Developments in infrared spectral histopathology using machine learning algorithms

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

2019

Jiayi Tang

School of Engineering

Department of Chemical Engineering and Analytical Science

List of Contents

Final word count: 63405

List of Tables

Lists of Figures

List of Abbreviations

Normal associated tissue	NAT
Principal component analysis	PCA
Area under curve	AUC
Receiver operating characteristic	ROC
Principal component discriminant function analysis	PC-DFA
Prostate adenocarcinoma	CaP
Formalin fixed paraffin embedded	FFPE
Extended multiplicative signal correction	EMSC
Electric Field Standing Wave	EFSW
Principal Component Analysis	PCA
Tissue microarray	TMA
Haematoxylin and eosin	H&E
Focal plane array	FPA
Quantum Cascade Lasers	QCLs
Mercury cadmium telluride	MCT
Support vector machine	SVM
Artificial Neural Network	ANN
Linear discriminant analysis	LDA
Partial least squares discriminant analysis	PLS-DA
Fourier Transform Infrared	FTIR
Readout integrated circuits	ROIC
Field of View	FOV
True positive	TP
True negative	TN
False positive	FP
False negative	FN
True positive rate	TPR
False positive rate	FPR
Out of bag	OOB
Caveolin-1	Cav-1
Caveolin scaffolding domain	CSD
Principal component	PC
Adaptive Boosting	AdaBoost
Canonical variate analysis	CVA
Principal component canonical variates analysis	PC-CVA
Quadratic discriminant analysis	QDA
Calcium fluoride	CaF_2
Barium fluoride	BaF_2

Abstract

Fourier Transform infrared spectroscopy, in particular, infrared microspectroscopy, has great potential for clinical applications in the flow of cancer diagnosis. Using large focal plane array detectors and with advancements in computer power, infrared hyperspectral imaging has significant advantages in both accuracy and speed of diagnosis. In light of previous research on cancer diagnosis and digital histopathology using infrared imaging, further studies combined with machine learning algorithms have been conducted and are presented in this thesis. Human tissue samples including breast and prostate have been studied.

Initial studies have been conducted on breast tissue on CaF_2. Infrared images were obtained and analysed using two machine learning algorithms namely Random Forest and AdaBoost. This demonstrated that good classification results, classification accuracies of 89% and 92%, could be obtained to distinguish cancerous from normal associated tissue (NAT). The caveolin-1 stain was applied as a possible breast cancer diagnosis correlated stain. Classification accuracies on cancerous and NAT spectra were 100% and 71.4% respectively in the independent test, which indicates the great potential of caveolin-1 as a biomarker correlated with breast cancer diagnosis. For further implementation of infrared spectroscopy into clinical field, glass substrates, which are cheap and robust, are selected as potential new substrate for infrared disease diagnosis. Studies related to the performance of cancer diagnosis and digital H&E staining using infrared spectra collected from glass slides were conducted on breast tissue. Excellent separation between cancerous and NAT spectra was obtained with classification accuracies of 81.3% and 83.2% on cancer and NAT classes in the independent test. In addition, unbalanced classes are commonly observed in breast tissue analysis, as the epithelium cells are often much fewer in number compared with the stroma cells. A study using different sampling methods and classification methods to solve the problem and boost the classification results was conducted on the spectra collected from breast tissue on CaF_2. Lastly, to test whether similar performance of classification can be observed from other types of tissue, studies on prostate tissue with glass substrates were also conducted. Reasonable classification results, classification accuracies, 72% and 68% were obtained with threshold (85% top scored testing spectra) added in the independent test (10 cores).

Declaration

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

Jiayi Tang

Copyright Statement

Acknowledgement

First of all, I would like to thank my parents for their support during my four-year study. Without their encouragement, I would not accomplish this honour.

Huge thanks to my supervisor Professor Peter Gardner, who has continuously guided me since I started my undergraduate study and lead me to my PhD degree

Special thanks to Doctor Alex Henderson for all the tutorials on Matlab coding and all the brilliant ideas about data processing

I would also like to give heartfelt thanks to everyone in my group. Without their kind help and support, it would be much more difficult for me to finish all the data analysis.

Finally, I would like to thank my husband Hao Yu and my dear daughter Qinxuan Yu for their love and support through my PhD time.

Chapter 1

Introduction

Cancer is one of the most arduous challenges faced by the human race. In 2018, there were 18.1 million new cancer cases and 9.6 million death cases due to cancer all over the world [1]. Three most common cancer types for both sexes were lung cancer (11.6% of the total cases), breast cancer (11.5%), and prostate cancer (7.1%). Among them, lung cancer was the leading cause of death among males followed by prostate cancer while breast cancer was the most commonly diagnosed and leading death cause for females [1,2].

The survival rate of cancer is largely related to early diagnosis. In the current cancer diagnosis flow, remained largely unchanged for over 140 years [3], in order to diagnose cancer, in most cases, biopsies are required. Each biopsy taken from the patient is manually examined by pathologists, who make decision about whether cancer occurred in the biopsy. The examination process involves possible preparation artefacts, unusual cases and subjective judgments about the disease severity [4,5]. In addition to that, the whole process is time-consuming and only has limited outputs.

To avoid the problem and improve the reliability of cancer diagnosis, developing computer-based automatic cancer diagnosis approaches which limit unnecessary human errors is essential. Such approaches, which can reduce variance in diagnosis and financial burden on the public health system, can be used to provide a second opinion for patients or as a pre-screen tool for pathologists. Vibrational spectroscopic imaging techniques are competitive candidates as automatic cancer diagnosis approaches. Between Raman and infrared tissue imaging, infrared holds its advantage in large scan area and short data collection time for the same size of measuring area. Many research has been conducted using infrared tissue imaging approach to distinguish cancerous and normal samples, involving prostate [6–8], lung [9], colon [10], breast [11–13] tissues. With the advantages in time spending and diagnosis accuracy, infrared imaging is a potentially powerful tool for clinical use.

In addition to cancer diagnosis, with the collected spectra of sample tissues for cancer diagnosis, digital histopathology can be performed without extra biopsy or additional measurements. Without any physical staining, based on the biological information collected from sample with the combination of classification or pattern recognition algorithms digital staining can be performed [3,5]. Unlike the traditional biomarker staining methods, multiple stained images of sample can be generated based on the same set of data.

To achieve cancer diagnosis classification and histological pattern recognition, reasonable mathematical methods need to be applied. State-of-art data mining method, machine learning algorithms, is vital for the idea of developing computer-based automatic cancer diagnosis and histological pattern recognition approaches. Multiple groups have already introduced machine learning methods into the field [3,8,14–16]. However, only a few machine learning approaches were applied in the previous research. With more algorithm introduced into the field, choosing the most suitable methods for each dataset, models based on infrared spectra will be more robust and accurate.

Inspired by previously researches conducted on infrared tissue imaging combining with machine learning methods, in this thesis, breast and prostate cancer diagnosis and digital histopathology studies were discussed using FTIR imaging system. For breast cancer diagnosis, tissues cores from the same patients on different types of substrates were applied to further validate the possibility of introducing the FTIR imaging technique into the current clinical cancer diagnosis working flow using the slides currently used by pathologists. Combining with the popular machine learning algorithms, classification models on digital histology and cancer diagnosis were constructed based on spectra collected. In addition to that, whether using stained marker would improve the classification accuracy of cancer diagnosis was also discussed. In term of prostate cancer diagnosis, unstained prostate tissue on glass was applied to construct classification models using machine learning algorithms to validate the possibility of establishing cancer diagnosis models by directly measuring current used pathological slides.

1.1. Reference

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA. Cancer J. Clin. (2018). doi:10.3322/caac.21492.

[2] S. John, J. Broggio, Cancer survival in England : national estimates for patients followed up to 2017, (2019) 1–27.

[3] D.C. Fernandez, R. Bhargava, S.M. Hewitt, I.W. Levin, Infrared spectroscopic imaging for histopathologic recognition, Nat. Biotechnol. (2005). doi:10.1038/nbt1080.

[4] A. De la Taille, A. Viellefond, N. Berger, E. Boucher, M. De Fromont, A. Fondimare, V. Moliníé, D. Piron, M. Sibony, F. Staroz, M. Triller, E. Peltier, N. Thiounn, M.A. Rubin, Evaluation of the interobserver reproducibility of Gleason grading of prostatic adenocarcinoma using tissue microarrays, Hum. Pathol. (2003). doi:10.1016/S0046-8177(03)00123-0.

[5] R. Bhargava, Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology, Anal. Bioanal. Chem. (2007). doi:10.1007/s00216-007-1511-9.

[6] E. Gazi, M. Baker, J. Dwyer, N.P. Lockyer, P. Gardner, J.H. Shanks, R.S. Reeve, C.A. Hart, N.W. Clarke, M.D. Brown, A Correlation of FTIR Spectra Derived from Prostate Cancer Biopsies with Gleason Grade and Tumour Stage, Eur. Urol. (2006). doi:10.1016/j.eururo.2006.03.031.

[7] M.J. Baker, E. Gazi, M.D. Brown, J.H. Shanks, N.W. Clarke, P. Gardner, Investigating FTIR based histopathology for the diagnosis of prostate cancer, J. Biophotonics. (2009). doi:10.1002/jbio.200810062.

[8] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, Analyst. 142 (2017) 1258–1268. doi:10.1039/c6an02224c.

[9] X. Mu, M. Kon, A. Ergin, S. Remiszewski, A. Akalin, C.M. Thompson, M. Diem, Statistical analysis of a lung cancer spectral histopathology (SHP) data set, Analyst. (2015). doi:10.1039/c4an01832j.

[10] C. Kuepper, F. Großerueschkamp, A. Kallenbach-thieltges, A. Mosig, Label-free classi fi cation of colon cancer grading using infrared spectral histopathology, Faraday Discuss. 187 (2016) 105–118. doi:10.1039/C5FD00157A.

[11] H. Fabian, N.A.N. Thi, M. Eiden, P. Lasch, J. Schmitt, D. Naumann, Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy, Biochim. Biophys. Acta - Biomembr. (2006). doi:10.1016/j.bbamem.2006.05.015.

[12] D.M. Mayerich, M. Walsh, A. Kadjacsy-Balla, S. Mittal, R. Bhargava, Breast histopathology using random decision forests-based classification of infrared spectroscopic imaging data, Proc. SPIE. 9041 (2014) 904107. doi:10.1117/12.2043783.

[13] P. Bassan, M.J. Weida, J. Rowlette, P. Gardner, Large scale infrared imaging of tissue micro arrays (TMAs) using a tunable Quantum Cascade Laser (QCL) based

microscope, Analyst. (2014). doi:10.1039/c4an00638k.

[14] S. Mittal, T.P. Wrobel, L.S. Leslie, A. Kadjacsy-balla, A four class model for digital breast histopathology using High- Definition Fourier transform infrared (FT-IR) spectroscopic imaging, 9791 (n.d.) 1–8. doi:10.1117/12.2217358.

[15] T.P. Wrobel, J.T. Kwak, A. Kadjacsy-balla, R. Bhargava, High-Definition Fourier Transform Infrared Spectroscopic Imaging of Prostate Tissue, 9791 (n.d.) 5–11. doi:10.1117/12.2217341.

[16] P. Bassan, J. Mellor, J. Shapiro, K.J. Williams, M.P. Lisanti, P. Gardner, Transmission FT-IR chemical imaging on glass substrates: Applications in infrared spectral histopathology, Anal. Chem. (2014). doi:10.1021/ac403412n.

Chapter 2

Literature survey: The use of hyperspectral imaging for infrared spectral histopathology

2.1. The 1990s

Research using the combination of tissue and FTIR has been conducted for around 60 years. The very first attempt was performed by Blout and Mellors [1]. A paper was published which concluded that the sensitivity of infrared techniques was not enough for tissue differentiation and identification. During the next 40 years, scientists focused on protein secondary structural motifs and the DNA double-helical structure using infrared [2,3]. Molecular biology and disease drew the most of attention [4].

In 1995, Fabian *et al*. [5] first successfully applied infrared spectroscopy in finding differences between xenograft and human tumour cells. Infrared was applied in distinguishing different cell lines which can be considered as the foundation of infrared histology. In 1998, Chiriboga *et al.* published their work on detection of cell maturation in cervical tissue [6]. In the research, tissue architecture was correlated with the collected spectra through pathological analysis. Spectra of different epithelium layers, representing different cellular maturation stages, from cervical tissue, were collected. Differences among layers were observed. The authors concluded that cell maturation and differentiation in cervical tissue can be monitored through their cell life cycles. Furthermore, spectra from individual layers can be used as a health inspection reference for exfoliated cells. This was the first time, which infrared spectroscopy was proved to be a powerful tool in tissue investigations. In the same year, Lasch and Naumann also published their work on colon tissue [7]. Unstained human melanoma and colon carcinoma tissues were measured on an infrared microscope with an automatic stage moving in x and y directions. For the very first time, the obtained spectra were analysed using clustering methods and principal component analysis. Infrared images were constructed based on the results of recombining classifiers with spatial information of each spectrum. This method was considered to be helpful in improving discriminations among types of structures in the tissue. The constructed false colour infrared images were compared with pathological results, which showed the outstanding sensitivity of infrared microscopy as a diagnostic tool for variations in the tissue.

From that time, infrared imaging was considered having the potential to obtain the same information produced by pathologists through tissue staining. Automatically detect changes in chemical compositions without stains and special preparations were considered to be the future goal of the field.

In terms of instrumentation, in the 1990s, FTIR spectrometers [5] and FTIR microscopes [6] were both used. Spectrometers with both DTGS and HgCdTe were both used for research [4]. For the FTIR microscope, a computerized microscope stage and visual capture software were added [4,6,7]. In that time, using the infrared mapping data collection method, general measuring size of 50x50 μm^2 could be covered with a combination of automatic microscope stage and predetermined coordination. In alternative, infrared microscopes with confocal array detectors could simultaneously acquire thousands of spectra arrayed pixels [8].

In 1999, a review paper was published, in which a data treatment routine was proposed [4]. The data processing process included water vapour correction, optional smoothing, background correction and removal of residual substrate spectral features. Possible data scaling was desired based on the following tests, as there were subtle differences in the concentration of cellular components. If the thickness of the samples were considered to be uniform, the scaling might not be necessary. A normalization based on the intensity of amide I was considered as a solution to the scaling problem.

2.2. The 2000s

Research into using infrared in tissue histology and disease diagnosis really began in the 1990s. Although there were technical limitations, hopes and expectations had been raised. From 2000 to 2010, people worked in two main directions, digital histology and cancer diagnosis.

Several tissues were analysed through digital histology. In 2000, Diem *et al*. [9] demonstrated the analysis of cirrhotic liver tissue spectra comparing with the normal liver tissue by FTIR and infrared spectral mapping. The results presented by using infrared spectral mapping were in good agreement with the histopathological information. In their research, spectra were pre-processed using water vapour subtraction, background correction, smoothing and the rubber band algorithm baseline correction. The processed spectra were stored in a compact format suitable for subsequent calculations. The processed spectra were then correlated and clustered using similar cluster computational method to principal component analysis (PCA) producing a coloured infrared map and compared with histological figures of liver samples. The analysis methods applied were considered to be cutting-edge at that time.

With the foundation created from previous research, Bhargava *et al*. published a paper on automated stainless histology on prostate tissue with well-defined tests of statistical significance in 2005 [10], which separated epithelium and stroma cells and outputted accurate histologic core images. The areas under curve (AUC) of individual receiver operating characteristic (ROC) curves were calculated for each class. AUC values for epithelium (0.994), stone (0.991), ganglion (0.954), nerve (0.971), lymphocytes (0.990), blood (0.932) and endothelium (0.983) display a high segmentation capability. Stroma subtypes were 0.983, 0.961 and 0.894 for fibrous stroma, mixed stroma and smooth muscle, respectively. AUC and ROC were first introduced to the field, in which they are still commonly used nowadays. This research provided ideas both in data processing of tissue spectra classification and stainless classification of tissue samples which could be a breakthrough regarding clinical application of tissue analysis and molecular stainless histopathology.

In terms of cancer diagnosis, groups of researchers started their cancer diagnosis research in these ten years with many types of human tissue being studied including prostate and breast. For prostate tissue, exciting results have been published in diagnosis and cancer classification [11].

In 2003, Gazi *et al*. [12] separated prostate cancer cell lines derived from different metastatic sites using FTIR spectroscopy for the first time. The ratio of peak areas at 1030 and 1080 cm^{-1} was found corresponding to the glycogen and phosphate vibrations, which could potentially be used to identify malignant cells. With a linear discriminant algorithm, classification of benign and cancerous tissue was conducted. In 2005, FTIR imaging of microarrays was coupled with statistical pattern recognition techniques in order to demonstrate histopathologic characterization of prostatic tissue and to differentiate benign from malignant prostatic epithelium.

Fernandez *et al*. [10] tried separation of cancerous and normal prostate epithelium cells from the same patients. 50 samples with both cancerous and normal matching sets were used for the research. 100% discrimination was observed with patient-matched normal and malignant acini when 20% threshold was applied. Cancerous tissue was considered contain more variation than normal tissue. Both these studies showed that spectroscopic imaging has a promising future for cancer diagnosis and potential prognosis.

In 2008, Baker *et al*. [13] published a paper on FTIR cancer diagnosis of prostate cancer. This paper testified the potential of using FTIR as a prostate cancer diagnosis tool for clinical usage. By combining with a principal component discriminant function analysis (PC-DFA) algorithm, overall sensitivity of 92.3% and specificity of 99.4% grading classification can be achieved in a three-band Gleason score criterion diagnosis. In addition, they also correlated FTIR spectral characteristics with clinically aggressive behaviour in prostate adenocarcinoma (CaP) manifest as local and/or distal spread for the first time. Apart from further validation of the possibility of using FTIR as a cancer diagnosis tool, this paper provided the possibility of correlating spectra to the cancer metastasis.

Focusing on breast tissue, in 2001, Eckel *et al*.[14] published a paper on characteristic infrared spectroscopic patterns in the protein bands of human breast cancer tissue. They collected spectra from 96 patients who had breast hyperplasia, fibroadenoma and carcinoma, and compared them with normal breast tissues and each cancer type itself. This research provides spectral characteristics of different breast cancer types, which were later used as references for breast cancer features or FTIR biomarkers.

FTIR imaging was applied to malignant and benign breast tumour tissue sections [15–17]. In 2003, Infrared microspectroscopic imaging had been applied to analysis breast cancer tissue by Fabian *et al*.[17] the functional group mapping and cluster analysis were applied to process data. Infrared images were produced and compared with the stained tissue images. This paper focused on differentiating different tumour types of breast cancer. The identification of fibroadenoma caused a problem because of spectral averaging of all tissue components present in the corresponding micro-areas. In 2006, Fabian *et al*. [15] classified benign lesions fibroadenoma from 14 patients, malignant ductal carcinoma in situ from 8 patients, connective tissue and adipose tissue by artificial neural network analysis. A four-class classifier was constructed. After training the classifier, an independent test was conducted. All spectra (a total of 386) taken from micro areas inside the epithelium of fibroadenomas from 4 patients were correctly classified. Out of the 421 spectra taken from micro areas of the in situ component of invasive ductal carcinomas of 3 patients, 93% were correctly identified. Although the number of patients in this study would now be considered as being very low the works demonstrated the ability to classify tumour types.

These papers revealed that breast tissue micro-heterogeneity is a particularly severe problem for FTIR microspectroscopy and that high-quality spectra with the high spatial resolution are required in order to detect subtle, cancer-related alterations in the biochemical and morphological composition of tissue at the microscopic level. These data also suggested that earlier reported spectral changes between malignant and normal tissues which were observed in single spectra at higher spatial resolution may have an alternative explanation. A significantly extended database of spectra of histopathologically well-defined tissues, which spans the natural variability of healthy and tumour breast tissues, is required in order to fully explore the possibilities of IR microspectroscopy.

Reviews [18–21] were published for further summarization and education of the field. These reviews address instrumentation, the applicability of various systems, spectroscopic bases and classification algorithms for decision making, and controversial aspects in the backdrop of the evolution of the field.

Protocols concerning cancer histopathology [11] and data processing [20] were published. In Bhargava’s paper [11], approaches to optimal data acquisition, classification and validation were all discussed. A systemic method was proposed with each component discussed quantitatively with the effects on the final classification results. Variation of spatial resolution, spectral resolution and signal to noise ratio were all considered in the paper. Prostate tissue was used as an illustrative example in developing protocols for automatic cancer histopathology. From paper published by Diem *et al*. [20] a protocol dealing with data analysis was proposed after a review of development of the field from 1994 to 2004. In the paper, a uniform data pre-processing protocol was used for all computational procedures. First, the data set is cropped in both spatial and spectral dimension (typically between 1800–800 cm^{-1}, since at the time this was considered to be the main diagnostic region). Next, pixels with too high or too low absorbance values, or with poor signal/noise ratio, are removed from the data set. The remaining spectra are then smoothed or derivatized using a Savitsky–Golay algorithm, and baseline corrected within this region. Finally, all spectra were vector normalized between 1800 and 800 cm^{-1} [20].

2.3. The 2010s

More research has been conducted from 2010 to 2019 with better-developed procedures analysing underlain information from infrared spectra [22–24].Apart from focusing more on biological differences, people started focusing more on the analysis methods. Machine learning and artificial neural network became the main methods of data analysis. With more and more conferences and field networking, protocols [10,25–29] of research combining FTIR and biological samples were increasingly clear. Experimenting and analysing methods were more systematic and standardized.

Despite many advances in the biological application of FTIR spectroscopy, there remain challenges in sample preparation, instrumentation and data handling.

2.3.1. Review on sample preparation methods

In terms of sample preparation, the main sample formats are fixed cells, biofluids, live cells and tissue. In the case of tissue samples, formalin fixed paraffin embedded (FFPE) is considered as one of the most common methods to preserve both the biochemical components and tissue architecture. Sample thickness of FFPE tissue should not exceed 8–12 μm to avoid total absorption. Before infrared measuring, FFPE tissue could be de-waxed for a minimum of 5 min in xylene with three washes performed [26]. Not all dewaxing is done with Xylene, some use hexane or commercial agents such as cytoclaer [30]. It is noticed that using solvent to remove paraffin can potentially leach out lipids native to the tissue samples, which questions the reliability of subsequent studies on the remaining lipids. Some research groups claimed that using FFPE sections without dewaxing does not affect the discriminative potential of the technique, in terms of using FTIR imaging [31–34]. Extended multiplicative signal correction (EMSC) algorithm can be used to correct the influence caused by FFPE [31,33]. In addition, using the wax embedded sample results in less spectral scattering since the refractive index change between wax and tissue is smaller than air and tissue. Considering the advantages of directly using FFPR tissue sections, in the research described in this thesis, all tissue samples were directly measured without de-waxing. The paraffin affected ranges were removed in data processing steps.

For other tissue sample types, for example, cryosectioned tissue samples, the sample must be thoroughly thawed before IR analysis. Once a sample is thawed, components may start to degrade [35]. To avoid excessive degradation, the data collection process should start as soon as possible after sample thawing. A dark environment helps to slow the effect of sample degradation for both sample drying and data collection. Under dry conditions, cryosections can be stored for months. However, the lipid in tissue starts oxidation after two weeks, which can be avoided by storing in an N_2 atmosphere [26].

In term of substrates, for transmission, the most commonly used substrates are BaF_2 and CaF_2. For reflection, IR-reflective substrates, which are usually Ag/SnO_2 coated glass slides, are used. Comparing with substrates used in transmission, low emissivity slides are much cheaper and robust. However, studies [36–39] have suggested that the Electric Field Standing Wave (EFSW) effect, which is a spectral distortion, can be raised in transflection model of operation. In 2015, Pilling *et al*. [40] published comparison of chemical images measured in transmission and transflection modes and found in transflection spectra undergo a non-linear distortion. Significant differences in spectra measured from the same area of tissue depending on the mode of operation were observed. Principal Component Analysis (PCA) highlighted the poorer discrimination between benign and cancerous tissue in transflection mode. To better implemented infrared microscopy in the clinical environment, other cheaper and reliable substrates should be introduced.

Recently, conventional glass was tested as a new substrate for transmission by Bassan *et al*. [41] and Pilling *et al*. [42]. It was believed that combining with fast infrared measurement, using conventional glass substrates could help the technology fit better in the current workflow of cancer diagnosis by pre-screening slides already used in clinics to reduce the workload of pathologists.

In 2014, Bassan *et al*.[41] published their research on infrared spectral histopathology of human Breast tissue with glass substrates in transmission mode. The glass substrate had a high-wavenumber transmission window allowing measurement of the C-H, N-H, and O-H stretches in the range of 2500-3800 cm^{-1}. FT-IR chemical images were measured at a spectral resolution of 8 cm^{-1} using 32 and 8 coadded scans for the background and sample, respectively. A breast tissue microarray (TMA) containing cores from 71 patients which remained embedded in paraffin on the glass slide was used to construct a model which classified four basic tissue cell types. Collected spectra were baseline corrected between 3100 and 3600 cm^{-1} and then normalised to the absorbance value at 3298 cm^{-1}. Using random forest classification algorithm, with a probability estimate acceptance threshold of 0.95 (95%), the correctly classified percentage for the independent test data set was epithelium = 98.25%, stroma = 99.94%, blood = 100.00%, and necrosis = 97.22%. They also used epithelial cells to discriminate normal and cancerous tissue and obtained a reasonable separation using PCA.

In the light of research of Bassan *et al*., in 2017, Pilling *et al*. [42] published their work on Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides. In the paper, they presented a study using results obtained from an extended patient sample set consisting of 182 prostate tissue cores, including 95 cancerous tissue samples and 87 NAT samples. These cores were obtained from 100 patients from 18 H&E slides. The microscope utilises ×15 Cassegrain optics with a corresponding. The hyperspectral images were obtained using a pixel size of 5.5 µm at 5 cm^{-1} resolution with 96 and 256 for the sample and background scans respectively. The spectra were quality checked on the intensity level of amide A band (3298 cm^{-1}). Spectra with an intensity range of 0.1-1 were kept. Only wavenumbers from 3125 to 3700 cm^{-1} were kept from further processing and classification. Random forest classification was followed by PCA noise reduction (15 PC kept), vector normalisation and first derivative using 19 point Savitzky–Golay smoothing. Four main classes of histology were classified and tested. A high degree of accuracy (>90%) was observed for each class with a probability of acceptance threshold of 0.6. Nine separate prostate sections with different degrees of staining were investigated to ensure the discrimination observed was on biomarkers, not the presence of stain. Finally, using four-class model discrimination among the normal epithelium, malignant epithelium, normal stroma and cancer-associated stroma was observed with classification accuracies over 95% using a probability of acceptance threshold of 0.5. This research pointed towards the possibility of implementing automated pre-screening of samples in the clinic.

2.3.2. Review on the instrumentation of Fourier transform infrared spectrometry

In terms of instrumentation, detectors have had a vast influence on the quality of spectra collected. Detectors can be generally divided into three groups, including single-element, linear array and focal plane array (FPA) detectors. In term of the single-detector instruments, infrared radiation impinges on a prescribed spatial area through a set of condensing optics passing from the rapid-scan interferometer after modulation. The infrared images of this area are obtained by controlling the size of the aperture. An optical image will also be collected through the optical path. Single-detector microscope works well for studies on small-sized samples. However, it has limited utility in collecting data from large and heterogeneous samples. To cope with this disadvantage, mapping spectra collecting methods and corresponding instruments were developed. Mapping is achieved by adding an automated microscope stage which is capable of precise movements to maintain registration between the same point through observation and collection. Larger aperture size and number of acquisition scans can both improve the data quality.

However, the spectra obtained will be in a trade-off between collecting time and data quality [18].

Linear array detectors are a row of detector elements defining a rectangular spatial region on the sample. It can be used to measure relatively large-sized sample with precise and sequential movements. It is also known as push-broom mapping or raster scanning [18]. In term of imaging, linear array detectors have a similar concept with the mapping process using a single detector, but it has an advantage in multiple channels of detection. The characteristics of multiple detecting elements bring several benefits, including faster speed for data collection and better signal to noise ratio compared with single detector instruments. Also, a linear array of detectors eliminates the need for geometrical apertures. Optics in the system determines the nominal spatial resolution. Furthermore, how precise the microscope stage move is crucial to the imaging results, as the final image consists of steps in small increments [18].

Nowadays, the most advanced detector in the FTIR microspectroscopic instrumentation is an FPA detector [43], which contains several thousand individual detecting elements forming a two-dimensional detecting plane. It further increases the advantage of multiple detecting channels compared with linear arrays. For a 128 x 128 FPA detector, it is 16384 times better than a single detector and 1024 times better compared with a 16-element linear array in term of the multichannel advantages. It enables the imaging of the whole field of view. Among the available infrared micro-spectroscopic methods, FPA-based global imaging represents the most comprehensive approach for recording spectra over large sample areas. The spatial resolution of an FPA based instrument is determined by both the optics which determines the magnification and the size of each element of the FPA. Even though the nominal spatial resolution is determined by detector and optics of the system, the actual resolution limit is still governed by the diffraction limit [26]. With an interferometer, the advantage of FPA is further enlarged in spectra acquiring time as the time cost of data collection throughout the full field of view is almost the same as taking one spectrum at only one point on the sample using a single

detector. A typical FPA detector chip is a few millimetres in size with each element in tens of microns [18,44].
Detector arrays are usually made of several types of infrared-sensitive materials. For mid-IR imaging, MCT FPAs [45] has been the most popular. This is due to its high sensitivity and good spectral range.

Light sources are also one of the most essential elements in term of the instrumentation of FTIR. The conventional thermal (globar) is by far the most common light source used for FITR instruments. This is particularly the case when doing FTIR microscopy using an FPA detector. For FTIR microscopy Synchrotron radiation can be coupled to the FTIR. Alternate light sources including QCLs and filters can also be used for infrared spectroscopy but these discrete frequency sources are not used in combination with FTIR [26]. The most commonly used light source is a globar source, which contents the silicon carbide rod generating infrared radiation. With the correct optics it can A synchrotron radiation facility consists of a particle accelerator and electron storage ring [46]. Electrons are emitted from the electron gun and are accelerated in the particle accelerator consisting of a linear accelerator and ring booster. When the electrons nearly research the speed of light, electromagnetic radiation will be produced. Most of the radiation produced is in the UV and X-ray region of the spectrum (depending on the energy of the ring but the radiation extends into the infrared. The IR radiation produced is 100 to 1000 times brighter compared with the conventional globar and concentrated in a small area. Therefore, synchrotron sources have natural spot size at the sample of 10 to 20 microns and can generate high signal to noise spectra [47]. It enables researchers to obtain information at the single-cell level or even nucleus level [48,49].

There are around 50 synchrotron facilities all over the world [50], and as of 2018, there are 28 IR beamlines most of which are currently operational [51]. In 2006 independent research groups [52–54] firstly combined single synchrotron beam with FPA detector. Improved signal to noise ratios were observed in a local region of the FPA detector compared with a conventional globar. They suggested that due to the inhomogeneous illumination of the FPA, small FPAs (and sample area) should be used for other studies in the future. In 2011, Nasse *et al.* [44] used multiple synchrotron beams during data collection. High spatial-resolution images of the interested area were obtained using a pixel size of 0.54 × 0.54 microns. They claimed that the results were around 100 times better compared with other instrumentation at that time with the lowest detection limit. After that, reviews published by Miller and Dumas [54] and Hirschmugl *et al.* [55] described the development, advantages in biological applications and potential future directions of synchrotron-based FTIR. Not only studies on cells, but also researches have been conducted on tissues, where each cell could be probed in subcellular resolution. For example, the structure of misfolded protein aggregates has been identified in the brain tissue of Alzheimer's disease patients [56,57], infectious prion proteins have been characterized in scrapie [58] and variations in bone composition have been observed in osteoporosis [59]. Synchrotron radiation was also adopted as a light source with FTIR microscope spectrometer for studying live cells [60–62] with different prototypes of IR compatible microfluidic devices.

Another alternative light source for infrared spectra collection is Quantum Cascade Lasers (QCLs), which were a new type of semiconductor laser. Because of their characteristics of narrow line widths [63,64], they have been used as gas sensors. With the improvement in wavelength range, modern tuneable QCLs was first applied as a light source for mid-infrared in 2007 [65]. Images obtained by both QCL and globar FTIR spectrometer were compared. The advantages and limitations of the new laser resource were explored in the study [66]. Further application on biological samples was conducted in 2013 by Liakat *et al*. [67]. Study of glucose concentration in biological fluids was presented. With the combination of mathematical model, the concentration of glucose in the fluid was successfully predicted using QCL as light source in transmission mode. A following the study on the glucose *in vitro*, the level of glucose in humans was measured by QCL fibre probe [68]. Spectra data were collected by illuminating light on human palm and only 2% error rate was found after comparing the detected results with the commercially used glucose detector. In principal therefore infrared glucose monitoring could be put into use but the cost of QCLs compared with other methods of blood testing at present makes this unlikely. In light of previous studies, Bassan *et al*. [69] applied QCL on human tissue. In 2014, image of a breast tissue array was presented by the QCL microscope system, which covered range of 900-1800 cm^{-1} with a 480x480 elements un-cooled micro-bolometer detector. It was reported that a tissue image at a single wavenumber of a whole TMA with the size of 20x24 mm^2 was acquired within 9 min, which was 126 times faster compared with a conventional FTIR instrument. This, of course, is not a like for like comparison since the FTIR generates a hyperspectral image but is useful if the wavelength of interest is known. Later, Clemens *et al*. [70] also published tissue images produced by application of QCL. Later that year, Pilling *et al*. utilised QCL imaging microscope with both prostate [71] and breast cancer [72] diagnosis, between 900 and 1800 cm^{-1} with a 480 × 480 un-cooled microbolometer FPA. In terms of prostate tissue digital histology, sensitivity and specificity of 93.39% and 94.72% respectively were presented with only 25 discrete frequency collections at key diagnostic wavenumbers. In the independent test, 72.14% sensitivity and 80.23% specificity were obtained. This research showed that discrete frequency infrared chemical imaging has the potential to provide high-resolution, high-throughput chemical images on a timescale that could revolutionise spectral histopathology [71]. Regarding the research on breast tissue, high classification rates, sensitivity (93.56%) and specificity (85.64%), for cancer and NAT tissue were observed in an independent test with continuous spectra acquired between 912 and 1800 cm^{-1}. Classification accuracy obtained had similar sensitivity and specificity using conventional FTIR imaging [72].

In order to obtain the best signal to noise ratio, several optimised instrumental and operational settings are suggested by Baker. *et al*. [26] in 2014. Single-element detector required a larger number of scans to achieve similar signal to noise ratio compared with FPA detector. For FTIR single-element detector the optimal number of co-additions scan is 512, while for FPA detector with the same globar source, the suggested number of co-additions is 64 or 128 scans. For both single and FPA detector, the optimal spectral resolution is 4 or 8 cm^{-1}, although it is possible to use numbers other than multiples of two. In much of the work in the Gardner group 5 cm^{-1} is often used [26]. For single-element detector, signal to noise ratio can be influenced by aperture size of the instrument during point or mapping measurements while interferometer mirror velocity may cause effect on signal to noise ratio when it comes to FPA imaging.

In general, the square root of the number of co-additions is proportional to the signal to noise ratio, and therefore an increased number of scans will enhance the signal to noise ratio [73]. IR spectroscopy has a spatial resolution that is limited by the diffraction limit. Therefore, when the resolution approaches the limit, the signal to noise ratio is reduced to a point where there is no further gain in image quality [74]. The choice of the type of detector (for example, thermal detector and quantum detector) also effects on the signal to noise ratio. A mercury cadmium telluride (MCT) quantum detector usually provides a superior signal to noise ratio than thermal detectors (usually deuterated triglycine sulfate detectors). An optimized cooling system in the detector, such as thermo-electrical cooling, will also reduce the dark current produced by the detector, which has been shown to have a detrimental effect on the signal to noise ratio [75].

2.3.3. Review on data process methods

In terms of data analysis methodology, there are generally four typical data processing goals, including pattern finding, biomarker identification, imaging and diagnosis. In order to achieve these objectives, techniques including quality control, pre-processing, feature extraction, clustering and classification have been applied based on the conditions of raw data in the various studies.

Quality control is commonly the first step of data analysis, during which outliers were identified and removed from the spectra dataset. Going through literature, many quality control methods have been proposed including signal-to-noise test [76,77], thickness test [76–78], and maximum and minimum absorbance threshold [42,69,72].

In pre-processing steps, usually, spectra will be denoised, baseline corrected, removing variation caused by difference in thickness or concentration, and removing sample contaminants (e.g., water, paraffin). Afterwards, first or secondary derivate of data can be obtained additionally to further highlight features of spectra. Regarding denoising, PCA denoising [69,72]and Savitzky–Golay Smoothing [79,80] can both be applied to the dataset. PCA denoising was claimed to have the best effect for removing noise [81], while the Savitzky-Golay method is more focused on smoothing the first derivative of spectra. In term of baseline correction, differentiation in 1st and 2nd order [79,80,82,83] and rubber band baseline correction [9,22,23,84] are mostly used in field. Extended multiplicative signal correction [82] and resonant Mie scattering correction [85,86] are used for further fitting of spectra. The effect, intensity difference crossing sample, caused by thickness variations is usually corrected by normalisation methods. Typical methods involved normalisation to the Amide I peak [22,84,87] and vector normalization [13,23,42,69,72]. The sequence and choices of pre-processing methods should be picked depending on the data and purpose of studies. Some methods are specifically designed to solve a certain problem. Recently, studies have been conducted on classification-based pre-processing steps for a given dataset [88], for example, feature selection. In terms of feature extraction, principal component analysis [80,82,87,89], partial least squares [88,90], linear discriminant analysis [82] and PCA-LDA [13,23,82] could all be used to select features from data set.

These methods all generate loadings (features generated by a linear operator), factors (features), scores (value of each factor) and scores plots (visualised presentation of scores). Loadings carried information about the contribution of each wavenumber to the feature pattern. The differences among these methods are about ways of obtaining the loading matrices [88].

Clustering is an unsupervised pattern recognition methodology. K-means [9,77,91], Fuzzy *c*-means [77], and HCA [20,76,77,83,87] are commonly used clustering methods in the field. These methods were designed to cluster the spectra into groups, based on the similarity between spectra. Cluster analysis is exploratory and used to find if there are hidden structures in the dataset, and whether these structures (clusters) correlate with data classes or other physical knowledge (*e.g.*, tissue structure) [88].

Classification is usually conducted by building up a mathematical model which predicts the class of an unknown sample based on the information previously studied from a similar type of data, known as training data. Several models are usually used in the field nowadays, including multiple linear classifiers, support vector machine (SVM), Artificial Neural Network (ANN) and ensemble classification models. In the case of linear classifiers, for example, linear discriminant analysis (LDA) and partial least squares discriminant analysis (PLS-DA) [92], they have simple structure and easy to construct. However, they are not complicated enough to accurately finding boundaries between classes [88]. SVM, which can be either linear or kernel-based, is a family of classifiers increasingly popular in bioinformatics, where non-linear decision boundary between classes can be found. Multiple studies using SVM as a classification method have been conducted. For example, Kelly *et al.* used SVM in classification of low-grade cervical cytology in 2010 [22], and Baker *et al*. [82] applied SVM as one of their approaches to finding differences among RWPE human prostate epithelial cell lines from the same source but differ in their mode of transformation and their mode of invasive phenotype. In addition, Sattlecker *et al*. [93] used SVM for breast cancer type predications in 2011, which obtained 66.7% predication accuracy with single classifier while 88.9% correct with ensemble system in the independent test.

ANN is widely used as a classification method. An ANN is a computational model based on the structure and functions of biological neural networks [88]. The artificial neural network is a branch of deep learning, a division of machine learning. Machine learning algorithms use computational methods to teach computers learning from experiences like humans. Information is directly learnt from data without predetermined equations as models. Deep learning is sufficiently appropriate for image recognition, which has been used widely for facial recognition, motion detection and driver assistance technologies [94]. Deeping learning uses neural networks to learn the characteristics of features directly from data with multiple nonlinear processing layers. The parallel working structure was inspired by biological nervous systems. In an example of applying ANN in the research, Kelly *et al*. classified low-grade cervical cytology with a classification accuracy of around 80% per spectra [22].

Classifier ensembles have been increasingly used recently [42,95,96]. A combination of multiple single model classifiers is considered as a more accurate option when it comes to classification problems. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions [77]. Aggregation strategies include Bagging [97], Boosting [97,98], Hierarchical Aggregation [99], and multiple two-class Classifiers [100]. There are studies using ensembles to successfully overcome the instability that results from a small dataset being available. Among these types of methods, random forests have been most commonly used [42,71,72,101–103]. Random forest is a classification method which operates by constructing decision trees to vote on the class membership [104,105]. Each tree gives a classification. Each decision is collected and the forest chooses the most popular class-based on decisions made by all trees in the forest [105]. This algorithm was developed by Breiman [104] in 1996 based on bagging predictors [106] with a random selection of features [107] to construct decision trees with higher accuracy of classification.

To sum up, variable researches have been conducted since the 1990s in term of application of infrared spectroscopy on biological samples. Even though systemic steps and protocols in many aspects of the research have been published, still, there are gaps to fill and problem to solve in the field. To achieve the objective of implementation of infrared into the clinical environment, there are still barriers to overcome, ideas to validate and further studies to conduct.

In the light of previous research, in this thesis, both the validation of previously proposed ideas and further research about new ideas in term of infrared clinical transition is discussed. Both standard and advanced data analysis methods in the field are applied to extract biological information from samples. Thanks for the effort of previous researchers, using their experiences and knowledge avoidable errors and mistakes in this thesis can be minimised, as from sample selection to final data analysis, steps were conducted carefully following previous protocols.

2.4. References

[1] E.R. Blout, R.C. Mellors, Infrared spectra of tissues, Science (80-.). (1949). doi:10.1126/science.110.2849.137.

[2] S. Boydston-White, T. Gopen, S. Houser, J. Bargonetti, M. Diem, Infrared spectroscopy of human tissue. V. Infrared spectroscopic studies of Myeloid Leukemia (ML-1) cells at different phases of the cell cycle, Biospectroscopy. (1999). doi:10.1002/(SICI)1520-6343(1999)5:4<219::AID-BSPY2>3.0.CO;2-O.

[3] J.R. Durig, Chemical, biological, and industrial applications of infrared spectroscopy, Wiley, 1985. https://books.google.co.uk/books/about/Chemical_Biological_and_Industrial_Appli.html?id=sVLwAAAAMAAJ&redir_esc=y (accessed September 11, 2019).

[4] M. Diem, S. Boydston-White, L. Chiriboga, Infrared spectroscopy of cells and tissues: shining light onto a novel subject, Appl. Spectrosc. (1999). doi:10.1366/0003702991946712.

[5] H. Fabian, M. Jackson, L. Murphy, P.H. Watson, I. Fichtner, H.H. Mantsch, A comparative infrared spectroscopic study of human breast tumors and breast tumor cell xenografts, Biospectroscopy. (1995). doi:10.1002/bspy.350010106.

[6] L. Chiriboga, P. Xie, V. Vigorita, D. Zarou, D. Zakim, M. Diem, Infrared spectroscopy of human tissue. II. A comparative study of spectra of biopsies of cervical squamous epithelium and of exfoliated cervical cells., Biospectroscopy. (1998). doi:10.1002/(SICI)1520-6343(1998)4:1&lt;55::AID-BSPY6&gt;3.0.CO;2-R.

[7] P. Lasch, D. Naumann, FT-IR microspectroscopic imaging of human carcinoma thin sections based on pattern recognition techniques., Cell. Mol. Biol. (Noisy-Le-Grand). (1998).

[8] E.N. Lewis, L.H. Kidder, J.F. Arens, M.C. Peck, I.W. Levin, Si: As focal-plane array detection for Fourier transform spectroscopic imaging in the infrared fingerprint region, Appl. Spectrosc. (1997). doi:10.1366/0003702971940602.

[9] M. Diem, L. Chiriboga, H. Yee, Infrared spectroscopy of human cells and tissue. VIII. Strategies for analysis of infrared tissue mapping data and applications to liver tissue, Biopolym. - Biospectroscopy Sect. (2000). doi:10.1002/1097-0282(2000)57:5<282::AID-BIP50>3.0.CO;2-R.

[10] D.C. Fernandez, R. Bhargava, S.M. Hewitt, I.W. Levin, Infrared spectroscopic imaging for histopathologic recognition, Nat. Biotechnol. (2005). doi:10.1038/nbt1080.

[11] R. Bhargava, Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology, Anal. Bioanal. Chem. (2007). doi:10.1007/s00216-007-1511-9.

[12] E. Gazi, J. Dwyer, P. Gardner, A. Ghanbari-Siahkali, A.P. Wade, J. Miyan, N.P. Lockyer, J.C. Vickerman, N.W. Clarke, J.H. Shanks, L.J. Scott, C.A. Hart, M. Brown, Applications of Fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer. A pilot study, J. Pathol. (2003). doi:10.1002/path.1421.

[13] M.J. Baker, E. Gazi, M.D. Brown, J.H. Shanks, P. Gardner, N.W. Clarke, FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer, Br. J. Cancer. (2008). doi:10.1038/sj.bjc.6604753.

[14] R. Eckel, H. Huo, H.W. Guan, X. Hu, X. Che, W.D. Huang, Characteristic infrared spectroscopic patterns in the protein bands of human breast cancer tissue, Vib. Spectrosc. (2001). doi:10.1016/S0924-2031(01)00134-5.

[15] H. Fabian, N.A.N. Thi, M. Eiden, P. Lasch, J. Schmitt, D. Naumann, Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy, Biochim. Biophys. Acta - Biomembr. (2006). doi:10.1016/j.bbamem.2006.05.015.

[16] H. Fabian, P. Lasch, M. Boese, W. Haensch, Mid-IR microspectroscopic imaging of breast tumor tissue sections, in: Biopolym. - Biospectroscopy Sect., 2002. doi:10.1002/bip.10088.

[17] H. Fabian, P. Lasch, M. Boese, W. Haensch, Infrared microspectroscopic imaging of benign breast tumor tissue sections, J. Mol. Struct. (2003). doi:10.1016/j.molstruc.2003.07.002.

[18] I.W. Levin, R. Bhargava, FOURIER TRANSFORM INFRARED VIBRATIONAL SPECTROSCOPIC IMAGING: Integrating Microscopy and Molecular Recognition, Annu. Rev. Phys. Chem. (2005). doi:10.1146/annurev.physchem.56.092503.141205.

[19] C. Krafft, V. Sergo, Biomedical applications of Raman and infrared spectroscopy to diagnose tissues, Spectroscopy. (2006). doi:10.1155/2006/738186.

[20] M. Diem, M. Romeo, S. Boydston-White, M. Miljković, C. Matthäus, A decade of vibrational micro-spectroscopy of human cells and tissue (1994-2004), in: Analyst, 2004. doi:10.1039/b408952a.

[21] C. Petibois, G. Déléris, Chemical mapping of tumor progression by FT-IR imaging: towards molecular histopathology, Trends Biotechnol. (2006). doi:10.1016/j.tibtech.2006.08.005.

[22] J.G. Kelly, P.P. Angelov, J. Trevisan, A. Vlachopoulou, E. Paraskevaidis, P.L. Martin-Hirsch, F.L. Martin, Robust classification of low-grade cervical cytology following analysis with ATR-FTIR spectroscopy and subsequent application of self-learning classifier eClass, Anal. Bioanal. Chem. (2010). doi:10.1007/s00216-010-4179-5.

[23] I.I. Patel, J. Trevisan, P.B. Singh, C.M. Nicholson, R.K.G. Krishnan, S.S.

Matanhelia, F.L. Martin, Segregation of human prostate tissues classified high-risk (UK) versus low-risk (India) for adenocarcinoma using Fourier-transform infrared or Raman microspectroscopy coupled with discriminant analysis, Anal. Bioanal. Chem. (2011). doi:10.1007/s00216-011-5123-z.

[24] T. S., B. R., Extracting knowledge from chemical imaging data using computational algorithms for digital cancer diagnosis, Yale J. Biol. Med. (2015).

[25] S. Tiwari, R. Bhargava, Extracting knowledge from chemical imaging data using computational algorithms for digital cancer diagnosis, Yale J. Biol. Med. (2015).

[26] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, P.R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P.L. Martin-Hirsch, B. Obinaju, G.D. Sockalingum, J. Sulé-Suso, R.J. Strong, M.J. Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, Nat. Protoc. (2014). doi:10.1038/nprot.2014.110.

[27] F.N. Pounder, R.K. Reddy, R. Bhargava, Development of a practical spatial-spectral analysis protocol for breast histopathology using Fourier transform infrared spectroscopic imaging, Faraday Discuss. (2016). doi:10.1039/c5fd00199d.

[28] M. Diem, A. Ergin, S. Remiszewski, X. Mu, A. Akalin, D. Raz, Infrared micro-spectroscopy of human tissue: Principals and future promises, Faraday Discuss. (2016). doi:10.1039/c6fd00023a.

[29] G. Bellisola, C. Sorio, Infrared spectroscopy and microscopy in cancer research and diagnosis, Am. J. Cancer Res. (2012).

[30] C. Hughes, L. Gaunt, M. Brown, N.W. Clarke, P. Gardner, Assessment of paraffin removal from prostate FFPE sections using transmission mode FTIR-FPA imaging, Anal. Methods. (2014). doi:10.1039/c3ay41308j.

[31] F. Lyng, E. Gazi, P. Gardner, Preparation of Tissues and Cells for Infrared and Raman Spectroscopy and Imaging, RSC Anal. Spectrosc. Monogr. 01 (2011) 147–185. http://arrow.dit.ie/radrep.

[32] A. Tfayli, O. Piot, A. Durlach, P. Bernard, M. Manfait, Discriminating nevus and melanoma on paraffin-embedded skin biopsies using FTIR microspectroscopy, Biochim. Biophys. Acta - Gen. Subj. (2005). doi:10.1016/j.bbagen.2005.04.020.

[33] A. Tfayli, C. Gobinet, V. Vrabie, R. Huez, M. Manfait, O. Piot, Digital dewaxing of Raman signals: Discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies, Appl. Spectrosc. (2009). doi:10.1366/000370209788347048.

[34] C. Gobinet, V. Vrabie, A. Tfayli, O. Piot, R. Huez, M. Manfait, Pre-processing

and Source Separation methods for Raman spectra analysis of biomedical samples, in: Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc., 2007. doi:10.1109/IEMBS.2007.4353773.

[35] D.M. Stitt, M.Z. Kastyak-Ibrahim, C.R. Liao, J. Morrison, B.C. Albensi, K.M. Gough, Tissue acquisition and storage associated oxidation considerations for FTIR microspectroscopic imaging of polyunsaturated fatty acids, Vib. Spectrosc. 60 (2012) 16–22. doi:10.1016/j.vibspec.2011.10.016.

[36] P. Bassan, J. Lee, A. Sachdeva, J. Pissardini, K.M. Dorling, J.S. Fletcher, A. Henderson, P. Gardner, The inherent problem of transflection-mode infrared spectroscopic microscopy and the ramifications for biomedical single point and imaging applications, Analyst. 138 (2013) 144–157. doi:10.1039/C2AN36090J.

[37] J. Filik, M.D. Frogley, J.K. Pijanka, K. Wehbe, G. Cinque, Electric field standing wave artefacts in FTIR micro-spectroscopy of biological materials, Analyst. 137 (2012) 853. doi:10.1039/c2an15995c.

[38] B.J. Davis, P.S. Carney, R. Bhargava, Theory of midinfrared absorption microspectroscopy: I. Homogeneous samples., Anal. Chem. 82 (2010) 3474–86. doi:10.1021/ac902067p.

[39] B.J. Davis, P.S. Carney, R. Bhargava, Theory of mid-infrared absorption microspectroscopy: II. Heterogeneous samples., Anal. Chem. 82 (2010) 3487–99. doi:10.1021/ac902068e.

[40] M.J. Pilling, P. Bassan, P. Gardner, Comparison of transmission and transflectance mode FTIR imaging of biological tissue, Analyst. (2015). doi:10.1039/c4an01975j.

[41] P. Bassan, J. Mellor, J. Shapiro, K.J. Williams, M.P. Lisanti, P. Gardner, Transmission FT-IR chemical imaging on glass substrates: Applications in infrared spectral histopathology, Anal. Chem. (2014). doi:10.1021/ac403412n.

[42] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, Analyst. 142 (2017) 1258–1268. doi:10.1039/c6an02224c.

[43] P. Colarusso, L.H. Kidder, I.W. Levin, J.C. Fraser, J.F. Arens, E.N. Lewis, Infrared spectroscopic imaging: From planetary to cellular systems, Appl. Spectrosc. (1998). doi:10.1366/0003702981943545.

[44] M.J. Nasse, M.J. Walsh, E.C. Mattson, R. Reininger, A. Kajdacsy-Balla, V. MacIas, R. Bhargava, C.J. Hirschmugl, High-resolution Fourier-transform infrared chemical imaging with multiple synchrotron beams, Nat. Methods. (2011). doi:10.1038/nmeth.1585.

[45] L.H. Kidder, I.W. Levin, E.N. Lewis, V.D. Kleiman, E.J. Heilweil, Mercury cadmium telluride focal-plane array detection for mid-infrared Fourier-

transform spectroscopic imaging, Opt. Lett. (2008). doi:10.1364/ol.22.000742.

[46] P.D. Johnson, Synchrotron Radiation, in: Exp. Methods Phys. Sci., 1997. doi:10.1016/S0076-695X(08)60611-0.

[47] W.D. Duncan, G.P. Williams, Infrared synchrotron radiation from electron storage rings, Appl. Opt. (2009). doi:10.1364/ao.22.002914.

[48] J.K. Pijanka, A. Kohler, Y. Yang, P. Dumas, S. Chio-Srichan, M. Manfait, G.D. Sockalingum, J. Sulé-Suso, Spectroscopic signatures of single, isolated cancer cell nuclei using synchrotron infrared microscopy, Analyst. (2009). doi:10.1039/b821112d.

[49] P. Dumas, G.D. Sockalingum, J. Sulé-Suso, Adding synchrotron radiation to infrared microspectroscopy: what's new in biomedical applications?, Trends Biotechnol. (2007). doi:10.1016/j.tibtech.2006.11.002.

[50] F.L. Martin, Shining a new light into molecular workings, Nat. Methods. (2011). doi:10.1038/nmeth.1594.

[51] L.M. Miller, P. Dumas, Infrared Spectroscopy Using Synchrotron Radiation, in: Encycl. Biophys., Springer Berlin Heidelberg, Berlin, Heidelberg, 2018: pp. 1–9. doi:10.1007/978-3-642-35943-9_128-1.

[52] D. Moss, B. Gasharova, Y.L. Mathis, Practical tests of a focal plane array detector microscope at the ANKA-IR beamline, Infrared Phys. Technol. (2006). doi:10.1016/j.infrared.2006.01.033.

[53] R. Bhargava, I.W. Levin, Spectrochemical Analysis Using Infrared Multichannel Detectors, 2007. doi:10.1002/9780470988541.

[54] L.M. Miller, P. Dumas, Chemical imaging of biological tissue with synchrotron infrared light, Biochim. Biophys. Acta - Biomembr. (2006). doi:10.1016/j.bbamem.2006.04.010.

[55] C.J. Hirschmugl, K.M. Gough, Fourier transform infrared spectrochemical imaging: Review of design and applications with a focal plane array and multiple beam synchrotron radiation source, Appl. Spectrosc. (2012). doi:10.1366/12-06629.

[56] L.P. in. Choo, D.L. Wetzel, W.C. Halliday, M. Jackson, S.M. LeVine, H.H. Mantsch, In situ characterization of β-amyloid in Alzheimer's diseased tissue by synchrotron Fourier transform infrared microspectroscopy, Biophys. J. (1996). doi:10.1016/S0006-3495(96)79411-0.

[57] L.M. Miller, Q. Wang, T.P. Telivala, R.J. Smith, A. Lanzirotti, J. Miklossy, Synchrotron-based infrared and X-ray imaging shows focalized accumulation of Cu and Zn co-localized with β-amyloid deposits in Alzheimer's disease, J. Struct. Biol. (2006). doi:10.1016/j.jsb.2005.09.004.

[58] L. Miller, Q. Wang, A. Kretlow, M. Beekes, D. Naumann, In situ

characterization of prion protein structure and metal accumulation in scrapie-infected cells by synchrotron infrared and X-ray imaging, Vib. Spectrosc. (2005). doi:10.1016/j.vibspec.2005.02.023.

[59] R.Y. Huang, L.M. Miller, C.S. Carlson, M.R. Chance, In situ chemistry of osteoporosis revealed by synchrotron infrared microspectroscopy, Bone. (2003). doi:10.1016/S8756-3282(03)00233-3.

[60] D.A. Moss, M. Keese, R. Pepperkok, IR microspectroscopy of live cells, in: Vib. Spectrosc., 2005. doi:10.1016/j.vibspec.2005.04.004.

[61] M.J. Tobin, L. Puskar, R.L. Barber, E.C. Harvey, P. Heraud, B.R. Wood, K.R. Bambery, C.T. Dillon, K.L. Munro, FTIR spectroscopy of single live cells in aqueous media by synchrotron IR microscopy using microfabricated sample holders, Vib. Spectrosc. (2010). doi:10.1016/j.vibspec.2010.02.005.

[62] J. Doherty, A. Raoof, A. Hussain, M. Wolna, G. Cinque, M. Brown, P. Gardner, J. Denbigh, Live single cell analysis using synchrotron FTIR microspectroscopy: Development of a simple dynamic flow system for prolonged sample viability, Analyst. (2019). doi:10.1039/c8an01566j.

[63] L. Menzel, A.A. Kosterev, R.F. Curl, F.K. Tittel, C. Gmachl, F. Capasso, D.L. Sivco, J.N. Baillargeon, A.L. Hutchinson, A.Y. Cho, W. Urban, Spectroscopic detection of biological NO with a quantum cascade laser, Appl. Phys. B Lasers Opt. (2001). doi:10.1007/s003400100562.

[64] J.J. Valle, J.R. Eyler, J. Oomens, D.T. Moore, A.F.G. Van Der Meer, G. Von Helden, G. Meijer, C.L. Hendrickson, A.G. Marshall, G.T. Blakney, Free electron laser-Fourier transform ion cyclotron resonance mass spectrometry facility for obtaining infrared multiphoton dissociation spectra of gaseous ions, Rev. Sci. Instrum. (2005). doi:10.1063/1.1841953.

[65] B.G. Lee, M.A. Belkin, R. Audet, J. MacArthur, L. Diehl, C. Pflügl, F. Capasso, D.C. Oakley, D. Chapman, A. Napoleone, D. Bour, S. Corzine, G. Höfler, J. Faist, Widely tunable single-mode quantum cascade laser source for mid-infrared spectroscopy, Appl. Phys. Lett. (2007). doi:10.1063/1.2816909.

[66] K. Yeh, M. Schulmerich, R. Bhargava, Mid-infrared microspectroscopic imaging with a quantum cascade laser, in: Next-Generation Spectrosc. Technol. VI, 2013. doi:10.1117/12.2015984.

[67] S. Liakat, K.A. Bors, T.-Y. Huang, A.P.M. Michel, E. Zanghi, C.F. Gmachl, In vitro measurements of physiological glucose concentrations in biological fluids using mid-infrared light, Biomed. Opt. Express. (2013). doi:10.1364/boe.4.001083.

[68] S. Liakat, K.A. Bors, L. Xu, C.M. Woods, J. Doyle, C.F. Gmachl, Noninvasive in vivo glucose sensing on human subjects using mid-infrared light, Biomed. Opt. Express. (2014). doi:10.1364/BOE.5.002397.

[69] P. Bassan, M.J. Weida, J. Rowlette, P. Gardner, Large scale infrared imaging

of tissue micro arrays (TMAs) using a tunable Quantum Cascade Laser (QCL) based microscope, Analyst. (2014). doi:10.1039/c4an00638k.

[70] G. Clemens, B. Bird, M.J. Weida, J. Rowlette, M.J. Baker, Quantum cascade laser-based mid-infrared spectrochemical imaging of tissues and biofluids, Spectrosc. Eur. (2014).

[71] M.J. Pilling, A. Henderson, B. Bird, M.D. Brown, N.W. Clarke, P. Gardner, High-throughput quantum cascade laser (QCL) spectral histopathology: A practical approach towards clinical translation, Faraday Discuss. (2016). doi:10.1039/c5fd00176e.

[72] M.J. Pilling, A. Henderson, P. Gardner, Quantum Cascade Laser Spectral Histopathology: Breast Cancer Diagnostics Using High Throughput Chemical Imaging, Anal. Chem. 89 (2017) 7348–7355. doi:10.1021/acs.analchem.7b00426.

[73] M. Tahtouh, P. Despland, R. Shimmon, J.R. Kalman, B.J. Reedy, The application of infrared chemical imaging to the detection and enhancement of latent fingerprints: Method optimization and further findings, J. Forensic Sci. (2007). doi:10.1111/j.1556-4029.2007.00517.x.

[74] P. Lasch, D. Naumann, Spatial resolution in infrared microspectroscopic imaging of tissues, Biochim. Biophys. Acta - Biomembr. (2006). doi:10.1016/j.bbamem.2006.06.008.

[75] R. Bhargava, D.C. Fernandez, M.D. Schaeberle, I.W. Levin, Effect of focal plane array cold shield aperture size on Fourier transform infrared micro-imaging spectrometer performance, Appl. Spectrosc. (2000). doi:10.1366/0003702001949069.

[76] B. Bird, M. Miljkovic, M.J. Romeo, J. Smith, N. Stone, M.W. George, M. Diem, Infrared micro-spectral imaging: Distinction of tissue types in axillary lymph node histology, BMC Clin. Pathol. (2008). doi:10.1186/1472-6890-8-8.

[77] P. Lasch, W. Haensch, D. Naumann, M. Diem, Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis, Biochim. Biophys. Acta - Mol. Basis Dis. (2004). doi:10.1016/j.bbadis.2003.12.006.

[78] C. Beleites, G. Steiner, M.G. Sowa, R. Baumgartner, S. Sobottka, G. Schackert, R. Salzer, Classification of human gliomas by infrared imaging spectroscopy and chemometric image processing, in: Vib. Spectrosc., 2005. doi:10.1016/j.vibspec.2005.02.020.

[79] D. Ami, P. Mereghetti, A. Natalello, S.M. Doglia, M. Zanoni, C.A. Redi, M. Monti, FTIR spectral signatures of mouse antral oocytes: Molecular markers of oocyte maturation and developmental competence, Biochim. Biophys. Acta - Mol. Cell Res. (2011). doi:10.1016/j.bbamcr.2011.03.009.

[80] D.R. Whelan, K.R. Bambery, P. Heraud, M.J. Tobin, M. Diem, D. McNaughton, B.R. Wood, Monitoring the reversible B to A-like transition of DNA in

eukaryotic cells using Fourier transform infrared spectroscopy, Nucleic Acids Res. (2011). doi:10.1093/nar/gkr175.

[81] M.K. Raczkowska, P. Koziol, S. Urbaniak, C. Paluszkiewicz, M. Kwiatek, T.P. Wrobel, Denoising influence on classification results in the context of hyperspectral data : High Definition FT-IR imaging, Submitted. (n.d.).

[82] M.J. Baker, C. Clarke, D. Démoulin, J.M. Nicholson, F.M. Lyng, H.J. Byrne, C.A. Hart, M.D. Brown, N.W. Clarke, P. Gardner, An investigation of the RWPE prostate derived family of cell lines using FTIR spectroscopy, Analyst. (2010). doi:10.1039/b920385k.

[83] S. Garip, E. Yapici, N.S. Ozek, M. Severcan, F. Severcan, Evaluation and discrimination of simvastatin-induced structural alterations in proteins of different rat tissues by FTIR spectroscopy and neural network analysis, in: Analyst, 2010. doi:10.1039/c0an00540a.

[84] K.T. Cheung, J. Trevisan, J.G. Kelly, K.M. Ashton, H.F. Stringfellow, S.E. Taylor, M.N. Singh, P.L. Martin-Hirsch, F.L. Martin, Fourier-transform infrared spectroscopy discriminates a spectral signature of endometriosis independent of inter-individual variation, Analyst. (2011). doi:10.1039/c0an00972e.

[85] P. Bassan, A. Kohler, H. Martens, J. Lee, H.J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, P. Gardner, Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples, Analyst. (2010). doi:10.1039/b921056c.

[86] P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J.H. Shanks, M. Brown, N.W. Clarke, P. Gardner, FTIR microscopy of biological cells and tissue: Data analysis using resonant Mie scattering (RMieS) EMSC algorithm, Analyst. (2012). doi:10.1039/c2an16088a.

[87] J.G. Kelly, T. Nakamura, S. Kinoshita, N.J. Fullwood, F.L. Martin, Evidence for a stem-cell lineage in corneal squamous cell carcinoma using synchrotron-based Fourier-transform infrared microspectroscopy and multivariate analysis, in: Analyst, 2010. doi:10.1039/c0an00507j.

[88] J. Trevisan, P.P. Angelov, P.L. Carmichael, A.D. Scott, F.L. Martin, Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: Current practices to future perspectives, Analyst. (2012). doi:10.1039/c2an16300d.

[89] E. Ly, O. Piot, A. Durlach, P. Bernard, M. Manfait, Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition, Analyst. (2009). doi:10.1039/b820998g.

[90] S. Duraipandian, W. Zheng, J. Ng, J.J.H. Low, A. Ilancheran, Z. Huang, In vivo diagnosis of cervical precancer using Raman spectroscopy and genetic algorithm techniques, Analyst. (2011). doi:10.1039/c1an15296c.

[91] C. Krafft, M. Kirsch, C. Beleites, G. Schackert, R. Salzer, Methodology for fiber-optic Raman mapping and FTIR imaging of metastases in mouse brains, Anal. Bioanal. Chem. (2007). doi:10.1007/s00216-007-1453-2.

[92] R. Liu, G. Lv, B. He, K. Xu, Discriminant analysis of milk adulteration based on near-infrared spectroscopy and pattern recognition, in: Opt. Diagnostics Sens. XI Towar. Point-of-Care Diagnostics; Des. Perform. Valid. Phantoms Used Conjunction with Opt. Meas. Tissue III, 2011. doi:10.1117/12.873260.

[93] M. Sattlecker, R. Baker, N. Stone, C. Bessant, Support vector machine ensembles for breast cancer type prediction from mid-FTIR micro-calcification spectra, Chemom. Intell. Lab. Syst. (2011). doi:10.1016/j.chemolab.2011.05.007.

[94] MathWorks, Directed acyclic graph (DAG) network for deep learning - MATLAB - MathWorks United Kingdom, (2019). https://uk.mathworks.com/help/deeplearning/ref/dagnetwork.html (accessed July 11, 2019).

[95] D.M. Mayerich, M. Walsh, A. Kadjacsy-Balla, S. Mittal, R. Bhargava, Breast histopathology using random decision forests-based classification of infrared spectroscopic imaging data, Proc. SPIE. 9041 (2014) 904107. doi:10.1117/12.2043783.

[96] S. Mittal, T.P. Wrobel, L.S. Leslie, A. Kadjacsy-balla, A four class model for digital breast histopathology using High- Definition Fourier transform infrared (FT-IR) spectroscopic imaging, 9791 (n.d.) 1–8. doi:10.1117/12.2217358.

[97] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification 2nd Edition, New York John Wiley, Sect. (2001).

[98] R.L. Somorjai, Creating robust, reliable, clinically relevant classifiers from spectroscopic data, Biophys. Rev. (2009). doi:10.1007/s12551-009-0023-6.

[99] J.A. Fernández Pierna, V. Baeten, A.M. Renier, R.P. Cogdill, P. Dardenne, Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds, J. Chemom. (2004). doi:10.1002/cem.877.

[100] R.L. Somorjai, M.E. Alexander, R. Baumgartner, S. Booth, C. Bowman, A. Demko, B. Dolenko, M. Mandelzweig, A.E. Nikulin, N.J. Pizzi, E. Pranckeviciene, A.R. Summers, P. Zhilkin, A Data-Driven, Flexible Machine Learning Strategy for the Classification of Biomedical Data, in: Artif. Intell. Methods Tools Syst. Biol., 2007. doi:10.1007/978-1-4020-5811-0_5.

[101] T.P. Wrobel, J.T. Kwak, A. Kadjacsy-balla, R. Bhargava, High-Definition Fourier Transform Infrared Spectroscopic Imaging of Prostate Tissue, 9791 (n.d.) 5–11. doi:10.1117/12.2217341.

[102] S. Mittal, T.P. Wrobel, S. Leslie, A. Kadjacsy-Balla, R. Bhargava, A four class

model for digital breast histopathology using High- Definition Fourier transform infrared (FT-IR) spectroscopic imaging, Prog. Biomed. Opt. Imaging - Proc. SPIE. 9791 (2016) 1–8. doi:10.1117/12.2217358.

[103] L.S. Leslie, T.P. Wrobel, D. Mayerich, S. Bindra, R. Emmadi, R. Bhargava, High definition infrared spectroscopic imaging for lymph node histopathology, PLoS One. (2015). doi:10.1371/journal.pone.0127238.

[104] L.E.O. Breiman, Random Forests, (2001) 5–32.

[105] A. Cutler, D.R. Cutler, J.R. Stevens, Random Forests, (2012). doi:10.1007/978-1-4419-9326-7.

[106] L. Breiman, Bagging predictors - Springer, Mach. Learn. (1996). doi:10.1007/BF00058655.

[107] T. Hastie, R. Tibshirani, J. Friedman, Elements of Statistical Learning 2nd ed., Elements. (2009). doi:10.1007/978-0-387-84858-7.

Chapter 3

Instrumentation and Methodology

3.1. The electromagnetic radiation

Molecular spectroscopy can be defined as the study of the interaction of electromagnetic waves and matters. The nature of electromagnetic radiation can be described as the perpendicular propagation of one electric field and one magnetic field. Both fields are made up of harmonic waves raised by the motion of photons [1,2].

Electromagnetic radiation has been divided into different segments based on their distinct properties, which can be described using wavelength, frequency and energy. A plot of the regions of the electromagnetic spectrum is shown in Figure 3.1. The mathematical relationships among these parameters motioned are shown in Equation 1.

$$\lambda = \frac{c}{\gamma} \qquad \text{Equation 1}$$

Where λ defines the wavelength, c defines the speed of light, and γ defines the frequency of the electromagnetic radiation.

