The Development of FTIR-Imaging for the Study of Human Prostate Cancer Biopsies

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Engineering and Physical Sciences

2012

Konrad Matthew Dorling

School of Chemical Engineering and Analytical Science
1. INTRODUCTION

1.2 PROSTATE CANCER

1.2.1 Statistics and Risk Factors
1.2.2 The Anatomy and Function of the Prostate Gland
1.2.3 Fundamentals of Prostate Cancer
1.2.4 Symptoms of Prostate Cancer
1.2.5 Prostate Cancer Screening Methods
  1.2.5.1 Digital Rectal Examination (DRE)
  1.2.5.2 Trans-Rectal Ultrasound Scan (TRUS)
  1.2.5.3 Prostate Specific Antigen (PSA) Test
1.2.6 Prostate Cancer Assessment
  1.2.6.1 Gleason Grading
  1.2.6.2 Histopathological Features and Immunohistochemistry of the Prostate Tissue
    1.2.6.2.1 Epithelium
    1.2.6.2.2 Stroma
  1.2.6.3 TNM Classification
1.2.7 Prostate Cancer Treatment Methods
1.3 PROJECT AIMS

2. THEORY OF EXPERIMENTATION AND DATA ANALYSIS

2.1 EXPERIMENTAL THEORY

2.1.1 Infrared (IR) Theory

2.1.1.1 The IR region of the Electromagnetic (EM) Spectrum

2.1.1.2 Energy of a Molecule

2.1.1.3 Molecular Vibrations

2.1.1.4 Stretching Vibrations

2.1.2 Fourier Transform Infrared (FTIR) Spectroscopy

2.1.2.1 Benchtop FTIR spectrometry

2.1.2.1.1 FTIR spectrometer

2.1.2.1.2 Microscope

2.1.2.1.3 Detectors

2.1.2.2 Fourier Transforms, FTIR Spectra and Peak Assignment

2.1.2.3 Synchrotron FTIR

2.1.3 FTIR Imaging and Focal Plane Array Detectors

2.1.4 Other Spectroscopic Techniques

2.1.4.1 Raman Spectroscopy

2.1.5 Applications of Vibrational Spectroscopy and Imaging in Cancer Diagnosis

2.2 PRE-PROCESSING AND MULTIVARIATE ANALYSIS TECHNIQUES

2.2.1 Pre-Processing Analysis

2.2.1.1 Linear Derivatives

2.2.1.2 Vector Normalisation

2.2.1.3 Extended Multiplicative Scatter (Signal) Correction (EMSC)

2.2.1.4 Resonant MIE Scattering EMSC (RMieS-EMSC)

2.2.1.5 Noise Reduction

2.2.1.6 Quality Testing

2.2.2 Multivariate Analysis (MVA)
2.2.2.1 Principal Component Analysis (PCA) 71
2.2.2.2 Linear Discriminant Analysis (LDA) 72
2.2.2.3 Hierarchical Cluster Analysis (HCA) 73
2.2.2.4 K-Means Clustering 75
2.2.2.5 Artificial Neural Networks (ANNs) 77
2.2.3 Methods Used 78

3. EXPERIMENTAL METHODS, DEVELOPMENT AND OPTIMISATIONS 79

3.1 SAMPLE PREPARATION 79
3.1.1 Prostate Tissue Biopsy Sections 79
3.1.2 Window Substrate 80
3.1.3 De-waxing Protocol 81

3.2 FTIR IMAGING PROTOCOL AND DEVELOPMENT 82
3.2.1 FTIR Microspectroscopy 82
3.2.2 Software Parameters 83
3.2.3 Early Imaging Results 85
3.2.4 FTIR Imaging Protocol Refinements 88

3.3 OPTIMISED FTIR IMAGING RESULTS 90
3.3.1 Initial HCA Results 90
3.3.2 Refined HCA Results 93
3.3.3 Consultation with Histopathologist 97
3.3.4 Further Development of Analysis and Return to the Histopathologist 98

3.5 CONCLUSIONS 105

4. PROSTATE CANCER PROGNOSIS STUDY 106

4.1 FOUNDATIONS OF STUDY 106
4.1.1 Introduction 106
4.1.2 Foundation Work 106
4.2 CONTINUATION OF RESEARCH

4.2.1 Using the Correct Number of Principal Components

4.2.2 Results and Reanalysis

4.2.3 Outlook

4.3 FTIR IMAGING OF ORIGINAL SAMPLE SET

4.3.1 Data Collection

4.3.2 Imaging Results and Analysis

5. LARGE-SCALE METASTATIC AND NON-METASTATIC STUDY

5.1 PROJECT OUTLINE

5.1.1 Importance of Sample Set Consistency

5.1.2 Samples and Sample Preparation

5.1.3 Experimental Protocol and Parameters to Use

5.2 DATA COLLECTION

5.3 MANAGING THE DATA

5.3.1 Immediate Pre-Processing

5.3.1.1 Noise Reduction

5.3.1.2 Application of RMieS-EMSC Using High-Throughput Computing

5.3.2 Image Overlaying and Spectral Extraction

5.3.3 Further Quality Control and Pre-Processing

5.4 DATA ANALYSIS

5.4.1 PCA Results

5.4.2 ANN Training and Testing Results

5.4.2.1 Comparing Non-Metastatic Cancerous to Non-Cancerous Epithelium

5.4.2.2 Comparing Non-Metastatic to Metastatic Cancerous Epithelium

5.5 FURTHER DISCUSSION OF RESULTS

5.5.1 Non-Metastatic Cancerous and Non-Cancerous Epithelium Results Comparison

5.5.2 Non-Metastatic and Metastatic Cancerous Epithelium Results Comparison
5.6 SUMMARY AND CONCLUSIONS

6. PROJECT CONCLUSIONS AND FUTURE WORK POTENTIAL
   6.1 FTIR Imaging Development and Tools
   6.2 Metastatic and Non-Metastatic Cancer Study

7. REFERENCES
ABSTRACT

The potential of using FTIR imaging as an analytical technique combined with pre-processing and multivariate analysis methods was investigated. FTIR spectroscopy has been used in the past to investigate aspects of prostate cancer cells and tissues, successfully showing the separation of spectral data taken from benign prostate samples and cancerous prostate samples of varying Gleason grade. This work was ground-breaking, diagnosing different grades of cancer in the same way to the original histopathologist-assigned grades of the tissue. The advent of FTIR imaging and its recent commercial availability has allowed the much more specific collection of FTIR spectra in the form of infrared images corresponding to hyperspectral cubes of data.

Optimised protocols for FTIR imaging were developed for the collection of such images from prostate cancer tissue samples, so that the highest quality data could be obtained as time efficiently as possible. With the recent development of a resonant Mie scattering correction algorithm, the pre-processing of data could be done rigorously, eliminating all physical effects from spectral data for the first time. Hierarchical cluster analysis and K-means cluster analysis were employed as image clustering methods to classify the tissue based on morphology. Imaging data that had RMieS-EMSC, vector normalisation and a second derivative applied showed the best cluster assignment as advised by an experienced histopathologist.

A large scale study was devised based on the author’s previous work to try and classify metastatic from non-metastatic prostate cancer epithelium using FTIR images. A method for the isolation of epithelial spectra was devised by the immunohistochemical staining of the tissue sample after data collection to highlight the epithelium, and overlaying the optical image of the stained tissue with the FTIR image. Resulting epithelial spectra were extracted from the FTIR images of the two tissue classes. Principal component analysis was applied to the data, and artificial neural network were constructed using training and test sets of patient-associated spectra. Experiments were done investigating whether non-metastatic cancer epithelium classified differently to non-cancerous epithelium, and whether epithelial spectra from patients with metastatic cancer would classify differently to patients with non-metastatic cancer. The non-metastatic data did not separate well from the non-cancer data. PCA results showed the metastatic data separated from the non-metastatic data very well, and seemingly robust ANNs were also developed to classify the data.
DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

K. Dorling  30/09/2012
COPYRIGHT STATEMENT

(i) The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

(ii) Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

(iii) The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

(iv) Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DoculInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.
DEDICATION

This PhD brings to a close my long 9 year stint at the University of Manchester. It has not been easy, and I would never have got here if not for a lot of people, that have offered me so much. Thanks to...

Pete, for your ongoing support and belief in me, and for the opportunities you have given me.

The rest of the Gardner group, in particular to Caryn, Alex and Graeme, for your friendship and all your help with everything throughout the years

John, your crafty pints, and Mind Betrayal, for the Rock, and some of the most memorable fun times I will never forget.

Erik, for your comradery and natural pessimism, you would always lighten my mood in my hard times, even if it wasn’t the intention. You are truly the only one in the ‘real world’ who really understood.

Johan, for always being my best buddy, regardless of anything else. I know I can count on you always, no matter what I need.

Morsan, Papps and Sebastian for always giving me the encouragement and support I needed, and all the love a family can give. I love you all so much.

Becki, my beautiful wife. You have stood by me through thick and thin, and I know it hasn’t been easy. You have given me the strength to achieve this fantastic feat, and kept me calm and sane along the way. I love you always, forever and ever.

All my other friends and family in England, Sweden, Poland, and around the world - there are too many to list! I have been blessed with such an amazing collection of you.

Finally, I have to thank God for giving me the will to finish what I started, my faith will never falter.

Thankyou all for always being there.
1. INTRODUCTION

1.1 CANCER

Cancer is a term for diseases that are characterised by the development of abnormal cells that continue to divide uncontrollably[1]. The uncontrollable cellular development usually results in a formation of a mass of cells known as a tumour. If a tumour turns malignant, it grows rapidly, becoming dangerously invasive and is often life-threatening. Benign tumours are the opposite and have a slow progression rate and are non-invasive, meaning that people can usually live with them comfortably.

The various existing forms of cancer are broadly classified according to the type of cell in which the primary tumour originates. The four major classes of cancer are carcinomas (originating in the epithelial cells in organs and glands), sarcomas (originating in the mesenchymal cells found in bone, muscle, fat and cartilage), lymphomas (originating in the lymphocytes found in part of the make-up of the immune system and lymphatic tissue) and leukemias (cancers of the blood caused by the abnormal production of immature blood cells in bone marrow). The class of cancer can be further identified by an appropriate prefix given denoting the exact cell type such as Oesteo- (referring to cancer originating in the bone) or Adeno- (referring to a glandular cell type)[2, 3].

Cancer development is related to an imbalance of homeostasis. Normal cells divide in an orderly manner due to the balance of growth-promoting genes (oncogenes) and growth-inhibiting genes (tumour suppressor genes). In the case of cancer, the balance is lost due to the activation of oncogenes and the inactivation of tumour suppressor genes. This abnormal process happens as a result of certain defects and changes in the cell genome. The mutated cells acquire a number of biological capabilities during their development in tumorigenesis, giving them an advantage over normal cells and allowing them to become stronger and exhibit parasitic
behaviour. In 2000, Hanahan and Weinberg first described the six biological capabilities cells acquire during the multistep process of tumour development and are represented in figure 1.1[4].

![Diagram showing the six hallmark capabilities of cancer.](image)

**Figure 1.1.** The six hallmark capabilities of cancer. Reproduced from Hanahan and Weinberg[5].

The six steps relate to the induction and sustenance of positively acting growth-stimulatory signals, the reduction of response to growth suppression cell signalling pathways, the resistance to apoptotic signalling (cell death), the reacquisition of cell immortality, the abilities to induce angiogenesis (blood vessel growth) for tumour expansion and avoid anoikis (a trigger for apoptosis in cells that become separated from their native extra cellular matrix (ECM)) allowing metastatic cancer cells to move freely through the body and invade other organs[4-6].

There are many structural differences between normal and cancer cells. Normal cells have a single nucleus and nucleolus, while cancer cells can have large, multiple nuclei and nucleoli. Normal cells also tend to have a built-in blood transport mechanism, a large cytoplasm and fine chromatin, whereas cancerous cells have a smaller cytoplasm, coarse chromatin and do not have a natural built-in blood transport mechanism.
1.2 PROSTATE CANCER

1.2.1 Statistics and Risk Factors

Unless otherwise mentioned, all facts and statistics on prostate cancer in this subsection are the most recent recorded data provided through the Cancer Research UK information website[7].

Prostate cancer is the most common cancer among men in the UK, making up a quarter of all new cancers diagnosed. It is difficult to know how common the incidence of prostate cancer is globally as cancer registration systems are scarce in many parts of the world; however, the International Agency for Research on Cancer does routinely use available global data to estimate worldwide cancer incidence[8]. Prostate cancer is the fifth most common cancer in both sexes combined and is estimated to account for 7% of all new cancer cases worldwide in 2008. Prostate cancer incidence has been increasing steadily since the 1970’s, although rates have become more stable over the last few years[9]. This can be attributed to increased detection through better and more frequent tests and diagnosis methods.

No easily modifiable risk factors have been identified for prostate cancer, and with insufficient evidence it is difficult to develop good prevention strategies. There are, however, known non-modifiable risk factors. Age is the strongest known risk factor for prostate cancer, with a very low risk in men under the age of 50, and rising risk with increasing age thereafter. Prostate cancer mortality is also highly related to age. In the UK between 2007 and 2009, an average of 93% of prostate cancer related deaths were in men the age of 65 or over. Although this number seems high, survivability has vastly improved in recent years. In the 1970’s, only two in ten men diagnosed with prostate cancer would survive longer than 10 years and now it is nearly seven in ten. This is also widely attributed to increased detection rates, specifically through the increased use of Prostate Specific Antigen (PSA) tests in the UK.
Global prostate cancer incidence rates suggest risk is also affected by ethnicity. The highest incidence is in men of African descent, and the lowest is in men of Asian origin[8]. Diets high in fat, especially animal fat, are known to increase prostate cancer risk. It has also been suggested that a low selenium and vitamin E intake may contribute to a higher risk, although studies have shown that giving doses of the two to a population of relatively healthy men does not help to prevent prostate cancer[10]. Diet may somehow be related to the ethnicity risk, as traditionally a lot of fish (high in vitamin E) and low-fat foods are eaten in Asia.

Family history is also known to have an effect on prostate cancer risk. Risk increases two to three times for men whose father or brother has been diagnosed with prostate cancer[11].

Other studies suggest that working regularly with pesticides[12, 13], as well as the amount of alcohol consumption can result in a higher risk of cancer[14]. More research proposes that men regularly taking aspirin or other Non-Steroidal Anti-Inflammatory Drugs (NSAIDs)[15, 16], as well as men suffering from diabetes mellitus have a lower chance of developing prostate cancer[17, 18].

1.2.2 The Anatomy and Function of the Prostate Gland

The prostate gland is an important part of the male reproductive system. A healthy prostate is about the size of a walnut and is situated just in front of the rectum and under the bladder. The prostate can be classified into three main zones of pathological interest. The transition zone (TZ) is the area that surrounds the first part of the urethra leading from the bladder, the central zone (CZ) is the cone-shaped region of the prostate that surrounds the ejaculatory ducts, and the peripheral zone (PZ) is the majority of the remaining glandular prostatic tissue. There is also an anterior fibromuscular stroma[19]. Prostate cancer can begin or spread from any of these zones, although most cancer originates from the
peripheral zone. Figures 1.2 and 1.3 below show the location and anatomy of the prostate.

Figure 1.2. (a) The male human anatomy of the prostate and its surrounding organs. (b) A zoomed in view of the inside of the prostate, bladder and rectum. Adapted from National Cancer Institute website[20].

Figure 1.3. The anatomical zones of the prostate. PZ = Peripheral zone, TZ = Transition zone, CZ= Central zone, AFM = Anterior fibromuscular zone. Reproduced from Umbreit et al.[21].
The main function of the prostate gland is to produce and secrete prostatic fluid, one of the main components that make up semen. Prostatic fluid is a thin, milky liquid that protects the sperm from the acidic environment of the vagina. It contains a clotting enzyme which makes the ejaculate clot weakly, with the effect of holding the semen in the deeper regions of the vagina where the cervix is located. Another constituent of the fluid is prostate specific antigen (PSA), an enzyme responsible for dissolving the clot and liquefying the ejaculate. Once the clot is dissolved, the sperm are highly motile and most active with increased chances of fertilisation. Amongst other proteins, nutrients and hormones, prostatic fluid is very rich in zinc.

The prostate is useful in other ways. The muscular contractions of the prostate help to expel the semen during ejaculation, and also play a part in controlling urine flow. The growth and function of the prostate is related to the hormone, testosterone, produced in the testes.

1.2.3 Fundamentals of Prostate Cancer

Over 90% of prostate tumours are adenocarcinomas that occur in the luminal epithelial tissue of the prostate, more specifically the outer-most layer of cells lining the ducts within the prostate gland[3]. This abnormal growth of cells in the glandular tissue is relatively slow compared to other cancers, although malignant tumours can spread quickly and metastasize into other nearby tissues of the body commonly starting with nearby lymph nodes, bones, the liver and rectum. Once the cancer spreads to other organs or parts of the body, there is a much lower chance of successfully treating the cancer.

1.2.4 Symptoms of Prostate Cancer

Symptoms among men are not always very clear; some men may not notice any at all, especially in the early stages of the cancer. If symptoms do arise, they are similar to symptoms felt from non-aggressive diseases such as benign prostatic hyperplasia (BPH). In both cases, the prostate enlarges and presses on the urethra, blocking
urine flow. Most symptoms include urinary problems including the frequent need to urinate, difficulty starting and maintaining a steady flow of urine, increased urination at night, painful urination, having weak flow of urine, and blood in the urine. Achieving or maintaining an erection may also be difficult. In advanced stages of prostate cancer, patients may feel pain in the lower back and thighs, bones and may become incontinent. Most of these more serious symptoms however are uncommon.

1.2.5 Prostate Cancer Screening Methods

1.2.5.1 Digital Rectal Examination (DRE)

The DRE procedure is one of the first checks a medical examiner will do for prostate cancer. It involves them inserting a gloved, lubricated finger into the rectum to check the size, shape and texture of the prostate gland directly. Areas that appear irregular, hard or lumpy need further evaluation, as they may contain cancer. If cancer is present, it may feel hard and knobbly, whereas somebody suffering from BPH usually has an enlarged, smooth and firm prostate. The prostate can also feel normal, even when there are cancer cells present. Prostate cancer that is apparent from a DRE is generally at a more advanced stage[22]. Thus, it is important to do further tests.

1.2.5.2 Trans-Rectal Ultrasound Scan (TRUS)

The TRUS is used to determine the size and density of the prostate gland. A small probe is inserted into the rectum and the ultrasound waves emitted produce a sonogram of the prostate on a screen. The procedure may feel uncomfortable, but is relatively quick. The TRUS also helps to determine which exact areas of the prostate to biopsy, if needed[23].
1.2.5.3 Prostate Specific Antigen (PSA) Test

A PSA test involves taking a sample of blood from the patient to measure the levels of PSA in the bloodstream. PSA is a protein produced by both cancerous and non-cancerous prostate cells. Men with prostate cancer usually have a higher level of PSA in their blood. The test is not completely reliable because there is no ‘normal’ level of PSA; it varies from man to man. PSA levels also naturally get higher as men get older, and raised levels can also be attributed to urinary or prostate infections, recent prostate biopsies/surgery, and BPH. Rough guidelines to PSA readings are 3ng/ml for men under 60, 4ng/ml for men aged 60-69, and 5ng/ml for men over the age of 70[24, 25].

Occasionally, patients with ‘normal’ PSA readings may still have prostate cancer, but in general, the higher the PSA level above the guided amounts, the more likely it is for the patient to have prostate cancer. High readings can also just be attributed to benign prostate diseases, so it is important that a PSA test is not used as the only indicator for diagnosing a case of prostate cancer[26]. Some men can have PSA levels in the hundreds or thousands once diagnosed with prostate cancer; and the higher the PSA level, the more likely it is the cancer will spread quickly.

PSA tests may also be used to monitor prostate cancer treatments and how well they are working. If the levels are stable, it is an indication that the cancer has stopped growing or spreading. Successful treatments will most likely cause PSA levels in the blood to fall[27].

1.2.6 Prostate Cancer Assessment

If a further evaluation of the prostate is needed, a biopsy will be taken. In most cases, a needle biopsy will be performed to obtain the samples necessary. The biopsy is normally transrectal (taken through the rectum), transperineal (taken through a small incision in the perineum), or transurethral (taken through the
urethra). A prognosis will then be given determining whether the cancer is benign or malignant; and a grading can then be done on the tissue sample, followed by a staging classification of the tumour. The American Joint Commission on Cancer (AJCC) has a recommended defined system of grading tumours according to tissue architecture[28]. The histologic grading system is as follows.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GX</td>
<td>Grade cannot be assessed</td>
<td>G1</td>
<td>Well differentiated</td>
</tr>
<tr>
<td></td>
<td>(Undetermined grade)</td>
<td>G2</td>
<td>Moderately differentiated</td>
</tr>
<tr>
<td></td>
<td>(Low grade)</td>
<td>G3</td>
<td>Poorly differentiated</td>
</tr>
<tr>
<td></td>
<td>(Intermediate grade)</td>
<td>G4</td>
<td>Undifferentiated</td>
</tr>
<tr>
<td></td>
<td>(High grade)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There are also some more specific cancer grading systems for different cancers. The widely used grading systems are the Bloom-Richardson system for breast cancer[29], the Fuhrman system for kidney cancer[30], and the Gleason grading system for prostate cancer[31].

**1.2.6.1 Gleason Grading**

Donald Gleason was the lead in a urological research group that proposed a grading system for cancerous cells in the prostate in 1966. The system is still in use today and is the sole foundation behind prostatic tumour assessment. Gleason grades range from 1 (lowest grade) to 5 (highest grade). The prognosis is based on the two most predominant histological patterns of cells in a prostate tissue sample, and their relative Gleason grades are added together to obtain a Gleason score. For example, if the most common pattern was Gleason grade 4, and the second most common pattern was Gleason grade 3, the Gleason score would be 4+3=7. It is notable that a cancer with a 4+3 Gleason score is more aggressive than one with a 3+4 Gleason score, even though both add up to 7. The original Gleason grading criteria are shown in figure 1.4.
Figure 1.4. The architectural pattern of cells determining Gleason grades 1 to 5. The illustration is Gleason’s own simplified drawing of the 5 grades of prostate cancer. Reproduced from Gleason[32].

Gleason briefly described the differences between the grades. Grade 1 is represented by simple round glands, close packed in rounded masses with well-defined edges. Grade 2 tissue structure is similar but glands are less closely packed and have loosely-defined edges. Grade 3 cancer is made up of glands of irregular shapes and spacing with badly-defined infiltrating edges. Grade 4 tissue has less glandular structure with different sized glands starting to fuse together into cords, chains or ragged, infiltrating masses. In grade 5 pattern tissue, separate glands are no longer distinguishable having fused together into smooth, solid, rounded masses of epithelium with central necrosis or ragged sheets of anaplastic adenocarcinoma[32]. Examples of the Gleason grading system are shown in figure 1.5.
Pathological grading systems, including the Gleason grading system, have a major recognised flaw. They are substantially subjective. Both intra-observer and inter-observer variation exist, as well as issues of both undergrading and the over-diagnosis of Gleason grades 2-4 because the patterns are sometimes difficult to clearly differentiate from each other.

Lattouf et al. conducted a study of 390 prostate cancer patients showing that Gleason grading by up to 15 pathologists resulted in 38.2% of tumours being undergraded and 32.6% overgraded[34]. Identical Gleason grades were assigned by the pathologists to only 29.2% of the tumours. This clearly shows the limitations of the system in clinical decision making.

A UK-based study performed by Melia et al. in 2006 showed that in the circulation of 81 prostate cancer slides amongst 9 pathologists, there was an inter-observer agreement of 78% on first readings. Intra-observer agreement reached 79%
between the pathologists for 47 of the slides that were re-circulated and read more than once[35].

Attempts to propose an objective method of prostate cancer diagnosis have been made, which are non-dependant on a histopathologist. One of the first studies to directly compare results obtained using spectroscopic techniques to Gleason grading was conducted Gazi et al[36]. They used Fourier transform infrared (FTIR) spectroscopy and assessed its relevance in the grading of prostate cancer, finding a correlation between the tissue biochemistry from FTIR-LDA results and the tissue architecture used to determine Gleason grade.

