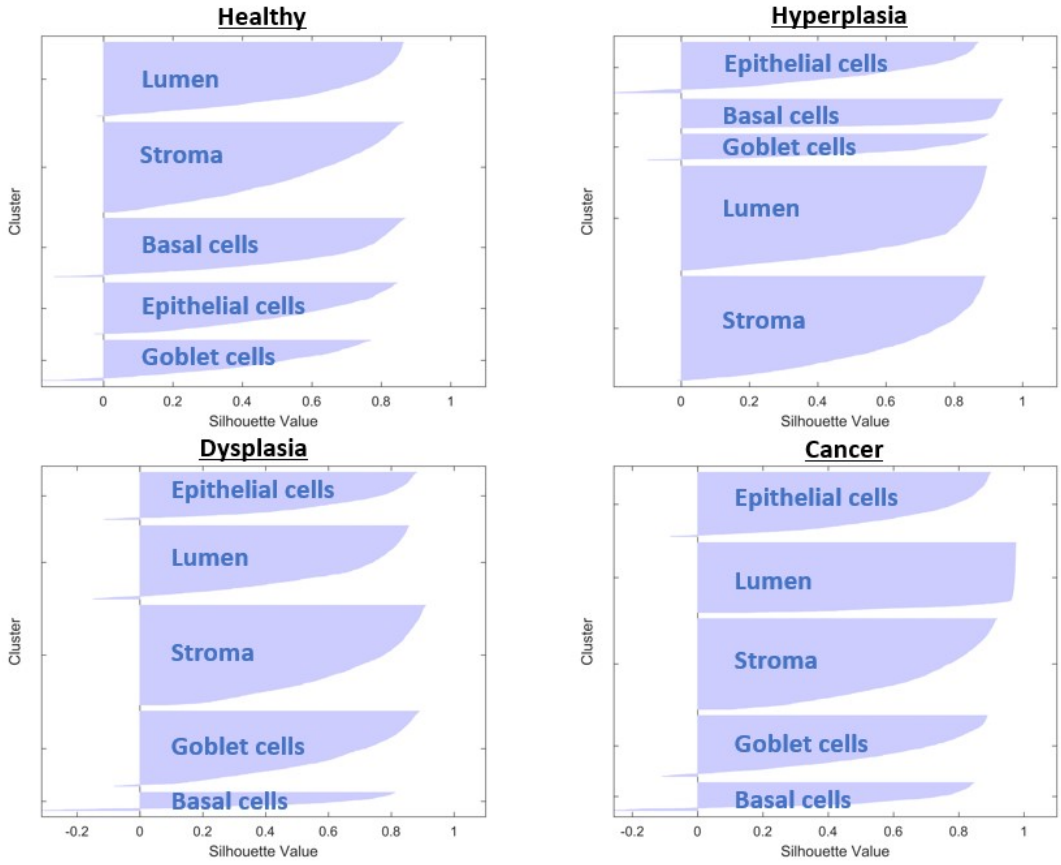


Figure 4.60: Pure spectra extracted from each k-means cluster identified (cluster number is given in the legend)



(a) Small sample area (70 × 70 μm^2)



(b) Large sample area obtained with mapping

Figure 4.61: Silhouette plots from k-Means clustering of spectral datasets. The plots show the measured distance between points in any one cluster to its neighbouring clusters. The distances are scaled and represented in the range of -1 to +1. The greater the value, i.e. when the value is close to +1, the point is distinctly different from its neighbouring clusters and a negative value on the silhouette plots indicates points that are probably misassigned to a wrong cluster.

4.6.8 Classification of spectral datasets with PLS

PLS analysis was used on the second derivative spectra to discriminate the colon specimens based on their degree of malignancy. PLS was chosen in this study over PCA as the former performed better with a mean squared error of 0.18 (R2 = 0.90) versus that of 0.36 (R2 = 0.76) for the latter. The optimum number of PCs for this separation is three, based on the 'elbow' of the graph where the mean squared prediction error seems to level off (see Fig. 4.62). The scores of the PLS are plotted in Fig. 4.63 and as can be seen, the data for each tissue are very well separated. When the datasets are projected along PC1 and PC2, cancer is well separated from the other data, followed by dysplasia with a small degree of overlap; on the other hand, healthy specimens are best separated from other data clusters when they are projected along PC1 and PC3, likewise for hyperplasia, dysplasia, and cancer.

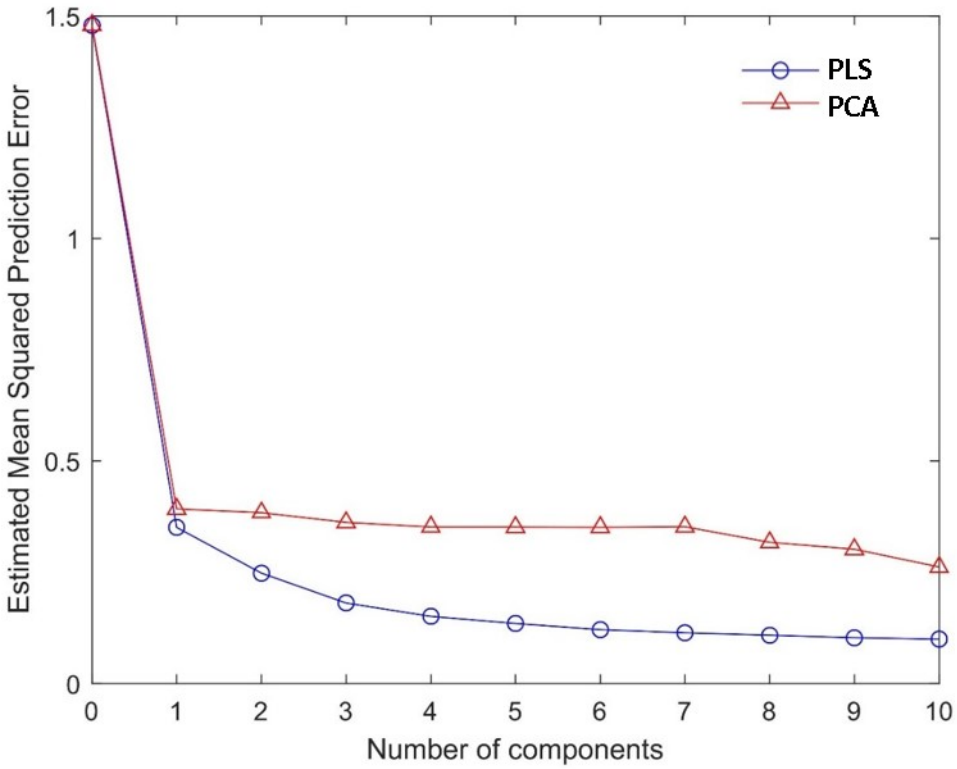


Figure 4.62: Plot of mean squared error against the number of components using PLS and PCA analysis, which shows that PLS performs better in the classification of the spectral datasets

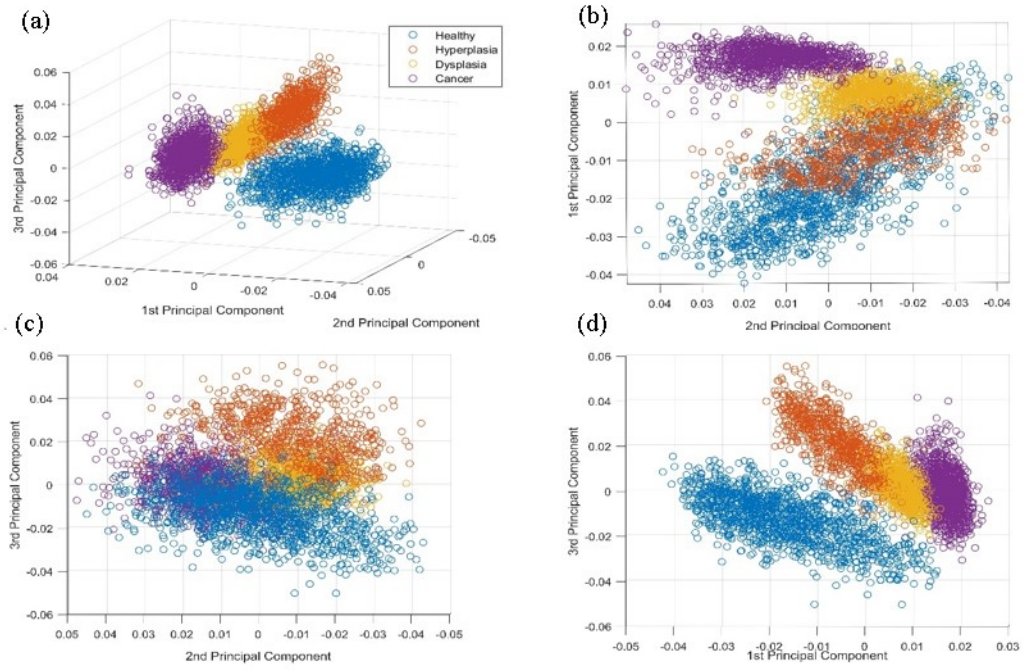


Figure 4.63: Score plots obtained from PLS presented in (a) 3D plot along PC1, PC2 and PC3; (b) 2D plot along PC1 and PC2; (c) 2D plot along PC2 and PC3 and (d) 2D plot along PC1 and PC3. Data in different colours come from tissue of different malignancy (blue - healthy; red - hyperplasia; yellow - dysplasia; and purple - cancer)

Significant features (or absorption bands at various wavenumbers) of the tissue datasets, based on the average weight (or loadings) of features along PC1 and PC3, are given in Fig. 4.64. The average weight of the feature (normalised between 0 and 1) along PC1 and PC3, annotated with blue dots, represents the significance of each wavenumber recorded for the classification of colon cancer. The important spectral biomarkers are mostly found in the fingerprint and amide II region. Orange dots showing the colon spectrum overlaid on the feature weights for clearer illustration.

A total number of 467 features were fed into the model and it is found that only $\sim$ 58 features govern $>$ 90% of the total variance for the classification of the spectral data from the first 3 PCs, thus it can be inferred that only $\sim$ 12 % of the all features of the datasets contains important information on the changes of the spectral biomarkers between different stages of colon cancer, the rest are either noise or spectral features that are identical between different stages of colon tissues. The features in this study are not independent of one another, the same biomolecules may give rise to the absorption of several spectral bands. Similar spectral biomarkers were observed for spectral data collected in transmission (Song et al. 2019). The important features are associated with either a spectral shift or a change in peak absorbance.

PLS loading plot (Fig. 4.64) gives the important biomarkers for differentiation of colon based on its malignancy. The amide I spectral region is excluded from the de-

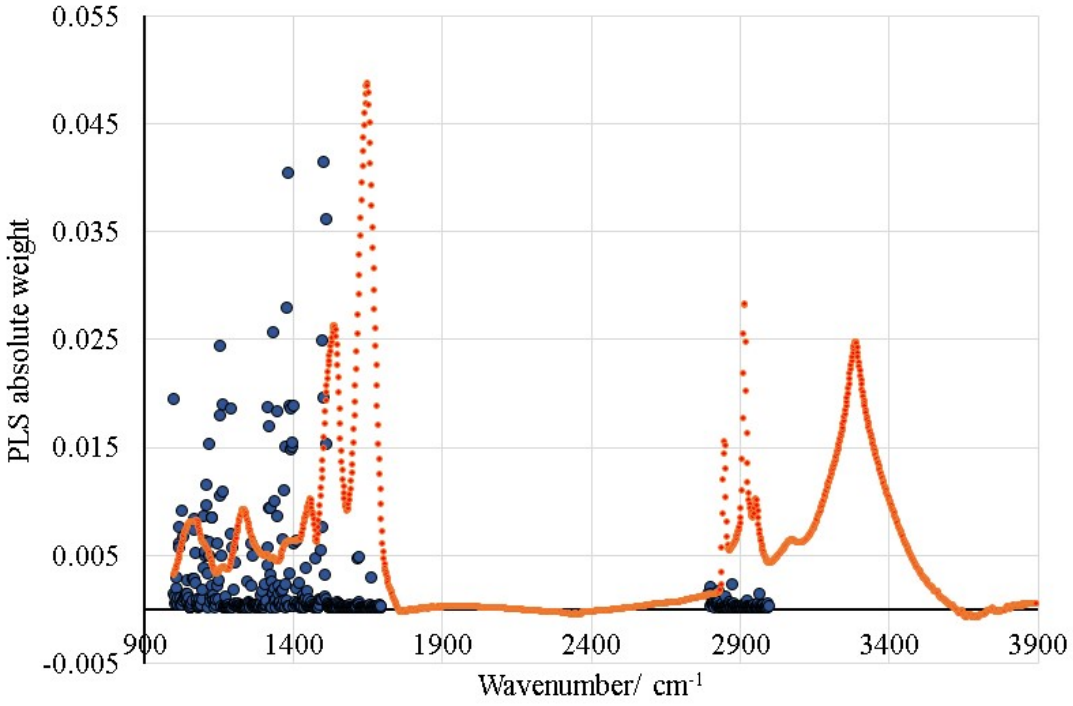


Figure 4.64: The average weight of the feature along PC1 and PC3, annotated with blue dots, represents the significance of each wavenumber recorded for the classification of colon cancer. The important spectral biomarkers are mostly found in the fingerprint and amide II region. Orange dots showing the colon spectrum overlaid on the feature weights for clearer illustration

termining features identified by PLS in this study, consistent with the results obtained in transmission mode on the same tissue samples (Song et al. 2019), further confirming that amide I band does not play a significant role in determining the malignancy of the colon biopsies studies. This surprisingly holds true for ATR-FTIR spectra which are free from the Mie scattering effect and water vapour (water vapour subtraction is applied to the spectra, shown in Section 3.3.2). Amide I band does not constitute as an important spectral biomarker in our study regardless of the probing depth, for analyses of spectra obtained in both transmission and ATR modes. The true reason for this is still unknown, it could be that the change in protein secondary structure is subtle and is not as detectable in relative comparison to change of N-H vibration in protein reflected in amide II band. The amide I band is mainly associated with the CO stretching vibration (70 – 85 %) with some contribution from C-N vibration and is directly related to the backbone conformation. Amide II band originates from from the NH bending vibration (40 – 60 %) and from the C-N stretching vibration (18 – 40%).

From Fig. 4.64, it should also be noted that the spectral region of 3000 – 2800 cm^{-1} plays very little role in the classification of the colon malignancy in this study with micro ATR-FTIR spectroscopic imaging. This is different from the results obtained in transmission mode on the same tissue sections. The most plausible explanation for

Table 4.7: Assignment of vibrational modes to the important spectral biomarkers identified from PLS loading plots, alongside the peak positions in different tissue specimens.

Spectral bands/ cm^{-1}	Band assignment	Peak position/ cm^{-1}			
		Healthy	Hyperplasia	Dysplasia	Cancer
1336	δ CH in a phenyl ring (of polysaccharides) or the CH2 side chain of collagen	1340	1388	1388	1340
1384 - 1382	δ CH3, δ NH, ν_s CO of carbon particle	1388	1387	1383	1387
1506 - 1502; 1513 - 1511	δ CH in a phenyl ring	1518	1512	1512	1506
1120	ν_s CO of carbohydrates or ribose (RNA)	1120	1113; 1128	1120	1120
1160 - 1155	ν_s CO of proteins or carbohydrates	1156	1157	1154	1154
1320 - 1310	Amide III of protein (minor contribution from collagen)	1315	1310	1315	1315
1350	ν_s CO, δ NH, δ CH	1354	1350	1350	1354
1375	δ NH, δ H, ν_s CN of cytosine or guanine	1375	–	–	–

this is the dissimilarity in the measured thickness of the sample in transmission and the effective thickness in ATR. The effective thickness in ATR depends on wavenumber – it is approximately 0.3 μm at wavenumber 3000 cm^{-1} (Section 2.1); whilst the path length in transmission is sample-thickness dependent, in this case, 3 μm. Apart from that, prior to sample measurements with spectroscopic imaging, the tissue is de-paraffinised. This is very likely to wash off any lipids or fatty acids components together with the paraffin from the surface of the tissue specimens but any lipid components away from the surface will remain. The loss of lipid components on the surface of the tissues could explain why the C-H stretching bands (assigned to lipids) are not important biomarkers for colon cancer classification in this micro ATR-FTIR imaging study. The important biomarkers in the fingerprint regions are given in Table 4.7 alongside their band assignments.

4.6.9 Summary

A new 'large-area' Ge ATR crystal was utilised with an FTIR microscope to improve the acquired images of de-paraffinized colon biopsy sections, without recourse to synchrotron source. The large crystal (ϕ = 28 mm) offered significant improvements compared to slide-on small germanium crystal (ϕ = 3.5 mm); for example, it facilitated a more uniform distribution of higher signal intensity within the FOV and rapid acquisition time. Mapping of a larger sample area with this new set-up, coupled with imaging using an FPA detector, was demonstrated for the first time on biological specimens. The performance of k-means clustering algorithm applied to classify the different anatomical structures of the colon biopsies was greatly improved with mapping. Comparison of H&E stained adjacent tissue sections with false-colour k-means images strongly supported the differentiation of five distinct regions within tissues. The efficiency of the methodology to categorise colon tissues at various stages of malignancy was analysed via multivariate chemometrics. The second derivative spectra extracted from the crypt region of the colon were subjected to PLS classification. Good separation between data in clusters occured when projecting spectra onto a PLS score plot on a plane constructed from the first three PCs. Important spectral biomarkers for colon malignancy classification were identified to exist mostly in the fingerprint region of the ATR-FTIR spectrum based on the chemometrics analysis.

Chapter 5

Conclusions and outlooks

5.1 General conclusions

This research started with the conventional dispersive IR micro-spectroscopic imaging of prostate biopsy samples (Song et al. 2018). Biochemical differences between cancer and benign areas within the specimens were identified in the spectra in the high wavenumber (predominantly lipid) regions ($3000 - 2800\,\mathrm{cm}^{-1}$). Side-by-side comparison of H&E stained adjacent tissue sections with IR images constructed showed that the differentiation of healthy and malignant regions within the tissues were possible from this high wavenumber spectral region. A systemic methodology was implemented to process the data, first by k-means clustering on the second derivative spectra, followed by PCA analysis. Four distinct regions within the tissue samples were successfully classified based on the anti-symmetric stretching mode of the methylene functional group (mostly from lipid). Separation between data in clusters occured when projecting the FTIR spectra on a PCA score plot on a plane made by first two principal components.