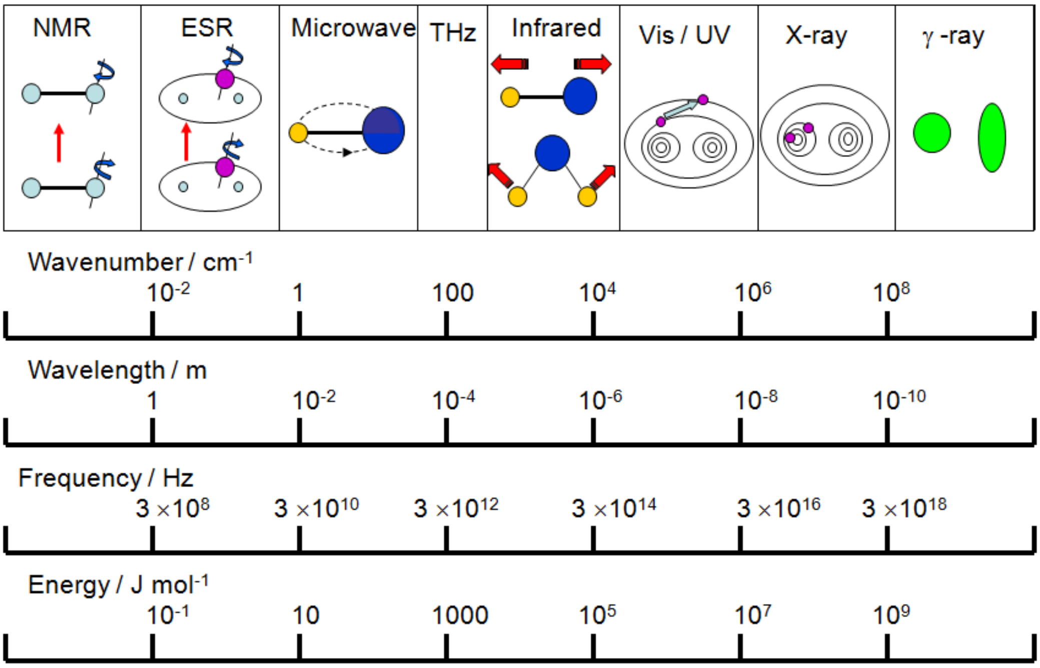


Figure 3.1 The regions of the electromagnetic spectrum

A more commonly used parameter to describe the wavelength of the electromagnetic radiation in Equation 2.

$$\bar{\upsilon} = \frac{1}{\lambda} \qquad \text{Equation 2}$$

Where $\bar{\upsilon}$ denotes wavenumber of the electromagnetic radiation, which determines the number of wave cycles in unit length.

The energy level of electromagnetic radiation can be expressed using the parameters mentioned above by the following equation.

$$E = h\gamma = \frac{hc}{\lambda} = hc\bar{\upsilon} \qquad \text{Equation 3}$$

Where *h* denotes the Plank's constant, which is $6.6 \times 10^{-34} J^{-1} s$.

Therefore, the energy level of the electromagnetic radiation is proportional to its wavenumber value. The increasing value of wavenumber will lead to an increase in radiation energy level.

3.1.1. The infrared radiation

Infrared is defined as the electromagnetic wave between visible light and microwave regions [2]. The infrared region is divided into the far-infrared, mid-infrared and near-infrared regions. The detailed information about wavenumber and wavelength are shown in Table 3.1. Among three sub-regions, the mid-infrared region is used for most spectrometers, for example, the FT-IR spectrometers.

Table 3.1 Sub-division of infrared spectra in term of wavelength and wavenumber

Infrared Sub-divisions	Wavelength /μ m	Wavenumber / cm^{-1}
Far-infrared	25-1000	400-10
Mid-infrared	2.5-25	4000-400
Near-infrared	0.7-2.5	14000-4000

When infrared radiation passes through substances, some radiation is absorbed by molecules and some are transmitted. The absorbed energy is provided by photon associated with the infrared. This energy is not large enough to induce direct electronic transitions. However, it conducts vibrational excitation of covalent bonds between atoms and groups. Molecular vibrations (stretching and bending) and rotations will cause net changes in the dipole moment of molecules. The larger the dipole change, more intense bands are observed in the respective infrared spectrum [1,3–5].

3.1.2. Interaction with molecules

Molecular spectroscopy can be defined as the study of the interaction of electromagnetic waves and particles [1,2]. Quantum theory allows the conception of a bond-vibrational nature as a possible explanation of the molecular absorption in the infrared region. This concept evolved the current understanding of the infrared spectra as a result of energetic transitions between quantized vibrational states. This is based on three premises. Firstly, molecules are constituted of covalently bonded atoms with linkages caused by electronic rearrangements [2]. The vibration must give rise to a dipole change which magnitude is correlated with each individual effect of the dipole moments within the whole molecule [2]. Last but not the least, because of the constant movement of molecules, the dipole moment has a continuous oscillating pattern with a defined frequency which can interact with the electric field of the IR radiation. When the frequency of a photon is equal to the frequency linked to the dipole moment of a specific molecule, the molecule will absorb IR radiation.

3.1.3. Molecular vibration

The molecular motion of a poly-functional molecule includes rotational, translational and vibrational movements. This means a molecule containing *n* atoms has 3*n* degrees of freedom on the x, y, and z-axes of the Cartesian plane respectively. 6 degrees of freedom are classified into translational and rotational motion. Therefore, an equation for non-linear molecules can be summarized as follow:

$$\text{Number of degrees of vibrational freedom} = 3n - 6 \qquad \text{Equation 2}$$

For linear molecules, there is no rotation along the bond axis. Thus, the equation can be expressed as follow:

$$\text{Number of degrees of vibrational freedom} = 3n - 5 \qquad \text{Equation 3}$$

Molecular vibrations are different for each molecule according to their specific chemical composition (except the enantiomers), structures and surroundings. Nevertheless, the general modes of vibration described below have been reported leading the vibrational motion of a molecule in most situations [2–4,6].

To understand molecular motion, the atoms can be assumed as individual masses, and the covalent bond works like a spring and undergoes simple harmonic oscillation. Then, the movement of a diatomic molecule can be described by Hook's law under these conditions [2], which is shown in Equation 5.

$$v = \frac{1}{2\pi}\sqrt{\frac{f}{\mu}} \qquad \text{Equation 4}$$

where

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \qquad \text{Equation 5}$$

However, rather than following Hook's law, real molecules perform a more complex behaviour which can be considered as anharmonic oscillators [2], as in reality the restoring force is not proportional to the displacement. Both elastic anharmonicity and damping anharmonicity can be used to describe real-world situations [7]. A Morse potential is used to approximate the vibrational structure in this case, as it accounts for the anharmonicity of real bonds. This explanation has the intrinsic properties of a specific molecule like the internuclear distance and the dissociation energy included.

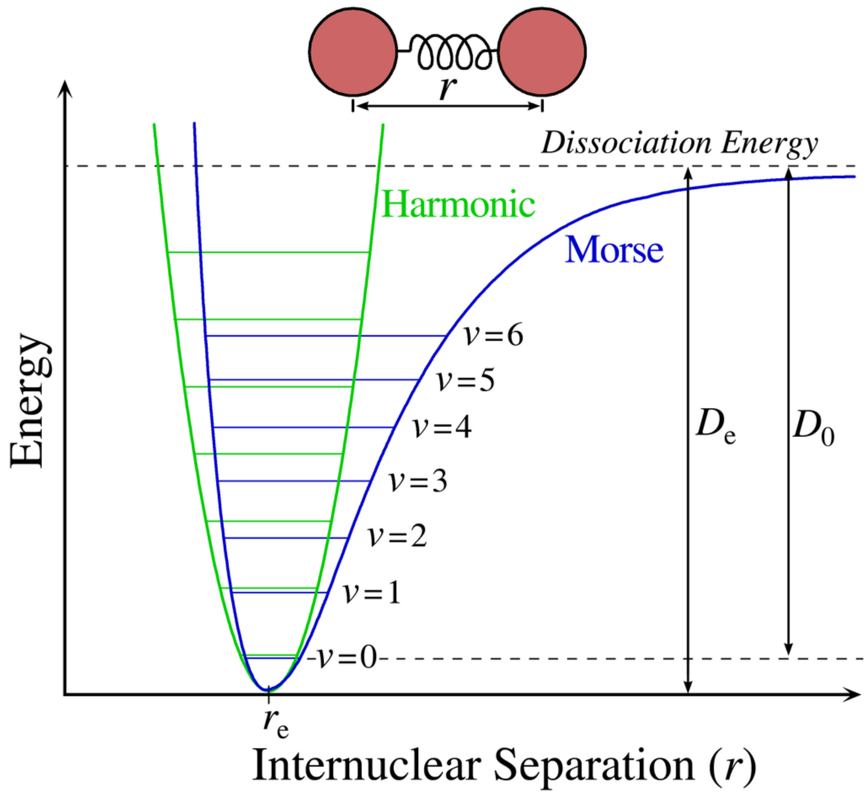


Figure 3.2 Vibrational energy levels for a diatomic molecule, The Morse potential (blue) and harmonic oscillator potential (green).

As shown in Figure 3.2, the energy levels of the harmonic oscillation are equally distributed, while they reduce when the energy reaches the dissociation energy for the anharmonic oscillation.

3.1.4. Modes of vibration

To describe the molecular motion, the changes of the conformational shape such as inter-atomic distance and bond angles for stretching and a bending motion respectively have to be taken into account. For the given molecular group, two stretching vibrations including symmetric and asymmetric stretch, and four deformation vibrations containing bending, wagging, twisting and rocking, are compared.

In a biological context, vibrational modes of peptide bonds, which is formed when amino acids linked to form protein known as one of the most important reactions in biochemistry [8], including Amide I, Amide II, and Amide A are shown in Figure 3.3. The peptide group as structural unit of proteins has 9 characteristic bands, among them amide I and amide II are the two most intense bands of proteins in the infrared spectrum [9]. The Amide I and Amide II are commonly used for protein structure analysis [10]. Between them, Amide I is the more widely used band to investigate the secondary structure of proteins [10,11], as it is the most sensitive amide band to differences in secondary structures.

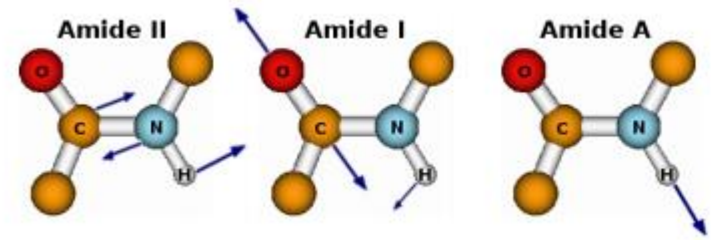


Figure 3.3 Vibrational modes for the peptide bonds

The type of vibrational mode of each peptide bond is presented in Table 3.2. From that table, it can be seen that Amide A in the wavenumber range of 3230-3300 cm^{-1} is 100% caused by N-H stretching, Amide I (wavenumber range: 1600-1700 cm^{-1}) is mainly raised by C=O stretching and coupled with C-N stretching, and Amide II, between wavenumber of 1510 and 1580 cm^{-1}, are mainly stemmed from the N-H bending and C-N stretching weakly coupled to the C=O stretching mode.

Table 3.2 wavenumber and bond assignment and vibrational modes of Amide A, Amide I and Amide II

Designation	Wavenumber (cm^{-1})	Assignment
Amide A	3300-3230	ν(NH) 100%
Amide I	1700-1600	ν(CO) 70-85% ν(CN) 10-20%
Amide II	1580-1510	δ(NH) 40-60% ν(CN) 18-40% ν(C-C) ~10%

3.2. FTIR spectroscopy

Fourier Transform Infrared (FTIR) techniques have been wildly used for the no-destructive analysis of biological specimens in recent years. Especially in the area of cytological and histological diagnosis by generating spectral images [4,12,13].

Molecular bonds are IR active because of the presence of an electric dipole moment which is able to change by atomic displacement due to natural vibrations. These vibrational modes can be measured quantitatively by IR spectroscopy. It provides a unique, label-free tool for exploring molecular composition and dynamics without damaging the samples. For interrogating biological materials, the most significant spectral regions measured are typically the fingerprint region (600–1450 cm^{-1}) and the amide I and amide II (amide I/II) region (1,500–1,700 cm^{-1}). The higher wavenumber region (2,550–3,500 cm^{-1}) is related to stretching vibrations such as S-H, C-H, N-H and O-H, while the lower-wave number regions typically associated with bending and carbon skeleton fingerprint vibrations [13].

FTIR spectrometry is a relatively advanced technique in the field. The first chemical IR spectroscopy was invented by Sir William Herschel in 1800. In the very beginning, most IR instrumentation was based on a prism or grating monochromator. In 1881, Michelson invented interferometer, which many years later turned out to be a revolutionary break-through in term of infrared spectroscopy development. In 1949, Peter Fellgett generated the first IR spectrum by using FTIR spectrometer and improved the quality of the spectrum collected. Commercial FTIR spectrometer appeared in 1960s. The first commercial FT-IR spectrometer has been developed in 1969 [3]. FT-IR is designed to overcome the limitations of dispersive instruments, which are slow-scanning time and limited energy throughout. The core element of an FT-IR spectrometer is interferometer, which enables measuring all of the infrared frequencies simultaneously. Infrared frequencies are encoded into this optical device [14]. The speed of infrared spectrometry is largely increased. The time spent to measure an element per sample is reduced to a matter of a few seconds rather than several minutes [3,14].

A beam splitter is integrated with the interferometer. It splits the incoming light into two beams. One of the beams is reflected by a fixed flat mirror while the other one is reflected by a moving flat mirror. Two reflected beams will recombine, as one travels a fixed distance while the other one adjusts its travelling track. The combined beam passes through the interferometer. The interfered signal results in every frequency coming from the source. FT-IR spectrometer can process extremely fast measurements.

The signal from a single frequency measured at a detector as a function of time is a sine wave. The combination of multiple sine waves from all frequencies applied during measuring gives an interference pattern, interferogram. An example plot of individual sine wave before interference and the resultant interferogram is shown in Figure 3.4. The interferogram generated by the detector is in the time domain spectrum. It needs to be translated into the frequency domain spectrum.

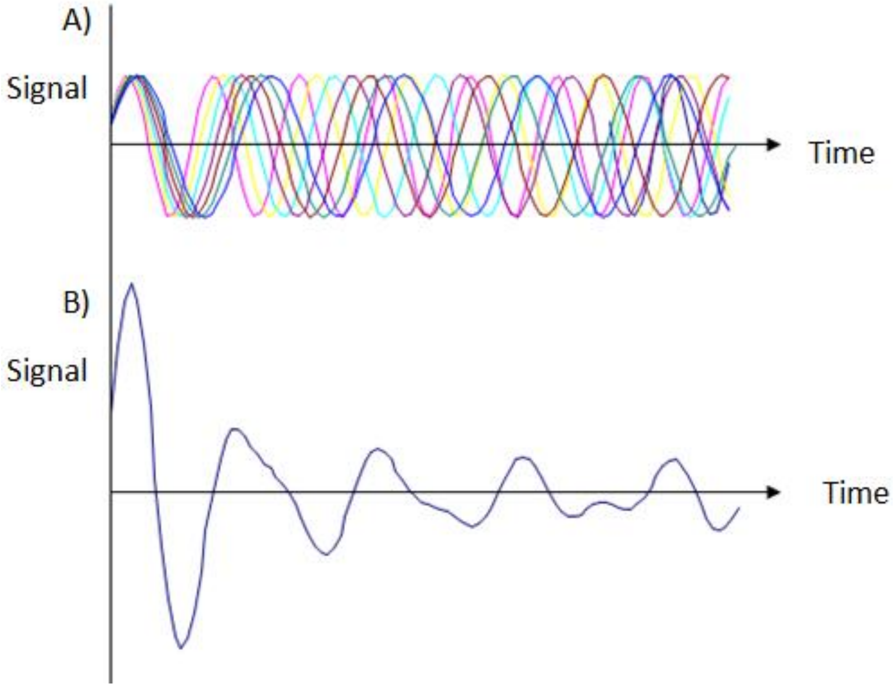


Figure 3.4 An example plot of A) individual sine wave before interference and B) the resultant interferogram functions of time

Despite the complexity of the interferogram, the separate frequencies can be extracted via the mathematical process known as Fourier transformation [14]. The transform can be expressed in equation 7 [15]:

$$\hat{f}(\xi) = \int_{-\infty}^{+\infty} f(x)e^{-2\pi i x\xi}\, dx \qquad \text{Equation 7}$$

Where ξ stands for frequency, x represents time.

An example plot of the transformation process is shown in Figure 3.5, where interferogram based on time is transferred into the wavenumber based spectrum.

The main advantage of Fourier transform methods is that all the frequencies are measured simultaneously rather than sequentially, which is also referred as the multiplex advantage. This calculation is performed by computers in which the users can choose desired sections of the overall results.

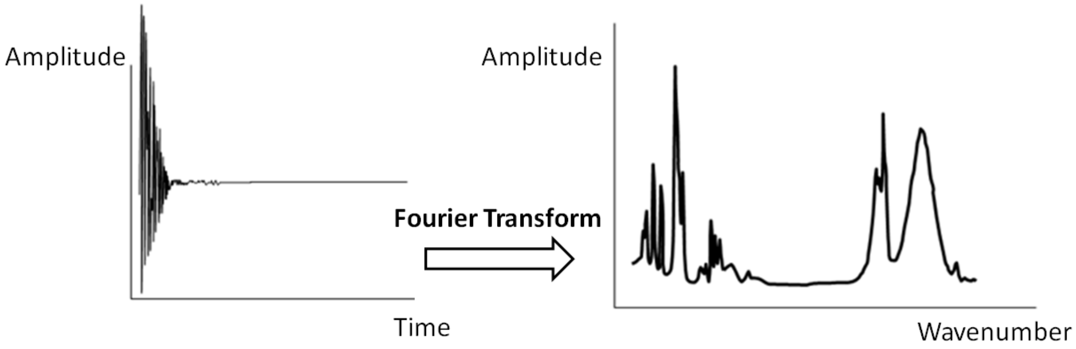


Figure 3.5 Example diagram of Fourier Transform from time-domain interferogram n to wavenumber based spectrum

3.2.1. FTIR microscopy

Compare with the traditional FT-IR spectrometer, infrared microscopes are able to record the transmission or reflectance spectra at each spatial point of samples [17]. The recorded data is processed by a spectral analysis calculation to determine the specific chemical composition of samples.

The combination with microscopy (micro-spectroscopy) permits the examination of complex tissues and heterogeneous samples. Detection by microscopy may be accomplished by raster-scanning a point illuminated on the sample or by using wide-field illumination and FPA or linear array detectors [16].

The FPA is an advanced infrared chemical imaging technique. It largely improved the quality and speed of infrared image, which provides the essential technology for high-quality infrared-microscopy. FPA is created by organising individual detecting elements in a lattice-like array [17]. Normally each pixel (one individual detector) has one independent contact and one-second contact sharing with other pixels in FPA. FPA can be made of both photon and thermal detectors, for example, MCT detector [18].

The most common detector used in the FPA and FT-IR spectrometer is MCT, which has a chemical formula HgCdTe. MCT detectors give a large spectral range, zero-bias resistance and have high quantum efficiency. These advantages make MCT the most popular photon detector. However, it also has disadvantages. The major limitation of MCT detectors is that it needs low temperature (77K) to reduce the thermal noises caused by excited current carriers [19]. These qualities make MCT the most popular intrinsic photovoltaic [20]. The MCT detects electrons which are exited from chemical bonds to conduction bands by photon. An external readout integrated circuits (ROIC) gathers those electrons and transforms them into electrical signals [21].

3.3. References

[1] W. Struve, I. Mills, Fundamentals of Molecular Spectroscopy, Vib. Spectrosc. (1990). doi:10.1016/0924-2031(90)80014-u.

[2] C.N. Banwell, E.M. McCash, Fundamentals of Molecular Spectroscopy, McGraw Hill, 2013. https://books.google.co.uk/books?id=xyHZAgAAQBAJ&source=gbs_book_similarbooks (accessed September 19, 2019).

[3] P.R. Griffiths, J.A. De Haseth, Fourier Transform Infrared Spectrometry: Second Edition, 2006. doi:10.1002/047010631X.

[4] P.R. Griffiths, J.A. de Haseth, Introduction to Vibrational Spectroscopy, in: Fourier Transform Infrared Spectrom., 2007. doi:10.1002/9780470106310.ch1.

[5] D. Steele, Infrared Spectroscopy: Theory, in: Handb. Vib. Spectrosc., 2006. doi:10.1002/0470027320.s0103.

[6] B.H. Stuart, Infrared Spectroscopy: Fundamentals and Applications, 2005. doi:10.1002/0470011149.

[7] R.M. Harris, Physical chemistry for the life sciences, Nature. (1980). doi:10.1038/284084b0.

[8] J. Parker, Peptide Bond, Encycl. Genet. (2001) 1429–1430. doi:10.1006/RWGN.2001.0969.

[9] U. Langel, B.F. Cravatt, A. Graslund, N.G.H. Von Heijne, M. Zorko, T. Land, S. Niessen, Introduction to Peptides and Proteins., Taylor and Francis, 2009. https://books.google.co.uk/books?id=GA3SBQAAQBAJ&printsec=frontcover&dq=proteins+peptide+introduce&hl=en&sa=X&ved=0ahUKEwjqiPv009_kAhWsgVwKHUmeBLkQ6AEIKjAA#v=onepage&q=proteins peptide introduce&f=false (accessed September 20, 2019).

[10] P.I. Haris, Infrared Spectroscopy of Protein Structure, in: Encycl. Biophys., Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 1095–1106. doi:10.1007/978-3-642-16712-6_135.

[11] A. Barth, P.I. (Parvez I.. Haris, Biological and biomedical infrared spectroscopy, IOS Press, 2009.

[12] J. Trevisan, P.P. Angelov, P.L. Carmichael, A.D. Scott, F.L. Martin, Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: Current practices to future perspectives, Analyst. (2012). doi:10.1039/c2an16300d.

[13] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, P.R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P.L. Martin-hirsch, B. Obinaju, G.D. Sockalingum, J. Sulé-suso, R.J. Strong, M.J.

Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, 9 (2014) 1771–1791. doi:10.1038/nprot.2014.110.

[14] M.A. Ganzoury, N.K. Allam, T. Nicolet, C. All, Introduction to Fourier Transform Infrared Spectrometry, Renew. Sustain. Energy Rev. (2015). doi:10.1016/j.rser.2015.05.073.

[15] J.-B.-J. Fourier, The analytical theory of heat, Dover Publications, 2003. https://books.google.co.uk/books/about/The_Analytical_Theory_of_Heat.html?id=YARW_BgJvrwC&redir_esc=y (accessed September 20, 2019).

[16] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, P.R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P.L. Martin-Hirsch, B. Obinaju, G.D. Sockalingum, J. Sulé-Suso, R.J. Strong, M.J. Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, Nat. Protoc. (2014). doi:10.1038/nprot.2014.110.

[17] A. Rogalski, Progress in focal plane array technologies, Prog. Quantum Electron. (2012). doi:10.1016/j.pquantelec.2012.07.001.

[18] R. Bhargava, Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology, Anal. Bioanal. Chem. (2007). doi:10.1007/s00216-007-1511-9.

[19] G.J. Bowen, I.D. Blenkinsop, R. Catchpole, N.T. Gordon, M.A.C. Harper, P.C. Haynes, L. Hipwood, C.J. Hollier, C. Jones, D.J. Lees, C.D. Maxey, D. Milner, M. Ordish, T.S. Philips, R.W. Price, C. Shaw, P. Southern, HOTEYE: a novel thermal camera using higher operating temperature infrared detectors, in: Infrared Technol. Appl. XXXI, 2005. doi:10.1117/12.603305.

[20] L.J. Kozlowski, Infrared detector arrays, Handb. Opt. Vol. II - Des. Fabr. Testing; Sources Detect. Radiom. Photom. (2010).

[21] A. Rogalski, HgCdTe infrared detector material: History, status and outlook, Reports Prog. Phys. (2005). doi:10.1088/0034-4885/68/10/R01.

Chapter 4

Cancer diagnosis of breast tissue on CaF2

4.1. Chapter overview

Breast cancer is the most common cancer for females in the world. It is also the second most common and dangerous cancer with 2,088,849 (11.6%) new cases and 626,679 (6.6%) cases of death for both males and females in 2018, across 20 countries and regions[1]. In the UK, breast cancer is the most commonly diagnosed cancer in 2016. In total 45,960 cases including 45656 females, it accounted for 15.2% of all malignant cancers and 30.8% of all female malignant cancer cases[2].

The survival rates of breast cancer generally decrease with the increase of cancer stages. Based on information provided by office for national statistics, 207,548 women who were diagnosed having breast cancer in 2011 had a five-year survival rates of 98.8% (stage I), 90.1% (stage II), 72.2% (stage III) and 27.9% (stage IV) followed up in 2017 in the UK[3]. Therefore, early detection of breast cancer could largely increase the survival rate of patients.

In order to reduce the mortal rate of breast cancer, reliable and efficient early diagnosis systems need to be established and applied for clinic uses. Currently, three screening methods are available for early breast cancer detection-containing mammography, ultrasound scan, and magnetic resonance imaging. In most cases, biopsies are required and taken to assess the presence of a disease and its degree of progression. Tissues taken from patients and sliced into thin slides are carefully investigated by professional pathologists using optical microscopes. This cancer finding and observing process is a time-consuming and expensive process. More importantly, this process is subjective and prone to observer error [4], as decisions are made by a sole pathologist or several individuals. Comparing with traditional pathology, FTIR spectrometry has its advantages including time efficiency, cheaper and having promising classification accuracies [5].

The research was conducted using on breast cancer diagnosis. Eckel *et al.*[6] first applied FTIR micro-spectroscopy for the analysis of breast tissues. In 2001, they published paper on finding characteristic infrared spectroscopic patterns in protein bands of human breast cancer tissues. Spectra from 96 patients were collected. This research provided spectral characteristics of different breast cancer types, which were later used as references for breast cancer features or biomarkers in the field. Since numerous research has been conducted by different research groups.[7–11] Most groups only analyse epithelial cell spectra since breast cancer occurs in epithelium cells in the vast majority of all breast cancers. However, in 2016, Verdonck *et al*. [12] found that stroma located close or at a distance from cancer cells shows distinct spectral characteristics comparing with normal stroma. Classifiers could be constructed based on adjacent stroma when it came to breast cancer diagnosis.

In the light of previous research, in this chapter, FTIR analysis on breast tissue is combined with the recently popular data analysis method, machine learning, to further validate the reliability of FTIR breast cancer diagnosis and to testify whether stroma could be used as a substitution of epithelium cells working as inputs of breast cancer classification models.

4.2. Methodology

All data were pre-processed with MATLAB 2018a. Infrared spectra for each biopsy core were extracted from the mosaic as a 256 × 256 datacube. Each datacube consists of 65536 spectra each of which contains 416 data points.

4.2.1. Breast Tissue preparation

The sample used for this project is a breast TMA slide, BR20832 (US Biomax, Inc.). In this case, the breast tumour tissue microarray contains 190 cases of ductal carcinoma, 1 mixed lobular and duct carcinoma, 1 mucinous carcinoma, 13 NAT and 3 normal breast tissue. Each core stands for one patient. In total, the BR20832 TMA contains information from 208 patients both healthy and having cancer.

The TMA was prepared by Biomax according to the following protocol. Fresh biopsies taken from 208 patients were immersed in an aqueous formalin solution, which fixes the tissue. After that, the tissue is dehydrated in washes of xylene and ethanol. In the end, paraffin was used to embed the tissue to provide a protective barrier [13]. The core was cut into 5µm thickness and 1mm diameter and fixed on the CaF_2 slide, which was designed for FTIR analysis use. The adjacent TMA was flowed on a glass slide and H&E stained. The H&E stained TMA would be used to annotate epithelium and stroma. The actual FTIR scan was performed on the normal non-stained CaF_2 slide, as the level of stain could influence spectra containing subtle biological differences. Figure 4.1 shows the image of the H&E stained BR20832 slide.

The normal slide used for FTIR analysis was covered in wax, which was a formalin-FFPE breast tissue TMA slide. No dewaxing process was undertaken before the actual spectra collection. It has been shown by Untereiner *et al.* [14], that using FFPE sections without de-waxing does not affect the discriminative potential of the technique, in terms of using FTIR imaging [15]. Also, using the wax embedded sample could avoid spectral scattering to a large extent since the refractive index change between wax and tissue was smaller than air and tissue.

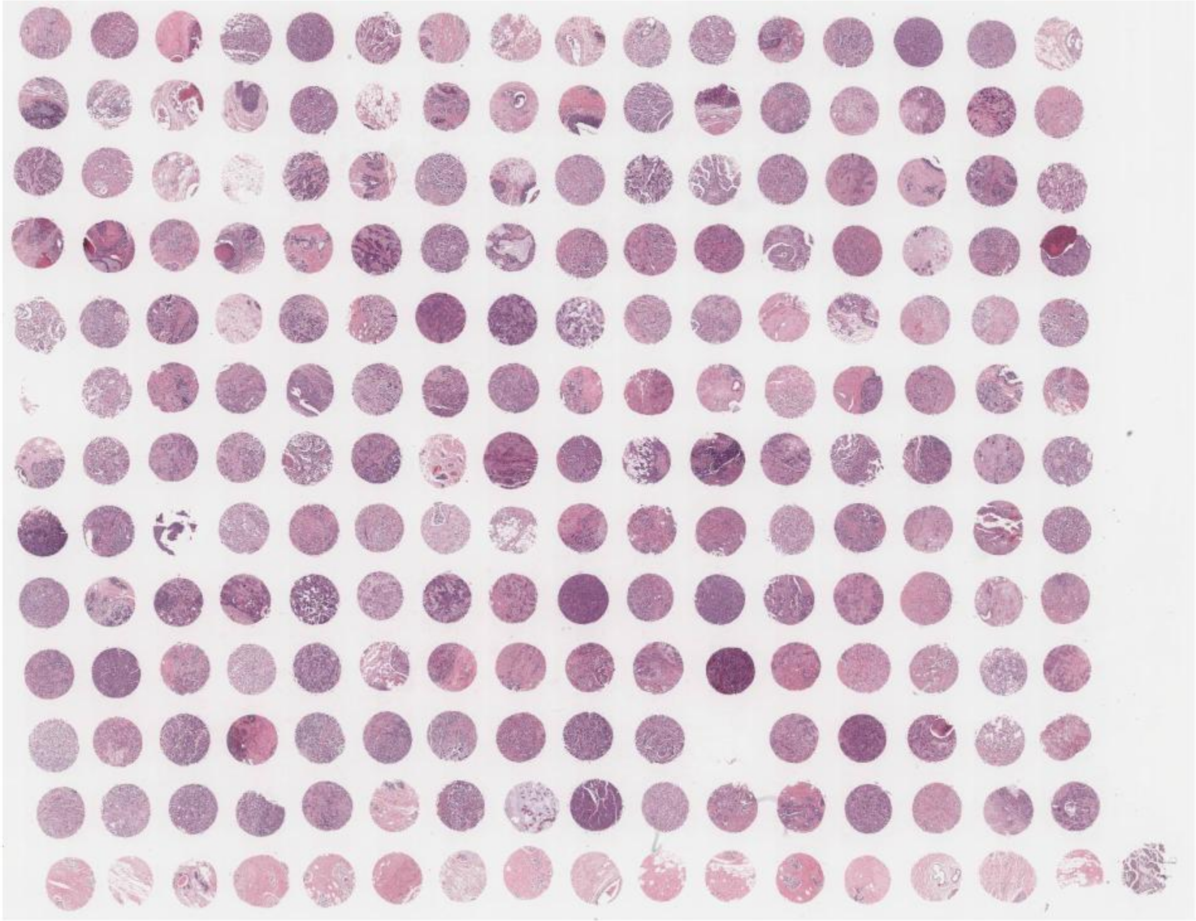


Figure 4.1 Optical image of BR20832 (breast tissue) H&E stained cores on glass slide

The whole TMA slide was measured twice. The first experiment (experiment one) was conducted in 2016 while the second experiment (experiment two) was finished in 2017. The reason for re-measuring was explained in the later section, 4.3.1.

For model one, 70 cores were selected from the sample set, which included 59 cores with grade I, II or III of breast cancer and 11 normal associate breast tissue cores. The model was constructed based on data collected from experiment one. Spectra collected in experiment two were used to establish model two. For model two, 70 cores were selected from sample set, which included 57 cores with grade I, II or III of breast cancer and 13 normal associate breast tissue cores.

4.2.2. Experiment procedure

The transmission was used to reduce the influence of the electric field standing wave effect.[16] The number of background scan was set to be 256, while the number of sample scans was 96. The resolution used was 5 cm^{-1} since it provided high-quality data and is significantly more time-efficient compared with 4 cm^{-1}. In experiment 1, a purge time of 20 minutes was adopted to reduce the influence of water vapour in the spectrum before loading TMA and after changing samples. In experiment 2, the whole TMA was placed in the purge box overnight to further reduce the influence of water vapour. Every day during measuring before data collection, the microscope was focused. The centre burst of spectrometer was found. The received energy intensity level distribution of pixels was adjusted to be as uniform as possible. The number of out of range pixel was ensured to be zero. The scan range was set from 1000 to 3800 cm^{-1} covering the important range of biological information of breast tissue. The spectrum collected was then saved and converted for data analysis process.

4.2.3. Instrumentation

For this project, an Agilent 670 FTIR spectrometer was used to conduct measurements. The spectrometer was connected to the Varian 620 infrared imaging microscope. A liquid nitrogen cooled 128×128 FPA detector was coupled to the microscope. A retro-reflection mirror was used in this equipment to increase the signal to noise ratio, which improved the quality of spectra. The Varian's Resolutions Pro software was used to adjust and control spectrometer and microscope during measurement. The system was designed for biotechnology applications. When ×15 optics was applied, 5.5 µm ×5.5 µm pixel sizes were obtained in low magnification mode of the instrument. The FTIR imaging system could be used to collect large chemical images across a Field of View (FOV) of 2.4 mm × 2.4 mm, as mosaics. In high magnification mode, the pixel size could be further reduced to only 1.1 µm ×1.1 µm, which could give high-quality images. The moving mirror of the spectrometer was not perpendicular to the static mirror as the common FTIR spectrometer. The numerical aperture equalled to 0.62 and the distance between optics and samples was 21 mm. Each core from TMA could be imaged as a 2×2 mosaic.

4.2.4. Data analysis methods

4.2.4.1. Annotation on Breast tissue Chemical Images

Chemical images of each of the prostate tissue cores were generated and compared to the H&E stained sections, and regions of epithelium were identified and annotated on chemical images generated based on WHO Classification of Tumours in the Breast[17].

Examples of core annotation were shown in Figure 4.2 B) and Figure 4.3 B). Cancerous epithelium cells are annotated in red (R255 G0 B0) and stroma parts in purple (R165 G55 B156), while in cancer-associated tissues, epithelium cells are coloured in green (R103 G193 B66) and stroma in yellow (R243 G143 B53). Figure 4.2 A) and B) are the chemical image and annotated image of core D5, which is a grade-2 and stage-II cancerous core from a 46-year old female. Red annotation represents cancerous epithelium cells while purple annotation indicates stroma region. M10 is shown Figure 4.3 A) and B) (chemical and annotated images respectively), which is a Cancer adjacent normal breast tissue from a 46-year old female. Green is used to annotate normal epithelium cells while yellow is used to annotated stroma area.

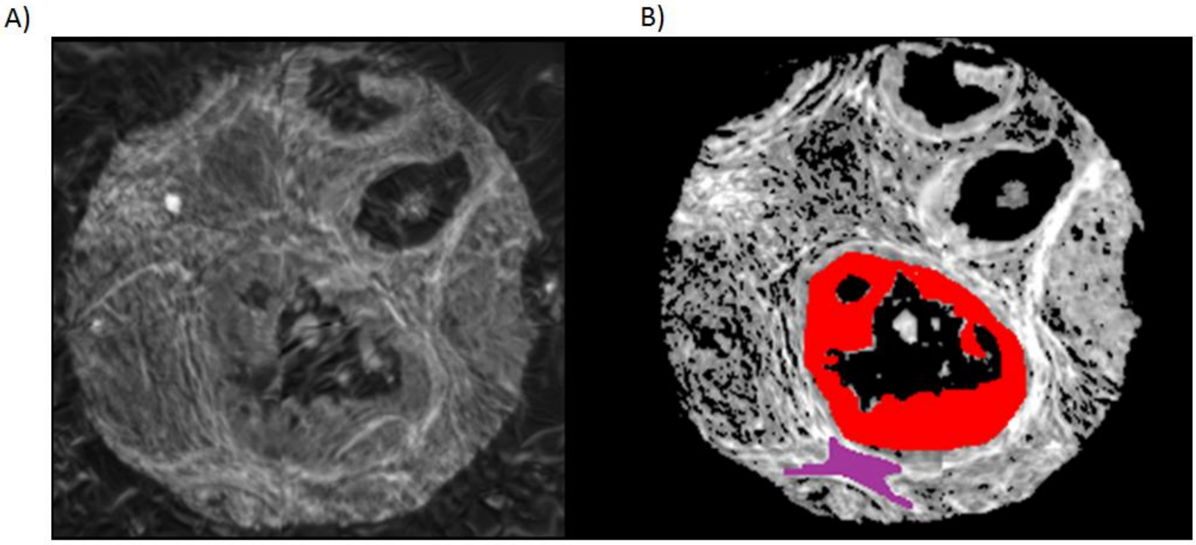


Figure 4.2 A) Chemical image (generated by the absorbance distribution of amide I band of the sample) and B) annotated chemical image of core D5 with red and purple stood for cancerous epithelium and stroma respectively

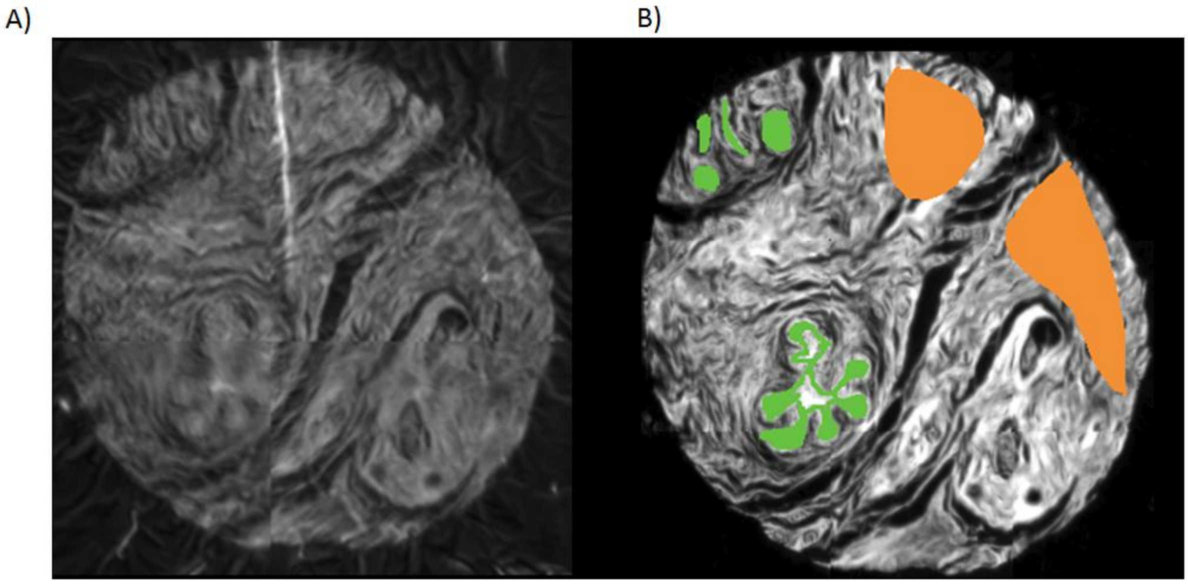


Figure 4.3 A) Chemical image which was plotted based on the distribution of Amide I absorbance and B) annotated chemical image of core M10 where green indicated normal epithelium and yellow showed normal stroma on the core

4.2.4.2. Pre-processing methods

Two models were constructed using data from two experiments with the same analysis routine. Principal component based noise reduction was used to improve the signal-to-noise ratio of raw spectra from annotated area.[18,19] PCA noise correction technique is defined as a statistical procedure concerned with elucidating the covariance structure of a set of variables[20]. PCA reduces dimensions of a large set of data based on vector space transformation. It develops artificial variables, which are known as the principal components based on variances of the dataset[21]. In terms of noise correction, only first few principal components, which best explains the dataset, were kept. The reason for that is that random noises also contribute to the variance of the spectrum, which can interfere with the actual subtle biological differences. Because of the nature of classification, differences apart from biological ones should be removed to improve the accuracy of the classifier. In this case, the first 80 principal components were kept to reach the best noise reduction results with the consideration of preserving most biological information.

Spectra were quality tested to remove data obtained from areas with little or no tissue based on the height of the amide I band. The eliminated spectra were either too weak or saturated, which provided unreliable results. Spectra have absorbance between 0.1 and 2 were selected and retained. Spectra with Amide I band intensity lower than 0.1 and higher than 2 were discarded.

Spectra ranges from 1000 to 1300 cm^{-1}, 1500 to 1750 cm^{-1}, and 3005 to 3550 cm^{-1} were taken. Regions describing the absorption bands of wax were removed.

Each spectrum was then vector normalized to correct for different thicknesses of breast tissue. In spectroscopic applications, scaling effects or differences could arise because of scattering effects, source and detector variations [7,22–25]. Vector normalisation was developed to correct all variables into the same scale, which gave all variables equal impact on the model. In this case, the vector normalisation was performed based on dividing each peak from the spectrum by the sum of the absolute value of the peaks from the same spectrum to scale every peak to the same level, which retained a vector with a unit area under the curve.

Finally, spectra were converted to the first derivative and performed a Savitzky–Golay smoothing using a window size of 19 data points. The advantage of analysing the first derivatives of spectra was that the original baselines were removed. The influence of baseline distortion was reduced largely, as the original information was converted to the variation on spectra slopes. Savitzky-Golay filtering was applied to further smooth the spectra. Savitzky-Golay smoothing algorithm was the most common method of improving the signal-to-noise ratio without distorting the signal through convolution of the spectrum [26]. This fitting method was first published in 1964 by Savitzky and Golay [27]. They proposed using the local least-squares polynomial to smooth data within the approximation interval [27,28]. They showed that the fitting result using the established model was equivalent to the discrete convolution with a fixed impulse response [27,28]. In this case, in order to improve the smoothing results, the least-squares polynomial smoothing was directly applied to the first derivatives of spectra with a window size of 19 data points, which could show subtle changes between cancerous and non-cancerous spectra, each data point was used to estimate the approximation interval.

The cancerous spectra were randomly selected to match the number of the spectra from the normal cores. An equal number of spectra could avoid the bias of the classifier caused by an unequal number of training sets.

Dataset was separated into two groups, a training set and an independent testing set. For model one, the training set contained 56 cores consisting of 47 cancerous cores and 9 NAT, while the independent group contained 14 cores consisting of 12 cancerous cores and two NAT. For model two, training set contained 55 cores consisting of 44 cancerous and 11 normal associated cores, while independent testing set contained 15 cores consisting of 13 cancerous and 2 normal associated cores. Each core was from a different patient. Cores used in the independent test were completely independent with the patients in the training.

4.2.4.3. Classification methods

4.2.4.3.1. Random forest classification

Random forest is a classification method operates by constructing decision trees to vote the class.[29][30] Each tree gives a classification. Each decision is collected and the forest chooses the most popular class-based on decisions made by all trees in the forest. This algorithm is developed by Breiman[29] in 1996 based on bagging predictors[31] with a random selection of features[32] to construct decision trees with higher accuracy of classification. Each tree is grown based on three rules. Firstly, if the number of cases is x in the training data set, sample *x* cases at random (can have or not have a replacement). Secondly, *m* features can be selected randomly out of the total number of features chosen for training decision trees, where *m* must smaller than the number of total features. The value of *m* is held constant during the forest growing. Lastly, pruning is not allowed during the training process.[29] All trees are grown to their largest possible extent. The forest error rate can be influenced by two aspects, including the correlation between two trees and the strength of individual tree in the forest.[31] An increase in the correlation between two trees leads to an increase of the forest error rate. In contrast, increasing strength of individual tree decreases the forest error rate of the classifier.

A flowchart of the mechanism of random forest classification is shown in Figure 4.4. In plot A, every decision tree in the forest is constructed based on a random bootstrap sample from the original data pool containing both cancerous (red) and non-cancerous spectra (blue). Plot B shows how the forest makes decisions. The class prediction for new samples is based on a majority voting process conducted by all individual trees. For example, the classification of a new sample X starts from the initial node, which is the root. Based on the features of X, the algorithm is carried out traverse down the tree. Each split nodes select the next branch to follow. This process is repeated until a leaf node, which makes a decision on classification, is reached. The node size was set to be 1 in this study, as small node size meant the trees would be grown deeper. Although it takes more memory and is slower, still it provides the best classification accuracy. After each tree has voted, the forest will assign the final prediction to the class has the most votes.

For this project, random forest classification was used to build a classifier which decided whether a core is cancerous or not, based on training spectra provided.

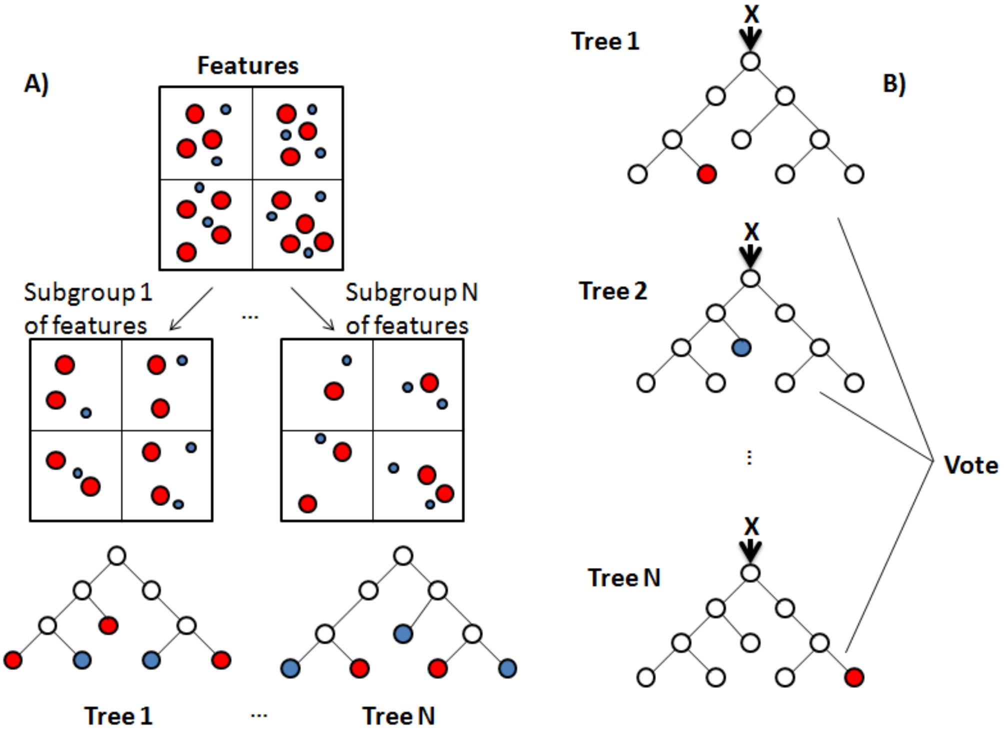


Figure 4.4 Random forest working plot A) feature selection working flow for each tree in the forest, where red and blue dots in feature pools represented different features extracted from samples while the red and blue filled circles in trees represented features used to train the branches. B) The voting flow of the forest where the red and blue coloured circles showed the class predictions of each tree. X is the input information to all the trees and the final decision is decided by each tree in the forest.

4.2.4.3.2. AdaBoost

Boosting is a machine learning approach based on the idea of creating a highly accurate prediction rule by ensemble many relatively weak and inaccurate rules.[33] The AdaBoost algorithm of Freund and Schapire [34], who won the 2003 Godel Prize for this work, was the first practical boosting algorithm, and the most widely used and studied boosting algorithm, which was applied to many different fields.

AdaBoost is usually applied to classification problem and the initial AdaBoost M1 works best on a binary classification problem. AdaBoost works by obtaining weighted majority votes of weak hypothesis where each hypothesis is assigned weight and conducted by every weak learner [35].

To be more specific, each weak learner is trained and tested on the randomly selected subset of training samples. Unlike random forest, AdaBoost focus on those features that contribute to less classification error, as evaluating every feature not only reduces the speed of model training process but also reduces the predictive power of the leaner. The misclassification errors are calculated and used to calculate the weight of samples in each training iteration. The misclassified samples are given higher weights, which increase the chosen possibility of this sample by the next classifier in its training subset with the aim of correctly classifying them in the next iteration. Weights are assigned to each trained weak classifiers according to their classification accuracy so that the more accurate classifiers have more impact in the final outcome. All learners are involved in weighted voting to allocate unknown samples.

AdaBoost can be combined with any other machine learning methods and boost their performance. However, it works best on weak leaners. It is claimed to have less chance of over-fitting comparing with other classification methods [33]. However, AdaBoost is sensitive to noise and outliers which can be considered as a disadvantage of the algorithm.

4.2.4.4. Statistical analysis methods

4.2.4.4.1. Confusion matrix

A confusion matrix is a table used for describing the performance of a classification model, which is often used in supervised testing, where true classes of samples are known.

Table 4.1 Sample diagram of confusion matrix

	Prediction	
Actual label	True positive	False positive
	False negative	True negative

Where predictions are labels predicted by the classifier on samples while the actual labels are the true labels of testing samples. Numerical labels are usually used to identify the class of samples during data processing. For example, in this chapter, the cancerous epithelium was labelled as class one (1) while the NAT epithelium was labelled as class two (2) in classification process. In this case, true positive (TP) means the predictions and true labels are the same and both belong to the main class, which is class one also known as positive class. True negative (TN) means the predictions are correct and also negative or belong to class two. False positive (FP) means that the positive predictions are wrong. These samples are from the negative class. False negative (FN) means that negative predictions are wrong. These classified samples are actually from the positive class, or class one.

Sensitivity and specificity can be calculated from the confusion matrix. The sensitivity (also known as the true positive rate) equals to the quotient of TP and the total number of actual positive samples. Sensitivity is often used to describe when it is actually yes, how often it predicts yes. The specificity (also known as the true negative rate) equals to the quotient of TN and the total number of true negative samples. It also equals to 1-the false positive rate. Specificity is used to describe when it is no, how often it predicts no.

4.2.4.4.2. ROC curve

Receiver operator characteristic (ROC) curves [36] are a common way of representing the inherent trade-off between sensitivity and specificity, and it helps to visualize the performance of a classifier. A ROC curve is plotted by the true positive rate (TPR) again the false positive rate (FPR). It can also be plotted by sensitivity against 1-specificity. The ROC curve virtualizes all predicted probabilities of the positive class by the model, although these probabilities are not directly shown on the curve. Construction of ROC curves for a binary system involves comparing the true positive rate to the false positive rate as the discrimination threshold is varied.

However, the ROC curve also has its disadvantage, which is that ROC curve is not sensitive to whether a full range of predicted probabilities is given. It means that the ROC curves of a model with the predicted probability ranges from 0.5 to 1 and 0 to 1 can be identical if the order of observations stays the same. The ROC curve is more focused on ranking of samples based on their predicted probabilities. A sample ROC curve plot was shown in Figure 4.5. When the ROC curve hugs the upper left corner of the plot, the model did a great job on classification. If the curve is close to the black diagonal, it usually indicates that the model did not perform well.

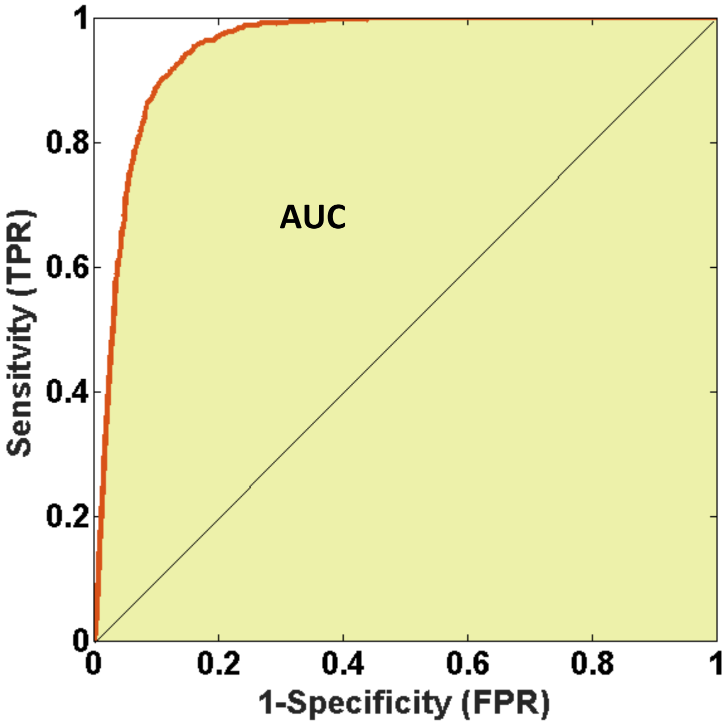


Figure 4.5 Sample plot of the ROC curve, where the red curve is the ROC curve and the yellow area under the red curve is the AUC of the sample ROC curve

4.2.4.4.3. AUC

The Area under curve (AUC) of the ROC curve is also used to describe the performance of a model in a single number. AUC is literally the percentage of area under the ROC curve over the whole area of the box, which is also shown in Figure 4.5. It quantifies the performance of a model. The range of AUC is from 0 to 1. Under the extreme circumstance, AUC equals to 0, the model reciprocates the classes. It means that the model predicts all the negative class as a positive class and vice versa. When AUC equals to 1, the classifier works perfectly and an ideal separation between two classes is obtained. Values close to 1 often indicates the relatively good performance of the model. If a classifier gets AUC value close to 0.5, it means the classifier works poorly on the test data set. As AUC is the area under ROC curve, it is also only sensitive to the ranking of samples based on their predicted probabilities.

4.3. Results and discussion

Models were constructed using data collected from both experiment one and two. Firstly, for data from experiment one, model was established on cancer diagnosis using epithelium spectra. In the light of model one, based on data collected from experiment two, models were constructed for both digital histology and cancer diagnosis using both epithelium and stroma spectra. Random forest and AdaBoost were used as main classification methods.

4.3.1. Experiment one on Br20832

In term of training a machine learning model, random forest, spectra subtracted from data cubes based on annotations were 78577 in total, which contained 59998 cancerous epithelium spectra and 18579 NAT epithelium spectra. Under-sampling was conducted to balance the number of spectra in each class. 18579 cancerous spectra were randomly selected from 59998 spectra to match the number of spectra in NAT class. Each class had the same number of spectra and no bias in either class would be made. 570 features were used to construct models.

In term of the independent test, 18819 spectra were pulled from data cubes, including 16455 cancerous epithelium spectra and 2364 NAT epithelium spectra.

Before classification started, the out-of-bag rate plot of the model was generated and shown in Figure 4.6. The out of bag (OOB) rate was also called out of bag error. It was a method to evaluate the prediction errors for machine learning models using subsets of training data. OOB was the mean prediction error on each training sample. It helps to select proper number of trees used for model training, which can largely avoid the unnecessary training time and achieve the best possible results.

It was shown in Figure 4.6 that the OOB error drops rapidly from 0 to 100 trees from around 0.04 to below 0.005. The error rate barely decreased from 100 to 500 trees, which indicated that the model was very close to its limits. To save training time, 100 trees were used to establish the model.

In terms of the number of features selected for each tree, the square root of the total number of features was applied, as it led to the best classification results [38]. To be more specific, the number of randomly selected features could affect the overall classification error rates. If too many features were picked for each tree, although the strength of that tree would increase, still more correlation between each individual tree would decrease the overall performance of the model as a forest. However, if too few features were selected, each individual tree would be weaker, but less correlation among trees would lead to a stronger model. In this case, the total number of features was 570. Therefore, the feature input for each tree was 23.

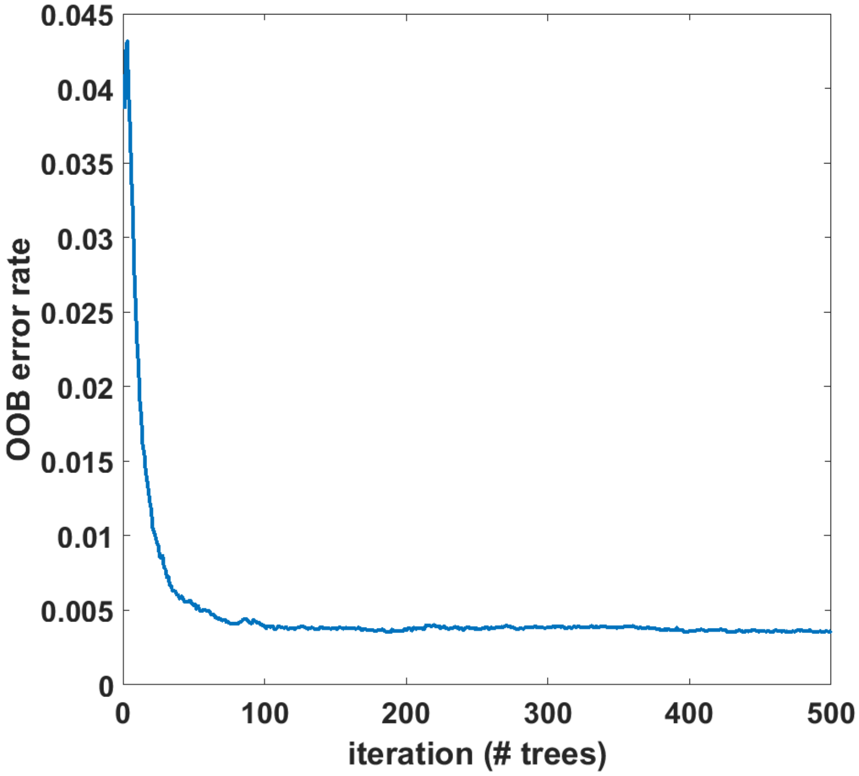


Figure 4.6 Out of bag rate plot of the random forest model based on data collected in experiment one

The model was constructed based on 100 trees with 23 feature input for each of them. The model was validated with 20% randomly selected spectra from the training dataset. To be more specific, 80% randomly selected spectra were used for model construction while the 20% which were not involved in the training process, was applied to validate the model. The confusion matrix of the training validation process was shown in Table 4.2. From the table, it could be seen that the true positive accuracy was 99.6% and the true negative accuracy was 99.7%, which were both very high. General classification accuracy of 99.7% was obtained, which indicates that the model worked perfectly on the subset of training data.

Table 4.2 Training validation test for model one

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	99.6	0.46
	NAT	0.3	99.7

An independent test was conducted based on spectra selected from patients different from the training process. The confusion matrix of the classification was shown in Table 4.3. From the table, it was clear that the true positive classification accuracy for cancer was 99% while the classification accuracy for the NAT class was 98.4%. The overall classification accuracy was 99.0%. A nearly perfect separation was observed.

Table 4.3 Independent test for model one

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	99.0	1.0
	NAT	1.6	98.4

The ROC curve of the model was plotted and was shown in Figure 4.7. Considering the similarities between malignant and non-malignant spectra, the ROC curves appeared surprisingly good. The resulting AUC value of 0.98 indicates that there was nearly perfect differentiation between malignant and non-malignant pixels.

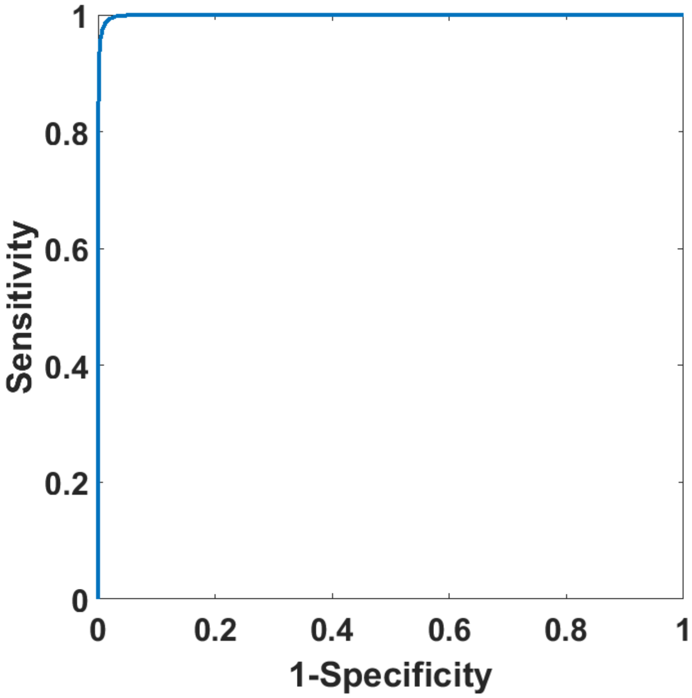


Figure 4.7 ROC curve of the random forest model trained on data collected in experiment one

To test whether the classification was valid, the importance of each feature was plotted, which was shown in Figure 4.8. The importance plot indicated the relative importance of each input feature for a particular model. It computed estimates of predictor importance by summing these estimates over all weak learners in the ensemble and was calculated by summing changes due to splits on every predictor and dividing the sum by the number of branch nodes. Figure 4.8, indicates that the classification was mainly caused by the variations of spectra in the range between 1500 and 1750 cm^{-1}. This region of the spectrum was enlarged and shown in Figure 4.8. It could be observed that the main peaks were 1734 (C=O stretching), 1717 (C=O stretching), 1699 (C=O guanine[39]), 1559 (Amide II), and 1507 cm^{-1} (Amide II). The main features responsible for this separation were mainly variations in the amide I and II. However, the number of sub-peaks in the importance plot was too many and the plot was too spiky. Water vapour was suspected to be involved in the classification. To further investigate, 10 randomly selected raw spectra from the training group in the range of 1500 and 1750 cm^{-1} were shown in Figure 4.10 and Figure 4.11.

Closely observed spectra randomly selected from cancerous cores in Figure 4.10. In the range of 1500 to 1750 cm^{-1}, it was clear that the Amide I peak spat into two sub-peaks, at 1649 and 1657 cm^{-1}. This observation indicated that there was appearance of water vapour during data collection process. From Figure 4.11, it could be observed that a bump appeared next to the Amide I peak around 1638 cm^{-1}, which also indicates the influence of water vapour on the spectra.

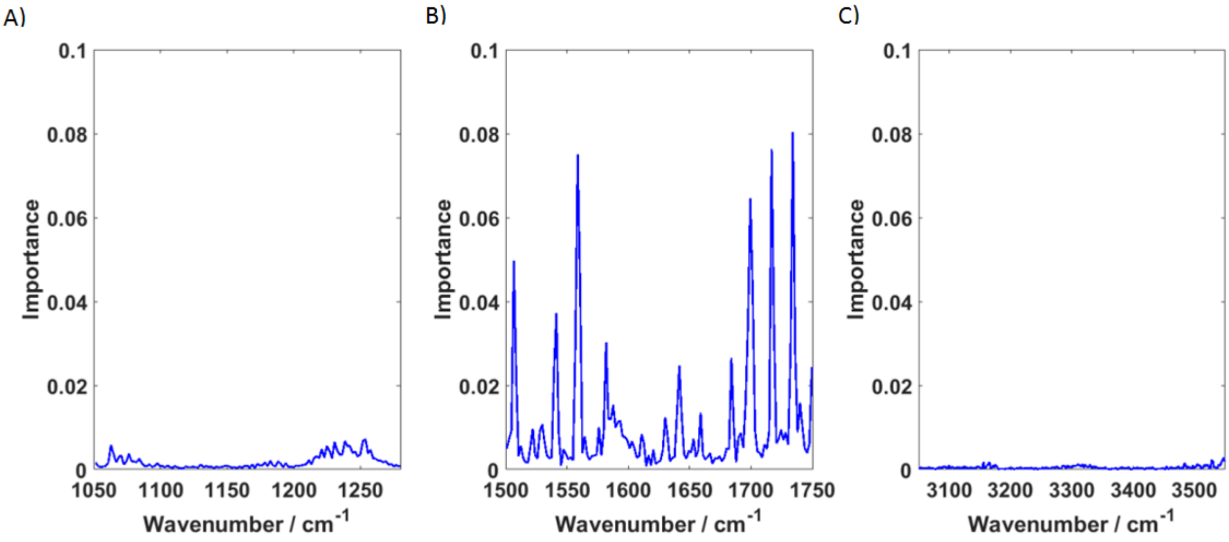


Figure 4.8 Importance plot of the random forest model trained on the data collected in experiment one in full scale, where A) wavenumber range 1050 to 1300 cm^{-1}, B) wavenumber range 1500 to 1750 cm^{-1}, and C) wavenumber range 3005 to 3550 cm^{-1}

The influence of water vapour on the NAT cores was slightly smaller than the cancerous cores, which proves that this water vapour content differences among cancerous and NAT cores were obtained during measurements as cancerous cores were collected before NAT cores based on the sequence of core arrangement during mosaic collection. The duration of taking a large mosaic in this experiment was around 5 hours.

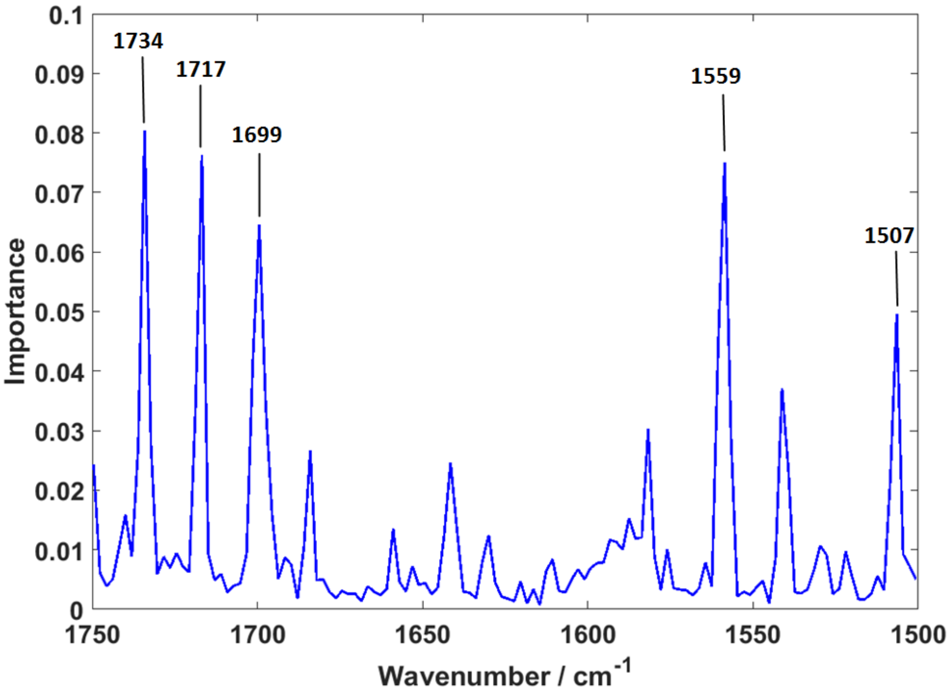


Figure 4.9 Enlarged importance plot of the random forest model trained on the spectra collected from experiment one in the range of 1500-1750 cm^{-1}

For each mosaic, cores were measured in columns. Cancerous cores were always taken first while the normal cores were in the last row of the slide therefore taken at the last. The large time difference between the first taken cancerous core and the last taken NAT core could further enhance the effect of water vapour appearance. Although the experiments started when the water vapour content was below 1%, still the long time gap between measurements caused problem. The perfect classification could be based on the difference between water vapour contents of cancerous and normal cores.

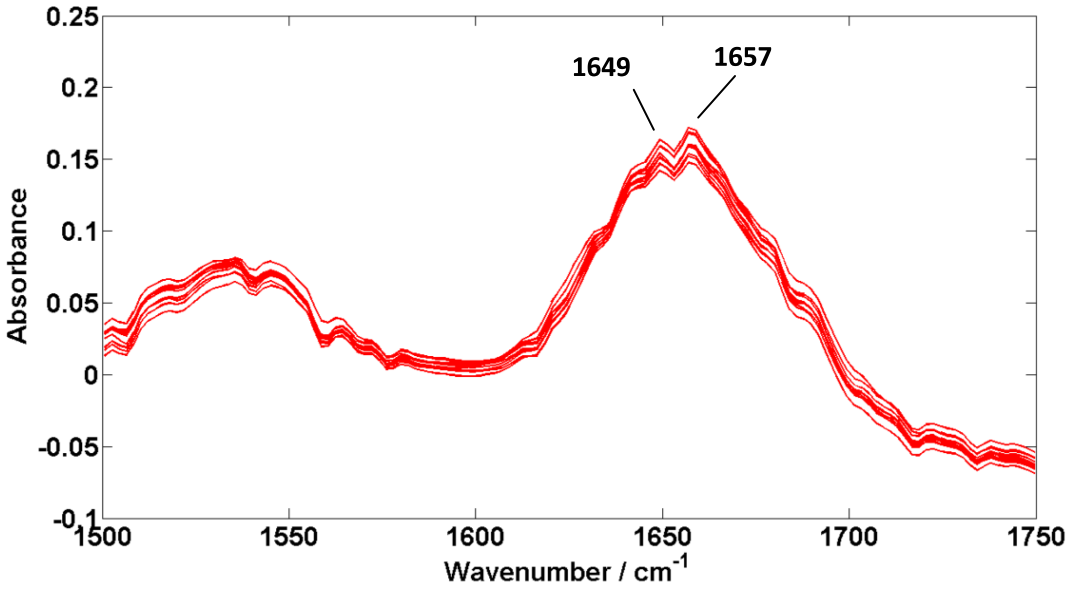


Figure 4.10 Ten randomly selected spectra collected in experiment one from cancerous cores in the range of 1500-1750 cm^{-1}

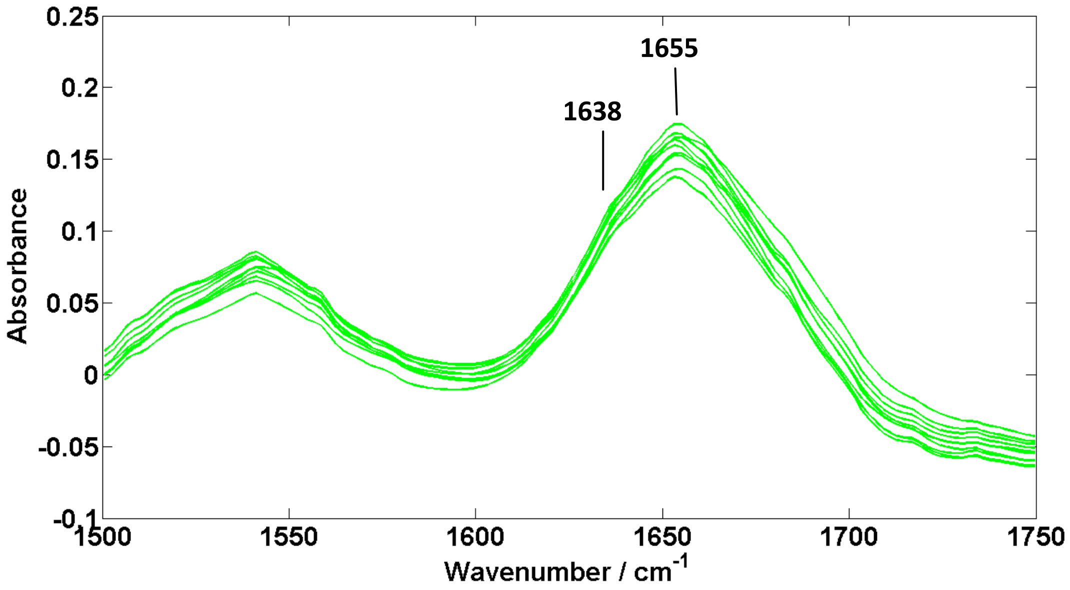


Figure 4.11 Ten randomly selected spectra collected in experiment one from NAT cores in the range of 1500-1750 cm^{-1}

4.3.1.1. Sectional conclusion

It could be concluded that the raw data used for analysis has variance in water vapour content of each core, which caused difficulty in picking up subtle biological information carried by spectra. Water vapour spectra covers range from 1400 to 1900 cm^{-1}, where Amide I and II [39] were also covered, and 3500 to 3900 cm^{-1}[40].