1.2.6.2 Histopathological Features and Immunohistochemistry of the Prostate Tissue

The prostate has a glandular structure and is made up of mucosal, submucosal and main prostatic glands that are arranged concentrically around the prostatic urethra[37]. The glands are lined with epithelial tissue and surrounded by fibromuscular stroma[21]. The interior of the ducts and glands is known as the lumen. Smooth muscle, veins and blood are often also identifiable within the stroma. The common components of prostate tissue are shown in figure 1.6.

![Figure 1.6. H&E stained prostate tissue labelling some morphological areas of interest where L is the lumen, S is stroma, and E is epithelium surrounding the glands. Image adapted from Peng et al.[38]](image-url)
1.2.6.2.1 Epithelium

The epithelium is one of the two major components of prostate tissue and is glandular. More than 95% of prostatic adenocarcinomas originate from the epithelial cells[39]. In benign prostate tissue, glands have two distinct layers of cells around them. These consist of a proliferative outer layer of basal epithelial cells, and a secretory luminal layer of epithelial cells. The glandular morphology deteriorates with higher degrees of prostate cancer, and one of the indicators for cancer aside from the glandular architectural changes is the absence of the layer of basal cells around the glands[40]. The lumen of the prostate glands contains prostatic secretions or other materials which may differ in benign and diseased prostates[41].

Because there are no absolute specific and sensitive markers solely for prostate cancer, various stains can be used on the prostate tissue that can be useful tools for the histopathologists. One of the most common stains is haematoxylin and eosin (H&E). It colours the nuclei of the cells dark blue and it very clearly separates the epithelium from the stromal compartment of the tissue. The nucleoli of the cells are visibly much clearer. Cells with multiple nucleoli are abnormal and are also usually a sign of cancer.

Additionally, antibody stains are commonly employed to study prostate tissue. Most normal basal cells express high molecular weight cytokeratin (HMWCK). A basal cell-specific cytokeratin stain would include antibodies such as 34βE12[42] or the more recently discovered p63 antibody[43]. The antibodies bind to the HMWCK, colouring it. There is no staining of the luminal or stromal cells. A negative stain would therefore indicate a lack of basal cells and probably cancer.

A keen interest developed in the field to also find an immunomarker that would identify prostatic adenocarcinoma through the positive staining of malignant
Xu et al. found that the gene for α-methylacyl coenzyme A racemase (AMACR), an enzyme that plays an important role in the β oxidation of branched chain fatty acids, is significantly over-expressed in prostate cancer tissue compared to in benign tissue [45]. A rabbit monoclonal antibody to AMACR was developed subsequently, and as an immunohistochemical (IHC) marker, was found to stain prostate cancer cells with little or no reactivity from benign prostate epithelial cells [46]. This stain is not perfect; however, as it sometimes does stain non-cancerous cells associated with high-grade prostatic intraepithelial neoplasia (PIN) and atypical adenomatous hyperplasia (AAH).

An ideal stain for histopathologists diagnosing prostate cancer would be one with a mix of one or multiple basal cell-specific markers such as 34βE12 or p63 antibodies, and a prostate cancer-specific marker like AMACR. Benign epithelial cells would be highlighted by the basal cell-specific markers, and any cancerous glands would be stained by the AMACR. The effect of these stains can be viewed in figure 1.7.

Figure 1.7. A stained section of prostate tissue biopsy using a cocktail of HMWCK, p63 and AMACR. The benign glands with the brown outline on the right of the image are positive for HMWCK and p63, showing weak immunoreactivity for racemase. The malignant red-stained glands on the left show a strong reaction to AMACR and a distinct lack of basal cells. Adapted from Web Pathology [33].
Another stain that is useful in highlighting the epithelial component of the prostate tissue contains the anti-pan cytokeratin antibody. This antibody is sensitive to all low and high molecular weight cytokeratins in virtually all epithelial cells[47].

1.2.6.2.2 Stroma

Prostate fibromuscular stroma comprises of stromal cells scattered across an extracellular matrix (ECM). The interactions between stromal cells and tumour cells have been linked with the growth and spread of cancer, making the differences between cancerous and benign stroma of particular interest[48]. Fibroblasts within the stroma have become of interest in recent years, as the cells are not only key in wound healing, but are also linked as prominent modifiers of tumour growth and progression[49]. In addition to fibroblasts, normal stroma is largely made up of smooth muscle cells (SMC) and to a lesser degree endothelial cells, red blood cells, lymphocytes and neurons[48]. The SMCs together with the fibroblasts are largely responsible for the synthesis of the ECM, made up of collagen and elastic fibres that provide flexibility and strength to the tissue[50].

Cancerous stroma, also known as reactive stroma, contains much fewer SMCs than in benign stroma, but has an increased amount of collagen fibres, and a new cell type known as a myofibroblast, more commonly known as a cancer- or carcinoma-associated fibroblast (CAF). CAF cells bear resemblance to both fibroblasts and smooth muscle cells and have been linked to the development and spread of tumours and metastasis[51]. Although there is now a reasonable understanding of the role of CAF cells in tumour progression, the signals that facilitate their transition from normal fibroblasts are still not fully understood[52, 53].

The variety of IHC stains is less common for the investigation of reactive stroma than for epithelium, although Vimentin is a commonly used, sensitive marker for fibroblasts and other components of stroma, and Masson’s trichrome stains can be successfully employed to distinguish between cancerous and normal stroma.
(staining the SMCs one colour and the collagen, fibroblasts and CAFs another)[50]. Research has shown BPH stroma to be fundamentally different from normal prostate fibromuscular stroma, typified by the emergence of a reactive stroma myofibroblast phenotype, highlighting the need for clear biomarkers of CAFs in cancerous stromal tissue to distinguish the stroma from benign diseases[54].

Studies have shown work towards finding potential biomarkers for CAFs in prostate tissue stroma and have included the potential of using stains such as CD90 and Fibroblast Specific Protein 1 (FSP1)[55], and recently a variety of other antibodies including Anti-Asporin (ASPN) and Stanniocalcin 1 (STC1)[56], but concluded that currently no reliable markers exist that uniformly identify CAFs.

1.2.6.3 TNM Classification

Staging is an important part of the evaluation process of any cancer, including prostate cancer, as it determines how far and how seriously the cancer has spread. It is normally only done after a histological classification of cancer. The most common system of determining cancer stage is the TNM classification method. Developed and maintained by the AJCC and the International Union Against Cancer (IUAC), it is based on the extent of the primary tumour (T), the extent of the spread to the lymph nodes (N), and the presence of any other metastases (M). Once the T, N and M are determined, they are combined and an overall stage of I, II, III or IV is assigned to the cancer. Specifically in prostate cancer, four levels of T-staging exist (T1-T4). The specific clinical staging of prostate cancer is depicted in the following table.
<table>
<thead>
<tr>
<th>Primary Tumour (T)</th>
<th>Regional Lymph Nodes (N)</th>
<th>Distant Metastasis (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TX</td>
<td>NX</td>
<td>M0</td>
</tr>
<tr>
<td>T0</td>
<td>N0</td>
<td>M1</td>
</tr>
<tr>
<td>T1</td>
<td>N1</td>
<td>M1a</td>
</tr>
<tr>
<td>T1a</td>
<td></td>
<td>M1b</td>
</tr>
<tr>
<td>T1b</td>
<td></td>
<td>M1c</td>
</tr>
<tr>
<td>T1c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1. The clinical staging of prostate cancer[28].

Different imaging methods may be employed to aid in the staging of prostate cancer. Techniques commonly used can include (but are not limited to) the following. Computed tomography (CT) can be used for cross-section images of the
body, and can be useful to see whether the cancer has spread to nearby lymph nodes or other pelvic areas. Ultrasound scans can be useful for determining local spread of cancer, while magnetic resonance imaging (MRI) produces very detailed images of the prostate and surrounding areas which can indicate whether the cancer has spread to the seminal vesicles or nearby organs like the bladder. Positron emission tomography (PET)/CT fuses two tomography techniques to detect cancer by identifying changes happening at a cellular and functional level.

1.2.7 Prostate Cancer Treatment Methods

The choice of treatment for prostate cancer depends on the staging and Gleason scores of the tumour, as well as the PSA levels of the patient. If the prostate cancer is locally confined, surgery is a common treatment, where either the whole prostate or part of it is removed. Radiation therapy may also be used for cancer that has not spread outside of the prostate gland, as an alternative to surgery, or in combination with surgery to destroy any cancer cells remaining in the area.

Another common treatment for prostate cancer is hormone therapy. Hormone therapy aims to block the androgen supply to the prostate cancer cells. Cancer cells need these male hormones to grow. Specifically, testosterone, the hormone produced in the testicles, encourages the growth and spread of prostate cancer. Hormone therapy is usually proposed when the cancer has spread from within the confines of the prostate gland. This is due to the fact that cancers that initially respond to hormone therapy typically develop resistance to the treatment after one to two years[57]. Hormone therapy can be carried out by either surgical or chemical castration, or by the introduction of oestrogen or anti-androgens to the system.

If hormone therapy stops becoming effective, or the prostate cancer has metastasised particularly to the bone, chemotherapy may be suggested. The therapy itself is commonly combined with radiation therapy, and rather than being used as a treatment to cure the cancer, it is more of a way to both alleviate pain
associated with the cancer and to attempt to extend and improve the quality of life of the afflicted patient. Some chemotherapy drugs are taken intravenously, while others are taken orally.

Another relatively newer treatment method is cryotherapy. The original and most commonly performed cryosurgery involves the ablation of the prostate[58]. The surgery, also known as cryoablation is minimally invasive, freezing the prostate and any cancer cells therein. However, there can be severe side effects such as problems with bladder control and/or impotence. The National Institute for Health and Clinical Excellence (NICE) has issued guidelines for prostate cancer treatment stating that cryotherapy should still only be used as a part of clinical trials, due to the relative newness of the treatment and the lack of knowledge and evidence on the long term effects of the treatment and the effects on quality of life[59].

Apart from these more drastic treatment methods, it is also common nowadays for patient and doctors to take a more laid back approach. Active surveillance is a modern approach that aims to identify men with early and low-risk prostate cancer who do not immediately require curative treatments, and spares patients the pain, stress and side-effects that come alongside them. It will delay necessary curative treatments until rapidly rising PSA levels or other tests indicate them necessary. Watchful waiting is another approach that is a management alternative, mainly for older men with slow-growing cancer or men who would struggle to cope with the effects of surgery and treatment. It involves a close monitoring of symptoms along with regular PSA tests and DREs, and if anything may flag an acute growth or aggressive state of the cancer, a palliative treatment such as hormone therapy will be offered. Doctors would typically only recommend a more aggressive/radical form of treatment if the patient would benefit from it for ten years or more. Watchful waiting may not be an option for younger patients or those with fast-growing tumours.
1.3 PROJECT AIMS

The research performed over the course of this thesis aims to develop and optimise the techniques of using FTIR-Imaging, pre-processing methods and multivariate analyses to analyse prostate tissue biopsies, the details of which are described in Chapter 2. The comparison and correlations between analysed images and tissue histopathology is important, and are achieved with the beneficial involvement and advice of an experienced histopathologist.

The research intends to use FTIR-imaging to build on previous work done by the Gardner group (The University of Manchester) and the work summarised in the writers own MSc thesis[60] in the aims of achieving an accurate and objective prostate cancer diagnosis and prognosis using FTIR spectroscopy and multivariate analysis. This work will specifically investigate a large, consistent sample set of prostate cancer tissue biopsies and use optimised data analysis techniques to attempt to classify cancerous epithelial tissue that was metastatic on presentation at the time of the biopsy and tissue that was non-metastatic on presentation, and stayed non-metastatic for at least five years after the biopsy was taken.
2. THEORY OF EXPERIMENTATION AND DATA ANALYSIS

2.1 EXPERIMENTAL THEORY

2.1.1 Infrared (IR) Theory

2.1.1.1 The IR region of the Electromagnetic (EM) Spectrum

Electromagnetic radiation is a periodically changing or oscillating electric field propagating in a certain direction with a magnetic field oscillating at the same frequency but perpendicular to the electric field[61]. This is shown in figure 2.1.

![Figure 2.1. Representation of an electromagnetic wave propagating in the x-direction at the speed of light, where \( \vec{E} \) is the electric field and \( \vec{B} \) is the magnetic field. Reproduced from Bauer and Westfall[62].](image)

EM radiation can be considered as a wave or particle (photons) travelling at the speed of light. The EM waves differ from one another by both wavelength and frequency. The frequency is the number of wave cycles per unit time and is directly related to the length of one complete wave cycle (wavelength), by an inverse proportionality.

\[
\nu = \frac{c}{\lambda}
\]

Equation 2.1
where $\nu$ is frequency, $\lambda$ is wavelength, and $c$ is the speed of light.

The change in energy of an EM wave is characterised by either frequency or wavelength using Planck’s Law.

$$E = h\nu = \frac{hc}{\lambda}$$

Equation 2.2

Infrared (IR) refers to the wavelength of light in the part of the electromagnetic spectrum between the visible and microwave regions. The IR region of the spectrum comprises of the near, mid and far IR regions. Most of the relevant chemical information of a compound is normally observed in the mid-IR region of the infrared spectrum sown in figure 2.2.

Figure 2.2. (a) The electromagnetic spectrum of light. (b) The IR regions of the electromagnetic spectrum and their corresponding wavelengths and wavenumber ranges. Adapted from the handbook for organic chemistry lab[63].
Spectral positions and intensities of IR absorptions allow the gathering of information about the structure of a compound. IR spectroscopy can be used as a helpful analytical tool to assess the purity and makeup of a compound. The information displayed in the IR spectra is gathered from various changes in energy due to molecular transitions between vibrational and rotational energy states.

2.1.1.2 Energy of a Molecule

The energies associated with molecular vibrations and rotations are quantised. Molecules can exist in a variety of vibrational, rotational and electronic energy levels, and can shift between them by a sudden jump involving a finite amount of energy. These energies are combined to give the total energy of a molecule.

\[ E_{\text{total}} = E_{\text{electronic}} + E_{\text{vibrational}} + E_{\text{rotational}} + E_{\text{translational}} \]

Equation 2.3

The electronic energy relates to the energy of the electrons of the molecule, the vibrational energy is associated with vibrations of the atoms within the molecule, the rotational energy is related to the rotation of the molecule, and the translational energy is associated with the movement of the molecule as a whole.

Within a given electronic energy level of a molecule, many possible vibrational energy levels exist. Within the vibrational energy states, a molecule has multiple possible rotational states. Raising the electronic energy state from the ground state requires more energy than raising the vibrational state. The energy levels of a molecule are represented by the diagram in figure 2.3.
Changing the electronic energy state involves the absorption (through excitation) or emission (through de-excitation) of a definite amount of energy. To achieve excitation, a source of radiation would be directed onto a molecule. It would absorb the energy, or photons, exciting the electron(s) enough to jump to a higher energy level. If the radiation were collected by a detector after the interaction, it would show an overall reduction in the intensity of the beam. If an electron was in a higher energy state, it could revert to a lower state or the ground (lowest) state by emitting a photon with enough energy to make the transition.

While visible light, ultraviolet and X-ray radiation can produce enough energy to induce electronic transitions, infrared radiation cannot. IR absorption is therefore restricted to exciting only the vibrational and rotational states of a molecule. For a molecule to absorb IR radiation, the molecular vibrations or rotations must cause a net change in the dipole moment of the molecule. This happens when the varying electric field of the radiation interacts with oscillations in the dipole moment. If the frequency of the radiation is the same as the vibrational frequency, the radiation will be absorbed, causing a change in the amplitude of the vibrations and a shift of vibrational energy state.
Vibrational transitions create absorption bands throughout the near and mid-IR regions of the electromagnetic spectrum. These IR regions are used for vibrational spectroscopy.

Rotational energy level transitions require only a very small amount of energy. Typically, the low energy radiation from the far-IR region and the microwave region are enough to cause these transitions and can be used for rotational spectroscopy.

2.1.1.3 Molecular Vibrations

A molecule with \( n \) number of atoms has a total of \( 3n \) degrees of freedom corresponding to the Cartesian coordinate system of the molecule. In non-linear molecules, three degrees of freedom correspond to rotational movement, three correspond to translational movement, and those remaining correspond to fundamental vibrations. In contrast, linear molecules have two rotational and three translational degrees of freedom. The net number of fundamental vibrations in any given molecule can be characterised by two equations.

\[
\text{Number of fundamental vibrations (non-linear molecule)} = 3n - 6
\]  
Equation 2.4

\[
\text{Number of fundamental vibrations (linear molecule)} = 3n - 5
\]  
Equation 2.5

Molecules can exhibit both stretching and bending types of vibrations. Stretching vibrations can be symmetric or asymmetric, and are seen as the consistent changing of the molecular bond lengths between the atoms. Bending vibrations can be classified into four groups. Scissoring, or bending, is characterised by a change in the angle of both bonds; rocking demonstrates a change of angle between a group of atoms; wagging exhibits a change in the angle between the plane of a group of atoms; and finally, twisting occurs when the angle between the planes of two
groups of atoms changes. Bending vibrations typically occur at lower frequencies than stretching vibrations associated with the molecule.

### 2.1.1.4 Stretching Vibrations

The stretching frequency of a molecular bond can be approximated by Hooke’s law. Considering two atoms and their connecting bond as two masses joined by a spring that exhibit simple harmonic motion, the motion of the atoms is determined by the expressed force.

\[ F = -kx \]  

Equation 2.6

where \( F \) is the force, \( x \) is the displacement of the spring, and \( k \) is a constant.

Combining Hooke’s law with Newton’s second law of motion and solving the resulting differential equations, an expression for the oscillating frequency can be obtained that relates to the reduced mass of the system.

\[ \nu = \frac{1}{2\pi} \sqrt{\frac{k}{\mu}} \]  

Equation 2.7

where \( \nu \) is the frequency of vibration, \( k \) is the force constant, and \( \mu \) is the reduced mass. The reduced mass is determined from the masses \( m_1 \) and \( m_2 \) of the two atoms.

\[ \mu = \frac{m_1 \cdot m_2}{m_1 + m_2} \]  

Equation 2.8
The potential energy of the spring is given as an expression that depends on the spring’s displacement.

\[ E = \frac{1}{2} k x^2 = \hbar \nu \quad \text{Equation 2.9} \]

If this equation were true for all molecules, it would mean that a molecule could absorb energy of any wavelength. This is not the case, however, as vibrational energy is quantised. The energy of an approximate quantum oscillator corresponds to specific discrete values for each allowed energy level characterised by vibrational quantum numbers.

\[ E_\nu = \left( n_\nu + \frac{1}{2} \right) \hbar \nu \quad \text{Equation 2.10} \]

where \( n_\nu \) are the vibrational quantum numbers, \( n_\nu = 0, 1, 2, \ldots \) The simple harmonic oscillator approximation can be represented by the potential energy curve below in figure 2.4.
The lowest viable vibrational energy can be calculated using $n_v = 0$, and is called the zero point energy, $E_0$, and corresponds to the ground state energy.

$$E_0 = \frac{1}{2} \hbar \nu$$ \hspace{1cm} \text{Equation 2.11}

Classically, for a harmonic oscillator undergoing vibrational changes, only transitions to the next energy level are allowed.

$$\Delta n_v = \pm 1$$ \hspace{1cm} \text{Equation 2.12}

It is not uncommon, however, to observe transitions in a molecule with energies of more than $\hbar \nu$. This is where the simple harmonic oscillator approximation does not hold true for molecules. Although energies at low energy states are still approximate to those predicted by the harmonic oscillator, at higher energy states
they are not, because a molecular bond can break and cannot be compressed beyond a certain point. A molecule is actually an anharmonic oscillator and can be shown as a function of energy against interatomic distance. The potential energy curve for an anharmonic oscillator is shown by figure 2.5.

As interatomic distance increases, the potential energy reaches a maximum value. Also, with increasing interatomic distance and quantum numbers, energy levels become more closely spaced, meaning the transitions between those levels become smaller in energy. Larger transitions can occur between energy levels, and correspond to bands in an IR spectrum called overtones. Overtones have a lower intensity than fundamental vibrational bands.

Infrared spectra are normally scaled by wavenumber (cm$^{-1}$) which is just the inverse of the wavelength (expressed in centimetres).
\[ \bar{\nu} = \frac{1}{\lambda} = \frac{\nu}{c} = \frac{1}{2\pi c} \frac{k}{\mu} \]  

Equation 2.13

where \( \bar{\nu} \) is the wavenumber.

### 2.1.2 Fourier Transform Infrared (FTIR) Spectroscopy

#### 2.1.2.1 Benchtop FTIR Spectrometry

#### 2.1.2.1.1 FTIR spectrometer

The main components of a modern, conventional FTIR spectrometer include an infrared radiation source, a Michelson interferometer, an IR detector, and of course a sample. Spectrometers may be coupled with microscopes for easier viewing of the sample.

The interferometer is the foundation of all modern FTIR systems. Its purpose is to split the beam of IR radiation into two beams, making one beam travel a different optical distance than the other beam in order to create alternating interference fringes. The first interferometer was constructed and explained by Michelson in the late 19\textsuperscript{th} century[64]. The function of the Michelson interferometer within the spectrometer set-up can be seen in figure 2.6.
The interferometer consists of a beam splitter, a fixed mirror, and a mirror that is able to move back and forth. The signal that exits the interferometer is a result of the two recombined beams (that have travelled different path lengths) interfering with each other, and is known as an interferogram. Fast measurements can be taken using FTIR spectroscopy because all frequencies are being measured simultaneously.

The typical sample analysis process using a FTIR system can be described with a number of steps. Firstly, IR radiation is emitted from a source inside the spectrometer. The beam passes through an aperture, controlling the amount of energy presented to the sample. The beam then enters the interferometer where it splits and is recombined, and the resulting interferogram signal exits. The beam normally continues through to a coupled infrared microscope where it enters the sample, where it is either transmitted through the sample and substrate
(transmission) or transmitted through the sample, reflected off the substrate and transmitted back through the sample again (transfection).

In transmission mode, the IR radiation is transmitted through the sample and the substrate it is mounted on, and is directed on to the detector. In transflection mode experiments, the IR beam transmits through the sample but is reflected back off the substrate, and transmits back through the sample before being directed to the detector. Different substrates are used for both modes. Some common window substrates for samples to be mounted on for transmission are calcium fluoride (CaF₂), potassium bromide (KBr), and barium fluoride (BaF₂). For transflection mode experiments, some common window substrates are gold, steel, or MirrIR (glass with a highly reflective coating). If a sample is particularly thick, transmission mode is better to use because the IR beam only travels through the sample once, meaning less radiation is absorbed. Although both techniques can be used efficiently, transflection mode experiments can be more prone to light scattering effects. A diagram showing the difference between operating in each mode is shown in figure 2.7.

Figure 2.7. (A) Transmission of an incident IR beam (a) through a sample (c) and transmission slide (d). The transmitted beam is indicated by (d). (B) An incident IR beam (a) shown as it transmits through the sample (c), reflects off the surface of the reflective slide (d), and transmits back through and out of the sample with reflected beam (b). The dashed arrow represents a small amount of reflected light off the surface of the sample that may be detected.
Another mode of obtaining spectra is through a sampling technique known as attenuated total reflection (ATR). An ATR crystal is put in the path of the IR beam as it approaches the sample. The crystal is then lowered until the point of contact with the sample. The IR beam enters the optically dense crystal at an angle greater than the critical angle $\theta_c$, and undergoes total internal reflection (TIR). The TIR creates an evanescent wave that extends a few microns beyond the surface of the crystal into the sample upon contact. In regions where the sample would absorb IR radiation, the evanescent wave is attenuated. This altered energy is passed back to the IR beam which exits the opposite end of the crystal and moves on towards the detector. For successful ATR measurements, the sample must be in direct contact with the ATR crystal, and the refractive index of the crystal must be significantly larger than that of the sample, otherwise TIR will not occur in the crystal. Diamond is an ideal material to use as an ATR crystal, but more affordable common alternatives include ones made from germanium (Ge) or zinc selenide (ZnSe).

Most of the IR radiation is absorbed by the sample before the beam passes to an IR detector for measurement of the interferogram signal. The measured signal is finally digitised and sent to a computer. The computer applies the FT and the final IR spectrum is displayed on screen for further manipulation and interpretation.