In addition to IR measurements, the dispersive microscope was designed to allow thermographic imaging and the effect of thermal radiation on the infrared absorption spectra of prostate biopsy samples to be studied as well. Regulation of signal to chopper and detector enabled simultaneous acquisition of infrared and thermal images of the tissues. Although the system is novel, the applications of thermal imaging of tissues were limited due to several reasons. First of all, the tissues were at room temperature – this meant that the IR emission of the surrounding (for instance, the light in the room) was non-negligible. Unlike thermographic imaging in clinics or hospitals, take thermography for breast cancer as an example, the cells in the tissues are not alive and there is no generation of heat by the tissues themselves. At room temperature, the difference captured with the thermal camera is dependent on the thermal emissivity of the samples, which in turn, depends on the morphology of the tissue. However, the difference in the tissue

emissivity is not very significant. This resulted in the thermal images obtained having a lower contrast compared to IR spectroscopic images. Secondly, the morphology of the tissues captured from their thermal emissivity was the same as the visual images observed under visible light microscope. In other words, the importance of having the chemical information, in particular for classifying diseased and healthy tissues, was lost. At this point, it was clear that supervised machine learning for the classification of tissues based on the thermal images had no additional values and should not be further pursued.

That said, the findings showed that the thermal emission was captured in IR spectra acquired from the dispersive microscopic spectrometer. This 'thermal noise' was not unfounded – it is one of the limitations of dispersive spectrometer that has been known before FTIR spectrometer was introduced. With the implementation of this system to simultaneously capture IR and thermal images, the 'thermal noise' can efficiently be subtracted from the IR images. The resultant IR images obtained after subtraction showed a much better contrast than before subtraction from the improved S/N ratio. The dispersive IR spectroscopic imaging can be improved via this implementation. There were several limitations associated with the system as well – the IR thermal imaging camera has a limited wavelength range between 3 – 5 μm; hence, only the stretching vibrational modes of CH functional groups were investigated. Although the results showed that this small wavelength range could differentiate malignant and healthy sections within the prostate issues; the fingerprint region which contains more information on the samples could not be investigated by the spectrometer used in this study. Not only that, when the IR images obtained from the dispersive IR spectrometer were compared with the FTIR spectroscopic images obtained from FTIR imaging on the same sections of the prostate tissues, the former had a lower S/N ratio, as the spectra were acquired from a 'single scan' – each single wavelength was taken one at a time which made 'averaging from multiple scans' too time exhausting. Spectra of good S/N ratio, however, could easily be achieved with FTIR spectroscopic imaging.

A general approach to classifying tissues with FTIR spectroscopic imaging was adopted next in the research. Colon tissues were first deparaffinised – the disappearance of the paraffin peak was used as a guidance for effective removal of the paraffin. The FTIR spectroscopic images of the tissue were acquired in the mid-IR range between 3900 – 900 cm^{-1} in the transmission mode at a spectral resolution of 4 cm^{-1} with 512 co-added scans (Song et al. 2019). The selection of the spectral resolution and co-added scans are dependent on the instrument used. From trials and errors, these were the most suitable parameters here using the Bruker Hyperion 3000 microscope. The parameters, however, were not a major issue in tissue classification. There was flexibility in the choice of the parameters as long as they were kept consistent throughout when the chemometric analysis or machine learning algorithm was trained and tested on the spectral data. Compared to the selection of parameters, the transfer of machine learning model trained from spectra

obtained across different instruments and measurement modes was more challenging. This is discussed later when the same samples were studied with ATR-FTIR spectroscopy.

This research has found that the pre-processing steps of the spectra obtained were important. The framework followed, that resulted in optimum chemometrics results, was first the removal of the water vapour bands by spectral subtraction; then the data was trimmed to retain the spectral range between 3000 – 2800 cm^{-1} and 1800 – 1000 cm^{-1}. CO_2 spectral bands were eliminated with this step. The data below 1000 cm^{-1} suffered from poor S/N ratio and the higher wavenumber above 3000 cm^{-1} was sensitive to the water content of the tissue. This was confirmed by investigating the effect of different levels of hydration on the spectra of the colon tissues. The tissues, despite being 'dead' and thin (few μm thick), were shown to be capable of being hydrated by absorbing the water from the surrounding.

To account for the remaining issue, which is the *resonant* Mie scattering effect, a physical correction with an added correcting lens and a computational correction were tested separately. Although the spectral data of colon biopsies obtained with the correcting lens for FTIR imaging showed a significant reduction in spectral aberrations, the physical correction could not 'completely' remove the effect of scattering, i.e. the baseline shift and the derivative-like shape on the high wavenumber side of amide I band could slightly be observed in the colon tissues in this study, most likely due to the use of a single pseudo-hemispherical lens, which partially corrects the chromatic aberration in transmission mode experiment, rather than two on both sides. However, this optical modification of the FTIR spectroscopic imaging with a CaF_2 correcting lens had the advantage that the Mie scattering correction algorithm did not need to be carried out, although the study found that the correction effect was not as good as that with computational method. The computational correction was successfully carried out with RMies-EMSC iterative correction algorithm. This process was computationally very demanding with a computational time of approximately 1 s per iteration, so for 64 $\times$64-FPA imaging (4096 spectra) with 10 iteration per spectrum would take $\sim$ 11 h to complete. Furthermore, in this study, the RMies effect was seen only at the edge of the tissue. In other words, a large amount of data further away from the edge of the tissues remained unaffected and freed from this spectral distortion, unlike cells. In brief, it can be concluded that the removal of Mie scattering is subjective to case-by-case studies. The findings showed that the disease states could be distinguished without resorting to the correction of Mie scattering effect. The second derivative spectra were more useful than the raw data as it eliminated the variation of baseline shifts and deconvolved overlapping spectral bands. To increase the S/N ratio, SG smoothing was implemented. 5- to 9-point smoothing were in the acceptable range; below/beyond which the classification accuracy dropped significantly.

Using k-means clustering and RF classifier with PCA reduction, this research has demonstrated that optimisation of the training model by refining the selected range

of FTIR spectral data could alter the prediction outcome. The best prediction outcome for the studied colon biopsy samples was obtained when unsupervised learning of the C-H stretching bands was coupled with supervised learning of the fingerprint region. Hence, while the C-H stretching region was useful for intra-tissue segmentation, only the fingerprint region within the spectral range of 1500 – 1000 cm^{-1} was important for supervised machine learning. The amide I band could be excluded from data analysis altogether, as evidenced in the Gini indices obtained. Similar to the study of prostate tissues with dispersive IR spectroscopic imaging, the significance of the differences in the C-H stretching (3000 – 2800 cm^{-1}) between healthy and malignant samples was seen, but the reliance on this spectral region alone was not sufficient. In supervised learning, the C-H stretching region alone gave the worst prediction.

Further experiment with the regulation of the humidity in a controlled humidity box with saturated salt solutions showed that the hydrational level of the colon tissues had an impact on the prediction outcome with the same machine learning algorithm (Song & Kazarian 2020). Developing and improving the mathematical aspect of the machine learning algorithm was out of the scope of this research study; however, a significant improvement in the sensitivity of the disease classification of the tissues was recorded from 92 % to 96 % at their de-hydrated state. Interestingly, the prediction accuracy decreased as the hydration increased (the lowest accuracy was recorded at humidity level near the room humidity in the UK at ∼45 %RH) but increased and reached a high value again at the most hydrated state. The reproducibility of the previous machine learning model depending only on the fingerprint region was warranted. The main spectral biomarkers were identified to be 1032 cm^{-1}, 1057 cm^{-1}, 1076 cm^{-1}, 1078 cm^{-1}, 1117 cm^{-1}, 1192 cm^{-1}, 1209 cm^{-1}, and also the ratio of the peak intensities between 1182 - 1140 cm^{-1}. When the tissue was de-hydrated, changes in peak intensities and peak shifts that corresponds to the vibrational motions of the phosphate group of nucleic acids, mainly DNA, were observed, leading to greater distinction between tissues of different malignancy at these spectral bands. The DNA was found to mostly present in the A-DNA form in the deparaffinised colon tissues, unlike those in live cells. From the findings, it was made clear that the humidity where the measurement was carried out should be strictly controlled, ideally at a very low humidity.

Due to the limitation in the availability of the tissue samples and the area sections that have been identified from H&E staining by the pathologist, the findings were based on a manageable number of datasets from eight different tissue sections. The limited number of images made it impossible to train a machine learning model based on the images, but the huge number of spectra (4096 spectra in one image) was utilized for the machine learning algorithm. More significantly, this research sets a framework for further application to an unknown colon biopsy sample in future work. The framework was completely straightforward and could be fully automated with simple programming

using MATLAB. K-means clustering at the first stage on the C-H stretching bands alone would pick up regions of high lipid absorbance which would subsequently be fed into the already trained RF model to predict the outcome of the malignancy stage of the specimen; all whilst maintaining the tissues at their de-hydrated state.

This research did not just involve establishing a groundwork for tissue classification, but also aimed to improve the quality of the spectra and images obtained and investigated the applications of novel implementation on biological tissues. One of such implementations was the use of additional correcting CaF_2 lens to minimize Mie scattering and providing extra magnification to the images obtained, which has been shown in literature and once again being shown here for imaging of colon tissues in transmission mode. Here in this research, a similar idea of an additional lens was adopted in ATR imaging. A 'large area' Ge crystal that was not attached to the microscope was used to replace the slide-on ATR Ge crystal that came with the microscope (Song & Kazarian 2019*a*). The study showed that combining mapping with ATR-FTIR spectroscopic imaging was improved with the novel set up of 'large area' Ge crystal compared to the conventional slide-on Ge crystal. This new crystal has the added benefits that it allows a larger amount of infrared light to probe the sample and be collected by the detector, thereby improving the overall S/N ratio of the spectra. Furthermore, the larger crystal ensures a uniform distribution of IR light intensity across the images area, pushing the limit of measurement area from the size of the crystal to that of the size of the FPA detector. This is a significant improvement to FTIR spectroscopic imaging with micro ATR at reduced acquisition time whilst keeping a high S/N ratio. This set-up was used for the first time to study colon cancer tissues on a conventional benchtop FTIR microscope without recourse to the use of a more powerful synchrotron source. The high-quality chemical images obtained was comparable to that of synchrotron, with spatial resolution around 6 μm. Combining imaging with mapping using the large crystal, this allowed large area to be taken; subsequently more data could be generated from a single specimen which allowed for more effective chemometric analysis. K-Means analysis of these large areas more effectively classified different morphology of the tissue when compared to the H&E stained images identified by pathologists. Classification of the spectral data of colon biopsies sections based on their degree of malignancy was obtained. Further data analysis with PLS revealed a good degree of data separation along the first three components. From PLS analysis, the important biochemical changes identified were mostly captured in the fingerprint (1400 – 1000 cm^{-1}) and amide II (1500 – 1400 cm^{-1}) region, most notably, these spectral ranges were associated with the change in C-H and C-O functional groups of the biomolecules present in colon biopsies. Likewise with imaging in transmission mode, there are hardly any significant differences in amide I band of the tissues. Furthermore, in this micro-ATR study, we provide evidence that the de-paraffinization process might lead to leaching of fatty acid components from the surface of the tissue as lipid bands (3000 – 2800 cm^{-1}) measured in ATR mode are not distinct in the progression of colon cancer, but have shown

to have secondary importance (after the fingerprint region) in transmission measurement. The demonstrated approach can be employed to study other biological samples, other than colon tissues.

The other modification to the ATR-FTIR system was the implementation of the concept of depth profiling on the differentiation of tissues (Song & Kazarian 2019*b*). The discriminant between cancer and non-cancer prostate tissues has been shown to vary as a function of penetration depths using micro ATR-FTIR imaging. In this experiment, introducing specially designed apertures into the micro ATR-FTIR imaging system fitted with slide-on Ge crystal allowed the experiment to be conducted to investigate the variation of tissue components in the z-direction, i.e. as a function of depth. The penetration depth or effective thickness was physically altered by changing the angle of incidence between 30.7° – 41.8°. Compared to previous work, the improved design of laser cut apertures at a high precision allows the angle to be varied at a difference of $\sim 2^{\circ}$. With this variable angle micro-ATR spectroscopic technique, the images obtained at each different depth has high spatial resolution and is free from chromatic aberration. Significantly, this experiment also highlighted the importance of probing at the surface of the tissue, consolidated by statistical student t-testing. Although limited by its shallow penetration depth on the μm scale, this work has demonstrated its applicability to destruction-free imaging of heterogeneous biological samples, allowing components embedded within the tissue sample to be studied. This work opens up the possibility to investigate multi-layers heterogeneous component without the need to microtome samples to obtain stacks of 2D chemical images at different probing depths. This approach could be applied to identification of embedded components in tissue for the construction of qualitative 3d model, for example, in the case of calcification of breast cancer, the calcium deposits within a tissue layer could potentially be probed by reducing the angle of incidence. The significance of the spectral biomarkers in tissue differentiation, i.e. the symmetric and asymmetric stretching of the phosphate group of nucleic acids, varied as a function of the penetration depth as can be seen on their PCA loading plot. This added complications to the machine learning process. Not only that, the S/N ratio of the spectra significantly suffered due to the reduced number of photons reaching the detector as the light beam was partially cut off to achieve the desired angle of incidence. All these contributed to increased in variation in the spectra for training of the machine learning algorithm, which would eventually lead to a low prediction accuracy.

Training a machine learning algorithm on spectra obtained from the colon tissues in transmission mode was successful here; but there has been a number of challenges faced when trying to transfer the algorithm across to spectra obtained from ATR-FTIR imaging. When the results obtained from both transmission and ATR imaging modes, it can be seen that the spectral biomarkers (or the important features) identified from both modes are different, especially in the high wavenumber spectral region, as explained in the paragraph

above. This means that transferring the machine learning model built on transmission spectra to predict the outcome of ATR spectra is impossible. This is also very likely to be true for all inter-mode-of-measurement machine learning application, although this cannot be tested and confirmed due to the limited availability of the FTIR spectrometers in the labs and also the constraint in time. In fact, it is intuitive to hypothesise that a new machine learning model needs to be trained on the known samples every time to test for the spectra obtained from unknown samples if the conditions change (including the use of a different measurement mode). When this is the case, the experiments carried out here in this research supported that building a machine learning model based on FTIR spectra in transmission could actually be better than ATR. There are several reasons to say this and they are explained in the following paragraph.

First of all, a large number of spectra needs to be obtained in order to efficiently trained a machine learning model; very often this would mean mapping needs to be carried out. By comparison, this is more difficult for ATR measurement than transmission as there is a need to make new contact every time the stage is moved to allow imaging of a new area on the tissues. A solution to this was provided here by introducing the use of 'large area' Ge ATR crystal. Mapping could be carried out, saving the hassle to have to make new contact each time, yet the images suffered from reduced resolution as the crystal was moved away from its center position perpendicular to the light beam in the mapping process. Essentially, the advantage of a higher spatial resolution in ATR compared to transmission imaging was lost when the mapping distance was far away from the center. Secondly, it was demonstrated that spectra obtained from ATR were susceptible to the deparaffinization process. This was not seen in transmission measurement. There are many different deparaffinization protocols and to standardize this (across different research groups and tissue banks etc) requires huge effort and is impractical at this point. From all the aforementioned reasons, it is thus suggested that transmission measurement is used for classification of the tissue, unless a particulate in a tissue section requires a higher spatial resolution to resolve it, then ATR-FTIR imaging would be useful.

Last but not least, the objectives of this research were reached through a series of experiments. It was demonstrated that healthy colon tissues and colon tissues of different malignancy, namely, hyperplasia, dysplasia, and cancer exhibited different FTIR spectra, which could be successfully captured using chemometrics analytical method, including both supervised and unsupervised machine learning. The instrumentation and methodology set out a framework for effective tissue classification based on the FTIR spectra and images obtained. Two types of tissues were studied here, strongly supporting the hypothesis that the FTIR imaging can capture important information on a wide variety of tissue samples from different parts of the human bodies. New modifications were also introduced to the FTIR microscopes for novel application on other biological system in the future. As always, there is more work that can be done to advance the research in

this field, which are briefly outlined in the next section.

5.2 Future work

The experimental studies presented in this thesis were focused on the development of spectroscopic approaches and the novel findings. A machine learning model was also developed for colon tissue classification. The obtained results show huge clinical potential of FTIR spectroscpic imaging technique, but this is just the first step towards developing a generalised automated processing framework for efficient and effective routine clinical use. To achieve that, more research needs to be carried out in the future, in the following areas.

5.2.1 Further experiments with different IREs, substrates, and tissue samples

One of the suggestions to analyse the sample with ATR-FTIR spectroscopic imaging with regards to the depth profiling is by combining the apertures with another IRE, for example, ZnS that would have a greater probing depth. This could provide a means to assess the heterogeneity within a thicker layer of the tissue sections, compared to Ge. Furthermore, in the transmission experimental study with corrected lens, using different types of substrates (BaF_2, ZnS, or ZnSe) of lower cut-off wavenumber than CaF_2 may allow spectra of a good S/N ratio at the low wavenumber, $\sim$1000 cm^{-1} to be collected and analysed. Besides, it would be interesting to compare the results attained in this thesis using different tissue types, such as breast, lung, and oesophagus, to name a few.

5.2.2 Establishing a reliable model with a large patient cohort

The number of samples in this study is too small to arrive at any generalization towards a larger pool of samples or at a generally applicable diagnostic algorithm. In this proof-of-concept study only six samples from three patients were used for training and testing the machine learning model. Given the biochemical variability within a patient population, it is unlikely that three patients are a sufficiently large dataset for identifying the key biomarkers. Given the limited samples and patient numbers used in this study for training and testing, it is highly likely that the model did not have sufficient variability built in to enable good discrimination between normal and cancerous tissue for testing on unknown new patients. Besides, variability in sample and substrate thickness, and whether the samples are left in wax or dewaxed, and other spectral processing parameters discussed in section 3.3 that could potentially affect the classifier performance are not investigated

in details in this work. Although the preliminary results in this study show great clinical potential, it is recommended that a much larger study investigating the effect of each parameter involving a large patient cohort is conducted in future research to build a robust diagnostic model for high performance classification.