The original biological information carried by the sample in Amide I and II would be interfered and disrupted by water vapour spectra. Even through the background should ratio out the water vapour influence, it only applies if the water vapour concentration in the purge box stays at the same level. Therefore, for future experiments extra care should be paid on monitoring the humidity level during data collection process.

Although classifiers constructed could not separate cancerous and normal cores based on biological differences, still construction process itself was effective. New data with less water vapour influence collected in the experiment two was used for the next section.

4.3.2. Experiment two on Br20832

4.3.2.1. Epithelium and stroma annotation check using AdaBoost model

A quick check on the accuracy of annotation used for the research was conducted by constructing a classification model on stroma and epithelium spectra. The correctly classified results were projected back to their original position on each core to visualise if the classification results were consistent with the initial annotation. AdaBoost was used to establish the model with 500 trees and 0.1 learning rate and trained for 500 iterations.

Spectra after pre-processing were passed into the model, which contains 90270 epithelium and 123104 stroma spectra. Under-sampling was applied and 90270 stroma spectra were randomly selected to balance the number of the epithelium spectra. An AdaBoost model was established and 20% of the training data was left from training and used for the training validation test of the epithelium and stroma classification model. The training validation test was presented in the Table 4.4. It could be seen that the classification accuracy of epithelium spectra was 98.7% and 98.5% stroma spectra in the training validation set were correctly identified. The overall classification accuracy was 98.6%, which was an excellent result considering the training validation group has a large number of spectra in it.

Table 4.4 Training validation test of the epithelium and stroma classifier using AdaBoost

Training Validation		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	98.7	1.3
	Stroma	1.5	98.5

An independent test was conducted to further validate the reliability of the model. 20041 epithelium and 27141 stroma spectra were passed and classified by the model. The confusion matrix of results was shown in the Table 4.5. The classification accuracy of epithelium spectra was very high, 97.2%. The accuracy of separating stroma spectra was slightly lower, 90.3%. However, considering these spectra were from independent patients from training group this was still a good result. Overall classification accuracy of 93.2% indicates great separation.

Table 4.5 Independent test of the epithelium and stroma classifier using AdaBoost

Independent Test		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	97.2	2.8
	Stroma	9.7	90.3

The ROC curve and feature importance plot were plotted in the Figure 4.12 and Figure 4.13 respectively. The ROC curve was smooth and hugged the top left corner of the plot box, which shows the model performed well. The AUC of the curve was 0.99, which again shows the model works excellently. In Figure 4.13, the importance of the features used to separate epithelium and stroma spectra in the model was shown covering the full spectral the full range. It could be seen that main peaks were observed in the range of 1050 -1300 cm^{-1} and 3000 - 3550 cm^{-1}. The enlarged plots of these two ranges were shown in Figure 4.14 and Figure 4.15.

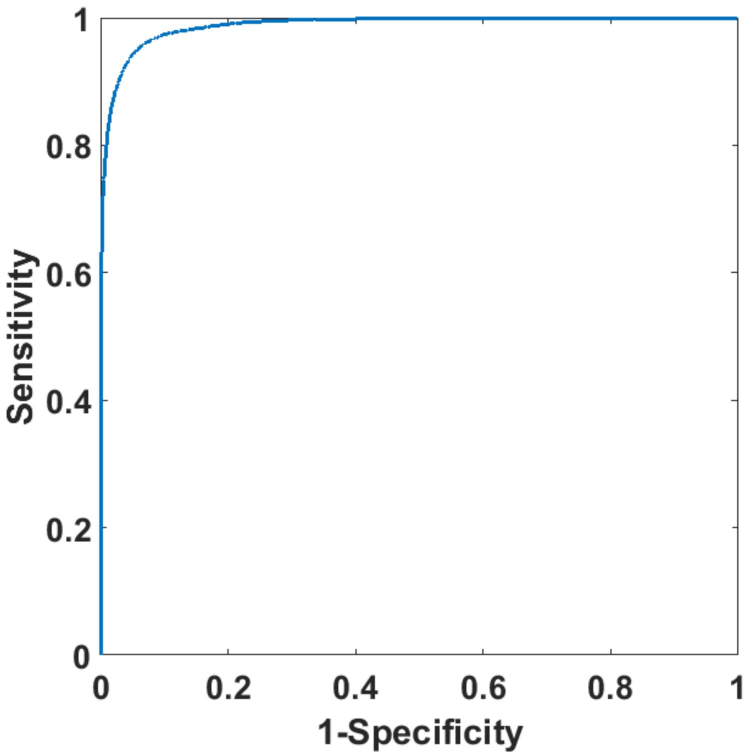


Figure 4.12 ROC curve of the epithelium and stroma classifier using AdaBoost trained on the data collected in experiment two

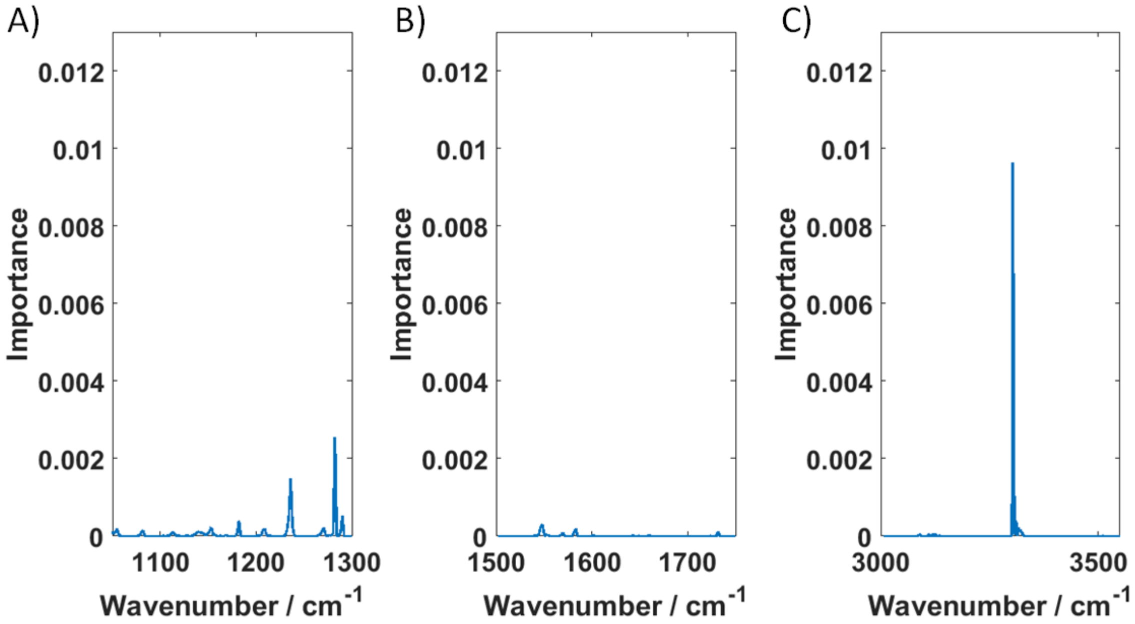


Figure 4.13 Feature importance of the classifier-full range for the epithelium and stroma classifier using AdaBoost trained on spectra collected in experiment two, where A) wavenumber range 1050 to 1300 cm^{-1}, B) wavenumber range 1500 to 1750 cm^{-1}, and C) wavenumber range 3005 to 3550 cm^{-1}

In Figure 4.14, only one major peak was observed at the position of 3303 cm^{-1}, which represents Amide A band. From Figure 4.13, it could be seen that peak at 3303 cm^{-1} contributed mostly on the separation between epithelium and stroma. There was no trace of water vapour observed. Minor peaks at 1282 (Amide III band components of proteins) and 1236 cm^{-1} (Stretching PO_2^- asymmetric) were observed, which also largely contributed to the model classification.

The correctly classified epithelium and stroma spectra were projected back to their original positions in each of the independent test cores and the resultant image shown in the Figure 4.16. The lighter blue represented the correctly classified pixels while the darker blue indicated spectra that passed the quality test but were misclassified by the model. For better comparison, the annotated version of cores in the independent test was also shown in Figure 4.17. It could be seen that the epithelium annotations were very consistent with the classification results. Stroma spectra also showed good performance, however, not as good as epithelium. This might be due to mislabelled stroma annotations for some cores or the complexity of stroma tissue as it contained different types of cells. The general features of stroma could be harder to find comparing with the single cell typed epithelium spectra.

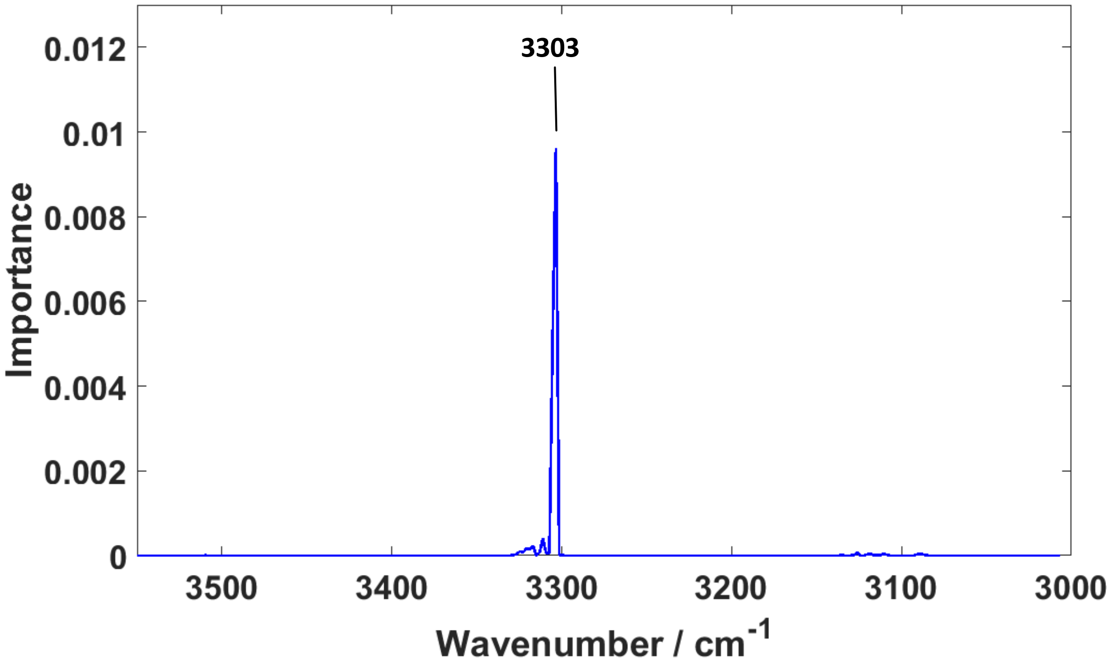


Figure 4.14 Feature importance of the classifier (3000 -3550 cm^{-1}) for epithelium and stroma using AdaBoost based on data collected in experiment two

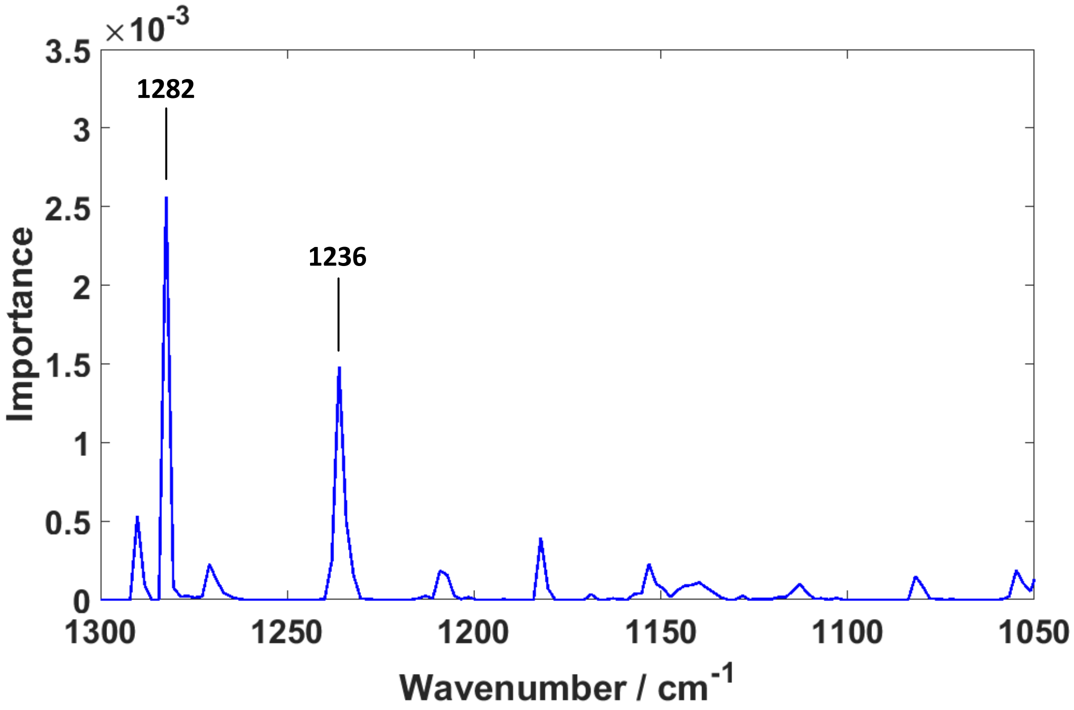


Figure 4.15 Feature importance of the classifier (1050 -1300 cm^{-1}) for epithelium and stroma using AdaBoost based on the spectra collected in the experiment two

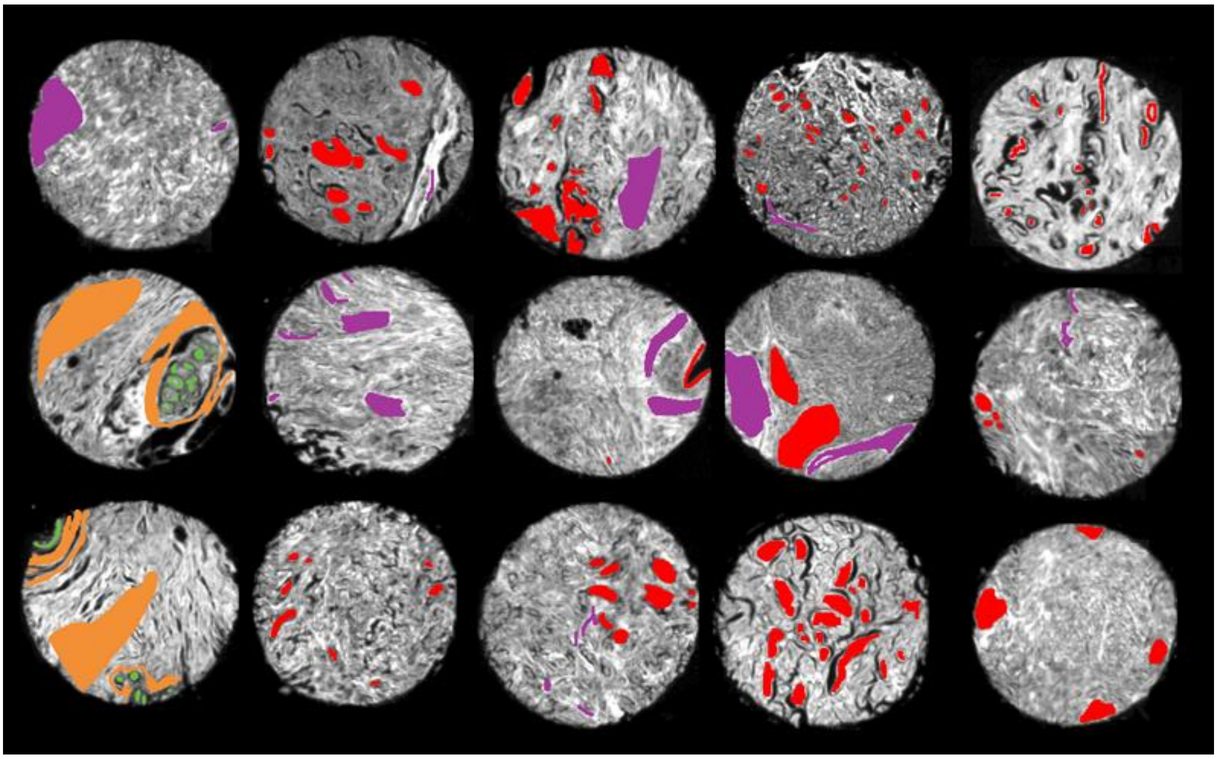


Figure 4.17 Original annotation of each core in the independent test, where red and green indicated cancerous and normal epithelium respectively while purple and yellow showed the cancerous and normal stromal regions of each core

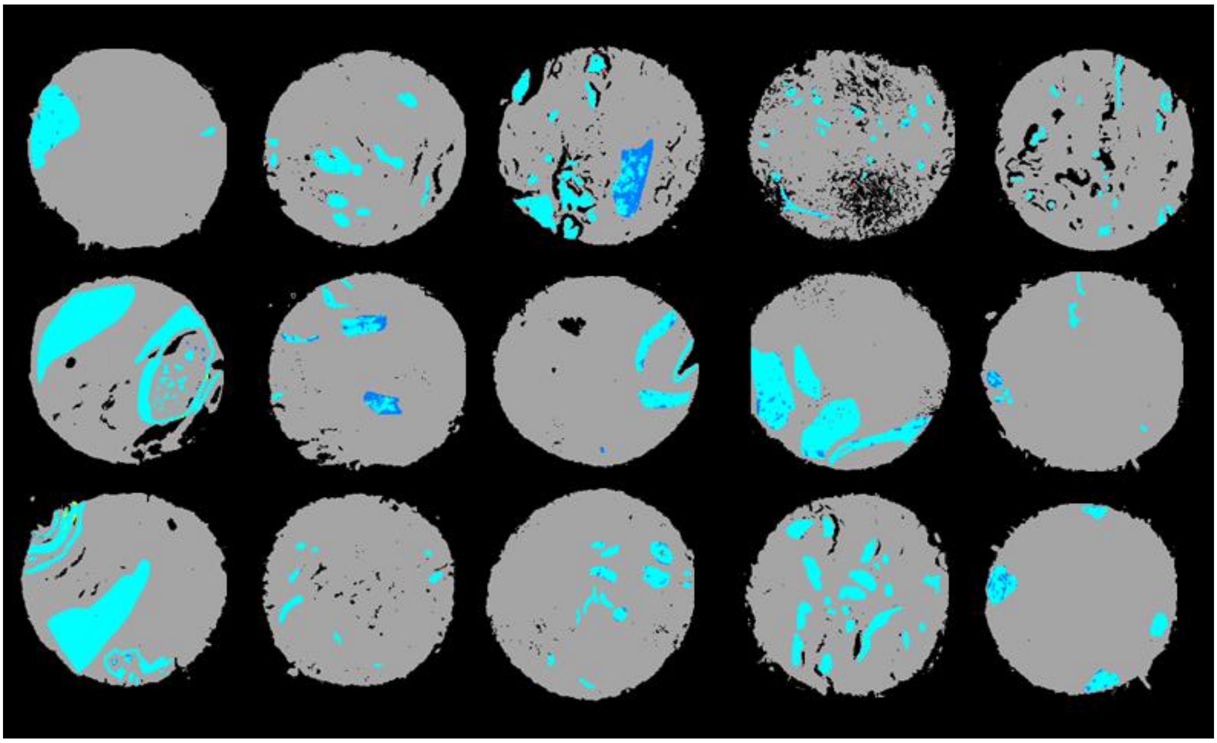


Figure 4.16 Correctly classified spectra projected back to their original location in each core, where the bright blue indicated the correctly classified pixels while the dark blue showed the locations of misclassified spectra

All spectra in each core were pre-processed and passed into the epithelium and stroma classification model. A digital staining plot of cores in the independent test was generated and presented in Figure 4.18. The plot was generated by projecting the classification results (epithelium or stroma) of each spectrum from every pixel on all cores in the independent test group. Red colour represented appearance of epithelium cells and purple indicated stroma content for cancerous cores, while yellow stood for stroma and green showed the presence of epithelium cells in NAT cores. This digital stain could be further improved tuning the model and adding more cell types as input for future studies.

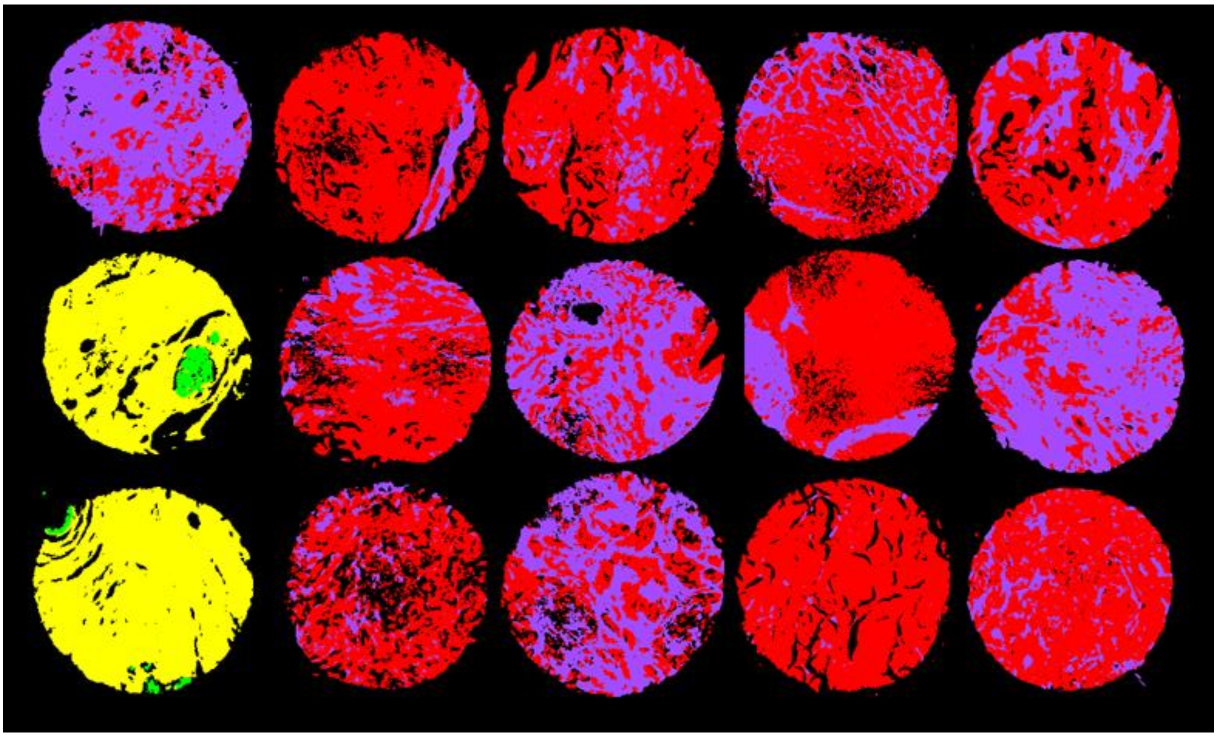


Figure 4.18 Digital histology of cores in the independent test group, where red and green stood for cancerous and normal epithelium while purple and yellow showed cancerous and normal stroma in each core

4.3.2.2. Cancer prediction using classification models

After checking that the annotations were mostly reliable, models were trained on cancer diagnosis using both random forest and AdaBoost. Models were tuned for research proposes. However, considering the time cost and dependence of fine turned hyper-parameters of each machine learning models on the training data structures, a trade-off has been made between the relatively small accuracy increases and considering large training time consumptions. The objective of tuning was to maximise the classification accuracy of models on the same independent test sets for each method. After comparing the performance of each model, class prediction of independent cores were presented in the traffic light cancer diagnosis system using the most effective model between two methods (both based on epithelium and stroma spectral information).

4.3.2.2.1. Random forest models

Preliminary models based on the random forest classification method were trained and independent tested on both epithelium and stroma spectra.

4.3.2.2.1.1. Based on epitheliumspectra

In term of model training, 70932 cancerous and 19338 NAT epithelium spectra were selected based on the annotations tested previously. Under-sampling was conducted and 19338 cancerous spectra were randomly selected. The total number of spectral input was 38676. 570 features were passed onto the model training as well. In order to obtain reasonable classification results, number of features input was set to be 23, which was the square root of total number of features [38]. To both save time and ensure model performance, the OOB plot of the model was generated and shown in the Figure 4.19. It could be seen that the classification error rate remains stable around 0.01 after 100 trees were used. No further error reduction was observed with an increase of number of trees used in training.

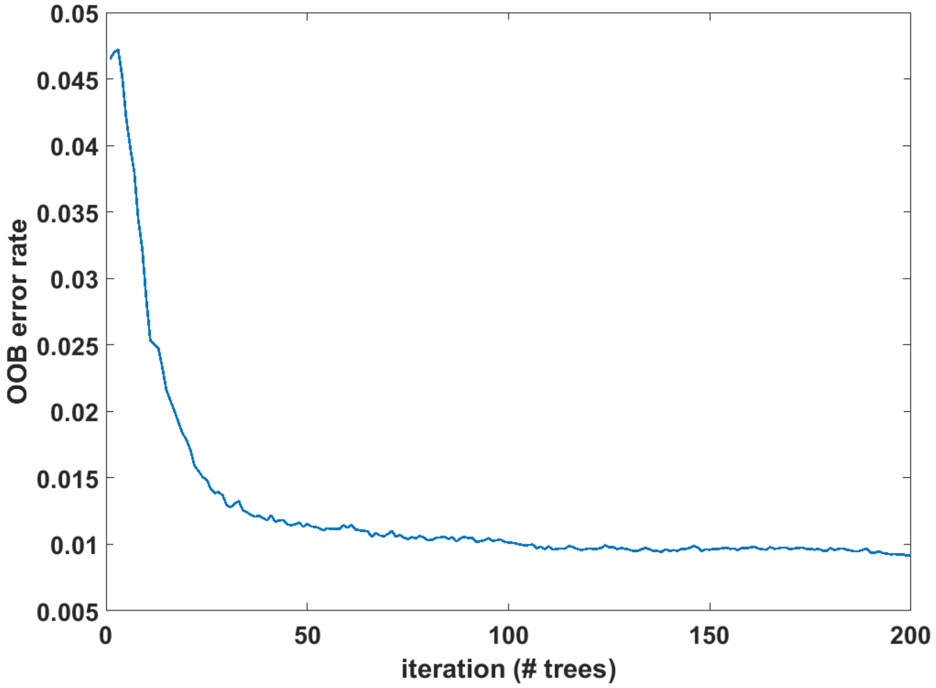


Figure 4.19 OOB rate plot of the random forest model based on epithelium spectra collected in experiment two

The random forest model was constructed using 100 trees and 23 feature input for each of the tree. It was validated with 20% randomly selected spectra from the training dataset. To be more specific, 80% randomly selected spectra were used for model construction while the 20% which was not involved in the training process was applied to validate the model. The confusion matrix of the training validation process was shown in the Table 4.6. From the table, it could be seen that the true positive accuracy was 98.8% and the true negative accuracy was 99.7%, which were both very high. A general classification accuracy of 99.3% was obtained, which indicated that the model worked very well on the subset of training data.

Table 4.6 Confusion matrix of the training validation for cancer NAT classification using random forest based on data collected in experiment two

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	98.8	1.2
	NAT	0.3	99.7

An independent test was conducted on the model constructed previously to further validate it. 20041 test spectra which contained 19291 cancerous spectra and 750 normal spectra were pre-processed and tested by the model. The confusion matrix of classification results was shown in Table 4.7. From the table, it was clear that the true positive classification accuracy for cancer was 77.5% while the classification accuracy for the NAT class was 86.4%. The overall classification accuracy was 77.8%. A reasonable separation was observed between cancerous and NAT classes.

Table 4.7 Confusion matrix of the independent test for cancer NAT classification using random forest based on data collected in experiment two

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	77.5	22.5
	NAT	13.6	86.4

The ROC curve of the model was plotted and was shown in Figure 4.20. Considering the similarities between malignant and non-malignant spectra, the ROC curves appeared relatively good. The curve was seen to approach very close to the top left corner of the plot, which indicated good classification. In addition, the resulting AUC value of 0.89 further demonstrated that a reasonable separation was observed between malignant and non-malignant pixels.

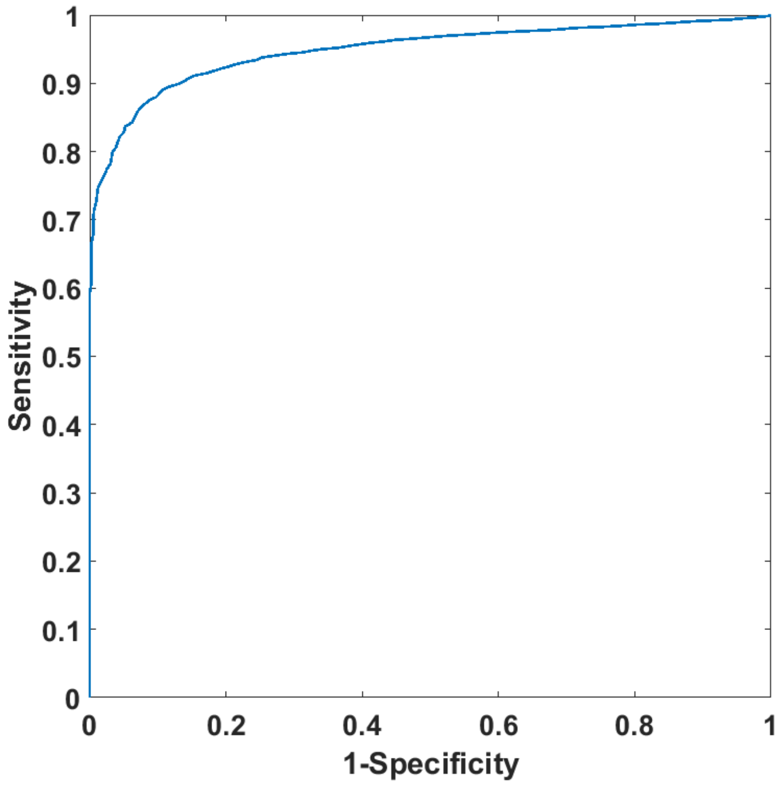


Figure 4.20 ROC curve of the independent test for cancer NAT classification using random forest based on epithelium spectra collected in experiment two

The importance of each feature was plotted, and was shown in Figure 4.21. Figure 4.21, it indicates that the classification was mainly caused by the variations of spectra in the range between 1050-1300 cm^{-1} and 1500-1750 cm^{-1}, which were enlarged and shown in Figure 4.22 and Figure 4.23 respectively.

It could be seen in Figure 4.22, one main peak at 1211 cm^{-1} (phosphate I) and in Figure 4.23 the main peaks at 1658 (Amide I), 1643 (Amide I), and 1543 cm^{-1} (Amide II). The main features responsible for this separation were mainly variations in amide I and II, especially amide I. Overall, although the main peaks of the feature importance plots were based on biological information, still many minor peaks were observed which could indicate noise or other interferences.

The independent test result was not as good as expected but the overall classification accuracy was still over 75.4% [41] which was the pathologist cancer diagnosis accuracy in 2010.

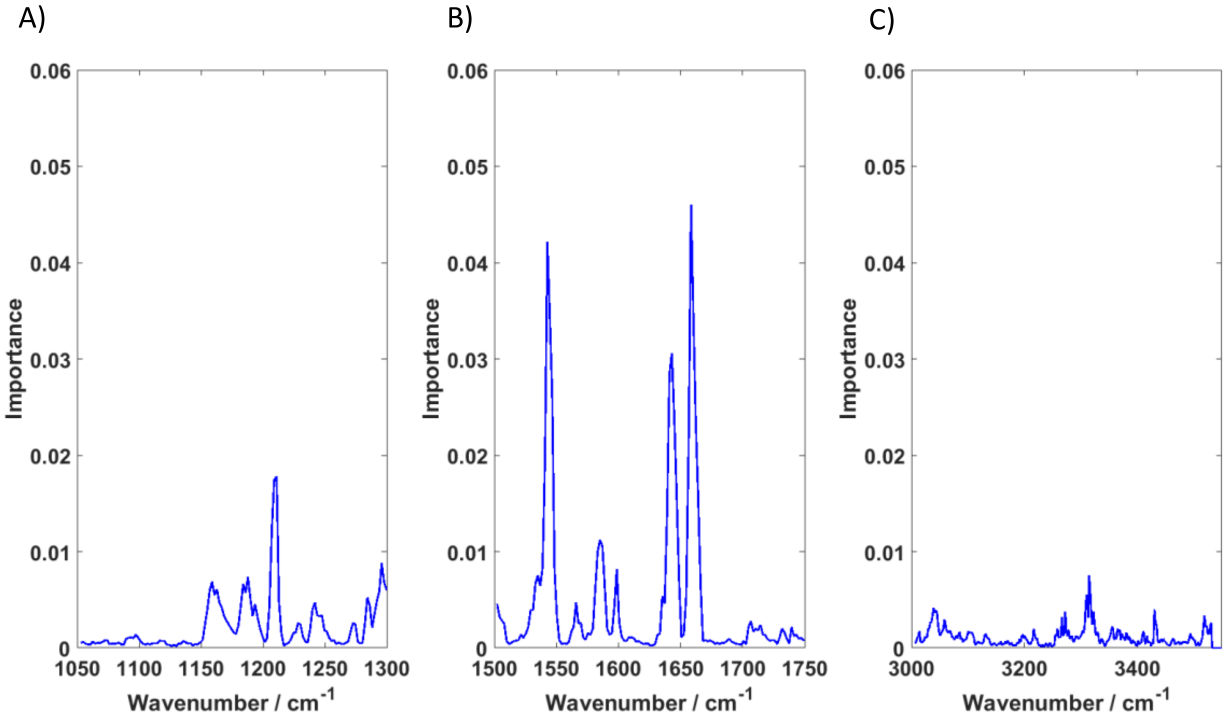


Figure 4.21 Feature importance plot (full range) of cancer NAT model based on epithelium spectra collected in experiment two using random forest, where A) wavenumber range 1050 to 1300 cm^{-1}, B) wavenumber range 1500 to 1750 cm^{-1}, and C) wavenumber range 3005 to 3550 cm^{-1}

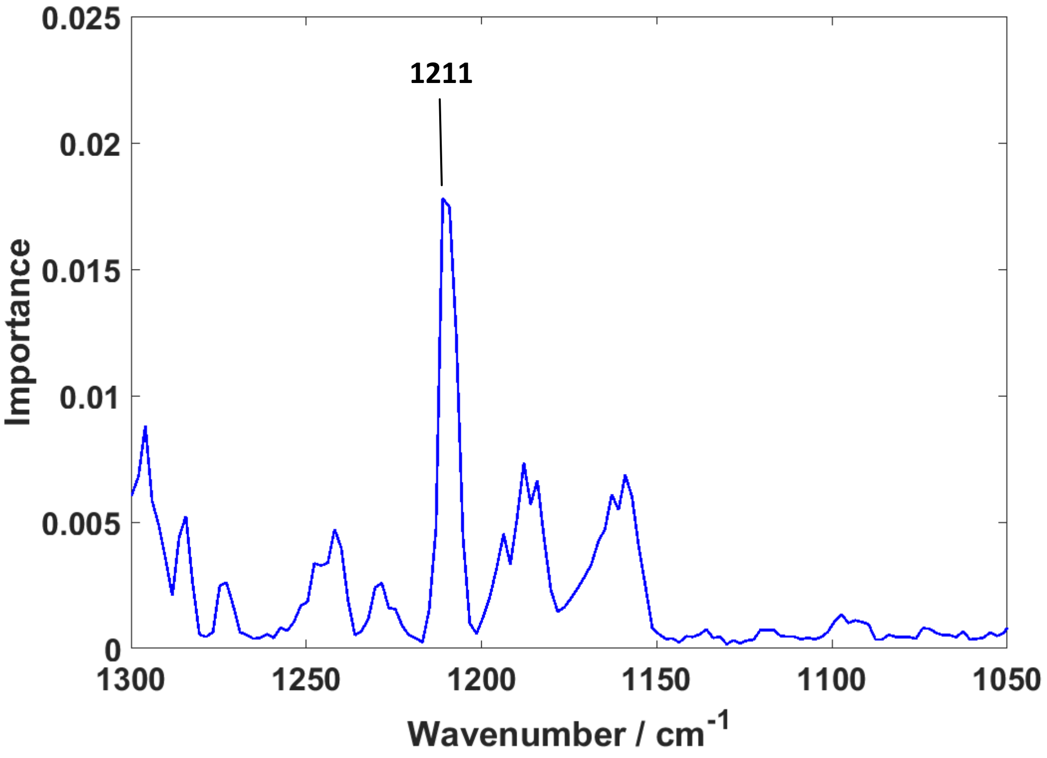


Figure 4.22 Feature importance plot (1050-1300 cm^{-1}) of cancer NAT model based on epithelium spectra collected in experiment two using random forest

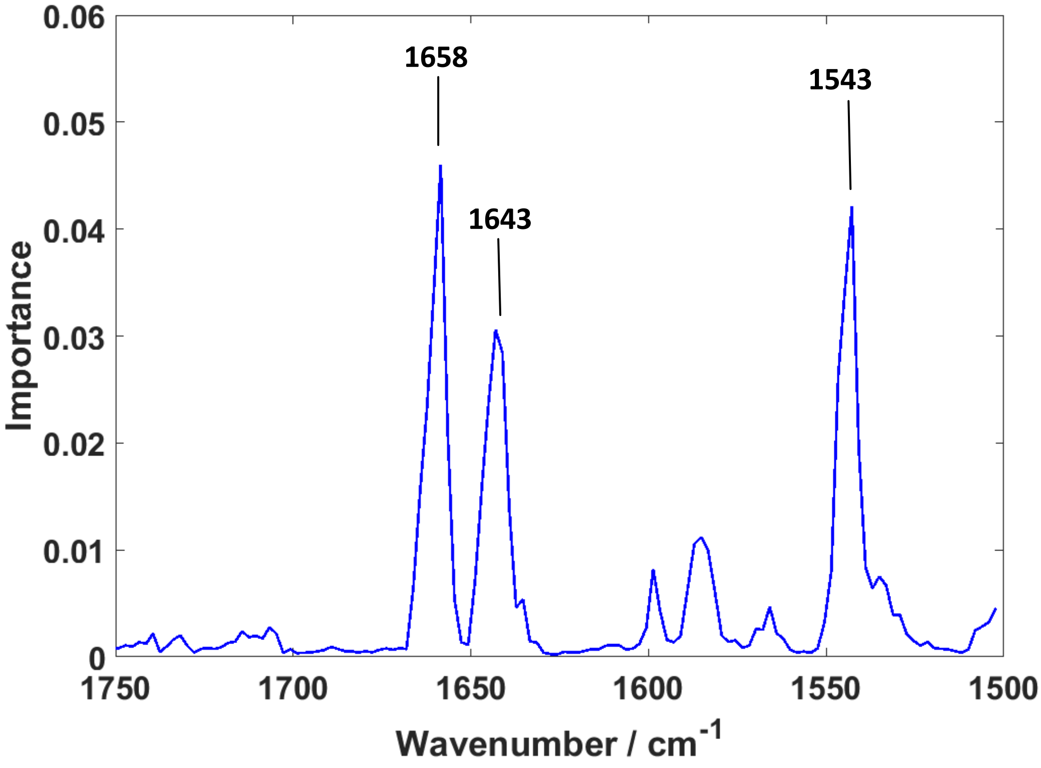


Figure 4.23 Feature importance plot (1500-1750 cm^{-1}) of cancer NAT model based on epithelium spectra collected in experiment two using random forest

4.3.2.2.1.2. Based on stroma spectra

Recent studies have suggested that stroma, specifically adjacent stroma, may have a key role to play in the initiation and progression of cancer. Because normal breast tissue was typically not particularly glandular, this limits the number of epithelium spectra which could be used for training and testing the classifier. A model was constructed based on stroma using random forest to investigate whether it could be used as a replacement of epithelium for cancer diagnosis.

A training database was constructed from the 55 training cores which consisted of 123104 stroma spectra (50918 normal spectra, and 72186 cancerous spectra). The optimal situation to prevent classifier bias was for each class to consist of equal numbers of spectra. Bias was minimized in the training set by selecting equal numbers of spectra from class with the number randomly selected being the size of the smallest class. A Random Forest was then constructed using 50918 spectra per class. There were two classes, including normal class and cancerous class. Each spectrum was noise reduced, quality checked, cut to the correct range and first derived in this sequence prior to being used for training.

To both save time and ensure model performance, the OOB plot of the model was generated and shown in the Figure 4.24. It could be seen that the classification error rate remained stable around 0.02 after 200 trees were used. No further distinct error reduction was observed with an increase of number of trees used in training.

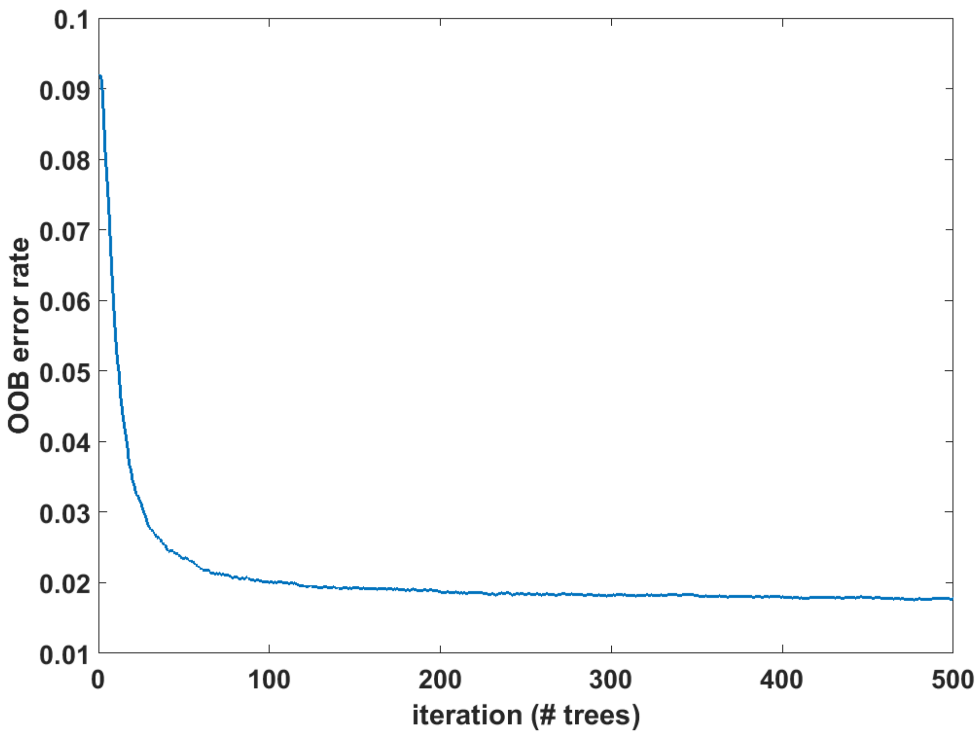


Figure 4.24 OOB plot for cancer diagnosis model based on stroma spectra using random forest

The random forest model was constructed by using 200 trees and 23 features per tree to balance the trade-off between training time and classification accuracy.

Confusion matrices provided a quantitative measure of performance and enable the correctness of classification for each class to be determined. Furthermore, the sources of misclassification could be easily identified from a confusion matrix, revealing which classes were difficult to discriminate between. The confusion matrix of training validation test of the model was shown in the Table 4.8. It could be seen that a 98.8% classification accuracy of cancerous class was observed and 96.4% stroma spectra from training dataset were correctly classified. Considering the training validation dataset was a relatively large data set containing 20367 randomly selected spectra, the training validation results were acceptable comparing with the classification accuracies obtained using epithelium spectra, shown in Table 4.6.

Table 4.8 Confusion Matrix of training validation test for cancer diagnosis model based on stroma using random forest

Training validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	98.8	1.2
	NAT	3.6	96.4

The Random Forest classifier was then tested on the independent test set which consisted of 27141 epithelium spectra (13157 normal spectra and 13985 cancerous spectra) from the 15 testing cores. Each tree in the Random Forest "voted" for the class which it predicted an unknown spectrum belongs to. The Random Forest then chose the class having the most votes over all the trees in the forest. The confusion matrices for independent test was generated and shown in the Table 4.9.

Table 4.9 Confusion Matrix of the independent test for cancer diagnosis model based on stroma using random forest

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	91.2	8.8
	NAT	45.4	54.6

From Table 4.9, it could be seen that the classification accuracy of the cancer class was very high, 91.2%, while only 54.6% NAT spectra were correctly classified. This was due to that the general feature of NAT was much harder to find compared with the general features of cancerous spectra. The constructed model could focus on the general similarities among training patients but not the general features of NAT cores. In term of cancer diagnosis, this model was very sensitive to cancerous spectra.

The ROC curve of the model was plotted and presented in the Figure 4.25. Considering the similarities between malignant and non-malignant stromal spectra, the ROC curves appeared acceptable. The resulting AUC value of 0.87 indicated that there was good differentiation between malignant and non-malignant pixels, considering the AUC obtained using epithelium spectra was 0.89, no large difference observed.

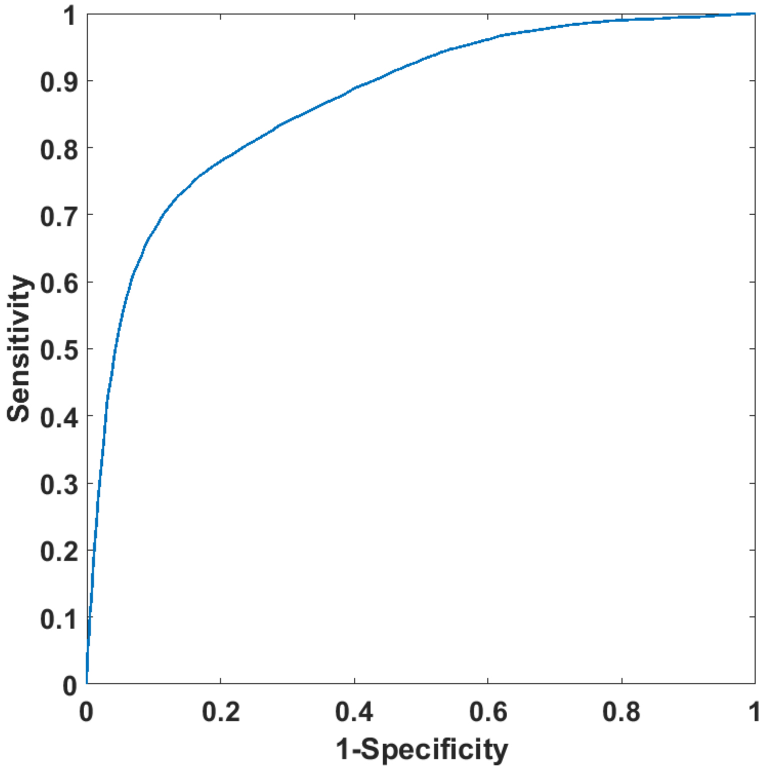


Figure 4.25 ROC curve of cancer diagnosis random forest model based on stroma spectra using random forest

The contribution of each feature in the classification process was plotted and shown in Figure 4.26. In Figure 4.26, it could be seen that the classification was mainly caused by the variations of spectra in the range between 1050-1300 cm^{-1} and 1500-1750 cm^{-1}, which were enlarged and shown in Figure 4.27 and Figure 4.28 respectively.

It could be seen in Figure 4.27, three main peaks were observed, including 1298 (N-H band), 1232 (Composed of amide III as well as phosphate vibration of nucleic acids[39]), and 1207 cm^{-1} (phosphate I). Four main peaks were 1543 (Amide II), 1585, 1658 (C=O) and 1732 cm^{-1} (C=O stretching). The main features responsible for this separation were mainly variations in amide II. Overall, although the main peaks of the feature importance plots were based on biological information, still many minor peaks were observed which could indicate noise or other interferences.

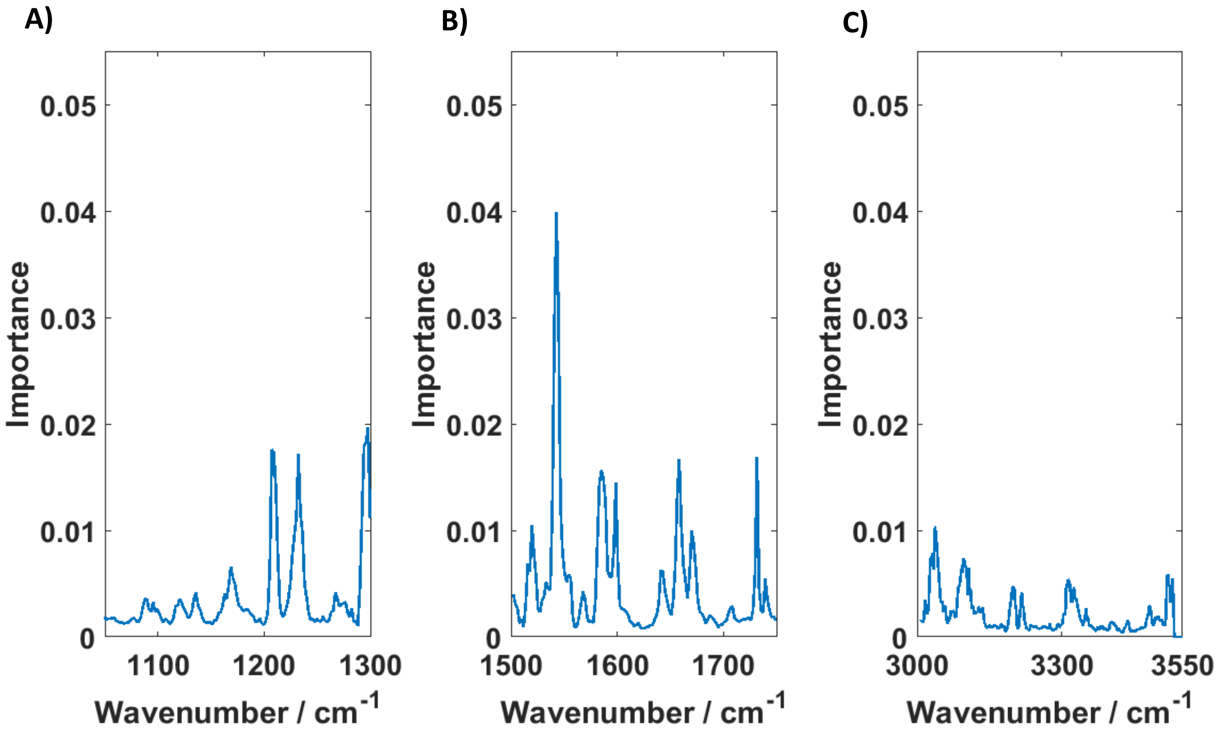


Figure 4.26 Feature importance plot (full-range) of cancer diagnosis model based on stroma using random forest, where A) wavenumber range 1050 to 1300 cm^{-1}, B) wavenumber range 1500 to 1750 cm^{-1}, and C) wavenumber range 3005 to 3550 cm^{-1}

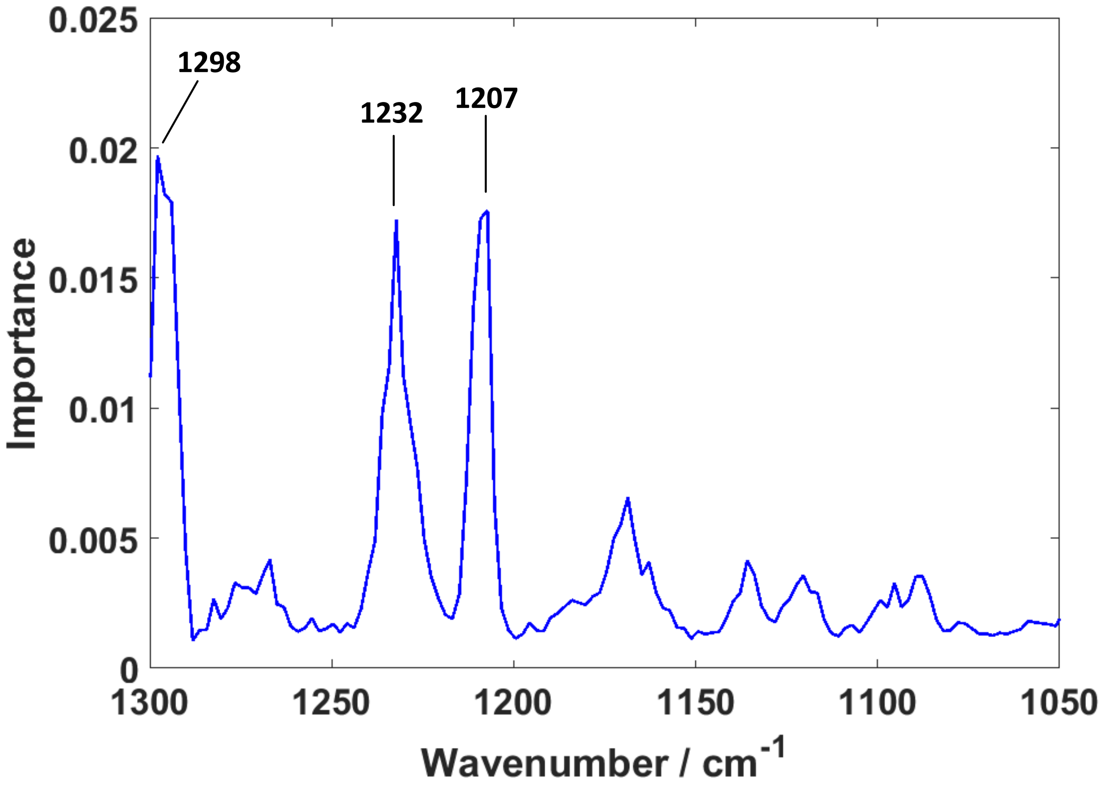


Figure 4.27 Feature importance plot (1050-1300 cm^{-1}) of cancer diagnosis model based on stroma using random forest

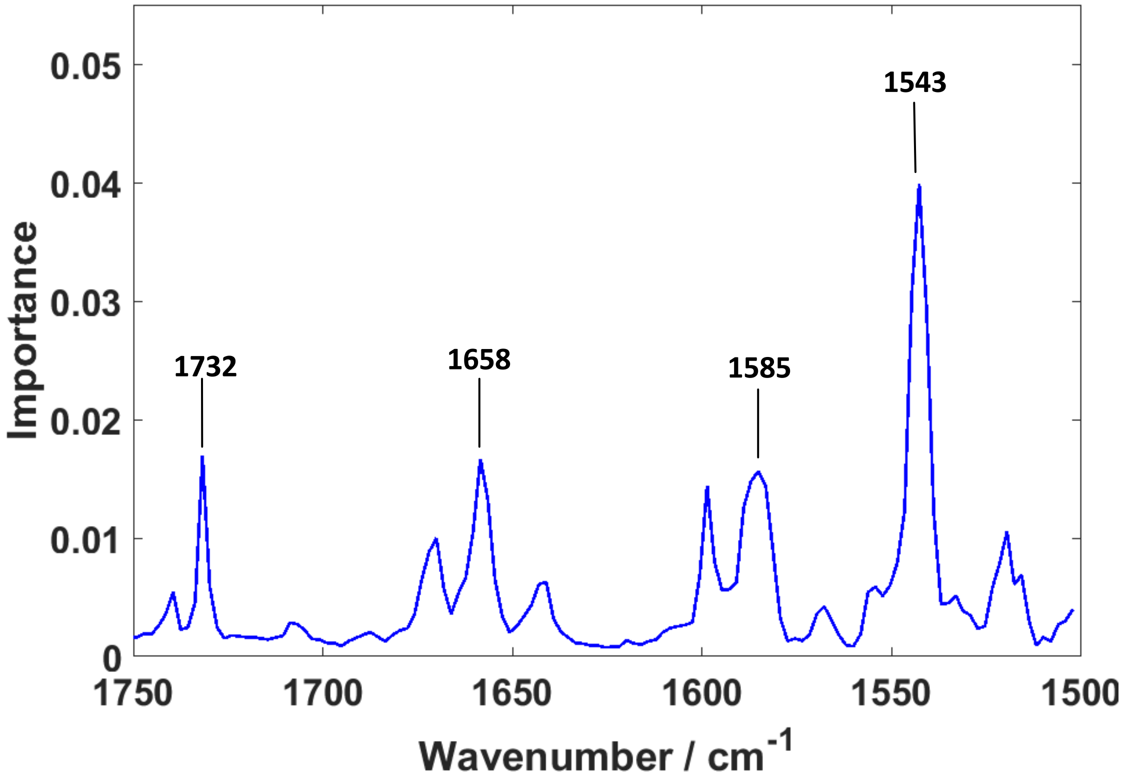


Figure 4.28 Feature importance plot (1500-1750 cm^{-1}) of cancer diagnosis model based on stroma using random forest

4.3.2.2.1.3. Sectional conclusion

Both classification results on the independent test group were not as good as one would wish for but within the acceptance range considering test spectra were from different patients and in large numbers, considering the cancer diagnosis accuracy of pathologist was 75.4%, in the research conducted by Kasraeian S, *et al.* [41] in 2010. The results observed from epithelium were more balanced compared with those from the stroma. However, the stroma model was more sensitive to cancerous spectra compared with epithelium one.

4.3.2.2.2. AdaBoost models

AdaBoost was used to construct preliminary models using annotated epithelium and stroma spectra.

4.3.2.2.2.1. Model constructed using epithelium spectra

In terms of model training, the same spectra used to construct the random forest model were used here as well. 70932 cancerous and 19338 NAT epithelium spectra were selected based on the annotations tested previously. 19338 cancerous spectra were randomly selected using the under-sampling method to match the number of spectra in NAT class. The total number of spectral input in both cancerous and NAT class was 38676. 570 features were used for the model training process.

In order to obtain reasonable classification results, AdaBoost M1 was used as the classification method for this binary problem. 500 trees were selected and trained for 500 iterations with a learning rate of 0.1, which was percentage shrinkage to learn new information. To train an ensemble using shrinkage, the Learning rate needed to be set to a value between 0 and 1, for example, 0.1 was a popular choice [42]. Training an ensemble classifier with shrinkage needed more learning iterations, but often results better accuracies in the independent tests. The tree template of each weak learner for the model based on epithelium spectra was tuned to better classification results as well. 5 Fold Cross-validation was conducted and the average classification accuracy was used on tuning the tree-complexity for each weak classifier. The number of tree split was exponentially increased for ensemble models from the decision stumps to trees contained 59049 splits. The averaged cross-validated misclassification rate was estimated for each ensemble model. The tree template, 81 splits, with the minimal misclassification rate was used for the final model construction.

An AdaBoost model was established based on epithelium spectra from the training group of patients. It was validated with 20% randomly selected spectra from the training dataset. To be more specific, 80% randomly selected spectra were used for model construction while the 20% which were not involved in the training process was used to validate the model. The confusion matrix of the training validation process was shown in the Table 4.10. From the table, it could be seen that the true positive accuracy was 99.9% and the true negative accuracy was 99.7%, which was nearly complete separation between two classes. A general classification accuracy of 99.8% was obtained, which indicated that the model works almost perfectly on the subset of training data.

Table 4.10 Confusion matrix of the training validation for cancer NAT classification using AdaBoost based on epithelium spectra collected in experiment two

Training validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	99.9	0.1
	NAT	0.3	99.7

An independent test was conducted on the model constructed. 20041 test spectra which contained 19291 cancerous spectra and 750 non-cancerous spectra were pre-processed and tested by the model. The confusion matrix of classification results was shown in Table 4.11. From the table, a true cancer diagnosis accuracy of 89.4% was observed and the classification accuracy for the NAT class was 92.8%. The overall classification accuracy was 89.5%, which indicates good separations between cancerous and NAT classes.

Table 4.11 Confusion matrix of the independent test for cancer NAT classification using AdaBoost based on epithelium spectra collected in experiment two

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	89.4	10.6
	NAT	7.2	92.8

The ROC curve of the model was plotted to investigate the performance of the constructed model and was shown in Figure 4.29. The resulting AUC value of 0.97 further shows that a reasonable separation was observed between malignant and non-malignant pixels.

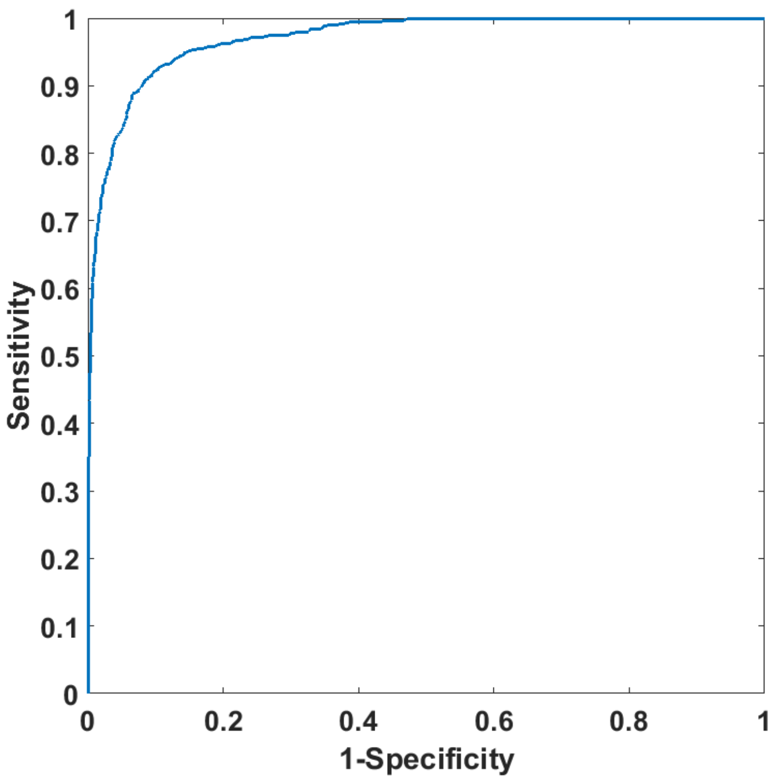


Figure 4.29 ROC curve of the model for cancer NAT classification using AdaBoost based on epithelium spectra collected in experiment two

The importance of each feature was plotted, and was shown in Figure 4.30. It indicates that the classification was mainly caused by the variations of spectra in the interval of 1500-1750 cm^{-1}, which was enlarged and shown in Figure 4.31.

It could be seen in Figure 4.31, main peaks at 1658 (Amide I), 1643 (Amide I), and 1545cm^{-1} (Amide II). The main features responsible for this separation were mainly variations in amide I and II, especially the amide I. Overall, although the main peaks of feature importance plot were based on biological information, many minor peaks were observed which could indicate noise or other interferences. However the nature of AdaBoost was such that these features were weighted lower meaning they contribute less to the classification.

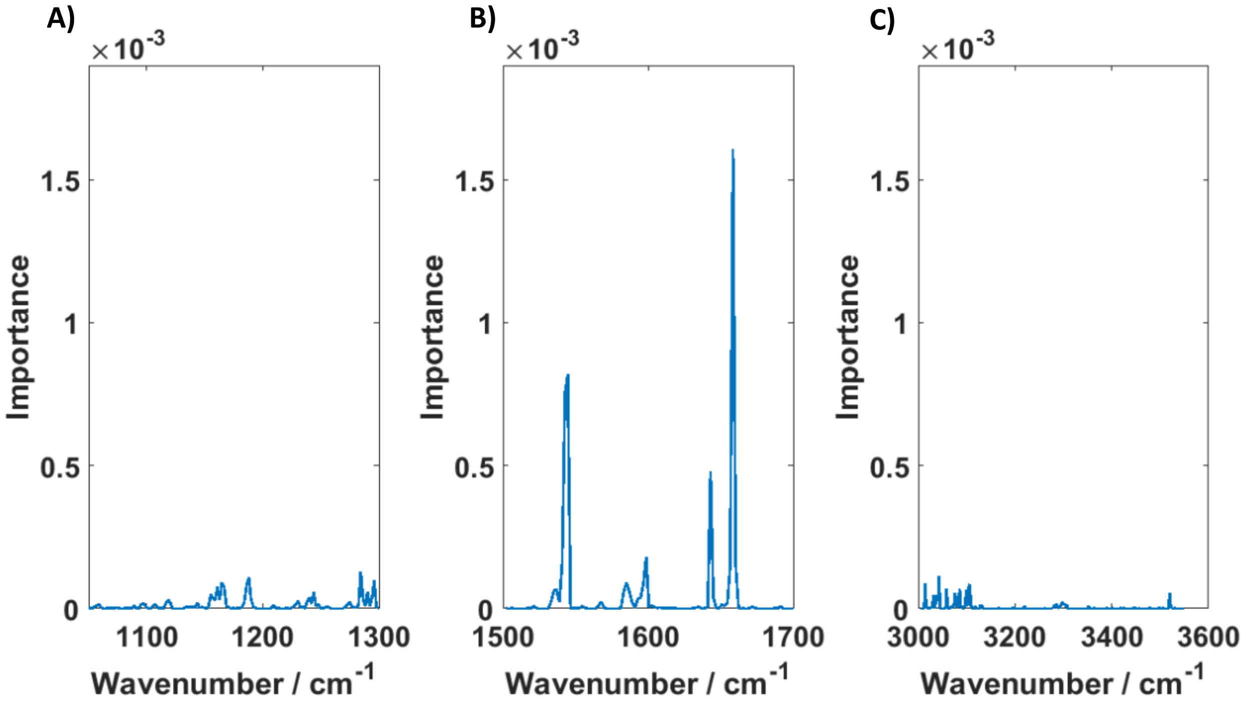


Figure 4.30 Feature importance plot of the model (full range) for cancer NAT classification using AdaBoost based on epithelium spectra collected in experiment two, where A) wavenumber range 1050 to 1300 cm^{-1}, B) wavenumber range 1500 to 1750 cm^{-1}, and C) wavenumber range 3005 to 3550 cm^{-1}

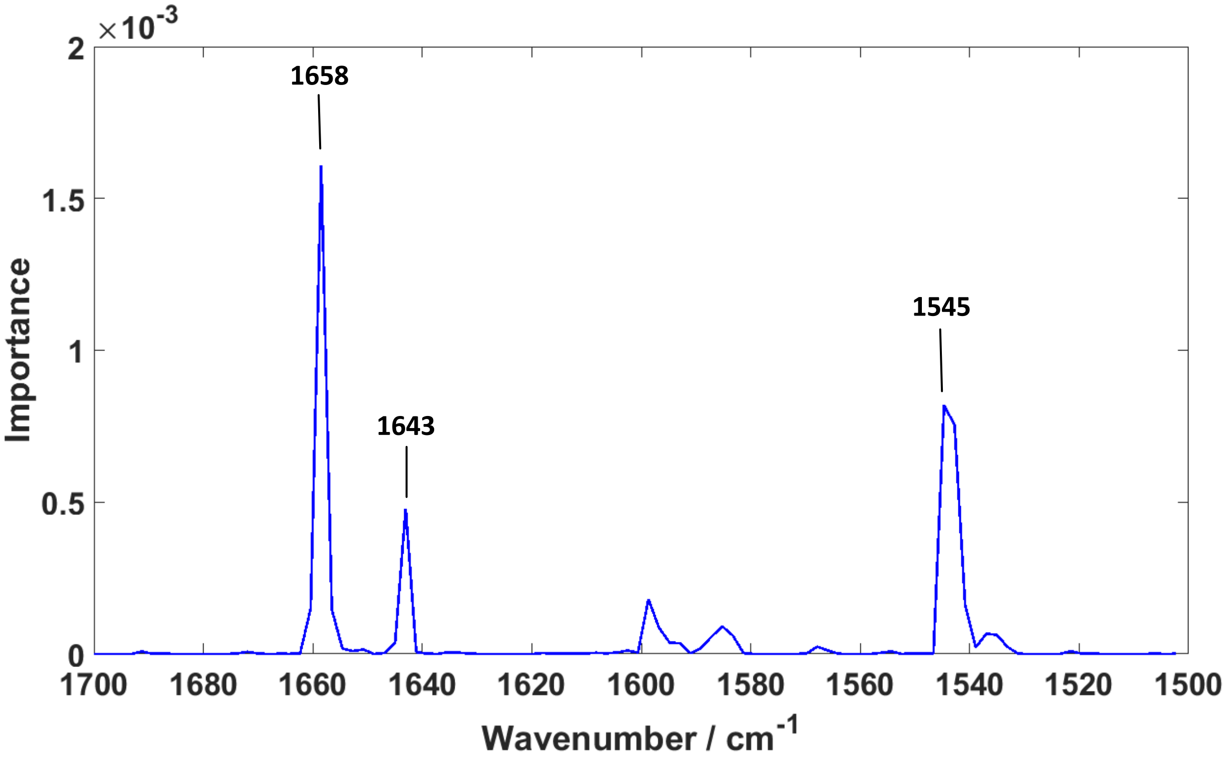


Figure 4.31 Feature importance plot of the model (1500-1700cm^{-1}) for cancer NAT classification using AdaBoost based on the epithelium spectra collected in experiment two

4.3.2.2.2.2. Model constructed using stroma spectra

The model was constructed based on stroma using AdaBoost to investigate whether stroma spectra could produce similar classification results as epithelium spectra in cancer diagnosis tests.

A training database was constructed from the same 55 training cores which consist of 123104 stroma spectra (50918 normal spectra, and 72186 cancerous spectra). Under-sampling was conducted to balance the number of spectra in each class. A Random Forest was then constructed using 50918 spectra per class, including a normal class and a cancerous class. Each spectrum was noise reduced, quality checked, cut to the correct range and first derived prior to being used for training.

The same configuration of the model construction for epithelium spectra using AdaBoost was applied. AdaBoost M1 with 500 trees which were selected and trained for 500 iterations. A learning rate of 0.1 was used to train the ensemble model. 5 Fold Cross-validation was conducted and the average classification accuracy was used on tuning the tree-complexity for each weak classifier. The tree template, 243 splits, with the minimal misclassification rate was used for the final model construction.

20% training validation data was selected randomly from the training dataset and applied on the model to test whether it worked on data with similar structure to the training data. The confusion matrix of the training validation test was shown in Table 4.12, in which classification accuracies of 98.0% for cancerous spectra and 95.8% for NAT spectra were observed. This could be considered as a relatively good training validation test result with over 20 thousands spectra tested in the validation group.

Table 4.12 Confusion matrix of the training validation for cancer NAT classification using AdaBoost based on stroma spectra collected in experiment two

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	98.0	2.0
	NAT	4.2	95.8

An independent test was conducted on the model constructed. 27141 test spectra which contained 13157 cancerous spectra and 13985 non-cancerous spectra were pre-processed and tested by the model. The confusion matrix of the classification results was shown in and shown in Table 4.13. From the table, a classification accuracy of cancer diagnosis, 89.6%, and classification accuracy for the NAT class, 85.7%, were observed. The overall classification accuracy was 87.7%, which indicated good separations between cancerous and NAT classes. When the training iteration of the model was increased to 5000, the overall classification accuracy increased to 89.2% (91.5% for cancerous class and 86.9% for NAT class). A 1.5% increase overall accuracy was obtained. However, the time spent on training was at least double from the previous time (more than two hours).

Table 4.13 Confusion matrix of the independent test for cancer NAT classification using AdaBoost based on stroma spectra collected in experiment two

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	89.6	10.4
	NAT	14.3	85.7

ROC curve of the model was plotted to investigate the performance of model and was shown in Figure 4.32. The resulting AUC value of 0.93 further shows that a good separation was observed between malignant and non-malignant pixels, comparing with the previous results obtained using a combination of random forest and stroma spectra, 0.87.