2.1.2.1.2 Microscope

The integration of a microscope with an IR spectrometer results in a very powerful tool that allows for the detection of chemical species from specified locations within a sample. Being able to point the IR beam at areas of interest, and size the aperture to the area of interest makes it extremely useful when looking at biological samples such as cells or tissue. The integrated system is known as FTIR microspectroscopy and is shown in figure 2.8.
When exiting the spectrometer, the IR beam is directed through the microscope via a series of mirrors and condensers so that the radiation hits the sample at the central point of focus and then returns up through the microscope to the detector. In a typical setup, transmission experiments will see the IR beam make use of both the top and bottom condensers, while reflection mode FTIR uses only the top condenser.

There are two main types of optical microscopes. These are finite tube length microscopes and infinity corrected microscopes. Infinity corrected microscopes have many advantages, especially for FTIR spectroscopy. ‘Infinity corrected’ refers to the collimation of IR and visible beams of light which allow for higher precision or sharper focus for both visible and IR analysis. The design is a modern and superior design, adding flexibility and maximising performance.

### 2.1.2.1.3 Detectors

All FTIR systems are equipped with infrared detectors, which may vary in sensitivity and composition. Most standard FTIR spectrometers, however, are fitted with a
deuterated tryglycine sulphate (DTGS) detector operating at an ambient
temperature. While this is acceptable to measure the IR signal within the
spectrometer, the sensitivity of the detector is simply too low to measure the signal
after the IR beam passes through the aperture of a microscope. As a result of this
shortcoming, liquid nitrogen cooled mercury cadmium telluride (MCT, HgCdTe)
detectors, which have a higher sensitivity are often used. These detectors operate
in photoconductive mode, using the incident photons from the IR beam to promote
electrons from the valence band to the conduction band. The increase in electrical
conductivity generates a change in the resistance. The output signal can then be
extracted as a voltage. Photovoltaic detectors operate differently, and generate a
measurable voltage and current in response to photon bombardment, much like a
solar cell.

Properties of MCT detectors can vary greatly depending on their mercury-cadmium
(Hg:Cd) ratio. Narrow-band, mid-band, and wide-band MCT detectors all have a
trade-off between sensitivity and cut-off radiation[65]. Narrow-band detectors are
very sensitive but do not respond to radiation below 750 cm\(^{-1}\), mid-band detectors
have a cut-off of about 600 cm\(^{-1}\) but about half the sensitivity of a narrow-band
MCT, while wide-band detectors are much less sensitive but have a cut-off of 450
\(\text{cm}^{-1}\).

Few FTIR spectra of biological samples contain meaningful peaks at bands below
700 cm\(^{-1}\), therefore most FTIR microscopes are fitted with narrow-band, or in some
cases, mid-band MCT detectors[65]. Detectors with other compositions can also be
used for IR analysis, such as indium antimonide (InSb), but none are as good or as
common as the MCT detector for reaching the lower wavenumber range.

Once the encoded interferogram signal reaches the detector, computations must
then be carried out to interpret the meaningful chemical information of the sample
and produce an IR spectrum.
2.1.2.2 Fourier Transforms, FTIR Spectra and Peak Assignment

When spectra are recorded using an FTIR spectrometer, each data point contains information about the entire IR spectrum. All the infrared frequencies are encoded in the signal that passes through the interferometer within the system. In review, the interferogram is generated by making measurements of the IR signal passing through the interferometer at many discrete positions of the moving mirror. Lord Rayleigh was the first to recognise that an infrared spectrum was related to the interferogram by a mathematical operator known as a Fourier transform\[66\]. To reveal the IR spectrum, the signal must be transformed from a function in the time-domain (\(f(t)\)) to one in the frequency-domain (\(F(\nu)\)). This Fourier transform (FT) is done using the following integral and the effect is shown in figure 2.9.

\[
F(\nu) = \int_{-\infty}^{\infty} f(t) \cdot e^{-2\pi i \nu t} \cdot dt
\]

Equation 2.14

Figure 2.9. Illustration of a function in the (a) time-domain, and (b) frequency-domain following a Fourier transform.

After the FT, the IR spectrum is normally obtained as a function of percent transmittance against wavenumber. Although these transmittance spectra can be useful as they provide a good contrast between the intensities of strong and weak
bands, it can also be useful to view the percent transmittance as absorbance, because the spectrum develops a clearer shape and it can be easier to interpret as a whole. This conversion can easily be applied in computer software before spectral output, and uses the principles of the Beer-Lambert law for the conversion. Transmittance and absorbance (using different scales) essentially represent how much of the original IR intensity is left after passing through or reflecting off the sample (after being absorbed).

\[ T = \frac{\%T}{100} = \frac{l}{I_0} \]

where \( T \) is the transmittance, \( \%T \) is percent transmittance, \( I_0 \) is the intensity of the incident light, and \( l \) is the intensity of the transmitted light.

\[ A = \log_{10}\left(\frac{1}{T}\right) = \epsilon l C \]

where \( A \) is the absorbance, \( \epsilon \) is the molar absorptivity, \( l \) is the path length, and \( C \) is the concentration of the absorbing species in the sample.

To obtain an interpretable spectrum of a sample, a reference spectrum (background) is primarily taken. This spectrum shows the IR intensity with no evident sample, and is usually taken of the surface of a blank slide, the same kind of slide the sample of interest is mounted on. Once this has been done, a sample spectrum can be measured. The resulting output spectrum is normally computed and can be viewed on screen as the ratio of the sample spectrum to the background spectrum. This removes all environmental influences (unless they are not stable), and just leaves the chemical and physical information of the sample. Example spectra are shown in figure 2.10.
The quantitative relationships between signal-to-noise ratio (SNR), resolution, and measurement time are known as the FTIR trading rules.

SNR is an important concept because it improves the overall quality of spectra, and the greater the SNR, the greater the ability to detect weak absorbances. FTIR systems benefit from Fellgett’s advantage, or the multiplex advantage[67]. The rapid-scanning of the moving mirror gives the system a huge speed and sensitivity advantage over older, step-scan spectrometer technology, as a complete spectrum can be obtained during a single scan of the moving mirror and the detector observes all frequencies simultaneously. FTIR spectrometers can achieve an equivalent SNR to that of a dispersive spectrometer in a fraction of the time.
An increased spectral acquisition time, or number of scans, is the most common way of achieving a higher SNR. SNR is proportional to the square root of the total number of measurements (N).

\[ SNR \propto \sqrt{N} \quad \text{Equation 2.17} \]

\[ SNR \propto RES \quad \text{Equation 2.18} \]

SNR is also proportional to resolution (RES). High resolution spectra have a worse SNR because the signal is reduced, but may contain a lot of information. An improved resolution better detects signals very close together. Smoothing the data, and averaging the signal can also improve the SNR by reducing the noise.

The band regions of a spectrum are important to understand. Different frequencies are associated with different peaks, each representing a different type of molecular vibration. The most chemically diagnostic region of a spectrum is known as the ‘fingerprint region’, and is between roughly 500 cm\(^{-1}\) and 1800 cm\(^{-1}\). Vibrations in this region are coupled and there are many of them. Both stretching and bending vibrations occur in this region, while in the wavenumber region between 1800 cm\(^{-1}\) and 4000 cm\(^{-1}\) only stretching vibrations of molecular bonds will give rise to bands. Movasaghi et al. have done a very comprehensive assessment of the FTIR spectroscopy of biological tissues and summarise the notable infrared bands and their associated modes of vibration[68].

2.1.2.3 Synchrotron FTIR

FTIR microspectroscopy can also be performed using an IR beamline at a synchrotron. The purpose of a synchrotron is to produce synchrotron radiation (SR), which is emitted from charged particles as they accelerate and contain said radiation in well-defined beams[69]. SR is emitted over a wide frequency range
from the far-IR to near-\(\gamma\) rays, entering synchrotrons into the class of broadband sources. The emitted SR is highly collimated and polarised\[70\].

A modern synchrotron (such as the Diamond Light Source in Oxfordshire, UK, henceforth being used as the example) is a very large structure comprising of five key components: and electron gun, a linear accelerator (Linac), a booster ring, a storage ring, and the beamline itself.

The electron gun fires out accelerated electrons, produced via thermoionic emission, into the Linac. The Linac houses a series of radiofrequency (RF) cavities that operate to accelerate the electrons even further, up to relativistic energies of around 100 MeV. The electrons are also grouped into bunches within the RF cavities, and focusing element within the Linac can reduce the size of the electron beam to increase the SR brightness\[72\]. RF cavities are also present in the final component of the injection system, the booster ring. Together with a series of
bending magnets, the electrons are accelerated round the bends of the booster ring until they reach an energy of about 3 GeV, at which point they are injected into the storage ring where the SR is generated[71].

The storage ring is a complex arrangement of several components, namely bending magnets, RF units and insertion devices. The storage ring is a polygon, allowing the electrons to undergo a large orthogonal acceleration as the beam is curved round the bends by the bending magnets. This results in the production of intense SR[70]. SR can also be produced by insertion devices, located on the straight sections of the storage ring between the bending magnets just before the beamline of desired usage[73]. A series of RF units are also placed around the storage ring to replace the energy lost by the electrons when they emit the SR.

The beamline is where the emitted SR is focused into analytical instruments, such as FTIR spectrometers, for sample analysis. Synchrotron FTIR has a distinct advantage over benchtop FTIR systems when using small apertures, as the brightness advantage of a synchrotron leads to a dramatically improved SNR[65].

2.1.3 FTIR Imaging and Focal Plane Array Detectors

FTIR imaging uses the same principles as FTIR spectroscopy, with one main difference. The only thing that differs from the conventional FTIR setup is that rather than using a normal IR detector, an imaging system relies on a focal-plane array (FPA) detector to absorb the infrared signal transmitted through (or reflected off) the sample and produce an image. The signals absorbed undergo the same FT processes before viewing the image and spectra on screen.

The resulting infrared image is a thermal map of the entire windowed sample. The data is in the form of a hyper-spectral cube of information. Each pixel of the image represents an entirely separate, complete IR spectrum corresponding to that area of the sample. The number of pixels, and therefore, spectra, in the image depend
on the pixel resolution of the FPA. The highest pixel resolution, commercially available FPA IR detectors in use today have a pixel resolution of 128x128 pixels. This corresponds to each image being made up of 16384 pixels and individual spectra. Using an FPA instead of a single point detector means an incredibly fast spectral acquisition time in comparison, with it being able to take hundreds of spectra simultaneously.

The FPA is a two-dimensional device sensitive to the IR region of the EM spectrum and consists of an array of photon-sensitive pixels on the focal-plane of a lens. The FPA operates in photovoltaic mode, detecting photons at certain wavelengths and generating a voltage in relation to the number of photons absorbed at each pixel. The output voltage is then digitised and used to construct an image of the sample. FPA detectors normally have the same composition as single point detectors, so are commonly MCT or InSb, and are liquid nitrogen cooled.

Spectra obtained via a single point detector are slightly different from those that make up a hyperspectral image. A single point spectrum represents an average of all materials in a sample within an aperture controlled area, and spectral signatures from trace components may become lost after the averaging process. In imaging, due to each pixel having its own corresponding spectrum, trace components are more likely detectable if they are of the order of the pixel size imaged on the sample. This is, unless the trace particles are dispersed homogenously (relative to pixel size) throughout the sample. Thus, IR imaging has detection limits than can depend on the spatial resolution, particle size and the chemical and spatial heterogeneity of the sample. The FPA also has a limited spectral acquisition range of about 850 cm⁻¹.

Just over a decade ago, the technology that allowed the use of FPA detectors with a rapid-scan interferometer did not exist due to the very high frame rates needed by the FPA. This was until 2001, when Digilab launched a breed of second-generation FPA detectors, specifically designed for spectroscopic chemical imaging. Technology
has continued to improve, now allowing for very large FPA detectors (128x128) to operate optimally at rapid-scan speeds.

FTIR imaging systems have also been coupled to synchrotron beamlines. The original coupling of the technology used a single beamline to illuminate the FPA, resulting in a localised bright spot at the point where the SR beam touched the detector[74-76]. Although the illuminated region of the FPA yields a very high SNR, the spot size is small and greatly limits the maximum usable FPA size. To compensate for the inhomogeneous illumination, a relatively small FPA and sample area must be used, or acquisition time must be increased. The trade-off between FPA coverage and SNR has limited the use of FPA and synchrotron-based technology[77].

The Synchrotron Radiation Centre (SRC) (Stoughton, WI, USA) however, boasts the overcoming of such trade-off problems. The infrared environmental imaging (IRENI) beamline implements the use of a FTIR-imaging microscope with a 128x128 FPA to obtain a high enough spatial resolution for diffraction limited images or better[77]. The beamline draws 12 IR beams from a single bending magnet on the SRC. It then recombines them as a 3x4 bundle of beams and they are refocused to fully illuminate the FPA.

2.1.4 Other Spectroscopic Techniques

2.1.4.1 Raman Spectroscopy

Raman spectroscopy is another useful spectroscopic technique for the investigation of biological (but not limited to) samples. It can operate in the UV, visible light and NIR ranges of the EM spectrum. While IR spectroscopy is based on the theory of the absorption of IR radiation by a molecule, Raman spectroscopy measures the scattered light or radiation by the molecule. Most light scatters elastically and has the same energy as the incident light. When it does not, the process is mainly known as Raman scattering, or the Raman effect, and is the inelastic scattering of
monochromatic light (photons). These scattered photons have frequencies different to those of the incident photons. The incident photons impact the molecule, causing an energy level shift by either the molecule absorbing energy (excitation) or losing energy (de-excitation). The scattered, re-emitted photons then leave the source of impact with a lower energy (after excitation) or with a higher energy (after de-excitation). A Raman spectrum is a measure of intensity against what is called the Raman shift (cm$^{-1}$). The Raman shift is the difference between the incident light energy and the scattered light energy. Raman spectra from benign and malignant prostate tissue are shown in figure 2.12.

![Figure 2.12. Raman spectra of benign (dashed) and malignant (solid) prostate tissue, taken from research done by Stone et al.[78]](image)

Raman spectroscopy is a complementary technique to FTIR, but Raman does have some advantages over FTIR. To name a few advantages, little or no sample preparation is required, purging is unnecessary due to water vapour and CO$_2$ being weak scatterers of light, aqueous solutions are able to be measured (also due to the weak scattering properties of water), better spatial resolution, spectra can be easier to interpret as bands are narrower, sharper and sometimes more isolated and there is less noise. Raman systems however, are limited to their low spectral range, and
are relatively expensive. Fluorescence can contaminate Raman spectra, and if a high energy laser is used, it can be destructive to the sample.

Surface enhanced Raman scattering (SERS) is a technique that operates on the same fundamental Raman spectroscopy principles, but can provide a significantly enhanced Raman scattering signal by using specific rough metal surfaces as substrates for experiments. The utility of the technique is limited, with results sometimes unreliable due to the variable EM field enhancement across the surface caused by the heterogeneity of the metallic substrates causing localised ‘hot spots’. As an alternative to SERS, the Raman signal can also be enhanced by using a modified atomic force microscopy (AFM) tip to take measurements through when brought down to touch the surface of the sample. This technique is known as tip-enhanced Raman spectroscopy (TERS).

2.1.5 Applications of Vibrational Spectroscopy and Imaging in Cancer Diagnosis

The wide field of using vibrational spectroscopic techniques for clinical applications, especially in cancer diagnosis, is ever expanding, with new and improved technology allowing the correct approaches to the clinical need for optical diagnosis methods (real-time, non-invasive techniques to study the abnormalities in tissue).

One of the foundation studies for the aims of the work carried out over the course of this thesis was done by Gazi et al. in 2006. Gazi et al. were already at the forefront of prostate cancer research using FTIR spectroscopy[79, 80], but in this paper they showed that FTIR spectroscopy coupled with data processing analytical techniques is actually a promising means of reinforcing the Gleason scoring system of grading prostate cancer[36]. FTIR spectra from 40 prostate cancer (CaP) tissue biopsy samples were collected in transmission mode. Linear discriminant analysis (LDA) was applied to the spectra, a grading model was constructed and statistical analysis was done. Results initially showed a low correlation of 20% between the biospectroscopic FTIR-LDA score and the histologic Gleason score, which was not
highly unexpected due to the natures of the grading systems. However, when results were compared with a histologic score based on a three-band Gleason criterion (bands classified as Gleason score of <7, ≥7, or >7) that aims to distinguish between the aggressiveness of tumours, a much higher correlation was obtained with ≥70% sensitivity and ≥81% specificity. The results also showed that the modelling used could differentiate less from more aggressive tumours more accurately than the pathologist.

This paper initiated a follow-up study by the same research group, expanding research into additionally correlating results of TNM classification of tissue with a two-band criterion for staging (bands classified as tumour being clinically confined to the prostate, or tumours that have been deemed clinically invasive)[81]. Results showed that using principal component-discriminant function analysis (PC-DFA), the correlation between histologic Gleason scoring and the three-band Gleason criterion improved with respective sensitivities and specificities of 92.3% and 98.9%. They also highlighted the success and potential of the two-band criterion-based system for identifying characteristics that differentiate between tumours confined to the prostate and clinically invasive tumours.

Other key work in the field of the biospectroscopic analysis of prostate cancer tissue has been published, including the use of Raman spectroscopy to accurately grade prostatic adenocarcinoma[82], and its use to investigate the biochemical basis for the spectral differentiation of pathology in both the bladder and prostate[83]. The combination of FTIR with synchrotron radiation induced X-ray emission (SRIXE) methods has also been used to investigate prostate cancer biopsies[84],

FTIR Imaging has also been used to investigate prostate cancer tissues, in studies aiming to develop an automated histopathological characterisation of prostate tissue[85, 86], and the computer based diagnosis of prostate cancer using genetics-based machine learning[87].
The applications of FTIR spectroscopy are not limited to prostate cancer diagnosis by any means.

Tobin et al. combined FTIR with synchrotron radiation to investigate oral tumours sections and cultured cervical cells. The multivariate analysis of their data showed the detection of subtle chemical differences between cell types within a tumour, furthering the potential benefits of FTIR spectroscopy in screening for cancer[88].

As Gazi et al. compared FTIR spectroscopy results to the current gold standard of Gleason grading for prostate cancer, El-Tawil et al. did the same for cervical cancer comparing results to Pap smear cytology. They were able to accurately differentiate normal from abnormal cervical cells, thus concluding that FTIR spectroscopy can be used as an alternative screening method for cervical cancer[89]. Similar studies have also been done[90, 91].

FTIR microspectroscopy has also been used in studies where it has successfully discriminated between tumour sites and healthy tissue in different anatomical sites such as the colon[92], breast[93], skin[94, 95], and brain[96, 97].

FTIR imaging has also been widely applied to the field of cancer diagnosis. Research has been done using imaging data to analyse neck tumours[98], to investigate lung cancer cells growing in an artificial membrane system to study tumour cell invasion[99], and to image areas of squamous and glandular cervical cancer epithelium to establish spectral patterns for different tissue types in different stages of the disease[100].

In 2005, Levin and Bhargava explained that in the emerging field of chemical imaging research, lots of work must be done concentrating on instrumentation development, theoretical analyses, data processing and the introduction of the technique to new fields. They also go on to show the clinical applications and
potential of FTIR imaging, especially in cancer pathology, proving that it is a versatile, helpful, and interdisciplinary tool to the medical world. They conclude, however, that we are still at the initial stages of designing effective strategies for the new and novel applications of the technique[101].

Seven years on, it may be safe to say that after some years of the publishing of many successful results obtained through the use of FTIR imaging in biomedical research, we are no longer in the initial development stages, although the technology keeps improving and the learning processes must be repeated, and techniques and protocols refined.
2.2 PRE-PROCESSING AND MULTIVARIATE ANALYSIS TECHNIQUES

2.2.1 Pre-Processing Analysis

2.2.1.1 Linear Derivatives

Baseline drift is inherent in most spectra generated through FTIR spectroscopy. This noise is an undesirable addition to the important chemical information of the sample measured, as it can change the shape of the spectrum and mask important features of certain peaks and associated shifts. Factors associated with drift noise are mainly dependent on instrumental factors such as the flickering of the IR source, variations in detector response, or temperature fluctuations.

Drift noise typically has a low-frequency dominance in the noise power spectrum. This implies that the errors in measurements at different wavenumbers are correlated, the phenomenon of which is typically referred to as error covariance[102]. Error covariance indicates that the errors corrupting the observed signals do not have the behaviour of independent random variables, but rather are related to errors observed at other wavenumbers.

Most multivariate regression methods assume uncorrelated noise across a spectrum, meaning uncorrected drift noise can be detrimental to these analyses when the observed properties of the data may represent a significant difference to the analysis model’s assumed properties of the data. Linear derivatives are commonly employed methods to correct for the baseline drift before or during subsequent analysis. In principle, taking the first derivative of the data will eliminate any constant factors of the baseline and across the spectrum. Second (and higher) derivative spectra should reduce baseline effects that can be modelled as polynomial functions of the ordinal variable.
Derivative spectra naturally de-emphasise low-frequency signals and enhance high frequency signals. This highlights the main drawback to derivatising data, as it tends to reduce the SNR by accentuating noise (high-frequency signal). For this reason, usually derivatives are applied while a smoothing of the data is performed simultaneously. The most common smoothing method uses the Savitzky-Golay algorithm first described in 1964 [103], and can be applied as a low-pass filter to yield a least-squares estimate of the derivative over a window of points in the spectrum[102]. This smoothing function essentially assumes that variables near to each other in the data matrix contain similar information and averages them together to reduce noise without significant loss of the signal of interest. The algorithm fits individual polynomials to windows around each data point of the spectrum, which are then used to smooth the data. The process normally requires a user input of the filter or window width, the order of the polynomial and the order of the derivative. The larger the window and the lower the order of polynomial, the more smoothing occurs. The filter width used should be no larger than on the order of the nominal width of non-noise features of interest of the spectrum. Otherwise, crucial information may be smoothed and lost in the derivatised representation of the spectrum. Smoothing may also be applied independent of derivatising data.

Derivatives change the spectrum shape, and are not very easy to read themselves, although second derivatives display sharp troughs in the locations of the original spectral peaks. Derivatives are linear operators and do not affect any linear relationships within the data, allowing the removal of the predominant background variation, without eliminating chemical variations in the data. The effects of applying derivatives to raw FTIR spectra are shown in figure 2.13.
Figure 2.13. (a) A raw FTIR spectrum, and its (b) first derivative spectrum, and (c) second derivative spectrum.

### 2.2.1.2 Vector Normalisation

Vector normalisation is a commonly employed pre-processing technique and can be applied in combination with most other methods. If spectra exhibit much absorbance variance, larger intensities are magnified more than lower intensities. Vector normalisation aims to remove these sources of systematic variation between spectra caused by factors such as varying amounts of sample (thickness), sample degradation over time or variations in detector sensitivity.

It works by first calculating the average of the absorbance values in the spectrum by squaring all variables, summing them and calculating the square root of the result. This average value is known as the vector length and is equal to 1. The spectrum is then scaled by dividing each absorbance value by the vector length. Figure 2.14 shows the typical application of vector normalisation on FTIR spectra.
Spectra can also be scaled using a minimum-maximum normalisation process, where all data points of the spectrum are scaled between 0 (the minimum absorbance value) and 1 (the maximum absorbance value). This method can be useful when analysing a single band of the spectrum such as the amide I band.

Another form of normalisation that may be applied to IR spectra is an offset correction. This is a non-scaling method that can correct a vertical offset between a group of spectra, performing a linear correction of each entire spectrum and making a certain point of each spectrum equal the same value (most often zero).

2.2.1.3 Extended Multiplicative Scatter (Signal) Correction (EMSC)

So far, the explained pre-processing methods have namely been filtering methods, transforming data or spectra into presumably improved versions of themselves by removing undesired variations. The other category of pre-processing can be characterised by model-based methods that may allow for quantifying and separating between different physical and chemical vibrations in spectra. The different effects can be studied separately after statistical estimations of the
filtering parameters are obtained. The original model-based pre-processing technique for use with vibrational spectroscopy data will now be explained.

Spectroscopic measurements can suffer from a variety of scatter effects including those caused by background effects, varying optical path length, temperature and pressure variations and those attributed to the scattering of light. These effects in general are composed of both a multiplicative effect, and an additive effect[104]. Multiplicative scatter correction (MSC) was a model-based process first proposed by Geladi et al. 1985 to correct for both multiplicative and additive effects, and was specifically designed to deal with light scattering effects in data[105]. The method is now termed multiplicative ‘signal’ correction as a generalisation of the original term because it is also applicable to other types of data.