5.2.3 Study of other sample forms

The study of FTIR spectroscopic imaging on biological samples can also be extended to studying other sample forms, for example, live cells and biofluids. In general, it has been demonstrated in various research works that chemical imaging of live cancer cells can be carried out in the natural aqueous environment by micro ATR-FTIR spectroscopic imaging, mainly because of the high spatial resolution and the shallow penetration depth (Kuimova et al. 2009). Due to the different nature of the cancer samples (tissue versus cells), it would be interesting to compare the spectra difference between them and employing machine learning to map tissue spectra with the spectra of the cells of the same type. This could potentially lead to a generalized machine learning model for *in vivo*, *ex vivo*, and *ex situ* FTIR measurements. Besides, the corrected lens approach described in this thesis could be used to study live cells (Chan et al. 2020). There is also a possibility to measure adhered cells on the surface of the large Ge crystal for high-throughput analysis.

In another latest ongoing research, it was suggested that a new blood test coupled with artificial intelligence can achieve more than 50 types of early-cancer detection, primarily focusing on the chemical changes to this DNA, known as the methylation patterns (Liu et al. 2020). In the light of this new research findings which suggest that categorization of different cancer types is possible from blood plasma, FTIR spectroscopy can be used to study the blood samples to complement their technique to obtain chemical information and potentially improve the diagnostic accuracy.

5.2.4 Multimodal imaging in combination with Raman spectroscopy

One of the other most popular vibrational spectroscopic techniques is the Raman spectroscopy. Unlike IR absorption, Raman is based on the inelastic scattering of photons by matter. In Raman spectroscopy, the tissue sample is illuminated with a monochromatic laser and the scattered light is subsequent collected .The light scattered are of different wavelengths than the incident light, which depend on the chemical structure of the analyte [1]. Raman spectroscopy has a high spatial resolution (< 1 μm) (Wrobel et al. 2012) and down to $<$ 5nm for Tip Enhanced Raman spectroscopy (TERS) (Deckert et al. 2015),

[1] An exchange of energy occurs when the photons in the incident laser light undergo inelastic collisions with molecules, resulting in a change in frequency of the scattered light. The difference in the frequency between incoming and scattered photons is recorded as the Raman shift. A larger Raman shift indicates that a larger amount of energy is required in a particular vibrational motion.

hence it has the potential to be used as a complementary technique to FTIR spectroscopy for research in cancer diagnosis (Cui et al. 2018). In future study, it would be interesting to combine the results from mutimodal imaging and mapping with Raman and IR spectroscopy to generate a machine learning model that could have a higher accuracy at disease classification.

5.2.5 *In vivo* probing of disease for translation to clinical use

In reacent years, fiber-optic probes have been developed for spectroscopic measurements in living systems, which open up the opportunity for *in vivo* cancer detection, which should be considered in the future work. For the study of colorectal disease, Raman spectroscopy in conjunction with specialised fibre-optic probes has the potential to provide rapid, objective diagnosis of dysplastic areas prompting tissue biopsy or polyp resection (Kallaway et al. 2013). There are many readily available probes that can be used to move the study of cancer a step ahead. For example, the 'Visionex probe' (Gaser Light Management System, Enviva Biomedical Raman Probes; Visionex Inc., Atlanta, GA) used for *in vivo* measurement of the Raman spectra from the colon (Song et al. 2005). These fibre-optic probes have enabled Raman spectroscopy instruments to be developed for routine clinical use. Furthermore, ATR-FTIR probe can also be developed and implemented in future research to provide real-time *in vivo* results, and by complementing the data from FTIR and Raman spectroscpic measurements, a model of high sensitivity and specificity is very likely to be developed. For colonoscopy, the commonly available endoscope channel has an internal diameter of 6 mm. As a result, in future work, a probe with a different IRE, i.e. diamond tip which is more expensive but smaller, needs to be designed for use in endoscopy for more comfort use (Mackanos & Contag 2010). Apart from the hardware advances, there are still many other challenges to making *in vivo* FTIR spectroscopic imaging a reality, such as the determination of the key wavelengths for measurements and tissue preparation, which need to be further investigated.

The other novel technique that can be employed to improve the work in this thesis is the 'Deep Raman spectroscopy', including the spatially offset Raman spectroscopy (SORS) [2], is another application that is being developed for non-invasive diagnosis of solid organs that can be achieved from outside of the body. It can be applied to study the complex scattering samples at depths from μm up to limits of 5 – 6 cm (Matousek et al. 2005). At this stage of development it can only be applied to specific applications where the Raman signature is very strong and distinct as there is a trade off with the intensity of the signature obtained from depth (Chakraborty et al. 2020), such as the chemical

[2]The basic SORS method involves making at least two Raman measurements; one at the source and one at an offset position of typically a few millimetres away. The two spectra can be subtracted using a scaled subtraction to produce two spectra representing the subsurface and surface spectra (Matousek et al. 2005).

characterisation of 'stone-like' materials in urology and cancer detection in a number of organs without the need for microtoming the samples (Matousek & Stone 2009).

5.2.6 Experimenting with advanced infrared source

The source of radiation for the IR spectrometers used for the experimental work in this thesis is the thermal (globar) radiation. One of the major highlights in recent year in the field of IR micro-spectroscopy is regarding the advancements in infrared sources, most notably the synchrotron radiation and quantum cascade lasers (QCLs) (Bhargava 2012). Compared to thermal sources, a synchrotron provides high-brightness radiation of a small emitted spot size[3]. There have been more than a dozen of the high-brightness third-generation synchrotron radiation facilities over the world, including the UK's national synchrotron facility, the Diamond Light Source, located at the Harwell Science and Innovation Campus in Oxfordshire (Diamond Light Source 2020) which makes research with this synchrotron source accessible. Significantly, improved image quality can also be observed when the synchrotron sources are coupled with FPA detector for imaging (Bhargava 2012).The high brightness of the synchrotron radiation source is also ideally suited to single-cell analysis (Doherty et al. 2019). The higher resolution that can be achieved enables diffraction-limited sub-cellular information to be obtained. Although synchrotron radiation gives added value to using IR spectroscopy for disease detection, for a wider applicability of IR micro-spectroscopy, commercial bench-top infrared instrumentation is still preferable, especially in a clinical environment.

The other high-power light source is the QCLs. QCLs are semiconductor lasers that utilises the emission of the photon tunnels into the next quantum well to generate multiple photons from a single electron, thereby making them extremely efficient (Faist et al. 1994). Chemical imaging with QCL based microscopes has been shown to give promising results in the field of biological systems (Kole et al. 2012, Kroger-Lui et al. 2015) and in particular, for the automated cancer classification in tissue sections (Kuepper et al. 2018). Therefore, using QCLs with micro ATR-FTIR imaging or corrected lens approach could enable fast and efficient measurements to be carried out (Kimber & Kazarian 2017). In this aspect, the groundwork that was described in this research thesis would be of great importance. Recently, an improvement to the QCLs, namely a miniaturized quantum cascade laser frequency comb that would achieve a much broader spectral coverage across mid- and far-IR was designed (Consolino et al. 2019). In future research, complementing this laser comb with microscope would make more advanced IR research on biological samples possible.

[3]Synchrotron radiation is an electromagnetic radiation emitting in the tangential direction of the track during the acceleration of charged particles (or electrons) at near light-speed (Zhu et al. 2017, Miller & Dumas 2013).

Bibliography

Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L. & Pounds, J. G. (2002), 'Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry', *Mol Cell Proteomics* **1**(12), 947–55.

Alturkistani, H. A., Tashkandi, F. M. & Mohammedsaleh, Z. M. (2015), 'Histological stains: A literature review and case study', *Glob J Health Sci* **8**(3), 72–9.

American Cancer Society (2020*a*), 'About colorectal cancer', `https://www.cancer.org/cancer/colon-rectal-cancer/about/what-is-colorectal-cancer.html`. Accessed Mar 12, 2020.

American Cancer Society (2020*b*), 'Key statistics for prostate cancer', `https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html`. Accessed Mar 12, 2020.

American Society of Clinical Oncology (ASCO) (2018), 'The genetics of cancer', `https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer`. Accessed Mar 11, 2020.

Argov, S., Sahu, R. K., Bernshtain, E., Salman, A., Shohat, G., Zelig, U. & Mordechai, S. (2004), 'Inflammatory bowel diseases as an intermediate stage between normal and cancer: a FTIR-microspectroscopy approach', *Biopolymers* **75**(5), 384–92.

Arneth, B. M. (2009), 'Clinical significance of measuring prostate-specific antigen', *Laboratory Medicine* **40**(8), 487–491.

Ashford, M. (2020), 'What does the prostate gland do?', `https://www.livescience.com/32751-what-does-the-prostate-gland-do.html`. Accessed Mar 11, 2020.

Atkins, P. W., De Paula, J. & Keeler, J. (2019), *Atkins' physical chemistry*, eleventh edn, OUP Oxford, Oxford, United Kingdom.

Averett, L. A., Griffiths, P. R. & Nishikida, K. (2008), 'Effective path length in attenuated total reflection spectroscopy', *Anal Chem* **80**(8), 3045–9.

Bailey, J. A., Dyer, R. B., Graff, D. K. & Schoonover, J. R. (2016), 'High spatial resolution for IR imaging using an IR diode laser', *Applied Spectroscopy* **54**(2), 159–163.

Baker, M. J., Byrne, H. J., Chalmers, J., Gardner, P., Goodacre, R., Henderson, A., Kazarian, S. G., Martin, F. L., Moger, J., Stone, N. & Sule-Suso, J. (2018), 'Clinical applications of infrared and Raman spectroscopy: state of play and future challenges', *Analyst* **143**(8), 1735–1757.

Baker, M. J., Clarke, C., Démoulin, D., Nicholson, J. M., Lyng, F. M., Byrne, H. J., Hart, C. A., Brown, M. D., Clarke, N. W. & Gardner, P. (2010), 'An investigation of the RWPE prostate derived family of cell lines using FTIR spectroscopy', *Analyst* **135**(5), 887–894.

Baker, M. J., Gazi, E., Brown, M. D., Shanks, J. H., Clarke, N. W. & Gardner, P. (2009), 'Investigating FTIR based histopathology for the diagnosis of prostate cancer', *J Biophotonics* **2**(1-2), 104–13.

Baker, M. J., Gazi, E., Brown, M. D., Shanks, J. H., Gardner, P. & Clarke, N. W. (2008), 'FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer', *Br J Cancer* **99**(11), 1859–66.

Baker, M. J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H. J., Dorling, K. M., Fielden, P. R., Fogarty, S. W., Fullwood, N. J., Heys, K. A., Hughes, C., Lasch, P., Martin-Hirsch, P. L., Obinaju, B., Sockalingum, G. D., Sule-Suso, J., Strong, R. J., Walsh, M. J., Wood, B. R., Gardner, P. & Martin, F. L. (2014), 'Using Fourier transform IR spectroscopy to analyze biological materials', *Nat Protoc* **9**(8), 1771–91.

Balbekova, A., Lohninger, H., Tilborg, G. A. F., Dijkhuizen, R. M., Bonta, M. & Limbeck, A. (2018), 'Fourier transform infrared (FT-IR) and laser ablation inductively coupled plasma–mass spectrometry (LA-ICP-MS) imaging of cerebral ischemia: Combined analysis of rat brain thin cuts toward improved tissue classification', *Appl Spectrosc* **72**(2), 241–250.

Balkwill, F. R., Capasso, M. & Hagemann, T. (2012), 'The tumor microenvironment at a glance', *Journal of Cell Science* **125**(23), 5591–5596.

Ball, D. W. (2003), *Physical chemistry*, first edn, Thomson-Brooks/Cole, Pacific Grove, CA.

Bambery, K. R., Wood, B. R. & McNaughton, D. (2012), 'Resonant Mie scattering (RMieS) correction applied to FTIR images of biological tissue samples', *Analyst* **137**(1), 126–32.

Bansil, R. & Turner, B. S. (2018), 'The biology of mucus: Composition, synthesis and organization', *Adv Drug Deliv Rev* **124**, 3–15.

Bassan, P., Byrne, H. J., Bonnier, F., Lee, J., Dumas, P. & Gardner, P. (2009), 'Resonant Mie scattering in infrared spectroscopy of biological materials–understanding the 'dispersion artefact", *Analyst* **134**(8), 1586–93.

Bassan, P. & Gardner, P. (2010), Chapter 8, scattering in biomedical infrared spectroscopy, *in* 'Biomedical Applications of Synchrotron Infrared Microspectroscopy: A Practical Approach', The Royal Society of Chemistry, pp. 260–276.

Bassan, P., Kohler, A., Martens, H., Lee, J., Byrne, H. J., Dumas, P., Gazi, E., Brown, M., Clarke, N. & Gardner, P. (2010), 'Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples', *Analyst* **135**(2), 268–77.

Bassan, P., Kohler, A., Martens, H., Lee, J., Jackson, E., Lockyer, N., Dumas, P., Brown, M., Clarke, N. & Gardner, P. (2010), 'RMieS-EMSC correction for infrared spectra of biological cells: extension using full Mie theory and GPU computing', *J Biophotonics* **3**(8-9), 609–20.

Bassan, P., Lee, J., Sachdeva, A., Pissardini, J., Dorling, K. M., Fletcher, J. S., Henderson, A. & Gardner, P. (2013), 'The inherent problem of transflection-mode infrared spectroscopic microscopy and the ramifications for biomedical single point and imaging applications', *Analyst* **138**(1), 144–57.

Bassan, P., Mellor, J., Shapiro, J., Williams, K. J., Lisanti, M. P. & Gardner, P. (2014), 'Transmission FT-IR chemical imaging on glass substrates: Applications in infrared spectral histopathology', *Anal Chem* **86**(3), 1648–1653.

Bassan, P., Sachdeva, A., Kohler, A., Hughes, C., Henderson, A., Boyle, J., Shanks, J. H., Brown, M., Clarke, N. W. & Gardner, P. (2012), 'FTIR microscopy of biological cells and tissue: data analysis using resonant Mie scattering (RMieS) EMSC algorithm', *Analyst* **137**(6), 1370–7.

Bast, R. C., Ravdin, P., Hayes, D. F., Bates, S., Fritsche, H., Jessup, J. M., Kemeny, N., Locker, G. Y., Mennel, R. G. & Somerfield, M. R. (2001), '2000 Update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology', *Journal of Clinical Oncology* **19**(6), 1865–1878.

Bath, M. (2019), 'The prostate gland?', `https://teachmeanatomy.info/pelvis/the-male-reproductive-system/prostate-gland/`. Accessed Mar 12, 2020.

Bel'skaya, L. V. (2019), 'Use of IR spectroscopy in cancer diagnosis. A review', *Journal of Applied Spectroscopy* **86**, 187–205.

Berg, J., Tymoczko, J. & Stryer, L. (2002), Section 27.1, DNA can assume a variety of structural forms, *in* 'Biochemistry', fifth edn, W H Freeman, New York.

Berisha, S., Lotfollahi, M., Jahanipour, J., Gurcan, I., Walsh, M. & Bhargava, R. (2019), 'Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks', *Analyst* **144**(5), 1642–1653.

Berrueta, L. A., Alonso-Salces, R. M. & Heberger, K. (2007), 'Supervised pattern recognition in food analysis', *J Chromatogr A* **1158**(1-2), 196–214.

Bertie, J. E. & Lan, Z. (1996), 'An accurate modified Kramers–Kronig transformation from reflectance to phase shift on attenuated total reflection', *The Journal of Chemical Physics* **105**(19), 8502–8514.

Bhargava, R. (2012), 'Infrared spectroscopic imaging: the next generation', *Applied spectroscopy* **66**(10), 1091–1120.

BioChain Institute, Inc. (2018), 'FFPE vs frozen tissue samples', `https://www.biochain.com/general/ffpe-vs-frozen-tissue-samples/`. Accessed Mar 20, 2020.

Bird, B., Miljkovic, M., Romeo, M. J., Smith, J., Stone, N., George, M. W. & Diem, M. (2008), 'Infrared micro-spectral imaging: distinction of tissue types in axillary lymph node histology', *BMC clinical pathology* **8**(8).

Blumel, R., Bagcioglu, M., Lukacs, R. & Kohler, A. (2016), 'Infrared refractive index dispersion of polymethyl methacrylate spheres from Mie ripples in Fourier-transform infrared microscopy extinction spectra', *J Opt Soc Am A Opt Image Sci Vis* **33**(9), 1687–96.

Bogomolny, E., Huleihel, M., Suproun, Y., Sahu, R. K. & Mordechai, S. (2007), 'Early spectral changes of cellular malignant transformation using fourier transform infrared microspectroscopy', *Journal of Biomedical Optics* **12**(2), 024003.

Bolin, F. P., Preuss, L. E., Taylor, R. C. & Ference, R. J. (1989), 'Refractive index of some mammalian tissues using a fiber optic cladding method', *Appl Opt* **28**(12), 2297–303.

Bonnier, F., Petitjean, F., Baker, M. J. & Byrne, H. J. (2014), 'Improved protocols for vibrational spectroscopic analysis of body fluids', *Journal of Biophotonics* **7**(3-4), 167–179.

Bose, R., Kavuri, S. M., Searleman, A. C., Shen, W., Shen, D., Koboldt, D. C., Monsey, J., Goel, N., Aronson, A. B., Li, S., Ma, C. X., Ding, L., Mardis, E. R. & Ellis, M. J. (2013), 'Activating HER2 mutations in HER2 gene amplification negative breast cancer', *Cancer Discovery* **3**(2), 224.

Boulet-Audet, M., Buffeteau, T., Boudreault, S., Daugey, N. & Pezolet, M. (2010), 'Quantitative determination of band distortions in diamond attenuated total reflectance infrared spectra', *J Phys Chem B* **114**(24), 8255–61.