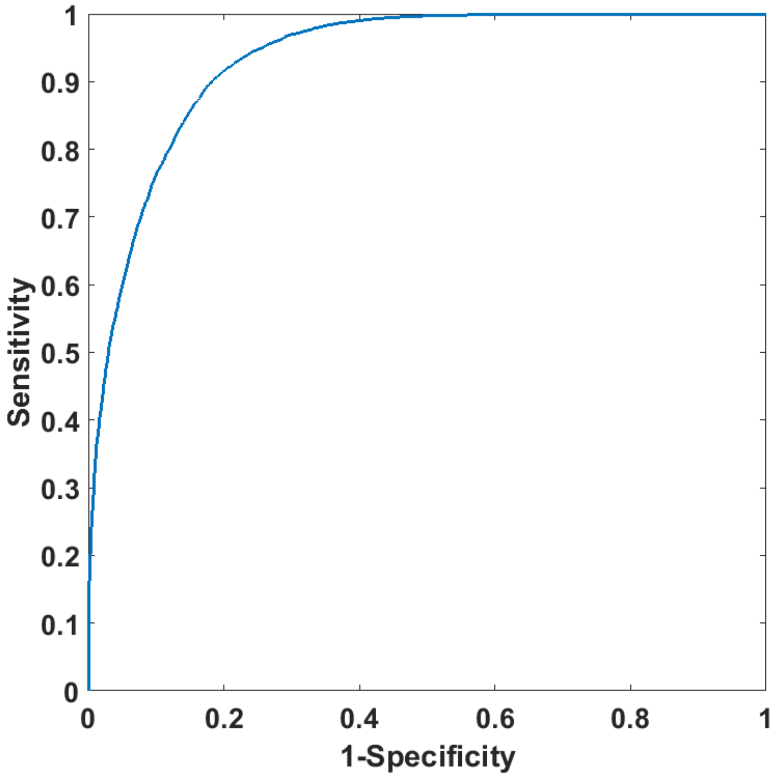


Figure 4.32 ROC curve of the model for cancer NAT classification using AdaBoost based on stroma spectra collected in experiment two

The importance of each feature was plotted and was shown in Figure 4.33. It could be seen that in the plot that the classification was mainly caused by the features from all three spectral intervals. A standard was set to 0.0001, which meant that only intensity of importance of a feature was larger than 0.0001 would be considered as main peaks. Six main peaks were found in all the ranges, which were enlarged and plotted in Figure 4.34 (1050 -1300 cm^{-1}), Figure 4.35 (1500 -1700cm^{-1}) and Figure 4.36 (3000 -3550cm^{-1}).

Most main peaks were found in the Figure 4.35, including 1660 (Amide I band), 1599 (Amide I), 1587 and 1545cm^{-1} (Amide II). Amide I and II variations were observed. One main peak was observed in Figure 4.34, which was 1294 cm^{-1} (N-H cytosine). It could relate to the DNA composition differences between cancer and NAT spectra. Another main peak was observed at 3037 cm^{-1} (N-H stretching) in Figure 4.36. These differences all showed that protein composition between two classes caused mainly the classification. Although many minor peaks were observed which could indicate noise or other interferences, AdaBoost gives them lower weighting and so they would contribute less to the classification.

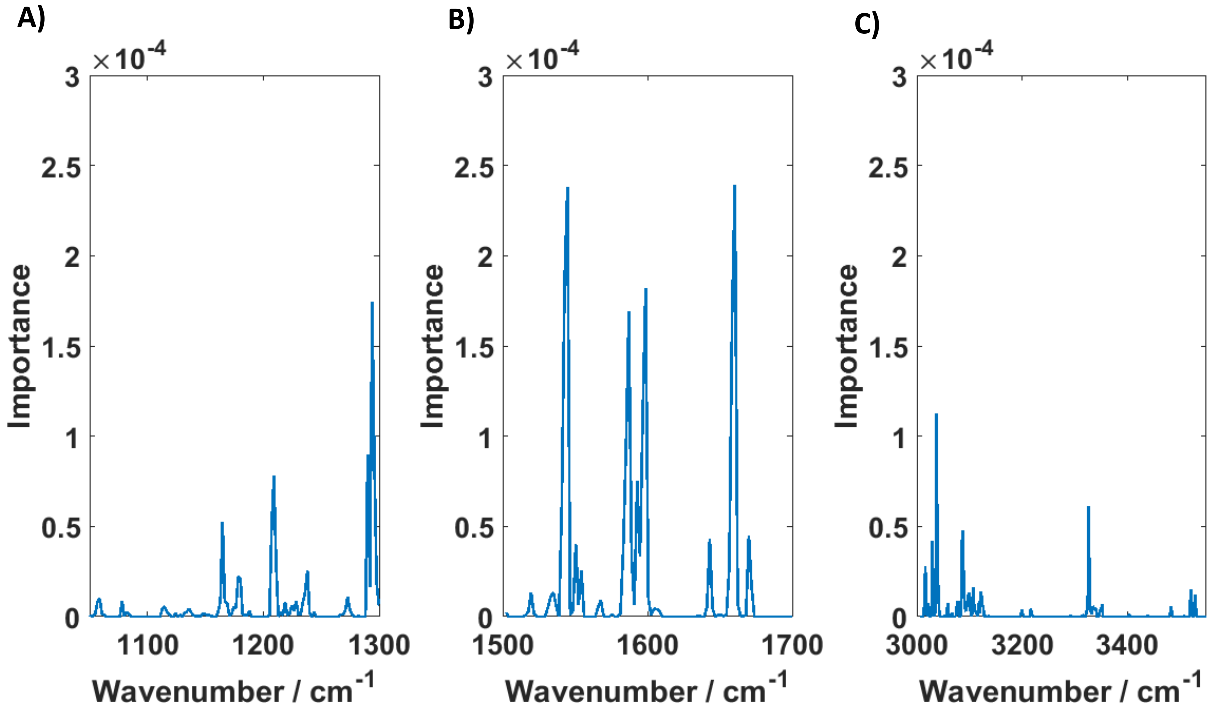


Figure 4.33 Feature importance plot of the model (full-range) for cancer NAT classification using AdaBoost based on stroma spectra collected in experiment two, where A) wavenumber range 1050 to 1300 cm^{-1}, B) wavenumber range 1500 to 1750 cm^{-1}, and C) wavenumber range 3005 to 3550 cm^{-1}

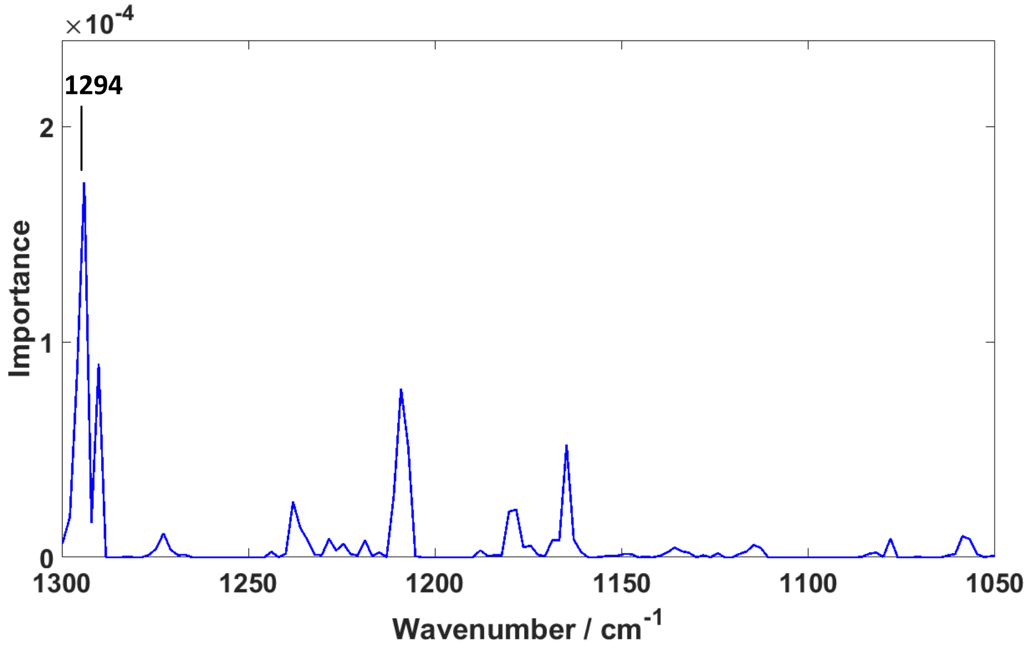


Figure 4.34 Feature importance plot of the model (1050-1300 cm^{-1}) for cancer NAT classification using AdaBoost based on stroma spectra collected in experiment two

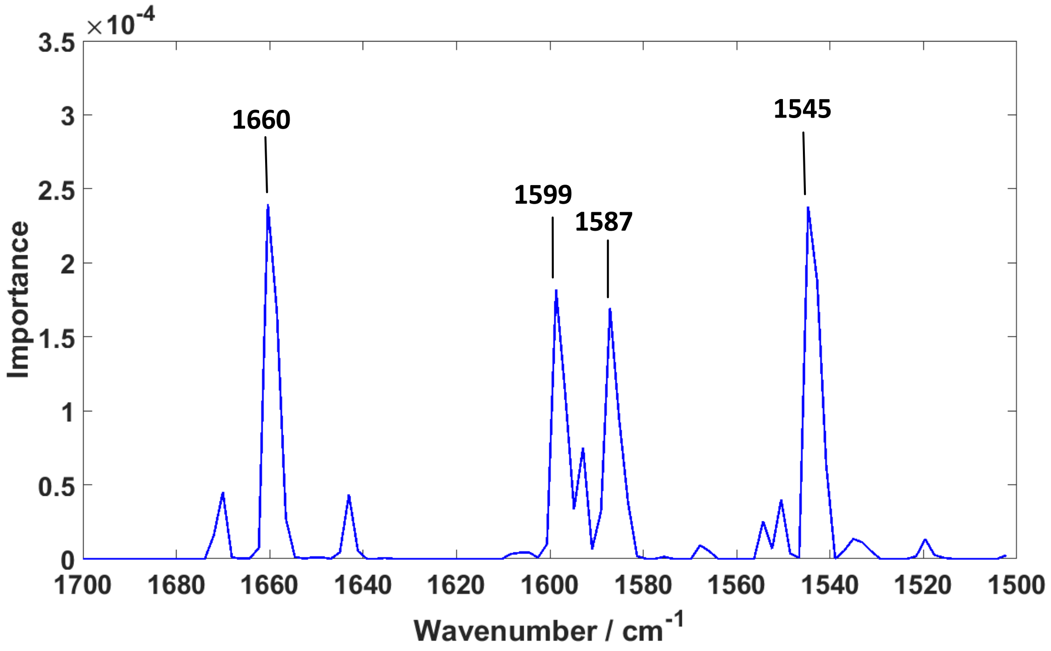


Figure 4.35 Feature importance plot of the model (1500-1700cm^{-1}) for cancer NAT classification using AdaBoost based on stroma spectra collected in experiment two

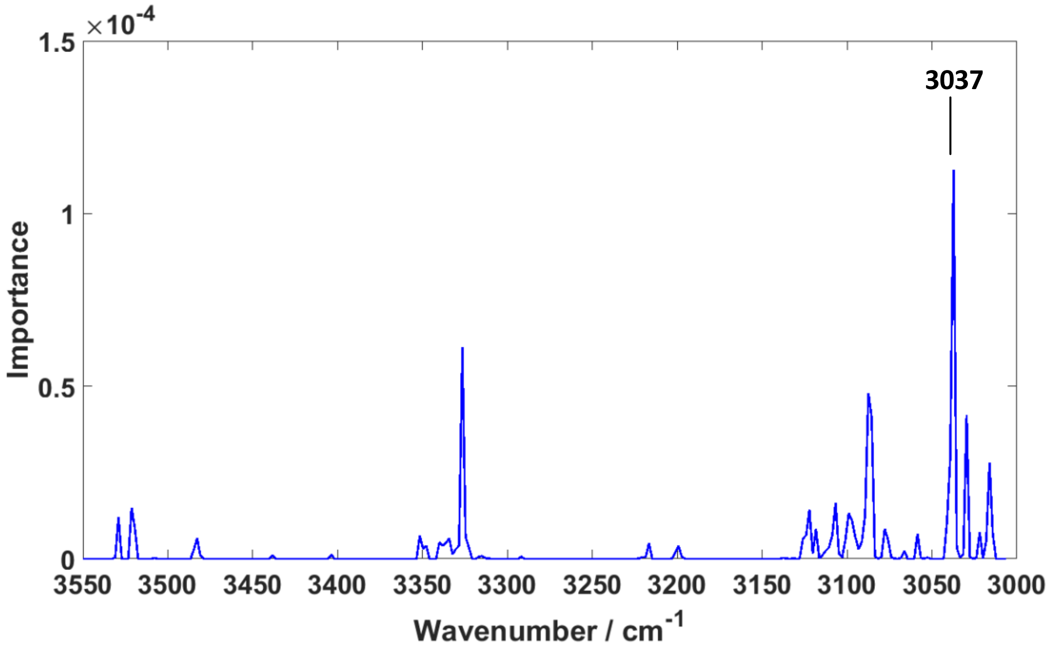


Figure 4.36 Feature importance plot of the model (3000-3550cm^{-1}) for cancer NAT classification using AdaBoost based on stroma spectra collected in experiment two

4.3.2.2.2.3. Sectional conclusion

Both classification results on the independent test group were considering good, comparing with the results obtained using random forest. Large number of spectra was used as independent test dataset, which increased the reliability of the model, as larger independent dataset could potentially contain lager variety of features. Using AdaBoost, similar independent test results were obtained by applying both epithelium and stroma spectra. Still epithelium had a small advantage, as it has higher overall accuracy. These independent test results could be improved by further by increasing the number of iteration of training, but as mentioned previously and in literature there was an additional cost in time[34] so this had not been considered here.

4.3.2.2.3. Comparisons on the results obtained using random forest and AdaBoost models

Comparing both AdaBoost and random forest models based on epithelium, applying the same training and independent test datasets, the overall classification accuracy of AdaBoost on independent test was 89.5% while the separation rate of random forest was 77.8%. An over 10% advantage was observed. Especially on detection of cancerous spectra, AdaBoost had an enhancement of 12.5%. When models constructed by features selected from stroma spectra, an even larger gap, over 20% for the overall classification accuracy, was observed between AdaBoost and random forest in the independent test. For random forest, 54.6% NAT stroma spectra could be successfully classified while for AdaBoost 85.7% NAT stroma spectra could be separated from cancerous spectra. A nearly 30% difference was observed when it came to the true negative classification accuracy of models. However, it was interesting that the random forest model was more sensitive to the cancerous spectra than AdaBoost model, as the true positive classification accuracy for random forest model was 91.2% while for AdaBoost model the accuracy was 89.6%. Overall, the AdaBoost models had a better performance on the independent test when the same test sets were provided.

There could be many reasons for these observations. One possible reason was that AdaBoost weighted features and weak classifiers favoured the classification process. To be more specific, for epithelium based models, four main peaks were observed in the feature importance plot of random forest model. Three main features among them were exactly the same with all three main peaks (1658, 1643, and 1543 cm^{-1}) in the feature importance plot of AdaBoost with different importance intensities. These intensity differences could be reason behind classification performance differences. In addition to that, for both feature importance plots, many minor peaks could be observed which might be noise or other interferences in the data. Because of the nature of AdaBoost, these not very helpful features were weighted much lower comparing with these minor features in random forest model. They would contribute less to the final classification. This explained, to a large extent, why both models worked well on the cancerous class, as the main features used for classification for both models were very similar. One possible reason for AdaBoost model having a much better performance on the stroma class could be that the general similarities among NAT tissues were different to find even without interruptions, with distortions from minor peaks, the difficulty further increased. AdaBoost model weighted down these distortions but the random forest did not. Therefore the AdaBoost model worked better on the stroma class.

Not only for epithelium spectra based models but also for stroma spectra based models, similar observations were observed. Features in amide I and II regions were both observed in the feature importance plot of both random forest and AdaBoost. Again, the weighting process in AdaBoost could be the main reason responsible to the difference in the independent test.

In addition, AdaBoost also dug deeper into the data by focusing more on the most contributed elements during model training process. Not just features and input sample, the weak classifiers which had less contribution to the classification error rate were also weighted to have more control of the final classification results.

In terms of comparing performance of epithelium and stroma based models, using the random forest classification method, the epithelium did have a more balanced and higher accuracy than stroma based model. One possible reason for it could be that the epithelium spectra were spectra of a single type of cell. Although the common similarities among cancerous spectra were difficult to find, they were easier to be found among spectra of the single type of cell compared with multiple type of cells in stroma. Applying the AdaBoost method, the advantage of using epithelium spectra was reduced. The overall classification accuracy for epithelium spectra based model was 89.5% and for stroma spectra based model was 87.7%. Only a less than 2% difference was found. The classification accuracies using both epithelium and stroma spectra for both cancerous and NAT classes were relatively balanced and reasonable.

The better performance models based on epithelium and stroma spectra were selected from models constructed using both classification methods. The selected models were used to predict the status of each independent test core without guidance of annotations. A traffic light colouring system was used to present the final result of classification, where the cancerous cores were coloured in red, the NAT cores were coloured in green and the cores needed further investigation were coloured in yellow. The traffic light colouring system was combined with both epithelium-stroma classifier and cancer-NAT classifier. To be more specific, all the spectra from each core were passed into the epithelium-stroma classifier. Only the 10% top scored spectra which were predicted as epithelium spectra were passed into the cancer-NAT classifier (epithelium spectra based). Again, only the top scored 10% spectra were kept and used for class assignment for each core. Among these 10% top scored spectra, if the number of predicted cancerous spectra was larger than the number of classified NAT spectra, and the number of diagnosed cancerous spectra was larger than half of the total number of kept spectra, the core was picked as a cancerous core. If the number of diagnosed cancerous spectra was less than the number of diagnosed stroma spectra and the number of predicted stroma spectra was larger than half of the total number of kept spectra, the core was named as a NAT core. In any other cases, the core was coloured in yellow which meant not determined and further classification required. The same rules and conditions were applied to the better performed model constructed based on stroma spectra.

Both models trained on AdaBoost classification methods were applied. The traffic light colouring plot of both model based on epithelium and stroma spectra were shown in Figure 4.37 and Figure 4.39 respectively. For better comparison, the annotation of each independent test core was shown in the Figure 4.38 and Figure 4.40. Yellow boxes on plots indicated that no spectra were annotated for the corresponding models, i.e. for epithelium based model, no epithelium cells were annotated in the core, but after epithelium-stroma classification corresponding spectra were still obtained. To ensure the annotation accuracy, areas which were difficult to annotate or having higher uncertainties were ignored. With the epithelium-stroma classifier, these hard-to-annotate spectra could still be selected and applied for further cancer diagnosis. Both models provided 100% classification accuracy in the unit of individual core, which further validated the success in both model training process.

Overall, epithelium spectra do have advantages in cancer diagnosis. However, stroma also had similar function. When appropriate methods and hyperparameters were applied, stroma could perform as well as epithelium spectra. It could be considered as a replacement of epithelium spectra when resource of epithelium spectra was lacking.

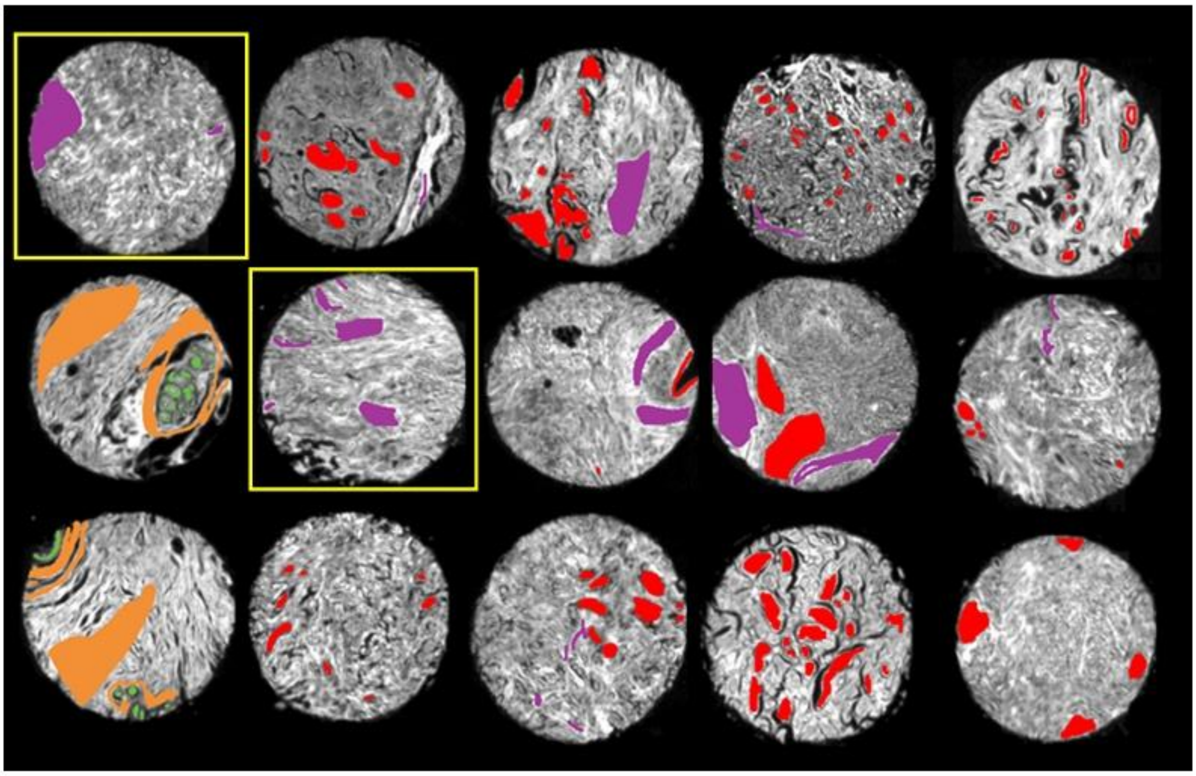


Figure 4.38 Annotation of each core in the independent test, where yellow boxes indicated that no epithelium cells were annotated in that core. Red and green showed cancerous and normal epithelium while purple and yellow indicated cancerous and normal stromal regions in each core.

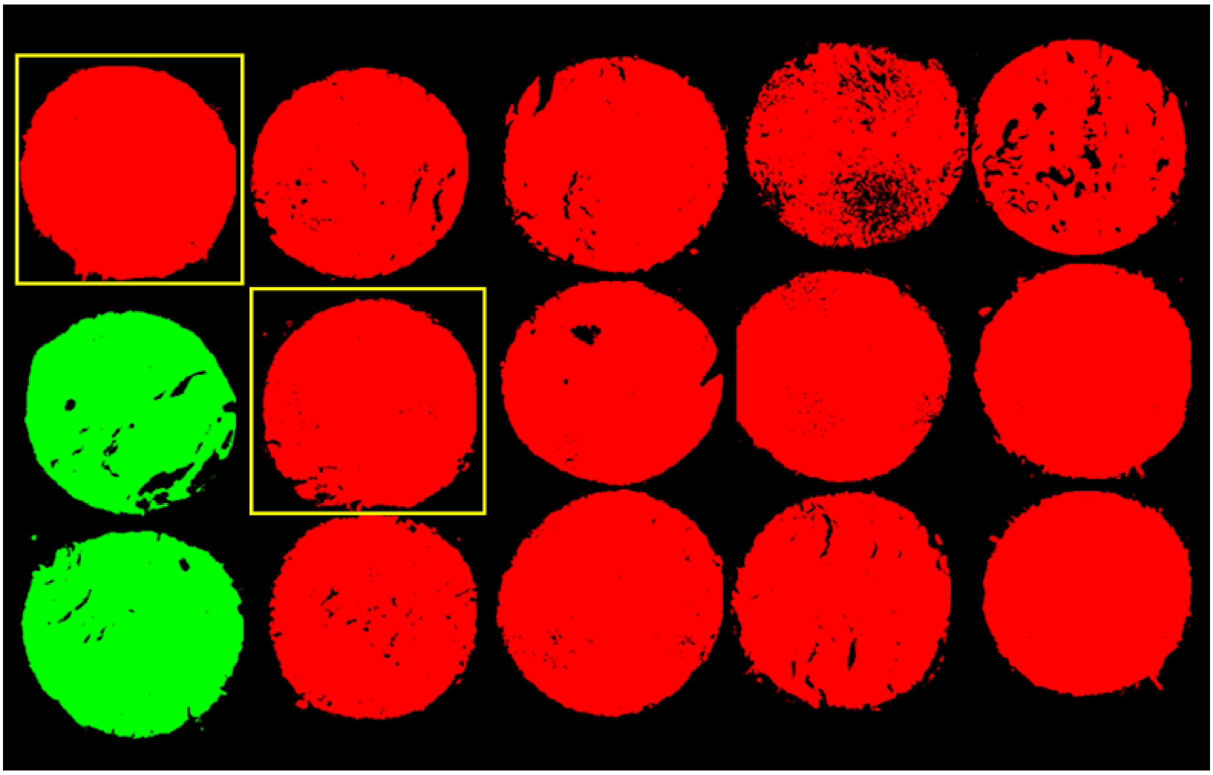


Figure 4.37 Traffic light colouring plot of the better performed model based on epithelium spectra, where yellow boxes indicated that no epithelium cells were annotated and no epithelium spectra were selected to train the classification model in that core. Cores in red were predicted as cancerous cores while cores in green were predicted as normal cores.

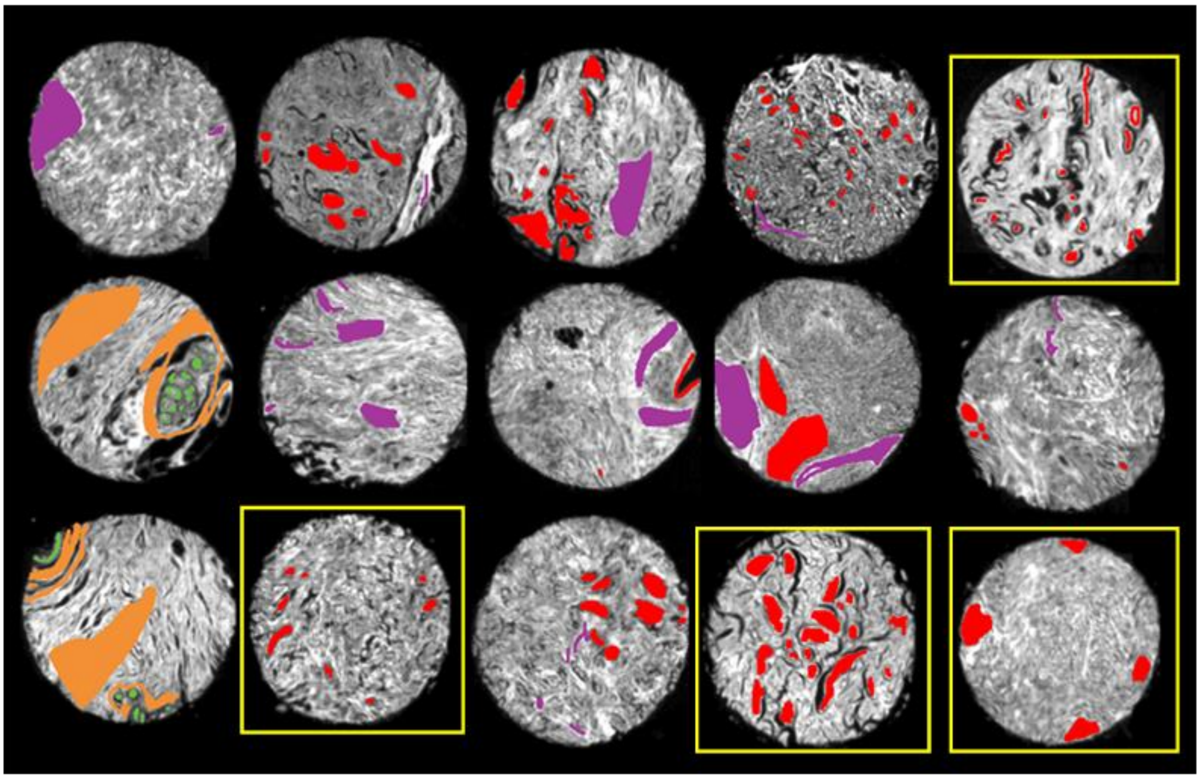


Figure 4.40 Traffic light colouring plot of the better performed model based on stroma spectra, where yellow boxes indicated that no stroma cells were annotated in that core. Red and green showed cancerous and normal epithelium while purple and yellow indicated cancerous and normal stromal regions in each core.

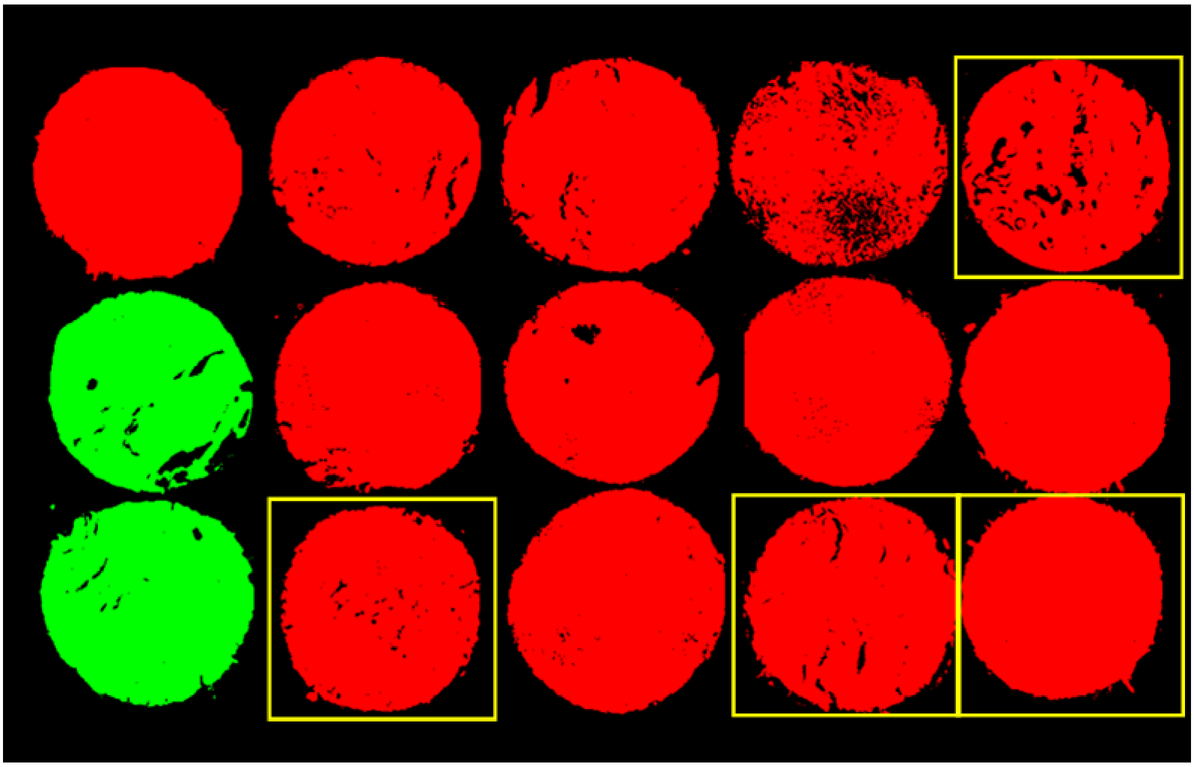


Figure 4.39 Annotation of each core in the independent test, where yellow boxes indicated that no stroma region was annotated and no stroma spectra were selected to train the classification model in that core. Cores in red were predicted as cancerous cores while cores in green were predicted as normal cores.

4.4. Conclusion and future work

To sum up, in experiment one, data was corrupted by water vapour. Although great care was taken with the purge and there was a bespoke purge box surrounding the microscope stage, the time delay between first and last measurements when scanning large mosaics made water vapour interference a potential problem. Note that this is a common interference factor for all researchers during FTIR data collection and has led to a number of erroneous conclusions be drawn in the literature [43][44]. For any other experiments in the future, better actions in water vapour prevention should be considered.

In terms of comparing random and AdaBoost, surprisingly AdaBoost worked better than random forest in term of the independent test in this section. AdaBoost was good at selecting features that favoured classification. When a large number of iteration was applied, good classification accuracies could be obtained. The results obtained in this section could be further boosted by optimising models. However, this would be a trade-off between time and increase of classification accuracy.

Stroma spectra can be used for cancer diagnosis when appropriate classification methods and parameters are applied. As in breast tissues stroma often much more than epithelium cells, developing classifiers on stroma can reduce the limitation of the shortage of training samples, which can further increase the reliability and accuracy of the model. Therefore comparing with epithelium cells, stroma has its own advantages in term of clinical translation. Further researches with lager stroma sample size and other more advanced data mining methods should be conducted.

4.5. Reference

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA. Cancer J. Clin. (2018). doi:10.3322/caac.21492.

[2] A. King, J. Broggio, Office for National Statistics: Cancer registration statistics, England 2016: release date 2018. Available at: https://www.ons.gov.uk/ (last accessed 03 August 2018), (2018) 1–22. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/final2016.

[3] S. John, J. Broggio, Cancer survival in England : national estimates for patients followed up to 2017, (2019) 1–27.

[4] J.B. Lattouf, F. Saad, Gleason score on biopsy: Is it reliable for predicting the final grade on pathology?, BJU Int. (2002). doi:10.1046/j.1464-410X.2002.02990.x.

[5] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, P.R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P.L. Martin-hirsch, B. Obinaju, G.D. Sockalingum, J. Sulé-suso, R.J. Strong, M.J. Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, 9 (2014) 1771–1791. doi:10.1038/nprot.2014.110.

[6] R. Eckel, H. Huo, H.W. Guan, X. Hu, X. Che, W.D. Huang, Characteristic infrared spectroscopic patterns in the protein bands of human breast cancer tissue, Vib. Spectrosc. (2001). doi:10.1016/S0924-2031(01)00134-5.

[7] P. Bassan, M.J. Weida, J. Rowlette, P. Gardner, Large scale infrared imaging of tissue micro arrays (TMAs) using a tunable Quantum Cascade Laser (QCL) based microscope, Analyst. (2014). doi:10.1039/c4an00638k.

[8] H. Fabian, N.A.N. Thi, M. Eiden, P. Lasch, J. Schmitt, D. Naumann, Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy, Biochim. Biophys. Acta - Biomembr. (2006). doi:10.1016/j.bbamem.2006.05.015.

[9] D.M. Mayerich, M. Walsh, A. Kadjacsy-Balla, S. Mittal, R. Bhargava, Breast histopathology using random decision forests-based classification of infrared spectroscopic imaging data, Proc. SPIE. 9041 (2014) 904107. doi:10.1117/12.2043783.

[10] H. Fabian, P. Lasch, M. Boese, W. Haensch, Infrared microspectroscopic imaging of benign breast tumor tissue sections, J. Mol. Struct. (2003). doi:10.1016/j.molstruc.2003.07.002.

[11] M. Sattlecker, R. Baker, N. Stone, C. Bessant, Support vector machine

ensembles for breast cancer type prediction from mid-FTIR micro-calcification spectra, Chemom. Intell. Lab. Syst. (2011). doi:10.1016/j.chemolab.2011.05.007.

[12] M. Verdonck, A. Denayer, B. Delvaux, S. Garaud, R. De Wind, C. Desmedt, C. Sotiriou, K. Willard-Gallo, E. Goormaghtigh, Characterization of human breast cancer tissues by infrared imaging, Analyst. (2016). doi:10.1039/c5an01512j.

[13] C. Hughes, L. Gaunt, M. Brown, N.W. Clarke, P. Gardner, Assessment of paraffin removal from prostate FFPE sections using transmission mode FTIR-FPA imaging, Anal. Methods. (2014). doi:10.1039/c3ay41308j.

[14] V. Untereiner, O. Piot, M.D. Diebold, O. Bouché, E. Scaglia, M. Manfait, Optical diagnosis of peritoneal metastases by infrared microscopic imaging, Anal. Bioanal. Chem. (2009). doi:10.1007/s00216-009-2630-2.

[15] F. Lyng, E. Gazi, P. Gardner, Chapter 5. Preparation of Tissues and Cells for Infrared and Raman Spectroscopy and Imaging, in: Biomed. Appl. Synchrotron Infrared Microspectrosc., 2010. doi:10.1039/9781849731997-00145.

[16] T.P. Wrobel, B. Wajnchold, H.J. Byrne, M. Baranska, Electric field standing wave effects in FT-IR transflection spectra of biological tissue sections: Simulated models of experimental variability, Vib. Spectrosc. (2013). doi:10.1016/j.vibspec.2013.09.008.

[17] Sunil R. Lakhani; Ian O. Ellis;, S.J.P.H.T. Schnitt, V.M.J. van De, WHO Classification of Tumours of the breast, 2012. doi:10.1017/CBO9781107415324.004.

[18] D.L. Massart, B.G.M. Vandeginste, J.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verberke, L.M.C. Buydens, S. De Jong, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, 1997. doi:10.1016/S0922-3487(97)80056-1.

[19] R. Salzer, H.W. Siesler, Infrared and Raman Spectroscopic Imaging, 2009. doi:10.1002/9783527628230.

[20] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/C3AY41907J.

[21] I.T. Jolliffe, Principal Component Analysis, Second Edition, Encycl. Stat. Behav. Sci. (2002). doi:10.2307/1270093.

[22] I.I. Patel, J. Trevisan, P.B. Singh, C.M. Nicholson, R.K.G. Krishnan, S.S. Matanhelia, F.L. Martin, Segregation of human prostate tissues classified high-risk (UK) versus low-risk (India) for adenocarcinoma using Fourier-transform infrared or Raman microspectroscopy coupled with discriminant analysis, Anal. Bioanal. Chem. (2011). doi:10.1007/s00216-011-5123-z.

[23] M.J. Baker, E. Gazi, M.D. Brown, J.H. Shanks, P. Gardner, N.W. Clarke, FTIR-

based spectroscopic analysis in the identification of clinically aggressive prostate cancer, Br. J. Cancer. (2008). doi:10.1038/sj.bjc.6604753.

[24] M.J. Pilling, A. Henderson, P. Gardner, Quantum Cascade Laser Spectral Histopathology: Breast Cancer Diagnostics Using High Throughput Chemical Imaging, Anal. Chem. 89 (2017) 7348–7355. doi:10.1021/acs.analchem.7b00426.

[25] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, Analyst. 142 (2017) 1258–1268. doi:10.1039/c6an02224c.

[26] P.R. Griffiths, J.A. De Haseth, Fourier Transform Infrared Spectrometry: Second Edition, 2006. doi:10.1002/047010631X.

[27] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. (1964). doi:10.1021/ac60214a047.

[28] R.W. Schafer, What is a savitzky-golay filter?, IEEE Signal Process. Mag. (2011). doi:10.1109/MSP.2011.941097.

[29] L.E.O. Breiman, Random Forests, (2001) 5–32.

[30] L. Breitman, A. Cutler, A. Liaw, M. Wiener, randomForest: Breiman and Cutler's Random Forests for Classification and Regression, Https://Www.Stat.Berkeley.Edu/~breiman/RandomForests/. (2018). doi:10.1023/A.

[31] L. Breiman, Bagging predictors - Springer, Mach. Learn. (1996). doi:10.1007/BF00058655.

[32] T. Hastie, R. Tibshirani, J. Friedman, Elements of Statistical Learning 2nd ed., Elements. (2009). doi:10.1007/978-0-387-84858-7.

[33] R.E. Schapire, Explaining adaboost, in: Empir. Inference Festschrift Honor Vladimir N. Vapnik, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 37–52. doi:10.1007/978-3-642-41136-6_5.

[34] Y. Freund, R.E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, J. Comput. Syst. Sci. (1997). doi:10.1006/jcss.1997.1504.

[35] B. Schölkopf, Z. Luo, V. Vovk, Empirical inference: Festschrift in honor of Vladimir N. Vapnik, Empir. Inference Festschrift Honor Vladimir N. Vapnik. (2013) 1–287. doi:10.1007/978-3-642-41136-6.

[36] J.A. Swets, Measuring the accuracy of diagnostic systems, Sci. Sci. (1988). doi:10.1126/science.3287615.

[37] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, 2013. doi:10.1007/978-1-4614-7138-7.

[38] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer New York, New York, NY, 2009. doi:10.1007/978-0-387-84858-7.

[39] A.C.S. Talari, M.A.G. Martinez, Z. Movasaghi, S. Rehman, I.U. Rehman, Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 52 (2017) 456–506. doi:10.1080/05704928.2016.1230863.

[40] O.B. Rodimova, Carbon Dioxide and Water Vapor Continuum Absorption in the Infrared Spectral Region, Orig. Russ. Text © O.B. Rodimova. 31 (2018) 564–569. doi:10.1134/S1024856018060143.

[41] S. Kasraeian, D.C. Allison, E.R. Ahlmann, A.N. Fedenko, L.R. Menendez, A comparison of fine-needle aspiration, core biopsy, and surgical biopsy in the diagnosis of extremity soft tissue masses, in: Clin. Orthop. Relat. Res., 2010. doi:10.1007/s11999-010-1401-x.

[42] MathWorks, Fitcensemble, MathWorks Doc. (2019). https://uk.mathworks.com/help/stats/fitcensemble.html (accessed July 27, 2019).

[43] M.C.D. Santos, Y.M. Nascimento, J.D. Monteiro, B.E.B. Alves, M.F. Melo, A.A.P. Paiva, H.W.B. Pereira, L.G. Medeiros, I.C. Morais, J.C. Fagundes Neto, J. V. Fernandes, J.M.G. Araújo, K.M.G. Lima, ATR-FTIR spectroscopy with chemometric algorithms of multivariate classification in the discrimination between healthy: Vs. dengue vs. chikungunya vs. zika clinical samples, Anal. Methods. (2018). doi:10.1039/c7ay02784b.

[44] M.C.D. Santos, Y.M. Nascimento, J.M.G. Araújo, K.M.G. Lima, ATR-FTIR spectroscopy coupled with multivariate analysis techniques for the identification of DENV-3 in different concentrations in blood and serum: A new approach, RSC Adv. (2017). doi:10.1039/c7ra03361c.

Chapter 5

Correlation of FTIR analysis and caveolin-1 staining in human breast tissue for cancer diagnosis

5.1. Chapter Overview

Recently, it has been proposed by Lisanti *et al.* that caveolin-1 can be used as a biomarker of breast cancer. It is believed that loss of Caveolin-1 (Cav-1) expression may be a key factor in driving the development of cancer-associated fibroblasts, which promote tumour growth. The idea of testifying whether caveolin-1 can be used as a diagnostic predictive biomarker for breast cancer progression from pre-malignancy has been conducted by many research groups with conflict results [1–4].

Caveolin-1 is the structural coat protein of caveolae, which involved in vesicular trafficking and signal transduction [5,6]. Caveolin-1 consists of a central 33 amino acid, the amino and carboxyl termini [6]. A plot of the caveolin-1 structure including domains is shown in Figure 5.1. The central 33 amino acids, the hydrophobic domain, forms a hairpin structure and spans the membrane (shown in Figure 5.1 B), while the amino and carboxyl termini remain hydrophilic and cytoplasmic [5]. It behaves as a scaffold protein, which controls the multi-molecular signalling process [6]. The Caveolin-1 has a structure that consists of four types of domain (shown in Figure 5.1 A) [1]. Firstly, N- and C-terminal cytosolic domains are present. Between residuals 61 and 102 of caveolin-1, an oligomerization domain occurs [2–4]. The caveolin scaffolding domain (CSD), which is known to play a role in dimerization and interaction with signalling partners, belongs to the oligomerization domain [1,3]. In addition, a central transmembrane domain and a C-terminal membrane attachment domain are also present in caveolin-1 [1–4].

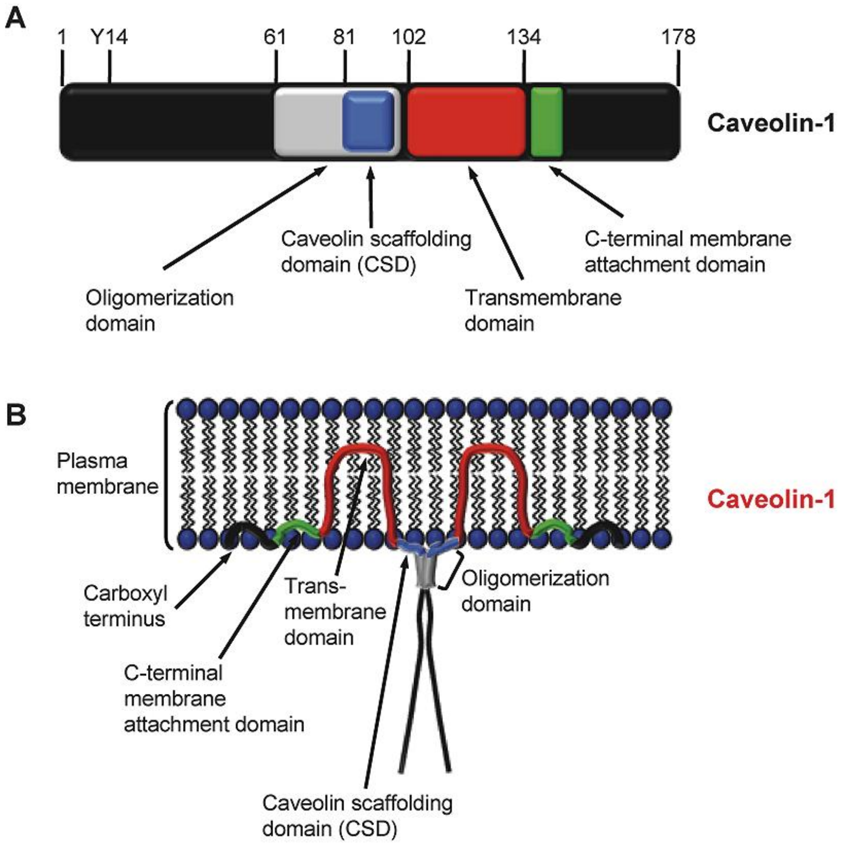


Figure 5.1 Structure of Caveolin-1[1] A). Functional domains and sites of caveolin-1. B). The structure of caveolin-1 in the plasma membrane.

In the last decade, it has been suggested that caveolin-1 plays a vital role in cancer progression and metastasis [1,6]. It is a candidate for a cancer suppressor protein. Also, it acts as a negative regulator of the Ras-p42/44 MAP kinase cascade [7]. Research regarding the function of caveolin-1 as a cancer suppressor has been carried out in different human body tissues, including breast cancer [8] colon cancer [9], ovarian carcinoma [10] and soft-tissue sarcomas [11]. The tumour suppressor ability of caveolin-1 was strongly supported by a down-regulation of expression of it [6].
Caveolin-1 is also considered as having a connection with the DNA repair processes, based on experiments performed by Cordes *et al*.[12]. Their results showed that silencing of caveolin-1 led to an increase of double-strand breaks of residual DNA in irradiated 3D cell cultures [12].

Caveolin-1 is also believed to increase the drug resistance of cell lines [9,13]. Many experiments were conducted on both colon and breast cancer cell lines [9,13]. The results showed that the drug resistance was largely increased with the up-regulation of caveolin-1 expression.

However, many still believe that caveolin-1 is frequently overexpressed in a large range of tumour entities. The data used to support functions of caveolin-1 is not enough to confirm that it plays a critical role in carcinogenesis, tumour progression, metastatic spread and therapy resistance [5]. In the aspect of breast cancer, based on the work of Wiechen et. al., it is clear that loss of caveolin-1 present in the stroma of breast tissue leads to a reduction of patient survival rate in HER2-negative breast cancers [11].

It seems like a new approach to study the properties of caveolin-1 needs to be researched, as the traditional methods have produced conflicting results. It will be interesting and potentially valuable to see if the possibility of using caveolin-1 as a clinical biomarker for breast cancer diagnosis can be testified using FTIR spectroscopy. This is one of the main objectives of the project, which includes two aspects. Firstly, it is crucial to show that caveolin-1 can be correlated with stroma in breast tissue using FTIR. Secondly is important to establish if caveolin-1 can act as a cancer suppressor and can be considered as a biomarker regarding breast cancer diagnosis also using FTIR.

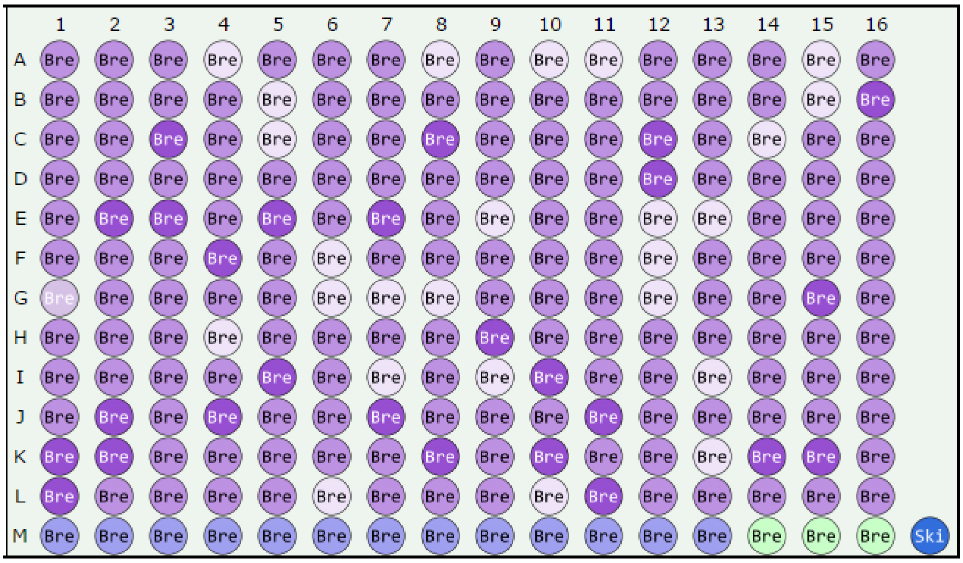
Early detection of cancer has vital significance since cancer treatment is often simpler and more effective when diagnosed at an early stage. There is a burning need to develop early detection techniques requiring non-invasive or minimally invasive procedures. Regarding breast cancer, mammography is the first step performed in the process of diagnosing cancer but at times mammograms are not able to provide clear and true results. A biopsy is suggested to confirm the presence of a tumour, which is also associated with false-positive results [14] due to human errors as diagnosis judgements are made by pathologists. Thus a desire to develop a more effective and safe technique has inspired researchers to develop a range of optical imaging and spectroscopy diagnosis techniques to improve breast cancer diagnosis results and efficiency. The technical improvement will reduce the time and economic costs of patients. One such promising technique, Infrared chemical imaging employs infrared light and has the potential to serve as an efficient diagnostic procedure.

The general objective of this project is to distinguish benign lesions from malignant ones using FTIR and also correlate caveolin-1 in stroma with breast cancer based on the classification produced. This objective includes two aspects. Firstly, it is crucial to show that caveolin-1 can be correlated with stroma in breast tissue using FTIR. The other objective is to testify whether caveolin-1 can be used as a clinical biomarker for breast cancer using FTIR.

5. 2. Methodology

5. 2.1 Caveolin-1 stained sample preparation

The sample used for this project is a breast TMA slide, BR20832. The tissue microarray is one of the most recent innovations regarding pathology. The microarray can be used to analyse multiple tissue samples at the same time since it consists of small representative tissue samples from different cases or patients assembled on a single histologic slide [15]. The coordinates of a tissue microarray are single recipient blocks extracted and re-embedded from donor paraffin blocks [15]. It is a practical and effective tool for tissue studies using FTIR and other high throughput analysis, helping to identify new diagnostic and prognostic markers and targets in human cancers [15,16].



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
A	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
B	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
C	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
D	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
E	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
F	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
G	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
H	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
I	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
J	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
K	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
L	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	
M	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Bre	Ski

Figure 5.2 Location of stages of cancer and non-cancer cores in BR20832

● - Malignant tumor, ● - Malignant tumor (stage 0), ○ - Malignant tumor (stage I), ● - Malignant tumor (stage II), ● - Malignant tumor (stage III), ● - NAT, ● - Normal tissue

In this case, the breast tumour tissue microarray, BR20832, contains 190 cases of ductal carcinoma, 1 mixed lobular and duct carcinoma, 1 mucinous carcinoma, 13 NAT and 3 normal breast tissue. The location of different stages of cancerous cores is shown in Figure 5.2. Each core stands for one patient. In total, the BR20832 TMA contains information from 208 patients both having cancer and healthy.

Fresh biopsies were immersed in an aqueous formalin solution, which fixed the tissue. The tissue was dehydrated in washes of xylene and ethanol. After that, paraffin was used to embed the tissue to provide a protective barrier [17]. The core was cut into 5 µm thickness and 1 mm diameter and fixed on the CaF_2 slide, which is for FTIR measurement use. There are also two adjacent TMAs on the glass which was H&E stained, for stroma annotation, and caveolin-1 stained, for correlation of caveolin-1 protein in stroma with breast cancer.

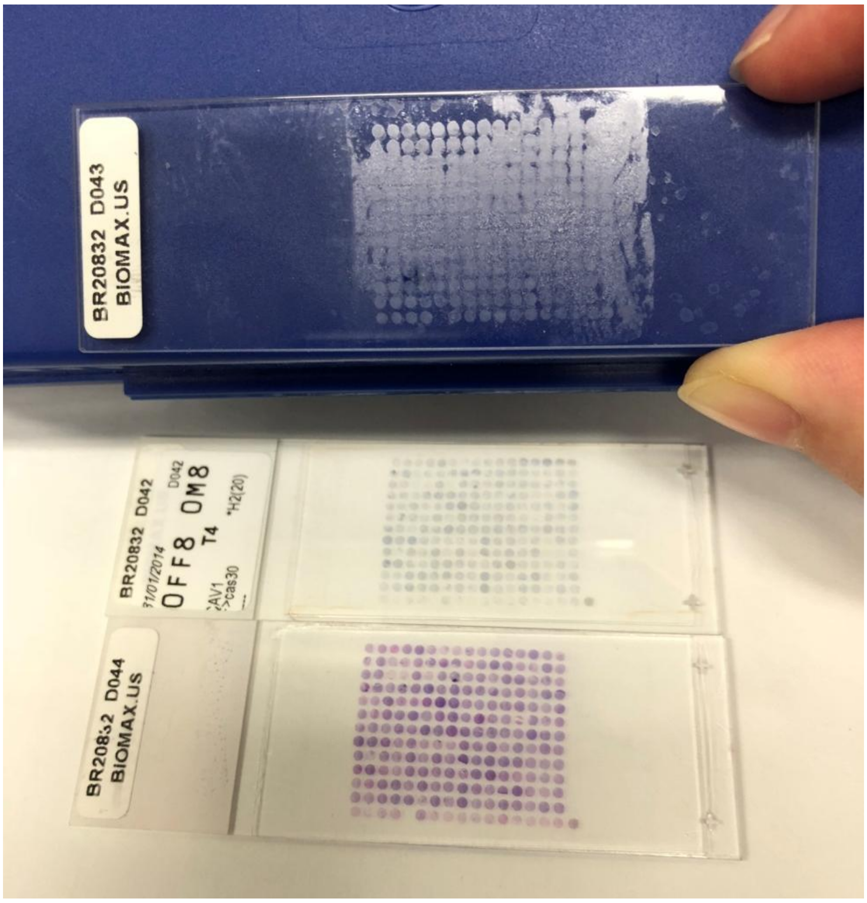


Figure 5.3 Photo of sample slides used. Slide BR20832 D042 (centre) is the caveolin-1 stained glass slide, D043 (top) is the paraffin-embedded CaF_2 slide, and D044 (bottom) is the H&E stained glass slide

The slide used for FTIR measurement was covered in wax and no dewaxing process was undertaken before actual spectra collection, as a number of studies have shown that non-dewaxed FFPE section can be used for skin and colon tissue scans successfully [18–23]. The influenced lipid bands [24] were removed in data processing steps.

5.2.2 Instrumentation and experiment Procedure

The same instrument was used with Chapter 4.Before imaging, background scans were taken as a single tile with 128 co-added scans at a spectral resolution 5 cm^{-1}. The area taken was selected to be clean and paraffin-free. 96 co-added sample scans were measured on each core from TMA. Interferograms were processed into absorption spectra using Happ-Genzel apodisation with a region between 900 and 3800 cm^{-1}.

A purge time of 1 hour was adopted to reduce the influence of water vapour in the spectrum before loading TMA and after changing samples. The scan range was set from 900 to 3800 cm^{-1} covering the important range of biological information of breast tissue.

5.2.3 Data Analysis Procedure

5.2.3.1 Annotation

Annotation was conducted on the chemical image generated using MatLab based on caveolin-1 stained images. Both epithelium cells and stroma cells were identified in this case. Stroma is mainly consisted of collagen and other connective tissues [25,26], which are mainly stained pink around epithelium cells. An example is given in Figure 5.4. In this case, yellow indicates stroma while orange marks locations of caveolin-1. All cores, selected to construct the model and conduct independent tests, were annotated in a similar fashion to that of Figure 5.4. Only cells that were ensured to be the stroma were annotated to control the accuracy of the classification model.

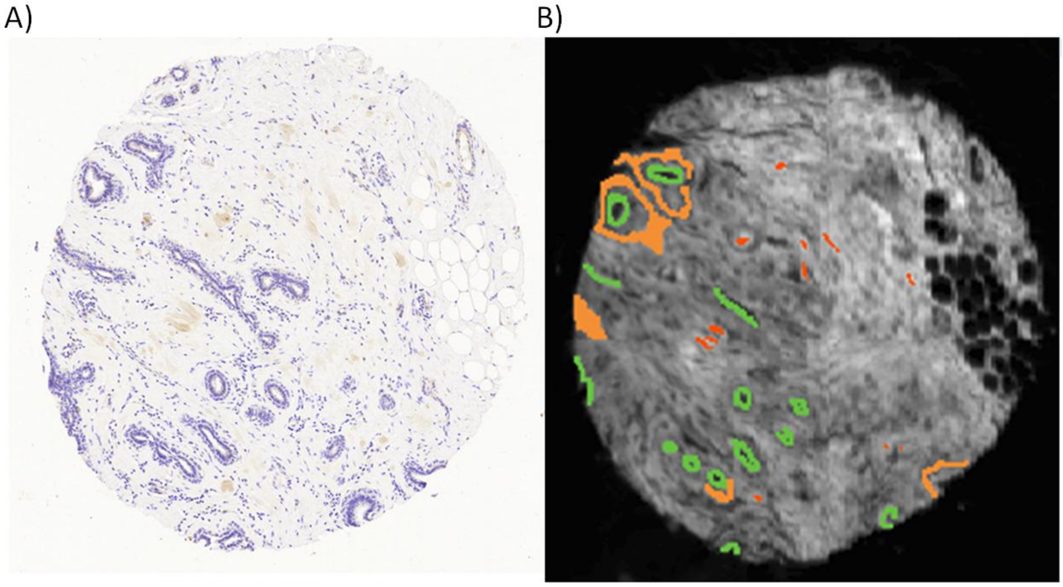


Figure 5.4 A) Bright field image and B) annotated chemical image (generated using the absorbance distribution of Amide I region) of core M6, where yellow indicates stroma, green indicates epithelium and red indicates caveolin-1 stain

5.2.4.2 Data Pre-processing Methods

Raw data collected was firstly run through a spectral quality test, and only spectra that had an intensity threshold between 0.07 and 2 were retained. The cancerous spectra were randomly selected to match the number of the spectra from the normal cores. In this case, 20 principal components were kept to reach the best noise reduction results with the consideration of preserving most biological information.After noise reduction, the only wavenumber ranges: 1000 cm^{-1} to 1300 cm^{-1}, 1500 cm^{-1} to 1750 cm^{-1} and 3005 cm^{-1} to 3550 cm^{-1} were kept, as these ranges contain most of the biological information related to cancer diagnosis and not influenced by the paraffin bands which is the material used to embedded tissue samples. A vector normalisation was performed. Savitzky-Golay first derivative filtering was applied to further smooth the spectra. A 19-data-point system was applied to reach the best result. The detailed information about pre-processing steps was mentioned in Chapter 4.

5.2.4.3 Classification Methods

Both PCA and random forest were employed to separate cancerous and normal spectra. PCA on pre-processed datasets was conducted to visualize trends and patterns of data. The output of PCA contains scores, loadings and the residuals [27]. For spectral data, the residuals corresponding to the residual spectra provide important biological and chemical information, which spectral variation has not been explained. Scores are readings in the same way as variables, which can be plotted and interpreted in many different ways. It is common to plot scores in scatter plots indicating the pattern of the raw data [28]. Loadings define the principal components. Scores and loadings are usually plotted on the same biplots, which both indicate the pattern and the reason behind the variance.

Random forest is a classification method operates by constructing decision trees to vote the class of unknown samples [29–31]. Each tree which was previously trained on similar data performs a classification. Each decision is collected and the forest chooses the most popular class based on decisions made by all trees in the forest. The favoured class will be assigned as the predicted class of unknown samples. Random forest classification is an ensemble method based on tree classifiers. It is proved to give better and more stable performance compared with single tree classifiers. The detailed working mechanism was presented in Chapter 4 (section 4.2.4.3.1).

5.2.4.4 Statistical Analysis Methods

Receiver operator characteristic (ROC) curves are used to access the performance of models. It presents the inherent trade-off between sensitivity and specificity. The Area under curve (AUC) of the ROC curve was also used to describe the performance of models in numbers. Together with the confusion matrix, which describes performance of models in tables, sensitivity, specificity, ROC curve and AUC of classification results will be used to access the performance of each classifier.

The detailed equation and explanations of each method were presented in Chapter 4 (section 4.2.4.4).

5.3. Results and discussion

Since the focus of the research was on caveolin-1 stained stroma regions, to eliminate possible separation caused by differences between cancerous and non-cancerous stroma spectra, comparisons were made as two separate groups: a cancerous caveolin-1 stained stroma group and a non-cancerous caveolin-1 stained stroma group. The cancerous group only contains spectra extracted from cancerous cores of different patients, and the non-cancerous group only contains spectra selected from sample cores of non-cancerous patients. Both principal component analysis and random forest were conducted for each group. PCA was a dimension reducing unsupervised method finding patterns in data while random forest focuses more on the details of data variations with supervised learning of data features. Combination of these two methods could produce a more comprehensive understanding of the spectra and their classification results.

5.3.1 Results produced using principal component analysis

5.3.1.1 PCA results produced using spectra selected from cancerous cores

A database was constructed from the 18 training cores which consist of 17464 spectra (17054 non-stained stroma spectra, and 410 caveolin-1 stained stroma spectra). PCA was applied to the database constructed after each spectrum was quality checked, noise reduced, vector normalized, and derivatized. A PCA score plot and loading plots of non-caveolin-1 stained stroma and caveolin-1 stained stroma selected from cancerous cores were plotted in Figure 5.5. Caveolin-1 stained stroma was plotted in blue while unstained stroma was plotted in red. It could be seen that no clear separation was observed in three-dimensional score plot in the axes of principal component (PC) 1 which consists 55.2% variance of the dataset, PC 2 (26.4%), and PC 3 (9.5%).

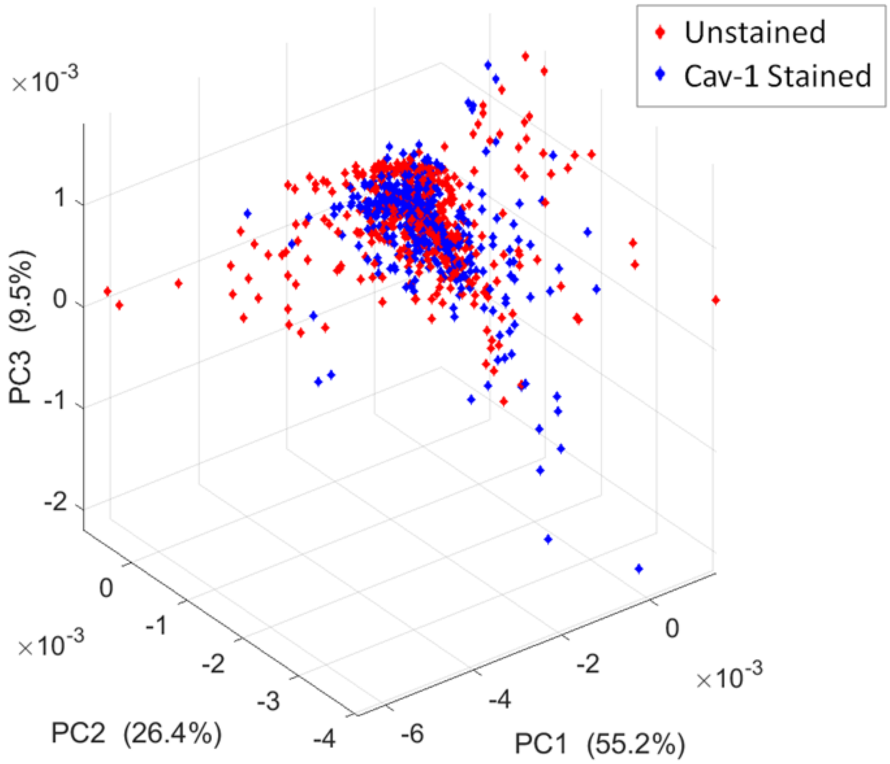


Figure 5.5 Three dimensional PCA score plot between unstained and caveolin-1 stained stroma with the axes of PC 1 (55.2%), PC 2 (26.4%), and PC 3 (9.5%) in cancerous cores, where unstained stroma data was plotted in red diamonds while caveolin-1 stained stroma data was plotted in blue diamonds

The Loadings of each principal component was plotted in Figure 5.6. PC 1 was mainly formed by Amide I (peaks at 1630 and 1655 cm^{-1}), which weighs 55.2% of the total variance of the spectra. PC 2 was mainly composed by Amide I (peaks at 1665 and 1675 cm^{-1}) and Amide II (1557 cm^{-1}), which counts 26.4% of the spectra variation. PC 3 was mainly made by variation of Amide II (1550 cm^{-1}) intensity, which counts 9.5% of the total variance. The variance of this set of spectra was mainly on Amide I and II, which indicates there were possible protein content differences although these differences were not just between caveolin-1 stained stroma and surrounded unstained stroma.

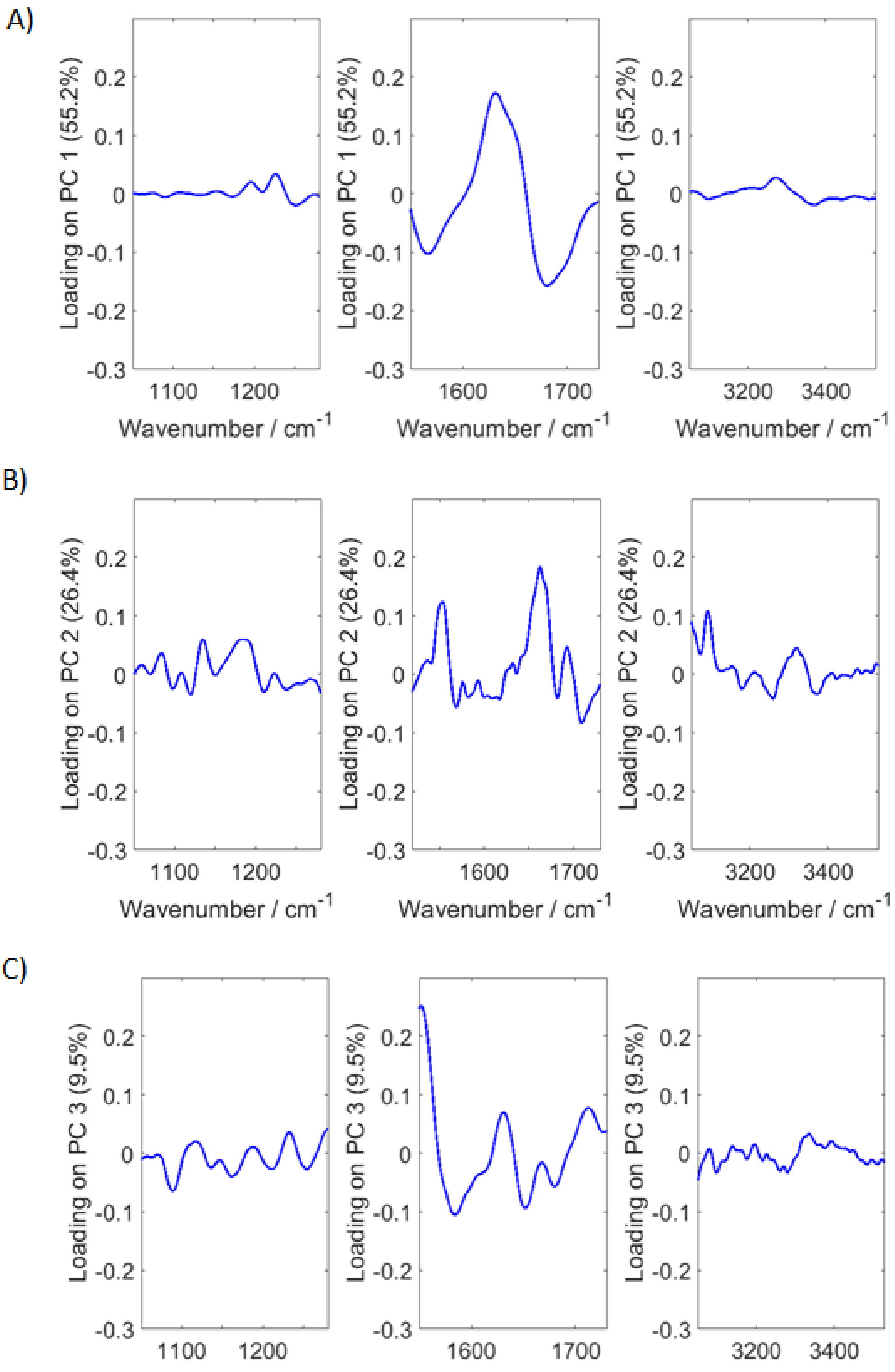


Figure 5.6 The Loading plots of A) PC1 (55.2%) B) PC 2 (26.4%) C) PC 3 (9.5%) in the ranges of 1000 cm^{-1} to 1300 cm^{-1}, 1550 cm^{-1} to 1750 cm^{-1} and 3005 cm^{-1} to 3550 cm^{-1}

5.3.1.2 PCA results produced using spectra selected from normal cores

PCA score plot and loading plots of unstained stroma and caveolin-1 stained stroma selected from normal cores were plotted in Figure 5.7. Caveolin-1 stained stroma was plotted in red while unstained stroma was plotted in blue. It could be seen that there was no clear separation between caveolin-1 stained stroma and unstained stroma from normal cores in PCA score plot as a large number of red scattering points overlapped with the blue dots. Spectra in the unstained stroma have more variations as scores points were more spread comparing with caveolin-1 stained stroma with clustered score points.