The MSC process aims to remove the amplification (multiplicative) and offset (additive) to prevent them from dominating over the important chemical signals in the data that can be of a lesser magnitude. It essentially corrects the baseline and amplification effects to the same average level in every spectrum. The principles of the correction are based on one of the basic laws of vibrational spectroscopy.

Remembering the Beer-Lambert law, an absorbance spectrum is directly proportional to the effective optical path length. The variations in spectra caused by optical path length variations can be labelled as the multiplicative variations. The original law is applicable for transparent samples containing a single light-absorbing chemical species and is given by equation 2.16, repeated here in vector form.

\[
A(\bar{v}) = k(\bar{v}) \cdot c \cdot b
\]  \hspace{1cm} \text{Equation 2.199}

where \( A(\bar{v}) \) is the absorbance, \( k(\bar{v}) \) is the characteristic absorptivity for a specific component at a specific wavenumber \( \bar{v} \), \( b \) is the optical path length, and \( c \) is the concentration of the light-absorbing chemical species in the sample.
Biological samples however, are too complex for this model. They have multiple light-absorbing species \((k(\nu))\), as well as many different biomolecule component spectra \((k_j(\nu))\), where \(j\) denotes the different biomolecules from \(j= 1\ldots J\) that have strongly overlapping characteristics\[106\]. The Beer-Lambert law can be modified to accommodate for this, rewritten as a superposition of absorbances of multiple species.

\[
A(\nu) = \left( \sum_{j=1}^{J} c_j \cdot k_j(\nu) \right) \cdot b
\]

Equation 2.20

where \(k_j(\nu)\) is the spectrum of constituent \(j\), and \(c_j\) is the concentration of constituent \(j\). It is assumed that the optical path length \(b\) is comparable for all constituents, and will hold for sufficiently homogenous samples. The constituent spectra of biological samples are sometimes difficult to obtain, and are often not known so to describe them fully would be only for theoretical interest. In general however, the overall shapes of IR spectra obtained from biological samples are very similar. This allows a very good approximation for the constituent spectra to be the same as the average spectrum of a sample set of spectra. As a result, the constituent spectrum \(k_j(\nu)\) can be expressed as the mean \(\bar{x}(\nu)\) of all spectra in a given data set plus any deviations \(\Delta k_j(\nu)\) from that mean.

\[
k_j(\nu) = \Delta k_j(\nu) + \bar{x}(\nu)
\]

Equation 2.21

The measured raw absorbance spectrum \(A(\nu)\) can then be elaborated inserting equation 2.21 into equation 2.20.
\[ A(\tilde{\nu}) = \left( \sum_{j=1}^{J} c_j \cdot \Delta k_j(\tilde{\nu}) \right) + \left( \sum_{j=1}^{J} c_j \cdot \bar{x}(\tilde{\nu}) \right) \cdot b \]  

Equation 2.22

To fully explain the correction that will be applied to the raw measured absorbance spectrum, the requirement is that the pure biochemical absorbance spectrum \( z(\tilde{\nu}) \) (which will be estimated), can be represented as the sum of all theoretically available constituent spectra.

\[ z(\tilde{\nu}) = \sum_{j=1}^{J} c_j \cdot k_j(\tilde{\nu}) \]  

Equation 2.23

where the sum of all constituent concentrations equals 1.

\[ \sum_{j=1}^{J} c_j = 1 \]  

Equation 2.24

This condition has significance, essentially stating that the constituent spectra are normalised such that, if multiplied with the respective concentrations, they will sum up to 100% in each measured spectrum. By inserting equation 2.24 into equation 2.22, another expression for the measured spectrum is obtained.

\[ A(\tilde{\nu}) = \left( \bar{x}(\tilde{\nu}) + \sum_{j=1}^{J} c_j \cdot \Delta k_j(\tilde{\nu}) \right) \cdot b \]  

Equation 2.25
When the overall shape of the measured spectrum $A(\bar{\nu})$ is close to the average spectrum $\bar{x}(\bar{\nu})$, equation 2.25 can be replaced by the statistical model:

$$A(\bar{\nu}) = \bar{x}(\bar{\nu}) \cdot b + e(\bar{\nu}) \quad \text{Equation 2.26}$$

where the un-modelled residual information $e(\bar{\nu})$ is defined as:

$$e(\bar{\nu}) = b \cdot \sum_{j=1}^{J} c_j \cdot \Delta k_j(\bar{\nu}). \quad \text{Equation 2.27}$$

The final MSC model is an extension of the Beer-Lambert expansion model by an additive component given by the constant baseline $a$, and can be expressed as:

$$A(\bar{\nu}) = a + \bar{x}(\bar{\nu}) \cdot b + e(\bar{\nu}). \quad \text{Equation 2.28}$$

The unknown parameters can be estimated using a linear least squares regression. When the additive and multiplicative effects, $a$ and $b$ respectively, are estimated, the corrected spectrum $A_{corr}(\bar{\nu})$ can be obtained according to the following calculation.

$$A_{corr}(\bar{\nu}) = \frac{(A(\bar{\nu}) - a)}{b} \quad \text{Equation 2.29}$$

In order to achieve a more effective separation of chemical effects and physical effects in IR spectra, Martens and Stark developed MSC into the extended multiplicative signal correction (EMSC)[107]. EMSC offers a much improved flexibility over the preceding technique, using knowledge about the analyte spectra
and interference effects to improve the estimated level of scattering. The basic version of the algorithm follows all the steps of MSC, but also adds a baseline effect with an arbitrary slope. The expression for the simple EMSC model can be described as:

\[
\mathbf{A}(\mathbf{\nu}) = a + \mathbf{x}(\mathbf{\nu}) \cdot b + d\mathbf{\nu} + \mathbf{e}(\mathbf{\nu}).
\]

Equation 2.30

Once the unknown parameters are estimated, as with MSC, spectra are corrected using:

\[
\mathbf{A}_{corr}(\mathbf{\nu}) = \frac{(\mathbf{A}(\mathbf{\nu}) - a - d\mathbf{\nu})}{b}.
\]

Equation 2.31

The EMSC model can be further extended with polynomial extensions, by constituent spectra and orthogonal subspace models[106], but due to the complexity of the models and irrelevance to the work in this thesis, they will not be discussed.

To interpret the EMSC model more easily in terms of inputs and outputs, the spectral data terms can be renamed as raw and reference spectra getting:

\[
\mathbf{Z}_{Raw} = a + b\mathbf{Z}_{Ref} + d\mathbf{\nu} + \mathbf{E}
\]

Equation 2.32

where a half arrow above a term denotes the vector form, and \(\mathbf{Z}_{Raw}\) is the raw spectrum, \(\mathbf{Z}_{Ref}\) is the chosen reference spectrum, \(a\) is the constant value for spectrum offset, \(b\) is the multiplicative scaling factor, \(d\) is the gradient of the sloping baseline, \(\mathbf{\nu}\) is the wavenumber, and \(\mathbf{E}\) is the un-modelled residual information.
In recent years, EMSC has been developed to deal with light scattering effects in IR spectra even further.

Most IR spectra of biological samples, particularly single cell spectra have been deemed to contain some sort of physical contribution due to scattering. Until recently, the precise explanation of this scattering artefact was unknown. However, findings have now shown and defined the predominant contribution to the scattering as resonant Mie scattering (RMieS) caused by a changing real refractive index near an absorption band\[108\], and attempts have been made to adequately estimate and correct the scattering by modifying the EMSC equation\[109-111\]. The culmination of the RMieS research produced an iterative algorithm to correct for the scattering based on EMSC and a Mie approximation formula, for the first time using full Mie Theory and GPU computing\[112\].

The improved understanding of the scattering efficiency came as a needed improvement as the approximation of the scattering was previously based on correcting for non-absorbing spherical particles\[113\]. However, biological samples are absorbing materials. Mie theory describes the total loss of light caused by a scattering particle, due to both the absorption properties of the particle and scattering, and requires knowledge of the particle radius and the complex refractive index ratio of the scattering particle to air\[109\]. Using this knowledge, the approximation was modified by using a refractive index spectrum of a reference, instead of a constant value\[113\].

Scattering efficiency curves were simulated to cover all possible parameter values for varying refractive indices and radii of scatterers. The curves were then decomposed using a non-mean-centred principal component analysis to find the first few principal components that cumulatively explained 99.9% of the variance of
the data[110]. The loading vectors associated with these principal components were precisely weighted, the values of which, along with the all other model parameters, are estimated by least squares linear regression.

The completed RMieS-EMSC algorithm uses an appropriate reference spectrum, such as Matrigel, which is a very good approximation for the pure absorbance spectrum of a biological material. The initially corrected spectrum exhibits a certain amount of spectral information from the reference, so the RMieS corrected output is iteratively fed back into the algorithm as the new reference spectrum. After a number of iterations, the corrected spectrum will only resemble the pure absorbance part of the raw spectrum.

The effect of applying RMieS-EMSC on a raw FTIR spectrum is shown in figure 2.15.

Figure 2.15. (a) A raw FTIR spectrum. (b) The corrected spectrum after applying RMieS-EMSC.

2.2.1.5 Noise Reduction

Noise reduction methods can be employed after data collection to improve the SNR of the spectra. This is extremely useful in practice, as it can vastly reduce data collection time by being able to use a lower number of scans or worsening the
resolution. Insight was already given by Bhargava et al. into this area more than a decade ago, proposing coaddition schemes with PCA-based filtering procedures to vastly improve the SNR of spectral data[114]. Since then, Bhargava et al. have always been at the forefront of this work, recently publishing a paper concerning the problems with the trade-off between the speed and accuracy of data collection with the SNR[115]. In this research, they demonstrate how an algorithm they have written for use post-acquisition using a minimum noise fraction (MNF) approach can reduce the SNR tenfold. Their results imply that obtaining data to the same high standard as one would normally expect is therefore possible with a reduction in acquisition time of almost two orders of magnitude. The automated nature of the approach allows the routine application of the algorithm to enhance data quality and optimise efforts in data acquisition.

2.2.1.6 Quality Testing

The quality control of data is important, especially when dealing with large spectral images and analysing them as a whole. Any ‘bad’ or unwanted data in a data set uses up valuable space and time during consequent analyses. Some form of quality test should be applied on all data to get rid of spectra that are of no use (for example areas of an FTIR image corresponding to areas where there is no tissue). A quality test for FTIR data can take many forms ranging from controls using maximum and minimum absorbance values, integrations of specific peaks, or noise levels to limit the accepted data. The quality control methods used in this project were often simple limitations of the maximum and minimum absorbance range of a spectrum, but these will be covered in more detail when relevant.
2.2.2 Multivariate Analysis (MVA)

2.2.2.1 Principal Component Analysis (PCA)

PCA is a cluster analysis tool, arguably forming the basis of many multivariate analysis methods\[116\]. It is a mathematical procedure that, when applied to spectroscopic data, aims to recognise subtle relationships in spectra that may be used to characterise those specific spectra to a particular disease state.

The origins of PCA can be traced back to 1901, when Pearson formulated the statistical analysis as finding the lines and planes of closest fit to systems of points in space\[117\]. However, the technique was first described in its modern form using principal components by Hotelling in 1933\[118\].

Principal components are a result of an orthogonal linear transformation of a data set. Given a set of observable variables (such as spectra), \(m\), and latent variables (such as wavenumbers), \(n\), a data matrix can be formed of \(m \times n\) dimensions. PCA would act as if plotting the \(m\) variables in a \(n\) -dimensional space. Any correlations in the \(m\) set of variables would hypothetically result in the correlated variables clustering closely together. Since the \(n\) -dimensional space cannot be visualised, a new set of coordinates must be assigned. PCA identifies a new set of orthogonal axes through the data, and by finding the direction of the maximal variance of the data through the coordinates in the \(n\) -dimensional space, the first principal component (PC) is obtained. The coordinates of the data points can then be mapped onto this new axis using orthogonal projection, obtaining a score plot\[119\]. Further principal components are then calculated on the condition that they are on an orthogonal axis to the previous component, that they account for the maximal variance in the \(m\) variables that has not been described by any previous PCs, and that they are uncorrelated with all preceding PCs. With a second principal component and orthogonal axis, the coordinates can again be projected, but this time into a 2D-space.
Each new computed PC accounts for less and less amounts of variance in the data set, and the cumulative percentage explained variance (summed over all PCs) will always equal 1. Thus, there can never be more PCs than variables. Usually the first 5-10 PCs of a typical spectral data set obtained in this thesis will explain the majority of the variance in the data (≈99%), which is why normally only the first handful of PCs are of any use for data analysis and interpretation.

The correlation between a PC and a variable estimates the information they share, and is known as a loading[119]. Loadings can help to determine what the variables are separating on in a multidimensional scores plot. For example, if a 2D scores plot of FTIR data reveals clear clustering, interpreting the loadings may show what areas or peaks of the spectrum are down to the clustering and separation of variables for each PC.

2.2.2.2 Linear Discriminant Analysis (LDA)

LDA is a similar technique to PCA, also looking for linear combinations of variables that best explain a data set. However, where PCA aims to find a subspace whose basis vectors correspond to the direction of the maximal variance of the variables in the data set, LDA instead searches for those vectors in the underlying space that best discriminate among classes of data[120]. LDA is one of the techniques under the umbrella term of discriminant function analysis (DFA). It is a supervised projection method used for classifying a set of variables into predefined classes. Information about sample grouping in the data set is used to produce measures of within-group variance and between-group variance[121]. This information is used to define a set of discriminant functions that optimally separate the predefined groups. A training set of the data is used to predict where data from a blind data set should fall, using a determined number of principal component scores as its basis. A test data set can then projected into the system, with the hope of correctly discriminating the data into the predefined groups.
LDA cannot be performed on raw spectral data. This is because neither co-linear variables nor too many variables can be fed into the LDA algorithm. PCA must be performed on data prior to attempting LDA because it removes co-linearities from a set of data as well as reducing the number of spectra[122]. If too many PCs are used in the LDA, clusters may seem very tightly packed and artificial, but if too few are used then no clusters may be evident that relate to the predefined groups.

### 2.2.2.3 Hierarchical Cluster Analysis (HCA)

HCA is an unsupervised clustering technique that is sensitive to very subtle spectral differences within a data set, meaning it can be very useful in separating between tissue type and disease state. Hierarchical clustering techniques use either bottom-up (agglomerative) or top-down (divisive) approaches, but both methods will assign data to clusters based on previously established clusters. Divisive hierarchical cluster analysis is uncommon, while agglomerative clustering approaches have been used successfully to classify spectral data in biomedical FTIR spectroscopy and imaging[123-126]. From this point forward, the term HCA will refer to the agglomerative approach.

The HCA algorithm has a number of steps. Primarily, each variable (spectrum) in the data set is assigned as a separate cluster, so the analysis will assume \( n \) number of clusters for a data set with \( n \) number of spectra. A distance matrix between all spectra is then calculated, containing the complete set of interspectral distances, or the measures of dissimilarity between the data. For a group of \( n \) number of spectra, the distance matrix has \( n \times n \) dimensions, and is symmetrical along the diagonal. The interspectral distances can be calculated using a variety of different methods, although using the Euclidean distance calculation (the direct geometric distance in the multi-dimensional space between spectra) is perhaps the most commonly used[127, 128]. HCA using a calculation of the distances using D-values with Pearson’s correlation coefficient has also been successfully used to achieve the best correlation between FTIR images and histology[126].
The algorithm then looks at which two clusters are least dissimilar to one another, and merges them together into another cluster. In this way, the distance matrix will have its dimensions reduced to \((n - 1) \times (n - 1)\). The distances of the new cluster to all other remaining clusters are subsequently reevaluated, and the next two least dissimilar clusters are merged. Computing the distances between newly formed clusters is not always straightforward as they sometimes contain multiple objects and a specific linkage criterion must be used to specify which point of the clusters measurements will be made from. Again, there are a selection of ways to do this[127]. The linkage using Ward’s algorithm[129] is regarded as the most efficient and most commonly employed method in FTIR HCA imaging and uses an analysis of variance approach to evaluate distances between the clusters, tending to produce denser clusters[123, 126, 130].

The HCA process is repeated in this manner until all original clusters (spectra) have been merged into one cluster. The merging process can be visualised in the form of tree-like structure known as a dendrogram[126]. In practice, a pre-determined desired number of clusters is fed into the algorithm, and the output will show information about the clustering as if the dendrogram was ‘cut’ at the point where the number of branches would equal the desired number of clusters.

Applying HCA to FTIR Imaging data will produce a pseudocolour map with as many different colours as chosen clusters. HCA outputs are shown below in figure 2.16.
Figure 2.16. (a) Typical dendrogram that has been ‘cut’ to reveal the classification of 6 different clusters, a-f. (An accurate dendrogram of FTIR imaging data taken with a 128x128 FPA would have a base of 16384 branch clusters, whereas for easier interpretation, this example shows 36.) (b) H&E stained section of prostate that corresponds to the area of tissue that will be FTIR imaged. (c) HCA image showing 6 clusters of different colours that have classified data from the associated FTIR hyperspectral image.

One drawback to HCA is the intensive computing of the multiple calculations of distance and linkage matrices, requiring lots of computer power and is very time consuming.

2.2.2.4 K-Means Clustering

K-means clustering is another cluster imaging tool that has proved its potential and use in the analysis of biospectroscopic images of tissue[126, 130-132]. It is a non-hierarchical clustering method that, like HCA, uses a ‘hard’ or two-class membership system to classify the spectral data into clusters (spectra either belong to a cluster or they do not).
K-means uses an iterative algorithm that begins by randomly selecting a number of spectra to use as initial cluster centres. The number of spectra will be the same as the number of desired clusters. The process computes a distance matrix (as with HCA) of interspectral distances. The spectra closest to a particular cluster centre will become part of that cluster. Next, new centroids of each cluster will be calculated and distance values between the centroids and each spectrum are recalculated. Clusters are reassessed, and if a spectrum is closer to the new centroid of a different cluster, it will switch membership to that cluster. This process repeats itself, recalculating new cluster centroid positions every time a spectrum changes cluster membership, until the solution is deemed stable by reaching a convergence criterion or when spectra no longer switch cluster membership[126, 130].

Figure 2.17. Simple illustration of the K-means clustering process. (a) Initial set-up. 2 data points (filled black point and filled pink point) are randomly chosen as centroids of different clusters. Each other data point is assigned to the nearest cluster centroid. (b) Results of first iteration. Cluster centroids are recalculated, with the old centroid being removed. If data points are now closer to the centroid of the other cluster, they are reassigned to belong to that cluster. (c) Results of second iteration. Cluster centroids are once again recalculated, and data points undergo cluster reassignment. In this example, the new clustered data has reached a stable state; data will not be reassigned during any further iterations. Figure adapted from Faber[133].

Fuzzy C-means is another clustering method that is very similar to K-means; however, it operates using a ‘soft’ membership system to classify data, assigning each spectral data point a degree of membership, dictating how well it fits in the
cluster to which it has been assigned. Fuzzy C-means is useful when assigning a low number of clusters that correspond to specific tissue structures, but results are more ambiguous when increasing the number of clusters in terms of the corresponding known histopathology[126].

2.2.2.5 Artificial Neural Networks (ANNs)

ANNs are mathematical models or algorithms that attempt to simulate the structure and different functional aspects of the human brain. Neural networks have significant advantages over standard pattern recognition algorithms as they learn from training sets of data rather than following a set of specific rules. ANNs can learn directly from data, and can process data that only broadly resembles that on which they were trained[134].

An ANN consists of an interconnected group of artificial neurons, or nodes. In most cases, ANNs are adaptive systems that change their structure based on internal and external information that flows through the network during the learning phase[135]. Practically, they are useful, non-linear, statistical data-modelling tools that can be used to model complex relationships between inputs and outputs, as well as finding patterns in data.

The technique can be optimised by pre-selecting appropriate spectral features from the data[136]. ANNs are typically trained on thousands of spectra, and the learning process is time consuming, although on development of the algorithm, spectra can be analysed very quickly. To set up an ANN, the spectral data will be fed into the network split in three parts; a training data set to estimate the network model parameters, an internal validation set to check the generalisation ability and stability of the network, and an external validation data set to project in and classify[137].
2.2.3 Methods Used

Over the span of this thesis, many different pre-processing and data analysis techniques were used. During the first two years of this project, most of the work was done on the development of FTIR-imaging protocols and investigation into pre-processing methods. At this point, no form of noise reduction was used due to the lack of a need for it, use of the RMieS-EMSC algorithm was limited due to its ongoing development, and the image clustering method of choice was HCA due to trends in the literature[126, 138, 139] and a lack of knowledge or testing of other methods. Once multiple images had been analysed in this way, it had become very difficult to continue using HCA on full data sets of many images purely due to the computation time and power needed. (A HCA with 5 clusters done on one image containing 16384 spectra took about 40 minutes). This problem was also apparent in the literature[130, 131], so a move towards using K-means clustering to analyse large datasets was made, as the computation time was reduced drastically (down to less than 1 minute per image). The RMieS-EMSC algorithm had been optimised for use via high throughput computing methods[112] and was also able to be applied to future data sets. PCA and ANNs were also used for the later data analysis in Chapter 5.
3. EXPERIMENTAL METHODS, DEVELOPMENT AND OPTIMISATIONS

3.1 SAMPLE PREPARATION

3.1.1 Prostate Tissue Biopsy Sections

All of the FTIR work carried out over the course of this thesis has been done looking at prostate tissue sections.

The prostate tissue specimens have been sourced as paraffin embedded blocks taken from the tissue banks organised and held by both the Christie NHS Foundation Trust and Salford Royal NHS Foundation Trust Hospitals. All tissue samples were collected with full ethical committee approval and are all in accordance with the Human Tissue Act (2004). All samples were obtained from consenting patients and consisted of both needle-core biopsies after radical prostatectomies and sliced-tissue biopsies from patients undergoing trans-urethral resection of the prostate (TURP).

All tissue sections have been obtained after an initial sample preparation done by another team at the Genito-Urinary Cancer Research Group at the Paterson Institute for Cancer Research. The paraffin embedded blocks of tissue are sliced using a microtome, with multiple serial sections being cut, with all sections having a thickness of 4 μm. The thin slices of tissue are then floated onto a window and left to dry under ambient conditions. Usually, one section is mounted on a window suitable for IR analysis, and a serial section is mounted onto glass. The serial section is dewaxed and stained using Haemotoxylin and Eosin (H&E).

The H&E stained sections were passed on to a collaborating experienced histopathologist at the Christie NHS Foundation Trust, with a sub-special interest in genito-urinary pathology, including participation in the UK national prostate histopathology external quality assurance (EQA) scheme. The histopathologist...
assessed the prostate tissue sections and where possible, marked the slides using a simple colour-coded system to indicate malignant areas of differing Gleason grade (green for Gleason 3, blue for Gleason 4, and black for Gleason 5 patterns).

At this point, all stained and sample sections were retrieved by the author for further preparation and analysis.

### 3.1.2 Window Substrate

The majority of the prostate cancer tissue sections analysed were mounted onto 25 mm x 75 mm MirrIR slides (Kevley Technologies, OH, USA) for transflection mode experiments. MirrIR slides are a cheap alternative to most other substrates for transflection mode FTIR analysis. The slides are made of glass with a low emissivity coating, consisting of multiple layers; most importantly including a very thin layer of silver-doped tin oxide which provides a very high mid-IR reflectance[65]. Another advantage of these slides is that they are transparent, allowing the possibility of visual microscopy.

One subset of samples used during the FTIR imaging protocol development however, was mounted onto CaF$_2$ slides for transmission mode imaging. The real disadvantage with this set of samples was the size of the windows. Circular, with a diameter of only 13 mm, in most cases they would not fit a whole standard section of prostate tissue. This caused some problems as the tissue areas that were marked for analysis on the serial sections, were often missing on the sample windows. Due to the developmental nature of the initial research, and the very long turnaround of obtaining samples (2-5 months), the work was still done using this sample set as a foundation, proof of principle study. Windows of a larger diameter have been used by the Gardner group in the past, but they are a lot thinner and very brittle. Custom sized windows are also available, but are expensive, and were unjustified for the nature of this study.
3.1.3 De-waxing Protocol

For the successful biospectroscopic analysis of paraffin embedded tissues, usually the paraffin is removed from the sample due to unwanted spectral contributions that may affect the overall chemical signatures of the biomolecules within the tissue. There is a lack of consensus however, with regard to a standard protocol for the deparaffinisation of paraffin embedded tissue sections, and many different methods and approaches have been used such as bathing samples in agents such as xylene, hexane, cyclohexane, and Citroclear [140]. To date, the most efficient method of paraffin removal (for use in a biospectroscopic study) is the protocol proposed by Fernandez et al. whereby paraffin embedded sections are deparaffinised by immersing them in hexane at 40°C with continuous stirring for 48 hours [85, 140, 141]. This method is extremely rigorous as it requires the constant washing of sections and refilling of solution over the two day period.