Bradford, A. (2016), 'Colon (large intestine): facts, function & diseases', https://www.livescience.com/52026-colon-large-intestine.html. Accessed Mar 12, 2020.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A. & Jemal, A. (2018), 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: A Cancer Journal for Clinicians* **68**(6), 394–424.

Breiman, L. (2001), 'Random forests', *Mach Learn* **45**(1), 5–32.

Brigham, E. O. (1988), *The fast Fourier transform and its applications*, Prentice Hall, Englewood Cliffs, N.J.

Brown, M. L., Lipscomb, J. & Snyder, C. (2001), 'The burden of illness of cancer: economic cost and quality of life', *Annual Review of Public Health* **22**(1), 91–113.

Brown, R. E., Short, S. P. & Williams, C. S. (2018), 'Colorectal cancer and metabolism', *Current Colorectal Cancer Reports* **14**(6), 226–241.
URL: *https://doi.org/10.1007/s11888-018-0420-y*

Brugel, W. (1965), 'Chemical applications of infrared spectroscopy', *Angewandte Chemie* **77**(8), 391–391.

Bruun, S. W., Kohler, A., Adt, I., Sockalingum, G. D., Manfait, M. & Martens, H. (2006), 'Correcting attenuated total reflection-Fourier transform infrared spectra for water vapor and carbon dioxide', *Appl Spectrosc* **60**(9), 1029–39.

Burch, C. R. (1947), 'Reflecting microscopes', *Proceedings of the Physical Society* **59**(1), 41–46–2.

Butler, H. J., Brennan, P. M., Cameron, J. M., Finlayson, D., Hegarty, M. G., Jenkinson, M. D., Palmer, D. S., Smith, B. R. & Baker, M. J. (2019), 'Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer', *Nature Communications* **10**(1), 4501.

Byrnes, J. M., Ramsey, M. S., King, P. L. & Lee, R. J. (2007), 'Thermal infrared reflectance and emission spectroscopy of quartzofeldspathic glasses', *Geophysical Research Letters* **34**(1), L01306.

Cameron, J. M., Butler, H. J., Palmer, D. S. & Baker, M. J. (2018), 'Biofluid spectroscopic disease diagnostics: A review on the processes and spectral impact of drying', *Journal of Biophotonics* **11**(4), e201700299.

Canadian Cancer Society (2020), 'The prostate', https://www.cancer.ca/en/cancer-information/cancer-type/prostate/prostate-cancer/the-prostate/?region=on. Accessed Mar 11, 2020.

Cancer Research UK (2017*a*), 'Bowel cancer statistics', https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero. Accessed Mar 12, 2020.

Cancer Research UK (2017*b*), 'Prostate cancer statistics', https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer#heading-Zero. Accessed Mar 12, 2020.

Cappuzzo, F., Varella-Garcia, M., Shigematsu, H., Domenichini, I., Bartolini, S., Ceresoli, G. L., Rossi, E., Ludovini, V., Gregorc, V., Toschi, L., Franklin, W. A., Crino, L., Gazdar, A. F., Bunn, P. A. & Hirsch, F. R. (2005), 'Increased HER2 gene copy number is associated with response to gefitinib therapy in epidermal growth factor receptor–positive non–small-cell lung cancer patients', *Journal of Clinical Oncology* **23**(22), 5007–5018.

Carson, Culley, I. & Rittmaster, R. (2003), 'The role of dihydrotestosterone in benign prostatic hyperplasia', *Urology* **61**(4), 2–7.

Castro, E. & Eeles, R. (2012), 'The role of BRCA1 and BRCA2 in prostate cancer', *Asian journal of andrology* **14**(3), 409–414.

Catalona, W. J., Smith, D. S., Ratliff, T. L., Dodds, K. M., Coplen, D. E., Yuan, J. J. J., Petros, J. A. & Andriole, G. L. (1991), 'Measurement of prostate-specific antigen in serum as a screening test for prostate cancer', *New England Journal of Medicine* **324**(17), 1156–1161.

Centers for Disease Control and Prevention (CDC) (2020), 'Hereditary colorectal (colon) cancer', https://www.cdc.gov/genomics/disease/colorectal_cancer/lynch.htm. Accessed Mar 12, 2020.

Chaber, R., Łach, K., Depciuch, J., Szmuc, K., Michalak, E., Raciborska, A., Koziorowska, A. & Cebulski, J. (2017), 'Fourier Transform Infrared (FTIR) spectroscopy of paraffin and deparafinnized bone tissue samples as a diagnostic tool for Ewing sarcoma of bones', *Infrared Physics & Technology* **85**, 364–371.

Chaffer, C. L. & Weinberg, R. A. (2011), 'A perspective on cancer cell metastasis', *Science* **331**(6024), 1559.

Chakraborty, A., Ghosh, A. & Barui, A. (2020), 'Advances in surface-enhanced Raman spectroscopy for cancer diagnosis and staging', *Journal of Raman Spectroscopy* **51**(1), 7–36.

Chalmers, J. M. & Griffiths, P. R. (2002), *Handbook of vibrational spectroscopy. Theory and instrumentation. Vol. 1*, John Wiley & Sons Ltd, Chichester.

Chalmers, J. M. & Griffiths, P. R. (2006), *Handbook of Vibrational Spectroscopy*, John Wiley & Sons Ltd, Chichester.

Chan, J. K. C. (2014), 'The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology', *International Journal of Surgical Pathology* **22**(1), 12–32.

Chan, K. L. A., Altharawi, A., Fale, P., Song, C. L., Kazarian, S. G., Cinque, G., Untereiner, V. & Sockalingum, G. D. (2020), 'Transmission Fourier transform infrared spectroscopic imaging, mapping, and synchrotron scanning microscopy with zinc sulfide hemispheres on living mammalian cells at sub-cellular resolution', *Applied Spectroscopy* **74**(5), 544–552.

Chan, K. L. A., Fleming, O. S., Kazarian, S. G., Vassou, D., Chryssikos, G. D. & Gionis, V. (2004), 'Polymorphism and devitrification of nifedipine under controlled humidity: a combined FT-Raman, IR and Raman microscopic investigation', *Journal of Raman Spectroscopy* **35**(5), 353–359.

Chan, K. L. A. & Kazarian, S. G. (2003), 'New opportunities in micro- and macro-attenuated total reflection infrared spectroscopic imaging: spatial resolution and sampling versatility', *Appl Spectrosc* **57**(4), 381–9.

Chan, K. L. A. & Kazarian, S. G. (2004), 'Visualisation of the heterogeneous water sorption in a pharmaceutical formulation under controlled humidity via FT-IR imaging', *Vibrational Spectroscopy* **35**(1-2), 45–49.

Chan, K. L. A. & Kazarian, S. G. (2006), 'High-throughput study of poly(ethylene glycol)/ibuprofen formulations under controlled environment using FTIR imaging', *J Comb Chem* **8**(1), 26–31.

Chan, K. L. A. & Kazarian, S. G. (2007*a*), 'Attenuated total reflection fourier transform infrared imaging with variable angles of incidence: a three-dimensional profiling of heterogeneous materials', *Appl Spectrosc* **61**(1), 48–54.

Chan, K. L. A. & Kazarian, S. G. (2007*b*), 'Chemical imaging of the stratum corneum under controlled humidity with the attenuated total reflection fourier transform infrared spectroscopy method', *J Biomed Opt* **12**(4), 044010.

Chan, K. L. A. & Kazarian, S. G. (2008), 'Attenuated total reflection-fourier transform infrared imaging of large areas using inverted prism crystals and combining imaging and mapping', *Appl Spectrosc* **62**(10), 1095–101.

Chan, K. L. A. & Kazarian, S. G. (2013), 'Correcting the effect of refraction and dispersion of light in FT-IR spectroscopic imaging in transmission through thick infrared windows', *Anal Chem* **85**(2), 1029–36.

Chasteen, T. G. (2009), 'A double beam spectrometer', `https://www.shsu.edu/~chm_tgc/primers/spect.html`. Accessed Mar 19, 2020.

Chen, H., Lin, Z., Wu, H., Wang, L., Wu, T. & Tan, C. (2015), 'Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest', *Spectrochim Acta A Mol Biomol Spectrosc* **135**.

Chen, N. & Zhou, Q. (2016), 'The evolving Gleason grading system', *Chinese journal of cancer research = Chung-kuo yen cheng yen chiu* **28**(1), 58–64.

Chiriboga, L., Xie, P., Yee, H., Vigorita, V., Zarou, D., Zakim, D. & Diem, M. (1998), 'Infrared spectroscopy of human tissue. I. Differentiation and maturation of epithelial cells in the human cervix', *Biospectroscopy* **4**(1), 47–53.

Christensen, P. R., Bandfield, J. L., Hamilton, V. E., Ruff, S. W., Kieffer, H. H., Titus, T. N., Malin, M. C., Morris, R. V., Lane, M. D., Clark, R. L., Jakosky, B. M., Mellon, M. T., Pearl, J. C., Conrath, B. J., Smith, M. D., Clancy, R. T., Kuzmin, R. O., Roush, T., Mehall, G. L., Gorelick, N., Bender, K., Murray, K., Dason, S., Greene, E., Silverman, S. & Greenfield, M. (2001), 'Mars global surveyor thermal emission spectrometer experiment: Investigation description and surface science results', *Journal of Geophysical Research: Planets* **106**(E10), 23823–23871.

Christou, C., Agapiou, A. & Kokkinofta, R. (2018), 'Use of FTIR spectroscopy and chemometrics for the classification of carobs origin', *J Adv Res* **10**, 1–8.

Clancy, S. (2008), 'Genetic mutation', *Nature Education* **1**(1), 1.

Coleman, P. B. (1993), *Practical Sampling Techniques for Infrared Analysis*, first edn, CRC Press, Taylor & Francis, Boca Raton, Florida.

Consolino, L., Nafa, M., Cappelli, F., Garrasi, K., Mezzapesa, F. P., Li, L., Davies, A. G., Linfield, E. H., Vitiello, M. S., De Natale, P. & Bartalini, S. (2019), 'Fully phase-stabilized quantum cascade laser frequency comb', *Nature Communications* **10**(1), 2938.

Cooper, G. M. (2000), *The cell : a molecular approach*, Sinauer Associates, Massachusetts.

Coulson, C. A. & Robertson, G. N. (1974), 'A theory of the broadening of the infrared absorption spectra of hydrogen-bonded species. I', *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **337**(1609), 167–197.

Cui, S., Zhang, S. & Yue, S. (2018), 'Raman spectroscopy and imaging for cancer diagnosis', *Journal of Healthcare Engineering Journal of Healthcare Engineering* **2018**, 1–11.

de Lima, F. A., Gobinet, C., Sockalingum, G., Garcia, S. B., Manfait, M., Untereiner, V., Piot, O. & Bachmann, L. (2017), 'Digital de-waxing on FTIR images', *Analyst* **142**(8), 1358–1370.

DeBerardinis, R. J. & Chandel, N. S. (2016), 'Fundamentals of cancer metabolism', *Science Advances* **2**(5), e1600200.
URL: *http://advances.sciencemag.org/content/2/5/e1600200.abstract*

Deckert, V., Deckert-Gaudig, T., Diegel, M., Götz, I., Langelüddecke, L., Schneidewind, H., Sharma, G., Singh, P., Singh, P., Trautmann, S., Zeisberger, M. & Zhang, Z. (2015), 'Spatial resolution in Raman spectroscopy', *Faraday Discussions* **177**(0), 9–20.

Delahunt, B., Srigley, J. R. & Lamb, D. S. (2009), 'Gleason grading: consensus and controversy', *Pathology* **41**(7), 613–614.

Depciuch, J., Kaznowska, E., Koziorowska, A. & Cebulski, J. (2017), 'Verification of the effectiveness of the fourier transform infrared spectroscopy computational model for colorectal cancer', *J Pharm Biomed Anal* **145**, 611–615.

Depciuch, J., Kaznowska, E., Szmuc, K., Zawlik, I., Cholewa, M., Heraud, P. & Cebulski, J. (2016), 'Comparing paraffined and deparaffinized breast cancer tissue samples and an analysis of Raman spectroscopy and infrared methods', *Infrared Physics & Technology* **76**, 217–226.

Diamond Light Source (2020), 'Diamond light source', `https://www.diamond.ac.uk/Home.html;jsessionid=AFF5B67127B698436BE5F442B5442416`. Accessed Apr 1, 2020.

Diem, M., Romeo, M., Boydston-White, S., Miljkovic, M. & Matthaus, C. (2004), 'A decade of vibrational micro-spectroscopy of human cells and tissue (1994-2004)', *Analyst* **129**(10), 880–5.

Doherty, J., Raoof, A., Hussain, A., Wolna, M., Cinque, G., Brown, M., Gardner, P. & Denbigh, J. (2019), 'Live single cell analysis using synchrotron FTIR microspectroscopy: development of a simple dynamic flow system for prolonged sample viability', *Analyst* **144**(3), 997–1007.

Dorling, K. M. & Baker, M. J. (2013), 'Rapid FTIR chemical imaging: highlighting FPA detectors', *Trends Biotechnol* **31**(8), 437–8.

Dorée, M. & Galas, S. (1994), 'The cyclin-dependent protein kinases and the control of cell division', *The FASEB Journal* **8**(14), 1114–1121.

Dovbeshko, G. (2000), 'FTIR spectroscopy studies of nucleic acid damage', *Talanta* **53**(1), 233–246.

Downing, H. D. & Williams, D. (1975), 'Optical constants of water in the infrared', *Journal of Geophysical Research* **80**(12), 1656–1661.

Dyachenko, P. N., Molesky, S., Petrov, A. Y., Stormer, M., Krekeler, T., Lang, S., Ritter, M., Jacob, Z. & Eich, M. (2016), 'Controlling thermal emission with refractory epsilon-near-zero metamaterials via topological transitions', *Nat Commun* **7**, 11809.

El-Azazy, M. (2019), Introductory chapter: infrared spectroscopy - a Synopsis of the fundamentals and applications, *in* 'Infrared spectroscopy: principles, advances, and applications', IntechOpen.

Elliott, A. & Ambrose, E. J. (1950), 'Structure of synthetic polypeptides', *Nature* **165**(4206), 921–922.

English, R. S. (2018), 'A hypothetical pathogenesis model for androgenic alopecia: clarifying the dihydrotestosterone paradox and rate-limiting recovery factors', *Medical Hypotheses* **111**, 73–81.

Ewing, A. V., Gabrienko, A. A., Semikolenov, S. V., Dubkov, K. A. & Kazarian, S. G. (2015), 'How do intermolecular interactions affect swelling of polyketones with a differing number of carbonyl groups? An in situ ATR-FTIR spectroscopic study of CO2 sorption in polymers', *The Journal of Physical Chemistry C* **119**(1), 431–440.

Ewing, A. V. & Kazarian, S. G. (2017), 'Infrared spectroscopy and spectroscopic imaging in forensic science', *Analyst* **142**(2), 257–272.

Ewing, A. V. & Kazarian, S. G. (2018), 'Current trends and opportunities for the applications of in situ vibrational spectroscopy to investigate the supercritical fluid processing of polymers', *The Journal of Supercritical Fluids* **134**, 88–95.

Fabian, H., Jackson, M., Murphy, L., Watson, P. H., Fichtner, I. & Mantsch, H. H. (1995), 'A comparative infrared spectroscopic study of human breast tumors and breast tumor cell xenografts', *Biospectroscopy* **1**(1), 37–45.

Fabian, H., Lasch, P. & Naumann, D. (2005), 'Analysis of biofluids in aqueous environment based on mid-infrared spectroscopy', *Journal of biomedical optics* **10**(3), 031103.

Fahmy, K. (2013), 'Fourier transform infrared spectroscopy for biophysical applications: Technical aspects', pp. 844–852.

Faist, J., Capasso, F., Sivco, D. L., Sirtori, C., Hutchinson, A. L. & Cho, A. Y. (1994), 'Quantum cascade laser', *Science* **264**(5158), 553.

Falk, M., Hartman, K. A. & Lord, R. C. (1963), 'Hydration of deoxyribonucleic acid. ii. an infrared study', *Journal of the American Chemical Society* **85**(4), 387–391.

Faolain, E. O., Hunter, M. B., Byrne, J. M., Kelehan, P., Lambkin, H. A., Byrne, H. J. & Lyng, F. M. (2005), 'Raman spectroscopic evaluation of efficacy of current paraffin wax section dewaxing agents', *J Histochem Cytochem* **53**(1), 121–9.

Feldman, A. T. & Wolfe, D. (2014), 'Tissue processing and hematoxylin and eosin staining', *Methods Mol Biol* **1180**, 31–43.

Fellgett, P. B. (1949), 'On the ultimate sensitivity and practical performance of radiation detectors', *J Opt Soc Am* **39**(11), 970–6.

Fischer, A. H., Jacobson, K. A., Rose, J. & Zeller, R. (2008), 'Hematoxylin and eosin staining of tissue and cell sections', *CSH Protoc* **2008**, pdb.prot4986.

Fringelli, U. (2000), ATR and reflectance IR spectroscopy, applications, *in* J. C. Lindon, G. E. Tranter & J. L. Holmes, eds, 'Encyclopedia of spectroscopy and spectrometry', Vol. 1, Academic Press, United Kingdom, pp. 58–75.

Fu, J., Jiang, Q. & Zhang, C. (2010), 'Collaboration of mitotic kinases in cell cycle control', *Nature Education* **3**(9), 1.

Fung, M. F. K., Senterman, M. K., Mikhael, N. Z., Lacelle, S. & Wong, P. T. T. (1996), 'Pressure-tuning fourier transform infrared spectroscopic study of carcinogenesis in human endometrium', *Biospectroscopy* **2**(3), 155–165.

Gaigneaux, A. & Goormaghtigh, E. (2013), 'A new dimension for cell identification by FTIR spectroscopy: depth profiling in attenuated total reflection', *Analyst* **138**(14), 4070–5.

Gao, Y., Huo, X., Dong, L., Sun, X., Sai, H., Wei, G., Xu, Y., Zhang, Y. & Wu, J. (2015), 'Fourier transform infrared microspectroscopy monitoring of 5-fluorouracil-induced apoptosis in SW620 colon cancer cells', *Molecular medicine reports* **11**(4), 2585–2591.