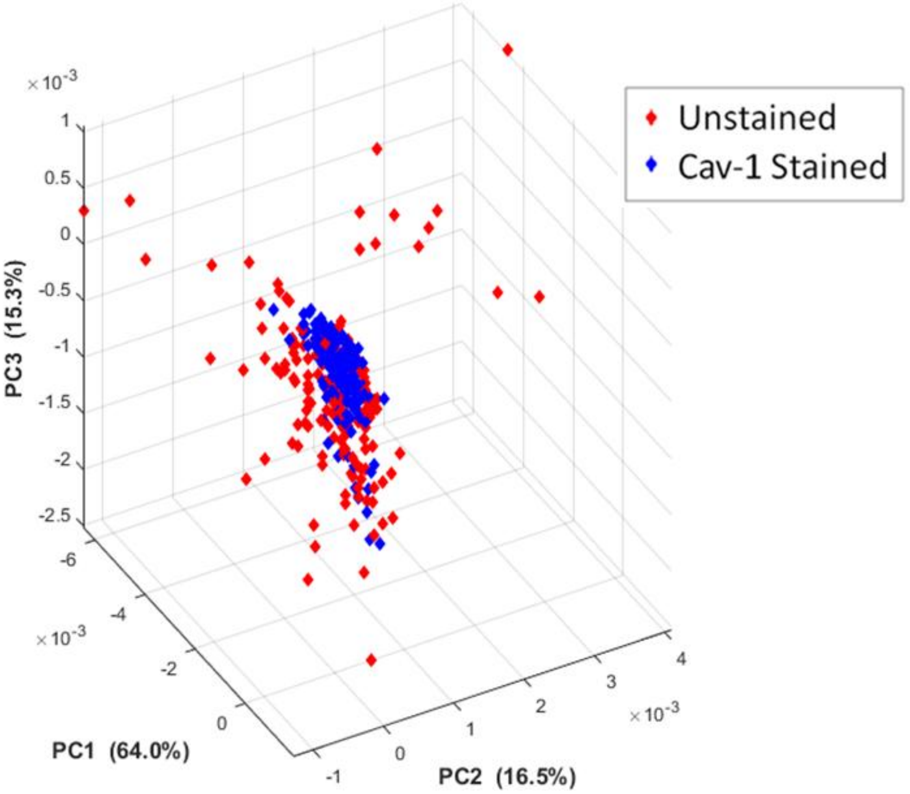


Figure 5.7 Three dimensional PCA score plot between unstained and caveolin-1 stained stroma with the axes of PC 1 (64.0%), PC 2 (16.5%), and PC 3 (15.3%) in normal cores, where unstained stroma data was plotted in red diamonds while caveolin-1 stained stroma data was plotted in blue diamonds

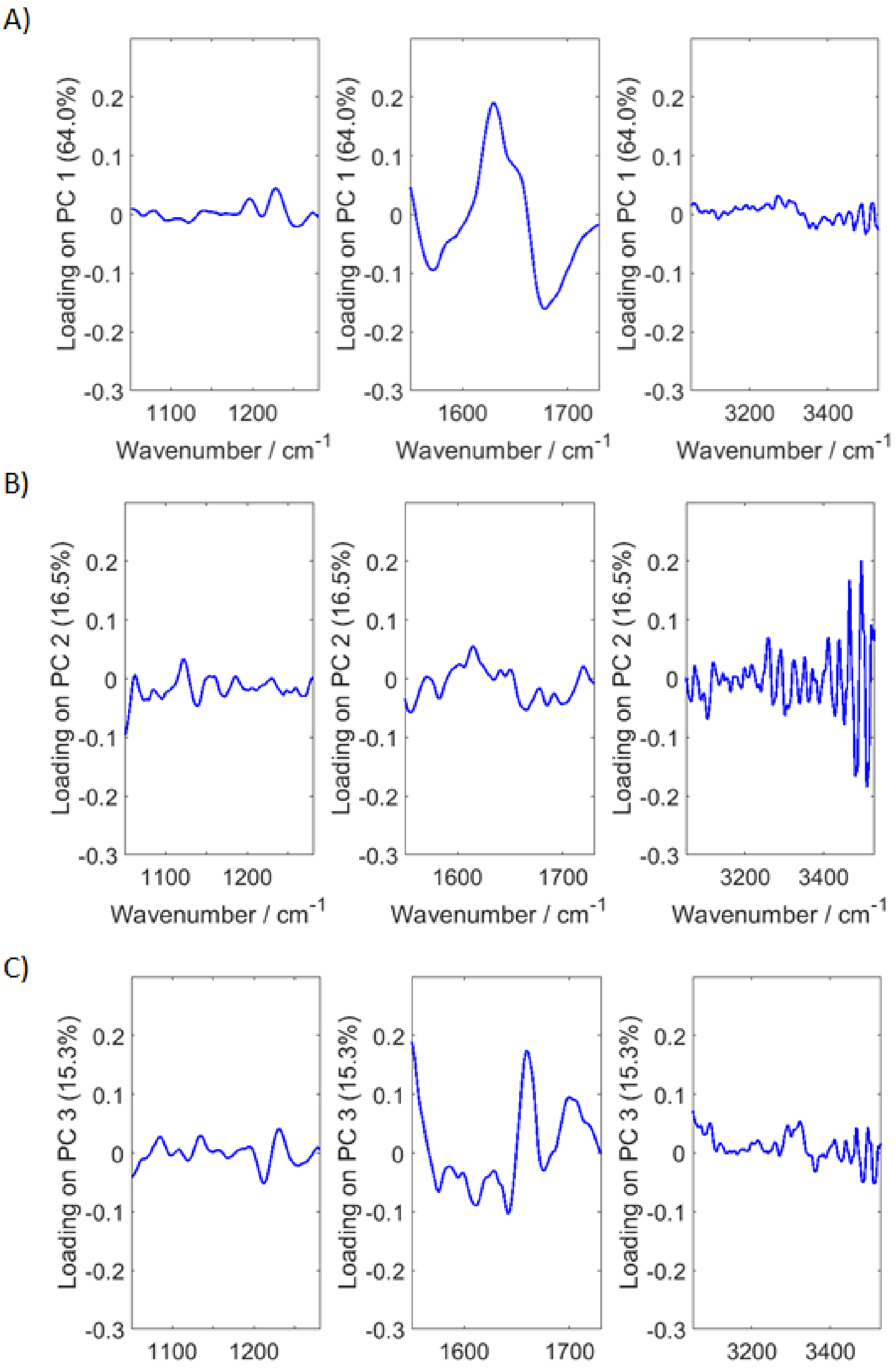


Figure 5.8 The Loading plots of A) PC1 (64.0%) B) PC 2 (16.5%) C) PC 3 (15.3%) in the ranges of 1000 cm^{-1} to 1300 cm^{-1}, 1550 cm^{-1} to 1750 cm^{-1} and 3005 cm^{-1} to 3550 cm^{-1}

The Loadings of each principal component was plotted. PC 1 was mainly formed by Amide I (peak at 1630 and 1670 cm^{-1}), which weighs 64% of the total variance of the spectra. PC 2 was also mainly composed of suspected water vapour, which counts 16.5% of the spectra variation. PC 3 was mainly made by variation of Amide I (1665 cm^{-1}) intensity, Amide II (1550 cm^{-1}) and C=O stretch peak at 1710 cm^{-1}, which counts 15.3% of the total variance. Variation types of stroma could be the main reason of Amide I and II differences. Differences between caveolin-1 stained stroma and other stroma were not clear using PCA.

5.3.1.3 PCA results produced using caveolin-1 to separate cancerous and normal spectra

The Loading plots of A) PC1 (55.2%) B) PC 2 (26.4%) C) PC 3 (9.5%) in the ranges of 1000 cm^{-1} to 1300 cm^{-1}, 1500 cm^{-1} to 1750 cm^{-1} and 3005 cm^{-1} to 3550 cm^{-1} and PCA score plot of caveolin-1 stained stroma from cancerous cores and caveolin-1 stained stroma selected from normal cores were plotted in Figure 5.9 and Figure 5.10. Caveolin-1 stained stroma from cancerous cores was plotted in red while normal caveolin-1 stroma was plotted in blue. It could be seen that normal caveolin-1 spectra were more tightly clustered while the caveolin-1 from cancerous patients were more spread. The widespread may because cancerous cores included all four stages of breast cancer and in each stage cells could have different variation from normal cells. A trend of separation was observed between two classes even through overlaps of some scattering points were observed. To validate the possible separation between cancerous and normal caveolin-1 stained spectra, a more specified classification method, random forest, would be applied later in this chapter.

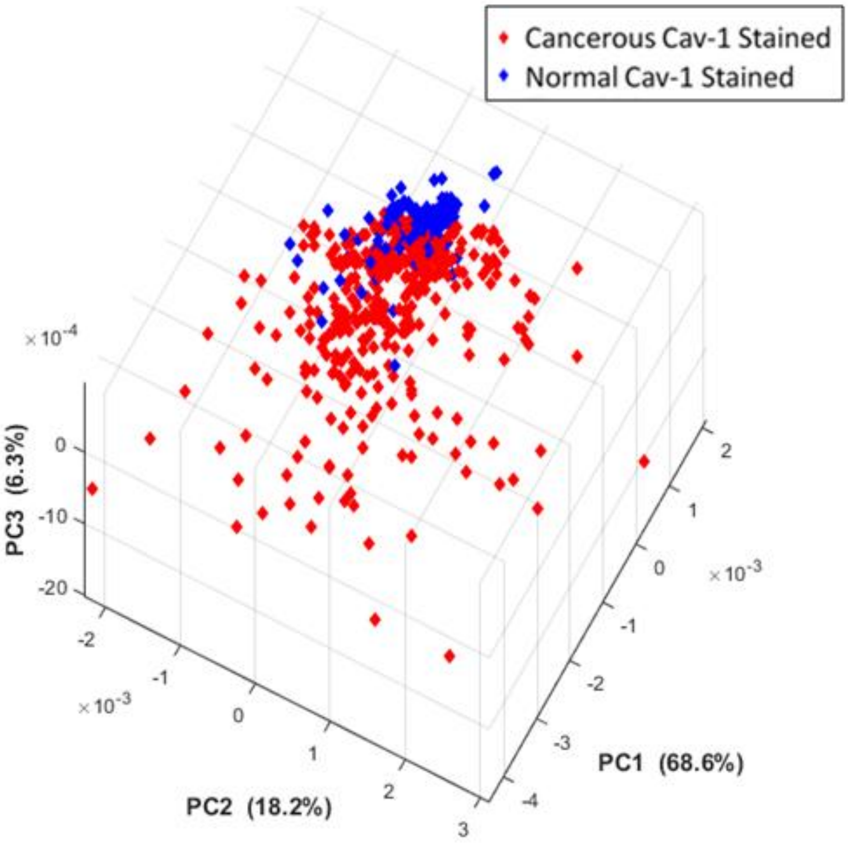


Figure 5.9 Three dimensional PCA score plot between caveolin-1 stained stroma from cancerous cores and caveolin-1 stained stroma from normal cores with the axes of PC 1 (68.6%), PC 2 (18.2%), and PC 3 (6.3%), where cancerous stained stroma data was plotted in red diamonds while normal caveolin-1 stained stroma data was plotted in blue diamonds

It could be seen in Figure 5.10, PC 1 was mainly formed by Amide II (peaks at 1550 and 1580 cm^{-1}), which weighs 68.6% of the total variance of the spectra. PC 2 counts 18.2% of the total variance which was caused by mainly Amide I (1630 cm^{-1} and 1680 cm$^{-1)}$ variations. PC 3, 6.3% of the total variance, was observed peaked at 1620, 1657 and 1707 cm^{-1}, which was mainly caused by Amide I. Difference between cancerous caveolin-1 stained stroma and unstained was tiny and difficult to be found using PCA. Random forest should be conducted to further classify cancerous and normal spectra.

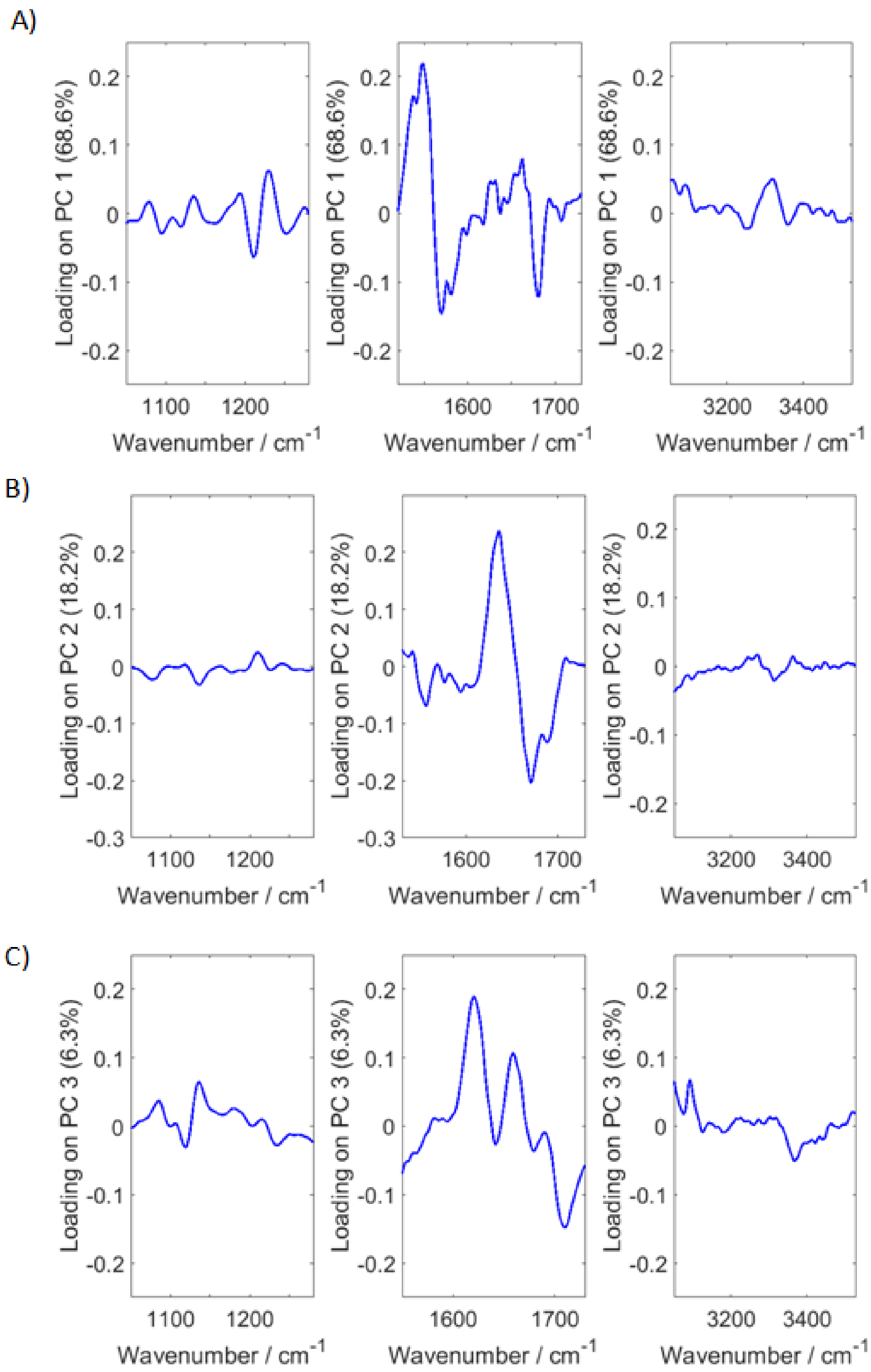


Figure 5.10 The Loading plots of A) PC1 (68.6%) B) PC 2 (18.2%) C) PC 3 (6.3%) in the ranges of 1000 cm^{-1} to 1300 cm^{-1}, 1500 cm^{-1} to 1750 cm^{-1} and 3005 cm^{-1} to 3550 cm^{-1}

5.3.2 Results produced using Random Forest

5.3.2.1 Random forest results produced using spectra selected from cancerous cores

A training database was constructed from the 18 training cores which consisted of 17464 spectra (17054 non-stained stroma spectra, and 410 caveolin-1 stained stroma spectra). The optimal situation to prevent classifier bias was for each class to consist of equal numbers of spectra. Bias was minimized in the training set by selecting equal numbers of spectra from the class with the number randomly selected being the size of the smallest class.

A Random Forest was then constructed using 410 spectra per class. There were two classes, including non-stained and caveolin-1 stained stroma. Each spectrum was quality checked, noise reduced, vector normalized, and derivatized prior to being used for training. 500 trees were used. The node size parameter, which limits how large each decision tree could grow, was set to 1.

Confusion matrices provide a quantitative measure of performance and enable the correctness of classification for each class to be determined. Furthermore, the sources of misclassification could be easily identified from a confusion matrix, revealing which classes were difficult to discriminate between. In Table 5.1, a random forest model validation training confusion matrix was shown, which tests the model constructed using the training database on other yet presented spectra taken from the same patient group. The validation test results were 69.2% and 84.9% for unstained and Cav-1 stained stroma respectively. These results were hot as high as expected possibly due to lack of training spectra. The ROC curve was plotted in Figure 5.11 and an AUC of 0.85 was found. It was noted that the AUC curves were not smooth, which also indicates a lack of training spectra.

Table 5.1 Confusion matrix of model validation for caveolin-1 stained and non-stained stroma from cancerous cores

Training Validation Cancer Group		Predicted / %	
		Unstained Stroma	Cav-1 Stained
True / %	Unstained Stroma	69.2	30.8
	Cav-1 Stained	15.1	84.9

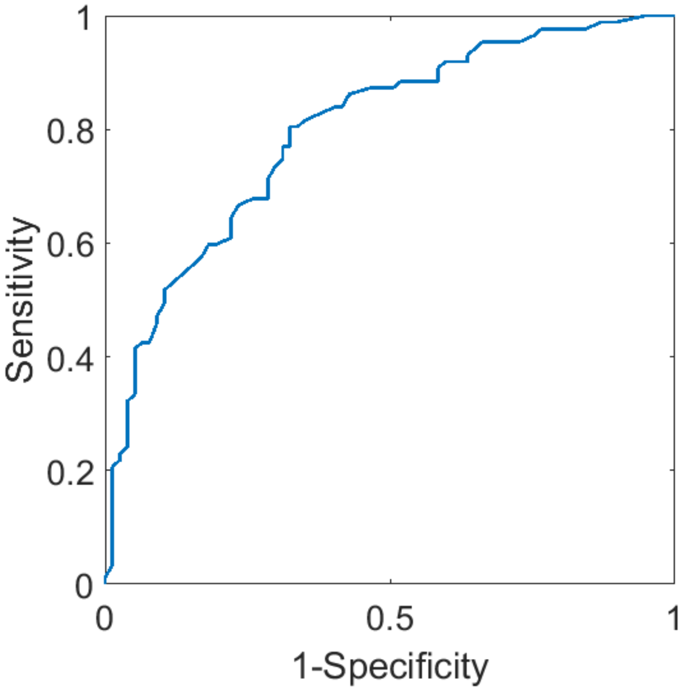


Figure 5.11 Training validation ROC curve for caveolin-1 stained and non-stained stroma from cancerous cores

The Random Forest classifier was then tested on the independent test set which consisted of 11842 spectra (11611 non-stained stroma spectra and 231 caveolin-1 stained stroma spectra) from the 10 testing cores. 231 spectra were selected randomly for each class. Each tree in the Random Forest decides the class of the unknown spectrum. The Random Forest then chooses the class having the most votes over all the trees in the forest. Confusion matrices and AUC curves for the independent test were plotted in Table 5.2 and Figure 5.12 respectively. The model was better at recognising caveolin-1 stained stroma than unstained stroma. The resulting AUC value of 0.79 indicates that there was a relatively good differentiation between stained and non-stained stroma. This result could be further improved by applying more spectra for model construction.

Table 5.2 Confusion matrix of independent test for caveolin-1 stained and non-stained stroma from cancerous cores

Independent Test Cancer Group		Predicted / %	
		Unstained Stroma	Cav-1 Stained
True / %	Unstained Stroma	64.5	35.5
	Cav-1 Stained	16.0	84.0

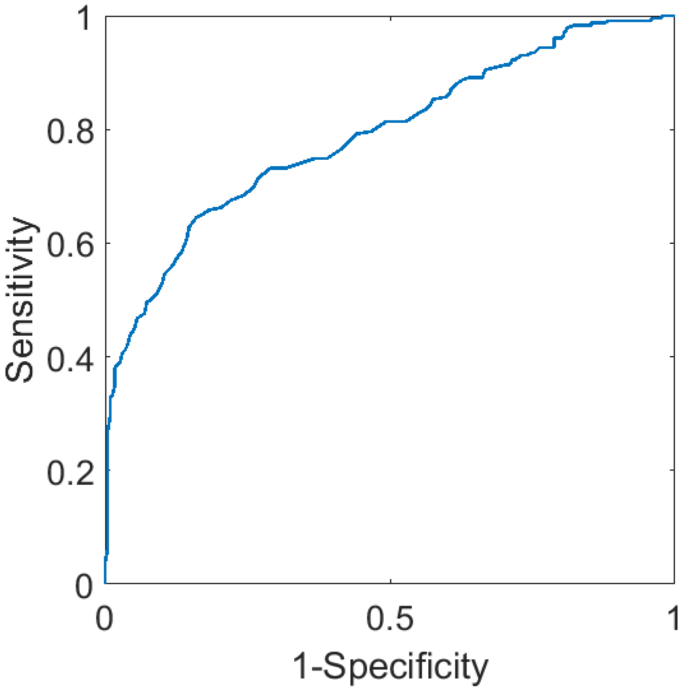


Figure 5.12 Independent test ROC curve for caveolin-1 stained and non-stained stroma from cancerous cores

5.3.2.2 Random forest results produced using spectra selected from normal cores

A training database was constructed from the 7 training cores which consisted of 13541 spectra (13368 non-stained stroma spectra, and 173 caveolin-1 stained stroma spectra). Equal numbers of spectra were selected from each class with the number randomly selected being the size of the smallest class. A Random Forest was then constructed using 173 spectra per class. Each spectrum was quality checked, noise reduced, vector normalized, and derivatized prior to being used for training. The same training configuration of 500 trees and 1 node size was applied.

In Table 5.3, a random forest model validation training confusion matrix was shown, which tests the model constructed using the training database on other spectra taken from the same patient group. The validation test results were 85% and 74%. The ROC curve was shown in Figure 5.13, and an AUC of 0.88 was found. Again the AUC curves were not smooth, which indicates lack of training spectra.

Table 5.3 Confusion matrix of model validation for caveolin-1 stained and non-stained stroma from normal cores

Training Validation Normal Group		Predicted / %	
		Unstained Stroma	Cav-1 Stained
True / %	Unstained Stroma	85.29	14.71
	Cav-1 Stained	25.71	74.29

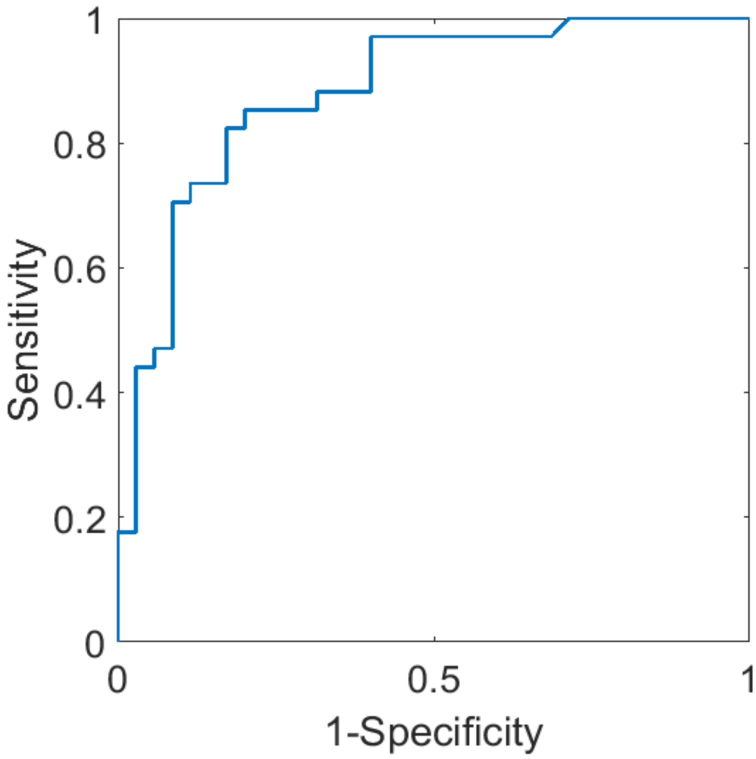


Figure 5.13 Training validation ROC curve for caveolin-1 stained and non-stained stroma from normal cores

The Random Forest classifier was then tested on the independent test set which consisted of 4288 spectra (4206 non-stained stroma spectra and 82 caveolin-1 stained stroma spectra) from the 10 testing cores. 82 spectra were selected randomly for each class.

The confusion matrices and AUC curves for the independent test were plotted in Table 5.4 and Figure 5.14 respectively. The model was better at recognising unstained stroma than caveolin-1 stained stroma. The resulting AUC value of 0.86 indicates that there was a relatively good differentiation between stained and non-stained stroma. This result could be further improved by applying more spectra for model construction.

Table 5.4 Confusion matrix of independent test for caveolin-1 stained and non-stained stroma from normal cores

Independent Test Normal Group		Predicted / %	
		Unstained Stroma	Cav-1 Stained
True / %	Unstained Stroma	89.9	10.1
	Cav-1 Stained	50.0	50.0

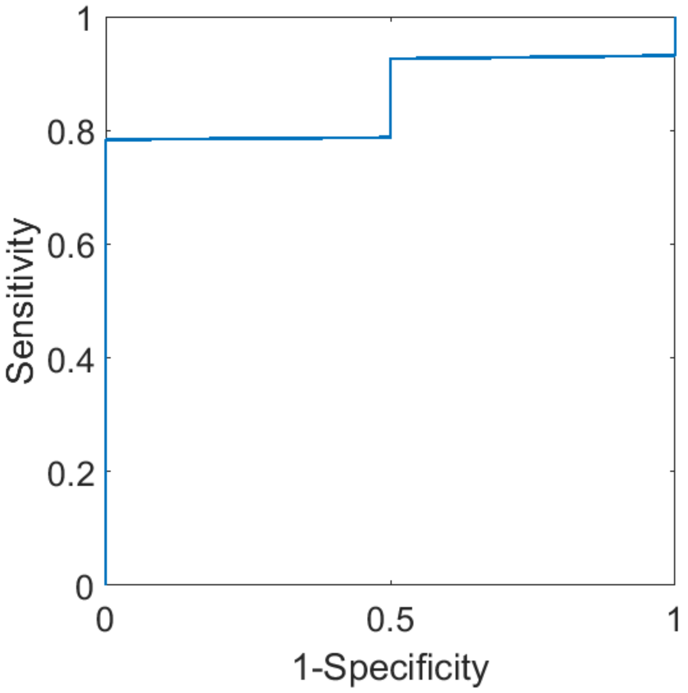


Figure 5.14 Independent test ROC curve for caveolin-1 stained and non-stained stroma from normal cores

5.3.2.3 Random forest results produced using caveolin-1 to separate cancerous and normal spectra

A training database was constructed of 550 caveolin-1 stained stroma spectra (410 from cancerous cores and 140 from normal cores). Equal numbers of spectra were selected from each class with the number randomly selected being the size of the smallest class. A Random Forest was then constructed using 140 spectra per class. The random forest model was constructed by 500 decision trees with node size equal to 1.

In Table 5.5, a random forest model validation training confusion matrix was shown, which tests the model constructed using the training database on other spectra taken from the same patient group. Validation test results, 96% and 88% were relatively good. The ROC curve was plotted in Figure 5.15, and an AUC of 0.96 was found. AUC curves could be further smoothed by adding more training spectra.

Table 5.5 Confusion matrix of model validation for caveolin-1 stained from cancerous and normal cancerous cores

Training Validation Cancer VS. Normal		Predicted / %	
		Cav-1 Stained Cancerous Spectra	Cav-1 Stained Normal Spectra
True / %	Cav-1 Stained Cancerous Spectra	95.7	4.3
	Cav-1 Stained Normal Spectra	12.1	87.9

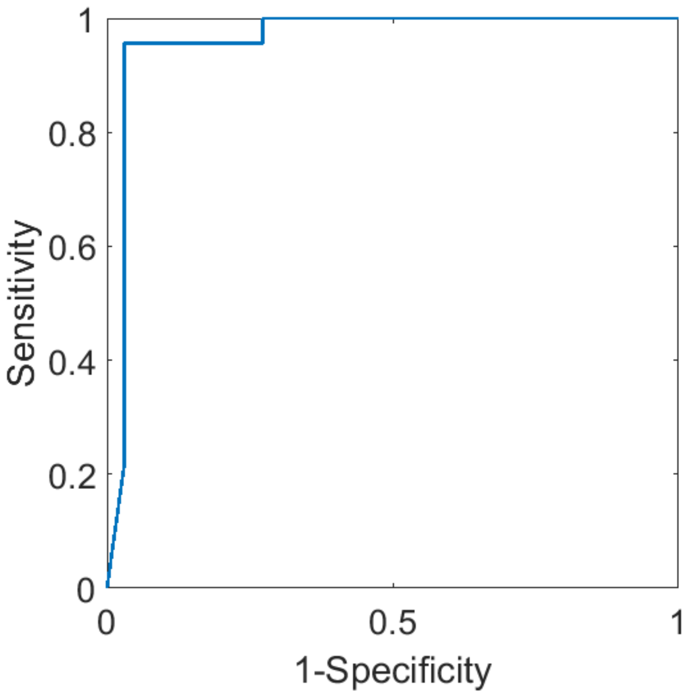


Figure 5.15 Training validation ROC curve for caveolin-1 stained stroma from cancerous and normal cores

The Random Forest classifier was then tested on the independent test set which consisted of 231 caveolin-1 stained cancerous spectra and 35 caveolin-1 stained normal spectra. 35 spectra were selected randomly for each class.

Confusion matrices and AUC curves for the independent test were plotted in Table 5.6 and Figure 5.16 respectively. The model was better at recognising unstained stroma than caveolin-1 stained stroma. The resulting AUC value of 0.96 indicates that there was a very good differentiation between stained cancerous and normal stroma. Considering the similarities between malignant and non-malignant spectra, the ROC curves appear surprisingly good.

Table 5.6 Confusion matrix of independent test for caveolin-1 stained stroma from cancerous and normal cores

Independent Test Cancer VS. Normal		Predicted / %	
		Cav-1 Stained Cancerous Spectra	Cav-1 Stained Normal Spectra
True / %	Cav-1 Stained Cancerous Spectra	100	0
	Cav-1 Stained Normal Spectra	28.6	71.4

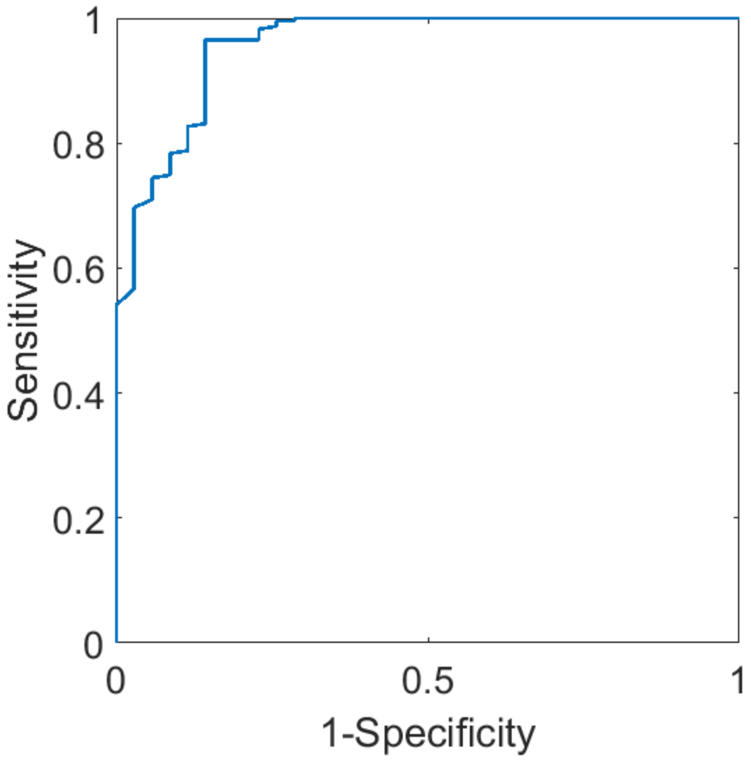


Figure 5.16 Independent test ROC curve for caveolin-1 stained stroma from cancerous and normal cores

5.3.2.3.1 Importance study on the results produced using random forest classifier

To further investigate whether the separation observed in the independent test of cancerous and normal caveolin-1 stained stroma spectra was caused by biological differences between cancerous and normal cores, an importance study of all wavenumbers was conducted. The importance plots show the importance of each spectral feature. Each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest, shown in Figure 5.17. The main peaks were observed at 1188, 1192, 1209, 1559, 3237 and 3493 cm^{-1}. Contribution peaks at 1188, 1192 and 1209 cm^{-1} could be related to changes in nucleic acid (PO_2^- stretch) between cancerous and normal stroma tissue. 1559 cm^{-1} was in the range of Amide II. 3237 and 3493 cm^{-1} were due mainly to the N-H stretch. Therefore this separation was mainly based on biological differences between cancerous and normal tissue.

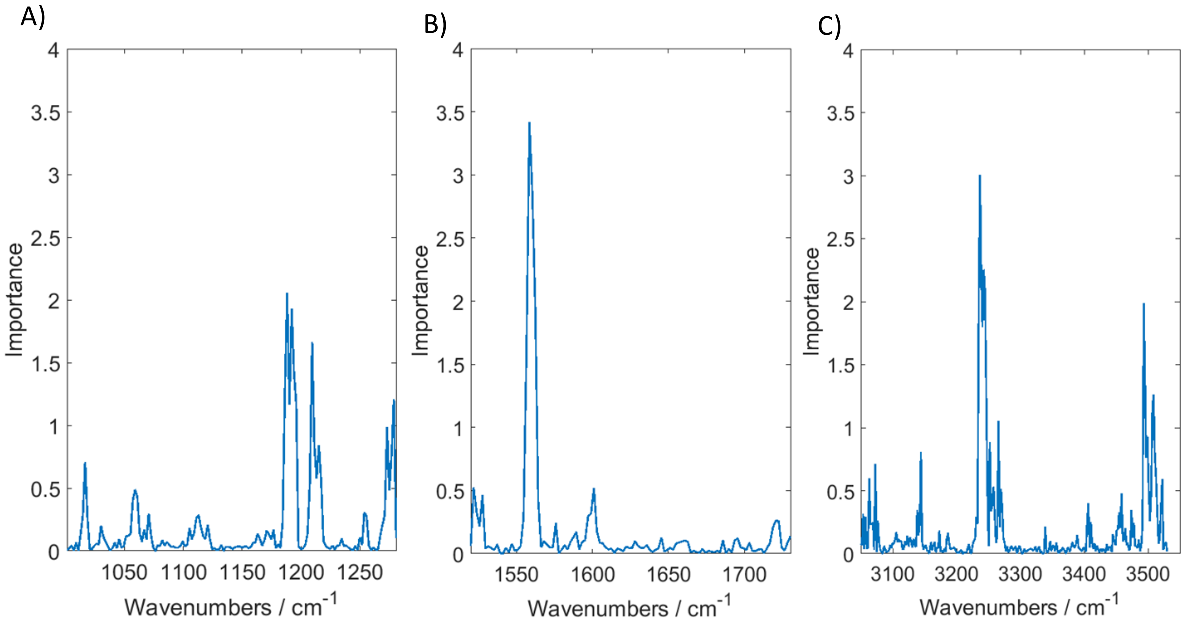


Figure 5.17 Importance of each wavenumber in the model, where in the ranges of A) 1000 cm^{-1} to 1300 cm^{-1}, B) 1500 cm^{-1} to 1750 cm^{-1} and C) 3005 cm^{-1} to 3550 cm^{-1}

5.4. Conclusion and future works

For both cancerous and normal cores, caveolin-1 could not be separated from the surrounding stroma. There were two possible reasons for this. The first one was the quantity of caveolin-1 spectra sample was too low for analysis. A low number of samples reduce the accuracy of separation between stroma and caveolin-1. The second cause could be the doubtful accuracy of annotation due to the nature of caveoline-1 stain. The stain was not particularly strong.

Caveolin-1 staining could be used for cancer classification, as it produced high accuracy (100% and 71.4%). The classification accuracy for cancerous spectra was even higher than the classification result obtained in Chapter 4 (89.6% for cancerous class). However, only a very small number of spectra from stained stroma were available in this study. Further experiments should be continued to ensure its reproducibility.

More information should be collected to ratify the results obtained from experiments, as a small sample size could not reflect actual correlations between stroma and caveolin-1 spectra. More sample cores should be provided for further research. Clearer caveolin-1 stained sample could improve the accuracy of classifiers. It could be seen that it was difficult to locate caveolin-1 on the chemical image because both of the sharpness of images and the colour of stains. Caveolin-1 stain was in light brown. It was hard to locate the correct area of it on its corresponding chemical image. This nature of caveolin-1 stain increased the difficulty of using it as an indicator of breast cancer since nature reduced the accuracy of annotations.

5.5. References

[1] S. Hehlgans, N. Cordes, Caveolin-1: an essential modulator of cancer cell radio-and chemoresistance., Am. J. Cancer Res. (2011).

[2] A.F.G. Quest, J.L. Gutierrez-Pajares, V.A. Torres, Caveolin-1: An ambiguous partner in cell signalling and cancer, J. Cell. Mol. Med. (2008). doi:10.1111/j.1582-4934.2008.00331.x.

[3] T.M. Williams, M.P. Lisanti, The caveolin proteins, Genome Biol. (2004). doi:10.1186/gb-2004-5-3-214.

[4] J.A. Engelman, X.L. Zhang, M.P. Lisanti, Genes encoding human caveolin-1 and -2 are co-localized to the D7S522 locus (7q31.1), a known fragile site (FRA7G) that is frequently deleted in human cancers, FEBS Lett. (1998). doi:10.1016/S0014-5793(98)01134-X.

[5] T. Bouras, M.P. Lisanti, R.G. Pestell, Caveolin in breast cancer, Cancer Biol. Ther. 3 (2004) 931–941. doi:10.4161/cbt.3.10.1147.

[6] C. Boscher, I.R. Nabi, CAVEOLIN-1: Role in Cell Signaling, in: 2012: pp. 29–50. doi:10.1007/978-1-4614-1222-9_3.

[7] A.M. Fra, N. Mastroianni, M. Mancini, E. Pasqualetto, R. Sitia, Human caveolin-1 and caveolin-2 are closely linked genes colocalized with WI-5336 in a region of 7q31 frequently deleted in tumors, Genomics. (1999). doi:10.1006/geno.1998.5723.

[8] S.W. Lee, C.L. Reimer, P. Oh, D.B. Campbell, J.E. Schnitzer, Tumor cell growth inhibition by caveolin re-expression in human breast cancer cells, Oncogene. 16 (1998) 1391–1397. doi:10.1038/sj.onc.1201661.

[9] F.C. Bender, M.A. Reymond, C. Bron, A.F.G. Quest, Caveolin-1 levels are down-regulated in human colon tumors, and ectopic expression of caveolin-1 in colon carcinoma cell lines reduces cell tumorigenicity, Cancer Res. (2000).

[10] K. Wiechen, L. Diatchenko, A. Agoulnik, K.M. Scharff, H. Schober, K. Arlt, B. Zhumabayeva, P.D. Siebert, M. Dietel, R. Schäfer, C. Sers, Caveolin-1 is down-regulated in human ovarian carcinoma and acts as a candidate tumor suppressor gene, Am. J. Pathol. (2001). doi:10.1016/S0002-9440(10)63010-6.

[11] K. Wiechen, C. Sers, A. Agoulnik, K. Arlt, M. Dietel, P.M. Schlag, U. Schneider, Down-Regulation of Caveolin-1, a Candidate Tumor Suppressor Gene, in Sarcomas, Am. J. Pathol. 158 (2001) 833–839. doi:10.1016/S0002-9440(10)64031-X.

[12] N. Cordes, S. Frick, T.B. Brunner, C. Pilarsky, R. Grützmann, B. Sipos, G. Klöppel, W.G. McKenna, E.J. Bernhard, Human pancreatic tumor cells are sensitized to ionizing radiation by knockdown of caveolin-1, Oncogene. 26 (2007) 6851–6862. doi:10.1038/sj.onc.1210498.

[13] Y. Lavie, M. Liscovitch, Changes in lipid and protein constituents of rafts and caveolae in multidrug resistant cancer cells and their functional consequences., Glycoconj. J. 17 (n.d.) 253–9. http://www.ncbi.nlm.nih.gov/pubmed/11201798 (accessed September 12, 2019).

[14] H.D. Nelson, E.S. O'Meara, K. Kerlikowske, S. Balch, D. Miglioretti, Factors Associated With Rates of False-Positive and False-Negative Results From Digital Mammography Screening: An Analysis of Registry Data, Ann. Intern. Med. 164 (2016) 226. doi:10.7326/M15-0971.

[15] N.M.T. Jawhar, Tissue Microarray: A rapidly evolving diagnostic and research tool, Ann. Saudi Med. 29 (2009) 123–127. doi:10.4103/0256-4947.51806.

[16] A.S. Singh, A.K.S. Sau, Tissue Microarray: A powerful and rapidly evolving tool for high-throughput analysis of clinical specimens, Int. J. Case Reports Images. 01 (2010) 1. doi:10.5348/ijcri-2010-09-1-RA-1.

[17] C. Hughes, L. Gaunt, M. Brown, N.W. Clarke, P. Gardner, Assessment of paraffin removal from prostate FFPE sections using transmission mode FTIR-FPA imaging, Anal. Methods. (2014). doi:10.1039/c3ay41308j.

[18] A. Tfayli, O. Piot, A. Durlach, P. Bernard, M. Manfait, Discriminating nevus and melanoma on paraffin-embedded skin biopsies using FTIR microspectroscopy, Biochim. Biophys. Acta - Gen. Subj. (2005). doi:10.1016/j.bbagen.2005.04.020.

[19] A. Tfayli, C. Gobinet, V. Vrabie, R. Huez, M. Manfait, O. Piot, Digital dewaxing of Raman signals: Discrimination between nevi and melanoma spectra obtained from paraffin-embedded skin biopsies, Appl. Spectrosc. (2009). doi:10.1366/000370209788347048.

[20] C. Gobinet, V. Vrabie, A. Tfayli, O. Piot, R. Huez, M. Manfait, Pre-processing and Source Separation methods for Raman spectra analysis of biomedical samples, in: Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc., 2007. doi:10.1109/IEMBS.2007.4353773.

[21] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, M. Manfait, Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies., Analyst. 133 (2008) 197–205. doi:10.1039/b715924b.

[22] E. Ly, O. Piot, A. Durlach, P. Bernard, M. Manfait, Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition, Analyst. (2009). doi:10.1039/b820998g.

[23] P. Bassan, A. Sachdeva, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Automated high-throughput assessment of prostate biopsy tissue using infrared spectroscopic chemical imaging, in: M.N. Gurcan, A. Madabhushi (Eds.), International Society for Optics and Photonics, 2014: p. 90410D. doi:10.1117/12.2043290.

[24] F. Lyng, E. Gazi, P. Gardner, Preparation of Tissues and Cells for Infrared and Raman Spectroscopy and Imaging, RSC Anal. Spectrosc. Monogr. 01 (2011) 147–185. http://arrow.dit.ie/radrep.

[25] F.N. Pounder, K. Reddy, R. Bhargava, Development of a practical spatial-spectral analysis protocol for breast histopathology using Fourier transform infrared spectroscopic imaging, Faraday Discuss. 187 (2016) 43–68. doi:10.1039/C5FD00199D.

[26] B. Weigelt, F.C. Geyer, J.S. Reis-Filho, Histological types of breast cancer: How special are they?, Mol. Oncol. (2010). doi:10.1016/j.molonc.2010.04.004.

[27] I.T. Jolliffe, Principal Component Analysis, Second Edition, Encycl. Stat. Behav. Sci. (2002). doi:10.2307/1270093.

[28] R. Salzer, H.W. Siesler, Infrared and Raman Spectroscopic Imaging, 2009. doi:10.1002/9783527628230.

[29] L. Breiman, Bagging predictors - Springer, Mach. Learn. (1996). doi:10.1007/BF00058655.

[30] L. Breiman, Random Forrest, Mach. Learn. (2001). doi:10.1023/A:1010933404324.

[31] L. Breitman, A. Cutler, A. Liaw, M. Wiener, randomForest: Breiman and Cutler’s Random Forests for Classification and Regression, Https://Www.Stat.Berkeley.Edu/~breiman/RandomForests/. (2018). doi:10.1023/A.

Chapter 6

Cancer diagnosis of H&E stained breast tissue on glass

6.1. Chapter overview

Nowadays, cancer diagnosis always involves biopsy which is removed from patients. Using biopsies to diagnosis cancer started from the 1930s [1] and developed together with pathologic diagnosis as a gold standard for cancer diagnosis [2]. Traditionally, after slicing into thin layers and floating onto slides, the biopsy samples are examined under an optical microscope. Trained pathologists will make decisions considering the deviations in the cell structures and/or the variations in the distribution of the cells across samples based on their personal experience. Even though a pathologist require over 12 years of training (5-year degree in medicine, 2-year general training and 5 or 6-year specialist training programme in pathology [3], the judgment is still subjective, and often leads to considerable variability in diagnosis[4]. The whole laboratory process is both time consuming and a huge workload for pathologists. With an increasingly gaining population and increase of cancer incidence world widely the situation is getting worse [5].

To alleviate the problem and improve the reliability of cancer diagnosis, the development of automated cancer diagnosis tools, which can eliminate human errors, is curial. Such automated tools can be used to provide a second opinion for patients or as a pre-screening tool for pathologists. Infrared spectroscopy has received particular attention recently, as it can interrogate tissue samples based on their chemical information in a label-free manner. Research has shown that infrared spectroscopy can be used to distinguish cancerous and normal samples in different tissue types including, prostate [6–8], lung [9], colon [10], breast [11–13].

Currently infrared imaging normally requires the tissue sections to be mounted on transmission slides most commonly calcium fluoride (CaF_2) or barium fluoride (BaF_2). Unfortunately, these are both expensive and particularly frangible. The high price and nature of substrates increase the difficulty of adopting infrared imaging into the current clinic cancer diagnosis flow. In an attempt to solve this issue, Pilling *et al.* proposed and subsequently tested applying infrared imaging directly to the H&E stained tissue on glass slides as used by the pathologists [8]. In light of this previous research, in this chapter, breast tissue was used to further investigate the feasibility of cancer diagnosis by combining H&E stained tissue on glass slides and infrared imaging.

6.2. Methodology

6.2.1 H&E stained breast tissue preparation

The sample used for this project is a breast TMA H&E slide, BR20832 (prepared by US BioMax.Inc). The breast tumour tissue microarray contains 192 cases of breast carcinoma, 13 NATs and 3 normal breast tissue cores. Each core is from one patient, so in total 208 patients. The paraffin embed slide was mounted onto a glass slide and H&E stained after de-waxing. A cover glass slide was applied to the TMA after H&E stain.

6.2.2 Instrumentation

The instrument and microscope used in the work reported here were those that have previously detailed in chapter 4.

6.2.3 Experiment Procedure

For data collection, transmission was used with 128 background scans and 96 sample scans. The spectral resolution used was 5 cm^{-1} as it has both time and spectral quality advantages. The H&E stained slide was loaded in the purge box overnight to minimize the effect of water vapour. Every day before data collection, the focus of instrument was calibrated. The received energy intensity level distribution of the pixels on the detector was adjusted to be as even as possible. The number of out of range pixels was ensured to be zero. The range was set from 2800 to 3800 cm^{-1} as the lower wavenumber range was saturated with glass slides. The spectrum collected was then saved and converted into mat files for data analysis process.

6.2.4 Data Analysis methods

All data were pre-processed with MATLAB 2018a. Infrared spectra for each biopsy core were extracted from the mosaic as a 256 × 256 data-cube. Each data-cube consists of 65536 spectra which contain 364 data points.

6.2.4.1 Annotation method

Chemical images of each of the breast tissue cores were generated and overlapped with the microscope image of H&E stained slide. Epithelium and stroma areas were identified and annotated on chemical images generated based on the WHO Classification of Tumours in the Breast [14]. The colour codes used are red (255, 0, 0), green (103,193, 66), purple (166, 55, 156) and yellow (244, 143, 53), which indicates cancerous epithelium, NAT epithelium, cancerous stroma and NAT stroma respectively. The example plot of annotated chemical images is shown in Figure 6.1.

6.2.4.2 Data Pre-processing Methods

Principal component based noise reduction was applied with 50 principal components kept. Spectra were quality tested based on the height of the amide A band. Spectra that have absorbance between 0.07 and 1.2 were retained. Spectral ranges from 3000 to 3700 cm^{-1} were used. Each spectrum was vector normalized and converted to first derivative with a Savitzky−Golay smoothing. The window size of smoothing was 19 data points, a conventional parameter for tissue analysis [8]. The detailed pre-processing information was presented in the Chapter 4.

The dataset was separated into two groups, a training set and an independent test set. Patients in the independent test group were independent of those in the training group for better validation.70 cores were separated into the training set, containing 55 cores (11 NAT and 44 cancerous cores), and an independent test group, containing 15 cores (2 NAT and 13 cancerous cores). The cores used for this chapter were matched with the cores used for model construction in chapter 4, where these cores were placed on CaF_2.

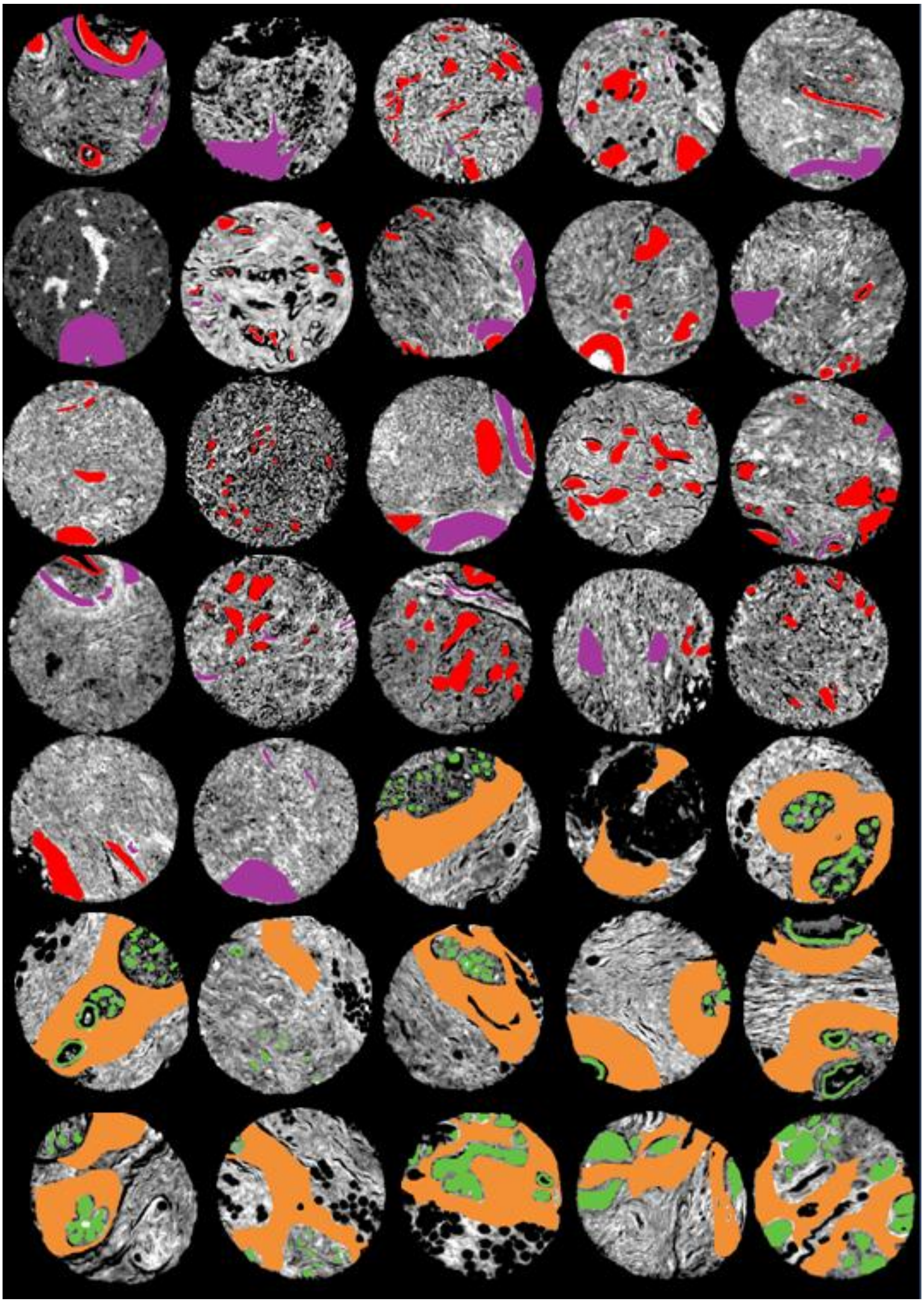


Figure 6.1 Examples of annotation on chemical images, where red indicated cancerous epithelium, purple indicated cancerous stroma, green indicated NAT epithelium and yellow indicated NAT stroma

6.2.4.3 Classification Methods

Both Random forest and AdaBoost were applied in order to separate cancerous and NAT breast tissues. In terms of random forest, a random forest grows many classification trees as required. All trees in the forest are involved in voting the most popular class as the final result from the forest [15]. For AdaBoost, all trees are involved in weighted voting to allocate unknown samples [16]. Detailed information regarding the operation of these algorithms is mentioned in Chapter 4.

6.2.4.4 Statistical Analysis Methods

Receiver operator characteristic (ROC) curves are a common way of representing the inherent trade-off between sensitivity and specificity. Sensitivity equals the true positive over the positive predictions while the specificity was the quotient of true negative and all negative predictions. The detailed equation and explanations are presented in Chapter 4.

Sensitivity, specificity, confusion matrix, ROC curve and AUC of classification results were used to present the performance of each classifier.

A threshold was also applied after the actual classification. It was applied based on the score of each test spectrum after classification. The higher the score the model was more confident. In this chapter, 50% and 10% of the top-scored test spectra were kept as confident predictions. These thresholds were used to exam the model performance at different angels as well.

6.3. Results and discussion

6.3.1. Histological classification results

A histological model was established using AdaBoost with 500 iterations and a learning rate of 0.1. The training dataset consisted of 104252 cancerous epithelium spectra and 22844 NAT epithelium spectra, 130384 cancerous stroma spectra and 112443 NAT stroma spectra. Two classes were balanced by randomly selecting 127096 spectra from the stroma class to match the number of spectra in the epithelium class.

An independent test set was extracted from independent cores, which contains 22360 cancerous and 1611 NAT epithelium spectra, and 17872 cancerous and 23112 NAT stroma spectra.

A training validation test was performed using a subset randomly selected from the training dataset which consisted of 20% of the whole training spectra. The confusion matrix of the validation test was shown in Table 6.1, where the classification accuracies on epithelium and stroma spectra were 79.6% and 53.2% respectively. Comparing with the training validation accuracies obtained from the previous chapter (Chapter 4), where exactly the same cores were used to construct a histological model, the classification accuracies on epithelium and stroma were 98.7% and 98.5% respectively (in Table 4.4). Over 20% difference in classification accuracies indicates poor performance of this model using spectra collected in glass from H&E stained tissue.

Table 6.1 Confusion matrix of training validation of the histological model constructed using AdaBoost classification method with both cancerous and NAT epithelium and stroma spectra

Training Validation		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	79.6	20.4
	Stroma	46.8	53.2

An independent test was conducted to further test the model. 23971 epithelium and 40984 stroma spectra were passed into the model. The confusion matrix obtained from the independent test was shown in Table 6.2. It could be seen that the model classified most of the spectra from the independent test as stroma spectra. For epithelium spectra, only 1.9% of them were correctly classified. The model completely fails on the independent test.

Table 6.2 Confusion matrix of an independent test of the histological model constructed using AdaBoost classification method with both cancerous and NAT epithelium and stroma spectra

Independent Test		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	1.9	98.1
	Stroma	1.4	98.6

The ROC curve also proves the poor performance of the model, which was shown in Figure 6.2. The curve was very close to the diagonal line of the plot box. The AUC value was 0.53 which further indicates poor classification.

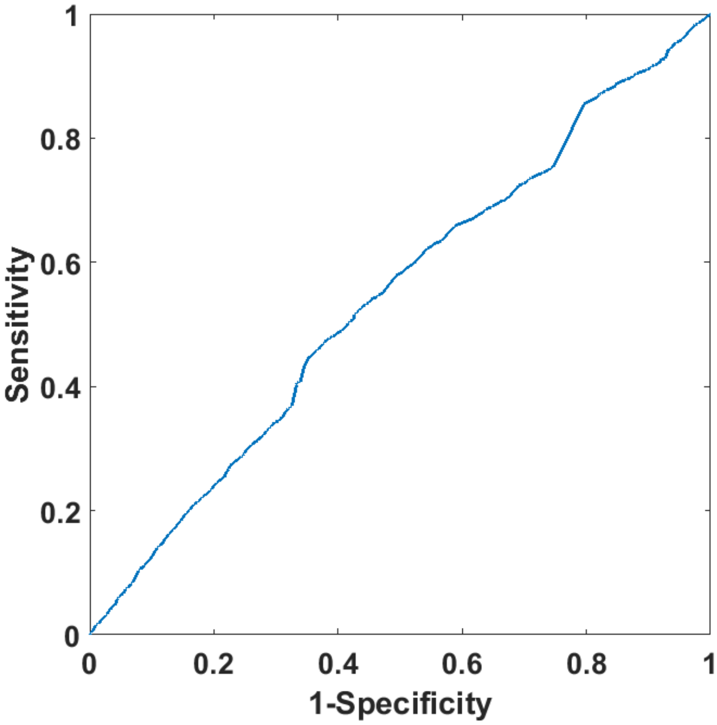


Figure 6.2 ROC of the histological model constructed using AdaBoost classification method with both cancerous and NAT epithelium and stroma spectra

The importance of features used for model training was plotted and presented in Figure 6.3. The main contribution was observed at the peak at 3317 (Amide A) cm^{-1}.

Comparing with the main features found in the same range on CaF_2, 3303 cm^{-1} (Figure 4.5), which represents Amide A band. One possible reason for this observation could be that even though both models focus on Amide A peak to separate epithelium and stroma spectra, without the influence of lower-wavenumber range the model completely fails. These lower-wavenumber features do not carry much weight but cannot be ignored. Another reason could be that the new annotations on these H&E stained cores on the glass were not accurate enough. Too many miss-labelled spectra ruin the model construction, therefore resulting in failure of the independent test. Alternatively the poor classification may be caused by inappropriate data processing steps. A different input data selection method was proposed in the later section to further investigate the problem.

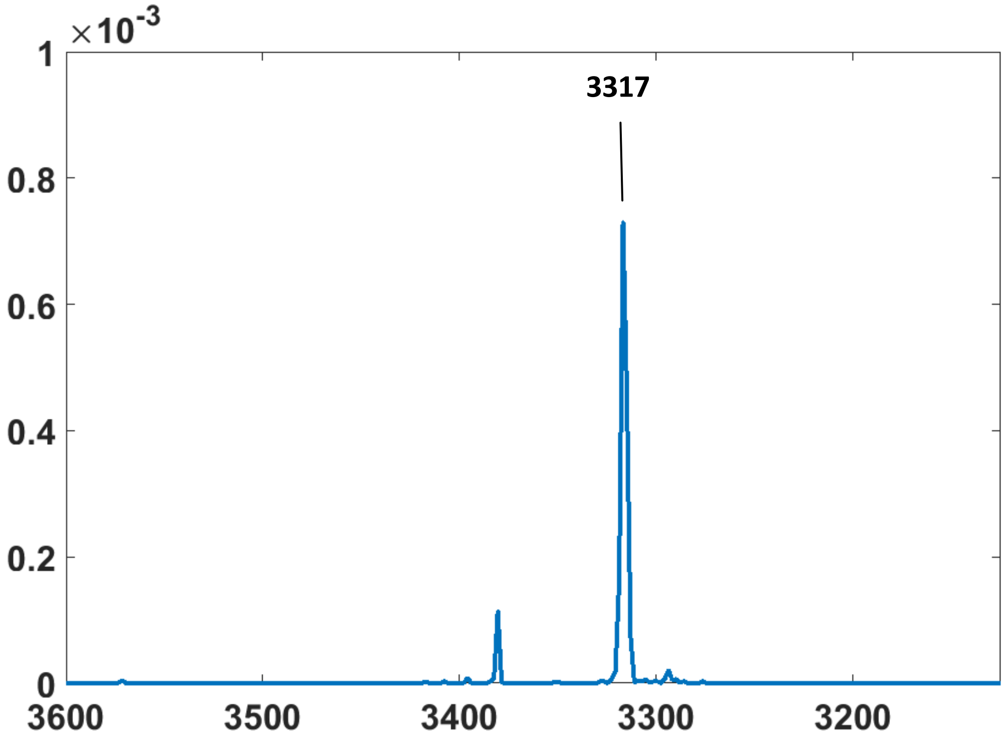


Figure 6.3 Feature importance of the histological model constructed using AdaBoost classification method with both cancerous and NAT epithelium and stroma spectra

6.3.2. Cancer predication classification results

Cancer diagnosis models were established based on both annotated epithelium and stroma spectra using either AdaBoost or random forest.

6.3.2.1. Classification results produced using spectra selected from epithelium areas

Epithelium spectra were extracted from 70 cores, which were consisted of training and independent test datasets. For training spectra, 104252 cancerous and 22844 NAT spectra were selected. To balance both classes, only 22844 cancerous epithelium spectra were randomly selected from the cancerous epithelium group. For independent test spectra, 22360 cancerous and 1611 NAT epithelium spectra were picked from the 15 independent test cores.

6.3.2.1.1. Classification results produced using AdaBoost classification method

A cancer diagnosis model was constructed with the AdaBoost classification method (500 iterations and learning rate 0.1). 20% of the balanced training spectra were randomly selected as the training validation dataset for the model. The confusion matrix of the model was generated and presented in Table 6.3. The classification accuracies on both cancerous and NAT spectra were 70.0% and 62.7% respectively, which were very poor compared with the classification accuracies obtained using the model constructed by epithelium spectra on CaF_2 (Table 4.10).

Table 6.3 Confusion matrix of training validation test of the cancer diagnosis model constructed using AdaBoost classification method with epithelium spectra

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	70.0	30.0
	NAT	37.3	62.7

Independent test spectra were tested by the constructed model. The confusion matrix of the independent test was generated and presented in Table 6.4. The classification accuracies on cancerous and NAT spectra were 57.0% and 60.3% respectively. Comparing with the independent classification accuracies obtained on

CaF_2, 89.4% and 92.8% (Table 4.11), the results were not great but it still could detect some differences between cancerous and NAT epithelium spectra.

Table 6.4 Confusion matrix of the independent test of the cancer diagnosis model constructed using AdaBoost classification method with epithelium spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	57.0	43.0
	NAT	39.7	60.3

The ROC curve and AUC of it were plotted and calculated. The curve was very close to the diagonal of the plot box, which indicates the poor performance of the model. In addition to that, the AUC value, 0.58, also further proves that the constructed model was not a great choice for cancer diagnosis.

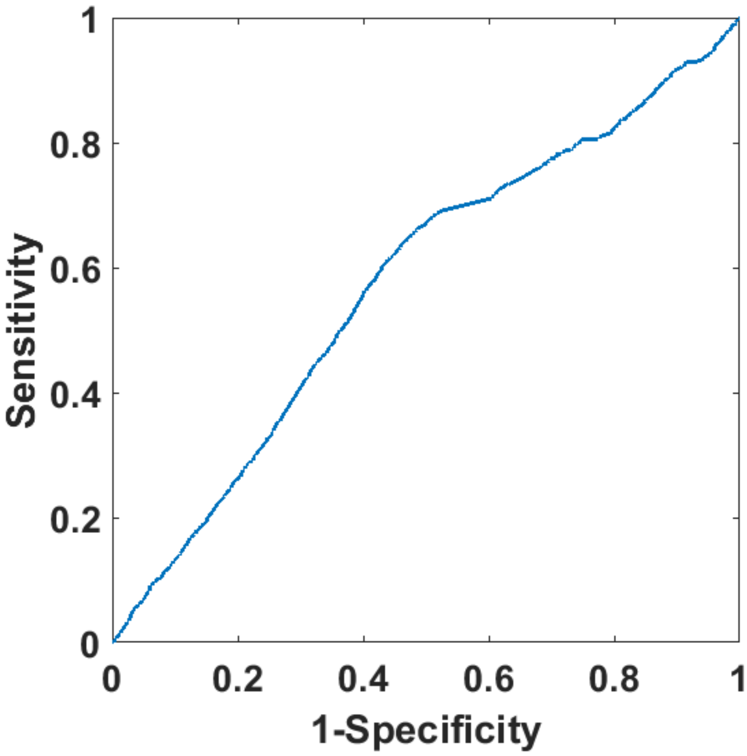


Figure 6.4 ROC curve of the cancer diagnosis model constructed using AdaBoost classification method with epithelium spectra

The importance of features used in the cancer diagnosis model was plotted and shown in Figure 6.5. Two main feature peaks were observed in the plot, 3296 (Amide A) and 3124 cm^{-1}. The main contribution to spectra separation was the

Amide A band, which was not the main feature for cancer diagnosis of samples on CaF_2. The main features for cancer diagnosis based on results obtained in CaF_2 were mainly focused on the range of 1500 to 1750 cm^{-1} (Figure 4.30). The poor performance could be raised by missing key cancer diagnosis features.

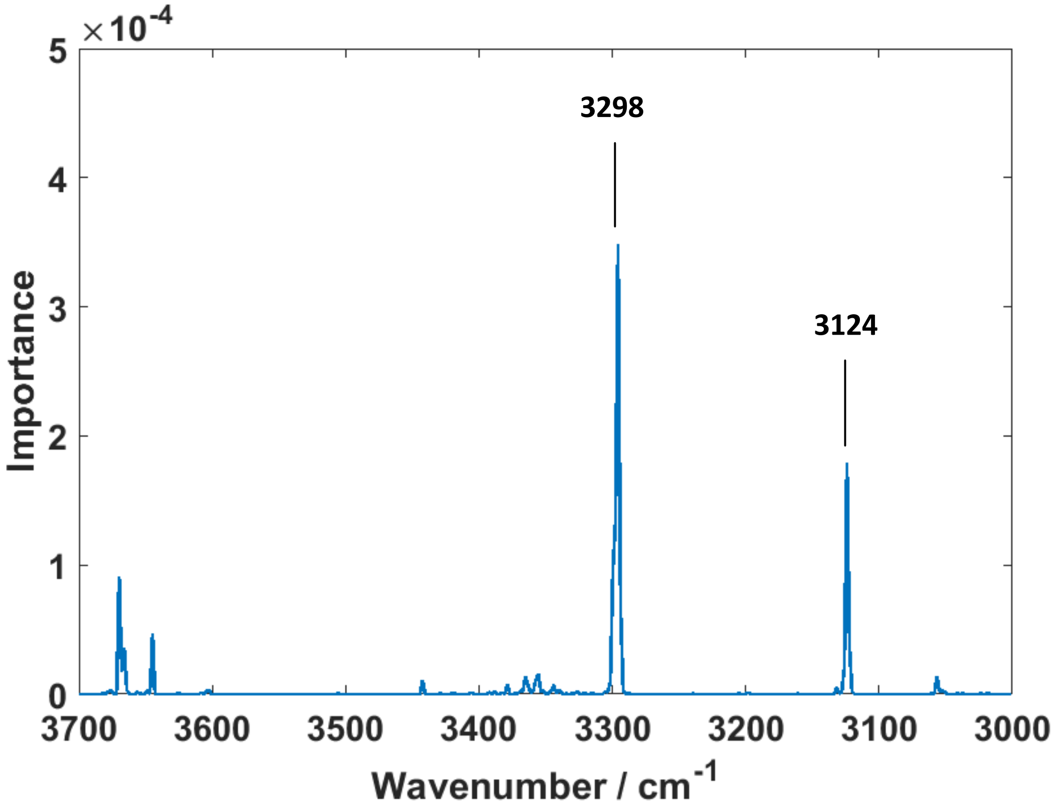


Figure 6.5 The importance of features of the cancer diagnosis model constructed using AdaBoost classification method with epithelium spectra

If a threshold was applied to the classification results of the constructed cancer diagnosis model, by only keeping half of the top-rated spectra (11986), the confusion matrix of these top-rated spectra was generated and shown in Table 6.5. The classification accuracy on cancerous spectra was slightly improved from 57.0% to 58.5% (by 1.5%), while the classification accuracy decreases by adding a threshold to the result. This observation shows that the model was not certain about the classification. The constructed model cannot find the real difference between cancer and NAT epithelium cells. The similarities found in two classes could be other similarities among patients.

Table 6.5 Confusion matrix of the independent test of the cancer diagnosis model constructed using AdaBoost classification method with 50% of the top-scored epithelium spectra

Independent Test (50%)		Predicted / %	
		Cancer	NAT
True / %	Cancer	58.5	41.5
	NAT	44.7	55.3

A higher threshold was applied to the results. Only the top-scored 10% spectra were kept, where 21574 spectra were defined as poorly classified and only 2397 spectra were kept. The confusion matrix after adding threshold was shown in Table 6.6. It could be seen that the model started to favour the cancerous class, where 62.2% of cancerous spectra were correctly classified while only 44.5% of NAT spectra were correctly classified. This observation further proves the previous assumption about the constructed model cannot find the real difference between cancer and NAT epithelium cells.

Table 6.6 Confusion matrix of the independent test of the cancer diagnosis model constructed using AdaBoost classification method with 10% of the top-scored epithelium spectra

Independent Test (10%)		Predicted / %	
		Cancer	NAT
True / %	Cancer	62.2	37.8
	NAT	55.5	44.5

6.3.2.1.2. Classification results produced using Random Forest classification method

A random forest model was constructed based on the epithelium spectra selected from the training cores, where 100 trees were applied (same with the study on CaF_2).

A subgroup of spectra was randomly selected from the training dataset, which consists of 20% of the total number of spectra. This subgroup was used for training validation test of the model. The obtained confusion matrix was calculated and presented in Table 6.7. The classification accuracies on cancerous and NAT class were 70.8% and 66.7% respectively. Comparing with the training validation test results of the previous model using AdaBoost, similar outputs were obtained. However, the model constructed using random forest worked slightly better on cancerous class, while the model constructed using AdaBoost focused more on the NAT class.

Table 6.7 Confusion matrix of training validation test of the cancer diagnosis model constructed using random forest classification method with epithelium spectra

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	70.8	29.2
	NAT	33.3	66.7

Independent test data was applied to the model. The obtained confusion matrix was shown in Table 6.8, where the classification accuracies on cancerous and NAT spectra were 66.9% and 43.3%. The model failed on the independent test as the classification accuracy on NAT was smaller than 50%.

Table 6.8 Confusion matrix of the independent test of the cancer diagnosis model constructed using random forest classification method with epithelium spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	66.9	33.1
	NAT	56.7	43.3

The ROC curve and AUC were plotted and calculated. Figure 6.6 indicates the ROC curve, which was close to the diagonal of the plotting box. This shows poor model performance. The AUC equals 0.53, which further indicates the poor performance of model as it was close to 0.5.