A successful, much less rigorous approach has been used by Gazi et al. involving Citroclear, a citric-extract based solution, to remove the wax from samples [36]. It is much more time efficient and less toxic than other methods and has been shown to adequately remove paraffin from FTIR spectra for the successful analysis and classification of data. Because of these reasons, the dewaxing protocol used for the work in this thesis is the one using Citroclear.

There have also been studies showing that paraffin contributions can be removed from FTIR spectral data post-acquisition, via digital dewaxing using an EMSC-based algorithm [130, 142].

To remove the paraffin from tissue sections, each mounted sample was washed on a rocking stage in Citroclear for 6 minutes, and then acetone (at 4°C) for a further 6 minutes before being air-dried for one hour under ambient conditions. Samples were then stored in a desiccator until required for data collection.
3.2 FTIR IMAGING PROTOCOL AND DEVELOPMENT

3.2.1 FTIR Microspectroscopy

The FTIR imaging system used consists of an Agilent Technologies 670-IR FTIR spectrometer with a KBr beamsplitter, and an Agilent Technologies 620-IR FTIR imaging microscope. The microscope is equipped with a liquid nitrogen cooled MCT detector for single-point spectroscopy, as well as a liquid nitrogen cooled 128x128 FPA MCT detector for FTIR imaging. The microscope has a fully automated joystick movable x-y stage, and a video camera to capture images of the sampling area. The 10x binoculars are used for viewing, coupled with a 15x IR condenser with a numerical aperture of 0.5 for FTIR experiments.

A low pass optical filter known as an undersampling ratio 4 (UDR4) filter is located in front of the face of the FPA detector. The purposes of this filter are to lower the sampling rate, to prevent aliasing by blocking non-IR sources of light, and greatly enhance the SNR during 128 x 128 FPA IR imaging.

A custom-designed pyrex purge box is attached to the front of the microscope and any further openings are sealed using foam and tape to minimise fluctuations in atmosphere during data collection. A purge line of compressed air is fed from a building supply to both the microscope and the spectrometer.

The microscope has a user-adjustable aperture, but this aperture must be fully open when FTIR imaging, to let as much IR radiation through as possible. Each FTIR image is made up of 16384 pixels, each corresponding to a separate spectrum, with the image covering an area of 700 \(\mu\text{m} \times 700 \mu\text{m}\) and each pixel equating to a 5.5 \(\mu\text{m} \times 5.5 \mu\text{m}\) sampling area.
The imaging system also has a mosaicing function, whereby the stage automatically moves and tiles together images according to how many are set and needed by the user.

### 3.2.2 Software Parameters

FTIR spectra and images are obtained using the Agilent Resolutions Pro. software package (v.5.0.0.700). Patching updates to the software were obtained over the course of the study.

Before the acquisition of imaging data, the FPA must first be cooled and calibrated. Using a hand held dewar, liquid nitrogen is steadily poured into the detector a number of times via means of a funnel, until the FPA has reached a stable temperature of 76K. The temperature can be monitored through the software. Once the detector is fully cooled, the calibration can be done.

To calibrate the FPA, the microscope should first be set into optical focus on the surface of a blank sample window, of the same type that samples to be measured are mounted on. Normally, detectors are connected to the electronics board of the spectrometer before they feed into the PC. The FPA detector however, bypasses any electronics and is hooked straight into the PC. This means it has no automatic means of registering the initial position of the moving mirror, needed to calculate the path difference to compute the FT calculations and process acquired data. For this reason, the user must find and register the centreburst position manually for the use of the FPA, by saving the centreburst of an interferogram calculated by another detector in the system (normally the in-built DTGS detector of the spectrometer). Once a centreburst has been saved, the user can complete calibration by setting the integration time of the FPA. The integration time essentially is a measure of the time that the shutter is open to collect the incoming photons. The aim is to optimise the SNR without saturating the detector. If the integration time is too high, the user will observe saturation effects in the FTIR
images; too low and the data quality and SNR will be reduced as the FPA has not been fully illuminated. The relationship between integration time and number of counts is shown in figure 3.1. The calibration is a non-uniformity correction, and results are shown with measures of high and low flux (in counts), and the number of out of range (OOR) pixels. If more than 6 pixels are displayed as OOR, it is an indication to the user that calibration is not optimised.

![Figure 3.1. The relationship between the number of counts and integration time. The optimised SNR is at the integration time at which the highest difference between low and high flux exists before the point of saturation.](image)

Once calibration is complete, a variety of parameters can be set in the software. The number of co-added scans, the wavenumber resolution and the wavenumber range can be adjusted. The wavenumber range is limited by the band-gap of the FPA detector, and the UDR4 filter. This makes the optimal usable wavenumber range for the system 900 cm\(^{-1}\) – 3800 cm\(^{-1}\). Depending on the experiment, different numbers of scans were used, ranging from 64 up to 512 in the testing phase of the instrument capabilities. The spectral resolution was set at either 4 cm\(^{-1}\) or 8 cm\(^{-1}\) during testing. Other modifiable software parameters were left at the defaults.
including a zero-filling factor of 2, and using the Blackman-Harris apodization function.

Once calibration and parameters are set, a background image of the blank, in-focus slide is taken to later ratio against the sample spectrum. The resulting image should be checked for a relatively even distribution of illumination and quality of spectra (SNR). The sample would then be moved into focus under the microscope, and the sample image acquisition would begin. The ratioed sample image should also then be checked for an even SNR across the image, and if any spectra give rise to atmospheric fluctuations such as water vapour and carbon dioxide, a new background image should be taken.

3.2.3 Early Imaging Results

At the beginning of the imaging work, there was no easy way of manipulating the spectral data after having acquired it other than from within the Agilent Resolutions Pro. software. Due to the novelty of the technique, there was no programming code readily available to import the imaging data into a programming environment such as MATLAB and to apply chemometrics on such a large scale. In time, this code was obtained and developed but not until after the initial image taking.

During the first imaging tests, there was also a distinct lack of samples. While awaiting fresh tissue sections that were to be cut onto CaF$_2$ windows, the only tissue samples available were those used in the author’s earlier MSc project. These consisted of a number of prostate cancer sections that were either sections taken from patient prostate biopsies where the cancer in the tissue was metastatic on presentation, and some where it was non-metastatic on presentation and stayed non-metastatic for at least five years after the biopsy was taken (according to patient history).
The images in figure 3.2 were novel at the time, showing some insight into what was being picked up and displayed by the FPA detector. Any optical images may seem slightly blurry towards the top. This was due to a manufacturing fault and the low resolution of the video camera. Another manufacturing fault means that the optical image is of a different size to the infrared image, and only represents part of the imaged area.

Figure 3.2. Optical images (a-c) and their respective corresponding infrared images (d-f) of areas of CaP tissue with metastatic Gleason grade 4 cancer.
The main problem with raw imaging data was quickly discovered, with the issue lying in the coloured representation of the image. The output is a false colour map, and it will always display the differences between image pixels (spectra) on a standard hot-cold (red-blue) scale. The image can be changed to display the differences between specific areas of the spectrum, and since tissue was being analysed the area under the amide I peak was used as the basis of the false colour map. This region was chosen as tissue is made up of many cells, and the proteins associated with the amide I band would give raw FTIR image more representative of the tissue structure, making it easier to compare with an optical image.

With this in mind, a piece of cancerous prostate tissue with marked areas of different malignancies that were close to one another (in distance) was used for a test to see what a raw FTIR image of this sort of area of tissue would look like, and whether it would show any differences between the Gleason grades. With the mosaicing function of the imaging spectrometer, a 2x2 tiled image was taken of a large section of tissue. An attempt to manually stitch together the optical image of the imaged area was also made and shown in figure 3.3.

Figure 3.3. (a) A 2x2 tiled FTIR image of a section of CaP tissue with a non-cancerous region, and regions of Gleason grade 3 and 4 cancer. (b) A manually stitched optical image of the imaged area of tissue, with regions of malignancy marked as assigned by the histopathologist.
Although the FTIR image seems to show a clear area of red and yellow different to the rest of the tissue which may be related to the high grade of cancer, this may also be something to do with a physical difference of the tissue, as the optical image clearly shows a difference in the unstained natural colour there as well. Before reading further into these results, the relevance of what is actually being portrayed by the raw image must be discussed.

A false colour map of spectral data, modified or not, cannot give a true representation of tissue morphology. Even though it may be useful to roughly view the distribution of a variety of absorption bands across the tissue, it is not based on a pure chemical signal from the tissue. Even when significant differences can be viewed between pixels across the image, they are not necessarily representative of any chemically significant differences.

Hypothetically, if an FTIR image was taken of a flat, single-element body with a homogeneous surface, the resulting image would still display a distribution of colour from red to blue, no matter how small the difference between spectra. Knowing that all areas of the sample used were chemically identical, any represented differences would solely be based on physical contributions such as light scattering, or even a difference in illumination of different areas of the 128x128 FPA. Biological samples are hardly ever completely homogeneous in thickness or in chemical make-up, so there are even more physical properties that contribute to the differences in spectral data, and the raw FTIR image. This highlights, more than ever, the need to obtain a pure chemical absorbance spectrum for each pixel of the FTIR image through the use of data pre-processing techniques.

3.2.4 FTIR Imaging Protocol Refinements

Over time, it became apparent that cooling the FPA detector multiple times per day was important. After about four hours, the temperature of the detector started
increasing above 76K, so it had to be refilled with liquid nitrogen to optimally stabilise again. The temperature changes were noticed by the worsening of the quality of spectral data. Topping up the liquid nitrogen three times a day was not practical, so the FPA was removed from the system and pumped out using a two stage rotary-turbo pump over the course of three days. When reinstalled, the detector held its temperature for longer and much less top up was needed on any given day, if at all.

In order to optimise the SNR and illumination of the FPA, the integration time was varied between 0.022 ms and 0.042 ms. If the detector was being saturated, artefacts would appear in the image originating from the centre in a cloud-like group of spots which would get bigger with a higher saturation. The lowest integration time where these saturation effects were not observed was 0.035 ms for transflection imaging, and this is the integration time that was used for the remainder of the experiments.

In the author’s previous work[60], before the lab was housed with an imaging spectrometer, single point FTIR spectra of prostate tissue were taken with 512 scans for a background spectrum and 256 scans for a sample spectrum. When the Agilent imaging microspectrometer was installed, the same parameters were initially used. However, after reducing the number of co-added scans when imaging by half, no visible detriment to SNR was observed. Furthermore, halving the number of scans again to 64 only slightly visibly decreased the SNR of the spectrum. In order to achieve as high SNR as possible without increasing data acquisition time too much, sample images were taken using 128 scans and background images with 256 scans. Obtaining an image with 128 scans and a spectral resolution of 4 cm⁻¹ took about four minutes. A background image was taken before every three sample images or whenever a different sample was used.
3.3 OPTIMISED FTIR IMAGING RESULTS

With the ability of importing the FTIR image files directly into MATLAB, they could be pre-processed and analysed using further code. The set of images analysed during this section were newly obtained prostate cancer sections at the time, on CaF$_2$. However, the aforementioned problem in section 3.1.2 affected the progress of this study. Without having the area of tissue corresponding to the marked sections of many different prostate samples, it was decided to image the low number of marked tissue sections anyway, and to optimise a data collection pre-processing protocol that would represent an accurate histology of the tissue when compared to a stained section.

3.3.1 Initial HCA Results

The very first attempts at obtaining meaningful results were by applying HCA and evaluating the technique for optimised use with FTIR imaging data. Before any pre-processing was applied to any imaging data, HCA was done on the raw data sets to see if any issues would arise from using the technique. Two CaP tissue biopsy samples were originally analysed. One focused area of tissue was highly cancerous as shown in figure 3.4, and the other was of more of a glandular structure with a lesser Gleason grade in figure 3.5.
Figure 3.4. Imaging analysis of an area of CaP tissue with Gleason pattern 4 cancer. (a) Optical image of analysed tissue area. (b) IR false-colour image of the same area. (c-f) HCA of the imaged area of tissue with 4-7 clusters respectively.
Figure 3.5. Imaging analysis of an area of CaP tissue with Gleason pattern 3 cancer. (a) Optical image of analysed tissue area. (b) IR false-colour image of the same area. (c-f) HCA of the imaged area of tissue with 4-7 clusters respectively.

It was immediately apparent from the HCA that differences between clusters were generally unclear due to the lack of a logical colour assignment scheme for the clusters. One of the initial challenges was to restructure the colour-coding and assignment part of the clustering code to something more clear and consistent.

Looking past the colour assignment, it is also apparent that cluster assignment is not very sensible. The optical images of the tissue show clear holes, or glands as
they may be, where there should be little to no chemical information present due to a lack of cells or biological material, and all such regions should be assigned to one clear cluster. In practice however, increasing the number of clusters in an analysis would in fact further separate these ‘empty’ regions of tissue into more clusters, meaning that the differences between spectra in those areas is the greatest. It becomes clear here that there are unwanted chemical contributions within the glandular and holey areas, and some means to ensure the assignment of all such spectra to a single cluster must be devised.

Although it is not easy to interpret anything from these raw data images, they have in fact given information about the issues one will face in upcoming analyses, and again calls for the pre-processing of data.

3.3.2 Refined HCA Results

At this point, the RMieS-EMSC algorithm had been developed for in-house testing, and its application on FTIR images had yet to be investigated. This original data processing of imaging data took an extremely long time, up to 4 full days per image. The computer used for the process had limited memory and power, and the computations and iterative process were very demanding. With severe time constraints on analysing data, a limited number of samples were looked at in this way. The default matrigel spectrum was used as a reference spectrum for the correction and the process was done using 20 iterations (under advice from the writer of the algorithm).

To account for any systematic variations in spectra, vector normalisation was always applied in conjunction with any other pre-processing method attempts. The culmination of the pre-processing work was presented in poster form at the SPEC 2010 international spectroscopy conference in Manchester, UK in June 2010 titled “A comparison of pre-processing methods on FTIR images of prostate cancer
tissue”. The developed process will be outlined first, and the results shall then be presented.

Before anything was done with the data sets, steps were made to fix the issue of certain areas of images (corresponding to places with a lack of tissue) from separating into many clusters. The best way of assuring this was to remove the ‘bad’ spectra from any form of analysis in the first place, and project them back into an image as one cluster at the final stage. A quality test was implemented on the data based on a threshold using a minimum and maximum accepted absorbance value for each spectrum within the measured IR range. The minimum to maximum threshold value used for this experiment was 0.3 A.U. Any spectrum failing to meet this threshold condition was removed from the data set before pre-processing.

The problem with the lack of a logical colour-coding system for clusters was also solved, by the use of some extra MATLAB code which reorganises the numbering system of the clusters to be the same every time, always assigning the same colours to each desired cluster. With the desired clustering setup optimised, imaging data from a section of BPH tissue and a section of CaP tissue were run through HCA after using a variety of combinations of pre-processing techniques. The resulting images are shown in figures 3.6 and 3.7.

This work is also unique in the sense that rather than using serial tissue sections for H&E comparison, the analysed tissue section was itself H&E stained after data collection so an optical image would be clear and identical to the FTIR sampled area. The idea of staining the same section of tissue overcame any problems normally noticed in using serial sections for comparison, namely the morphing and altering of glandular structures in places within the tissue. However, once the tissue had been stained it voided any recollection of FTIR data.
Figure 3.6. (a) Optical image of the H&E stained area of imaged BPH tissue. (b) Total intensity infrared image of sampled area. (c-f) HCA images of the tissue after (c) vector normalising, (d) first derivative followed by vector normalisation, (e) second derivative followed by vector normalisation, and (f) RMieS-EMSC to 20 iterations followed by vector normalisation.
Figure 3.7. (a) Optical image of the H&E stained area of imaged CaP tissue. (b) Total intensity infrared image of sampled area. (c-f) HCA images of the tissue with 6 clusters after (c) vector normalising, (d) first derivative followed by vector normalisation, (e) second derivative followed by vector normalisation, and (f) RMieS-EMSC to 20 iterations followed by vector normalisation.

With the HCA images now having a much clearer clustering representation of the tissue, to gain a better insight into whether the technique was working well, rather than any spectral interpretation that would not necessarily explain anything about how good the clustering was, the results were taken to the histopathologist who had no prior knowledge on pre-processing techniques. His advice was useful in determining whether any of the clustering patterns made logical sense and whether any method seemed better than another at describing tissue architecture.
3.3.3 Consultation with Histopathologist

Any views expressed in this subsection were agreed and advised by the histopathologist.

Figures 3.6(c) and 3.7(c) show the vector normalised analyses, and on inspection the clustering in both looks quite similar to the false colour patterns from the total intensity images in figures 3.6(b) and 3.7(b). This may mean that as a standalone technique, vector normalisation may not fully remove physical contributions to the data, as the other HCA images look more refined and representative of the true tissue architecture. The second derivative data shown in figures 3.6(e) and 3.7(e) are the most alike the tissue morphology, clearly showing different elements of the tissue.

In the BPH tissue in figure 3.6(e), the white cluster is the gland lumens, the dark purple cluster is the epithelium, and the pink, orange and green clusters make up the stroma. The orange cluster seems to represent the elements of stroma with a higher concentration of smooth muscle, whereas the light and dark green clusters indicate areas of stroma with less smooth muscle and more collagen. In the CaP tissue analysed in figure 3.7(e), the white cluster once again corresponds to the gland lumens, the pink and orange clusters make up the glandular epithelium, the light and dark green clusters comprise the stroma, and the edges of the biopsy are shown by the dark purple cluster.

Although this analysis shows that the second derivative data may look most like the tissue morphology, it still does not necessarily represent the tissue on a purely chemical level. After undergoing RMieS-EMSC, the data should show, or most closely show, the pure chemistry of the tissue, even if it may not be accurately represented by what is observed from the tissue morphology.
3.3.4 Further Development of Analysis and Return to the Histopathologist

Findings and discussions during the attended international spectroscopy conference SPEC2010, made it clear that Mie scattering contributions to spectral data must be corrected for in all FTIR data of biological samples to make valid assumptions and conclusions from analyses. The presented data and conclusions made in subsections 3.3.2 and 3.3.3 show that using RMieS-EMSC was not the best matchup with the tissue morphologically, however, the technique must be used in some form. Since those results were obtained, the algorithm had been further developed and now used a better matrigel reference spectrum (ie. one that more accurately resembles a zero-scattering spectral representation of biological matter). As a final experiment with pre-processing techniques, all previous tried pre-processing techniques used were applied in combination with the improved correction algorithm before being taken back to the histopathologist for a blind interpretation of the results.

For the reanalysis of the prostate tissue, a smaller wavenumber region of the spectral data was used, specifically the fingerprint region between 900 and 1800 cm\(^{-1}\), as it is not only the most chemically diagnostic region of an FTIR spectrum but doing so would reduce the size of data sets and allow for a quicker analysis. Please note that of the following images, figures 3.8 and 3.9 are analyses of the same tissue, as are figures 3.10 and 3.11. The only reason they are separated is due to size and spacing over pages.
Figure 3.8. 5 and 6 clusters HCA images of a CaP tissue section after using RMieS-EMSC on the data set and also (a,b) first derivative followed by vector normalisation, (c,d) second derivative followed by vector normalisation, (e,f) vector normalisation.
Figure 3.9. 5 and 6 clusters HCA images of a CaP tissue section after using RMieS-EMSC on the data set and also (a,b) vector normalisation followed by first derivative, (c,d) vector normalisation followed by second derivative.
Figure 3.10. 5 and 6 clusters HCA images of another CaP tissue section after using RMieS-EMSC on the data set and also (a,b) first derivative followed by vector normalisation, (c,d) second derivative followed by vector normalisation, (e,f) vector normalisation.
Figure 3.11. 5 and 6 clusters HCA images of another CaP tissue section after using RMieS-EMSC on the data set and also (a,b) vector normalisation followed by first derivative, (c,d) vector normalisation followed by second derivative.

It can be overwhelming at first glance to anybody who tries to immediately compare and make sense of all these images. This is why the images were printed out on a large scale in pairs, and methodically inspected with the histopathologist. Note that in this slightly different cluster colour scheme, in all cases, the orange cluster is the quality tested one showing the areas of no tissue. Simultaneously to looking at the HCA images, the H&E stained tissue was placed under the multi-binocular microscope in the histopathologist’s microscopy suite for direct comparison. Once again, the images were shown blindly, without mentioning which pre-processing methods had been used.
Initial reactions from the histopathologist were interesting, as in all cases, he stated that the images showing the HCA after RMieS-EMSC with vector normalisation followed by second derivative were the most likely representation of the tissue morphology (Figures 3.9 and 3.11 (c) and (d)).

The different cluster separations were discussed, and it was judged difficult to understand what was happening with the separation in figures 3.8 and 3.10 (e) and (f), specifically the images corresponding to RMieS-EMSC with just vector normalisation. This links in with the results obtained in subsection 3.3.2 and 3.3.3, having obtained the same conclusion there.

If particular focus is made on the images from the clearer CaP tissue section in figures 3.8 and 3.9, one can observe that the HCA images using vector normalisation followed by the first derivative technique (figures 3.8 and 3.9 (a) and (b)) tend to cluster tissue edge spectra separately from the main body of stromal tissue. This is not the case in the second derivative images in figures 3.8 and 3.9 (c) and (d). There is no reason why the chemistry behind these spectra should be any different from the main body of stroma which is why the second derivative images are truer to tissue architecture in that regard.

Both second derivative HCA images (figures 3.8 and 3.9 (c) and (d)) separate the smooth muscle tissue from the rest of the tissue very well. The main difference between doing vector normalisation before or after derivatising is how the methods cluster the stroma of the tissue. Comparing figure 3.8 (d) to figure 3.9 (d), the dark green cluster in both represents the stromal tissue. In figure 3.8 (d) the blue cluster is the smooth muscle tissue, while in figure 3.9 (d) it is shown by the pink cluster. The main difference between the images can be observed if closer attention is paid to the pink cluster in 3.8 (d). Matching this same clustered area to the image in figure 3.9 (d), one can see it fits in with the dark green cluster that is categorised with the rest of the stroma. It made more sense to the histopathologist that this
area should cluster with the stroma, not separately (from a morphological point of view).

The light green cluster in figure 3.9 (d) is somewhat comparable to the purple cluster in figure 3.8 (d) representing the epithelial cells around the glands. There seems to be more clustering separation on a whole around the glandular regions of tissue in figure 3.9 (d), which is where one may expect to see a difference in chemistry, possibly due to a lack of basal cells or additional nucleoli. This agrees with the histopathologist's assessment of the HCA images.
3.5 CONCLUSIONS

After this study, it can be argued that performing RMieS-EMSC followed by vector normalisation and a second derivative is the best way to pre-process FTIR-imaging data sets, if using real tissue morphology as a guide to how representative the clustering of HCA data is. It is also key to note that although the discussed analyses may show a good correspondence with the tissue morphology, the actual chemical differences between parts of the tissue may not correspond in the same way. This is an obstacle that is very difficult to overcome, as there is no understanding on how biological tissue should separate chemically. All that can be done is to set a standard that future work will be based on, backed up by this imaging study.
4. PROSTATE CANCER PROGNOSIS STUDY

4.1 FOUNDATIONS OF STUDY

4.1.1 Introduction

The diagnosis of prostate cancer has been shown to be possible through the means of biospectroscopic analysis, based on the guidelines set for the grading of cancer by the current gold standard of the Gleason grading system. Although the grading system is largely subjective, a pathologist should have little trouble in distinguishing a higher grade of cancer from a low grade cancer. However, under a microscope, metastatic cancer cells look identical to cells of the original cancer. This means that a histopathologist has no way of determining whether a cancer has spread to other organs or bones without doing further tests and scans to investigate those surrounding areas.