Gautam, R., Vanga, S., Ariese, F. & Umapathy, S. (2015), 'Review of multidimensional data processing approaches for Raman and infrared spectroscopy', *EPJ Techniques and Instrumentation* **2**(1), 8.

Gazi, E., Baker, M., Dwyer, J., Lockyer, N. P., Gardner, P., Shanks, J. H., Reeve, R. S., Hart, C. A., Clarke, N. W. & Brown, M. D. (2006), 'A correlation of FTIR spectra derived from prostate cancer biopsies with gleason grade and tumour stage', *Eur Urol* **50**(4), 750–60; discussion 760–1.

Gazi, E., Dwyer, J., Gardner, P., Ghanbari-Siahkali, A., Wade, A. P., Miyan, J., Lockyer, N. P., Vickerman, J. C., Clarke, N. W., Shanks, J. H., Scott, L. J., Hart, C. A. & Brown, M. (2003), 'Applications of Fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer. a pilot study', *J Pathol* **201**(1), 99–108.

Gazi, E., Dwyer, J., Lockyer, N., Gardner, P., Vickerman, J. C., Miyan, J., Hart, C. A., Brown, M., Shanks, J. H. & Clarke, N. (2004), 'The combined application of FTIR microspectroscopy and ToF-SIMS imaging in the study of prostate cancer', *Faraday Discuss* **126**, 41–59; discussion 77–92.

Geladi, P., MacDougall, D. & Martens, H. (2016), 'Linearization and scatter-correction for near-infrared reflectance spectra of meat', *Applied Spectroscopy* **39**(3), 491–500.

Geneticist Inc. (2018), 'The pros and cons of FFPE vs frozen tissue samples', `https://www.geneticistinc.com/blog/the-pros-and-cons-of-ffpe-vs-frozen-tissue-samples`. Accessed Mar 20, 2020.

German, M. J., Hammiche, A., Ragavan, N., Tobin, M. J., Cooper, L. J., Matanhelia, S. S., Hindley, A. C., Nicholson, C. M., Fullwood, N. J., Pollock, H. M. & Martin, F. L. (2006), 'Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell', *Biophys J* **90**(10), 3783–95.

Ghasemi, A. & Zahediasl, S. (2012), 'Normality tests for statistical analysis: a guide for non-statisticians', *Int J Endocrinol Metab* **10**(2), 486–9.

Ghimire, H., Venkataramani, M., Bian, Z., Liu, Y. & Perera, A. G. U. (2017), 'ATR-FTIR spectral discrimination between normal and tumorous mouse models of lymphoma and melanoma from serum samples', *Scientific Reports* **7**(1), 16993.

Gioacchini, G., Giorgini, E., Vaccari, L., Ferraris, P., Sabbatini, S., Bianchi, V., Borini, A. & Carnevali, O. (2014), 'A new approach to evaluate aging effects on human oocytes: Fourier transform infrared imaging spectroscopy study', *Fertil Steril* **101**(1), 120–7.

Giorgini, E., Sabbatini, S., Conti, C., Rubini, C., Rocchetti, R., Fioroni, M., Meme, L. & Orilisi, G. (2017), 'Fourier Transform Infrared Imaging analysis of dental pulp inflammatory diseases', *Oral Dis* **23**(4), 484–491.

Glassford, S. E., Byrne, B. & Kazarian, S. G. (2013), 'Recent applications of ATR FTIR spectroscopy and imaging to proteins', *Biochim Biophys Acta* **1834**(12), 2849–58.

Gleason, D. F. & Mellinger, G. T. (1974), 'Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and cinical staging', *Journal of Urology* **111**(1), 58–64.

Gonsales, A. (2018), 'An approach to choosing the number of components in a principal component analysis', `https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer`. Accessed Mar 25, 2020.

Goodacre, R. (2003), 'Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules', *Vib Spectrosc* **32**(1), 33–45.

Grabska, J., Ishigaki, M., Beć, K. B., Wójcik, M. J. & Ozaki, Y. (2017), 'Correlations between structure and near-infrared spectra of saturated and unsaturated carboxylic acids. Insight from anharmonic density functional theory calculations', *The Journal of Physical Chemistry A* **121**(18), 3437–3451.

Granato, D., Putnik, P., Kovačević, D. B., Santos, J. S., Calado, V., Rocha, R. S., Cruz, A. G. D., Jarvis, B., Rodionova, O. Y. & Pomerantsev, A. (2018), 'Trends in chemometrics: Food authentication, microbiology, and effects of processing', *Comprehensive Reviews in Food Science and Food Safety* **17**(3), 663–677.

Greenspan, L. (1977), 'Humidity fixed points of binary saturated aqueous solutions', *Journal of Research of the National Bureau of Standards Section A: Physics and Chemistry* **81A**(1), 89.

Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. (2000), *An introduction to genetic analysis*, seventh edn, W. H. Freeman and Company, New York.

Griffiths, P. R. & de Haseth, J. A. (2007), *Fourier Transform Infrared Spectrometry*, second edn, John Wiley & Sons, Inc.

Gutierrez, M. C. (1992), 'Derivative spectroscopy applied to the determination of alpha- and beta-acids in hops', *Journal of the Institute of Brewing* **98**(4), 277–281.

Hall, J. P., Sanchez-Weatherby, J., Alberti, C., Quimper, C. H., O'Sullivan, K., Brazier, J. A., Winter, G., Sorensen, T., Kelly, J. M., Cardin, D. J. & Cardin, C. J. (2014), 'Controlled dehydration of a ruthenium complex–DNA crystal induces reversible DNA kinking', *Journal of the American Chemical Society* **136**(50), 17505–17512.

Hanahan, D. & Weinberg, R. A. (2011), 'Hallmarks of cancer: the next generation', *Cell* **144**(5), 646–674.

Hancer, M., Sperline, R. P. & Miller, J. D. (2016), 'Anomalous dispersion effects in the IR-ATR spectroscopy of water', *Applied Spectroscopy* **54**(1), 138–143.

Hands, J. R., Clemens, G., Stables, R., Ashton, K., Brodbelt, A., Davis, C., Dawson, T. P., Jenkinson, M. D., Lea, R. W., Walker, C. & Baker, M. J. (2016), 'Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy', *Journal of neuro-oncology* **127**(3), 463–472.

Harrick, N. J. & Carlson, A. I. (1971), 'Internal reflection spectroscopy: validity of effective thickness equations', *Appl Opt* **10**(1), 19–23.

Harrick, N. J. & du Pre, F. K. (1966), 'Effective thickness of bulk materials and of thin films for internal reflection spectroscopy', *Appl Opt* **5**(11), 1739–43.

Hartmann, M. (1984), 'Light scattering by small particles', *Acta Polymerica* **35**(4), 338–338.

Harvey, T. J., Henderson, A., Gazi, E., Clarke, N. W., Brown, M., Faria, E. C., Snook, R. D. & Gardner, P. (2007), 'Discrimination of prostate cancer cells by reflection mode FTIR photoacoustic spectroscopy', *Analyst* **132**(4), 292–5.

Hatcher, D., Daniels, G., Osman, I. & Lee, P. (2009), 'Molecular mechanisms involving prostate cancer racial disparity', *American journal of translational research* **1**(3), 235–248.

Hayes, J. H. & Barry, M. J. (2014), 'Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence', *JAMA* **311**(11), 1143–1149.

Hermes, M., Morrish, R. B., Huot, L., Meng, L., Junaid, S., Tomko, J., Lloyd, G. R., Masselink, W. T., Tidemand-Lichtenberg, P., Pedersen, C., Palombo, F. & Stone, N. (2018), 'Mid-IR hyperspectral imaging for label-free histopathology and cytology', *Journal of Optics* **20**(2), 023002.

Hirsh, S. L., McKenzie, D. R., Nosworthy, N. J., Denman, J. A., Sezerman, O. U. & Bilek, M. M. M. (2013), 'The Vroman effect: Competitive protein exchange with dynamic multilayer protein aggregates', *Colloids and Surfaces B: Biointerfaces* **103**, 395–404.

Hormigo, A., Gu, B., Karimi, S., Riedel, E., Panageas, K. S., Edgar, M. A., Tanwar, M. K., Rao, J. S., Fleisher, M., DeAngelis, L. M. & Holland, E. C. (2006), 'YKL-40 and matrix Metalloproteinase-9 as potential serum biomarkers for patients with high-grade gliomas', *Clinical Cancer Research* **12**(19), 5698.

Horosh, M., Feldman, H., Yablonovich, A., Firer, M. A. & Abookasis, D. (2016), 'Broadband infrared spectroscopy for non-contact measurement of neurological disease biomarkers in cerebrospinal fluid', *Applied Spectroscopy* **71**(3), 496–506.

Horvath, H. (2009), 'Gustav Mie and the scattering and absorption of light by particles: Historic developments and basics', *Journal of Quantitative Spectroscopy and Radiative Transfer* **110**(11), 787–799.

Huang, B., Song, B.-l. & Xu, C. (2020), 'Cholesterol metabolism in cancer: mechanisms and therapeutic opportunities', *Nature Metabolism* **2**(2), 132–141.

Huang, J. B. & Urban, M. W. (1992), 'Evaluation and analysis of attenuated total reflectance FT-IR spectra using Kramers-Kronig Transforms', *Applied Spectroscopy* **46**(11), 1666–1672.

Hughes, C., Gaunt, L., Brown, M., Clarke, N. W. & Gardner, P. (2014), 'Assessment of paraffin removal from prostate FFPE sections using transmission mode FTIR-FPA imaging', *Anal. Methods* **6**(4), 1028–1035.

Humphrey, P. A. (2004), 'Gleason grading and prognostic factors in carcinoma of the prostate', *Modern Pathology* **17**(3), 292–306.

Ingle, J. D. & Crouch, S. R. (1988), *Spectrochemical analysis*, Prentice-Hall, Englewood Cliffs, N.J.

Jacques, S. L. (2013), 'Optical properties of biological tissues: a review', *Phys Med Biol* **58**(11), R37–61.

Jones, R. C. (1960), 'Proposal of the detectivity D** for detectors limited by radiation noise†', *Journal of the Optical Society of America* **50**(11), 1058.

Josef Marx, F. & Karenberg, A. (2009), 'History of the term prostate', *The Prostate* **69**(2), 208–213.

Kallaway, C., Almond, L. M., Barr, H., Wood, J., Hutchings, J., Kendall, C. & Stone, N. (2013), 'Advances in the clinical application of Raman spectroscopy for cancer diagnostics', *Photodiagnosis and Photodynamic Therapy* **10**(3), 207–219.

Kallenbach-Thieltges, A., Grosseruschkamp, F., Mosig, A., Diem, M., Tannapfel, A. & Gerwert, K. (2013), 'Immunohistochemistry, histopathology and infrared spectral histopathology of colon cancer tissue sections', *J Biophotonics* **6**(1), 88–100.

Kazarian, A., Blyuss, O., Metodieva, G., Gentry-Maharaj, A., Ryan, A., Kiseleva, E. M., Prytomanova, O. M., Jacobs, I. J., Widschwendter, M., Menon, U. & Timms, J. F. (2017), 'Testing breast cancer serum biomarkers for early detection and prognosis in pre-diagnosis samples', *British journal of cancer* **116**(4), 501–508.

Kazarian, S. G. & Chan, K. L. (2013), 'ATR-FTIR spectroscopic imaging: recent advances and applications to biological systems', *Analyst* **138**(7), 1940–51.

Kazarian, S. G. & Chan, K. L. A. (2006), 'Applications of ATR-FTIR spectroscopic imaging to biomedical samples', *Biochim Biophys Acta* **1758**(7), 858–67.

Kazarian, S. G. & Chan, K. L. A. (2010), 'Micro- and macro-attenuated total reflection fourier transform infrared spectroscopic imaging. plenary lecture at the 5th international conference on advanced vibrational spectroscopy, 2009, melbourne, australia', *Appl Spectrosc* **64**(5), 135A–152A.

Kaznowska, E., Depciuch, J., Szmuc, K. & Cebulski, J. (2017), 'Use of FTIR spectroscopy and PCA-LDC analysis to identify cancerous lesions within the human colon', *J Pharm Biomed Anal* **134**, 259–268.

Ke, X. & Shen, L. (2017), 'Molecular targeted therapy of cancer: The progress and future prospect', *Frontiers in Laboratory Medicine* **1**(2), 69–75.

Kempfert, K. D. (2004), 'Performance and application comparisons for single and three reflection diamond crystals for the MIRacle^{TM} ATR accessory', `http://files.alfresco.mjh.group/alfresco_images/pharma/2014/08/22/66df65af-9a7a-40c1-b724-b4741e89602b/article-125605.pdf`. Accessed Mar 11, 2020.

Kempfert, K. D., Jiang, E. Y., Oas, S., Coffin, J. & Thermo Nicolet Spectroscopy Research Center (2001), 'Detectors for Fourier transform spectroscopy', `http://kinecat.pl/wp-content/uploads/2012/11/IR_detectors.pdf`. Accessed Mar 20, 2020.

Kendix, E. L. (2009), Trasmission and Reflection (ATR) Far-Infrared Spectroscopy Applied in the Analysis of Cultural Heritage Materials, Doctoral thesis, Alma Mater Studiorum Università di Bologna.

Khan, S. A., Khan, S. B., Khan, L. U., Farooq, A., Akhtar, K. & Asiri, A. M. (2018), Fourier transform infrared spectroscopy: Fundamentals and application in functional groups and nanomaterials characterization, *in* S. K. Sharma, ed., 'Handbook of Materials Characterization', Springer International Publishing, Cham, pp. 317–344.

Kim, H.-D., Tomida, A., Ogiso, Y. & Tsuruo, T. (1999), 'Glucose-regulated stresses cause degradation of DNA topoisomerase IIα by inducing nuclear proteasome during G1 cell cycle arrest in cancer cells', *Journal of Cellular Physiology* **180**(1), 97–104.

Kimber, J. A., Foreman, L., Turner, B., Rich, P. & Kazarian, S. G. (2016), 'FTIR spectroscopic imaging and mapping with correcting lenses for studies of biological cells and tissues', *Faraday Discuss* **187**, 69–85.

Kimber, J. A. & Kazarian, S. G. (2017), 'Spectroscopic imaging of biomaterials and biological systems with FTIR microscopy or with quantum cascade lasers', *Analytical and Bioanalytical Chemistry* **409**(25), 5813–5820.

Kimber, J. A., Kazarian, S. G. & Štěpánek, F. (2012), 'Modelling of pharmaceutical tablet swelling and dissolution using discrete element method', *Chemical Engineering Science* **69**(1), 394–403.

Kobrina, Y., Rieppo, L., Saarakkala, S., Jurvelin, J. S. & Isaksson, H. (2012), 'Clustering of infrared spectra reveals histological zones in intact articular cartilage', *Osteoarthritis Cartilage* **20**(5), 460–8.

Kohler, A., Sulé-Suso, J., Sockalingum, G. D., Tobin, M., Bahrami, F. & Yang, Y. (2008), 'Estimating and Correcting Mie Scattering in Synchrotron-Based Microscopic Fourier Transform Infrared Spectra by Extended Multiplicative Signal Correction', *Appl Spectrosc* **62**(3), 259–266.

Kole, M. R., Reddy, R. K., Schulmerich, M. V., Gelber, M. K. & Bhargava, R. (2012), 'Discrete frequency infrared microspectroscopy and imaging with a tunable quantum cascade laser', *Analytical Chemistry* **84**(23), 10366–10372.

Kondepati, V. R., Heise, H. M., Oszinda, T., Mueller, R., Keese, M. & Backhaus, J. (2008), 'Detection of structural disorders in colorectal cancer DNA with Fourier-transform infrared spectroscopy', *Vibrational Spectroscopy* **46**(2), 150–157.

Konevskikh, T., Lukacs, R., Blumel, R., Ponossov, A. & Kohler, A. (2016), 'Mie scatter corrections in single cell infrared microspectroscopy', *Faraday Discuss* **187**, 235–57.

Krieger, J. N., Lee, S. W. H., Jeon, J., Cheah, P. Y., Liong, M. L. & Riley, D. E. (2008), 'Epidemiology of prostatitis', *International Journal of Antimicrobial Agents* **31**, 85–90.

Kroger-Lui, N., Gretz, N., Haase, K., Kränzlin, B., Neudecker, S., Pucci, A., Regenscheit, A., Schönhals, A. & Petrich, W. (2015), 'Rapid identification of goblet cells in

unstained colon thin sections by means of quantum cascade laser-based infrared microspectroscopy', *Analyst* **140**(7), 2086–2092.

Krycer, J. R. & Brown, A. J. (2013), 'Cholesterol accumulation in prostate cancer: A classic observation from a modern perspective', *Biochimica Et Biophysica Acta-Reviews on Cancer* **1835**(2), 219–229.

Kuepper, C., Kallenbach-Thieltges, A., Juette, H., Tannapfel, A., Großerueschkamp, F. & Gerwert, K. (2018), 'Quantum cascade laser-based infrared microscopy for label-free and automated cancer classification in tissue sections', *Scientific Reports* **8**(1), 7717.

Kuimova, M. K., Chan, K. L. A. & Kazarian, S. G. (2009), 'Chemical imaging of live cancer cells in the natural aqueous environment', *Applied Spectroscopy* **63**(2), 164–171.

Labrie, F., Dupont, A., Suburu, R., Cusan, L., Tremblay, M., Gomez, J.-L. & Emond, J. (1992), 'Serum prostate specific antigen as pre-screening test for prostate cancer', *The Journal of Urology* **147**(3, Part 2), 846–851.

Landau, L. & Lifshitz, E. (2013), *Statistical Physics*, third edn, Butterworth-Heinemann.

Lasch, P. (2012), 'Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging', *Chemometrics and Intelligent Laboratory Systems* **117**, 100–114.

Lasch, P., Boese, M., Pacifico, A. & Diem, M. (2002), 'FT-IR spectroscopic investigations of single cells on the subcellular level', *Vibrational Spectroscopy - VIB SPECTROSC* **28**, 147–157.