The feature importance plot of the random forest model was plotted and shown in Figure 6.7. The peaks mainly contributed to the spectra separation were 3298 (Amide A) and 3124 cm^{-1}, which were also found in the main feature in the importance plot of AdaBoost. Unlike the plot on AdaBoost, the feature importance plot using the random forest as classification method has many other minor peaks and fluctuations, which could be a possible reason for the failure in the independent test.

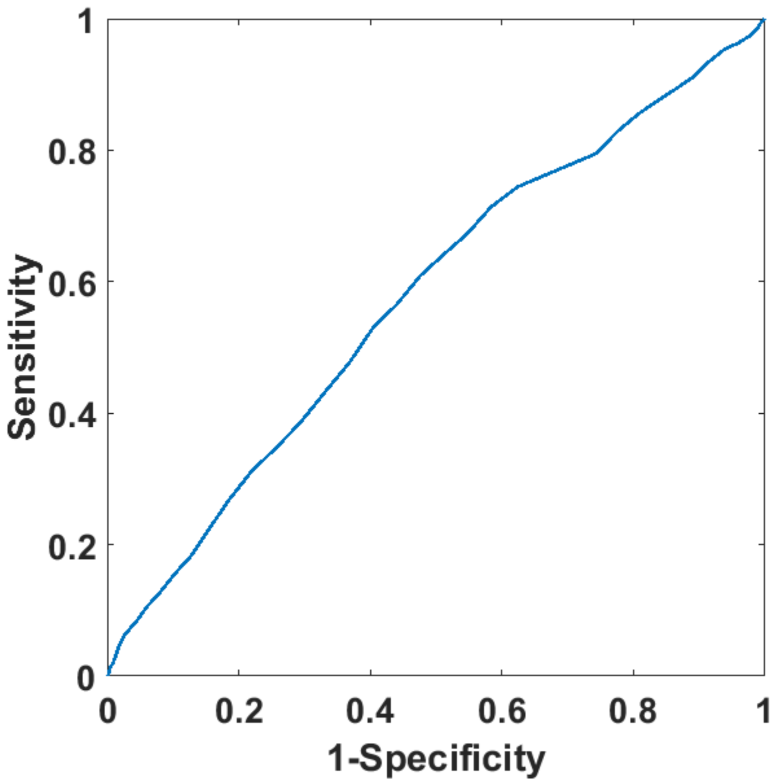


Figure 6.6 ROC curve of the cancer diagnosis model constructed using random forest classification method with epithelium spectra

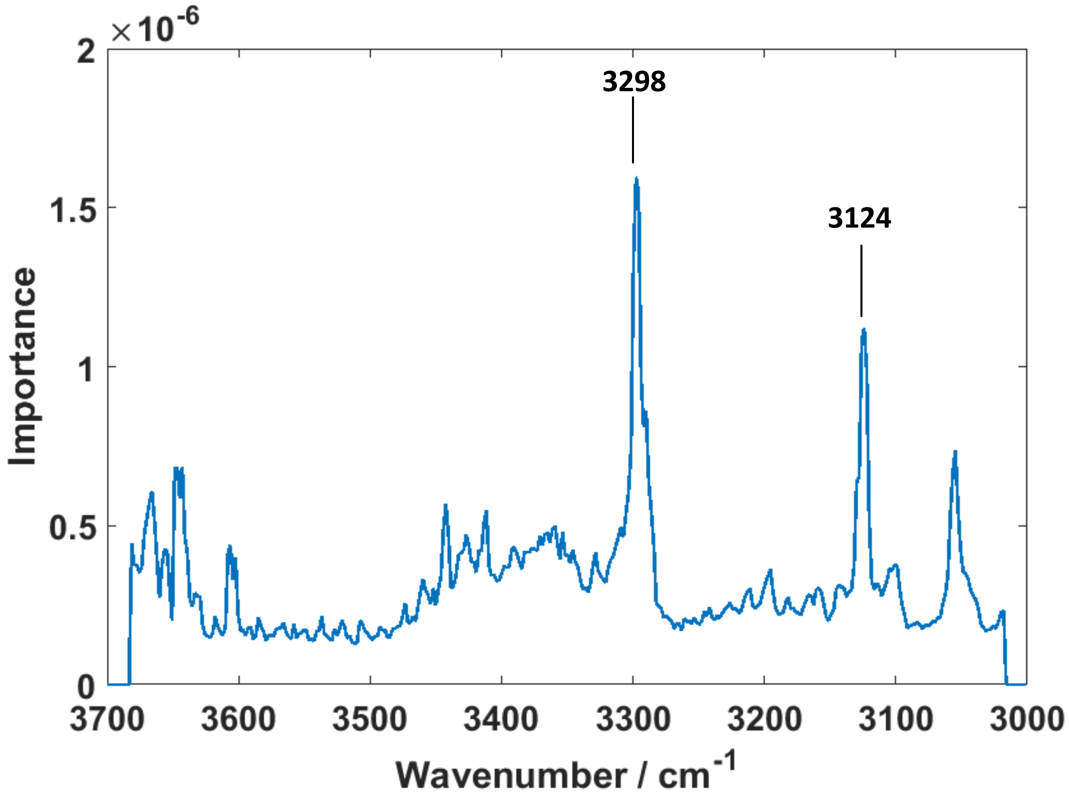


Figure 6.7 Feature importance plot of the cancer diagnosis model constructed using random forest classification method with epithelium spectra

A threshold of only keeping 50% top-scored spectra was applied to the results obtained in the independent test. 11074 spectra were defined as poorly classified and 12897 spectra were used to generate the confusion matrix shown in Table 6.9. The model favours the cancerous class even more after applying threshold, which means that even the confident predictions made by the model cannot tell the differences between cancerous and NAT spectra in the independent test.

Table 6.9 Confusion matrix of the independent test of the cancer diagnosis model constructed using random forest classification method with 50% top-scored epithelium spectra

Independent Test (50%)		Predicted / %	
		Cancer	NAT
True / %	Cancer	76.6	23.4
	NAT	61.0	39.0

An even higher threshold was applied, where only the top 10% of the spectra were kept. 20914 spectra were thrown as poorly classified samples. Only 3057 spectra were kept and a confusion matrix was calculated and showed in Table 6.10. The same trend observed previously was found, which further indicates that the model was not good enough for cancer diagnosis as it completely failed the independent test.

Table 6.10 Confusion matrix of the independent test of the cancer diagnosis model constructed using random forest classification method with 10% top-scored epithelium spectra

Independent Test (10%)		Predicted / %	
		Cancer	NAT
True / %	Cancer	86.3	13.7
	NAT	65.5	34.5

6.3.2.2. Classification results produced using spectra selected from stromal areas

Stroma spectra were selected from samples based on annotation and separated into two groups, training group and independent test group. The training group contained 130384 cancerous and 112443 NAT stroma spectra. Cancer and NAT classes were balanced by randomly selecting 112443 cancerous spectra to match the number of spectra in NAT class. The independent test dataset contains 17872 cancerous and 23112 NAT stroma spectra, which were exacted from the patient cores not related to the training cores.

6.3.2.2.1. Classification results produced using AdaBoost classification method

A cancer diagnosis model was constructed with the AdaBoost classification method. 500 trees were involved in the model training and the learning rate was set to 0.1, which was consistent with the model constructed using epithelium spectra.

A training validation test was performed using 20% randomly selected spectra from the training dataset. The training validation results were calculated and presented in the confusion matrix in Table 6.11. 78.0% cancerous and 61.9% NAT spectra were correctly classified by the constructed model using stroma spectra. Comparing with the training validation results obtained through the cancer diagnosis model constructed based on epithelium spectra, surprisingly higher classification accuracy on cancer class was observed, 8% lager. The difference in NAT class was less than

1%. Overall, similar performance was observed, which also was not great considering the results obtained in Chapter 4 on the training cores.

Table 6.11 Confusion matrix of training validation test of the cancer diagnosis model constructed using AdaBoost classification method with stroma spectra

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	78.0	22.0
	NAT	38.1	61.9

An independent test was conducted on the model, which predicts the labels of spectra exacted from the annotated area of the cores in the independent group. The predicted labels were compared with the actual label to calculate the classification accuracies for each class. The produced results were recorded and showed in Table 6.12, where the correct classification accuracy was only 2.3% for cancerous spectra. The model favours NAT class, as it tends to classify every spectrum as NAT spectrum.

Table 6.12 Confusion matrix of the independent test of the cancer diagnosis model constructed using AdaBoost classification method with stroma spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	2.3	97.7
	NAT	2.5	97.5

The ROC curve was plotted in Figure 6.8, where the curve fluctuated around the diagonal of the plotting box. The AUC value of the ROC curve was 0.52. Both the shape of the ROC curve and the value of AUC indicate the model acts poorly on the independent test set, which was consistent with the confusion matrix.

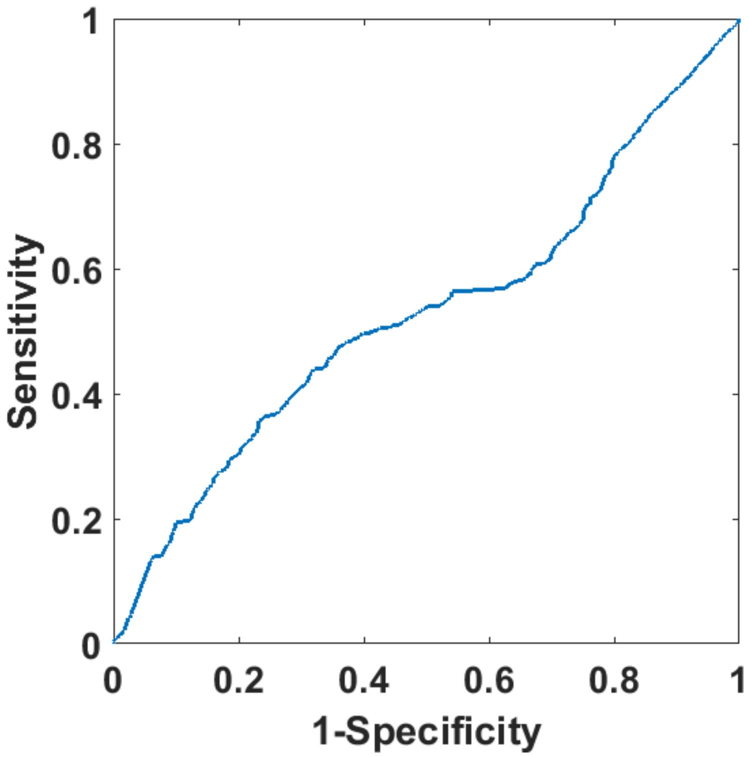


Figure 6.8 ROC curve of the cancer diagnosis model constructed using AdaBoost classification method with stroma spectra

Feature importance of the model was plotted in Figure 6.9. The major peaks were observed at 3020 cm^{-1}. In addition to that, some minor peaks with importance over 0.0001 were also found at 3041, 3066, 3321 (N-H band), and 3681cm^{-1}. Applying the same cores to train the model on CaF_2, 3037, 3087, and 3327 cm^{-1} (Figure 4.35) were observed as important features in the range of 3000 to 3550 cm^{-1}. These peaks were in similar positions with the obtained features in this case. However, the previous features found in the model on CaF_2 were not the main features responsible for the separation, which was mainly caused by major features in the lower wavenumber range, Amide I and II variations. Without features found from 1500 to 1750 cm^{-1}, with current level of data mining, classification cannot be successful with stroma spectra.

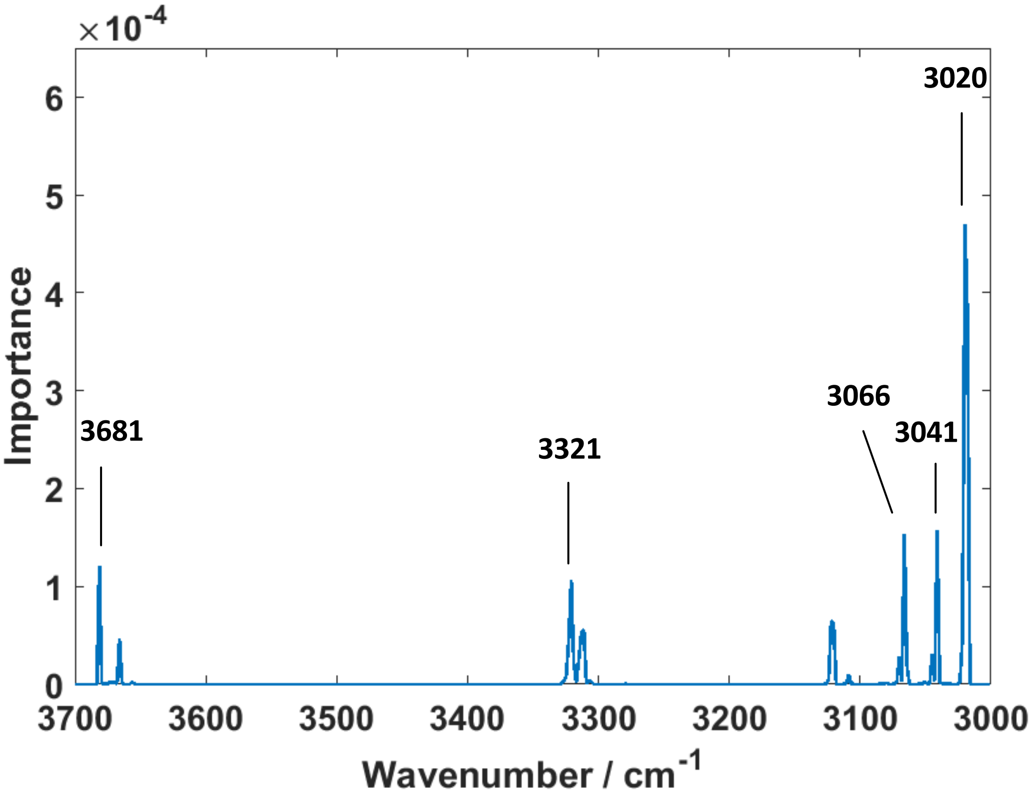


Figure 6.9 Feature importance plot of the cancer diagnosis model constructed using AdaBoost classification method with stroma spectra

A threshold was applied to the results of independent test spectra, where 20542 spectra were kept for prediction and 20442 spectra were thrown as poorly classified samples. With the threshold, the model predicts every left spectrum as NAT stroma spectrum. Even with confident predictions, the model cannot tell the differences between cancerous and NAT class in the independent test.

Table 6.13 Confusion matrix of the independent test of the cancer diagnosis model constructed using AdaBoost classification method with top 50% scored stroma spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	0	100
	NAT	0	100

Another even larger threshold was applied by only keeping 4099 spectra and defining the rest 36885 stroma spectra as poorly classified spectra. Both correct classification accuracies on cancer and NAT classes stays at 100%, which further indicates the complete failure of the model on the independent test.

Table 6.14 Confusion matrix of the independent test of the cancer diagnosis model constructed using AdaBoost classification method with top 10% scored stroma spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	0	100
	NAT	0	100

6.3.2.2.2. Classification results produced using Random Forest classification method

A random forest model was constructed with 100 trees which was consistent with the parameters used for model construction based on epithelium spectra.

A training validation test was performed and the confusion matrix of the classification was shown in Table 6.15, where 20% randomly selected stroma spectra from the training dataset was applied. The obtained classification accuracies were compared with the model constructed using epithelium spectra using random forest. Higher accuracies were observed in both classes, especially NAT class (increased by around 7%).

Table 6.15 Confusion matrix of training validation test of the cancer diagnosis model constructed using random forest classification method with stroma spectra

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	74.4	25.6
	NAT	26.6	73.4

An independent test was followed. The model favours the NAT class, as the correct classification accuracy on cancer class was only 25.0% while the false negative prediction of cancerous spectra was 75%. The same model preference was found in using AdaBoost and stroma spectra to establish cancer diagnosis model.

Table 6.16 Confusion matrix of the independent test of the cancer diagnosis model constructed using random forest classification method with stroma spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	25.0	75.0
	NAT	23.5	76.5

The ROC curve was plotted and shown in Figure 6.10, where the curve was very close to the diagonal of the box. The AUC value was 0.51. Both observations demonstrate the poor performance of the model.

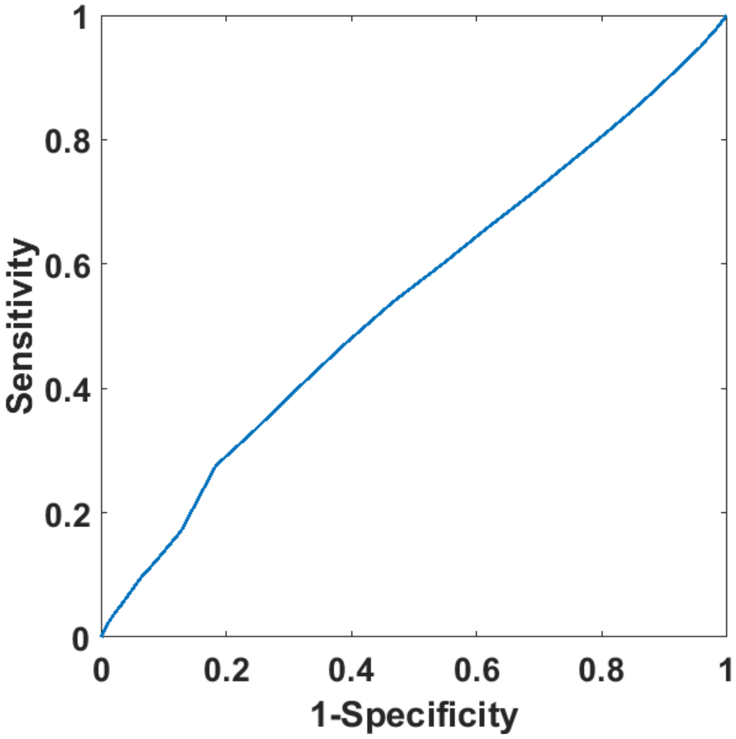


Figure 6.10 ROC curve of the cancer diagnosis model constructed using random forest classification method with stroma spectra

Feature importance of the model was plotted in Figure 6.11, where main peaks (importance over 2x10^{-7}) were found at 3018, 3041, 3070, 3120, and 3321 (N-H band) cm^{-1}. Among them two peaks, 3041and 3321 cm^{-1} were also found as important features in the model using AdaBoost. Other main peaks, for example 3018 and 3070 cm^{-1}, were also found in the model using AdaBoost however slightly shifted in positions.

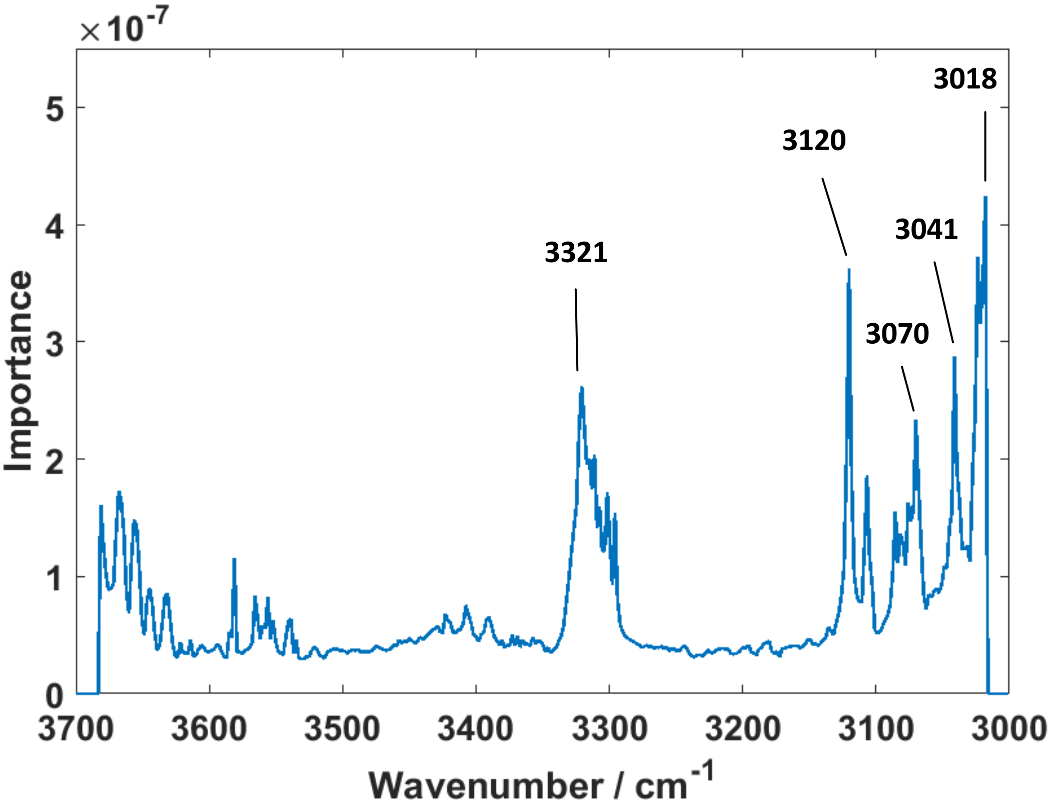


Figure 6.11 Feature importance plot of the cancer diagnosis model constructed using random forest classification method with stroma spectra

A threshold of only keeping 50% top-scored stroma spectra was applied to the results obtained in the independent test. 20442 spectra were defined as poorly classified and 20542 spectra were used as predictions to generate the confusion matrix shown in Table 6.17. The model favours the NAT class even more after applying the threshold.

Table 6.17 Confusion matrix of the independent test of the cancer diagnosis model constructed using AdaBoost classification method with top 50% scored stroma spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	11.5	88.5
	NAT	11.2	88.8

Another threshold of only keeping the top 10% stroma spectra in the independent test group was applied, where 36177 spectra were considered as poorly classified top-scored and only 4807 spectra were used to show the new confusion matrix,

Table 6.18. Almost every spectrum was classified as NAT spectrum. The model completely failed on the independent test.

Table 6.18 Confusion matrix of the independent test of the cancer diagnosis model constructed using AdaBoost classification method with top 10% scored stroma spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	5.4	94.6
	NAT	5.3	94.7

6.3.2.3. Sectional conclusion

Apart from the model constructed using epithelium spectra and the AdaBoost classification method, other models all fail in the independent test. The only model that survives the independent test barely identifies differences between cancerous and NAT epithelium spectra. Compared with the previously obtained classification accuracies on CaF_2, the performance of even the best models using H&E on glass needs to improve significantly. In terms of the features obtained during model construction, for epithelium spectra, in the range of 3000 to 3700 cm^{-1}, two main features, 3296 (Amide A) and 3124 cm^{-1}, were noticed in both models. Many minor fluctuations in terms of peak intensities were found in the importance plot of random forest, which could lead to failure in the independent test comparing with AdaBoost. Similar situations were observed in models based on stroma spectra, where main features in AdaBoost model were more defined and significant comparing with minor variations while the random forest model involves more voting depended on minor features. For models built on stroma spectra, major features were also occurred in similar wavenumbers with the features from previously constructed models using spectra on CaF_2. However, the previous separation in Chapter 4 was mainly caused by features in the lower wavenumber range, which could be the reason for the complete failure with both classification methods using spectra collected on glass slide.

6.3.3. Classification results produced using a different sampling method

To further investigate whether the previous failures were caused by inappropriate data processing and selection methods, a different data selection protocol and slight changes to the pre-processing configurations were applied.

In previous training data selection, equal numbers of spectra for both classes were randomly selected from 55 cores. In that case, each core could have a different contribution to the model construction as the size of the annotated area on each core was not the same. Some cores which have a larger annotated area could have a greater contribution. Their patient's similarities could be enhanced and considered as cancer-related features. To avoid this effect, a new sampling method of selecting the same number of spectra from each core was adopted. Thus there was equal opportunity for each patient to contribute to the model. To further simplify the question of cancer prediction, only grade II breast cancer patients were used. The difference between cancer and normal class would, therefore, be more focused on just difference between breast cancer and NAT tissue but not the biological information involved grade differences. The model in this section was constructed using AdaBoost trained on epithelium spectra, as this combination provides the best classification results in the independent test in the previous section.

The minimal number of pixel among all cores was selected to be the standard. Annotated pixels were randomly selected to match the number of the standard, which in this case were 82 pixels (each pixel stand for one spectrum, 82 spectra) per core. The total number of spectra used for training the model was 2706, including 2050 cancerous spectra and 656 NAT spectra. These spectra were picked from 25 cancerous cores and 8 NAT cores. For the independent test dataset, 77 pixels were randomly selected from the annotated areas of the cores. The total number of spectra used for the independent test were 456, including 342 cancerous spectra and 114 NAT spectra.

Principal component based noise reduction was applied with 80 principal components being retained. Spectra were quality tested based on the height of the amide A band. Spectra that had an absorbance between 0.07 and 1.2 were retained. The spectral range 3100 to 3600 cm^{-1} was used. Each spectrum was then vector normalized and converted to the first derivative using Savitzky−Golay smoothing with 19 points window size.

For machine learning algorithms, a training validation dataset was separated from the actual training dataset to test the established model. To make sure each core was actually involved in the model construction, 80% spectra were randomly selected from the training dataset provided by each core, and the rest of the spectra contributed by each core were used for training validation. The classification results of training validation test were presented in Table 6.19.

Table 6.19 Training validation confusion matrix of AdaBoost using a new sampling method to select epithelium spectra

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	99.5	0.5
	NAT	0	100

It could be seen that both classification rates boost from 70.0% and 62.7% (in Table 6.3) to 99.5% and 100% for cancerous and NAT class respectively. It means that the model had a nearly perfect performance on the validation data which contains similar biological information with the data used to construct the model.

Table 6.20 Independent test confusion matrix of AdaBoost using a new sampling method to select epithelium spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	81.9	18.1
	NAT	15.8	84.2

Further performance test of the model using the independent test dataset was performed and was shown in Table 6.20. 81.9% cancerous spectra and 84.2% NAT spectra were correctly classified. Comparing with the performance of previous epithelium cancer diagnosis model using AdaBoost nearly 25% increases were observed in both cancerous and NAT classes. Considering that using fine-needle biopsy to determine malignancy had an overall classification accuracy of 75.4% [17] based on the research conducted in 2010 by Kasraeian *et al*. . The classification accuracies in the independent test were considered excellent.

The ROC curve of the model was presented in Figure 6.12. The curve was close to the top left corner of the plotting box. The ROC curve was not smooth, which could be improved by adding more independent test spectra. To quantify the model performance in AUC value, 0.9 was obtained, which indicates good performance.

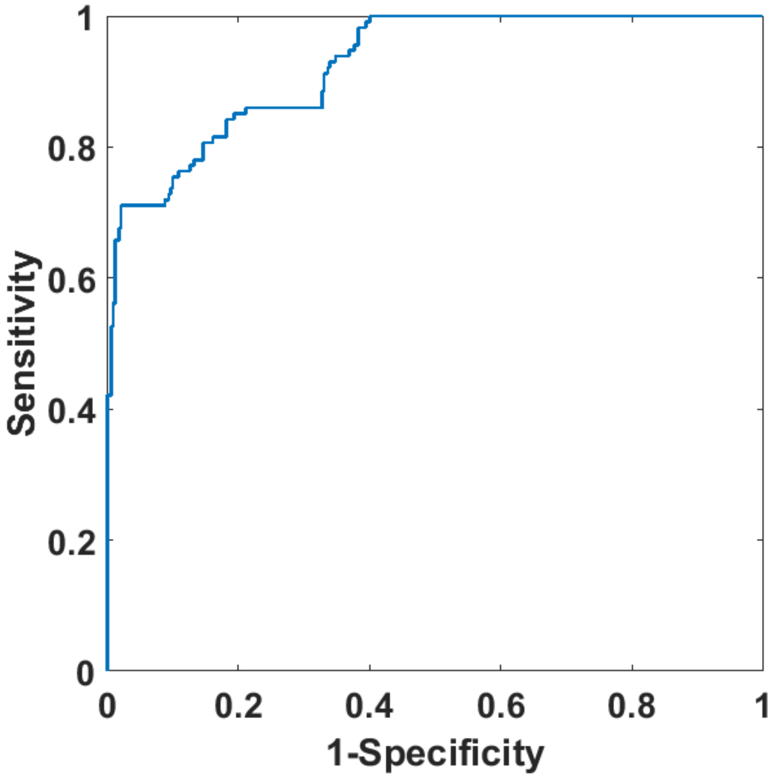


Figure 6.12 ROC curve of the model constructed based on AdaBoost using a new sampling method to select epithelium spectra

To find the key biological information used for cancer diagnosis, the feature importance plot of the model was constructed and shown in Figure 6.13. Three main peaks having intensity larger than 0.001 were labelled, which were 3118, 3300 (N-H stretching), and 3571 (O-H stretching) cm^{-1}. Comparing with the AdaBoost cancer diagnosis model constructed in the previous section, apart from 3571 cm^{-1} other major peaks were also found however with a different ratio. In the previous model, 3296 (Amide A) cm^{-1} has the highest intensity which was almost twice the intensity of importance at 3124 cm^{-1}. In this model, the highest peak was at 3118 cm^{-1} with importance intensity equals 0.0045, tripled the intensity of peak 3300 cm^{-1}. All these differences could lead to difference in the classification performance of the constructed model.

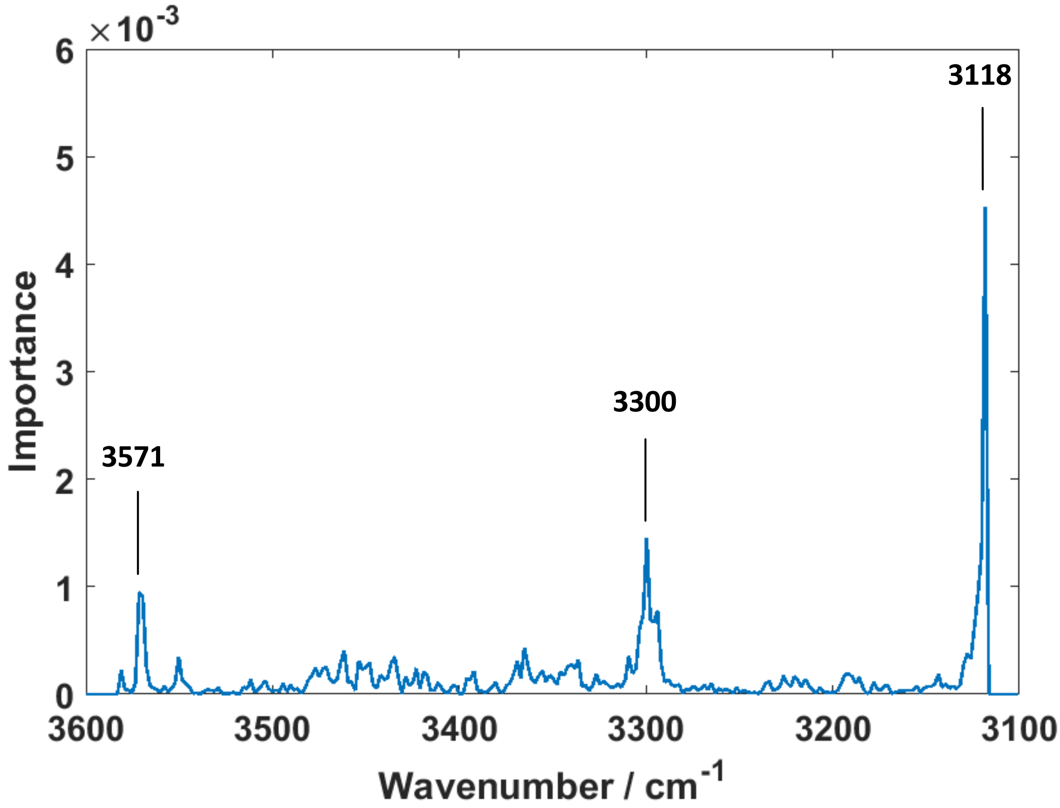


Figure 6.13 Feature importance plot of the model constructed based on AdaBoost using a new sampling method to select epithelium spectra

As pixels used for both model construction and independent test were randomly selected among a large number of spectra, the whole process, including model construction and independent test, was repeated five times. The independent test results were generated and were presented in Table 6.21. It could be seen that results on five corresponding models were stable, and the averaged classification accuracy on cancerous spectra was 81.3% and 83.2%.

Table 6.21 Independent test confusion matrix on the corresponding models with AdaBoost classification method using the new sampling method to select epithelium spectra and the averaged classification accuracy of independent test

Independent Test (Run one)		Predicted / %	
		Cancer	NAT
True / %	Cancer	81.9	18.1
	NAT	15.8	84.2
Independent Test (Run Two)		Predicted / %	
		Cancer	NAT
True / %	Cancer	85.4	14.6
	NAT	18.4	81.6
Independent Test (Run Three)		Predicted / %	
		Cancer	NAT
True / %	Cancer	67.8	32.2
	NAT	22.8	77.2
Independent Test (Run Four)		Predicted / %	
		Cancer	NAT
True / %	Cancer	87.7	12.3
	NAT	13.2	86.8
Independent Test (Run Five)		Predicted / %	
		Cancer	NAT
True / %	Cancer	83.6	16.4
	NAT	14.0	86.0
Average		Predicted / %	
		Cancer	NAT
True / %	Cancer	81.3	18.7
	NAT	16.8	83.2

To perform a traffic light colouring system on each independent core, a new histological model was established with the manner of extracting the same number of pixel from each core. To ensure the input accuracy of the cancer diagnosis model, only the top 10% scored epithelium spectra from each independent test core was passed to the next classification model. The classification accuracy of the new histological model was tested on the annotated pixels from the independent test with a threshold of only keeping the top 10% scored spectra, the classification accuracies were 100% for both cancerous and NAT class. The confusion matrix contained this information was shown in Table 6.22.

Table 6.22 Confusion matrix of the independent test with the top 10% scored spectra on the histological model between epithelium and stroma using AdaBoost classification method and epithelium spectra

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	100	0
	NAT	0	100

After selecting epithelium spectra using the histological model, these chosen spectra were passed to the cancer-NAT model. The top 10% scored spectra were used for a class assignment. If the number of spectra predicted as cancer was more than the number of spectra being diagnosed as NAT, the core would be predicted as cancerous core and coloured in red. If majority of spectra were predicted as NAT, then vice versa and coloured in green. The coloured core images of each core in the independent test set were presented in Figure 6.14. There was one exception, F6, which had only one epithelium spectrum scored in the top 10%. Although there was only one spectrum input, the cancer diagnosis model still predicts the core correctly. The overall classification accuracy in the unit of core was 100%. If F6 was removed as it does not fully complete the class assignment rules, the classification accuracy would be 87.5%, slightly higher than the averaged overall classification accuracy, 81.7%, in the unit of spectrum.

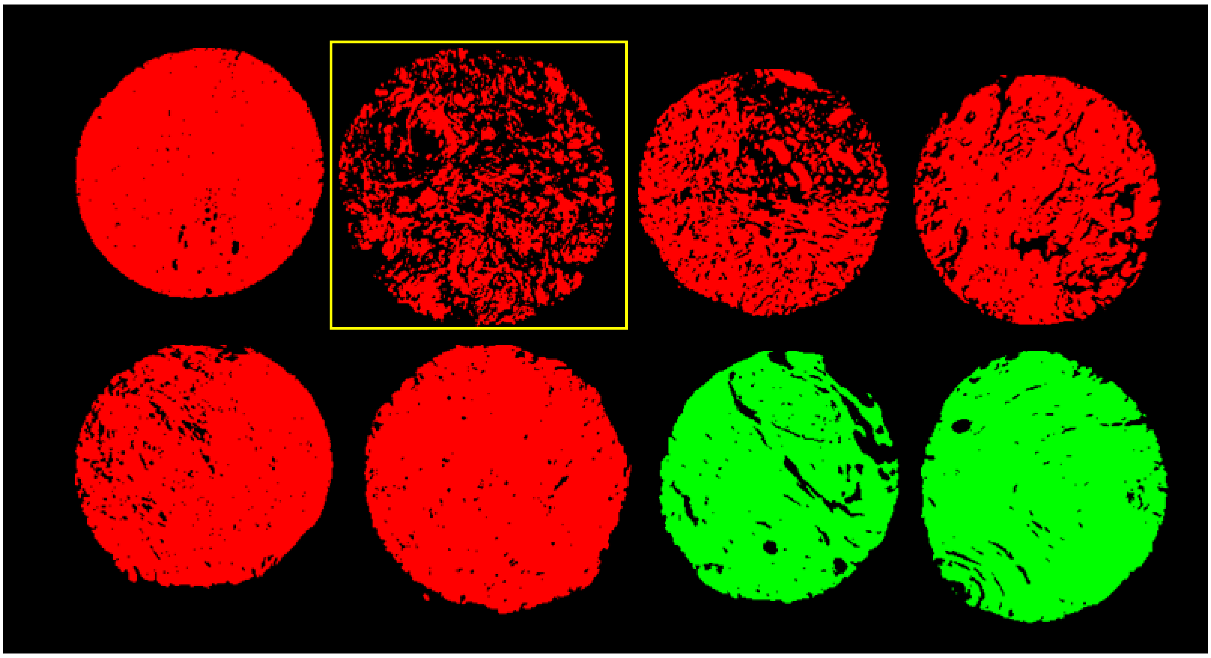


Figure 6.14 Traffic light coloured core images of the independent test set (red means cancerous core while green presents NAT core). The yellow box indicated F6 which only had 1 epithleium spectra used in classification

6.4. Conclusion and Future work

With careful pre-processing process and data selecting method, H&E stained breast tissue on glass could be used for cancer prediction, especially using epithelium spectra. However, this new model using a different sampling method did not consider all grades of breast cancer, as it only focused on finding the differences between the grade II breast cancer and NAT breast tissues. The main reason for only applying grade II breast cancer tissue was that the epithelium cells in these cores were easier to annotate and the number of annotated pixels was more consistent. For future studies, more annotated cores including all three grades of breast cancer (same number of cores in each grade) could be used to establish a more robust breast cancer prediction model.

6.5. References

[1] P.W. Johenning, A history of aspiration biopsy with special attention to prostate biopsy., Diagn. Cytopathol. (1988). doi:10.1002/dc.2840040318.

[2] L.B. Rorke, Pathologic diagnosis as the gold standard, Cancer. (1997). doi:10.1002/(SICI)1097-0142(19970215)79:4<665::AID-CNCR1>3.0.CO;2-D.

[3] National Careers Service, Pathologist | Explore careers, (n.d.). https://nationalcareers.service.gov.uk/job-profiles/pathologist (accessed August 19, 2019).

[4] C. Demir, B. Yener, Automated cancer diagnosis based on histopathological images: a systematic survey, Dept. Comput. Sci., Rensselaer Polytech. Inst., Troy, NY, USA, Tech. Rep. (2005). doi:10.1.1.61.1199.

[5] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA. Cancer J. Clin. (2018). doi:10.3322/caac.21492.

[6] E. Gazi, M. Baker, J. Dwyer, N.P. Lockyer, P. Gardner, J.H. Shanks, R.S. Reeve, C.A. Hart, N.W. Clarke, M.D. Brown, A Correlation of FTIR Spectra Derived from Prostate Cancer Biopsies with Gleason Grade and Tumour Stage, Eur. Urol. (2006). doi:10.1016/j.eururo.2006.03.031.

[7] M.J. Baker, E. Gazi, M.D. Brown, J.H. Shanks, N.W. Clarke, P. Gardner, Investigating FTIR based histopathology for the diagnosis of prostate cancer, J. Biophotonics. (2009). doi:10.1002/jbio.200810062.

[8] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, Analyst. 142 (2017) 1258–1268. doi:10.1039/c6an02224c.

[9] X. Mu, M. Kon, A. Ergin, S. Remiszewski, A. Akalin, C.M. Thompson, M. Diem, Statistical analysis of a lung cancer spectral histopathology (SHP) data set, Analyst. (2015). doi:10.1039/c4an01832j.

[10] C. Kuepper, F. Großerueschkamp, A. Kallenbach-thieltges, A. Mosig, Label-free classification of colon cancer grading using infrared spectral histopathology, Faraday Discuss. 187 (2016) 105–118. doi:10.1039/C5FD00157A.

[11] H. Fabian, N.A.N. Thi, M. Eiden, P. Lasch, J. Schmitt, D. Naumann, Diagnosing benign and malignant lesions in breast tissue sections by using IR-

microspectroscopy, Biochim. Biophys. Acta - Biomembr. (2006). doi:10.1016/j.bbamem.2006.05.015.

[12] D.M. Mayerich, M. Walsh, A. Kadjacsy-Balla, S. Mittal, R. Bhargava, Breast histopathology using random decision forests-based classification of infrared spectroscopic imaging data, Proc. SPIE. 9041 (2014) 904107. doi:10.1117/12.2043783.

[13] P. Bassan, M.J. Weida, J. Rowlette, P. Gardner, Large scale infrared imaging of tissue micro arrays (TMAs) using a tunable Quantum Cascade Laser (QCL) based microscope, Analyst. (2014). doi:10.1039/c4an00638k.

[14] Sunil R. Lakhani; Ian O. Ellis;, S.J.P.H.T. Schnitt, V.M.J. van De, WHO Classification of Tumours of the breast, 2012. doi:10.1017/CBO9781107415324.004.

[15] A. Cutler, D.R. Cutler, J.R. Stevens, Random Forests, (2012). doi:10.1007/978-1-4419-9326-7.

[16] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, RUSBoost : A Hybrid Approach to Alleviating Class Imbalance, 40 (2010) 185–197.

[17] S. Kasraeian, D.C. Allison, E.R. Ahlmann, A.N. Fedenko, L.R. Menendez, A comparison of fine-needle aspiration, core biopsy, and surgical biopsy in the diagnosis of extremity soft tissue masses, in: Clin. Orthop. Relat. Res., 2010. doi:10.1007/s11999-010-1401-x.

Chapter 7

Unbalanced Data Studies with Different Classification Methods

7. 1. Chapter Overview

Machine learning is known as a branch of artificial intelligence. It allows a computer to learn from data and improve with experience. It refines a model that can be used to predict outcomes of inquiry-based on previous learning. There are two types of machine learning, which can be categorised as either supervised or unsupervised learning. For supervised learning, a labelled set of input-output pairs is given up to learning mapping from inputs to outputs. The second type is unsupervised learning, which is more data mining than actual learning. Its objective is to find hidden patterns in data, which is also called knowledge discovery [1]. There are no defined answers, as there is no specific existing pattern to find [1,2].

Machine learning has become increasingly popular and more matured in the last decade. It has grown into a powerful tool in many different fields. Research has been done combining both FTIR and random forest, which is a supervised machine learning technique based on ensemble decision trees. A random forest grows many classification trees as required. Once an input data is put into the forest, each tree has its classification result. All trees in the forest are involved in voting the most popular class as the final result from the forest [3]. Random forest is considered to be an effective approach when it comes to spectral analysis. Pilling et al. showed that random forest can give high accuracy on a glass substrate with classification accuracies over 95% [4].

Unlike random forest, not many people have tried to combine another excellent decision tree ensemble method, boosting, with spectral analysis. Boosting can improve the performance of weak learners regardless of whether training data is balanced or imbalanced [5]. The most famous and commonly used boosting method is Adaptive Boosting (AdaBoost) [6]. The AdaBoost algorithm of Freund and Schapire [7] was published in 1997. It is the first practical boosting algorithm, and remains one of the most widely used and studied, with applications in numerous fields [8]. AdaBoost works by obtaining weighted majority votes of the weak hypothesis where each hypothesis is assigned weight and conducted by every weak learner [8]. To be more specific, during each iteration, weights of incorrectly classified samples are modified with the aim of correctly classifying them in the next iteration. All trees are involved in weighted voting to allocate unknown samples. It is considered more effective at handling imbalanced datasets than random forest, as the minority class which is much easier to be miss-classified can be given higher weights in subsequent iterations [5].

Conventional statistical approaches are compared with machine learning technique on the imbalanced datasets. Principal component analysis (PCA) is one of the most commonly and widely used in exploratory data analysis and predictive modelling. It is defined as a statistical procedure concerned with explicating the covariance structure of datasets [9]. PCA develops artificial variables, the principal components, which are variables explained variances of datasets [10]. PCA can be related to canonical variate analysis (CVA), which studies linear relationships between two vector variates [11]. One multivariate discrimination approach using principal components analysis to reduce the dimensionality and then performing canonical variates analysis is known as principal component canonical variates analysis (PC-CVA) [12]. PC-CVA is recommended rather than merely PCA, as this two-stage approach is considered to bring improvements on model stability compared with than just PCA and canonical variates extracting more detailed information than just principal components [13].

Another popular conventional approach for pattern recognition and classification is partial least squares discriminant analysis (PLS-DA), which can be regarded as a linear classifier with the objective of using a straight line dividing data into two or more regions. For a two-class problem, PLS1 is adopted. Samples are assigned into two classes labelled +1 or −1. Unclassified samples are allocated into different classes based on estimated numerical labels [14]. PLS-DA has several advantages, including widespread availability and insights of variables through weights and loadings, comparing with traditional statistical methods (linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA)) [15].

Both PC-CVA and PLS-DA are excellent tools when analyzing balanced datasets, but when applied to imbalanced data sets, conventional methods are not always a good choice, as imbalanced classes are not taken into account [15,16]. Unfortunately, medically related datasets are often unbalanced due to the nature of diseases. Normal or high-risk samples are usually less likely to be obtained, while samples of intermediate disease stages consist the majority of the overall dataset. Creating an effective classification model can be challenging, as imbalanced training data can cause bias [16]. When number of samples in one class largely exceeded number of samples in the other class, traditional data mining algorithms tend to favour majority class by allocating samples to it. The minority class, which is frequently the positive class (generally the class of interest), can have poor classification accuracy due to the biased model. Therefore, techniques are required to ensure that a model can efficiently identify these both important and rarely occurring samples. Different methods can be conducted to solve this, including different sampling methods [17]. There are two commonly used sampling methods, including under-sampling and over-sampling. Class distribution can be balanced by either duplicating selected members of the minority class (over-sampling) or removing selected examples from the majority class (under-sampling) [16]. Under-sampling and over-sampling can be performed in different ways [5]. The easiest method is randomly resampling datasets. Random under-sampling balances two classes by randomly removing extra data from the majority class to match the number of samples in minority class, while random over-sampling replicate samples in minority class until the number of samples match the desired amount. These are methods applied in this chapter.

In this chapter, the possibility of applying machine learning techniques (random forest and AdaBoost) into infrared data analysis on human breast cancer will be discussed. Classical statistical approaches will be compared with random forest and AdaBoost. Also, the application of different classification methods combining with class-balancing sampling methods will be mentioned as well.

7.2. Methodology

7.2.1. Breast tissue preparation

A formalin-fixed paraffin embedded breast TMA slide, ID BR20832, was used for this study. It was purchased from US Biomax, Rockville, MD. The TMA contains 15 pathological indicated non-malignant and 192 malignant cores, in total 207 breast tissue biopsy cores. Each core is from a different patient. A 5 µm thickness and 1 mm diameter section was floated onto a standard histology glass slide and H&E stained. An adjacent section was floated onto a BaF_2 slide without dewaxing, which reduces the possibility of having chemical changes of samples during de-paraffinization and decreases spectral scattering as matching refractive index between paraffin and sample. [18]

To construct high variance models 70 cores were selected from the sample set, which included 57 cores with grade I, II or III of breast cancer and 13 normal associate breast tissue cores.

7.2.2. Infrared Chemical Imaging instrumentation and experimental method

The instrument and microscope used in the work reported here were those that have previously detailed in chapter 4.

Before imaging, background scans were taken as a single tile with 128 co-added scans at a spectral resolution 5 cm^{-1}. The area taken was selected to be clean and paraffin free. 96 co-added sample scans were measured on each core from TMA. Interferograms were processed into absorption spectra using Happ-Genzel apodisation with a region between 900 and 3800 cm^{-1}.

7.2.3. Data preprocessing procedures

All data were pre-processed with MATLAB 2017a. Infrared spectra for each biopsy core were extracted from the mosaic as a 256 × 256 × 1478 datacube. Each datacube consists of 65536 spectra which contain 1478 data points.

Chemical images of each of the breast tissue cores were generated and compared to the H&E stained sections, and regions of epithelium were identified and annotated on chemical images generated based on WHO Classification of Tumours in the Breast[19]. Principal component based noise reduction was used to improve the signal-to-noise ratio of raw spectra from annotated area. The first 80 principal components were kept. Spectra were quality tested to remove data obtained from areas with little or no tissue based on the height of the amide I band. Spectra have absorbance between 0.1 and 2 were retained. Spectra ranges 1000 to 1300 cm$^{−1}$, 1500 to 1750 cm$^{−1}$, and 3005 to 3550 cm$^{−1}$ were taken. Region describing the absorption bands of wax was removed. Each spectrum was then vector normalized to correct for different thicknesses of breast tissue. Finally spectra were converted to first derivative and performed a Savitzky−Golay smoothing using a window size of 19 data points. The dataset was separated into two groups, a training set and independent test set. The training set contained 44 cancerous and 11 normal associated cores giving in total 55 cores, while independent test set contained 13 cancerous and 2 normal associated cores giving in total 15 cores. Each core was from a different patient. Information from 55 patients was involved in training process and spectra from 15 different patients were used for independent test.

7.2.3.1. Spectra sampling methods

Both under-sampling and over-sampling were applied to subgroups of the dataset, which contained a total number of 10000 spectra with different cancer-to-normal ratios from a data pool containing 70932 cancerous spectra and 19338 non-cancerous spectra extracted from annotated epithelium area of 55 cores. Ratios between cancerous and normal spectra of 9:1, 8:2, 7:3, 6:4 and 5:5, with a fixed total number of spectra, were generated based on a randomly selected index number list. Firstly, a 5:5 cancer-to-normal ratio was randomly generated, that contains 5000 cancerous and 5000 normal spectra. Subgroups with other proportions were generated based on it. A 6:4 subgroup contains 6000 cancerous spectra, which contained the first selected 5000 cancerous spectra from the previous group and another randomly selected 1000 cancerous spectra from data pool, and 4000 normal spectra which randomly choose from previous selected 5000 spectra. This method was applied to all 4 unbalanced subgroups, as it minimised the performance differences of unbalanced classes with different ratios caused by spectral information of randomly selected groups. Keeping as many as possible the same selected spectra in each subgroup minimised the effects of biological information differences causing variations of classification performance.

To balance training sets of cancerous and non-cancerous classes, two sampling methods were conducted, under-sampling and over-sampling. In terms of under-sampling, for each subgroup, spectra from the cancerous class were randomly selected and removed to match the number of spectra in the normal class. To over-sample each subgroup, spectra from normal class were randomly duplicated to match the number of spectra in cancerous class.

7.2.3.2. Classification methods

Four classification methods were used to show if there were differences among unbalanced classes with different cancer-to-normal ratios. Commonly used parameters were applied with each classification method, as the main focus of this chapter is to show the influences of unbalanced dataset on each method. All methods were trained on either balanced or unbalanced training datasets and tested on the same independent test dataset.

7.2.3.2.1. Random forest

A random forest algorithm (available from http://code.google.com/p/randomforest-matlab/) was used. 500 trees were used to train the classifier. The number of variables to split on at each node was the square root of number of features from variables, which was the number of wavenumber intervals, 570, after wavenumber range selection. The minimum node size to split was default to one.

7.2.3.2.2. AdaBoost

An AdaBoost model was constructed using MatLab 2017 built-in Statistical and Machine Learning Toolbox. AdaBoost M1 was used as an only two-class problem was studied. 500 iterations were applied with learning rate (to train an ensemble using shrinkage) equalled to 0.1.

7.2.3.2.3. PCCVA

A PCCVA model was constructed using MatLab 2017 (available from https://bitbucket.org/AlexHenderson/chitoolbox/src/master/ChiToolbox/). Principal components explained 95% of the dataset were used for analysis. PCA was conducted first and CVA was performed based on PC scores obtained previously. Independent dataset was projected to the model constructed based on a training dataset. A Mahalanobis distance was applied to measure the distance from projected independent test sample to training CV score distribution of both classes.

7.2.3.2.4. PLSDA

The PLDSA model was developed using cluster toolbox V2.0 (available from https://github.com/Biospec/cluster-toolbox-v2.0). There are numerical PLSDA and its enhanced algorithms. The algorithm used followed the paper published by Brereton et al [15], where PLS1 was used, which designed for two-class problems. In addition to that, 10-fold cross-validation was conducted to avoid over-fitting and the optimum principal component was used for model construction. As unbalanced training data sets were applied, according to Brereton et al [15], data centre was modified by subtracting the average of the means of the two unbalanced groups. It helped adopt unbalanced training sets better, as the new data centre considered the boundary was shifted towards the larger group which caused misclassification. Independent dataset was projected to the model constructed based on training dataset.

7.3. Results and discussion

7.3.1. Classification results produced using unbalanced training data

Applying unbalanced training data sets to construct classification models can lead to misclassification of the minor class as the classification model usually favours the dominant class. It is interesting to see if unbalance training data will cause a large influence on the classification results. To show the extent of problem unbalanced training data sets because when applying classification, several questions are proposed and addressed in this section. Namely (i) Will all four chosen classification methods be influenced by unbalanced training data? (ii) Will the same unbalanced data set lead to the same effects on all classification methods? (iii) Are there any methods significantly influenced to a greater or lesser degree influenced by unbalanced training data compared with than others? (iv) Will different unbalanced ratios have the same effect on classification results? (v) Is there a trend of influence with different unbalanced ratios?

To answer these questions, experiments were conducted using four classification methods which were applied to the same groups of unbalanced training data. Two of them were statistical methods, which were commonly used in the field. Two of them were machine learning methods which were less commonly used. Four unbalanced ratios were used in this case, including 6:4 (6000 cancerous spectra vs. 4000 non-cancerous spectra), 7:3 (7000 cancerous spectra vs. 3000 non-cancerous spectra), 8:2 (8000 cancerous spectra vs. 2000 non-cancerous spectra) and 9:1 (9000 cancerous spectra vs. 1000 non-cancerous spectra). The 5:5 (5000 cancerous spectra vs. 5000 non-cancerous spectra) subgroup was used as a perfect-world standard to access the extent of bias of each unbalanced model with different condition combinations.

7.3.1.1. Statistical approaches

Starting with statistical classification methods, PLS-DA and PCCVA, the same independent test dataset was applied to both models using different unbalanced training ratios. For reproducibility of results, each model construction process was repeated three times with randomly selected groups of training data and every time followed by an independent test consisting of the same group of spectra. All independent test results were plotted in Figure 7.1.

Classification accuracy of cancerous class for both methods was plotted in red colour while the NAT class was shown in green. All three attempts of the independent test were indicated as scatters and the averaged correctly classification accuracies were presented in lines. The averaged independent test classification accuracy of correctly separate cancerous spectra was plotted in dotted lines while for non-cancerous spectra was plotted in solid lines. Results of three independent test attempts were plotted in scattering points.

It could be seen that both methods were largely influenced by unbalanced ratios. With the increase of cancerous spectra, for PLSDA, the classification accuracy grew dramatically. From an average of 59% at ratio 6:4, the classification accuracy increased to 82% at ratio 9:1. With the decrease of non-cancerous spectra, the classification accuracy reduced largely as well. From average accuracy 95% at ratio 6:4, the classification accuracy decreased to 39%. It could be seen that although unbalanced data had been mean centred based on subtracting the average of the means of the two groups, still classification accuracy was influenced. If unbalanced datasets were applied to original PLSDA method without the new mean centring method, applying the same independent test data, the averaged classification accuracy of cancerous spectra was pulled from 70% (6:4) to 98% (9:1) while the averaged classification accuracy for non-cancerous spectra dropped from 78% (6:4) to 13% (9:1). The mean centring method proposed by Brereton et al [15] reduces to the influences of unbalanced ratios to PLSDA model. However, even after mean centring correction, with large unbalanced ratio, the model was still largely influenced.

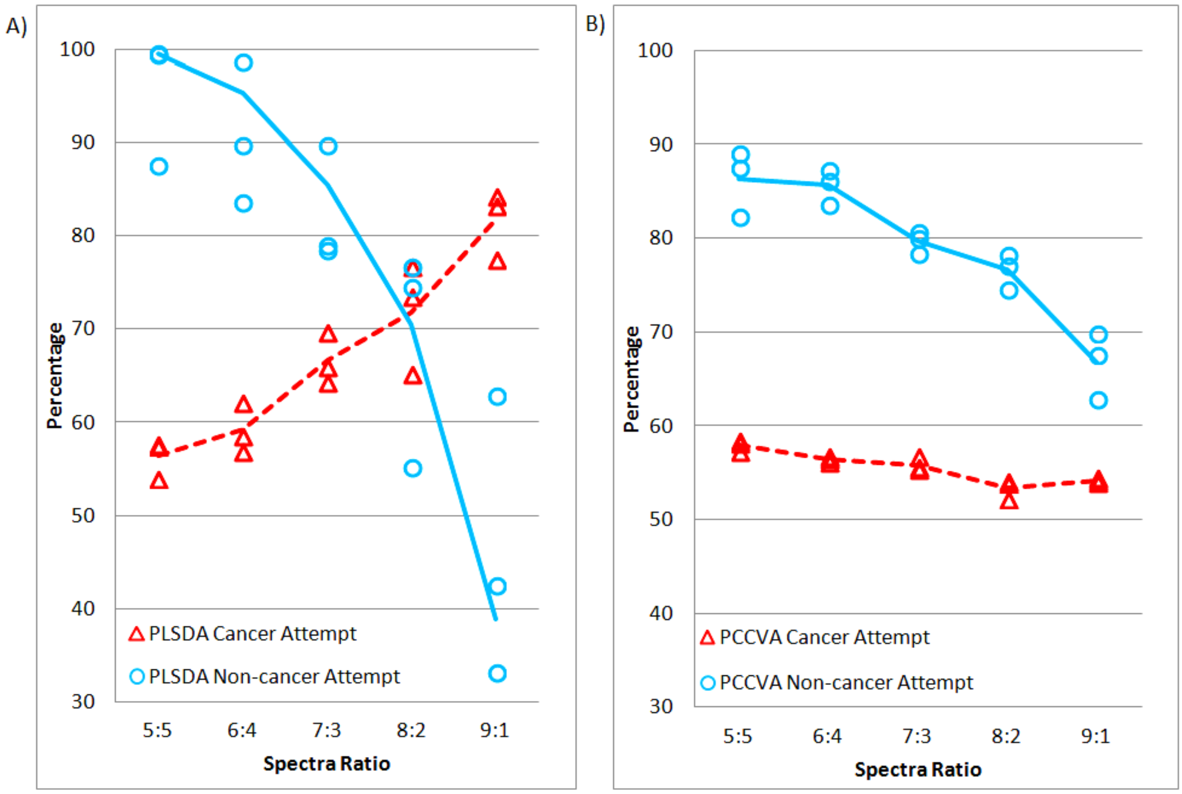


Figure 7.1 Classification accuracies using A) PLSDA B) PCCVA with unbalanced training datasets with different ratios, where spectra ratio stands for the ratio between the number of spectra in cancer class and NAT class. Average classification accuracies of NAT class was plotted in blue solid line while cancer class was plotted in dotted red line. Classification results conducted using both PLSDA and PCCVA were highly influenced by the unbalanced ratios between cancer and non-cancer classes. With PLSDA, more spectra led to higher classification rates and fewer spectra led to lower rates. With PCCVA, increase number of cancerous spectra did not lead to increase of classification accuracy which showed the model did not benefit from big data, but decease of the number of sample did lead to the reduction of classification accuracies.

For PCCVA, with the increase of cancerous spectra, the classification accuracy remained at approximately the same the level, from 56% (6000 cancerous spectra) to 54% (9000 cancerous spectra). For non-cancerous class, with the decrease of non-cancerous spectra, the classification accuracy reduces, dropping from 86% (with 4000 non-cancerous spectra) to 68% (with 1000 non-cancerous spectra).

Comparing these two methods, although they both largely biased by the major class, PCCVA was slightly more stable than PLSDA, as it had fluctuations during changing training datasets with different unbalanced ratios. One possible reason could be that the nature of PLSDA relies more on data centre and not every centroid was of equal significance [15].

7.3.1.2. Machine Learning

The same independent test dataset was applied to both AdaBoost and random forest using different unbalanced training ratios. For reproducibility of results, independent tests were repeated three times, and all independent test results were plotted in Figure 7.2. Classification accuracies for cancerous class with all ratios were plotted in red colour, while for the NAT class were plotted in green colour. The averaged independent test classification accuracy of correctly separate cancerous spectra was plotted in dotted lines while for non-cancerous spectra was plotted in solid lines. Results of three independent test attempts were plotted in scatters.

It could be seen that both methods were influenced by unbalanced ratios. With the increase of cancerous spectra, for AdaBoost, the classification accuracy of cancerous spectra increased but not too aggressively, compared with two statistical classification methods. From an average of 83% at ratio 6:4, the classification accuracy increased to 92% at ratio 9:1. With the decrease of non-cancerous spectra, the classification accuracy decreased. From average accuracy 96% at ratio 6:4, the classification accuracy decreased to 86%. For both classes the influences of unbalanced ratios had remained within the range of 10%.

For random forest, with the increase of cancerous spectra, the classification accuracy increased as well, from 81% (with 6000 cancerous spectra) to 94% (with 9000 cancerous spectra). For non-cancerous class, with the decrease of non-cancerous spectra, the classification accuracy decreased largely, dropping from 93% (with 4000 non-cancerous spectra) to 50% (with 1000 non-cancerous spectra).

Comparing two machine learning methods' performance on unbalanced datasets, they both adopted unbalanced training data better than statistical approaches. Both methods had similar fluctuations in cancerous class with the increase of ratios. However, random forest still failed to separate non-cancerous class, while AdaBoost still performed relatively well when 9:1 unbalanced ratio was applied. This could be raised by the nature of AdaBoost algorithm, which could give minority class higher weights in the next iteration [5].

It was clear that unbalanced training datasets caused an increase of classification accuracy of majority class and decrease of classification accuracy of minority class. The extent of increase and decrease was influenced by unbalanced ratios. The larger the ratio, the larger the difference would be made. From results mentioned previously, all four classification methods were influenced by unbalanced training data with different extent. Comparing four methods, AdaBoost was considered to have the best performance on unbalanced training datasets. It gave the smallest fluctuation when different unbalanced ratios were applied. It could because of the mechanism of AdaBoost. The minority class, non-cancer class, which was much easier to be miss-classified could be given higher weights in each iteration, as it corrected weights for miss-classified samples [5].

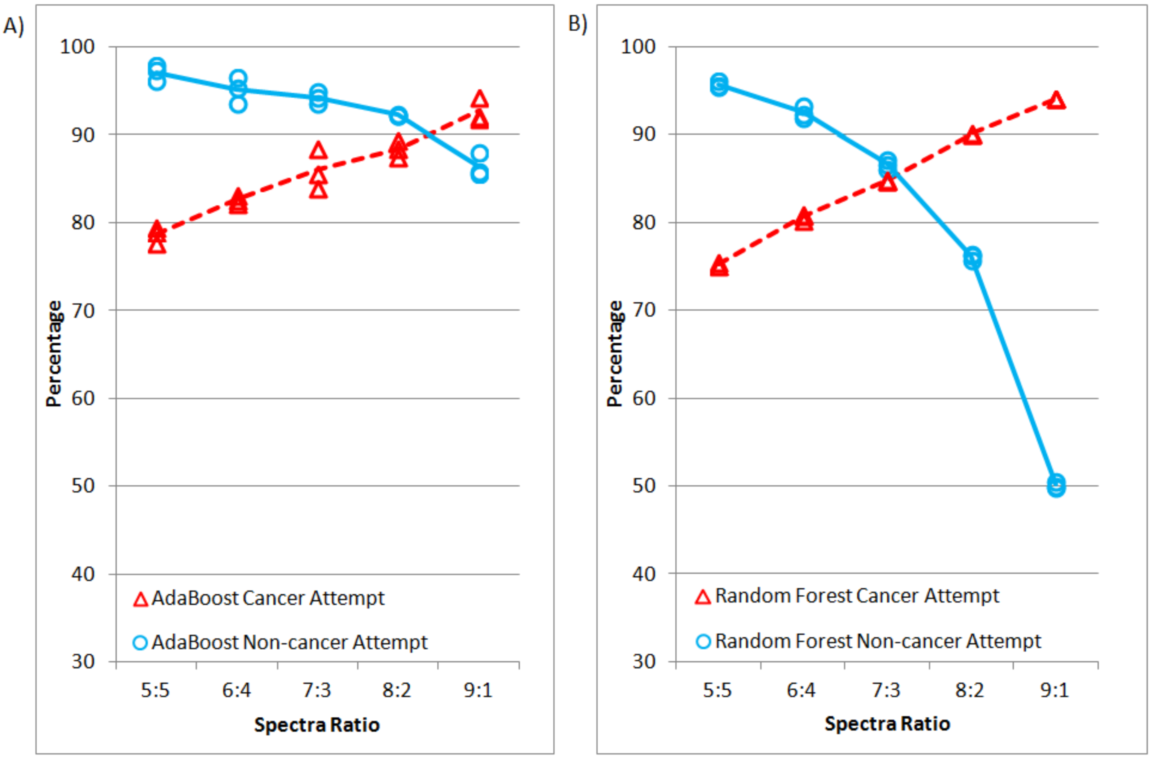


Figure 7.2 Classification accuracies using A) AdaBoost B) Random Forest with different unbalanced training datasets, where spectra ratio stands for the ratio between the number of spectra in cancer class and NAT class. Average classification accuracies of NAT class was plotted in blue solid line while cancer class was plotted in dotted red line

7.3.2. Classification results produced using balancing classes

To solve the unbalanced training data problem, one straight forward method was applied, namely sampling. There were two sampling types, under-sampling and over-sampling. To balance training sets of cancerous and non-cancerous classes, both sampling methods were conducted. They were both applied to groups of unbalanced training data with different unbalanced ratios. After sampling, balanced training groups were applied to four classification methods (PLS-DA, PCCVA, AdaBoost and random forest) to investigate performances of two sampling methods pairing with different classifiers. The same independent test data set was applied to all models to investigate responses of different classification methods to both sampling methods.

7.3.2.1. Statistical approaches

Starting with statistical classification methods, PLS-DA and PCCVA were used. Three attempts were applied to each unbalanced ratio and plotted as scatter points to show the possible fluctuations caused by random selection of spectra. The average classification accuracy of both cancer and non-cancer classes were plotted as trend guidance in Figure 7.3 (PLS-DA) and Figure 7.4 (PCCVA), where A) showed classification accuracy after under-sampling and B) showed after over-sampling with different unbalanced ratios respectively.

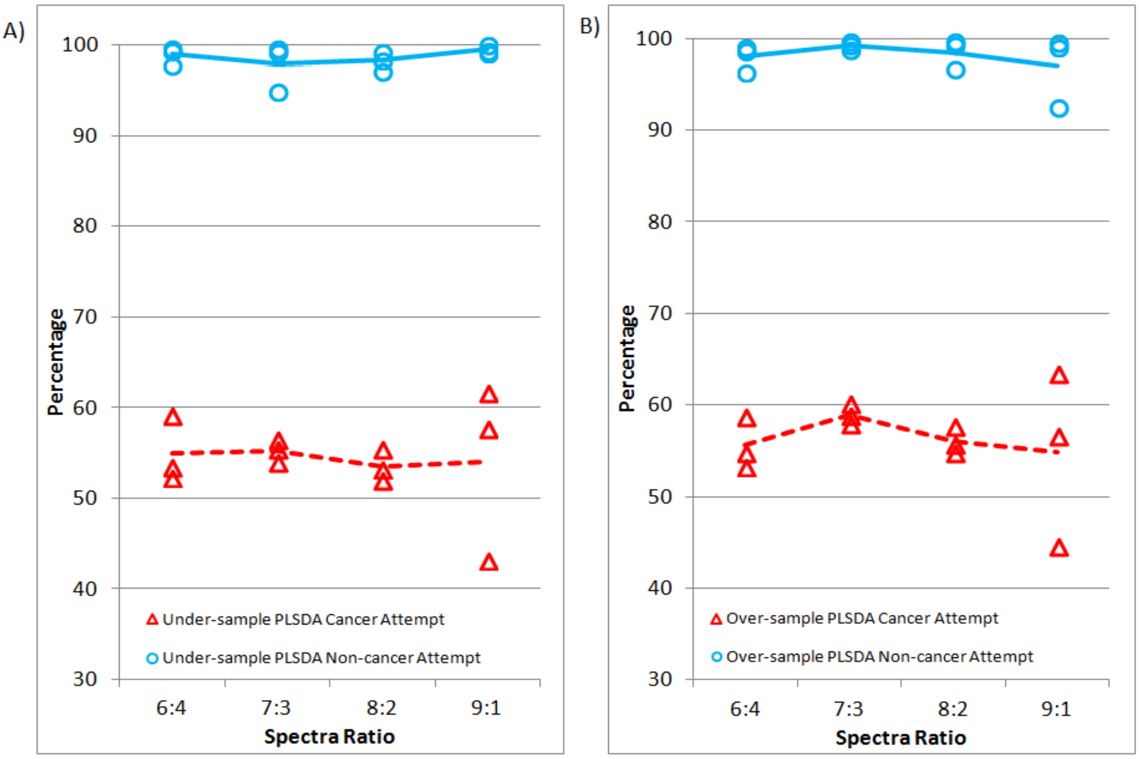


Figure 7.3 PLSDA classification accuracy with different unbalanced ratios A) under-sampling method was applied B) over-sampling method was applied, where spectra ratio stands for the initially unbalanced ratio between the number of spectra in cancer class and NAT class applied in training process. Average classification accuracies of NAT class was plotted in blue solid line while cancer class was plotted in dotted red line

For PLS-DA, surprisingly both sampling methods had similar performance with all unbalanced ratios. For under-sampling, the averaged cancerous spectra separation accuracy remained in the range of 53% (at 8:2 ratio, 2000 spectra per class) and 55% (at 7:3 ratio, 3000 spectra per class), while the non-cancerous spectra classification rates were between 98% (at 7:3 ratio, 3000 spectra per class) and 99% (at 9:1 ratio, 1000 spectra per class). For over-sampling, the averaged cancerous spectra separation accuracy remained between 55% (at 9:1 ratio, 9000 spectra per class) and 59% (at 7:3 ratio, 7000 spectra per class), while the non-cancerous spectra classification accuracies were remained in the range of 97% (at 9:1 ratio, 9000 spectra per class) and 99% (at 7:3 ratio, 7000 spectra per class). Both methods had only small fluctuations in term of averaged classification accuracy. Not much difference was observed when different unbalanced ratios were applied using both methods.

In general, it could see that PLSDA did not benefit from increasing or decreasing the number of spectra in the training dataset in this case. For all unbalanced ratios, the averaged results remained within a five percent range for both sampling methods. PLSDA kept its performance when only a small number of spectra were used, which could be considered as an advantage of PLSDA classification.

Unlike averaged classification accuracy, the scatter of three attempts of classification on independent test data was large for both under-sampling and over-sampling when the 9:1 unbalanced ratio was applied. For under-sampling, the spread of classification accuracy of cancerous spectra increased. One possible reason could be that with the decrease in the number of training spectra, the linear separator calculated did not bring as good separation as before. For over-sampling, the classification results were influenced by the purity of training data, since for the non-cancerous class only 1000 unique spectra were available in a total number of 9000 spectra. These observations could be considered as a decrease in reliability in the classification result.

To explain the obtained results, as cancerous spectra including different stages and grades of breast cancer, it was possible that the spread of spectra after data projection in cancer class was more spread or even surround the non-malignant class. Clear separation cannot be achieved by applying linear separator as it was different to draw a straight line between these two classes. This assumption would be validated in later of this section.