The specific aim of this study is to investigate whether it is possible to differentiate metastatic epithelial cells from non-metastatic epithelial cells purely on their chemistry using spectral data. Positive results in this regard would be ground-breaking, highlighting another advantage of using FTIR spectrometry for the analysis of prostate tissue. The potential for future work in the area would be big, being able to achieve diagnostic results from information obtained solely from prostate biopsies that a pathologist would not.

4.1.2 Foundation Work

The author’s original work in this area resulted in an MSc thesis entitled “Development of an objective spectroscopic based grading system for prostate cancer biopsies”[60]. Therein, single-point FTIR spectral data was obtained from a set of prostate cancer tissue biopsies, and multivariate analysis was employed to attempt to differentiate patients with metastatic prostate cancer, and patients with
non-metastatic cancer that stayed non-metastatic for at least 5 years after the initial biopsy.

The data analysis resulted in DFA plots of the pre-processed spectral data, for which 80% of the data was used as a training set and 20% was used as a blind testing set. 536 spectra were used from Gleason grade 3, 4 and 5 zones in 41 CaP biopsy samples where in 18 patients there was evidence of metastasis and in 23 the cancer was non-metastatic. Some of the main results are shown in figures 4.1 and 4.2. In figure 4.1 the CO₂ region of the spectrum was removed (1800-2700 cm⁻¹), and in figure 4.2 only the fingerprint region of the spectrum was used for analysis (800-1800 cm⁻¹).

![DFA projection showing 90% confidence ellipses for non-metastatic (1) and metastatic (2) sample data using a blind test set (blue) created from randomly selected patients. (Analysis is run over entire CO₂ region removed FTIR spectrum).](image)

Figure 4.1. DFA projection showing 90% confidence ellipses for non-metastatic (1) and metastatic (2) sample data using a blind test set (blue) created from randomly selected patients. (Analysis is run over entire CO₂ region removed FTIR spectrum).
Figure 4.2. DFA projection showing 90% confidence ellipses for non-metastatic (1) and metastatic (2) sample data using a blind test set (blue) created from randomly selected patients. (Analysis is run over the fingerprint region of the FTIR spectrum).

Sensitivities and specificities of data analyses can be useful in understanding both the strength of a statistical model and the relevance of the results obtained. Sensitivity is the probability of obtaining a positive test result from patients that have the disease. Specificity is the probability of a negative test result among patients without the disease. Ideally, the highest possible values for both sensitivity and specificity should be targeted. It can be argued, however, that a higher sensitivity is clinically more important than specificity, as it is much worse to completely miss a diagnosis than to diagnose a patient by accident.

For the data projection in figure 4.1, the respective sensitivities and specificities for the non-metastatic cancerous tissue are 58.3% and 75%, and for the metastatic cancerous tissue are 100% and 60.4%. The results in figure 4.2 show an improved sensitivity of the non-metastatic prediction which is 70.8%, but the non-metastatic specificity decreases to 53.1%. The respective sensitivities and specificities of the
metastatic cancer are 100% and 64.6%. Note that any apparent separation of data in the y-direction is irrelevant due to the scale of the second discriminant function. Both models suggest some degree of separation between the classes of tissue, being able to identify metastatic cancer extremely well. They are however, quite weak at specifying non-metastatic cancerous tissue from metastatic cancerous tissue in patients without metastatic cancer.

These results showed the potential for the further refined investigation of metastatic and non-metastatic cancerous tissue samples.
In order to improve on the results in section 4.1, a few key areas can be refined. Firstly, an assessment should be done of whether the analysis and pre-processing are being done correctly. A quality control of spectral data should be done, to reduce the chance of outlying data points. Knowledge of the origin of the samples would be useful, to see whether patient and biopsy information is similar, and a larger overall sample set would be useful.

Originally, plans were made to reanalyse the MSc data using an accepted method of choosing how many principal components to use in the DFA. Previously, the number of principal components used for analysis was 20. This equated to the principal component that explained 99.9% of the total data variance. This method was used based on other research that had proved successful using similar values[81]. A small study into methods of choosing principal components would be done anyway.

4.2.1 Using the Correct Number of Principal Components

A common ongoing debate with DFA techniques is how many principal components should be used in the process. If too few PCs are used, no clusters relating to the data groups will form. If too many are used, oversampling may occur, manipulating the data so they separate into highly packed artificial clusters.

There are a number of proposed methods to choose the correct number of principal components of a data set. Defernez et al. concern themselves with the importance of the overfitting of data[143]. Their study states how replicate data are by no means independent observations, and that any observed separation in the data clustering may be essentially worthless because of this fact, highlighting the importance of using spectra from separate samples for training data, and not just random spectra from the whole data set. The study also states how reasonable
separation and clustering may be achieved by noise in spectra, giving irrelevant results.

Overfitting of data should be strongly suspected when the dimensionality of the data exceeds \( \frac{(n-g)}{3} \), where \( n \) is the number of independent observations and \( g \) is the number of groups. Any given model must be questioned whether it can generalise and correctly classify observations other than those with which it was defined[143].

Another commonly employed method of deciding how many principal components to use is by looking at the cumulative percent variance (CPV) of a data set. The CPV is a measure of the percent variance captured by the principal components of a data set. The idea is to account for as much of the variance as possible while retaining as few PCs as possible. The desired CPV is completely subjective although normally ranges from 90-99%, with most of the literature quoting to aim for 95%[144].

Yet another method that can be used is by referencing a scree plot. It is an empirical procedure to select the number of PCs using the eigenvalues or percent variance as a basis. Residual percent variance (RPV) or the eigenvalues are plotted against the number of principal components, and the ‘elbow’ of the resulting curve is used to assign the number of PCs to use. The implementation of this method is relatively easy, but in some cases it can be difficult to find the exact point of the ‘elbow’ of the curve if it decreases smoothly.

The Kaiser rule is a popular approach that uses the average eigenvalue to determine the number of PCs to use. It accepts all eigenvalues with values above the average eigenvalue and rejects those below the average. A principal component contributing less than an ‘average’ variable is insignificant. For correlation-based PCA, the average eigenvalue is always equal to 1.
Cross-validation is again a popular statistical criterion to choose the number of principal components. The basis of the technique is to estimate the values of data that is removed from a model, and compare these estimates to the actual values of the data. The predicted error sum of squares (PRESS) is calculated from the predicted and actual values of the deleted data. The variance of the reconstructed error (VRE) method was developed to select the number of PCs based on the best reconstruction of the variables. The minimum found in the VRE calculation directly determines the number of PCs to use.

4.2.2 Results and Reanalysis

The various discussed methods in section 4.2.1 were applied to the data from the authors MSc, in order to obtain the correct number of PCs to use for the data analysis.

Using the condition suggested by Defernez et al., after analysing CaP biopsy tissue from 41 independent patients in two groups (metastatic and non-metastatic), the total number of PCs should not exceed the maximum of $\frac{41-2}{3} = 13$ PCs. The application of the other techniques are shown in figure 4.3.
Aiming for a 95% CPV, results in figure 4.3(a) show that 7 PCs should be used. The scree plot in figure 4.3(b) shows the smoothing out of the curve to occur between 5 and 8 PCs. This is a good indication of the ambiguity of this technique. It is useful only to use this technique in unison with another. Rejecting all eigenvalues below the value of 1, according to the Kaiser rule in figure 4.3(c), 8 PCs should be used. These methods all suggest similar values of principal components to use, and they all adhere to the Defernez et al. condition and do not suggest exceeding 13 PCs. In light of these new calculations and research, the previous DFA plots of the MSc data may be overfitting clusters due to too high a number of principal components used, although there is no real evidence of highly packed artificial clusters. The analysis would however, be redone using a more appropriate number of 7 PCs. Figures 4.4 and 4.5 show the DFA plots reanalysing the same data as used in the original analysis.
Figure 4.4. DFA projection showing 90% confidence ellipses for non-metastatic (1) and metastatic (2) sample data using a blind test set (blue) created from randomly selected patients. (Analysis is run over entire CO$_2$ region removed FTIR spectrum using 7 PCs).

Figure 4.5. DFA projection showing 90% confidence ellipses for non-metastatic (1) and metastatic (2) sample data using a blind test set (blue) created from randomly selected patients. (Analysis is run over the fingerprint region of the FTIR spectrum using 7 PCs).
The reader can already see that the separation of tissue class is worse than the original results. The clusters have a very strong overlap, and although sensitivities of the data may still be high, the specificities are much lower. The non-metastatic cancer data in figure 4.4 has a sensitivity of 80% and a specificity of only 8.3%, while the metastatic cancer can be explained with an equally low sensitivity of 76.1% and a specificity of 6.5%. Using the fingerprint region only analysis, the respective sensitivities and specificities of the non-metastatic cancer data are 86% and 3.5%, and of the metastatic cancer data are 80.9% and 10.6%.

4.2.3 Outlook

The results after reanalysing the data show very low sensitivities, and they may be attributed to a couple of things. Firstly, having done the DFA analysis with a more appropriate number of principal components, it can be discounted as a reason for the bad results, even if the separation was worse than before. The blame could lie either with the quality of the spectral data, or with the method of analysis. DFA may not be the best technique for attempting to achieve separation, as the data is colinear and there are only two different groups to be classified. It could also just mean that the model does not work reliably enough to classify metastatic and non-metastatic cancer data, but this could also be to do with the data collection of the wrong areas of tissue.

It was concluded that the FTIR spectra were not of a high standard enough to easily achieve separation using multivariate analysis, no matter how much pre-processing was applied. The raw spectra and the pre-processed counterparts with the \( \text{CO}_2 \) wavenumber region removed are shown in figure 4.6.
It is crucial to mention that all the single-point FTIR work and the reanalysis from the author’s MSc was done before delivery of the FTIR-imaging system, or the development of the scattering correction algorithm. Although it was originally planned to retake all the spectral data, taking care to collect the data from the correctly marked regions of epithelial tissue as accurately as possible for another attempt at improving the analysis, the FTIR-imaging system had arrived.

Although focus initially shifted to the development of the novel technique discussed in chapter 3, the metastatic versus non-metastatic cancer study was planned on a larger scale using optimised methods of FTIR-imaging and pre-processing.

However, with FTIR imaging capabilities now available, the author decided to retake data from the original metastatic and non-metastatic cancer tissue samples in image form prior to the large scale study with the aim of determining any problems to be expected in the future study.
4.3 FTIR IMAGING OF ORIGINAL SAMPLE SET

4.3.1 Data Collection

After being transported overseas, only 34 of the original 41 samples were able to be analysed. 15 of these samples were from patients exhibiting metastatic cancer, and 19 were non-metastatic on presentation.

Between one and three sample images were taken from each sample, with a single background image taken beforehand. Experimental parameters used were the same as those outlined in the previous work in section 3.2.4. 256 scans were taken for the background, and 128 scans were taken for each sample image. The spectral resolution was 4 cm\(^{-1}\) and the wavenumber range used was 900 - 3800 cm\(^{-1}\). Images were all taken using the 128 x 128 FPA coupled with the Agilent 670-IR spectrometer. The images taken resulted in a raw data set comprising of 90 imaging files corresponding to 38 metastatic cancer associated images and 52 non-metastatic cancer associated images.

4.3.2 Imaging Results and Analysis

A quality test was applied to the data sets based on a minimum-maximum threshold of the amide I wavenumber region, rejecting any spectra that failed to meet the required absorbance contribution. This was done to remove the bad spectra from areas of the images that contained no tissue.

Pre-processing was then done on the data based on conclusions from the study in section 3.3.3 and 3.5, so vector normalisation and second derivative processing were used. Due to time constraints, and the low amount of scattering contributions in the raw data, RMieS-EMSC was not used on the large data set. Also, because of the drastic reduction in computation time, K-means clustering was used instead of HCA.
K-means was applied to the 90 images using 6 clusters with the aim that the epithelium of the tissue would cluster together in one of the coloured groups. The clusters that most closely resembled the areas of epithelium when compared to the optical images were isolated, and the spectra associated with those clusters were extracted. This process was extremely subjective and quite difficult, but it must be noted that for this particular study it was acceptable to estimate certain parameters and be less thorough as it was being used as a rough, preliminary indicator for future work. Some examples of the cluster isolation process are shown in figure 4.7.

Figure 4.7. (a) and (c) show low resolution optical images of the imaged prostate tissue sections. (b) and (d) are the corresponding K-means analyses. The epithelium associated coloured clusters were red/brown for (b) and yellow/orange for (d).
The epithelial spectra from each image were averaged so that PCA could be done, resulting in 78 average spectra. The loss of spectra is because for 12 of the images, it was incredibly difficult to isolate the epithelial tissue associated cluster, so these images were removed from the analysis to reduce observer error. Spectra were also extracted from the entire quality controlled areas of the tissue in each image, and subsequently averaged for use as another data set. PCA was then done on the two separate data sets; firstly on the spectra from the whole tissue, and secondly on the epithelial spectra only. The PCA results are shown in figures 4.8 and 4.9.

Figure 4.8. PCA scatter plots of the average spectra from each whole FTIR image data set, where the red dots correspond to the metastatic cancer data and the black dots correspond to the non-metastatic cancer data. (a) PC1 versus PC2. (b) PC1 versus PC3. (c) PC2 versus PC3.
Figure 4.9. PCA scatter plots of the average spectra from the epithelial regions of tissue, where the red dots correspond to the metastatic cancer data and the black dots correspond to the non-metastatic cancer data. (a) PC1 versus PC2. (b) PC1 versus PC3. (c) PC2 versus PC3.

The results show different clustering patterns, depending on which spectra were used. Figure 4.8 shows hints of separation between average metastatic and non-metastatic cancer spectra of the whole imaged tissue. There are larger clustered areas of metastatic cancer spectra towards the left hand side, and non-metastatic cancer spectra on the right, even though there is a strong overlap. However, the separation between the clusters is much more refined and conclusive when just the average spectra of the epithelial regions of the tissue are used in figure 4.9. The separation was best when plotting PC1 against PC2 in figure 4.9(a) with the majority of the non-metastatic tissue spectra in a tightly packed area on the right hand side of the PCA plot. The error and overlap of the clusters can most likely be attributed to the rough estimation and extraction of the epithelial spectra, but the results
show that this process has worked to some extent, improving on the results in figure 4.8.

These results highlight the diagnostic and prognostic potential of using only epithelial spectra in the future study to compare metastatic cancer spectra to non-metastatic cancer spectra, but a clear, accurate method of isolating and extracting these spectra must be devised. Furthering this study by refining results using new samples from a controlled, similar and consistent patient group will also be important.
5. LARGE-SCALE METASTATIC AND NON-METASTATIC STUDY

5.1 PROJECT OUTLINE

This project was planned to be as accurate and consistent as possible in the collection, treatment and the analysis of data. Using optimised FTIR-imaging techniques, images were taken in the same manner, to a high standard from every sample. In order to analyse the correct areas of tissue, focus was to be on the investigation of the epithelial tissue areas. With the immunohistochemical staining of tissue to highlight these regions, an accurate comparison of stained image to FTIR image could be made, and associated spectra from the desired area were extracted. Spectra were quality tested and pre-processed using the RMieS-EMSC algorithm in combination with other pre-processing methods. Finally a model was built using PCA and ANNs.

5.1.1 Importance of Sample Set Consistency

In the planning of this large-scale study, it was quickly determined that one of the most important aspects of the research was to obtain samples that were as similar in biopsy and patient characteristics as possible. In the preliminary study presented in section 4.1, the only condition that was set for the samples was that there would be about an equal set of non-metastatic and metastatic cancer tissue samples that would all have a Gleason score of 6-7. In practice, although a good number of samples was obtained with records agreeing with the criteria, upon closer inspection of the samples, the histopathologist regraded some of the samples to include Gleason pattern 5 areas of tissue. While this was not ideal, the samples were still used in the study, but it did highlight the problems of subjective observer variability in the grading of tissue.

For this new project, the sample set was attempted to be as consistent as possible. The ages of the patients were to be within 10 years of another, the year the biopsy was taken was to be within 5 years of each other, and all biopsies obtained were to
be from as few tissue banks as possible. All biopsy tissue would also be classified with a Gleason score of 6-7.

5.1.2 Samples and Sample Preparation

The samples obtained met the conditions set out as well as possible. Tissue specimens, as in the previous study, were obtained as paraffin embedded blocks taken from tissue banks in both the Christie NHS Foundation Trust and Salford Royal NHS Foundation Trust Hospitals. All tissue samples were collected with full ethical committee approval and are all in accordance with the Human Tissue Act (2004). All samples were obtained from consenting patients and consisted of both needle-core biopsies after radical prostatectomies and sliced-tissue biopsies from patients undergoing trans-urethral resection of the prostate (TURP).

The tissue blocks were each sliced into five serial sections; the first was mounted on glass for H&E staining, the second and fourth slices were mounted onto MirrIR, the third and fifth slices were also mounted onto glass. All the tissue sections were sliced and mounted with a thickness of 4 μm, by the team at the Genito-Urinary Cancer Research Group at the Paterson Institute for Cancer Research. The reasons for this exact protocol were for accuracy and for backup reasons. Sections that are next to one another have the highest degree of similarity. For this reason, having a glass mounted sample for H&E staining and one for IHC staining either side of a MirrIR mounted sample for data collection would arguably would be the best way to achieve the most similarity between the sections. The fourth and fifth sections were sliced just in case something went wrong with the other sections and data had to be recollected or restained.

However, in order to attempt an even higher degree of accuracy when pinpointing the epithelium in the tissue, the MirrIR mounted sample that FTIR-imaging data was collected from was itself IHC stained (post data collection), so that the stained tissue morphology would be identical to that visible in the FTIR images.
The H&E stained sections were passed on to the histopathologist at the Christie NHS Foundation Trust, where he assessed the prostate tissue sections, marking the slides with the familiar colour-coding system to indicate the malignant areas of differing Gleason grade (green for Gleason 3, blue for Gleason 4, and black for Gleason 5 patterns).

All stained and sample sections (except the ones to be IHC stained post data collection) were then retrieved by the author for further preparation and analysis.

All samples were dewaxed using the Citroclear paraffin removal method, as in the previous study, bathing them in Citroclear on a rocking stage for 6 minutes, followed by washing them in acetone (at 4°C) for a further 6 minutes before being air-dried for one hour under ambient conditions. Samples were then stored in a desiccator until required.

5.1.3 Experimental Protocol and Parameters to Use

To achieve the best IHC staining results of the tissue, the staining using an anti-pan cytokeratin monoclonal antibody was performed as quickly as possible after the samples had been dewaxed. This required a quick turnaround, taking all the required FTIR-imaging data as quickly as possible, before sending the samples back to the lab for staining. The full process from dewaxing to staining was always accomplished within a two week window.

The data collection protocol was also set out in detail to ensure as similar conditions as possible over the collection period. There was an air conditioning unit above the spectrometer in the lab that was not always regulated. The first action was to close the flap of the A/C unit before data collection, so the atmospheric conditions around the lab would be similar. While working, the door to the lab was also kept closed as often as possible. The FPA was always cooled the same way with the same
amount of liquid nitrogen every time, and the detector was left to settle for the same time of at least 15 minutes every time it was filled.

In order to reference the exact position of the tissue being analysed, an England finder was used. An England finder graticule is a glass slide marked in such a way that a referenced position can be directly read relative to the locating edges[145]. Each MirrIR slide was positioned directly over the England finder (with the frosted corner of the slide in the same orientation for each sample), and stuck on as accurately as possible with a small amount of masking tape. The tape was trimmed to ensure there was no excess on the bottom of the England finder that would cause unevenness. The mounted England finder and sample were then moved onto the stage of the IR microscope and corners of the slides were pushed securely back into the positional sample holder pins. A blank, clean MirrIR slide was pushed up next to the sample, and the microscope was focused onto the blank surface of the slide before the purge box was closed.

After five minutes of letting the atmosphere within the closed purge box stabilise, the FPA was calibrated using the described method in section 3.2.2. The interferogram of the DTGS detector within the spectrometer was also used to find the centreburst, and the peak-to-peak voltage of the interferogram was logged. Ensuring the FPA temperature was 77K, the integration time was always set as 0.035 ms and calibration was completed. If the number of out of range pixels equated to anything more than 6, the whole process would be restarted.

Before the data collection, parameters were set to take 64 co-added scans for a sample image, 128 scans for a background image, the spectral resolution was set at 4 cm\(^{-1}\), the wavenumber range was set between 900-3800 cm\(^{-1}\), and the other parameters were left at their defaults.
One background image was to be taken for every three sample images, aiming for a total of at least six sample images from each patient. A file-naming protocol was devised for clarity. Each patient with non-metastatic cancer would follow the numbering scheme 101, 102, 103, etc. The metastatic cancer patient cases would follow the convention 201, 202, 203, etc. Each image taken from each patient was labelled with the letters a-f in most cases, where a-d were cancerous areas of the tissue and e-f were non-cancerous regions. Sometimes there were more than six images taken, and for some samples there were no easily identifiable regions without cancer.

Once the background image was taken, it was checked for quality consistency and the stage position was moved to a desired area of the sample. Once in focus, making sure the position associated correctly with the histopathologist’s markings, and that there was plenty of epithelium visible in the sampling area, a sample image was taken. Background images took roughly 8 minutes, and sample images took roughly 4 minutes.

Once the sample image had been collected, an optical image of the sampled area was taken using the microscope built-in camera. Without moving the stage in the x-y direction, the sample was defocused down to the surface of the England finder. An optical image of the England finder position was then taken, and the coordinate was logged along with all the other information in the log book. This naming convention would follow the suit of the letter, number, and quadrant of the coordinate, eg. F36-2.

The stage would then once again be raised to the sample surface, and the sampling area would be moved to a new location for imaging. This process was always followed in the same manner, unless for example in some cases where seven images were taken from a patient, a background image followed by three sample images would be taken before another background image and the final 4 sample
images. The spectral quality of the final images was always checked in case a new background needed to be taken.

When a new sample was to be imaged, the purge box was opened and the samples were exchanged and remounted onto the England finder in the aforementioned manner.

Once data had been collected and logged, a collaborating colleague took the samples to be IHC stained. Once they had been retrieved again, they were taken to a microscopy suite to obtain high resolution images of the stained tissue. The high resolution images were obtained using a Nikon Eclipse 90i fluorescence imaging microscope which allowed the quick collection and stitching of large areas of tissue using the NIS-Elements AR3.2 (64bit) software.

The stained samples were again mounted on the England finder in the same way as before, the logbook was referenced for the coordinates of the imaged areas, and with the help of these and the optical images of the tissue, the previously imaged location was identified, and a hi-res image was taken. The hi-res image was in the form of a 2x2 mosaic, making sure to capture more than the entire FTIR imaged area, as these optical images were later used to match to the epithelial areas of the tissue in the IR images.
5.2 DATA COLLECTION

A total of 147 FTIR images consisting of over 2.4 million infrared spectra were collected from 24 biopsies from patients with non-metastatic prostate cancer, that stayed non-metastatic for at least 5 years after the biopsy was taken. 15 biopsy sections from patients with metastatic prostate cancer were also obtained, and 94 FTIR images were collected from them.

Post-data collection, 235 successfully matched up high resolution images were taken from the 241 newly stained tissue sections. Figure 5.1 shows a couple of optical images of the IHC stained tissue.

Figures 5.1. (a-c) Stitched optical images of the corresponding FTIR imaged regions of prostate tissue, taken using the Nikon Eclipse 90i fluorescence imaging microscope.
5.3 MANAGING THE DATA

After all the imaging data and the optical images had been taken, the vast task of managing all the data and analysing it appropriately remained. The first step taken was to treat the data with forms of pre-processing.

5.3.1 Immediate Pre-Processing

5.3.1.1 Noise Reduction

Although the latest noise reduction proposed methods by Bhargava et al. seem extremely powerful[115], they are quite complex and excessive for the data sets used in this project. The FTIR-Imaging system used during the course of this thesis (Agilent 670-IR) was one of the top of the range spectrometers and used the largest commercially available FPA on the market, and the single tile image collection could be done relatively quickly with a high number of scans and resolution (4 minutes, using 128 scans and 4 cm⁻¹ resolution). This made the noise contribution to the spectra minimal, so only a basic PCA-based filtration noise reduction was performed.

The way this noise reduction tool works is by obtaining the first 20 principal components of each spectral data set, and reconstructing the data matrix using only the data associated with those principal components. This works very efficiently as those first 20 PCs describe the vast majority of the variance in the data set (99.9%).