Lasch, P., Haensch, W., Naumann, D. & Diem, M. (2004), 'Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis', *Biochim Biophys Acta* **1688**(2), 176–86.

Lasch, P., Mahadevanansen, A. & Diem, M. (2004), 'FT-IR microspectroscopic imaging of prostate tissue sections', *Proc SPIE* **5321**, 1.

Lasch, P. & Naumann, D. (1997), FT-IR microspectroscopic imaging of human carcinoma tissue thin sections, *in* P. Carmona, R. Navarro & A. Hernanz, eds, 'Spectroscopy of Biological Molecules: Modern Trends', Springer Netherlands, Dordrecht, pp. 441–442.

Lee, E. Y. H. P. & Muller, W. J. (2010), 'Oncogenes and tumor suppressor genes', *Cold Spring Harbor Perspectives in Biology* **2**(10), a003236.

Lee, S. Y., Yoon, K. A., Jang, S. H., Ganbold, E. O., Uuriintuya, D., Shin, S. M., Ryu, P. D. & Joo, S. W. (2009), 'Infrared spectroscopy characterization of normal and lung cancer cells originated from epithelium', *Journal of veterinary science* **10**(4), 299–304.

Levine, A. J., Hu, W. & Feng, Z. (2008), Tumor suppressor genes, *in* 'The molecular basis of cancer', Elsevier, pp. 31–38.

Lewis, L. & Sommer, A. J. (1999), 'Attenuated total internal reflection microspectroscopy of isolated particles: An alternative approach to current methods', *Applied Spectroscopy* **53**(4), 375–380.

Li, L., Bi, X., Sun, H., Liu, S., Yu, M., Zhang, Y., Weng, S., Yang, L., Bao, Y., Wu, J., Xu, Y. & Shen, K. (2018), 'Characterization of ovarian cancer cells and tissues by Fourier transform infrared spectroscopy', *Journal of ovarian research* **11**(1), 64–64.

Li, S., Chen, G., Zhang, Y., Guo, Z., Liu, Z. & Xu, J. (2014), 'Identification and characterization of colorectal cancer using Raman spectroscopy and feature selection techniques', *Opt Express* **22**(21), 25895–908.

Li, X., Li, Q. B., Zhang, G. J., Xu, Y. Z., Sun, X. J., Shi, J. S., Zhang, Y. F. & Wu, J. G. (2012), 'Identification of colitis and cancer in colon biopsies by Fourier Transform Infrared spectroscopy and chemometrics', *ScientificWorldJournal* **2012**, 936149.

Liberty, F., James, A. K., Katherine, V. O., James, M. B., Samuel, M. J., Tom, F., Sergei, G. K. & Peter, R. (2015), Assessing dysplasia of a bronchial biopsy with FTIR spectroscopic imaging, *in* 'Optical Diagnostics and Sensing XV: Toward Point-of-Care Diagnostics', Vol. 9332, Proc.SPIE.

Liotta, L. A. & Petricoin, E. F. (2006), 'Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold', *The Journal of Clinical Investigation* **116**(1), 26–30.

Litwin, M. S. & Tan, H.-J. (2017), 'The diagnosis and treatment of prostate cancer: a review', *JAMA* **317**(24), 2532–2542.

Liu, M. C., Oxnard, G. R., Klein, E. A., Swanton, C., Seiden, M. V. & on behalf of the CCGA Consortium (2020), 'Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA', *Annals of Oncology* **31**(6), 745–759.

Liu, Q., Zhang, H., Jiang, X., Qian, C., Liu, Z. & Luo, D. (2017), 'Factors involved in cancer metastasis: a better understanding to "seed and soil" hypothesis', *Molecular Cancer* **16**(1), 176.

Liu, X. & Padilla, W. J. (2017), 'Reconfigurable room temperature metamaterial infrared emitter', *Optica* **4**(4), 430.

Liu, Y. (2006), 'Fatty acid oxidation is a dominant bioenergetic pathway in prostate cancer', *Prostate Cancer and Prostatic Diseases* **9**(3), 230–234.

Locker, G. Y., Hamilton, S., Harris, J., Jessup, J. M., Kemeny, N., Macdonald, J. S., Somerfield, M. R., Hayes, D. F. & Bast, R. C. (2006), 'ASCO 2006 Update of Recommendations for the Use of Tumor Markers in Gastrointestinal Cancer', *Journal of Clinical Oncology* **24**(33), 5313–5327.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. & Darnell, J. (2000*a*), Section 1.4, the life cycle of cells, *in* 'Molecular Cell Biology', fourth edn, W. H. Freeman, New York.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. & Darnell, J. (2000*b*), Section 24.2, proto-oncogenes and tumor-suppressor genes, *in* 'Molecular Cell Biology', fourth edn, W. H. Freeman, New York.

Long, J., Zhang, C.-J., Zhu, N., Du, K., Yin, Y.-F., Tan, X., Liao, D.-F. & Qin, L. (2018), 'Lipid metabolism and carcinogenesis, cancer development', *American journal of cancer research* **8**(5), 778–791.

Lu, G. & Fei, B. (2014), 'Medical hyperspectral imaging: a review', *J Biomed Opt* **19**(1), 10901.

Lu, Z., Cassidy, B. M., DeJong, S. A., Belliveau, R. G., Myrick, M. L. & Morgan, S. L. (2017), 'Attenuated total reflection (ATR) sampling in infrared spectroscopy of heterogeneous materials requires reproducible pressure control', *Appl Spectrosc* **71**(1), 97–104.

Luder Ripoli, F., Mohr, A., Conradine Hammer, S., Willenbrock, S., Hewicker-Trautwein, M., Hennecke, S., Murua Escobar, H. & Nolte, I. (2016), 'A comparison of fresh frozen vs. formalin-fixed, paraffin-embedded specimens of canine mammary tumors via branched-DNA assay', *Int J Mol Sci* **17**(5).

Ly, E., Piot, O., Wolthuis, R., Durlach, A., Bernard, P. & Manfait, M. (2008), 'Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies', *Analyst* **133**(2), 197–205.

Lyng, F., Gazi, E. & Gardner, P. (2010), Chapter 5. Preparation of Tissues and Cells for Infrared and Raman Spectroscopy and Imaging, *in* G. Srinivasan, ed., 'Vibrational Spectroscopic Imaging for Biomedical Applications', The McGraw-Hill Companies, Inc, pp. 145–191.

Mackanos, M. A. & Contag, C. H. (2010), 'Fiber-optic probes enable cancer detection with ftir spectroscopy', *Trends in Biotechnology* **28**(6), 317–323.

Mahadevan-Jansen, A., Mordechai, S., Puppels, G. J., Salman, A. O., Argov, S., Cohen, B., Erukhimovitch, V., Goldstein, J., Chaims, O. & Hammody, Z. (2000), 'Fourier-transform infrared spectroscopy of human cancerous and normal intestine', *Proc. SPIE* **3918**, 66.

Malins, D. C., Gilman, N. K., Green, V. M., Wheeler, T. M., Barker, E. A. & Anderson, K. M. (2005), 'A cancer DNA phenotype in healthy prostates, conserved in tumors and adjacent normal cells, implies a relationship to carcinogenesis', *Proc Natl Acad Sci U S A* **102**(52), 19093–6.

Malins, D. C., Polissar, N. L. & Gunselman, S. J. (1997), 'Models of DNA structure achieve almost perfect discrimination between normal prostate, benign prostatic hyperplasia (BPH), and adenocarcinoma and have a high potential for predicting BPH and prostate cancer', *Proceedings of the National Academy of Sciences of the United States of America* **94**(1), 259–264.

Marechal, Y. (2011), 'The molecular structure of liquid water delivered by absorption spectroscopy in the whole IR region completed with thermodynamics data', *Journal of Molecular Structure* **1004**(1-3), 146–155.

Martens, H. & Stark, E. (1991), 'Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy', *Journal of Pharmaceutical and Biomedical Analysis* **9**(8), 625–635.

Matousek, P., Clark, I. P., Draper, E. R. C., Morris, M. D., Goodship, A. E., Everall, N., Towrie, M., Finney, W. F. & Parker, A. W. (2005), 'Subsurface probing in diffusely scattering media using spatially offset Raman spectroscopy', *Applied Spectroscopy* **59**(4), 393–400.

Matousek, P. & Stone, N. (2009), 'Emerging concepts in deep Raman spectroscopy of biological tissue', *Analyst* **134**(6), 1058–1066.

Mayo Foundation for Medical Education and Research (MFMER) (2020*a*), 'Colon cancer', `https://www.mayoclinic.org/diseases-conditions/colon-cancer/diagnosis-treatment/drc-20353674`. Accessed Mar 12, 2020.

Mayo Foundation for Medical Education and Research (MFMER) (2020*b*), 'Prostate cancer', `https://www.mayoclinic.org/diseases-conditions/prostate-cancer/diagnosis-treatment/drc-20353093`. Accessed Mar 12, 2020.

Memorial Sloan Kettering Cancer (2020), 'Inherited risk for prostate cancer', `https://www.mskcc.org/cancer-care/risk-assessment-screening/hereditary-genetics/genetic-counseling/inherited-risk-prostate`. Accessed Mar 12, 2020.

Mendelsohn, R., Flach, C. R. & Moore, D. J. (2006), 'Determination of molecular conformation and permeation in skin via IR spectroscopy, microscopy, and imaging', *Biochimica Et Biophysica Acta-Biomembranes* **1758**(7), 923–933.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. & Hamprecht, F. A. (2009), 'A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data', *BMC Bioinformatics* **10**, 213.

Messerschmidt, R. G. (1987), *Photometric Considerations in the Design and Use of Infrared Microscope Accessories:The Design, Sample Handling, and Applications of Infrared Microscopes*, ASTM International, West Conshohocken, PA.

Mie, G. (1908), 'Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen', *Annalen der Physik* **330**(3), 377–445.

Mignolet, A., Derenne, A., Smolina, M., Wood, B. R. & Goormaghtigh, E. (2016), 'FTIR spectral signature of anticancer drugs. can drug mode of action be identified?', *Biochim Biophys Acta* **1864**(1), 85–101.

Miljkovic, M., Bird, B. & Diem, M. (2012), 'Line shape distortion effects in infrared spectroscopy', *Analyst* **137**(17), 3954–64.

Miller, L. M. & Dumas, P. (2013), Infrared spectroscopy using synchrotron radiation, *in* G. C. K. Roberts, ed., 'Encyclopedia of Biophysics', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1106–1112.

Milosevic, M. (2012), *Internal reflection and ATR spetroscopy*, Chemical Analysis, John Wiley & Sons, Inc., New Jersey, USA.

Milosevic, M. (2013), 'On the nature of the evanescent wave', *Appl Spectrosc* **67**(2), 126–31.

Mohlenhoff, B., Romeo, M., Diem, M. & Wood, B. R. (2005), 'Mie-type scattering and non-Beer-Lambert absorption behavior of human cells in infrared microspectroscopy', *Biophys J* **88**(5), 3635–40.

Montironi, R., Cheng, L., Scarpelli, M. & Lopez-Beltran, A. (2016), 'Pathology and genetics: tumours of the urinary system and male genital system: clinical implications of the 4th Edition of the WHO classification and beyond', *European Urology* **70**(1), 120–123.

Mostaco-Guidolin, L. B., Murakami, L. S., Nomizo, A. & Bachmann, L. (2009), 'Fourier transform infrared spectroscopy of skin cancer cells and tissues', *Applied Spectroscopy Reviews* **44**(5), 438–455.

Movasaghi, Z., Rehman, S. & ur Rehman, D. I. (2008), 'Fourier transform infrared (FTIR) spectroscopy of biological tissues', *Applied Spectroscopy Reviews* **43**(2), 134–179.

Musumeci, G. (2014), 'Past, present and future: overview on histology and histopathology', *Journal of Histology and Histopathology* **1**(1), 5.

Nagase, Y., Yoshida, S. & Kamiyama, K. (2005), 'Analysis of human tear fluid by Fourier transform infrared spectroscopy', *Biopolymers* **79**(1), 18–27.

Nakamura, A., Koga, T., Fujimaki, M., Ohki, Y., Sota, T., Lipinska-Kalita, K., Nagae, T., Ishimaru, S. & Aizawa, K. (2000), 'Application of infrared attenuated total reflection

spectroscopy to in situ analysis of atheromatous plaques in aorta', *Japanese Journal of Applied Physics* **39**(Part 2, No. 6A), L490–L492.

Nakano, T. & Kawata, S. (1994), 'Evanescent-field scanning microscope with fourier-transform infrared spectrometer', *Scanning* **16**(3), 368–371.

Nallala, J., Gobinet, C., Diebold, M. D., Untereiner, V., Bouche, O., Manfait, M., Sockalingum, G. D. & Piot, O. (2012), 'Infrared spectral imaging as a novel approach for histopathological recognition in colon cancer diagnosis', *J Biomed Opt* **17**(11), 116013.

Nallala, J., Lloyd, G. R. & Stone, N. (2015), 'Evaluation of different tissue deparaffinization procedures for infrared spectral imaging', *Analyst* **140**(7), 2369–75.

National Cancer Institute (NIH) (2017), 'Prostate-specific antigen (PSA) test', `https://www.cancer.gov/types/prostate/psa-fact-sheet`. Accessed Mar 12, 2020.

National Cancer Institute (NIH) (2018), 'Cancer-causing substances in the environment', `https://www.cancer.gov/about-cancer/causes-prevention/risk/substances`. Accessed Mar 11, 2020.

Naumann, D. (2013), Infrared spectroscopy of cells, tissues, and biofluids, *in* G. C. K. Roberts, ed., 'Encyclopedia of Biophysics', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1057–1065.

Nembrini, S., Konig, I. R. & Wright, M. N. (2018), 'The revival of the Gini importance?', *Bioinformatics* **34**(21), 3711–3718.

Norton, R. H. & Beer, R. (1976), 'New apodizing functions for Fourier spectrometry', *Journal of the Optical Society of America* **66**(3), 259.

Nunes, C. A., Alvarenga, V. O., de Souza Sant'Ana, A., Santos, J. S. & Granato, D. (2015), 'The use of statistical software in food science and technology: Advantages, limitations and misuses', *Food Res Int* **75**, 270–280.

O Faolain, E., Hunter, M. B., Byrne, J. M., Kelehan, P., McNamara, M., Byrne, H. J. & Lyng, F. M. (2005), 'A study examining the effects of tissue processing on human tissue sections using vibrational spectroscopy', *Vibrational Spectroscopy* **38**(1-2), 121–127.

O'Brien, F. E. M. (1948), 'The Control of Humidity by Saturated Salt Solutions', *Journal of Scientific Instruments* **25**(3), 73–76.

Okobia, M. N., Zmuda, J. M., Ferrell, R. E., Patrick, A. L. & Bunker, C. H. (2011), 'Chromosome 8q24 variants are associated with prostate cancer risk in a high risk population of African ancestry', *The Prostate* **71**(10), 1054–1063.

Oliver, K. V., Maréchal, A. & Rich, P. R. (2016), 'Effects of the hydration state on the mid-infrared spectra of urea and creatinine in relation to urine analyses', *Applied Spectroscopy* **70**(6), 983–994.

Owen, A. J. (1995), 'Uses of derivative spectroscopy: Application note', https://www.whoi.edu/cms/files/derivative_spectroscopy_59633940_175744.pdf. Accessed Mar 25, 2020.

Paraskevaidi, M., Morais, C. L. M., Lima, K. M. G., Ashton, K. M., Stringfellow, H. F., Martin-Hirsch, P. L. & Martin, F. L. (2018), 'Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer', *Analyst* **143**(13), 3156–3163.

Patel, N. D. & Parsons, J. K. (2014), 'Epidemiology and etiology of benign prostatic hyperplasia and bladder outlet obstruction', *Indian J Urol* **30**(2), 170–176.

Patterson, B. M. & Havrilla, G. J. (2006), 'Attenuated total internal reflection infrared microspectroscopic imaging using a large-radius germanium internal reflection element and a linear array detector', *Appl Spectrosc* **60**(11), 1256–66.

Patterson, B. M., Havrilla, G. J., Marcott, C. & Story, G. M. (2007), 'Infrared microspectroscopic imaging using a large radius germanium internal reflection element and a focal plane array detector', *Appl Spectrosc* **61**(11), 1147–52.

Pevsner, A. & Diem, M. (2003), 'IR spectroscopic studies of major cellular components. III. Hydration of protein, nucleic acid, and phospholipid films', *Biopolymers* **72**(4), 282–289.

Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S., Schatz, C. R., Miller, S. S., Su, Q., McGrath, A. M., Estock, M. A., Parmar, P. P., Zhao, M., Huang, S.-T., Zhou, J., Wang, F., Esquer-Blasco, R., Anderson, N. L., Taylor, J. & Steiner, S. (2003), 'The human serum proteome: Display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins', *PROTEOMICS* **3**(7), 1345–1364.

PIKE Technologies, Inc. (2020), 'ATR – theory and applications', https://www.piketech.com/wp-content/uploads/2019/12/PIKE_ATR_theory_and_app_2020-1.pdf. Accessed Mar 19, 2020.

Pilling, M. J., Henderson, A., Shanks, J. H., Brown, M. D., Clarke, N. W. & Gardner, P. (2017), 'Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation', *Analyst.* **142**(8), 1258–1268.

Pinkley, L. W., Sethna, P. P. & Williams, D. (1977), 'Optical constants of water in the infrared: Influence of temperature*', *Journal of the Optical Society of America* **67**(4), 494.

Pisapia, C., Jamme, F., Duponchel, L. & Ménez, B. (2018), 'Tracking hidden organic carbon in rocks using chemometrics and hyperspectral imaging', *Scientific Reports* **8**(1), 2396.

Planck, M. (2013), *Theory of Heat Radiation*, second edn, Dover Publications, Inc.

Pleshko, N. L., Boskey, A. L. & Mendelsohn, R. (1992), 'An FT-IR microscopic investigation of the effects of tissue preservation on bone', *Calcified Tissue International* **51**(1), 72–77.

Prabhakar, S., Jain, N. & Singh, R. A. (2012), 'Infrared spectra in monitoring biochemical parameters of human blood', *Journal of Physics: Conference Series* **365**, 012059.