For PCCVA, shown in Figure 7.4, both sampling methods had similar averaged classification accuracies when the unbalanced ratios 6:4, 7:3 and 8:2 were applied. For under-sampling, the averaged cancerous spectra separation accuracy remained in the range of 57% (at ratio 6:4, 4000 spectra per class) and 59% (at 8:2 ratio, 2000 spectra per class), while the non-cancerous spectra classification rates were between 88% (at 7:3 ratio, 3000 spectra per class) and 89% (at 6:4 ratio, 4000 spectra per class). For over-sampling, the averaged cancerous spectra separation accuracy remained between 57% (at 6:4 ratio, 6000 spectra per class) and 58% (at 8:2 ratio, 8000 spectra per class), while the non-cancerous spectra classification accuracies remained in the range of 88% (at 8:2 ratio, 8000 spectra per class) and 90% (at 7:3 ratio, 7000 spectra per class). Both methods had only small fluctuations in terms of averaged classification accuracy. When the ratio 9:1 was applied, the non-cancerous classification accuracy of over-sampling remained 86% while classification accuracy after applying under-sampling dropped rapidly from average from 89% to 69%. Classification accuracy of the cancerous class after under-sampling was reduced from average 59% to 52%, while the averaged accuracy of over-sampling remained at a similar level.

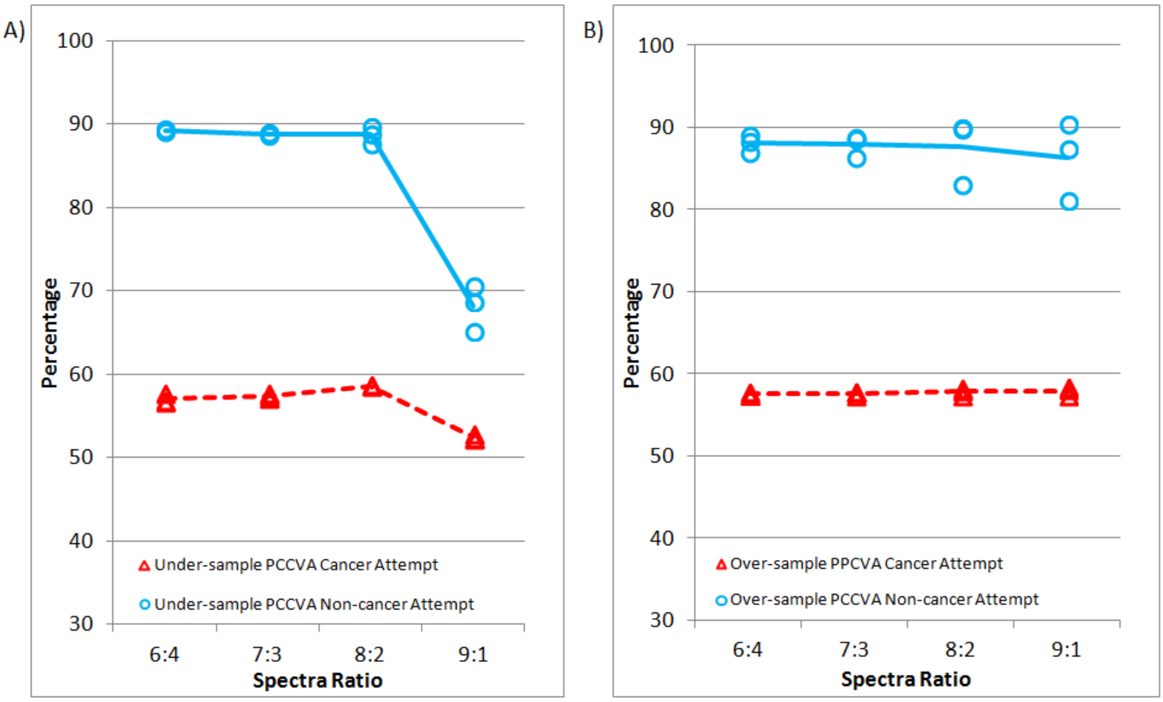


Figure 7.4 PCCVA classification accuracy with different unbalanced ratios A) under-sampling method was applied B) over-sampling method was applied, where spectra ratio stands for the initially unbalanced ratio between number of spectra in cancer class and NAT class applied in training process. Average classification accuracies of NAT class was plotted in blue solid line while cancer class was plotted in dotted red line

It could be observed that, for the non-cancer class, the increase in the number of spectra by over-sampling did not cause large increase in its classification accuracy but when number of spectra was reduced to certain level, (in this case, 1000 spectra), its classification accuracy was be largely reduced. It was possible that when the number of spectra was reduced to certain point, the model started to lose some information which was curial for linear separation. To test if this was the case CV loadings were plotted in Figure 7.6. A comparison of the loadings between the unbalanced ratios 6:4 and 9:1, showed that differences were observed in the ranges of 1134 and 1176 cm^{-1}, 1593 and 1604 cm^{-1}, and peak 1641 cm^{-1}. The shape of loading curve with fewer training spectra did not vary much from the curve with more spectra, which in a way explained the stable performance of PCCVA, as the linear separator changed only very little in terms of peak intensity. Most previously separated independent spectra would still be separated even through less information was provided. Looking at the distribution patterns of PCCVA scores of both 6:4 and 9:1 ratios in Figure 7.5, both A) and B) had a similar distribution pattern with different scales, where scores were more spread with 6:4 ratio while scores were more gathered when it came to 9:1 ratio, as more information was given during model construction. It could be seen that there was difference between medians of cancerous and non-cancerous scores.

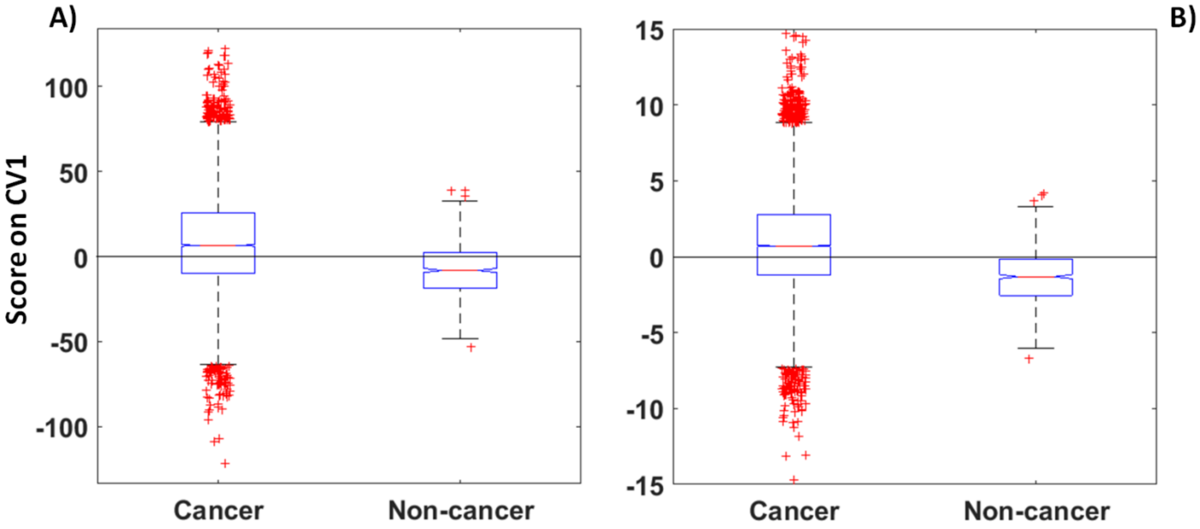


Figure 7.5 PCCVA scores of under-sampled independent test dataset. A) with unbalanced ratio of 6:4 B) with unbalanced ratio 9:1

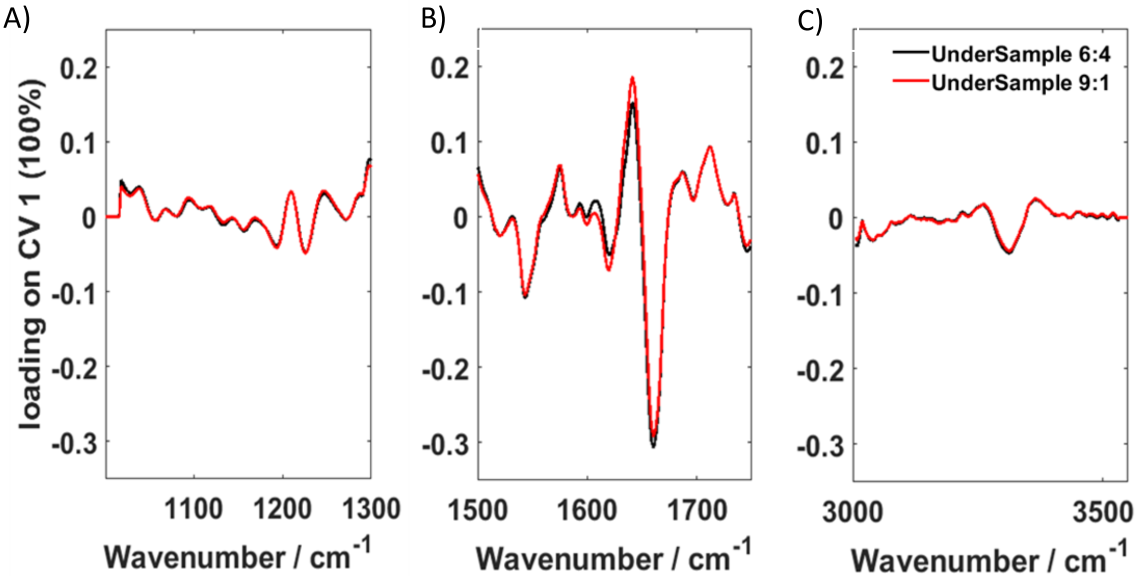


Figure 7.6 CV loadings of under-sampled independent test dataset, where unbalanced ratio of 6:4 was plotted in black and 9:1 was in red in the wavenumber range A) 1050 to 1300 cm^{-1}, B) wavenumber range 1500 to 1750 cm^{-1}, and C) wavenumber range 3005 to 3550 cm^{-1}

For the cancer class separation, it could be seen that PCCVA does not benefit from increasing the number of spectra in the training dataset. Going from 2000 spectra to 4000 spectra (under-sampling) and 6000 to 9000 spectra (over-sampling), the cancerous class classification accuracy stayed between average of 57% and 58%. In terms of scatter points representing repeating experiments, three attempts of classification on independent test data resulted in a large spread of ranges for both under-sampling and over-sampling when 9:1 unbalanced ratio was applied. For over-sampling, the spread of the scatter points showed that classification results were influenced by the purity of training data, as for non-cancerous class only 1000 unique spectra were available in total number of 9000 spectra. This observation could be considered as a decrease of reliability of the classification result, which could be seen in Figure 7.7, where although the information provided for model construction increases, the distance between median of CV scores decreases. More information was therefore not always good for classification.

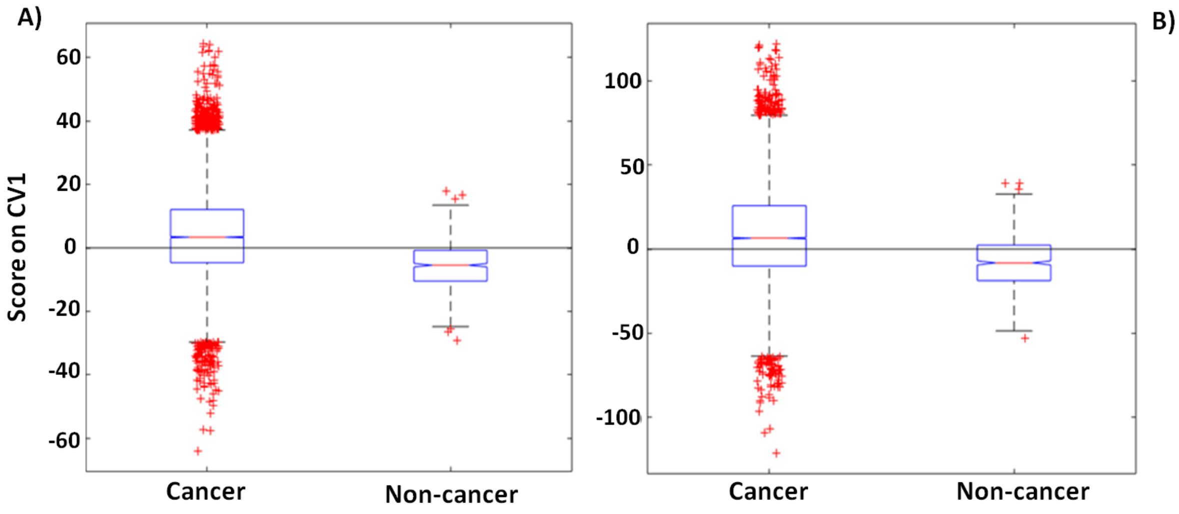


Figure 7.7 PCCVA scores of over-sampled independent test dataset. A) with unbalanced ratio of 6:4 B) with unbalanced ratio 9:1

PCCVA and PLSDA had poor classification of cancerous spectra but high accuracy to separate non-cancerous spectra. One possible reason could be that non-cancerous epithelium cells were more uniform than cancerous ones. Cancer cells exhibit more variability in terms of cell size and shapes compared to normal cells [20]. Non-cancerous spectra were more gathered as a group while cancerous spectra were more spread. In addition to that, individual differences among patients having different stages and grades of breast cancer could contribute to disjoint distribution of the cancerous class, which could not be separated by linear separators. Why do linear separators seem to fail? To answer this question, projected PCA score of an independent dataset was plotted and shown in Figure 7.8. It could be seen that there were two clusters of cancerous spectra, which could be considered as disjoint data distribution, and non-cancerous spectra was located in the middle of these two groups. If a disjoint data distribution was observed in one class, linear separators were not recommended, as it was difficult to separate data using one line. This explains why nearly half of the cancerous spectra were incorrectly classified while non-cancerous spectra were separated with high accuracy.

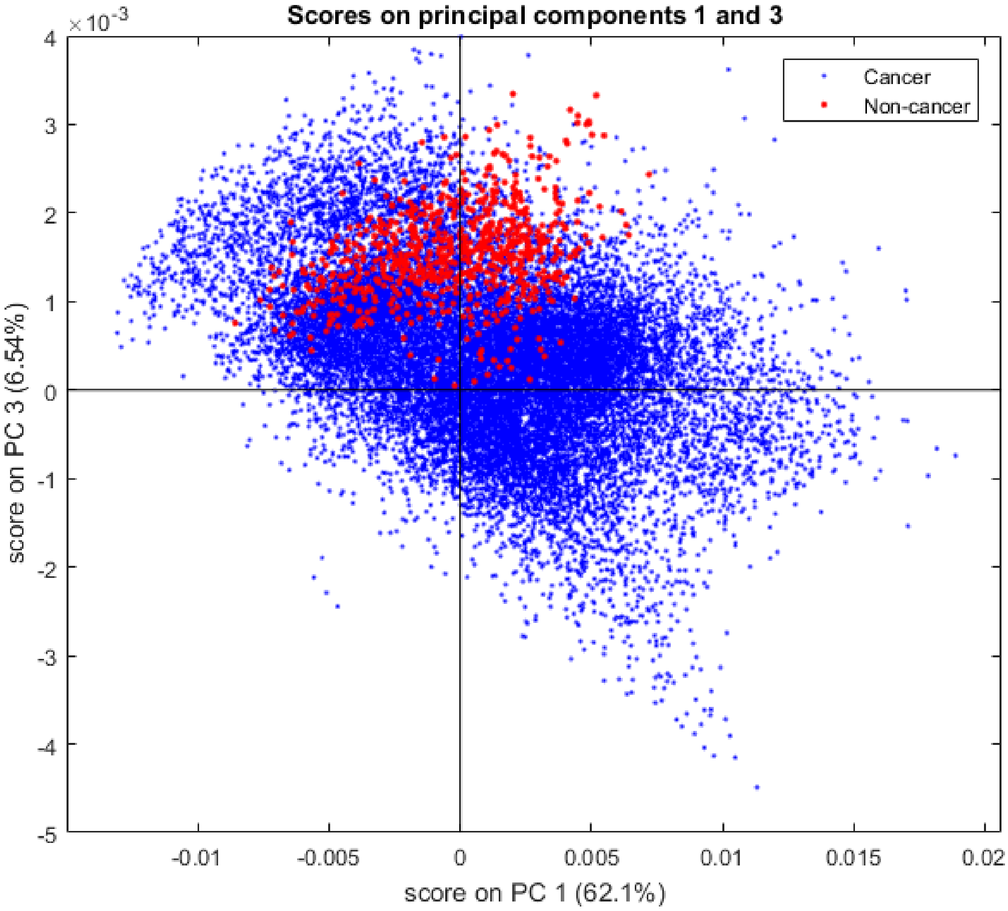


Figure 7.8 projected PCA score of independent dataset on PC 1 and 3, where data in cancer class was plotted in blue while the NAT class was plotted in red

7.3.2.2. Machine learning

Both under-sampling and over-sampling were conducted combining with both AdaBoost and random forest to balance data with different unbalancing ratios. Three attempts were applied to each unbalanced ratio. The average classification accuracy of both cancer and non-cancer classes were also plotted as this helps to identify possible trends.

7.3.2.2.1. Applying AdaBoost

Using AdaBoost it could be seen in Figure 7.9 that with small unbalanced ratios (6:4 and 7:3), the non-cancer classification accuracies were similar to those obtained using under-sampling, 98% and 95%, and over-sampling 95% and 96% respectively. With large ratios (8:2 and 9:1), under-sampling remained its performance to a similar level (98% at 8:2 and 97% at 9:1) while the accuracy of over-sampling dropped from 96% (at ratio 7:3) to 87% (reduced by nearly ten percent). One possible reason for the accuracy decrease with the number of spectra increase was that the purity of non-cancerous spectra was reduced, as the number of unique spectra in the group was reduced while the number of replicates increased. The model could pick up repeated features, which were similarities among these patients but not cancer diagnosis related, as key features and weighted. This could reduce the performance of model as separation favoured features would not be as outstanding as applying less repeated spectra to train the model. Traces could be found in Figure 7.10, where in B) peaks had more intensity in terms of importance of each feature with 9:1 ratio. If importance of a feature was larger than 0.0002 was considered as key features of the model, in plot A) 3 key features were found while in plot B) 5 features were spotted. In plots A) and B), both bands at 1543 and 1658 cm^{-1}, which were in amide II and amide I regions respectively, had the biggest contribution in terms of classification. In plot A), the amide I region had the dominant position while in plot B) amide II had the highest intensity.

Another possible reason for decreasing classification accuracy of non-malignant class with increasing unbalanced classes could be that although there was no lost of information, over-sampling still sometimes led to over-fitting [21]. By duplicating existing spectra, repeated common but not cancer versus NAT spectra classification related features among the training spectra were weighted more, which made the model perfect to the training set but performs poorly to other datasets. For the cancerous class, when the ratio increased to 9:1, over-sampling had 9000 unique cancerous spectra as training input, while after under-sampling only 1000 cancerous spectra were kept as training input. Large differences in number of spectra led to accuracy differences.

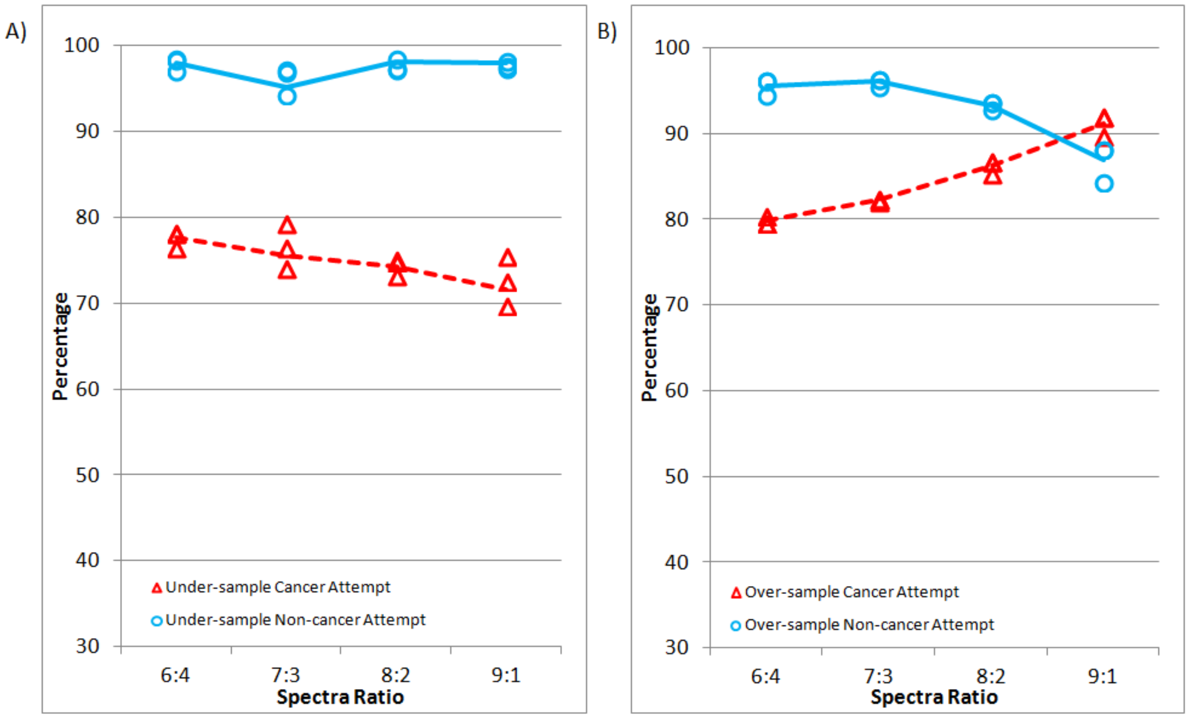


Figure 7.9 AdaBoost classification accuracy with different unbalanced ratios A) under-sampling method was applied B) over-sampling method was applied, where spectra ratio stands for the initially unbalanced ratio between number of spectra in cancer class and NAT class applied in training process. Average classification accuracies of NAT class was plotted in blue solid line while cancer class was plotted in dotted red line

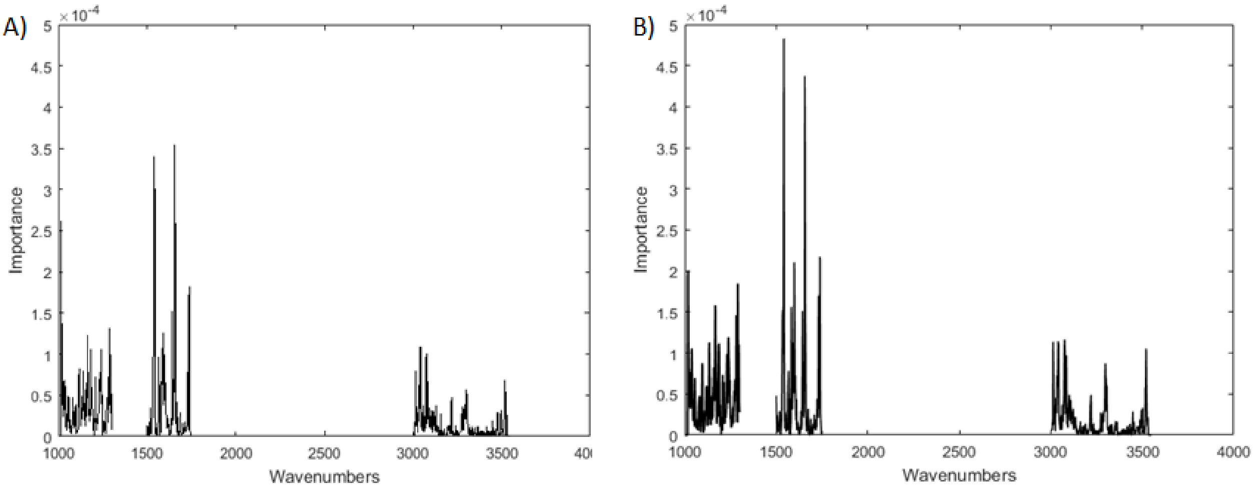


Figure 7.10 Importance Plot of over-sampled training data using AdaBoost. A) with 6:4 unbalanced ratio B) with 9:1 unbalanced ratio

For the cancerous class, which was the major class in this case, under-sampling decreased the averaged accuracy of classification while over-sampling increased the averaged classification accuracy. The reason was very straight forwards. Under-sampling reduced the number of spectra input with the increase of unbalanced ratio, which meant the training information had been reduced by deleting examples from the training data. Large differences in number of spectra led to accuracy differences, which could be observed from Figure 7.11. In Figure 7.11 B), fewer features were considered important in term of separation comparing with 6:4 unbalanced ratio (Figure 7.11, section A).

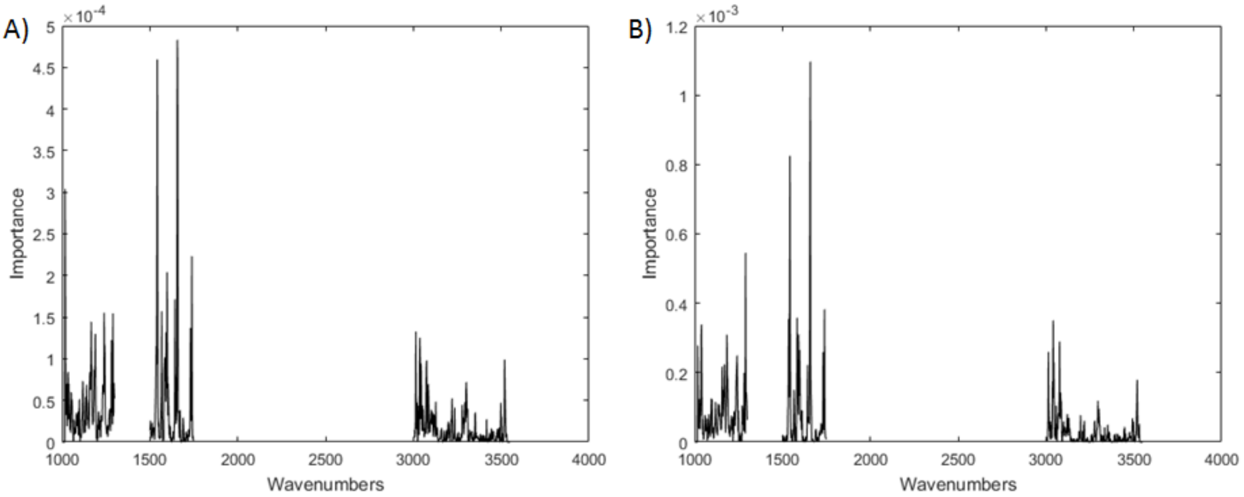


Figure 7.11 Importance Plot of under-sampled training data using AdaBoost. A) with 6:4 unbalanced ratio B) with 9:1 unbalanced ratio

To closely exam the impact of number of spectra in training on the classification accuracy of AdaBoost, groups of balanced classes with different number of spectra were applied to investigate how many spectra was enough to construct a relatively good machine learning model. From 5000 spectra per class, spectra in every smaller class were randomly selected from the previous class to maximally eliminate differences caused by random selection of spectra. Three attempts were applied to each group. The average classification accuracy of both cancer and non-cancer classes were plotted in Figure 7.12, where blue lines showed independent test accuracies of cancer separation while the red lines showed the accuracy of non-cancer classification. All results of three attempts were plotted as scatter points (blue for the cancer class and red for the non-cancer class) to show the possible fluctuations caused by random selection of spectra.

It could be seen that the non-cancerous spectra classification accuracy decreased gradually before completely dropping to zero, while cancer classification accuracy gradually decreased before increasing again to one hundred percent, where the model completely failed. At 240 spectra (120 per each class), AdaBoost started to fail, while at 260 spectra (130 per class), it could still give average of 62% (cancer) and 81% (non-cancer) classification accuracy.

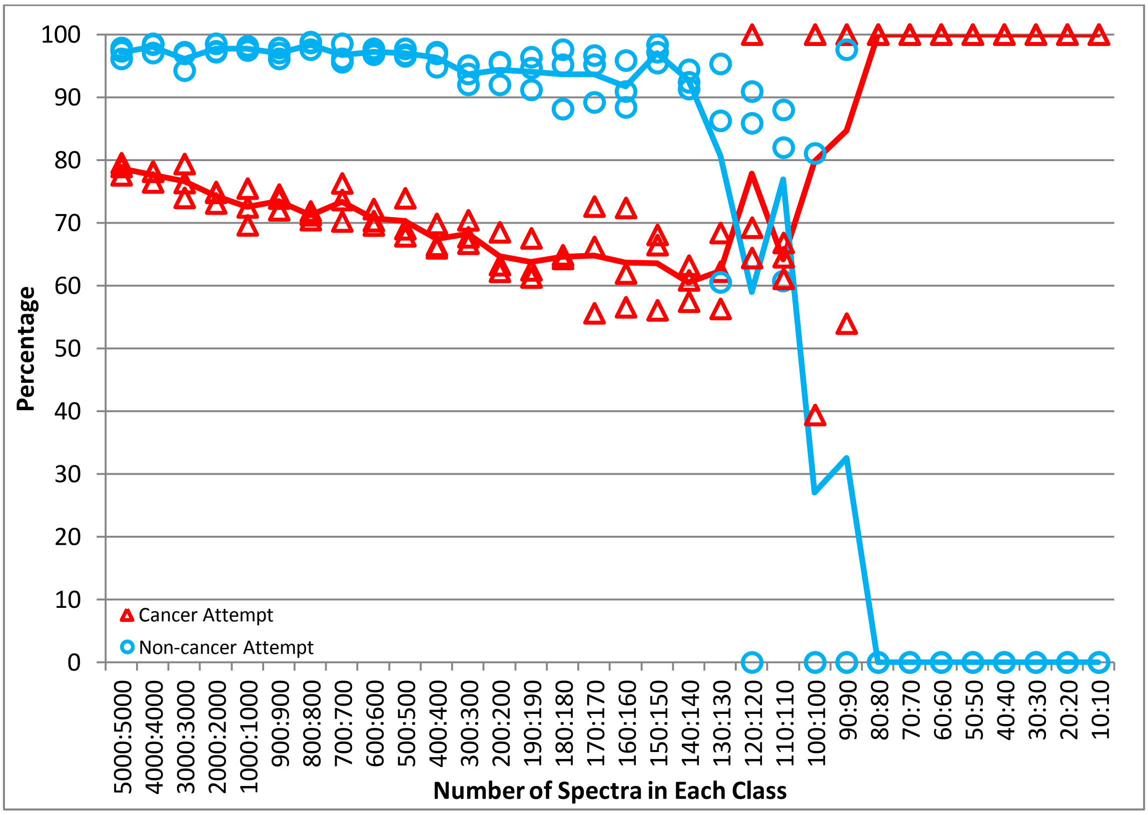


Figure 7.12 Classification accuracy of AdaBoost with different number of spectra, where average classification accuracies of NAT class was plotted in blue solid line while cancer class was plotted in solid red line

7.3.2.2.2. Applying Random forest

When random forest was applied to the data sets, shown in Figure 7.13, a similar, but more pronounced trend to that seen for the Application of AdaBoost (Figure 7.9) was observed. With small unbalanced ratios (6:4 and 7:3), non-cancer classification accuracies were close between under-sampling, 97% and 96%, and over-sampling, 95% and 93% respectively. With large ratios (8:2 and 9:1), under-sampling keep its performance (96% and 92% respectively) to a similar level. Large differences in number of spectra led to accuracy differences. The accuracy of over-sampling dropped from 93% (at ratio 7:3) to 72% (reduced by nearly twenty percent, over twice that seen when of applying AdaBoost). The reason for the trend was mentioned in the previous section.

For the cancerous class, again the same trend was spotted. With the increase of unbalanced ratios, the averaged classification accuracy decreased from 73% at ratio 6:4 to 70% at ratio 9:1 when applying under-sampling, while it increased from 78% at ratio 6:4 to 89% at ratio 9:1 when over-sampling was applied. Differences in the number of training spectra were still the main reason for the differences in classification as mentioned in the previous section and similar observations on the importance plots were shown in Figure 7.14.

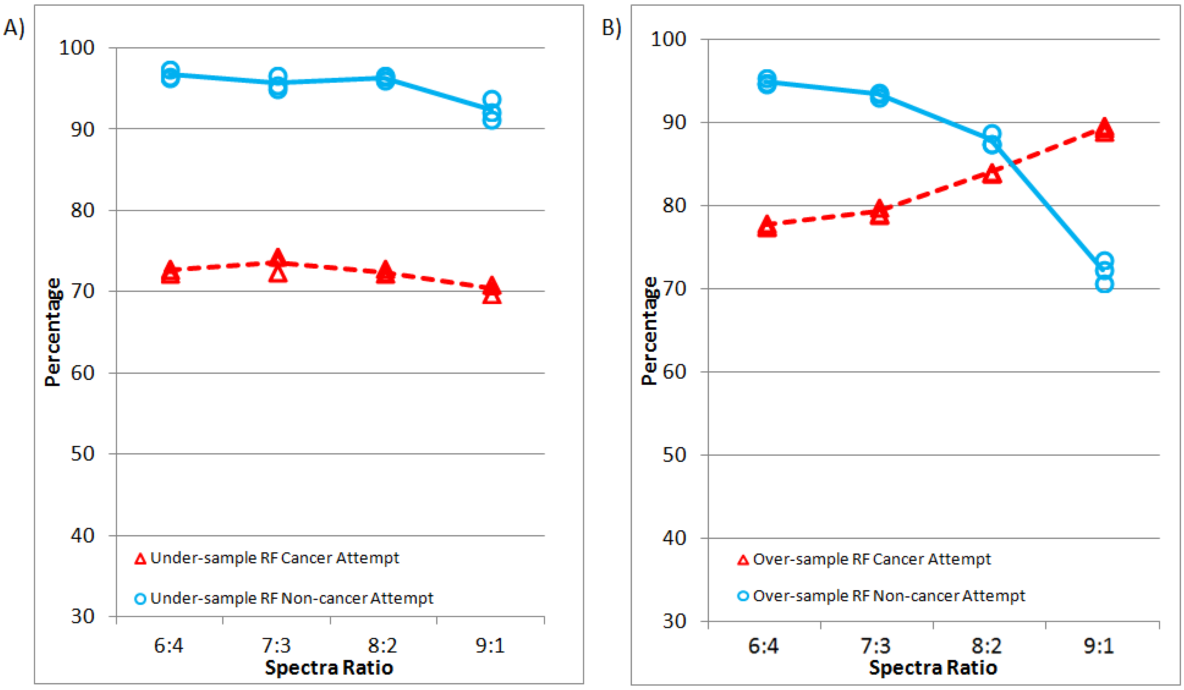


Figure 7.13 AdaBoost classification accuracy with different unbalanced ratios A) under-sampling method was applied B) over-sampling method was applied, where spectra ratio stands for the initially unbalanced ratio between number of spectra in cancer class and NAT class applied in training process. Average classification accuracies of NAT class was plotted in blue solid line while cancer class was plotted in dotted red line

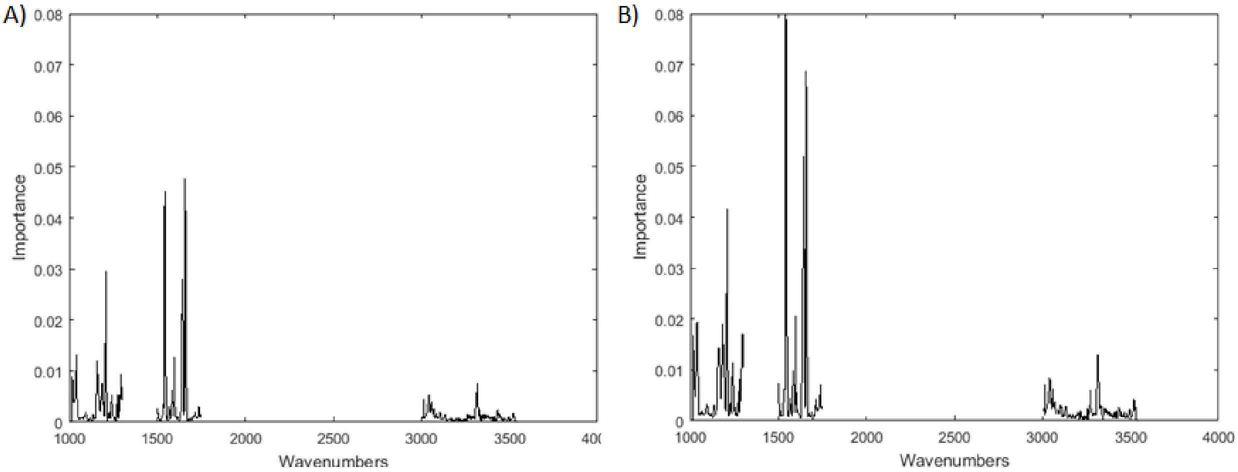


Figure 7.14 Importance Plot of over-sampled training data using random forest. A) with 6:4 unbalanced ratio B) with 9:1 unbalanced ratio

To closely exam the impact of the number of spectra in the training set on the classification accuracy of random forest, groups of balanced classes with different number of spectra were applied to investigate how many spectra was enough to construct a relatively good machine learning model. From 5000 spectra per class, spectra in every smaller class were randomly selected from the previous class to maximally eliminate differences caused by random selection of spectra. Three attempts were applied to each group. The average classification accuracy of both cancer and non-cancer classes were plotted in Figure 7.15, where blue lines showed independent test accuracies of cancer separation while the red lines showed the accuracy of non-cancer classification. All results of three attempts were plotted as scatter points (blue for the cancer class and red for the non-cancer class) to show the possible fluctuations caused by random selection of spectra.

Both cancer and non-cancer classification accuracies decreased but there was no complete failure, as there was no zero classification accuracy. However the accuracy ranges of three attempts were much more spread with the decrease of number of spectra. For non-cancerous spectra, from 400 spectra (200 per class) the largest difference between two of three attempts was over twenty percent. All attempts of cancerous class were more gathered before 40 spectra (20 per class). Although random forest was more consistent than AdaBoost as there was no complete failure, its stability reduced after a total number of 400 spectra.

If only a very small number of spectra could be used to construct models, random forest was more suitable than AdaBoost. However, many attempts should be conducted to ensure the reliability of the final classification accuracy obtained.

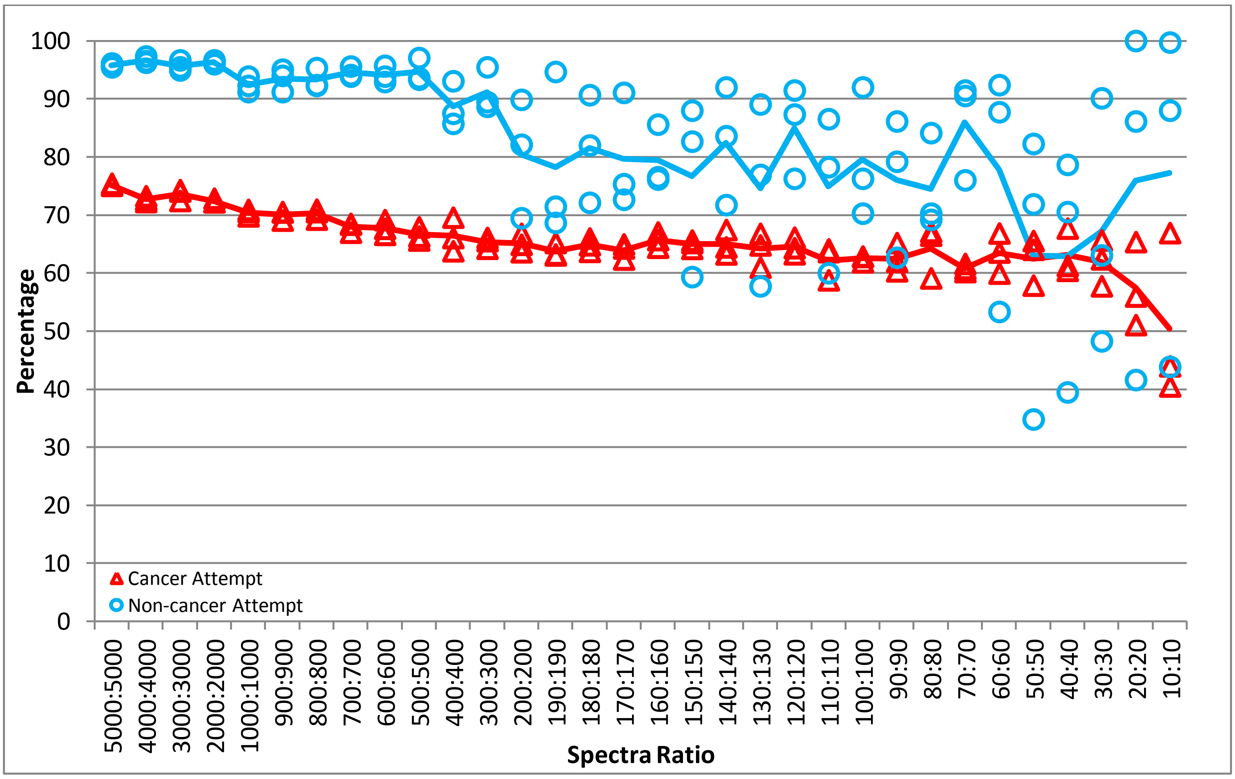


Figure 7.15 Classification accuracy of Random Forest with different number of spectra, where average classification accuracies of NAT class was plotted in blue solid line while cancer class was plotted in solid red line

To sum up, considering the classification accuracy of minority class, under-sampling had better performance in both machine learning models. For small unbalanced ratios, both sampling methods could be applied to balance the classification accuracies of two classes. However, when large unbalanced ratios were applied, under-sampling was more appropriate as over-sampling could over-fit and easier to be influenced by noise.

For unbalanced training data, considering the classification accuracy of minority class, under-sampling had better performance classification models. If unbalanced ratios were small, for example 6:4 or 7:3, both over-sampling and under-sampling could be adopted in term of balancing two classes, as both methods led to similar results. However, when large unbalanced ratios were applied, in this case, over 8:2, under-sampling was more appropriate than over-sampling. With the increase of replicates, purity of spectra in minority class would be reduced, which led to larger possibility of over-fit and classification would be more influenced by noise.

A flowchart can be generated as guidance when it comes to spectra classification in Figure 7.16. The recommended path will be using under-sampling to balance classes first, followed by machine learning classification methods. In this study it is shown that Random forest can be performed if there are more than 200 spectra in each class. AdaBoost can be used if there are more than 130 spectra in each class. The actual number of spectra, however, may differ for different data sets.

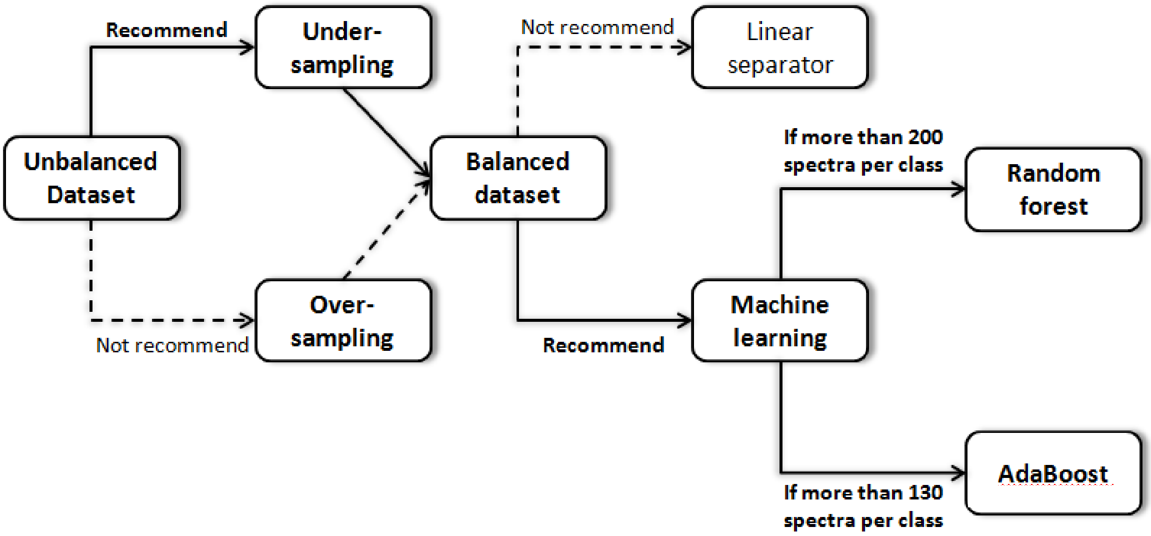


Figure 7.16 Guidance flowchart of spectra classification

7.4 Conclusion and future work

In term of picking the most appropriate classification method, data structure should be observed first. If disjoint data distribution is observed in one class, linear separators are not recommended, as it is difficult to separate data using one line. As showed previously, machine learning approaches are less influenced by data distribution than linear separators. However, there is no one method which is the best for everything. Exploring more methods is always better in order to find the best method which solves one particular problem. The flow chart is plotted as a little help guide when it comes to classification of human spectra.

7.5 References

[1] J. Bell, Machine Learning Technical Professionals, 2014.

[2] K.P. Murphy, Machine learning : a probabilistic perspective, MIT Press, 2012.

[3] A. Cutler, D.R. Cutler, J.R. Stevens, Random Forests, (2012). doi:10.1007/978-1-4419-9326-7.

[4] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, Analyst. 142 (2017) 1258–1268. doi:10.1039/c6an02224c.

[5] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, RUSBoost : A Hybrid Approach to Alleviating Class Imbalance, 40 (2010) 185–197.

[6] Y. Freund, R.E. Schapire, Experiments with a New Boosting Algorithm, (1996). https://cseweb.ucsd.edu/~yfreund/papers/boostingexperiments.pdf (accessed June 16, 2018).

[7] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, 139 (1995) 23–37. doi:10.1007/3-540-59119-2_166.

[8] B. Schölkopf, Z. Luo, V. Vovk, Empirical inference: Festschrift in honor of Vladimir N. Vapnik, Empir. Inference Festschrift Honor Vladimir N. Vapnik. (2013) 1–287. doi:10.1007/978-3-642-41136-6.

[9] A.S. Hess, J.R. Hess, Principal component analysis, Transfusion. (2018). doi:10.1111/trf.14639.

[10] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/C3AY41907J.

[11] A.J. Izenman, Modern Multivariate Statistical Techniques, Springer New York, New York, NY, 2008. doi:10.1007/978-0-387-78189-1.

[12] P.W. Yendle, H.J.H. Macfie ’, DISCRIMINANT PRINCIPAL COMPONENTS ANALYSIS, J. Chemom. 3 (1989) 589–600. https://onlinelibrary.wiley.com/doi/pdf/10.1002/cem.1180030407 (accessed June 17, 2018).

[13] N.A. Campbell, Some practical aspects of canonical variate analysis, J. Appl. Stat. 6 (1979) 7–18. doi:10.1080/02664767900000002.

[14] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis for chemometrics and metabolomics: How scores, loadings, and weights differ according to two common algorithms, J. Chemom. 32 (2018) 1–16. doi:10.1002/cem.3028.

[15] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis : taking

the magic away, (2014) 213–225. doi:10.1002/cem.2609.

[16] M.M. Rahman, D.N. Davis, Addressing the Class Imbalance Problem in Medical Datasets, Int. J. Mach. Learn. Comput. (2013) 224–228. doi:10.7763/IJMLC.2013.V3.307.

[17] G.M. Weiss, G. M., Mining with rarity, ACM SIGKDD Explor. Newsl. 6 (2004) 7. doi:10.1145/1007730.1007734.

[18] F. Lyng, E. Gazi, P. Gardner, Preparation of Tissues and Cells for Infrared and Raman Spectroscopy and Imaging, RSC Anal. Spectrosc. Monogr. 01 (2011) 147–185. http://arrow.dit.ie/radrep.

[19] Sunil R. Lakhani; Ian O. Ellis;, S.J.P.H.T. Schnitt, V.M.J. van De, WHO Classification of Tumours of the breast, 2012. doi:10.1017/CBO9781107415324.004.

[20] A.I. Baba, C. Câtoi, Chapter 3 TUMOR CELL MORPHOLOGY, in: Comp. Oncol., 2007: pp. 1–31. http://www.ncbi.nlm.nih.gov/books/NBK9557/.

[21] C. Drummond, R.C. Holte, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, Work. Learn. from Imbalanced Datasets II. (2003) 1–8. doi:10.1.1.68.6858.

Chapter 8

Prostate tissues on glass

8.1. Chapter Overview

In the UK, nearly 250 thousand people are diagnosed with cancer every year. Breast, lung, bowel, and prostate, which are the most common types of cancer, account for over half of all these cases. In 2015, around 47,200 men were diagnosed with prostate cancer in the UK [1]. Prostate cancer is strongly correlated with age and indeed is endemic in the older population. With incidence rates increasing at 3.7% annually in the period between 2010 and 2015 [2], prostate cancer occurrences in the community aged over 60 have caused substantial tension on the health care systems and has also lead to enormous health care costs.

In most cases of cancer diagnosis, biopsies are required and taken to assess the presence of a disease and its degree of progression. For most examples, these thin slices of tissue are carefully investigated by professional pathologists under the help of optical microscopes. This process is a time-consuming laboratory process. More importantly, this process is subjective and prone to observer error [3]. Comparing with the traditional diagnosis method, infrared micro-spectroscopy has the potential to be a faster, cheaper and more reliable method of human tissue analysis.

In the past ten years, research on using infrared micro-spectroscopy to study cancer and its diagnosis have increased significantly [4–6]. Although infrared micro-spectroscopy brought attractive positive results in terms of cancer diagnosis accuracy, the infrared technique requires biopsy samples with different standards from those currently used for pathologists. Not only do the samples need to be cut into different sizes and thickness, but the slices of tissue also have to be placed on mid-infrared transparent substrates (for example, calcium fluoride (CaF_2) or barium fluoride (BaF_2)) for spectra collection. These substrates used for transmission mode, which has less distortion from EFSW effect and provides less influenced biological information [7,8], are both expensive (typically greater than $40 per slide) and particularly fragile and frangible. Due to their extreme brittleness, they are not suitable for automatic tissue sample preparation instruments. All the sample preparation process has to be manually conducted [9]. Both the price and nature of substrates increase the difficulty of popularising the whole infrared cancer diagnosis technology into the clinical field.

In order to get the technology adopted by the clinical field, similar or even the same samples together with substrates used by pathologists, namely tissue on glass slides, could be considered as a possibility for infrared measurements. Glass has its advantage in both price and robustness. Most importantly it is already commonly used in the clinical field by pathologists. It will introduce the least interruption of the original clinical workflow and will make the new technology much easier to be accepted by the field. In this chapter, the possibility of using conventional glass as a transmission substrate combining with fast measuring (using high spectral resolution) is discussed. The wavenumber range chosen for the research was the only narrow transmission window from 3125 to 3600 cm^{-1} (referred as the high wavenumber range), which covers containing the N–H, and O–H stretching regions. Currently, most research on cancer diagnosis focuses on the fingerprint region (below 1800 cm^{-1}) and considers the high wavenumber range to contain very little diagnostic information However in the light of the previous work conducted by Bassan *et al*.[2] and Pilling *et al*.[10], it is believed that with sufficient data mining and large patient cohort it is possible to solve the problem. In order to achieve this goal, a high spectral resolution was used to rapidly analyze 17 prostate tissue microarrays (TMAs) on glass substrates. Machine learning models were constructed in order to discriminate cancerous and normal associate prostate tissue.

8. 2. Methodology

8. 2.1 Prostate tissue preparation

For this study 17 formalin-fixed paraffin embedded prostate TMA slides were used. They are borrowed through the Christie NHS Foundation Trust from UK BioBank. TMAs contain 1288 cores from 294 patients. Each core in the TMA was cut to 5 µm thickness and 1 mm diameter. The whole TMA section was floated onto a standard histology glass slide.

To construct high variance models 100 cores were selected from the sample set as a training group, which included 50 cores with different grades of prostate cancer and 50 normal associate prostate tissue cores from 60 patients. Another 10 cores from 10 independent patients were selected as independent test set.

Additional 21 cores, including 11 other cores from the previously selected patients in the training group and 10 cores from another independent 10 patients from the training group, were chosen as a larger test set for the study of cancer classification.

In total, 131 cores from 80 patients were picked and measured. Cores were selected based on both the condition and completeness of cores, and the stages and grades of prostate cancer of cores.

8.2.2 Instrumentation and experimental procedures

The instrument and microscope used in the work reported here were those that have previously detailed in chapter 4.

Before imaging, background scans were taken as a single tile with 32 co-added scans at a spectral resolution 9.3 cm^{-1} which gave nearly the same resolution as that commonly used i.e. 8 cm^{-1} but with an acquisition speed of similar to using 16 cm^{-1}. Interferograms were processed into absorption spectra using Happ-Genzel apodisation resulting in spectra with a region between 1000 and 3800 cm^{-1}.

A purge time of 60 minutes was adopted to reduce the influence of water vapour in the spectrum after loading the TMA on to the microscope stage and after changing samples. The microscope was focused before the spectrometer calibration, which was performed on the clear part of the slide. The absorbance level of each detector element was checked and adjusted to be relatively uniform before every mosaic measurement throughout the whole project to minimise the possible systemic error caused by the equipment.

8.2.3 Data analysis methods

All data were pre-processed with MATLAB 2018a. Infrared spectra for each biopsy core were extracted from the mosaic as a 256 × 256 datacube. Each datacube consists of 65536 spectra each of which contains 416 data points.

8.2.3.1 Annotation method

Chemical images of each of the prostate tissue cores were generated and compared to the H&E stained sections, and regions of epithelium were identified and annotated on chemical images generated based on WHO Classification of Tumours in the Prostate [11].

An example of 50 annotated cores is shown in Figure 8.1. Cancerous epithelium cells were annotated in red (R255 G0 B0) and the stroma parts in purple (R162 G77 B255), while in the cancer associated tissues, epithelium cells were coloured in green (R0 G255 B0) and stroma in yellow (R243 G254 B68).

8.2.3.2 Data Pre-processing Methods

Principal component based noise reduction was applied to improve the signal-to-noise ratio of raw spectra from the annotated area. The first 50 principal components were kept. Spectra were quality tested to remove data obtained from areas with little or no tissue based on the height of the amide A band. Spectra have absorbance between 0.07 and 1.2 were retained. Spectra ranges 3125 to 3600 cm^{-1} were taken. Each spectrum was then vector normalized to correct for different thicknesses of prostate tissue. Finally spectra were converted to first derivative and a Savitzky−Golay smoothing using a window size of 19 data points performed, which is considered to be a conventional parameter for tissue analysis [10]. The dataset was separated into two groups, a training set and independent testing set as previously described.

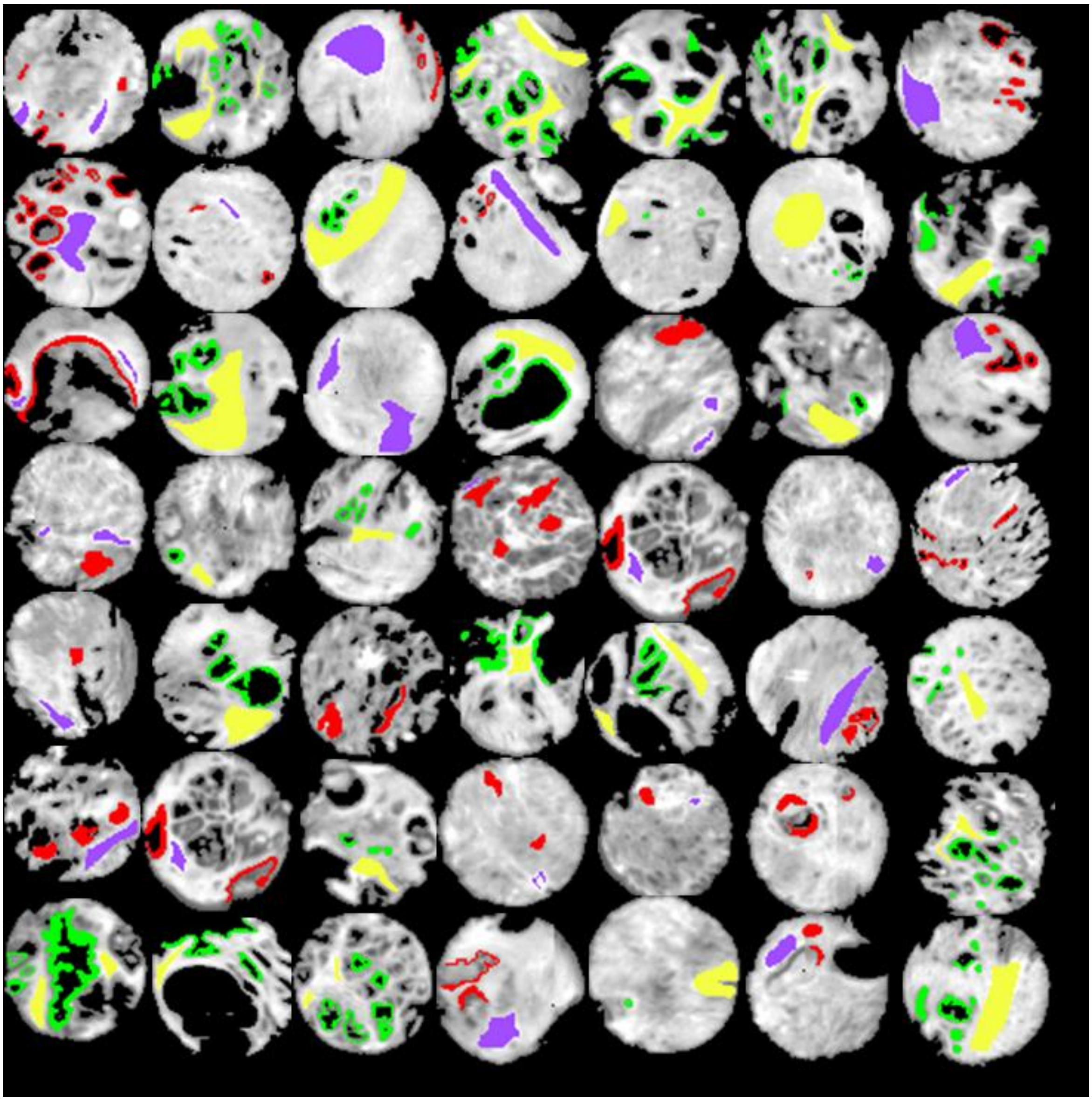


Figure 8.1 Examples of annotation on Infrared based chemical images of 49 prostate biopsy cores with red Cancerous epithelium, purple cancerous stroma, green non-cancerous epithelium, yellow noncancerous stroma

8.2.3.3 Classification Methods

Both Random forest and AdaBoost were applied in order to separate cancerous and NAT prostate tissues.

In terms of random forest, a random forest grows many classification trees as required. Once input data is put into the forest, each tree has its classification result. All trees in the forest are involved in voting the most popular class as the final result from the forest [12].

AdaBoost, however, works by obtaining weighted majority votes of the weak hypothesis where each hypothesis is assigned weight and conducted by every weak learner [13]. To be more specific, during each iteration, weights of incorrectly classified samples are modified with the aim of correctly classifying them in the next iteration. All trees are involved in weighted voting to allocate unknown samples. It is considered more effective at handling imbalanced datasets than random forest, as the minority class, which is much easier to be miss-classified, can be given higher weights in subsequent iterations [14].

Detailed explanation of both classification methods are in Chapter 4 (4.2.4.3).

8.2.3.4 Statistical Analysis Methods

Receiver operator characteristic (ROC) curves are a common way of representing the inherent trade-off between sensitivity and specificity. Sensitivity equalled the true positive over the positive predictions while the specificity was the quotient of true negative and all negative predictions. The detailed equation and explanations were presented in Chapter 4 (4.2.4.4).

Sensitivity, specificity, confusion matrix, ROC curve and AUC of classification results will be used to present the performance of each classifier.

8.3. Results and discussion

The results of classification between epithelium and the results of classification between cancerous and NAT prostrate spectra stroma using both random forest and AdaBoost were presented in this section.

8.3.1. Histology classification results

To test the quality of annotation and the possibility of conducting classifications on high spectral resolution data on glass, a histology classification was performed before the actual cancer diagnosis on the glass.

In terms of training both machine learning models, a total number of 26686 spectra subtracted from data cubes based on annotations. These consists of 14241 epithelium spectra (containing 5364 cancerous and 8877 NAT spectra) and 12445 stroma spectra (containing 3732 cancerous and 8713 NAT spectra). Under-sampling was conducted to balance the number of spectra in each class. 12445 epithelium spectra were randomly selected from 14241 spectra to match the number of spectra in stroma class. Each class therefore has the same number of spectra to ensure that there was no bias for either class. 124 features were used to construct classification models. 90% randomly selected spectra were used for model construction while 10% randomly selected spectra were adopted for a model validation test.

An independent test dataset consisted of 6647 spectra was extracted from data cubes, including 3430 epithelium spectra (1510 cancerous and 1920 non-cancerous spectra) and 3217 stroma spectra (1567 cancerous and 1650 NAT spectra).

8.3.1.1. Using AdaBoost classification method

In the case of separating epithelium and stroma, when AdaBoost was applied (500 iterations with 500 trees), 97% epithelium cells were correctly classified while 97% stroma cells were correctly classified in the model validation test, in which 10% randomly selected spectra from the training data set was used as test set to test the model performance on the training data. The confusion matrix was shown in Table

8.1. The overall classification accuracy was 97.1%, which was considered to be excellent for a model such as this.

Table 8.1 Confusion matrix of model validation test on the classification of epithelium and stroma cells using AdaBoost

Training Validation		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	97	3
	Stroma	3	97

When the independent test spectra were applied, without applying any threshold, epithelium cells were 90% correctly classified while 85% of stroma cells were correctly classified. The lower classification accuracy for stroma could be due to the variety of cells types in stroma. The general features of all types of cells in stroma were difficult to find compared with the specific features of epithelium cells. The confusion matrix of independent test was shown in Table 8.2.

Table 8.2 Confusion matrix of independent test on the classification of epithelium and stroma cells using AdaBoost

Independent Test		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	90	10
	Stroma	15	85

The ROC curve was plotted based on scores of each prediction (spectrum), shown in Figure 8.2. The AUC of each class was 0.92, which shows that the model works well on the independent test data set. To test whether the classification was valid, the importance of each feature was plotted, and was shown in Figure 8.3. Figure 8.3 , indicates that the classification was mainly caused by the variations in the spectra mainly associated with the peak at 3353 cm^{-1}, N-H stretching, which means that the classification between epithelium and stroma spectra was based on chemical differences other than water vapour or other physical conditions.

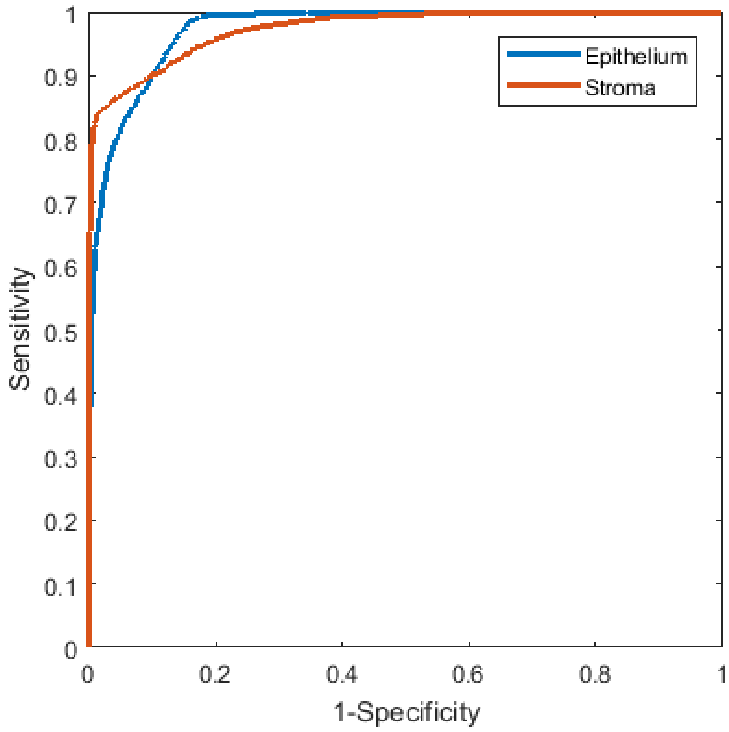


Figure 8.2 ROC curve of independent test on the classification of epithelium and stroma cells using AdaBoost

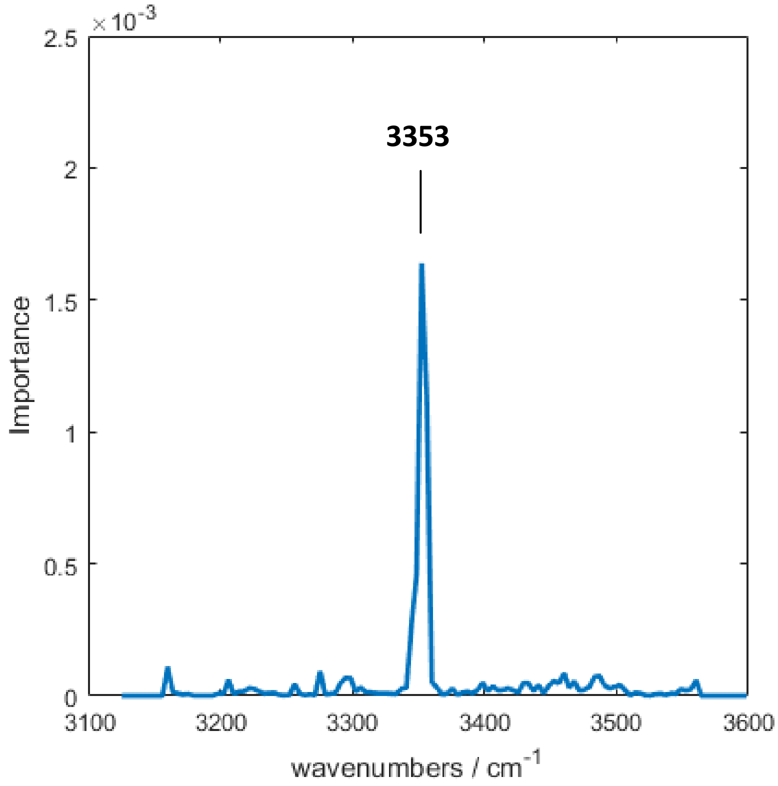


Figure 8.3 The feature importance of independent test on the classification of epithelium and stroma cells using AdaBoost

The performance of the model could be further boosted by adding a threshold after conducting classification. Only if the score of each spectrum was higher than a certain value, would it be considered to be an accurate prediction, as lower scores in most cases show that the model was not sure about the classification results. As long as, there were highly confident predictions, even though they might be fewer in number, the final decisions would still be more significant than those decisions made by keeping poor predictions, which could weight down the accuracy of final decision. Removing low scored predictions could help the model produce better results under the same configurations. If only half of the predictions were kept, 98% of epithelium spectra and 95% stroma spectra could be correctly classified, as shown in Table 8.3. In that case, 3287 spectra were considered as poorly classified. If only the top 85% scored independent test spectra were kept, meaning that 1141 spectra are thrown away, 93% epithelium and 92% stroma spectra were correctly classified. The confusion matrix table was shown in Table 8.4. These results show that firstly, stroma spectra were more difficult to classify than epithelium spectra. Secondly, the model was very confident about its decision when only half of the spectra were kept. Even without any threshold, the model performed excellently in the independent test.

Table 8.3 Confusion matrix of independent test on the classification of epithelium and stroma cells keeping half of the spectra based on score ranking using AdaBoost

Independent Test (Threshold Applied)		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	98	2
	Stroma	5	95

Table 8.4 Confusion matrix of independent test on the classification of epithelium and stroma cells keeping 85% of the spectra based on score ranking using AdaBoost

Independent Test (Threshold Applied)		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	93	7
	Stroma	8	92

For better visualisation, the predictions of all the spectra in 10 independent test cores were projected back to their original pixel locations. A projection plot of the epithelium and stroma classification results using AdaBoost was shown in Figure 8.4, where epithelium predictions were in red and green (cancerous and NAT epithelium respectively) and stroma predictions were in purple and yellow (cancerous and NAT stroma respectively).

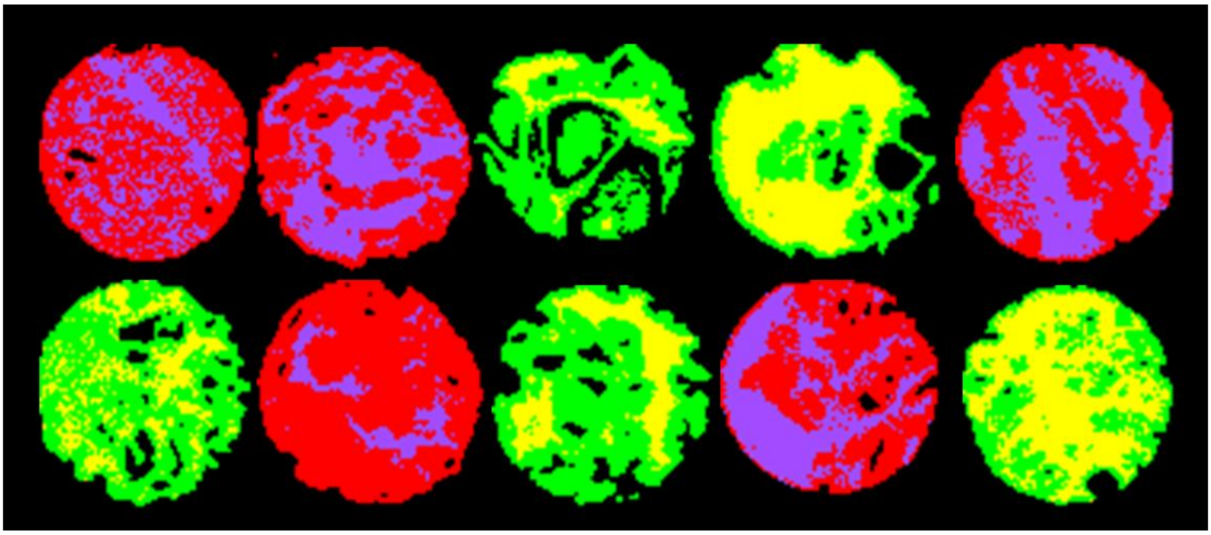


Figure 8.4 Projection plot of the epithelium and stroma classification results of 10 independent test cores using AdaBoost, where red represents cancerous epithelium, purple stands for cancerous stroma, green represents NAT epithelium and yellow shows NAT stroma.

8.3.1.2. Using Random forest classification method

The random forest classification method, with 500 trees, was also applied to separate epithelium and stroma cells. The confusion matrix of the training validation test of the model was shown in Table 8.5, where 98% epithelium cells were correctly classified while 97% stroma cells were correctly classified in the model validation test. The overall classification accuracy was 97.4%.

Table 8.5 Confusion matrix of model validation test on the classification of epithelium and stroma cells using random forest

Training Validation		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	98	2
	Stroma	3	97

When the model was applied to the independent test spectra, epithelium cells were 90% correctly classified while 85% of stroma cells were correctly classified. The same reason mentioned in the previous section, multiple types of cells, was found as the cause of the slightly lower classification accuracy of stroma comparing with the epithelium spectra in the independent test. The confusion matrix of independent test was shown in Table 8.6.