5.3.1.2 Application of RMieS-EMSC Using High-Throughput Computing

In previous work described in this thesis (chapter 3), the RMieS-EMSC algorithm had been used to some effect. One year on, the algorithm had been fully developed and was easily applied to all of the noise reduced data. The main reason behind the feasibility of applying the algorithm on such vast amounts of data was the availability of using the University of Manchester’s high-throughput computing
(HTC) resource known as Condor. Condor was originally developed by the University of Wisconsin-Madison nearly 15 years ago [146] as a specialised workload management system for highly compute-intensive jobs. It operates on the principle of eliminating wasted computation time of idle machines.

Jobs are assigned by the user to Condor, which splits them into many smaller jobs. Using a pool of networked computers (at Manchester it equates to over 700 desktop PCs with multiple cores), the smaller job packets are computed using the CPU cycles on the idle machines in the network. The most efficient time for these computations is when many of the computers become idle, such as at night time, weekends and during holidays. The system is a fantastic resource for the faculty and university as a whole, and is the only way that the RMieS-EMSC algorithm could be applied to all 241 FTIR images (this corresponds to almost 4 million spectra).

The RMieS-EMSC was applied using a Matrigel reference spectrum, and was done to 20 iterations.

5.3.2 Image Overlaying and Spectral Extraction

With the FTIR images and the high resolution IHC stained tissue images, work could be started on isolating the epithelial parts of the tissue. In order to extract only the epithelial spectra from the data, a graphical user interface (GUI) was devised using MATLAB to match and overlay the two corresponding images for each sampled area of tissue.

With the IR data and the high resolution images in a workspace, running the GUI would first load up the IHC image in the form of an RGB intensity image and convert it to a grayscale image. The user uses a slider to choose the desired intensity. This step is to highlight the epithelium as strongly as possible. In the IHC staining of the tissue, the epithelium turns a dark brown, while the rest of the tissue is a light yellow-brown colour. Making the dark brown contribution even brighter by
increasing its intensity in a grayscale image, will help to create a strong mask that can be laid over the FTIR image. The first step of the GUI is shown in figure 5.2.

Figure 5.2. Step 1 of the graphical user interface. The slider is moved across the intensity distribution curve to increase or decrease the intensity of the epithelial contribution from the input image. The output is shown in the segmented image on the top right.

Once the intensity has been set, the second step of the GUI is to perform a basic quality control and removal of pixels associated with areas of no tissue. The grayscale image is then recomputed using an amide I threshold between 1597 - 1701 cm⁻¹. The amide I band is a useful indicator in biological tissue, as there are normally strong contributions from the band in areas of tissue due to the high number of cells and proteins within. Areas with no tissue will have very little to no amide I contributions. Another slider is moved by the user to modify the strength and intensity of the quality test and spectral rejection process. The slider position set was always minimal, to make sure that no epithelial spectra were accidentally removed. This step is shown in figure 5.3.
With the slider positions set in steps 1 and 2 of the GUI, a binary mask is created using the information obtained which shows the outline of the suspected epithelium. It is then the user’s task to make sure that this mask lines up directly over the epithelium of the tissue of the FTIR image, using the glands and binary mask outline as a guide. To achieve this, the coordinate changes of the image position must be entered manually, with the user able to change the rotation of the mask from the IHC image, and the horizontal and vertical positions of the image. It was easy to see once the overlay had been applied correctly, with the green binary mask snugly fitting round the glands and epithelial regions of the tissue with no overlap into the wrong areas. Once the user was happy with the overlay, the parameters were saved and the epithelial spectra from the FTIR image corresponding to the areas under the mask were extracted. The final step is shown in figure 5.4.
Figure 5.4. Step 3 of the GUI. The IHC image is shown in the top left, and the FTIR image on the top right with the quality controlled areas overlaid. The binary mask for the epithelium is shown by the green outline in the bottom left image, and the bottom right image shows the final areas of the image that will be extracted as epithelial spectra.

The output at this stage of the analysis was in the form of a data matrix containing all the spectral information from the epithelium-associated pixels, and a binary matrix containing the information of the exact pixel positions of the extracted spectra to be able to reconstruct an image at a later time.

5.3.3 Further Quality Control and Pre-Processing

As a further step to ensure the accuracy of the analysis, the epithelial images and their IHC image counterparts were opened up again. Any regions in the epithelial tissue binary image that still did not correspond to the epithelium-stained regions in the IHC were removed from the data matrix.

At this stage, a final quality test was done on the remaining spectral data, using another minimum-maximum threshold of the amide I wavenumber region to reject
spectra on the edges of the epithelial areas that did not meet the quality standard set.

Before the chemometric analysis of the data, further pre-processing was applied to the data sets in the form of vector normalisation and second derivative. The differently processed data sets would be used in the statistical models for comparison.
5.4 DATA ANALYSIS

Analyses were planned for three separate spectral data sets. These were to be the comparing of areas of non-metastatic cancerous epithelium to areas of non-metastatic, non-cancerous epithelium; the areas of non-metastatic cancerous epithelium to areas of metastatic cancerous epithelium; and areas of non-metastatic, non-cancerous epithelium to metastatic, non-cancerous epithelium. This was unable to be achieved due to the very low number of collected images associated with metastatic, non-cancerous data. It was difficult to find regions of the metastatic cancer tissue that had not been marked by the histopathologist for malignancy. Because of this, analysis was focused on the two first comparisons.

5.4.1 PCA Results

A principal component analysis of such large data sets like we are dealing with, would not only be impossible to interpret, but would also be very computer intensive. For this reason, the average spectrum of each sample image would be used for the analysis instead.

Each patient sample was imaged in at least six different locations, with two of the locations corresponding to non-cancerous areas of the tissue, as marked by the histopathologist. This was the case for most samples, but not all. There was also some difficulty in the successful overlaying of images to extract epithelial spectra for some cases, so these images were removed from subsequent analyses. This left 34 images from 17 patients corresponding to non-cancerous areas of tissue with other areas that exhibited non-metastatic cancer. 40 images were used as the comparative set of non-metastatic cancerous epithelium from the same 17 patients.

Two principal component analyses were done; one on the entire vector normalised infrared spectral region with the CO$_2$ region removed (1000-1800 cm$^{-1}$ and 2700-
3800 cm$^{-1}$), and one using just the vector normalised fingerprint region of the infrared spectra (1000-1800 cm$^{-1}$). The PCA plots are shown in figures 5.5 and 5.6.

Figure 5.5. PCA scatter plots of non-metastatic cancerous epithelium (blue) and non-cancerous epithelium from tissue exhibiting non-metastatic cancer in other areas (green). Only the CO$_2$ wavenumber region has been removed from the data (1800-2700 cm$^{-1}$). (a) PC1 versus PC2, (b) PC1 versus PC3, and (c) PC2 versus PC3.
Figure 5.6. PCA scatter plots of non-metastatic cancerous epithelium (blue) and non-cancerous epithelium from tissue exhibiting non-metastatic cancer in other areas (green). Only the fingerprint region of the spectrum has been used in the analysis (1000-1800 cm\(^{-1}\)). (a) PC1 versus PC2, (b) PC1 versus PC3, and (c) PC2 versus PC3.

The reader can immediately see that there is little to no clustering in the PCA plots of figure 5.5. The plots in figure 5.6 are not much of an improvement, still showing a very large overlap of the clusters. The plot of PC 1 and PC3 in figure 5.6(b) shows the best separation on PC3 with the non-metastatic group clustering more to the
top, and the non-cancerous group more to the bottom, but the results are not good by any means.

The next data set that PCA was applied to was on the metastatic and non-metastatic cancerous epithelium spectra. Once again, due to the extremely large data set, the average spectra were taken for each image before being run through the analysis. The data used corresponded to a total of 40 metastatic cancer average epithelial spectra and 60 non-metastatic cancer average epithelial spectra. The PCA was done twice again; once on the full spectral wavenumber region with CO$_2$ removed, and once on the fingerprint region of the spectra. These results are shown in figures 5.7 and 5.8.
Figure 5.7. PCA scatter plots of non-metastatic cancerous epithelium (blue) and metastatic cancerous epithelium (red). Only the CO$_2$ wavenumber region has been removed from the data (1800-2700 cm$^{-1}$). (a) PC1 versus PC2, (b) PC1 versus PC3, and (c) PC2 versus PC3.
Figure 5.8. PCA scatter plots of non-metastatic cancerous epithelium (blue) and metastatic cancerous epithelium (red). Only the fingerprint region of the spectrum has been used in the analysis (1000-1800 cm\(^{-1}\)). (a) PC1 versus PC2, (b) PC1 versus PC3, and (c) PC2 versus PC3.

The separation observed in the above PCA plots is very clear. When using the whole wavenumber region of the spectrum, PC1 is the largest contributor to the cluster separation, as can be seen in figures 5.7 (a) and (b). PC2 show a small group of non-metastatic cancer spectra clustering away from the rest of the data which can be
seen in figures 5.7 (b) and (c), but is not useful alone. PC3 as can be viewed in figures 5.7 (a) and (c) shows no real effect on the data separation.

When looking at the PCA plots of just the fingerprint region of the spectra in figure 5.8, the clusters separate slightly differently depending on the principal components used. PC1 can no longer be used as a sole basis of separating the data, as there is a slight overlap when plotted against PC2 in figure 5.8 (a). However, when PC1 is plotted against PC3, the combination achieves a perfect separation of the two clusters. This is seen in figure 5.8 (b). In figure 5.8 (c), plotting PC2 and PC3 does not yield any obvious separation of data, apart from a select few non-metastatic cancer spectra moving to the bottom-left of the plot and some metastatic cancer spectra to the top-right. There is still a very significant overlap.

These results are very good, and also interesting when we investigate why the clusters are separating in such a way. This will be discussed later in section 5.5, after the results from the artificial neural networks have been presented.

5.4.2 ANN Training and Testing Results

To construct a useful neural net, it must be clearly outlined what the user wants the output from the ANN to tell them, and feed the net enough data for training and validating the model before testing it. With this in mind, ANNs were constructed to attempt to separate the same data sets as were analysed using PCA in section 5.4.1.

5.4.2.1 Comparing Non-Metastatic Cancerous to Non-Cancerous Epithelium

For the first ANN, epithelium spectral data was used from the data set consisting of non-metastatic cancerous epithelium and non-cancerous epithelium from the same patients. After image overlaying and spectral extraction, there were 21 patients from the original 24 that had undergone the previous processes successfully, and 17 patients also had acceptable images of non-cancerous epithelial regions. Another
aspect required for a clearer interpretation of results is the consistency of grouped analysed spectra. Because each image had a different number of extracted spectra associated with the epithelium, not all the spectra were used. Instead, 1000 random spectra from each set of epithelial data were extracted and saved as the new patient-associated data set. The number of 1000 spectra was used because across the range of images, 1000 spectra was close to the minimum number of epithelial spectra extracted for some images. In some other images, the actual number of associated epithelial spectra ranged up to around 8000. Also, because there were only a maximum of two images from each patient that were associated with non-cancerous regions of epithelium, two images were used from the non-metastatic cancerous regions as well. This resulted in 76000 spectra that would be used in the ANN.

Roughly 70% of the data was used to train and validate the network, using an equal number of patient-associated data from both groups. Data from 10 patients exhibiting non-metastatic cancer and 10 patients with non-cancerous epithelium were used for the training. The number of neurons in the hidden layer used for the training of the network was 20, which was the default setting for the algorithm. The training algorithm used was one available in the MATLAB toolbox, and used 70% of the training data for the training of the model, 15% for validating the model and 15% for internal testing of the model. The network is trained iteratively, and relies on a plot of the mean-squared error (MSE) performance to determine the end point of the iterations, or epochs. The MSE performance decreases rapidly during training, and when the validation curve reaches a global minimum, the training stops. The plot shows curves for the internal training, validation, and test sets.

The performance plot for the first ANN (from now on known as ANN1) is shown in figure 5.9. The spectra have been vector normalised, and have had the CO₂ wavenumber region removed (1800-2700 cm⁻¹).
Figure 5.9. MSE performance plot for ANN1 using vector normalised, non-metastatic cancerous and non-cancerous epithelium spectra with the CO2 wavenumber region removed.

Once the model has been created, the user can run a blind test set through the model to determine how well groups of data separate. 36000 spectra corresponding to 11 patient samples of areas with non-metastatic cancerous epithelium and 7 patient samples of epithelial areas that were non-cancerous were fed into the model. The output and success of the model can be represented by bars showing the class comparisons. The class prediction bars for this ANN are shown in figure 5.10.
Figure 5.10. Bars showing the true class of the spectral data fed in to ANN1 on the right, and the real class prediction on the left. The blue colour is associated with the non-metastatic cancerous epithelium, and the red colour is associated with the non-cancerous areas of epithelium.

The true positives associated with the results in figure 4.23 are 72.47% and 41.58% for non-metastatic cancerous and non-cancerous epithelium respectively. This makes their false positive projections 27.53% and 58.42% respectively. These results are not particularly promising in separating the two classes of data in question. The average spectra associated with each class of data is shown in figure 5.11.
Figure 5.1. (a) Average spectra of each class of spectra (solid line) plus and minus their standard deviations (dotted line). The green spectrum has an origin point of 1 absorbance unit higher than the blue spectrum. (b) Average spectra of each class of data. In both images, the blue spectrum is associated with the non-metastatic epithelium, and the green spectrum is associated with the non-cancerous areas of epithelium.

When the spectra are viewed on the same scale in figure 5.11 (b), very little difference between the spectra can be seen, with only slightly higher peaks at the amide I and amide II bands of the non-metastatic cancer spectrum apparent.

Another ANN (ANN2) was then devised using the same parameters for the data, but with a second derivative applied to the spectra as well. The MSE performance
curve, the predictive potential of the model, and the class associated average spectra are shown in figure 5.12.

Figure 5.12. (a) MSE performance plot for ANN2. (b) Class prediction and true class bars for ANN2. The blue colour is associated with the non-metastatic cancerous epithelium, and the red colour is associated with the non-cancerous areas of epithelium. (c) Average second derivative spectrum associated with non-metastatic cancer epithelium. (d) Average second derivative spectrum associated with non-cancerous epithelium.

The class prediction bar in figure 5.12 (b) shows equally sporadic results as in figure 5.10 for ANN1 with true positives of 65.08% and 28.16% for non-metastatic cancerous epithelium and non-cancerous epithelium respectively, giving respective false positives of 34.92% and 71.84%. Derivative spectra are provided separately in figures 5.12 (c) and (d) because when assigned to the same scale, they superimposed perfectly showing no obvious differences.

A third ANN (ANN3) was constructed in a final attempt to achieve class separation. The data fed into the network this time was vector normalised spectral data from
the fingerprint wavenumber region of the spectra (1000-1800 cm\(^{-1}\)). The performance curve and class predictions bars are shown in figure 5.13.

![Figure 5.13](image_url)

Figure 5.13. (a) MSE performance curve for ANN3. (b) Class prediction and true class bars for ANN3. The blue colour is associated with the non-metastatic cancerous epithelium, and the red colour is associated with the non-cancerous areas of epithelium.

Upon viewing, the class prediction bar seems very slightly better than those of ANN1 and ANN2 and gives true positives of 74.25% and 42.40% for non-metastatic cancerous epithelium and non-cancerous epithelium respectively, with respective false positives of 25.75% and 57.6%. The average class associated spectra are shown in figure 5.14.
Figure 5.14. Average spectra of each class of spectra (solid line) plus and minus their standard deviations (dotted line). The green spectrum has an origin point of 1 absorbance unit higher than the blue spectrum. The blue spectrum is associated with the non-metastatic epithelium, and the green spectrum is associated with the non-cancerous areas of epithelium.

As in the spectra of ANN1 in figure 5.11, the only visible difference between the average class spectra are the slightly higher peaks of the amide I and amide II bands in the non-metastatic cancerous epithelium spectrum.

5.4.2.2 Comparing Non-Metastatic to Metastatic Cancerous Epithelium

Due to the success of the PCA results in section 5.4.1, there were high hopes of developing a successful and robust model using artificial neural networks to separate metastatic from non-metastatic cancerous epithelial spectra. For this study, images corresponding to 21 patients with non-metastatic prostate cancer and 14 patients with metastatic prostate cancer were used. 1000 random spectra were extracted from three images for each patient, totalling 105000 spectra. For the first ANN (ANN4), 60000 total spectra from 10 non-metastatic cancer patients and 10 metastatic cancer patients were used to feed into the ANN as a training set. The remaining 45000 spectra from 4 metastatic cancer patients and 11 non-metastatic cancer patients was used as the blind test set and fed into the completed
model. All spectra used in ANN4 were vector normalised and had the CO$_2$ wavenumber region of the spectrum removed. The results are shown in figure 5.15.

![Figure 5.15](image)

Figure 5.15. (a) MSE performance plot for ANN4. (b) Class prediction and true class bars for ANN4. The blue colour is associated with non-metastatic cancerous epithelium, and the red colour is associated with metastatic cancerous epithelium. (c) Average spectra of each class of spectra (solid line) plus and minus their standard deviations (dotted line). The red spectrum has an origin point of 1 absorbance unit higher than the blue spectrum. The blue spectrum is associated with the non-metastatic epithelium, and the red spectrum is associated with the metastatic cancer epithelium.

The results show a very good model for separating non-metastatic from metastatic spectral data. In comparison to the ANN results of the previous study in section 5.4.2.1, the mean squared error in figure 5.15 (a) is about two orders smaller than in the previous models. The predicted class bar in figure 5.15 (b) is extremely close
to the true class representation, yielding true positives of 99.44% for non-metastatic cancer, and 99.97% for metastatic cancer. The false positives are therefore only 0.56% and 0.03% for non-metastatic and metastatic cancer respectively. These numbers are extremely good, and the average class spectra in figure 5.15 (c) also show much higher absorbance values for the metastatic data across the majority of the spectrum.

The next ANN (ANN5) used the same parameters for data selection and extraction, but all spectra had a second derivative applied before training and testing the ANN. Results from ANN5 are shown in figure 5.16.
Figure 5.16. (a) MSE performance plot for ANN5. (b) Class prediction and true class bars for ANN5. The blue colour is associated with non-metastatic cancerous epithelium, and the red colour is associated with metastatic cancerous epithelium. (c) Average second derivative spectrum associated with non-metastatic cancer epithelium. (d) Average second derivative spectrum associated with metastatic cancerous epithelium.

The results from ANN5 are equally as good as the ones from ANN4, although the MSE is almost a further order of magnitude smaller, as can be seen in figure 5.16 (a). The true positives are almost identical to ANN4 which are 99.44% and 99.97% for non-metastatic and metastatic cancer epithelium respectively. The second derivative spectra in figure 5.16 (c) and (d) show a similar structure, except that the
metastatic cancer-associated spectra are on a larger absorbance scale. The other aspect of the spectra that seems different is the make-up of the peaks across the fingerprint region between 1000 and 1800 cm\(^{-1}\). This wavenumber region will be investigated further in the next ANN model.

ANN6 was built using only the vector normalised fingerprint regions of the spectra to train and test the network. Results from this network, are shown below in figure 5.17.

Figure 5.17. (a) MSE performance plot for ANN6. (b) Class prediction and true class bars for ANN6. The blue colour is associated with non-metastatic cancerous epithelium, and the red colour is associated with metastatic cancerous epithelium. (c) Average fingerprint region spectra of each class of spectra (solid line) plus and minus their standard deviations (dotted line). The red spectrum has an origin point of 1 absorbance unit higher than the blue spectrum. The blue spectrum is associated with the non-metastatic epithelium, and the red spectrum is associated with the metastatic cancer epithelium.
ANN6 also shows a low degree of error in figure 5.17 (a), of the order of $10^{-3}$. While looking at a smaller wavenumber region, the class prediction bar is still very promising. The true positives obtained from the information given in figure 5.17 (b) are 98.41% and 99.96% for non-metastatic and metastatic cancer epithelium respectively. This makes the sensitivity for non-metastatic cancer a little lower than in the other models, but the metastatic sensitivity remains roughly the same as in ANN4 and ANN5. Once again, as can be seen in figure 5.17(c), the metastatic-cancer associated spectra have higher overall absorbances. With a closer view of this particular wavenumber region of the spectra, the first large peak from the left has a different shape for the two spectral classes. This will be looked at in further detail in section 5.5.

Thus far, the constructed neural nets for separating non-metastatic from metastatic epithelium have shown very good results. This was achieved using a training set that was made up of spectra from 20 patients. Because of the extremely good sensitivities so far, another set of ANNs were constructed using a much smaller training set that would allow the projection of most of the data into the model as a blind test set. All spectra were randomly extracted from each patient image, making up the same total of 105000 spectra from 21 patients with non-metastatic cancer and 14 patients with metastatic cancer. This time, a training set of 24000 spectra was constructed from the spectral data from 4 non-metastatic and 4 metastatic cancer patients. This allowed the use of the remaining 81000 spectra as the blind test set, nearing twice the size of the test set used in ANNs 4-6.

All spectra in ANN7 had the CO$_2$ region removed (1800 – 2700 cm$^{-1}$) and were vector normalised before the model was created. The results are represented in figure 5.18.
Figure 5.18. (a) MSE performance plot for ANN7. (b) Class prediction and true class bars for ANN7. The blue colour is associated with non-metastatic cancerous epithelium, and the red colour is associated with metastatic cancerous epithelium. (c) Average spectra of each class of spectra (solid line) plus and minus their standard deviations (dotted line). The red spectrum has an origin point of 1 absorbance unit higher than the blue spectrum. The blue spectrum is associated with the non-metastatic epithelium, and the red spectrum is associated with the metastatic cancer epithelium.

The model shows a relatively smooth MSE curve in figure 5.18 (a), although the class prediction bar in figure 5.18 (b) is worse than in the other set of ANNs. This is to be expected as the training set is a lot smaller, so less information has been used as a basis for the prediction of blind data. The true positives are 99.68% for the non-metastatic data, and 95.5% for the metastatic cancer data. This makes the false
positives 0.32% and 4.5% for non-metastatic and metastatic cancer epithelium respectively. The spectra in figure 5.18 (c) once again show the average metastatic cancer-associated spectra to have higher absorbance value across the spectra.

As before, another neural net was modelled using the second derivative spectra of the data. Results are shown after building ANN8 in figure 5.19.
Figure 5.19. (a) MSE performance plot for ANN8. (b) Class prediction and true class bars for ANN8. The blue colour is associated with non-metastatic cancerous epithelium, and the red colour is associated with metastatic cancerous epithelium. (c) Average second derivative spectrum associated with non-metastatic cancer epithelium. (d) Average second derivative spectrum associated with metastatic cancerous epithelium.
ANN8 gives the best results yet, with true positives for the non-metastatic cancer data and metastatic cancer data 99.64% and 99.71% respectively. However, the performance curves in figure 5.19 (a) show a sharp rise in the validation curve as the training curve drops off. There is also a difference in trend at the end of the validation and test curves, suggesting something irregular. These performance curve traits point towards the overfitting of data, so caution must be taken when interpreting the results, as they are better than any previous ones, and should not be because of the smaller training set used. Spectra in figure 5.19 (c) and (d) show no immediately obvious differences other than again, the scale of the absorbance axis.

A final artificial neural network, ANN9, was constructed with the smaller training set, this time looking at the vector normalised fingerprint region of the epithelial spectra.
Figure 5.20. (a) MSE performance plot for ANN9. (b) Class prediction and true class bars for ANN9. The blue colour is associated with non-metastatic cancerous epithelium, and the red colour is associated with metastatic cancerous epithelium. (c) Average fingerprint region spectra of each class of spectra (solid line) plus and minus their standard deviations (dotted line). The red spectrum has an origin point of 1 absorbance unit higher than the blue spectrum. The blue spectrum is associated with the non-metastatic epithelium, and the red spectrum is associated with the metastatic cancer epithelium.

Figure 5.20 (a) displays one of the smoothest and consistent performance curves of all the ANNs, suggesting a very stable model, even though the MSE is slightly larger. As expected, the class predictions in figure 5.20 (b) are worse than in ANN6, due to the smaller training set, but overall, the sensitivities are still very good at 95.89% for non-metastatic cancerous epithelium and 92.64% for metastatic epithelium. This
makes the false positives for non-metastatic and metastatic cancer-associated spectra 4.11% and 7.35% respectively. Spectra shown in figure 5.20 (c) are similar to their counterparts from ANN6, with higher absorbance values for the metastatic spectra and a slightly different shape of the amide I region.