Rath, P., Bousché, O., Merrill, A. R., Cramer, W. A. & Rothschild, K. J. (1991), 'Fourier transform infrared evidence for a predominantly alpha-helical structure of the membrane bound channel forming COOH-terminal peptide of colicin E1', *Biophysical Journal* **59**(3), 516–522.

Rawla, P. (2019), 'Epidemiology of prostate cancer', *World Journal of Oncology* **10**(2), 63–89.

Rayleigh (2009), 'XXXI. Investigations in optics, with special reference to the spectroscope', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **8**(49), 261–274.

Rebbeck, T. R. & Haas, G. P. (2014), 'Temporal trends and racial disparities in global prostate cancer prevalence', *The Canadian journal of urology* **21**(5), 7496–7506.

Riedl, M. (2001), *Optical Design Fundamentals for Infrared Systems*, second edn, SPIE Press, Bellingham, WA.

Rigas, B., Morgello, S., Goldman, I. S. & Wong, P. T. (1990), 'Human colorectal cancers display abnormal Fourier-transform infrared spectra', *Proc Natl Acad Sci U S A* **87**(20), 8140–4.

Rigas, B. & Wong, P. T. T. (1992), 'Human colon adenocarcinoma cell lines display infrared spectroscopic features of malignant colon tissues', *Cancer Research* **52**(1), 84.

Robbins, C. M., Hooker, S., Kittles, R. A. & Carpten, J. D. (2011), 'EphB2 SNPs and Sporadic Prostate Cancer Risk in African American Men', *PLOS ONE* **6**(5), e19494.

Robertson, C. W. & Williams, D. (1971), 'Lambert absorption coefficients of water in the infrared*', *Journal of the Optical Society of America* **61**(10), 1316.

Robinson, D. (2020), 'K-means clustering is not a free lunch', `http://varianceexplained.org/r/kmeans-free-lunch/`. Accessed Mar 30, 2020.

Robinson, J. W., Skelly Frame, E. M. & Frame II, G. M. (2005), IR spectroscopy, *in* 'Undergraduate Instrumental Analysis', sixth edn, Taylor and Francis, Hoboken, p. 235.

Rockland, L. B. (1960), 'Saturated salt solutions for static control of relative humidity between 5° and 40° C', *Analytical Chemistry* **32**(10), 1375–1376.

Rogalski, A. (2005), 'HgCdTe infrared detector material: history, status and outlook', *Reports on Progress in Physics* **68**(10), 2267–2336.

Rossmanna, C. & Haemmerich, D. (2014), 'Review of temperature dependence of thermal properties, dielectric properties, and perfusion of biological tissues at hyperthermic and ablation temperatures', *Critical reviews in biomedical engineering* **42**(6), 467–92.

Rowan-Robinson, M. (2013), *Night vision: Exploring the infrared universe*, Cambridge University Press, Cambridge.

Ryu, M., Kimber, J. A., Sato, T., Nakatani, R., Hayakawa, T., Romano, M., Pradere, C., Hovhannisyan, A. A., Kazarian, S. G. & Morikawa, J. (2017), 'Infrared thermo-spectroscopic imaging of styrene radical polymerization in microfluidics', *Chemical Engineering Journal* **324**, 259–265.

Sabbatini, S., Conti, C., Orilisi, G. & Giorgini, E. (2017), 'Infrared spectroscopy as a new tool for studying single living cells: Is there a niche?', *Biomedical Spectroscopy and Imaging* **6**(3-4), 85–99.

Sahu, R. K., Argov, S., Salman, A., Huleihel, M., Grossman, N., Hammody, Z., Kapelushnik, J. & Mordechai, S. (2004), 'Characteristic absorbance of nucleic acids in the mid-ir region as possible common biomarkers for diagnosis of malignancy', *Technol Cancer Res Treat* **3**(6), 629–38.

Sahu, R. K., Argov, S., Salman, A., Zelig, U., Huleihel, M., Grossman, N., Gopas, J., Kapelushnik, J. & Mordechai, S. (2005), 'Can Fourier transform infrared spectroscopy at higher wavenumbers (mid IR) shed light on biomarkers for carcinogenesis in tissues?', *J Biomed Opt* **10**(5), 054017.

Sahu, R., Salman, A. & Mordechai, S. (2017), 'Tracing overlapping biological signals in mid-infrared using colonic tissues as a model system', *World J Gastroenterol* **23**(2), 286–296.

Salisbury, J. W., Wald, A. & D'Aria, D. M. (1993), 'Thermal infrared remote sensing and Kirchhoff's law: 1. laboratory measurements', *Lunar and Planetary Science Conference* **99**(B8), 11897–911.

Saptari, V. (2003), *Fourier transform spectroscopy instrumentation engineering*, SPIE Optical Engineering Press, Bellingham, WA.

Savitzky, A. & Golay, M. J. E. (1964), 'Smoothing and differentiation of data by simplified least squares procedures', *Analytical Chemistry* **36**(8), 1627–1639.

Schofield, A. J., Blumel, R., Kohler, A., Lukacs, R. & Hirschmugl, C. J. (2019), 'Extracting pure absorbance spectra in infrared microspectroscopy by modeling absorption bands as Fano resonances', *J Chem Phys* **150**(15), 154124.

scikit-learn developers (2019), 'Selecting the number of clusters with silhouette analysis on KMeans clustering', `https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`. Accessed Mar 30, 2020.

Seyfried, T. N. & Huysentruyt, L. C. (2013), 'On the origin of cancer metastasis', *Crit Rev Oncog Critical Reviews in Oncogenesis* **18**(1-2), 43–73.

Shalygin, A. S., Kozhevnikov, I. V., Kazarian, S. G. & Martyanov, O. N. (2019), 'Spectroscopic imaging of deposition of asphaltenes from crude oil under flow', *Journal of Petroleum Science and Engineering* **181**, 106205.

Sijbers, J., Scheunders, P., Bonnet, N., Van Dyck, D. & Raman, E. (1996), 'Quantification and improvement of the signal-to-noise ratio in a magnetic resonance image acquisition procedure', *Magnetic Resonance Imaging* **14**(10), 1157–1163.

Siqueira, L. F. S., Morais, C. L. M., Araújo Júnior, R. F., de Araújo, A. A. & Lima, K. M. G. (2018), 'SVM for FT-MIR prostate cancer classification: An alternative to the traditional methods', *Journal of Chemometrics* **32**(12), e3075.

Skov, T., Honoré, A. H., Jensen, H. M., Næs, T. & Engelsen, S. B. (2014), 'Chemometrics in foodomics: Handling data structures from multiple analytical platforms', *TrAC Trends in Analytical Chemistry* **60**, 71–79.

Smith, B. C. (2002), *Quantitative Spectroscopy: Theory and Practice*, first edn, Academic Press.

Smith, B. C. (2011), *Fundamentals of Fourier Transform Infrared Spectroscopy*, second edn, CRC Press, Taylor and Francis Group, USA.

Smith, B. R., Ashton, K. M., Brodbelt, A., Dawson, T., Jenkinson, M. D. & Hunt, N. T. (2016), 'Combining random forest and 2D correlation analysis to identify serum spectral signatures for neuro-oncology', *Analyst.* **141**.

Sommer, A. J., Tisinger, L. G., Marcott, C. & Story, G. M. (2001), 'Attenuated total internal reflection infrared mapping microspectroscopy using an imaging microscope', *Applied Spectroscopy* **55**(3), 252–256.

Song, C. L. & Kazarian, S. G. (2019*a*), 'Micro ATR-FTIR spectroscopic imaging of colon biopsies with a large area Ge crystal', *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **228**, 117695.

Song, C. L. & Kazarian, S. G. (2019*b*), 'Three-dimensional depth profiling of prostate tissue by micro ATR-FTIR spectroscopic imaging with variable angles of incidence', *Analyst* **144**(9), 2954–2964.

Song, C. L. & Kazarian, S. G. (2020), 'The effect of controlled humidity and tissue hydration on colon cancer diagnostic via ftir spectroscopic imaging', *Analytical Chemistry* . **URL:** *https://doi.org/10.1021/acs.analchem.0c01002*

Song, C. L., Ryu, M., Morikawa, J., Kothari, A. & Kazarian, S. G. (2018), 'Thermal effect on dispersive infrared spectroscopic imaging of prostate cancer tissue', *J Biophotonics* **11**(12), e201800187.

Song, C. L., Vardaki, M. Z., Goldin, R. D. & Kazarian, S. G. (2019), 'Fourier transform infrared spectroscopic imaging of colon tissues: evaluating the significance of amide I and C-H stretching bands in diagnostic applications with machine learning', *Anal Bioanal Chem* **411**(26), 6969–6981.

Song, L. M. W. K., Molckovsky, A., Wang, K. K., Burgart, L. J., Dolenko, B., Somorjai, R. L. & Wilson, B. C. (2005), 'Diagnostic potential of Raman spectroscopy in Barrett's esophagus', *PROCEEDINGS- SPIE THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING* **5692**(23), 140–146.

Specac Limited (2019), 'Transmission vs ATR spectroscopy — Animated guides', `https://www.specac.com/en/news/calendar/2019/04/transmission-vs-atr`. Accessed Mar 19, 2020.

Specac Limited (2020), 'What is band distortion and band shift in ATR?', `https://www.specac.com/en/documents/instructional/what-is-band-distortion-and-band-shift-in-atr`. Accessed Mar 19, 2020.

Stahle, L. & Wold, S. (1987), 'Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study', *Journal of Chemometrics* **1**(3), 185–196.

Stark, G. (2020), 'Light', `https://www.britannica.com/science/light`. Accessed Mar 19, 2020.

Stuart, B. H. (2004), *Infrared Spectroscopy: Fundamentals and Applications*, John Wiley & Sons, UK.

Su, W.-H. & Sun, D.-W. (2018), 'Fourier transform infrared and Raman and hyperspectral imaging techniques for quality determinations of powdery foods: A review', *Comprehensive Reviews in Food Science and Food Safety* **17**(1), 104–122.

Svensson, T., Andersson-Engels, S., Einarsdottir, M. & Svanberg, K. (2007), 'In vivo optical characterization of human prostate tissue using near-infrared time-resolved spectroscopy', *J Biomed Opt* **12**(1), 014022.

Szymanska, E., Gerretzen, J., Engel, J., Geurts, B., Blanchet, L. & Buydens, L. M. C. (2015), 'Chemometrics and qualitative analysis have a vibrant relationship', *TrAC Trends in Analytical Chemistry* **69**, 34–51.

Takamura, A., Watanabe, K., Akutsu, T. & Ozawa, T. (2018), 'Soft and robust identification of body fluid using fourier transform infrared spectroscopy and chemometric strategies for forensic analysis', *Scientific Reports* **8**(1), 8459.

Tan, Z.-J. & Chen, S.-J. (2006), 'Nucleic acid helix stability: Effects of salt concentration, cation valence and size, and chain length', *Biophysical Journal* **90**(4), 1175–1190.

Taylor, R. A., Fraser, M., Rebello, R. J., Boutros, P. C., Murphy, D. G., Bristow, R. G. & Risbridger, G. P. (2019), 'The influence of BRCA2 mutation on localized prostate cancer', *Nature Reviews Urology* **16**(5), 281–290.

The MathWorks, Inc. (2020*a*), 'barttest', `https://uk.mathworks.com/help/stats/barttest.html#bt5e98y-2`. Accessed Mar 25, 2020.

The MathWorks, Inc. (2020*b*), 'Discriminant analysis', `https://uk.mathworks.com/help/stats/classification-discriminant-analysis.html`. Accessed Mar 30, 2020.

The MathWorks, Inc. (2020*c*), 'k-Means clustering', `https://uk.mathworks.com/help/stats/k-means-clustering.html`. Accessed Mar 30, 2020.

Theophanides, T. (2012), Introduction to infrared spectroscopy in life and biomedical sciences, *in* 'Infrared Spectroscopy: Life and Biomedical Sciences', INTECH Open Access Publisher.

Thermo Fisher Scientific Inc (2013), 'Introduction to Fourier transform infrared spectroscopy', `http://tools.thermofisher.com/content/sfs/brochures/BR50555_E_0513M_H_1.pdf`. Accessed Mar 19, 2020.

Thermo Nicolet Corp (2002), 'FT-IR vs. dispersive infrared: Theory of infrared spectroscopy instrumentation', `http://www.thermo.com.cn/Resources/200802/productPDF_21615.pdf`. Accessed Mar 19, 2020.

Thiam, A. R., Farese, R. V., J. & Walther, T. C. (2013), 'The biophysics and cell biology of lipid droplets', *Nat Rev Mol Cell Biol* **14**(12), 775–86.

Thompson, I. M., Pauler, D. K., Goodman, P. J., Tangen, C. M., Lucia, M. S., Parnes, H. L., Minasian, L. M., Ford, L. G., Lippman, S. M., Crawford, E. D., Crowley, J. J. & Coltman, C. A. (2004), 'Prevalence of prostate cancer among men with a prostate-specific antigen level $\leq$ 4.0 ng per milliliter', *New England Journal of Medicine* **350**(22), 2239–2246.

Travo, A., Paya, C., Déléris, G., Colin, J., Mortemousque, B. & Forfar, I. (2014), 'Potential of FTIR spectroscopy for analysis of tears for diagnosis purposes', *Analytical and bioanalytical chemistry* **406**(9-10), 2367–2376.

Trevisan, J., Angelov, P. P., Carmichael, P. L., Scott, A. D. & Martin, F. L. (2012), 'Extracting biological information with computational analysis of fourier-transform infrared (ftir) biospectroscopy datasets: current practices to future perspectives', *Analyst* **137**(14), 3202–3215.

Tseng, D. Y. (1997), 'Spectroscopic analysis (FT-IR) of polysaccharide degradation in a bench-scale solid waste treatment (composting) system', **66**, 497–509.

van den Driesche, S., Witarski, W., Pastorekova, S., Breiteneder, H., Hafner, C. & Vellekoop, M. J. (2011), 'A label-free indicator for tumor cells based on the CH2-stretch ratio', *Analyst* **136**(11), 2397–2402.

van der Beek, C. M., Dejong, C. H. C., Troost, F. J., Masclee, A. A. M. & Lenaerts, K. (2017), 'Role of short-chain fatty acids in colonic inflammation, carcinogenesis, and mucosal protection and healing', *Nutr Rev* **75**(4), 286–305.

Varmuza, K. & Filzmoser, P. (2009), Chemoinformatics-chemometrics-statistics, *in* 'Introduction to Multivariate Statistical Analysis in Chemometrics', CRC Press, Boca Raton.

Vasefi, F., MacKinnon, N. & Farkas, D. L. (2016), Hyperspectral and multispectral imaging in dermatology, *in* M. R. Hamblin, P. Avci & G. K. Gupta, eds, 'Imaging in Dermatology', Academic Press, Elsevier Inc., pp. 187–201.

Venyaminov, S. Y. & Prendergast, F. G. (1997), 'Water (H2O and D2O) molar absorptivity in the 1000–4000 cm^{-1} range and quantitative infrared spectroscopy of aqueous solutions', *Analytical Biochemistry* **248**(2), 234–245.

Verikas, A., Vaiciukynas, E., Gelzinis, A., Parker, J. & Olsson, M. C. (2016), 'Electromyographic patterns during golf swing: Activation sequence profiling and prediction of shot effectiveness', *Sensors (Basel, Switzerland)* **16**(4), 592.

Vichi, A., Eliazyan, G. & Kazarian, S. G. (2018), 'Study of the degradation and conservation of historical leather book covers with macro attenuated total reflection-fourier transform infrared spectroscopic imaging', *ACS Omega* **3**(7), 7150–7157.

Villa, E., Ali, E. S., Sahu, U. & Ben-Sahra, I. (2019), 'Cancer cells tune the signaling pathways to empower de novo synthesis of nucleotides', *Cancers* **11**(5), 688.

Vongsvivut, J., Perez-Guaita, D., Wood, B. R., Heraud, P., Khambatta, K., Hartnell, D., Hackett, M. J. & Tobin, M. J. (2019), 'Synchrotron macro ATR-FTIR microspectroscopy for high-resolution chemical mapping of single cells', *Analyst* **144**(10), 3226–3238.

Wang, Z., Tangella, K., Balla, A. & Popescu, G. (2011), 'Tissue refractive index as marker of disease', *J Biomed Opt* **16**(11), 116017.

WebMD (2020), 'Prostate cancer risk factors', https://www.webmd.com/prostate-cancer/guide/prostate-cancer-risk-factors#2. Accessed Mar 12, 2020.

Werner, M., Chott, A., Fabiano, A. & Battifora, H. (2000), 'Effect of formalin tissue fixation and processing on immunohistochemistry', *Am J Surg Pathol* **24**(7), 1016–9.

Whelan, D. R., Bambery, K. R., Puskar, L., McNaughton, D. & Wood, B. R. (2013*a*), 'Quantification of DNA in simple eukaryotic cells using Fourier transform infrared spectroscopy', *Journal of Biophotonics* **6**(10), 775–784.

Whelan, D. R., Bambery, K. R., Puskar, L., McNaughton, D. & Wood, B. R. (2013*b*), 'Synchrotron Fourier transform infrared (FTIR) analysis of single living cells progressing through the cell cycle', *Analyst* **138**(14), 3891–3899.

Wolthuis, R., Travo, A., Nicolet, C., Neuville, A., Gaub, M. P., Guenot, D., Ly, E., Manfait, M., Jeannesson, P. & Piot, O. (2008), 'IR spectral imaging for histopathological characterization of xenografted human colon carcinomas', *Anal Chem* **80**(22), 8461–9.

Wood, B. R. (2016), 'The importance of hydration and DNA conformation in interpreting infrared spectra of cells and tissues', *Chemical Society Reviews* **45**(7), 1980–1998.

World Cancer Research Fund (WCRF) International (2018), 'Worldwide cancer data', https://www.wcrf.org/dietandcancer/cancer-trends/worldwide-cancer-data. Accessed Mar 11, 2020.