Now the results have been presented in full, a further discussion linking trends and observations can be made.
5.5 FURTHER DISCUSSION OF RESULTS

5.5.1 Non-Metastatic Cancerous and Non-Cancerous Epithelium Results Comparison

When comparing the PCA results and the ANN models, there is a correlation with the poorness of the models’ abilities to successfully separate and classify the epithelial spectral data from the two classes. The representative average spectra of the classes viewed in figures 5.11, 5.12 and 5.14 show a high degree of similarity, with the only apparent difference being slightly higher peaks at the amide I and amide II regions. The similarity between the average spectra is made clearer when focusing on solely the fingerprint region in figure 5.14.

Prediction bars from ANNs 1-3 are erroneous in their classifications of data, and while the wrongful assignment of the classes may be random, it can be seen that large clumps of spectral data are assigned to the same class. For example, for ANN2 in figure 5.12, the class prediction bar shows relatively solid ‘blocks’ of metastatic data classifying as non-metastatic data in areas between 0 – 2000, 4000 – 6000, 15000 and 20000 to name a few. Although these represented areas are not completely solid, the majority of the spectra therein classify the same way. This is indicative of almost the entire extracted spectral data sets from specific patients classifying together. One can therefore assume that some of this data is separating by patient rather than anything else. However, where grouping does occur, it associates to the two images taken from each patient class. There is no conclusive evidence that all patient associated spectra do group together, because where the blind test data set shared both non-metastatic and non-cancerous spectra from the same patient, they sometimes, but not always classified the same way. In ANNs 1-3, 3 patients had spectral data in both classes. The areas on the class bars that are linked in this way are between 6000 – 8000 and 28000 – 30000 for patient 8, 10000 – 12000 and 30000 – 32000 for patient 10, and 16000 – 18000 and 26000 – 28000 for patient 4.
Research by Gazi et al. has shown that the ratio of the bands associating with vibrations of glycogen (1030 cm\(^{-1}\)) and phosphate (1080 cm\(^{-1}\)) can be used to separate benign from metastatic cells of the prostate\[79\]. Although the epithelial cells were of a non-metastatic nature, a rough calculation of the glycogen-phosphate ratio was calculated for the average spectra to see if there were any other suggestions that the class spectra may have subtle differences. Upon calculation, the ratios across all the spectra were similar giving almost identical minimum, maximum and mean values of the ratio (0.63, 0.82, and 0.72 respectively), which do not agree with Gazi’s findings. This may be to do with the tissue not being of a metastatic cancer nature, but nevertheless, no further evidence has been discovered to allow the separation of the spectral classes.

The similar spectra and lack of separation between the two classes of spectral data may suggest a couple of hypotheses. It could be that the non-cancerous epithelial cells have already developed the same chemical signatures as the cancerous areas of epithelium, even if not obvious to the histopathologist. On the other hand, the non-cancerous regions of epithelium could in fact be cancerous, and had not been marked accurately enough by the histopathologist. This is not an easy problem to overcome, as cancerous areas of tissue are not always so clear cut when distinguishing from areas of no cancer. In past research, on closer inspection of some samples after revisiting the histopathologist, it turned out that only the very clear, main areas of cancer tissue were marked for analysis and some unmarked areas were also cancerous. Without the histopathologist on hand at the time of data collection, or the very precise outlining of specific regions of cancer which would require a time-consuming histological analysis of the tissue, this will always remain a possible concern in the collection and analysis of data.

5.5.2 Non-Metastatic and Metastatic Cancerous Epithelium Results Comparison

The results from the analysis of the metastatic and non-metastatic cancer epithelium data classes were a lot more successful and interesting. Starting with the
PCA results, the separation will be looked at in more detail. Figure 5.7 shows the good separation of the data when plotting PC2 or PC3 against PC1, and while PC1 completely classifies the data into different groups, there is a small amount of separation on PC2 as well. PC2 shows a cluster of 9 non-metastatic spectra away from the rest of the data, almost as if it were a third cluster. In figure 5.21 below, the loadings of PC1 and PC2 are shown.

![Figure 5.21](image)

Figure 5.21. (a) Scatter plot of non-metastatic (blue) and metastatic (red) cancerous epithelium showing PC2 against PC1. (b) PC1 loading plot. (b) PC2 loading plot.

When checking the loading of PC2, one can see that the loading peak that is pushing the small cluster of data above 0.05 on the y-axis is at $\sim 1620 \text{ cm}^{-1}$. This peak is associated with nucleic acids and is due to the base carbonyl stretching and breathing mode of vibration[147]. The literature does not suggest any obvious
reasons why part of the data set should be separating on this peak, but interesting nonetheless that no metastatic spectra have noticeable contributions from this area. The PC1 loading shows the notable positive separation peaks as the majority of the higher fingerprint region, attributed mainly to N-H bending in the amide I and II regions, and the notable negative peaks corresponding to certain N-H and O-H stretching regions[68]. Focusing on the strong contribution from the fingerprint region in the positive direction of PC1, it was useful to look at the results from the fingerprint region spectral analysis.

The scatter plots in figure 5.8 showed a reasonably good separation of data using PC1 and PC2, but the best was when plotting PC3 against PC1. Once again, PC1 can almost fully correctly classify the data by itself, but not as well as when looking at the whole spectral region. Using PC3 in addition adds another dimension to the analysis and the clear clustering of the epithelial spectral data is complete. The scores plot and the loadings on PC1 and PC3 are shown below in figure 5.22.
It is difficult to see whether any features represented in the loading of PC3 may be classifying the data, as there is no clear separation using that principal component alone. However, the scale of the axis of the PC shows that any classification has been determined between the -0.05 and 0.05 range which holds the peak information of various carbohydrates and phospholipids in DNA and RNA [68]. The interesting separation occurs on PC1, where on closer inspection the red cluster of the metastatic data can be seen to move to the negative side of the principal component below around -0.05. Looking at the loading, one can clearly see the only notable peak below that value is at around 1080 cm\(^{-1}\) corresponding to the phosphate symmetric stretching band. As mentioned in section 5.5.1, this band has been shown to be one of the markers for metastatic prostate cancer in comparison.
to benign data[79]. Higher absorbance values at this band would reduce the value of the glycogen-phosphate ratio, provided the glycogen band at 1030 cm\(^{-1}\) remained consistent across both classes of data. The glycogen-phosphate ratios were calculated for the metastatic and non-metastatic cancer data, but they did not show any clear differences, and again did not agree with Gazi’s findings as the average ratios equated to about 0.73 for both classes of data. Further specific work would need to be done to accurately investigate this ratio of bands across the spectral data sets, as the rough calculation using only the average spectra of each image is not sufficient for qualitative analysis in this regard.

When looking at the average spectra of the two classes from ANN9 in figure 5.20, the scale differences make it difficult to analyse and subtle differences in peak heights and shifts that may be present. In order to slightly account for this, a constant was applied to the average non-metastatic spectrum, to make the scale more even. The amplified non-metastatic average spectrum, and the average metastatic cancer spectrum are presented in figure 5.23.
Two standout features are shown by the spectra in figure 5.23. Firstly, the shape and peak position of the amide I band is different, and the ratio of the amide II band and the phosphate band are also different.

The sharp peak of the amide I band in the average metastatic spectrum is at 1659 cm\(^{-1}\), and is probably due to a different secondary structure of the protein. Deconvolution of the amide I band and curve fitting has shown in studies that the sharp peak is highly likely to be attributed to a higher occurrence of \(\alpha\)-helical contributions, also showing that the maximum peak of the \(\alpha\)-helical substructure is more prevalent in malignant cells[93]. In comparison, the amide I region of the non-metastatic spectrum is broader, suggesting a lower contribution from the \(\alpha\)-helix structure relative to the other sub-structural components.

Other research has shown that the ratio of the symmetric to asymmetric stretching of the phosphodiester groups in nucleic acids (1084/1240 cm\(^{-1}\)) ratio can be
indicative of malignant epithelial cancer cells[148]. When looking at the overlaid spectra in figure 5.23, the non-metastatic spectrum clearly presents a different ratio of these peaks to the metastatic data spectrum. A rough calculation of the ratio was done over the entire range of spectral data and the average ratios were about 0.71 for the non-metastatic epithelium, and 0.78 for the metastatic epithelium. The earlier calculations of the glycogen-phosphate ratios only showed differences of about 0.01, but the phosphate symmetric to asymmetric stretching ratio is different by 0.07 which is significantly higher. This may also link to the conclusions that were made from the loading plot in figure 5.22 (b), with the 1080 cm$^{-1}$ band being significant in a different ratio of peak absorbances to originally thought.

The difference of absorbance values across the non-metastatic and metastatic spectra may also have a strong part to play in the classification of data. Any physical differences that may normally be attributed to affecting absorbances can be discounted however, as all spectra have been thoroughly pre-processed, eliminating all sources of light-scattering and the like using the RMieS-EMSC algorithm and vector normalising. All treatment of the samples during preparation was identical, even though the cutting and mounting of each class of samples took place at different times. The only other factor to consider as an inherent difference between the non-metastatic and metastatic cancer samples is the type of biopsy. The non-metastatic cancer patients had their biopsies taken during a trans-urethral resection of the prostate, while metastatic cancer patient biopsies were all radical prostatectomies. Although the samples were obtained from a shared tissue bank, they did come from different hospitals. It is important to note that previous unpublished research in the Gardner group has also used mixed samples in this way for analysis, and there was no impact from the different natures of the samples on the results. Higher absorbances across the metastatic cancer spectra were not observed in the raw spectra from the images, shown by an example in figure 5.24 below.
Figure 5.24. Average raw spectra from IR images collected from a non-metastatic cancer patient (blue) and a metastatic cancer patient (red).

These spectra are averaged over the whole tissue, while the spectra used in the ANN and PCA analyses were the isolated chemical signatures from the epithelium, so there is no reason to suggest that the raw spectra are influencing the observed increasing absorbance values. They are more likely to be influenced by the discrete chemical signatures of the metastatic cells, that show a higher degree of proliferative signalling and gene expression in the DNA[149].

The ANN results, including the ones using a smaller training set in ANNs 6-9, showed a very good classification of the non-metastatic and metastatic cancer data sets. Although powerful for non-linear modelling, there are disadvantages to using neural nets to classify data. First of all, although the MSE performance curves can give some indication of how well the model is performing, it is not too clear exactly how robust the model is. Overfitting may occur if the validation curve increases sharply as the training curve continues to drop. This was viewed in ANN8, and this was discussed earlier with suspicions of overfitting due to the better nature of the results in comparison to the ones using the larger training set in ANN5. Another
downside is that the ANNs have a binary output, which means that although true and false positives can be calculated to indicate the sensitivities of the data modelling, the specificities cannot be calculated because there is no way of testing the likelihood of a true negative in a patient without the disease.

The overall process used in this study, from data collection process to the analytical techniques used, has proved successful, although enhancements can definitely be made for any future work. First of all, the principle of IHC staining the analysed section of tissue after data collection is a very good one, reducing the error of any overlaying of images due to a difference of tissue structure. It does however, have its drawbacks as the staining did not always prove 100% successful as in some cases the tissue was ripped in areas, morphing the tissue architecture. These images were not used in the analysis but it did reduce the sample set sizes. It is crucial to stain as efficiently as possible because there is only one chance to do it correctly. The image overlaying system used was also successful, isolating the epithelial regions of tissue well. There will always be some degree of error though, with the manual input of orientating the images.
5.6 SUMMARY AND CONCLUSIONS

An optimised protocol for the FTIR imaging of prostate cancer biopsies was developed after months of prior research, resulting in the successful acquisition of thousands of high quality FTIR spectra from areas of the imaged samples with a minimum noise contribution. The immunohistochemical staining of the imaged tissue biopsy samples allowed the overlaying of optical IHC images and their FTIR image counterparts with minimum error in terms of the changing of tissue structure that is often present in serial tissue sections. The spectra in each image corresponding to the epithelium of the tissue were extracted, based on the results from the binary mask created from the tissue overlay process.

A thorough pre-processing analysis was applied to the very large sample sets including a quality control, noise reduction, vector normalisation and RMieS-EMSC using high-throughput computing, drastically reducing computation times and highlighting the potential of the university’s Condor resource.

Two separate analytical experiments were carried out, using PCA and artificial neural networks to attempt to describe the separation of the data. The first experiment aimed to classify spectra from both cancerous regions and non-cancerous regions of epithelium from images taken from the prostate biopsy sections of patients whose cancer was classified as non-metastatic on presentation, and stayed non-metastatic for at least five years after the biopsy was taken. The areas that were imaged were based on the marked guidance of an experienced histopathologist. Both the PCA results in section 5.4.1 and the first three ANN models in section 5.4.2.1 were quite poor in the separation of the classes, and some hypotheses were made. The extracted non-cancerous spectra were in fact taken from cancerous regions, an unfortunate error that could be explained by the vagueness and interpretation of the histopathologist’s markings. Another more likely alternative is that the non-cancerous epithelial spectra cannot be treated as benign spectra as they are still taken from biopsies of patients with prostate cancer,
and chemical changes and imbalances promoting cell growth (as inherently occur in cancer cells) may already have occurred in the ‘non-cancer’ cells but is not yet visible to a histopathologist. A comparison of average spectra from the two classes of data did not show any notable differences.

The second experiment, which was the culmination and main objective of the research presented in this thesis, was carried out on the epithelial spectra extracted from the FTIR images of prostate biopsy samples of one set of patients who were diagnosed with non-metastatic cancer on presentation, and one set of patients who were diagnosed with metastatic prostate cancer. Both the PCA and ANN results were very good at separating non-metastatic from metastatic data as can be seen in sections 5.4.1 and 5.4.2.2. When analysing the entire wavenumber region of the epithelial spectra, upon further investigation in section 5.5.2 shows the loadings of PC1 and PC2 to be the best indicators for the data clustering, with PC1 able to classify the data correctly alone. With the loading plot indicating contributions from the fingerprint region of the spectra as the likely classifier, analysis was done again using only the smaller wavenumber region of the spectra (1000 – 1800 cm\(^{-1}\)). The fingerprint region PCA plots show good separation again, with PC1 and PC3 classifying all the data correctly when plotted against each other. A closer look at the loadings suggest that on PC1, the metastatic data may be separating on the notable peak at 1080 cm\(^{-1}\), associated with the symmetric stretching mode of phosphates. Further analysis of the metastatic and non-metastatic spectra indicated a difference in the symmetric-asymmetric stretching mode ratios of phosphate, possibly linking with some results from other literature[148].

The majority of the developed ANN models seemed robust and results reliable, backing up the PCA cluster separation well, even if the information from the models themselves could not describe reasons for the separation. All in all, the proposed research was a success and lays the foundation for further work in the area, which will be outlined in Chapter 6.
6. PROJECT CONCLUSIONS AND FUTURE WORK POTENTIAL

6.1 FTIR Imaging Development and Tools

The development of FTIR Imaging as a technique for the analysis of prostate cancer tissue biopsies has been refined after many months of research, and has been presented over the course of this thesis, culminating in an optimised protocol for measuring FTIR images in a large scale study. Chapters 3 and 4 showed the in-depth analysis of these tissues, showing that HCA clustering and subsequently, K-means clustering are both good tools for matching the clustered FTIR images to the real tissue morphology. These conclusions were made under the collaborative advisement of an experienced histopathologist.

FTIR imaging has been shown to have many advantages over single-point spectroscopy, both qualitatively and quantitatively. The high number of spectra obtained from each FTIR image provide a very precise spectral representation of every pixel of the image, and with a good enough protocol for extracting the spectra of interest, more qualitative analyses can be done. This was shown by the success in Chapter 5 of the development of an image overlaying tool and extraction process for epithelial spectra.

Although the image overlaying system helped to achieve very reasonable spectral data sets, it was a very subjective process. There is room for a lot of future development of the protocol, and a more objective way of extracting spectra must be targeted. Currently, the manual input of orientating the IHC stained image to superimpose the FTIR image as well as possible is difficult and time-consuming as well as being subjective. A completely objective overhaul of the system would be very difficult, as some manual input would be required to set up parameters for the overlay. There are some available tools in MATLAB for automatic image registration, but they struggle dealing with images of different dimensions and resolutions. Image overlaying can also be done by assigning reference points in both images and
the overlay is computed thereafter. Any results would still have to be checked thoroughly, and the method may not actually prove any less time consuming if it does not immediately work efficiently.

The IHC staining of the prostate sections using the anti-pan cytokeratin antibody was a novel approach that was very good at highlighting the areas of epithelium in tissue. There were many cases however, where although the staining worked well, there were also tears to the tissue and some areas that were not very clean near small air bubbles under the coverslip of the slide. While the air bubbles had no impact when visually matching up images, it did cause trouble when overlaying because the bubble areas were sometimes darker than the stromal regions of tissue. When setting the threshold for the intensity of the image, these areas were grouped together with the epithelial areas. In these cases a much lower intensity was set so the non-epithelial regions of tissue were not used in the analysis. This also lowered the number of extracted epithelial spectra, but it was justified as overall it reduced the chance of using spectra of the wrong class in subsequent analysis. In the future, greater care must be taken in the IHC staining of tissue so there are no imperfections in the stained optical images. A better staining of tissue will produce better overlay results. The success of staining the specifically imaged tissue sections opens the way for other isolation work; with other stains, different areas of tissue can be highlighted and the corresponding images could be used for overlaying and spectral extraction. In FTIR cancer research however, there is no real application of using a different type of stain, as an epithelial marker is all that is needed.

The direction of future FTIR imaging work will slowly move towards the automated data collection of entire tissue biopsies. Although this was not done in the Chapter 5 study, the automated mosaic function of the Agilent Resolutions Pro. imaging software and microspectrometer show the potential for obtaining large multi-file mosaics of very large sample areas. Entire tissue biopsies can be imaged in this way, and although tiled images will take longer to collect, the resulting output images
will be incredibly useful in many ways. An advantage of imaging whole tissue sections is that one need not worry about any irreversible post-data collection staining of the tissue, as all the desired spectral information from all possible sampling areas of the tissue will have been recorded and the original sample should not be needed again. Any or all specific regions of interest of the tissue would then be available to use for analysis. If future data sets were all collected in this way, a databank filled with the imaging spectral information of all samples could be created for referencing and for use in further studies. For this protocol to work however, all image contributions to the databank would have to have been imaged using the exact same parameters in data collection, so that any data used would be fully compatible with each other.
6.2 Metastatic and Non-Metastatic Cancer Study

Analysing the epithelial spectra from metastatic and non-metastatic cancer tissue biopsy sections using PCA and artificial neural network modelling proved very successful, as the two classes of data separated well. Loading plots backed up by the associated class spectra suggested a possible separation of data on the symmetric to asymmetric stretching mode ratio of the phosphates present in the nucleic acids within the epithelial cells (1084 / 1240 cm\(^{-1}\)). These results call for a further investigation of these peaks across the whole spectral data sets, not just the average spectra. The ratio itself would be calculated for each spectrum in the data sets, and those ratios alone would be used in training and testing other PCA and ANN models. This would hopefully give more insight into whether this ratio is truly significant in separating metastatic from non-metastatic cancer epithelium.

The fingerprint region of the spectra was shown to be able the separate the classes of data alone. Closer inspection of the fingerprint wavenumber region of metastatic and non-metastatic spectra revealed not only higher absorbance values of the metastatic data, but also a difference in shape of the amide I band. The sharper peak suggested a higher contribution from the \(\alpha\)-helical substructure of the band, but a Fourier self-deconvolution of the region would have to be done to see the real qualitative differences between the substructures of the amide I in metastatic and non-metastatic cancer epithelium. New PCA and ANN models could be created using the wavenumber region of just the amide I band of the spectral data sets to see whether the different substructure features are consistent across the classes of data.

The results from the main study in Chapter 5 open the way for an expansion of the study. With successful ANNs already built, any future data could be projected into the models as an addition to the blind test data set. This would further test the robustness of the model. Tissue sections from more patients would be needed for further data collection. Any patient samples that would be able to be obtained from
tissue banks other than those already used as a resource would be beneficial. The successful classification of the data when fed through the existing models would completely eliminate any suspicion of class separation occurring due to the type of biopsy and source of the patient samples.

The non-metastatic prostate cancer biopsies used for data collection and analysis were diagnosed on first presentation of the biopsy, and patient follow-ups of up to five years after the initial biopsy date showed no metastatic changes of the cancer. Introducing a third class of data into the study, of non-metastatic cancer patients whose cancer turned metastatic would be very interesting to investigate. Any differences between the spectra from this class and the others may be viewed as the missing link to understanding the aggressive change in the chemistry of cancer cells, and give rise to certain markers in the infrared spectra of the epithelial cancer cells that may explain the change. The difficulty of obtaining a large set of such samples for analysis is high, so much liaising with clinicians who have access to the patient information and contacts managing other tissue banks would be necessary.

The possible avenues of continuing this research mentioned here will help to reinforce the fact that FTIR imaging coupled with rigorous pre-processing methods and multivariate analysis techniques is a useful tool for investigating the chemical differences between epithelial prostate cancer tissues of metastatic and non-metastatic natures. The successful data analysis of large patient sample sets may conclusively prove the spectral features that separate the data classes, and those results can potentially be developed to isolate the gene expressions or markers in cancer cells through new antibody stains.

The real clinical relevance however, would be in discovering the precursor to the metastatic cancer changes in epithelial cells from positive results obtained after analysing data from the suggested third class of patients who exhibited non-metastatic prostate cancer on biopsy presentation but went on to develop metastases at a later date. Since metastatic prostate cancer is currently only
discovered after the spread of the cancer to surrounding organs, and usually dealt with by the removal of part or all of the prostate gland, the early discovery of a high likelihood of metastatic cancer progression would be positively life-changing for a prostate cancer patient.

Before achieving any kind of transition of the FTIR imaging technology to the clinic, a number of obstacles and problems must first be overcome. One major, key issue in this transition is obtaining the understanding, cooperation, trust and full support from the relevant clinicians involved in the assessment of prostate cancer. The idea must be clearly put across that FTIR imaging is a useful complementary technique to aid pathologists in prostate cancer diagnosis and assessment, and is in no way threatening to their jobs. Further studies must be done involving large sample sets obtained from multiple tissue banks that must show conclusive results that demonstrate that with the FTIR tool at a pathologist’s disposal, it can be used whenever applicable to significantly strengthen the specificity and sensitivity of a diagnosis, and improve the rapidity of coming to a conclusive result.

It may still take a long time for the important studies mentioned in the previous paragraph to be realistically carried out. To convince clinicians of anything, the community must be sure, and in agreement of all data acquisition and processing protocols. Since the technology is constantly evolving, and new solutions to problems are being proposed, there is currently little consistency and stability in the field of FTIR data analysis.

Agreement is important in the methods of sample preparation including how to mount samples; in what medium (if any) to mount them; and on what substrate to mount them on.

The discoveries and explanations of physical effects that contribute to false features in spectra must be eliminated in an optimised manner that is approved by the FTIR
community as a whole. Methods of pre-processing spectral data must be optimised, as well as deciding on appropriate data analysis techniques and how best to interpret and explain results correctly. Interpreting and explaining results correctly may be the most important transitional step as far as the data analysis goes, as it must be outstandingly clear that the obtained results are not only correct, but more importantly, relevant.

These are not the only hindrances to the future of using FTIR imaging as a clinical tool. There is also a commercial aspect that should not be overlooked. First of all, costs and practicality of the instruments need to be carefully assessed. When costs are as low as can be acceptable, and instruments become more portable and user-friendly, only then would they potentially be considered for widescale use. IR-specific microscopes may also have to be purchased, and for complete sales viability, a manufacturer must show their equipment to be leaps and bounds above the competition.

The field of vibrational spectroscopy and imaging, specifically FTIR, is constantly growing, and the community is becoming more knowledgeable as promising research and ideas are shared. It will not be too long before there is full agreement on technical issues, problems with data quality and processing protocols, and their optimised solutions. It is then just a matter of time before those all-important large scale studies are undertaken to showcase the power and significance of the FTIR imaging technique in the important clinical world. When the need and want of the technology is apparent, there is no question that instrument manufacturers will meet customer demands by offering technological systems that are superior, more practical and cost effective than currently on the market today.
7. REFERENCES


60. Dorling KM, Development of an objective spectroscopic based grading system for prostate cancer biopsies, in The Department of Chemical Engineering and Analytical Science (CEAS)2008, The University of Manchester.