Wrobel, T. P., Marzec, K. M., Majzner, K., Kochan, K., Bartus, M., Chlopicki, S. & Baranska, M. (2012), 'Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy of a single endothelial cell', *Analyst* **137**(18), 4135.

Wrobel, T. P., Vichi, A., Baranska, M. & Kazarian, S. G. (2015), 'Micro-attenuated total reflection fourier transform infrared (micro ATR FT-IR) spectroscopic imaging with variable angles of incidence', *Appl Spectrosc* **69**(10), 1170–4.

Wu, J. G., Xu, Y. Z., Sun, C. W., Soloway, R. D., Xu, D. F., Wu, Q. G., Sun, K. H., Weng, S. F. & Xu, G. X. (2001), 'Distinguishing malignant from normal oral tissues using FTIR fiber-optic techniques', *Biopolymers* **62**(4), 185–92.

Xiang, J., Yan, H., Li, J., Wang, X., Chen, H. & Zheng, X. (2019), 'Transperineal versus transrectal prostate biopsy in the diagnosis of prostate cancer: a systematic review and meta-analysis', *World Journal of Surgical Oncology* **17**(1), 31.

Xiao, Y., Shahsafi, A., Wan, C., Roney, P. J., Joe, G., Yu, Z., Salman, J. & Kats, M. A. (2019), 'Measuring thermal emission near room temperature using Fourier-transform infrared spectroscopy', *Physical Review Applied* **11**(1), 014026.

Yeng, Y. X., Ghebrebrhan, M., Bermel, P., Chan, W. R., Joannopoulos, J. D., Soljacic, M. & Celanovic, I. (2012), 'Enabling high-temperature nanophotonics for energy applications', *Proc Natl Acad Sci U S A* **109**(7), 2280–5.

Yu, M.-C., Rich, P., Foreman, L., Smith, J., Yu, M.-S., Tanna, A., Dibbur, V., Unwin, R. & Tam, F. W. K. (2017), 'Label free detection of sensitive mid-infrared biomarkers of glomerulonephritis in urine using fourier transform infrared spectroscopy', *Scientific Reports* **7**(1), 4601.

Zhao, S., Zhu, L., Gao, L. & Li, D. (2018), Chapter 2 - limitations for microplastic quantification in the ocean and recommendations for improvement and standardization, *in* E. Y. Zeng, ed., 'Microplastic Contamination in Aquatic Environments', Elsevier, pp. 27–49.

Zhu, Y., Zhang, J., Li, A., Zhang, Y. & Fan, C. (2017), 'Synchrotron-based X-ray microscopy for sub-100nm resolution cell imaging', *Current Opinion in Chemical Biology* **39**, 11–16.

Zhuo, W., Gabriel, P., Krishnarao, V. T. & Andre, B. (2011), 'Tissue refractive index as marker of disease', *Journal of Biomedical Optics* **16**(11), 1–8.

Zuccheri, G. & Samorì, B. (2002), Chapter 17 - scanning force microscopy studies on the structure and dynamics of single DNA molecules, *in* B. P. Jena & J. K. Heinrich Hörber, eds, 'Methods in Cell Biology', Vol. 68, Academic Press, pp. 357–395.

Zucchiatti, P., Mitri, E., Kenig, S., Billè, F., Kourousias, G., Bedolla, D. E. & Vaccari, L. (2016), 'Contribution of ribonucleic acid (RNA) to the Fourier transform infrared (FTIR) spectrum of eukaryotic cells', *Analytical Chemistry* **88**(24), 12090–12098.

Chapter 6

Appendices

6.1 The life cycle of a cancer cell

Unlike cancer cells, normal cells follow a regular timing mechanism. The life cycle of a normal cells is controlled by a complex series of molecular and biochemical signalling pathways by which the cells grow, divide, and die (Lodish et al. 2000*a*). G1 phase is the first of the four phases in a cell cycle. During this period, a cell grows and synthesises mRNA, proteins, and other molecular building blocks in preparation for the subsequent steps leading to mitosis; after which DNA is duplicated during the synthesis (S) phase. The cell experiences more growth in the G2 phase before it enters the mitosis (M) phase where cell separation is accomplished. The regulation of the cell cycle is accomplished by a complex set of enzymes called protein kinases. Cyclin–dependent kinases (CDKs), which act as master regulators of the cell cycle was discovered by scientists decades ago (Dorée & Galas 1994). As the name suggests, they are activated by forming complexes with cyclins to provide the driving force for the cell cycle progression. Apart from the CDKs, scientists have also discovered the presence of other regulatory proteins, such as Polo-like kinases (PLKs) and Aurora kinases which are responsible for error mitigation during cell cycle (Fu et al. 2010). When functioning properly, they act as the body's own tumour suppressors by controlling cell growth and inducing the death of damaged cells. The first three phases (G1, S, and G2) together are called the interphase. They play paramount role in regulating the cell cycle. At the end of each phase is a vital checkpoint where the cell is checked for DNA damage to ensure it has all the necessary cellular machinery for successful cell division. A 'molecular switch' is toggled on or off depending on the outcome of the cellular check – cells with intact DNA progresses to the subsequent step whereas cells with damaged DNA are destroyed through internal programmed cell deaths, known as 'apoptosis'.

Normal cells differ from cancer cells in a number of other aspects, apart from its

ability to evade growth suppressor and resist cell death, which have already been mentioned in the paragraph above. These include sustaining proliferative signalling, inducing angiogenesis, activating invasion or metastasis, reprogramming of energy metabolism, and evading immune destruction (Hanahan & Weinberg 2011).

6.2 Anatomy of prostate

The prostate is made up of many branching ducts surrounded by the stroma. The stroma is made up of connective tissue and muscle fibres (Chaffer & Weinberg 2011). The tissue of the prostate gland is histologically divided by scientists into four zones, listed here from innermost to outermost, which encircle the urethra like layers of an onion (Bath 2019):

- Anterior fibromuscular stroma
  - A thickened area of tissue that surrounds the apex of the prostate. It is made of muscle fibres and fibrous connective tissue. This area of the prostate does not contain any glands. CaP is rarely found in this part of the prostate.
- Peripheral zone
  - The largest area of the prostate. It can easily be felt by the doctor during a digital rectal exam (DRE).
- Central zone
  - Lies behind the transition zone and surrounds the ejaculatory ducts, which run from the seminal vesicles to the prostatic urethra. Very few CaPs start in the central zone.
- Transition zone
  - Surrounds the part of the urethra that passes through the prostate (called the prostatic urethra).

6.3 Colon's TNM staging system

Table 6.1: Colon AJCC grading system (American Cancer Society 2020*a*).

Stage	Stage grouping	Description

0	Tis, N0, M0	The cancer is in its earliest stage. This stage is also known as carcinoma in situ or intramucosal carcinoma (Tis). It has not grown beyond the inner layer (mucosa) of the colon or rectum.
I	T1 or T2, N0, M0	The cancer has grown through the muscularis mucosa into the submucosa (T1), and it may also have grown into the muscularis propria (T2). It has not spread to nearby lymph nodes (N0) or to distant sites (M0).
IIA	T3, N0, M0	The cancer has grown into the outermost layers of the colon or rectum but has not gone through them (T3). It has not reached nearby organs. It has not spread to nearby lymph nodes (N0) or to distant sites (M0).
IIB	T4a, N0, M0	The cancer has grown through the wall of the colon or rectum but has not grown into other nearby tissues or organs (T4a). It has not yet spread to nearby lymph nodes (N0) or to distant sites (M0).
IIC	T4b, N0, M0	The cancer has grown through the wall of the colon or rectum and is attached to or has grown into other nearby tissues or organs (T4b). It has not yet spread to nearby lymph nodes (N0) or to distant sites (M0).
IIIA	T1 or T2, N1/ N1c, M0 OR	The cancer has grown through the mucosa into the submucosa (T1), and it may also have grown into the muscularis propria (T2). It has spread to 1 to 3 nearby lymph nodes (N1) or into areas of fat near the lymph nodes but not the nodes themselves (N1c). It has not spread to distant sites (M0).

	T1, N2a, M0	The cancer has grown through the mucosa into the submucosa (T1). It has spread to 4 to 6 nearby lymph nodes (N2a). It has not spread to distant sites (M0).
IIIB	T3 or T4a, N1/ N1c, M0	The cancer has grown into the outermost layers of the colon or rectum (T3) or through the visceral peritoneum (T4a) but has not reached nearby organs. It has spread to 1 to 3 nearby lymph nodes (N1a or N1b) or into areas of fat near the lymph nodes but not the nodes themselves (N1c). It has not spread to distant sites (M0).
	OR	
	T2 or T3, N2a, M0	The cancer has grown into the muscularis propria (T2) or into the outermost layers of the colon or rectum (T3). It has spread to 4 to 6 nearby lymph nodes (N2a). It has not spread to distant sites (M0).
	OR	
	T1 or T2, N2b, M0	The cancer has grown through the mucosa into the submucosa (T1), and it may also have grown into the muscularis propria (T2). It has spread to 7 or more nearby lymph nodes (N2b). It has not spread to distant sites (M0).
IIIC	T4a, N2a, M0	The cancer has grown through the wall of the colon or rectum (including the visceral peritoneum) but has not reached nearby organs (T4a). It has spread to 4 to 6 nearby lymph nodes (N2a). It has not spread to distant sites (M0).

	OR	
	T3 or T4a, N2b, M0	The cancer has grown into the outermost layers of the colon or rectum (T3) or through the visceral peritoneum (T4a) but has not reached nearby organs. It has spread to 7 or more nearby lymph nodes (N2b). It has not spread to distant sites (M0).
	OR	
	T4b, N1 or N2, M0	The cancer has grown through the wall of the colon or rectum and is attached to or has grown into other nearby tissues or organs (T4b). It has spread to at least one nearby lymph node or into areas of fat near the lymph nodes (N1 or N2). It has not spread to distant sites (M0).
IVA	Any T, Any N, M1a	The cancer may or may not have grown through the wall of the colon or rectum (Any T). It might or might not have spread to nearby lymph nodes. (Any N). It has spread to 1 distant organ (such as the liver or lung) or distant set of lymph nodes, but not to distant parts of the peritoneum (the lining of the abdominal cavity) (M1a).
IVB	Any T, Any N, M1b	The cancer might or might not have grown through the wall of the colon or rectum (Any T). It might or might not have spread to nearby lymph nodes (Any N). It has spread to more than 1 distant organ (such as the liver or lung) or distant set of lymph nodes, but not to distant parts of the peritoneum (the lining of the abdominal cavity) (M1b).

IVC	Any T, Any N, M1c	The cancer might or might not have grown through the wall of the colon or rectum (Any T). It might or might not have spread to nearby lymph nodes (Any N). It has spread to distant parts of the peritoneum (the lining of the abdominal cavity), and may or may not have spread to distant organs or lymph nodes (M1c).

6.4 Origin of ATR spectra distortion and the correction methods

The spectra distortion in ATR can generally be attributed to anomalous dispersion – a phenomenon caused by the sharp change of the refractive index of a sample in the vicinity of the absorption peak (Hancer et al. 2016), as depicted in Fig. 6.1. 'Anomalous dispersion' is not to be confused with 'normal dispersion' [1].

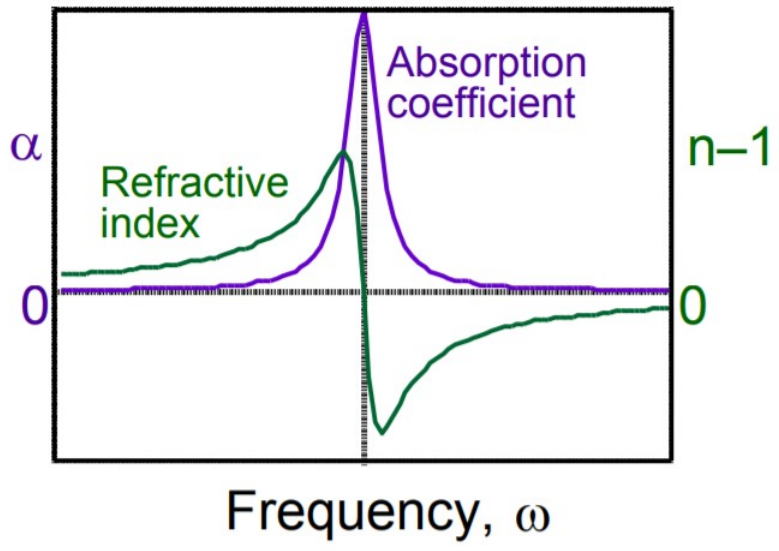


Figure 6.1: Variation of refractive index and absorption coefficient with frequency

[1] The term 'Anomalous dispersion' refers to the decrease of refractive index when frequency increases $\left(\frac{dn}{d\omega} < 0\right)$ as opposed to 'normal dispersion' when the refractive index increases with increasing frequency $\left(\frac{dn}{d\omega} > 0\right)$.

Although ATR based measurement has been favoured for biomedical applications for the lack of the scattering artefacts in the observed spectra caused by the RMieS mechanism, band shift in ATR is not uncommon (Miljkovic et al. 2012). A study of protein by (Boulet-Audet et al. 2010) with ATR-FTIR spectroscopy has demonstrated that a significant shift of the amide I and amide II bands towards low wavenumber is induced by anomalous dispersion effect at the interface between the sample and IRE and the effect is more pronounced when a low refractive index IRE such as diamond ($n_{diamond}$=2.4)was used. It was found that the frequency shifts in the amide I and II manifolds of protein films on a diamond IRE are as high as $20\,\mathrm{cm}^{-1}$, whereas these shifts were about $4\,\mathrm{cm}^{-1}$ on a Ge IRE (n_{Ge}=4.0). To obtain ATR spectra that are close to their transmittance counterparts, a high refractive index IRE is recommended so the angle of incidence of IR light is far from the critical angle. Similar observation in the shifting of the amide bands is also reported for ATR measurement on a historical collagen-based leather book in a more recent study (Vichi et al. 2018).

The increase in penetration depth with wavelength is corrected with most spectroscopic software by incorporating the term $\left(sin^2\phi - \frac{n_2}{n_1}\right)^{-1/2}$ to the mathematical formula of d_p in Harrick's equation (Griffiths & de Haseth 2007). On the other hand, the correction for the spectral shift is less straightforward. Advanced ATR correction algorithm is introduced to correct for the anomalous dispersion (Averett et al. 2008), as follows:

$$A_{corr} = \left(log_{10}\ e\ \frac{n_2}{n_1} \frac{E_0^2}{cos\ \phi} \frac{d_p}{2}\ \alpha \right) \tag{6.1}$$

where A_{corr} is the corrected ATR intensity; E_0 is the electric field of the evanescent wave, and α is the absorption coefficient of the sample. In order to apply the advanced ATR correction, the inputs required are the refractive indices of the sample and IRE, the angle of incidence, and the number of reflections. Alternatively, KK transformation can be used to transform reflectance spectra $R(\widetilde{\nu})$ to phase change (Bertie & Lan 1996). In an ATR-FTIR experiment, the measured reflectivity spectrum is a complex function of absorption $k(\widetilde{\nu})$ and refractive $n(\widetilde{\nu})$ index spectra (Huang & Urban 1992). The real and imaginary parts of $R(\widetilde{\nu})$ can be obtained from the following equations:

$$n(\widetilde{\nu}) = \frac{1 - R(\widetilde{\nu})}{1 + R(\widetilde{\nu}) - 2\sqrt{R}\ cos\ \phi} \tag{6.2}$$

$$k(\widetilde{\nu}) = \frac{-2\sqrt{R}\ sin\ \phi}{1 + R(\widetilde{\nu}) - 2\sqrt{R}\ cos\ \phi} \tag{6.3}$$

$$\text{where } R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2} \tag{6.4}$$

Comparing both advanced ATR correction algorithm with the KK transformation performed on ATR spectrum of cinnabar (HgS) in one study, it was found that the former gives a simulated resultant transmission spectrum that closely resembles the actual transmission spectrum; while the Kramer-Kronig transformation, despite successfully removing the ATR distortion features, overcompensates causing an considerable increase in wavenumbers (Kendix 2009).

6.5 Source and detector

6.5.1 Source

A globar is the simplest and cheapest broadband thermal light source to generate mid IR. The globar source is found in both spectrometers used for the work reported in this thesis, namely the Tensor 27 (Bruker Corp.) and Varian 670 (Agilent Technologies, Inc.). It consists of a silicon carbide rod which is electrically heated up to 1000 °C–1650 °C and radiates like a black-body radiator. The output radiation has a large spectral emission range (6000 – 50 cm^{-1}) but low spectral intensity (radiant intensity per wavelength) (Hermes et al. 2018). Recently, other more powerful sources to generate IR is becoming more popular such as that of Quantum Cascade Laser (QCL) or the synchrotron source (see 'Outlook').

6.5.2 Detector

The detector functions to convert light reaching it to an electrical signal proportional to the light intensity. The most common detector utilized in IR spectrometers are either thermal detectors or quantum detectors (Chalmers & Griffiths 2006). Out of these two, the former is the more common source. There are two most commonly used thermal detectors for the detection of mid-IR region (Stuart 2004). The temperature-sensitive pyroelectric device incorporating deuterated triglycine sulphate (DTGS) in an alkali halide window suitable for room temperature operation is the normal detector for routine use in commercial spectrometer. The working principle of a DTGS detector is simple – upon absorption of IR radiation, the crystal heats up, causing a change in its polarizability and consequently, a charge is generated which is detected by two parallel electrodes (Kempfert et al. 2001). For more sensitive work, mercury cadmium telluride (MCT) is preferred over DTGS; however, it requires low temperatures to operate properly; cooling with liquid

nitrogen bath is commonly featured although thermoelectric cooling may be found in modern spectrometers (Rogalski 2005). The sensitivity of the detectors can be compared in terms of their specific detectivity, D* [2] . The MCT detector has a D* of 6.4×10^{10}, over 200 times more sensitive than the DTGS detector with a D* of 2.7×10^{8}(Kempfert et al. 2001).

[2]Specific detectivity is given by $D^* = \frac{\sqrt{\text{Photon sensitive area} \times \text{Bandwidth}}}{\text{Noise equivalent power}}$, commonly expressed in Jones unit (Jones 1960).