Table 8.6 Confusion matrix of independent test on the classification of epithelium and stroma cells using random forest

Independent Test		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	90	2
	Stroma	15	85

The ROC curve was plotted based on scores of each prediction, shown in Figure 8.5. The AUC of each class was 0.92, which shows that the model works well on the independent test data set. To test whether the classification was valid, the importance of each feature was plotted, and shown in Figure 8.6, which indicates that the classification was mainly caused by the variations of spectra at the peak 3353 and 3360 cm^{-1} were N-H stretching, and the peak 3252 and 3272 cm^{-1} were O-H stretching. Comparing with AdaBoost, three extra main peaks were observed, at 3252, 3272 and 3360 cm^{-1}. These may be caused by different rules for feature processing of two models, as in random forest there was no extra weight added to each feature while in AdaBoost the weights of each feature would be re-evaluated based on training results. After 500 iterations, only 3353 cm^{-1} contributes the least error, in this case. Therefore, AdaBoost keeps it which favours the classification as an important feature and weights down other features [14].

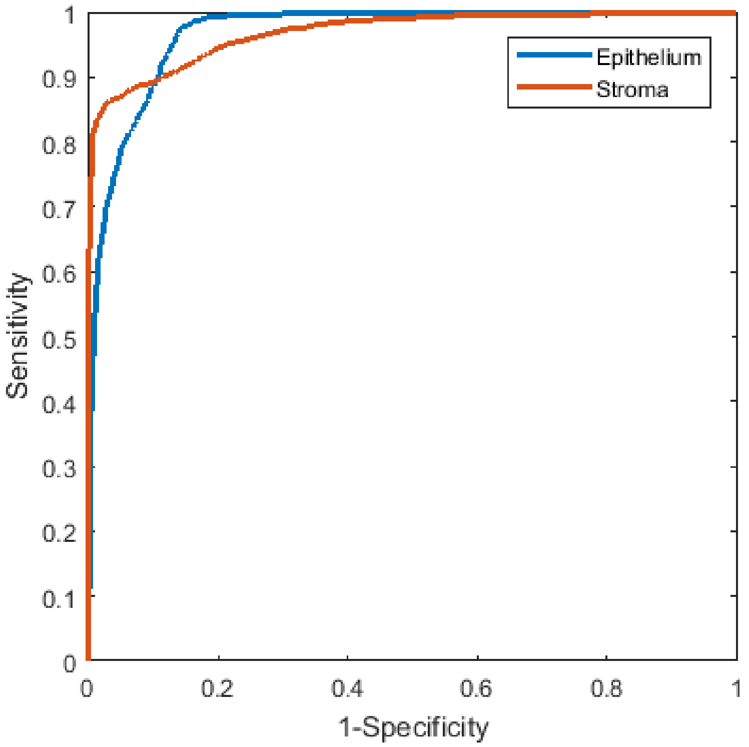


Figure 8.5 ROC curve of independent test on the classification of epithelium and stroma cells using random forest

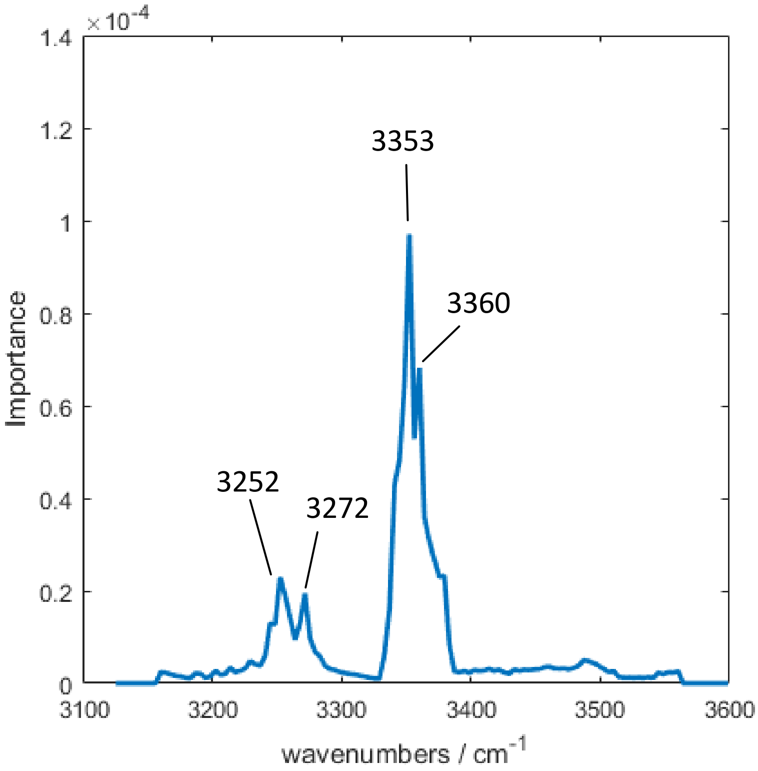


Figure 8.6 The feature importance plot of independent test on the classification of epithelium and stroma cells using random forest

The performance of the model could be further boosted by adding a threshold after conducting classification. Removing low scored spectra could increase the performance of model. If only half of the spectra which were top-rated were kept, 96% epithelium spectra and 94% stroma spectra could be correctly classified, as shown in Table 8.7. In that case, 3568 spectra were considered as poorly classified. If only top the 85% scored independent test spectra were kept, 93% epithelium and 90% stroma spectra were correctly classified. The confusion matrix table was shown in Table 8.8. The random forest model also shows that stroma spectra had slightly lower classification accuracy than epithelium spectra, which was still due to the nature of stroma regions.

Table 8.7 Confusion matrix of independent test on the classification of epithelium and stroma cells keeping half of the spectra based on score ranking using Random forest

Independent Test (Threshold applied)		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	96	4
	Stroma	6	94

Table 8.8 Confusion matrix of independent test on the classification of epithelium and stroma cells keeping 85%of the spectra based on score ranking using Random forest

Independent Test (Threshold applied)		Predicted / %	
		Epithelium	Stroma
True / %	Epithelium	93	7
	Stroma	10	90

Comparing the performance of AdaBoost and random forest classification approach, AdaBoost seems to have slightly better accuracy when it came to independent test in this case. However, only conventional hyperparameters of each model were applied for this research. The main difference causing the slight advantage in classification accuracy of AdaBoost could be that only 3353 cm^{-1}, N-H stretching, was considered to be the key feature of separation, while for random forest both N-H stretching (3353 and 3360 cm^{-1}) and O-H stretching (3252 and 3272 cm^{-1}) were considered as key classification features. AdaBoost requires less number of key features to achieve similar or even better classification results. If only discrete wavenumber and limited number of measurements could be taken, AdaBoost holds its advantage as it needs less number of key feature measurements to produce good classification with this dataset.

Predictions of spectra in the independent test group were projected back to their original pixel locations. A projection plot of the epithelium and stroma classification results using random forest was shown in Figure 8.4, where epithelium predictions were in red and green (cancerous and NAT epithelium respectively) and stroma predictions were in purple and yellow (cancerous and NAT stroma respectively). No large differences were observed between the histological predictions made by two classification methods, shown in Figure 8.4 (AdaBoost) and Figure 8.7 (random forest). It supports the observations found from the classification accuracies presented in confusion matrixes.

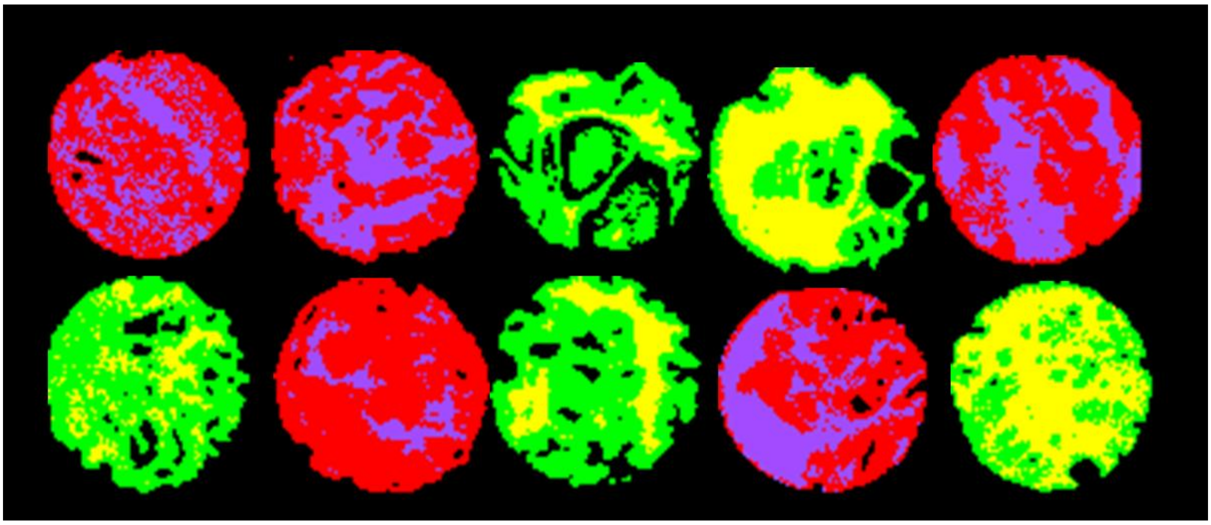


Figure 8.7 Projection plot of the epithelium and stroma classification results of 10 independent test cores using random forest, where red represents cancerous epithelium, purple stands for cancerous stroma, green represents NAT epithelium and yellow shows NAT stroma.

8.3.1.3. Section conclusion

Similar classification results, 98.25% (for epithelium) and 99.94 % (for stroma) when random forest, with a threshold, was applied, were obtained by Bassan *et al*. on breast tissue[2]. Previously, on prostate tissue classification accuracy of 97.27% (for epithelium) and 94.20% (for stroma) on using random forest models with threshold applied were also obtained by Pilling *et al.* [10]. The work presented here shows that using AdaBoost could deliver classification of epithelium and stroma on glass, as well as random forest or even better in the independent test.

Considering the results obtained by previous groups and the results obtained in this research, the histological classification on glass using infrared was completely possible and highly accurate classification rates could be obtained.

8.3.2. Cancer prediction classification results

Attempt to construct classifiers built by combined features of both epithelium and stroma spectra were conducted in this section. Currently in the field, in order to perform cancer diagnosis on prostate tissue, firstly a model needs to be built on separating epithelium and stroma. The spectra predicted as epithelium spectra were then passed to cancer diagnosis model [15]. This two-step process increases the processing time of cancer diagnosis and the classification errors produced by both models. In the clinical field, pathologists often draw the entire area contained cancer instead of individual cells specifically. Inspired by that, models in this section were constructed using spectra from annotated epithelium cells and the adjacent stroma. These models were not restricted by cell types of spectra, which reduce the errors and only require one-step classification, straight forward to use.

For training machine learning models (AdaBoost and random forest), a total of 26686 spectra, which contains 14241 epithelium spectra (containing 5364 cancerous and 8877 NAT spectra) and 12445 stroma spectra (containing 3732 cancerous and 8713 NAT spectra) were extracted from data cubes. In total, in the training spectra set, there were 9096 cancerous spectra and 17590 NAT spectra. Under-sampling was conducted to balance the number of spectra in each class. 9096 NAT spectra were randomly selected from 17590 spectra to match the number of spectra in cancerous class. Each class has the same number of spectra and no bias in either class would be made. 124 features were used to construct models.

For further investigation, to test the influences of human individual differences on cancer diagnosis, another 21 cores were chosen to perform further testing. From the previously selected patients, 11 different cores from the cores used in the model were chosen, and another 10 cores (in total, 20 cores, together with the previously selected independent test cores) from the different patients in the training of models were also picked.

8.3.2.1. Applying AdaBoost

Cancer and NAT prostate spectra classification model was constructed using the training dataset extracted from 90 cores. 500 trees were trained in 500 iterations. The learning rate of the model was set to 0.1, which was a commonly used value in the machine learning model construction [16].

The confusion matrix of the model validation test, in which 10% randomly selected spectra from the training data set was used as test set to validate the model performance, was shown in Table 8.9. 78% cancerous spectra were correctly classified while 85% NAT spectra were correctly classified in model validation test. The overall classification accuracy was 80.9%.

Table 8.9 Confusion matrix of model validation test on the classification of cancerous and NAT spectra using AdaBoost

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	78	22
	NAT	15	85

8.3.2.1.1 Different Cores from previously selected patients

11 sample cores selected from patients previously appeared in the training process but different from the cores used in the process were selected as a test group. 4321 spectra were included, which contains 2151 cancerous spectra (1267 epithelium and 884 stroma spectra) and 2170 NAT spectra (825 epithelium and 1345 stroma spectra).

When AdaBoost was applied, without any threshold, the classification accuracy on cancerous spectra was 72% while the accuracy on NAT spectra was 57%. The confusion matrix of this test was shown in Table 8.10.

Table 8.10 Confusion matrix of test on the classification of epithelium and stroma cells using AdaBoost with additional 11 different cores from same patients with training group

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	72	28
	NAT	43	57

The ROC curve was plotted based on scores of each prediction, shown in Figure 8.8. The AUC of each class was 0.72. The importance of each feature was plotted in Figure 8.9. with peaks at 3333 cm^{-1}, N-H stretching, 3401 cm^{-1}, O-H & N-H stretching vibrations, and 3418, 3453 and 3461 cm^{-1}, O-H stretching.

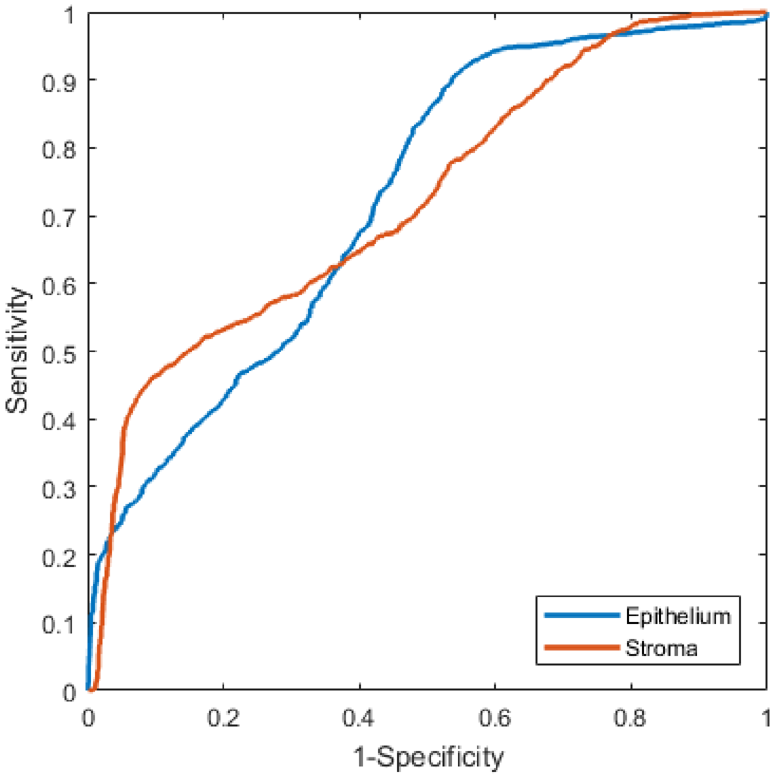


Figure 8.8 ROC curve of test on the classification of cancerous and NAT spectra using AdaBoost with additional 11 different cores from same patients with training group

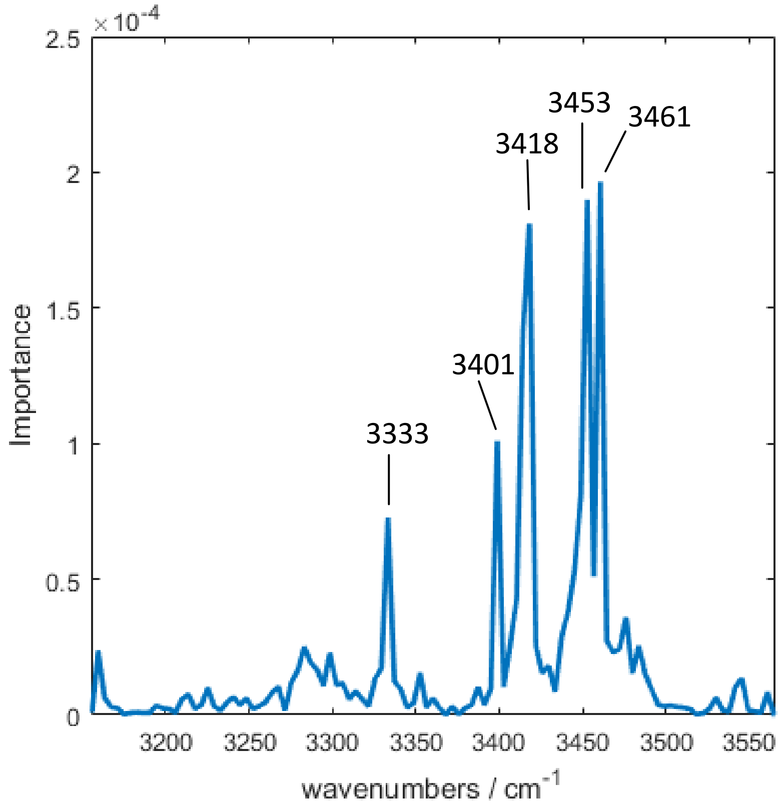


Figure 8.9 The feature importance plot of independent test on the classification of cancerous and NAT spectra using AdaBoost

Keeping half of the spectra which were top-rated, 77% cancerous spectra and 89% NAT spectra could be correctly classified, shown in Table 8.11. In that case, 2527 spectra were considered as poorly classified. If only top 85% scored independent test spectra were kept, 78% cancerous and 66% NAT spectra were correctly classified. The confusion matrix table was shown in Table 8.12.

With the increase of the number of poorly classified spectra, the classification accuracy of cancerous spectra remains the same while the accuracy of NAT spectra increases over 20%. With help from thresholds, the test results of 11 cores became very close to the training validation results which were shown in Table 8.9. This means the hidden patterns of prostate cancer found during training process by the model could also be identified from other cores of these patients. The test set extracted from different cores with previously appeared patients has similar data features and structures with the training data.

Table 8.11 Confusion matrix of test on the classification of cancerous and NAT spectra keeping half of the spectra based on score ranking using AdaBoost with additional 11 different cores from same patients with training group

Independent Test (threshold applied)		Predicted / %	
		Cancer	NAT
True / %	Cancer	78	22
	NAT	11	89

Table 8.12 Confusion matrix of test on the classification of cancerous and NAT spectra keeping 85% of the spectra based on score ranking using AdaBoost with additional 11 different cores from same patients with training group

Independent Test (threshold applied)		Predicted / %	
		Cancer	NAT
True / %	Cancer	78	22
	NAT	34	66

8.3.2.1.2 Independent Test with 10 cores

For the independent test with 10 cores, 6647 spectra were pulled from data cubes, including 3077 cancerous spectra (1510 epithelium and 1567 stroma spectra) and 3570 NAT spectra (1920 epithelium and 1650 stroma spectra).

When independent test spectra exacted from independent test cores were applied, without applying any threshold, cancerous spectra were 71% correctly classified while 62% NAT spectra were correctly classified. NAT spectra were more difficult to classify, which could because that the general features among patients with different physical conditions (i.e. health conditions, age and other factors) could be more difficult to find compared with finding the general features of prostate cancer. The human individual difference could be a potential barrier for cancer diagnosis. The confusion matrix of independent test was shown in Table 8.13.

Table 8.13 Confusion matrix independent test on the classification of cancerous and NAT spectra using AdaBoost

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	71	29
	NAT	38	62

The ROC curve of the independent test of the trained model was plotted in Figure 8.10. The AUC of each class was 0.67. The importance of each feature was plotted in Figure 8.9. Figure 8.9, it indicates that the classification was mainly caused by the variations of spectra at the peak 3333 cm^{-1}, N-H stretching, 3401 cm^{-1}, O-H & N-H stretching vibrations (hydrogen bonding network may vary in the malignant tissue) [17], and 3418, 3453 and 3461 cm^{-1}, O-H stretching.

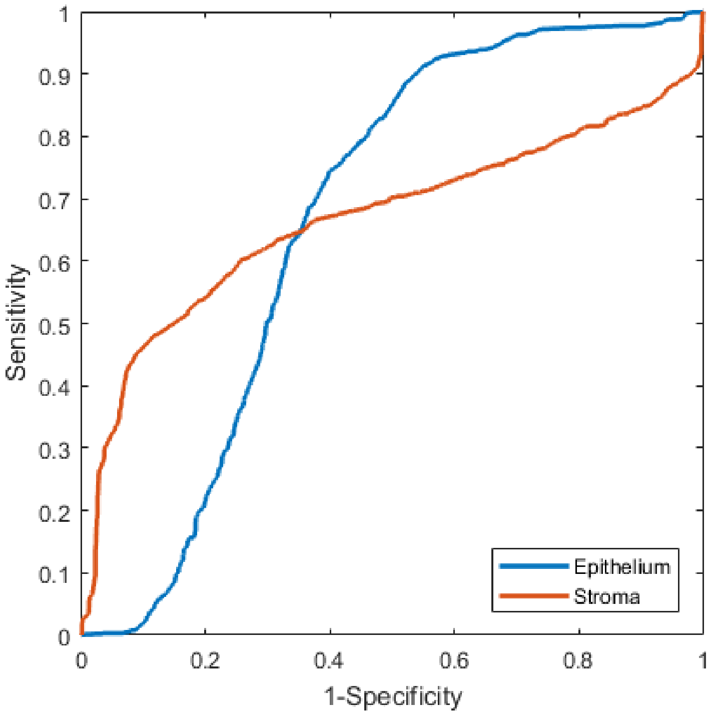


Figure 8.10 ROC curve of independent test on the classification of cancerous and NAT spectra using AdaBoost

The performance of the model could be further boosted by adding a threshold after conducting classification. Half of the spectra which were top-rated were kept, 72% cancerous spectra and 68% NAT spectra could be correctly classified, shown in Table 8.14. In that case, 3230 spectra were considered as poor classified. If only top

85% scored independent test spectra were kept, 76% cancerous and 62% NAT spectra were correctly classified. The confusion matrix table was shown in Table 8.15. With the increase of number of poorly classified spectra, the classification accuracy of cancerous spectra decreased while the accuracy of NAT spectra increases, which means that the model was not sure about its classifications. Most of the scores for spectra were relatively low. Another possible reason could be that the model captured features only occurred in the training group (noise or similarities not related to cancer diagnosis) and weighted these feature, which causes negative effect in the independent test group. It was considered as the main disadvantage of AdaBoost classification method [13].

Table 8.14 Confusion matrix of independent test on the classification of cancerous and NAT spectra keeping half of the spectra based on score ranking using AdaBoost

Independent Test (Threshold applied)		Predicted / %	
		Cancer	NAT
True / %	Cancer	72	28
	NAT	32	68

Table 8.15 Confusion matrix of independent test on the classification of cancerous and NAT spectra keeping 85% of the spectra based on score ranking using AdaBoost

Independent Test (Threshold applied)		Predicted / %	
		Cancer	NAT
True / %	Cancer	76	24
	NAT	38	62

8.3.2.1.3 Independent Test with 20 cores

20 patients different from the previous training process were selected including the previously 10 cores in the independent test. The size of an additional independent test was doubled to test with increasing human individual differences involved would the independent test results remain at a similar level. Additional 5892 spectra were added to the previous independent test dataset, which contained 2568 cancerous spectra (1285 epithelium and 1283 stroma spectra) and 3324 NAT spectra (1821 epithelium and 1503 stroma spectra).

When AdaBoost was applied, without any threshold, the classification accuracy on cancerous spectra was 40% while the accuracy on NAT spectra was 58%. The confusion matrix of this test was shown in Table 8.16. The model was in favoured of classifying spectra as NAT spectra and fails to identify cancerous spectra. The cancerous class identification accuracy was largely reduced compared with the independent test conducted on 10 cores (shown in Table 8.13). One possible reason for this observation could be that the smaller group of patients have more similarities with the training group compared with the larger independent patient group. The classification accuracy observed previously could be based on the similarities other than prostate cancer properties between training and independent test groups. With another independent group, these similarities were not detected or reduced, which leads to the poor classification on the cancerous class.

Table 8.16 Confusion matrix of independent test on the classification of epithelium and stroma cells using AdaBoost with additional 20 cores from different patients with training group

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	40	60
	NAT	42	58

The ROC curve was plotted based on scores of each prediction, shown in Figure 8.11. The AUC of each class was 0.47. The ROC curve supports the results observed in the independent test. Most of the parts of the ROC curves were below the 50% classification line (the diagonal line from bottom left corner to the top right corner), which indicates that the model acts poorly on the independent test group.

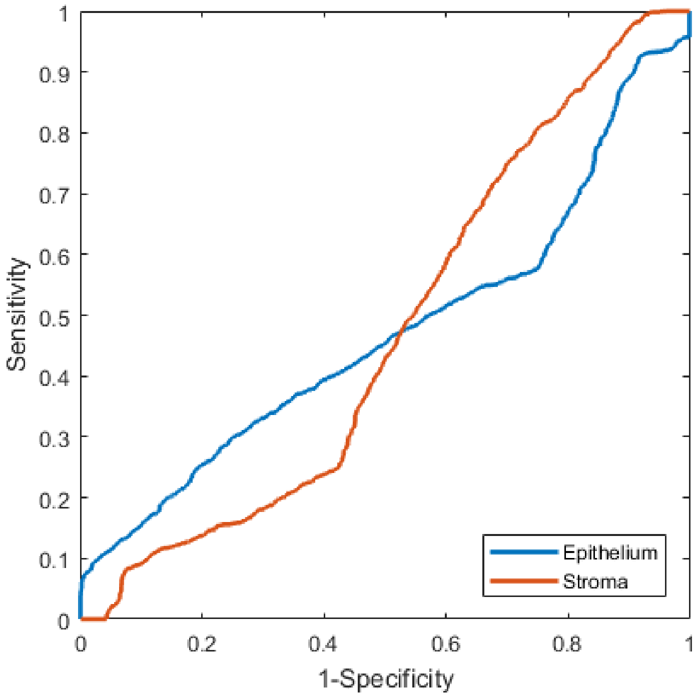


Figure 8.11 ROC curve of independent test on the classification of cancerous and NAT spectra using AdaBoost with additional 20 cores from different patients with training group

The threshold of keeping half of the spectra based on score ranking was applied to the classification results and shown in Table 8.17. Only 32% cancerous spectra and 67% NAT spectra could be correctly classified. In that case, 6274 spectra were considered as poor classified. The classification failure becomes even worse after applying the threshold, which meant that the model could not identify the cancerous spectra in the independent test group. It was confidently making wrong decisions. The main features found by the classifier were not sufficient to separate two classes when it came to the independent test group. There were many possible reasons for the poor performance. One of them was that the main features selected from the training group were not or fully related with prostate cancer. Some general features in each class could be caused by other common individual conditions, for example, age and other physical conditions. Another one could be that the number of feature input was too less. The information contained was not enough to separate cancerous from normal prostate spectra as only high wavenumber range was applied. Lack of data input could not be compensated by data mining at this level.

Table 8.17 Confusion matrix of independent test on the classification of cancerous and NAT spectra keeping half of the spectra based on score ranking using AdaBoost with additional 20 cores from different patients with training group

Independent Test (Threshold applied)		Predicted / %	
		Cancer	NAT
True / %	Cancer	32	68
	NAT	33	67

8.3.3.1.4 Classification per core

Classification results conducted using AdaBoost method were further integrated and presented in terms of individual cores, which could provide better understandings of the results in terms of visualising spectral locations.

Classified spectra were traced back into their pixel positions in each core, which were shown in Figure 8.12. Inside the marked box were 20 cores from different patients, while the cores, not in the box, were other cores from the same patients with the training group. The dark blue area of each core was made of pixels contained spectra passed quality checked and were used in the independent test. The pale blue area was made of pixel contained spectra were correctly classified in the independent test. It could be seen that there were some cores and certain regions on the cores were poorly classified. This may be raised by inaccurate annotation or strong individual differences between training data and test data.

When only half of the spectra were kept based on score ranking from high to low, the pale blue areas indicate the corrected classified pixels while the dark blue regions show the spectra passed threshold conditions. This was plotted in Figure 8.13. Again cores in the box were the different patient group while the others were from the same patients in the training group. Other colours (red, purple, green and yellow) were annotation colours for both epithelium and stroma from cancerous and NAT cores. For better comparison, the original annotations of these cores were shown in Figure 8.14. When standards were raised, the model was very confident on some cores where pale blue colour was shown concentrated in certain areas, which may contain more similar spectral patterns found in the training group.

These confident classification areas would increase the classification accuracy of individual cores.

After applying a threshold, it seems like the classification accuracy was boosted. However, the usage of core information was actually reduced as not every core had highly scored spectra during classification. With a high level of threshold, the model focused more on the cores which were easier to identify while giving up on the cores which were harder to separate. These difficult cores were actually the most challenging part of the research as these cores could be potentially difficult to identify for pathologist as well, which would thus make a good classification model useful to the current cancer diagnosis flow. Therefore, when a threshold was applied, the user should be cautious about how much information could be discarded from the data.

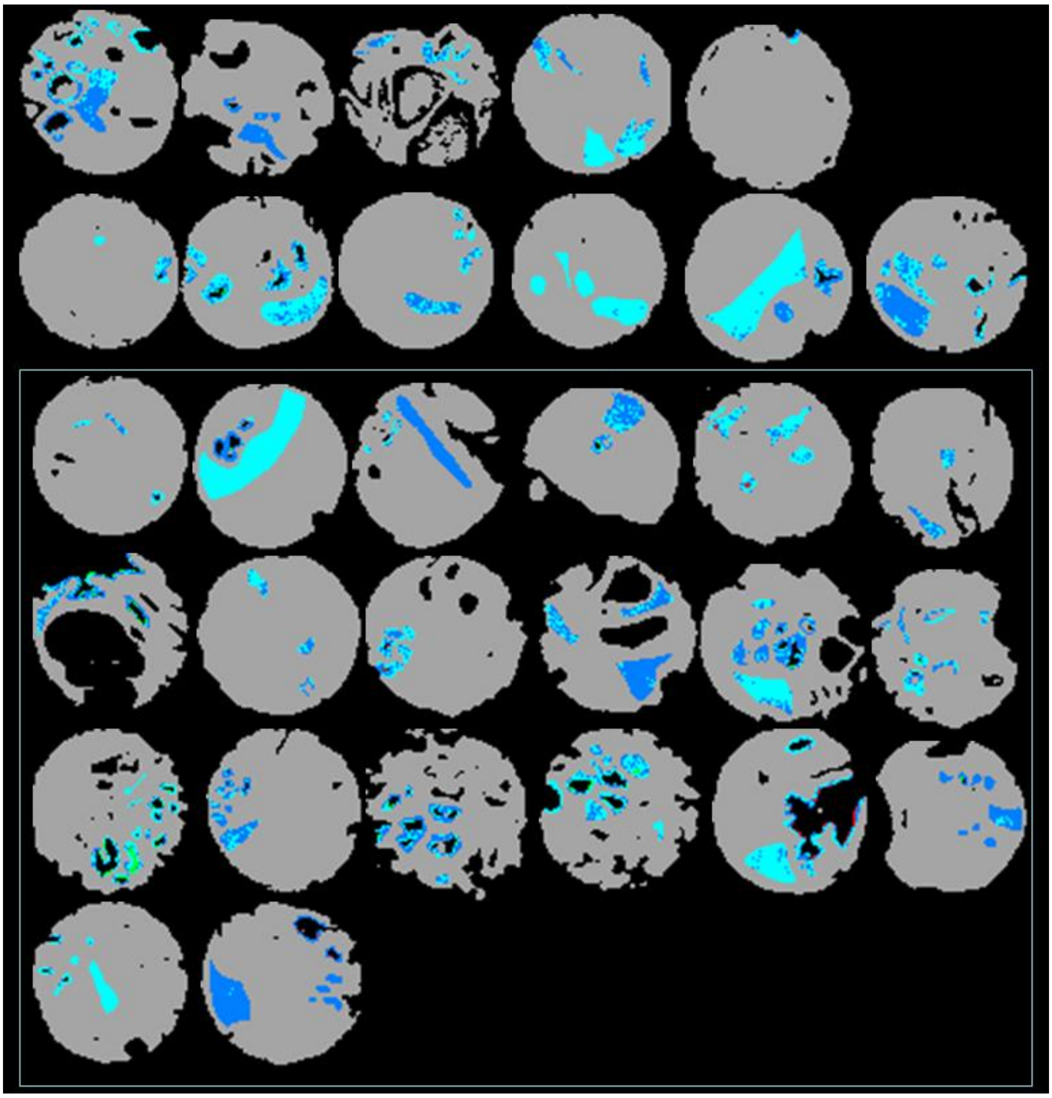


Figure 8.12 classification results of 31 cores without any threshold using AdaBoost, where the dark blue area of each core was made of pixels contained spectra pre-processed before classification and the pale blue area was made of pixel contained spectra were correctly classified in the independent test.

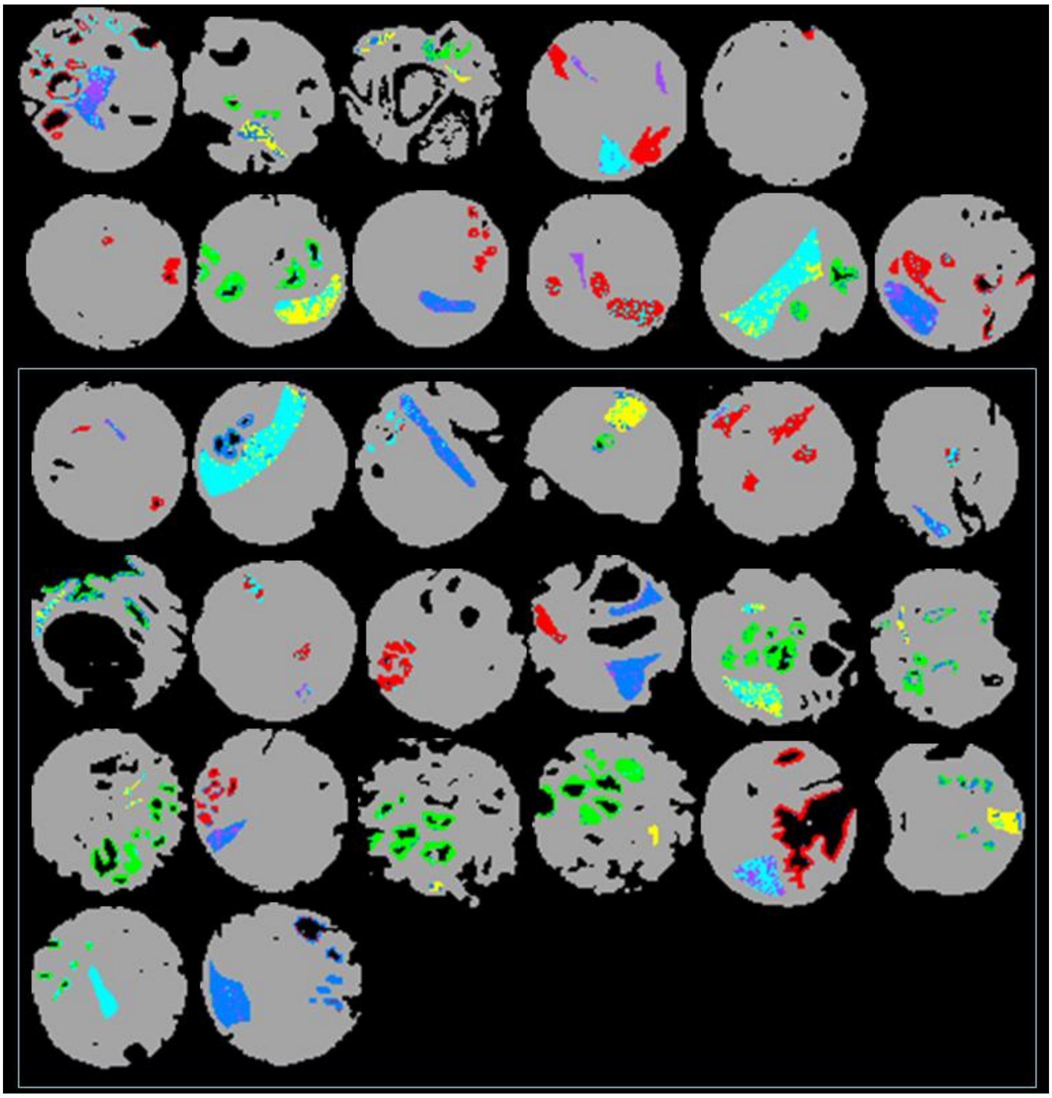


Figure 8.13 classification results of 31 cores without any threshold using AdaBoost, where dark blue area of each core was made of pixels contained spectra pre-processed before classification, the pale blue area was made of pixel contained spectra were correct, red represents cancerous epithelium annotation, purple stands for cancerous stroma annotation, green shows NAT epithelium annotation and the yellow indicates NAT stroma annotation

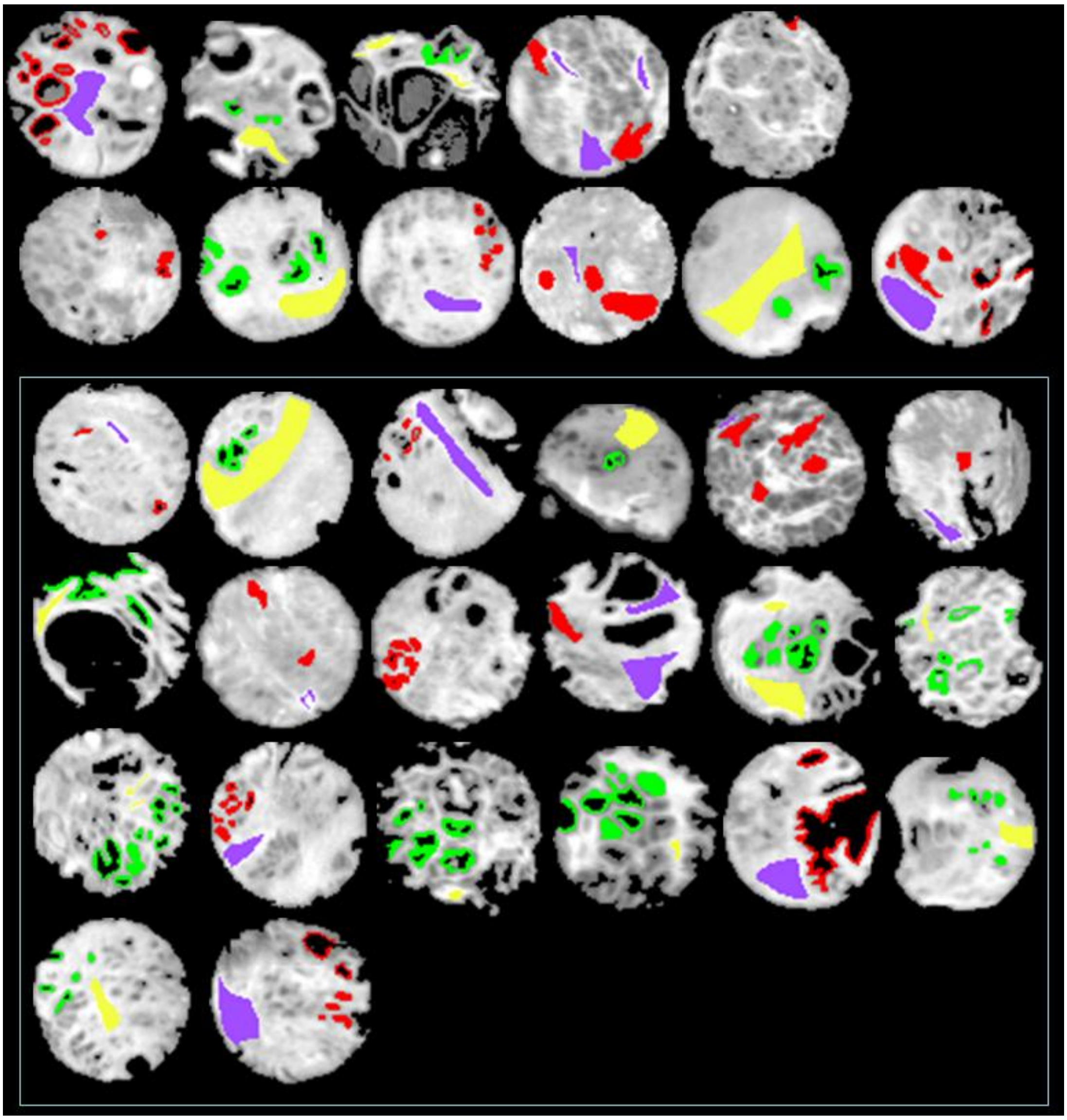


Figure 8.14 Annotations all 31 cores used for the test, where red represents cancerous epithelium annotation, purple stands for cancerous stroma annotation, green shows NAT epithelium annotation and the yellow indicates NAT stroma annotation

When an individual core was considered as a whole, a cancer diagnosis could be simplified by applying a traffic light system, which indicates any cancer core as red cores, any NAT cores as green and any not determined cores as yellow cores. The advantage of this system would be that it reduces the workload for pathologist as only yellow cores needed further investigation if the classification model could provide classification accuracy higher than human cancer diagnosis accuracy. According to recent research using fine-needle aspiration and core biopsy had an overall accuracy of 75.4% according to research conducted in 2010 by Kasraeian S, *et al*. [18] and the grading accuracy of pathologist was 76% according to research conducted in 2011 by Barqawi *et al*. [19].

Two result counting methods were applied after classification. The first one using spectra from the annotated area of independent test cores. Only the top-scored 10% spectra were kept for each core and among these spectra if one or more than one spectrum was classified as cancerous spectrum, this core would be classified as a cancerous core and coloured in red. If no spectrum was classified as cancerous ones and the number of diagnosed NAT spectra was larger than half of the total number of selected spectra, the tested core would be classified as NAT core and coloured in green. In any other cases, the cores would be coloured in yellow. The coloured cores of both other cores from the same patients and cores from different patients in the training group were shown in Figure 8.15 . In this plot, cores in the box were cores from different patients while the rest of cores were other cores from the same patients in the training process. The classification accuracy for cores from the same patients was 7 out of 11, which was 64%, while the accuracy for different patients was 10 out of 20, which was 50%. These accuracies were similar to the previously calculated classification accuracies in the unit of spectrum other than individual core.

The second method uses the previously constructed epithelium-stroma classification model as a filter, which only keeps the top 50% scored spectra after classification for both types of spectra. All these spectra, both epithelium and stroma, were then sent to the cancer NAT classification model, and only the top 10% scored spectra and their scores were used for class assignments. If the number of diagnosed cancerous spectra was larger than the number of classified NAT spectra, with the condition that the number of cancerous spectra was larger than half of the total number of kept spectra, the core was picked as a cancerous core. If the number of diagnosed cancerous spectra was less than the number of diagnosed NAT spectra and the number of NAT spectra was larger than half of the total number of kept spectra, the core was named as a NAT core. In any other cases, the core was coloured in yellow which means not determined and requires further investigations. For AdaBoost, the classification accuracy for cores from the same patients was 7 out of 11, which was 64%, while the accuracy for different patients was 9 out of 20, which was 45%. The detailed plot was shown in Figure 8.16. Similar accuracies were observed with the previous classification results in term of the spectrum.

The first method focused on finding cancerous spectra. As long as there was one top-scored spectrum was considered as cancerous spectrum, the entire core would be diagnosed as cancerous core. Although this method maximised the possibility of obtaining true positive cancer diagnosis, still it reduces the chance of a core being classified as NAT and could potentially increase the number false-positive classification of cancerous cores. This method was reasonable if the research was aimed at finding all the cancerous cores without any missing any one even if there might be NAT cores wrongly classified as cancerous ones.

The second method was a standard way of counting votes, in which the majority wins. It was a more balanced method for both true positive and true negative for cancer diagnosis. Apart from the counting methods, the sources of tested spectra were different. For the first method, spectra were collected from the annotated regions of cores, which were considered having a better chance to provide cancer diagnosis information. In addition, the second method shows the disadvantage of using multiple classification models one after one, as the error rates accumulate in each classification step.

Overall, these two methods obtain similar classification results. Considering individual core as a unit increases the classification accuracy, though into a small extent in a small number of samples. The research focus should be still on the classification model itself rather than result counting since classification model was the foundation of the whole research.

Comparing results obtained from other cores from the same patients as used for training and cores from different patients the classification accuracy from the same patients was 15% to 20% higher than classification accuracy from different patients. This indicates that either the classification considered common features of the training patients as the features of prostate cancer or the individual differences among people were difficult to overcome with limited number of patients. In order to obtain better classification accuracy, a larger training data set was required.

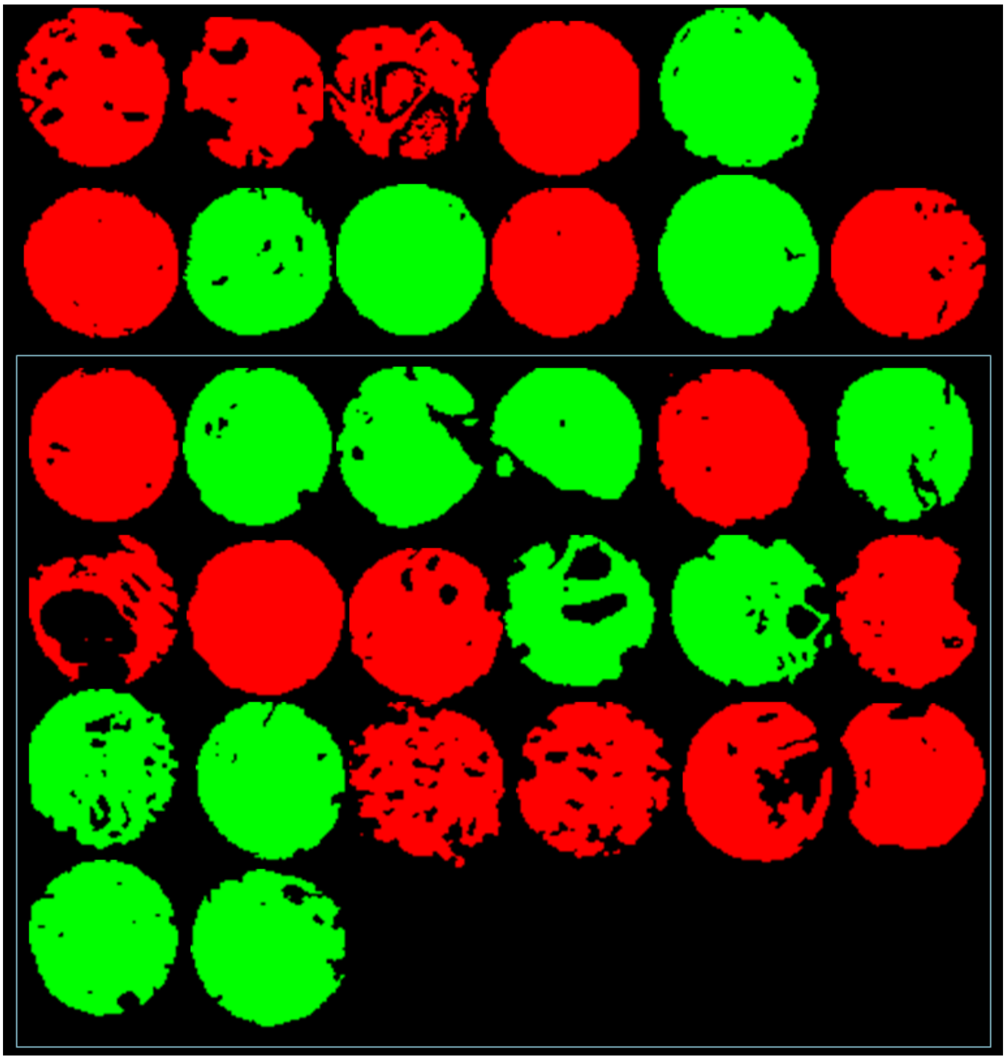


Figure 8.15 classification results presented in the traffic light system with the integration method one using AdaBoost, where red stands for cancer prediction and green indicates non-cancer prediction

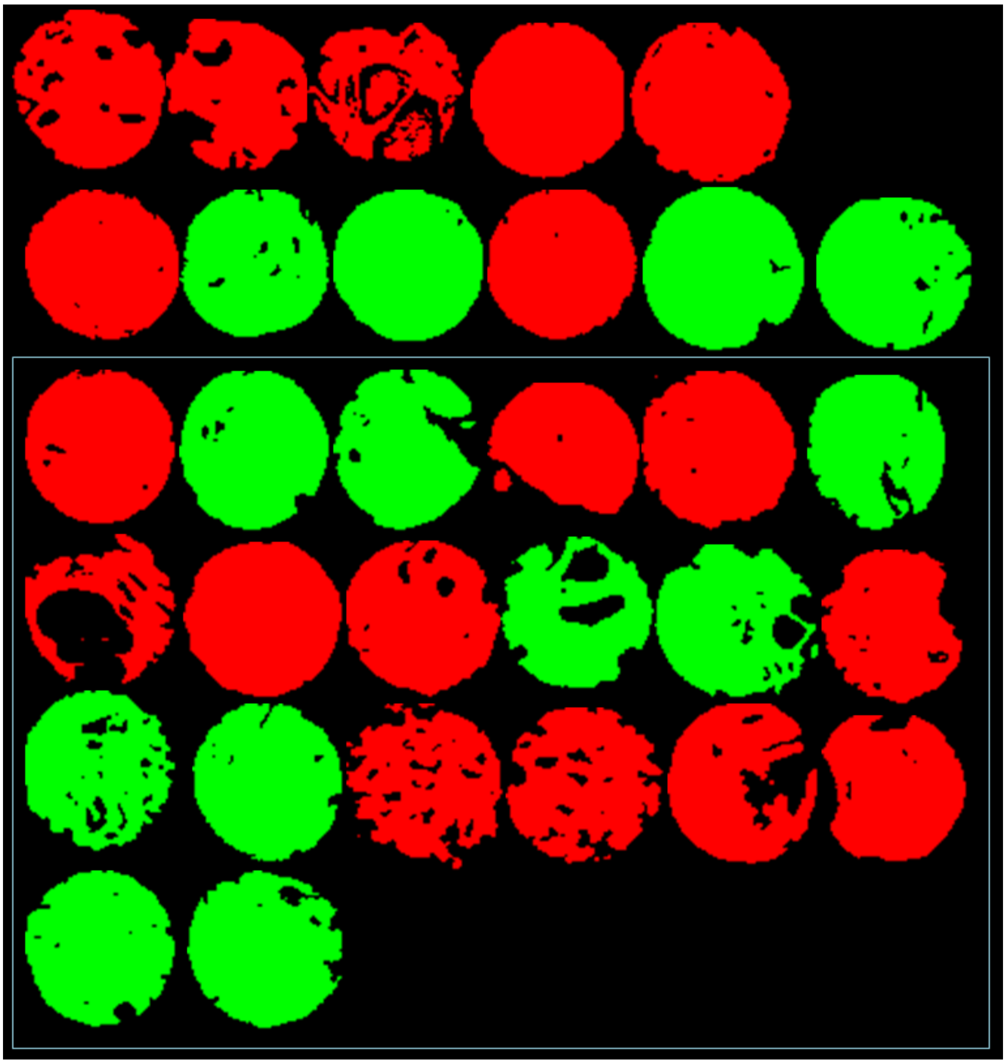


Figure 8.16 classification results presented in the traffic light system with the integration method two using AdaBoost, where red stands for cancer prediction and green indicates non-cancer prediction

8.3.2.2. Applying Random Forest

Applying random forest to the problem of separating cancerous and NAT spectra, using 500 decision trees, the confusion matrix of the training validation test, in which 10% randomly selected spectra from training data set was used as the test set to validate the model performance on training data, was shown in Table 8.18. 92% cancerous spectra were correctly classified while 93% of NAT spectra were correctly classified in the model validation test. The overall classification accuracy was 92.1%. This training validation test has higher classification rates compared with AdaBoost, as for each true positive rate it was 15% higher.

Table 8.18 Confusion matrix of model validation test on the classification of cancerous and NAT spectra using random forest

Training Validation		Predicted / %	
		Cancer	NAT
True / %	Cancer	92	8
	NAT	7	93

8.3.2.2.1 Different Cores from previously selected patients

When the random forest was applied, without any threshold, the classification accuracy on cancerous spectra was 64% while the accuracy on NAT spectra was 59%. The confusion matrix of this test was shown in Table 8.19. Comparing the previous independent test results, the classification accuracy increases, especially for cancerous class (nearly 10%). It meant that some general features collected by the model were the non-cancer related similarities among training patients. As the test group carried the same or part of these similarities, the classification results were higher than the previous independent test.

Table 8.19 Confusion matrix of test on the classification of epithelium and stroma cells using random forest with additional 11 different cores from same patients with training group

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	64	36
	NAT	41	59

The ROC curve was plotted based on scores of each prediction, shown in Figure 8.17. The AUC of each class was 0.65. The importance of each feature was plotted in Figure 8.18 with peaks at 3418 and 3453 cm^{-1}, which were both O-H stretching.

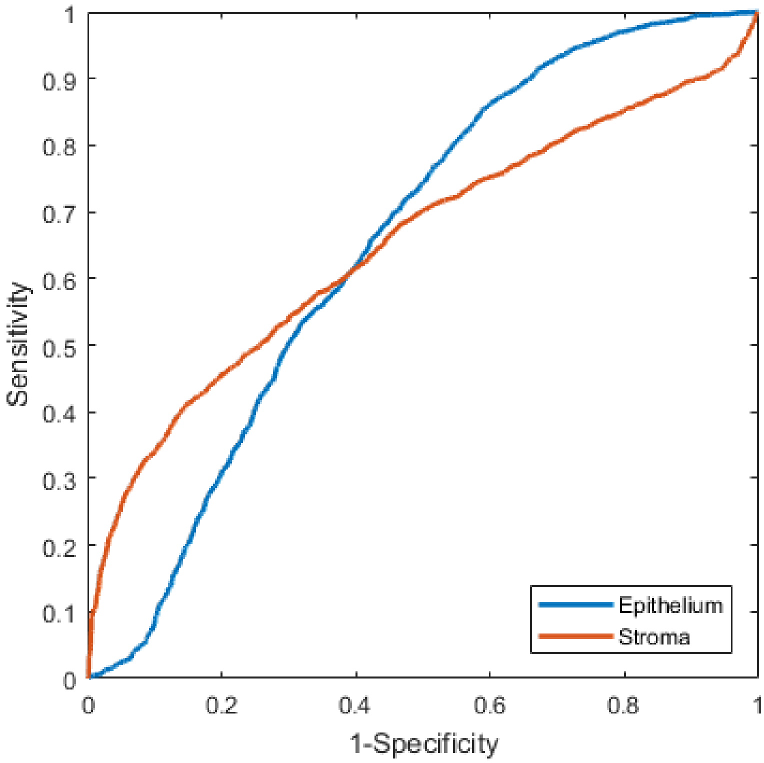


Figure 8.17 ROC curve of test on the classification of cancerous and NAT spectra using random forest with additional 11 different cores from same patients with training group

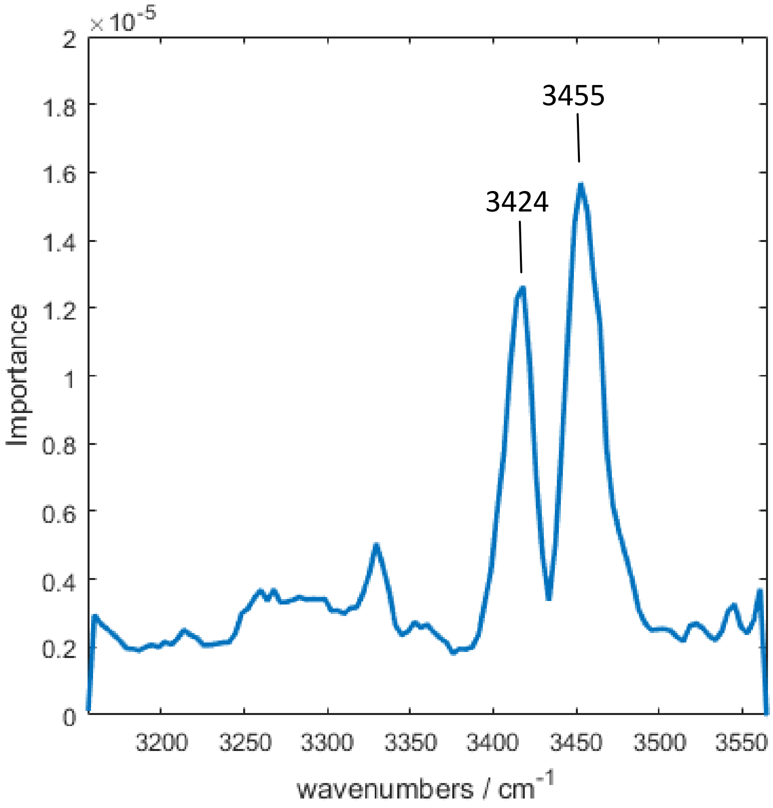


Figure 8.18 feature importance plot of independent test on the classification of cancerous and NAT spectra using random forest

A threshold was applied by removing half of the spectra with low scores. 68% cancerous spectra and 68% NAT spectra could be correctly classified, shown in Table 8.20. In that case, 2290 spectra were considered as poor classified. If only the top 85% scored test spectra were kept, 64% cancerous and 59% NAT spectra were correctly classified. The confusion matrix table was shown in Table 8.21. With the increase of number of poorly classified spectra, the classification accuracy of cancerous spectra decreases while the accuracy of NAT spectra slightly increases.

Table 8.20 Confusion matrix of test on the classification of cancerous and NAT spectra keeping half of the spectra based on score ranking using random forest with additional 11 different cores from same patients with training group

Independent Test (Threshold applied)		Predicted / %	
		Cancer	NAT
True / %	Cancer	68	32
	NAT	32	68

Table 8.21 Confusion matrix of test on the classification of cancerous and NAT spectra keeping 85% of the spectra based on score ranking using random forest with additional 11 different cores from same patients with training group

Independent Test (Threshold applied)		Predicted / %	
		Cancer	NAT
True / %	Cancer	64	36
	NAT	41	59

8.3.2.2.2 Independent Test with 10 cores

When independent test spectra were applied, without any threshold, cancerous spectra were 53% correctly classified while 63% NAT spectra were correctly classified. The model finds general features among patients slightly better than finding the features for prostate cancer. However, the overall classification accuracies were not satisfactory enough. The model struggles to cope with the independent test. Combining with the great result obtained in the training validation test, one possible conclusion could be that model was over-fitting itself. The confusion matrix of independent test was shown in Table 8.22.

Table 8.22 Confusion matrix of independent test on the classification of cancerous and NAT spectra using random forest

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	53	47
	NAT	37	63

The ROC curve was plotted based on scores of each prediction, shown in Figure 8.19. The AUC of each class was 0.62. The importance of each feature was plotted, and shown in Figure 8.18. In Figure 8.18, it was shown that the classification was mainly caused by the variations of spectra at the peak 3424 and 3455 cm^{-1}, O-H stretching. Comparing the main features selected by the random forest model and the AdaBoost model, AdaBoost model seems to have more complex feature importance plot than random forest, which could possibly explain the slightly better performance of AdaBoost in the independent test.

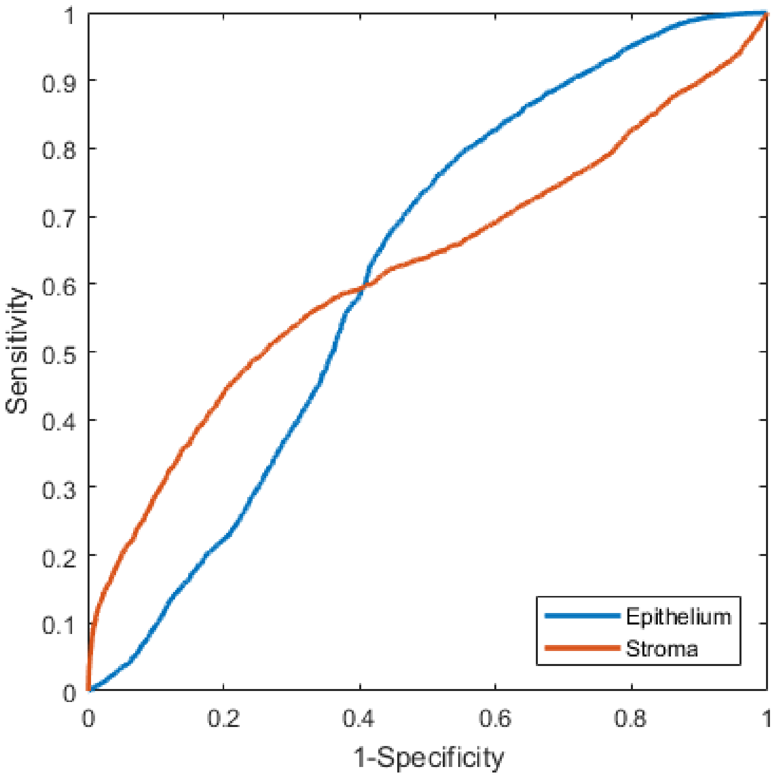


Figure 8.19 ROC curve of independent test on classification of cancerous and NAT spectra using random forest

The performance of the model could be further boosted by adding a threshold after conducting classification. Half of the spectra which were top-rated, in this case, were kept, 53% cancerous spectra and 70% NAT spectra could be correctly classified, shown in Table 8.23. In that case, 3230 spectra were considered as poor classified. If only top 85% scored independent test spectra were kept, 54% cancerous and 65% NAT spectra were correctly classified. The confusion matrix table was shown in Table 8.24. With the increase of number of rejected spectra, the classification accuracy of both cancerous and NAT spectra increases slightly, this means that it was close to the limits of this model. There was no classification determining features observed and most of the scores for spectra were relatively low.

Table 8.23 Confusion matrix of independent test on the classification of cancerous and NAT spectra keeping half of the spectra based on score ranking using Random forest

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	53	47
	NAT	30	70

Table 8.24 Confusion matrix of independent test on the classification of cancerous and NAT spectra keeping half of the spectra based on score ranking using Random forest

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	54	46
	NAT	35	65

8.3.2.2.3 Independent Test with 20 cores

When the random forest was applied, without any threshold, the classification accuracy on cancerous spectra was 35% while the accuracy on NAT spectra was 63%. The confusion matrix of this test was shown in Table 8.25.

The random forest model favoured classifying spectra as NAT spectra and failed to identify cancerous spectra. Similar results were observed with AdaBoost when it came to doubled sized independent test (shown in Table 8.16). The cancerous class identification accuracy was also reduced compared with the independent test conducted on 10 cores (shown in Table 8.22). The same reason for the failure of the AdaBoost model could behind the failure of using random forest as well. The smaller group of patients had more similarities with the training group compared with the larger independent patient group. The classification accuracy observed previously could be based on the non-cancer related similarities between training and independent test groups. With a different group, these similarities were not detected or reduced, which leads to the poor classification on the cancerous class.

Table 8.25 Confusion matrix of independent test on the classification of cancerous and NAT spectra using random forest with additional 20 cores from different patients with training group

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	35	65
	NAT	37	63

The ROC curve was plotted based on scores of each prediction, shown in Figure 8.20. The AUC of each class was 0.47. The ROC curve supports the results observed in the independent test. Most of the parts of the ROC curves were below the 50% classification line, which indicates that the model acts poorly on the independent test group.

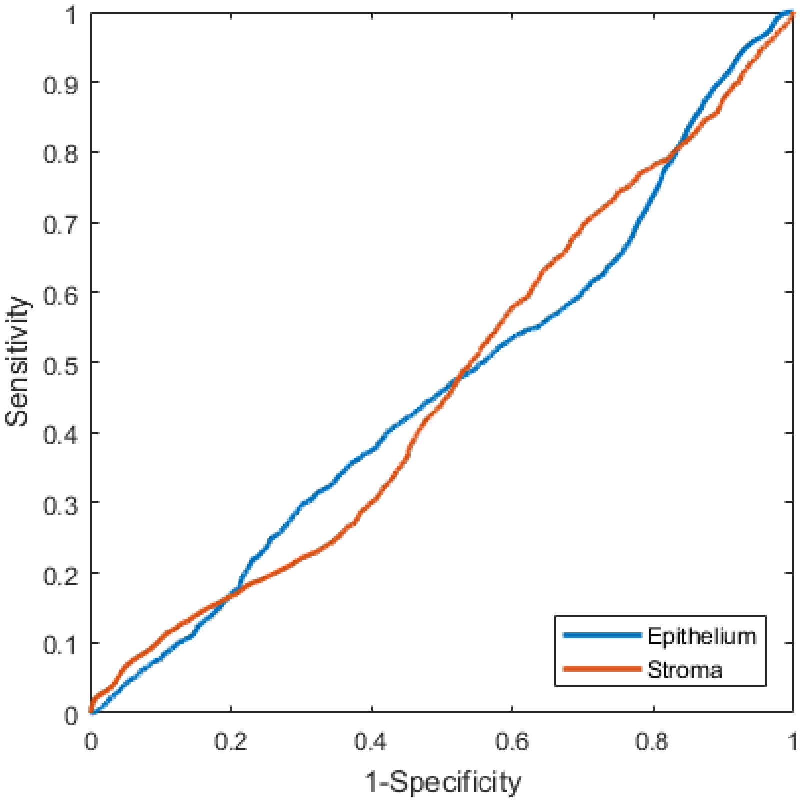


Figure 8.20 ROC of independent test on the classification of cancerous and NAT spectra using random forest with additional 20 cores from different patients with training group

Removing low scored spectra by applying threshold and keeping half of the spectra which was top rated. 35% cancerous spectra and 63% NAT spectra could be correctly classified, shown in Table 8.26. In that case, 6274 spectra were considered as poor classified. The classification does not change after applying the threshold, which the model already reaches its limits. Its performance cannot be boosted anymore.

Table 8.26 Confusion matrix of independent test on the classification of cancerous and NAT spectra keeping half of the spectra based on score ranking using random forest with additional 20 cores from different patients with training group

Independent Test		Predicted / %	
		Cancer	NAT
True / %	Cancer	35	65
	NAT	37	63

8.3.3.2.4 Classification per core

Classification results conducted using random forest method were further integrated and presented in the unit of single core, which visualises the classification results.

Tested spectra were traced back into their original pixel positions in each core, which were shown in Figure 8.21. The dark blue area of each core was made of pixels contained spectra that passed the quality check and were used in the independent test. The pale blue area was made of pixels that contained spectra that were correctly classified in the independent test. It could be seen that there were some cores and certain regions on the cores which were poorly classified. This may be raised by inaccurate annotation.

After applying the 50% threshold (only keeping half of the spectra based on score ranking from top to bottom), the pale blue areas indicate the correctly classified pixels while the dark blue regions show the spectra which were wrongly classified, Figure 8.22. Other colours (red, purple, green and yellow) were annotation colours for both epithelium and stroma from cancerous and NAT cores. For better comparison, the original annotations of these cores were shown in Figure 8.23. When highly scored spectra were considered, the model was confident on some cores where pale blue colour was shown concentrated in certain areas.

After applying the threshold, similar observations were obtained with AdaBoost, increase of classification accuracies. However, the usage of core information was actually reduced as not every core has many highly scored spectra present. With high level of threshold, the model focuses more on some cores which were easier to identify while giving up on the cores which were harder to separate. These cores which were difficult to be identified were actually the core part of the research as these cores could be potentially difficult to identify for pathologist as well. Therefore, when a threshold was applied, the user should be cautious about how much information could be discarded.

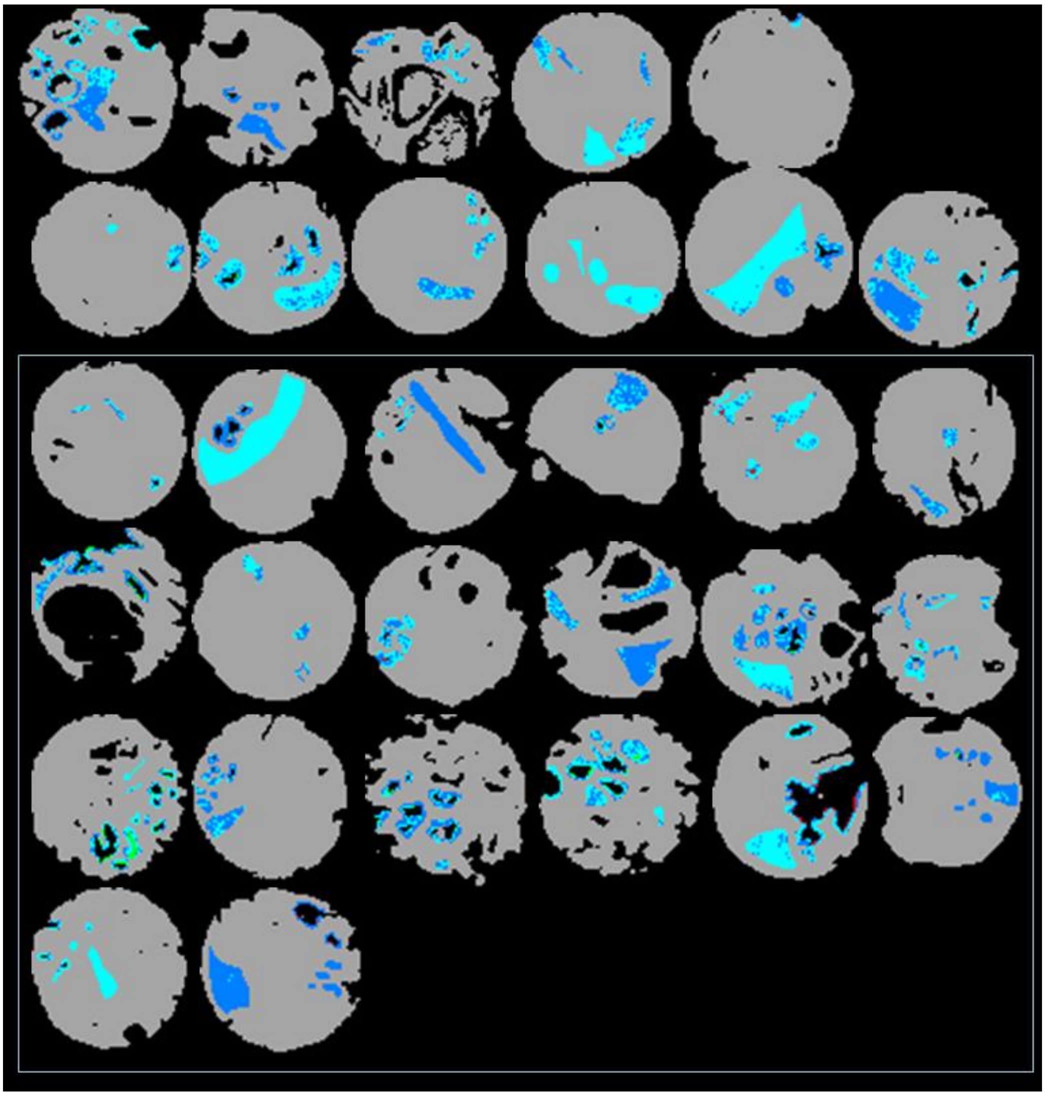


Figure 8.21 Classification results of 31 cores without any threshold using random forest, where the dark blue area of each core was made of pixels contained spectra pre-processed before classification and the pale blue area was made of pixel contained spectra were correctly classified in the independent test.

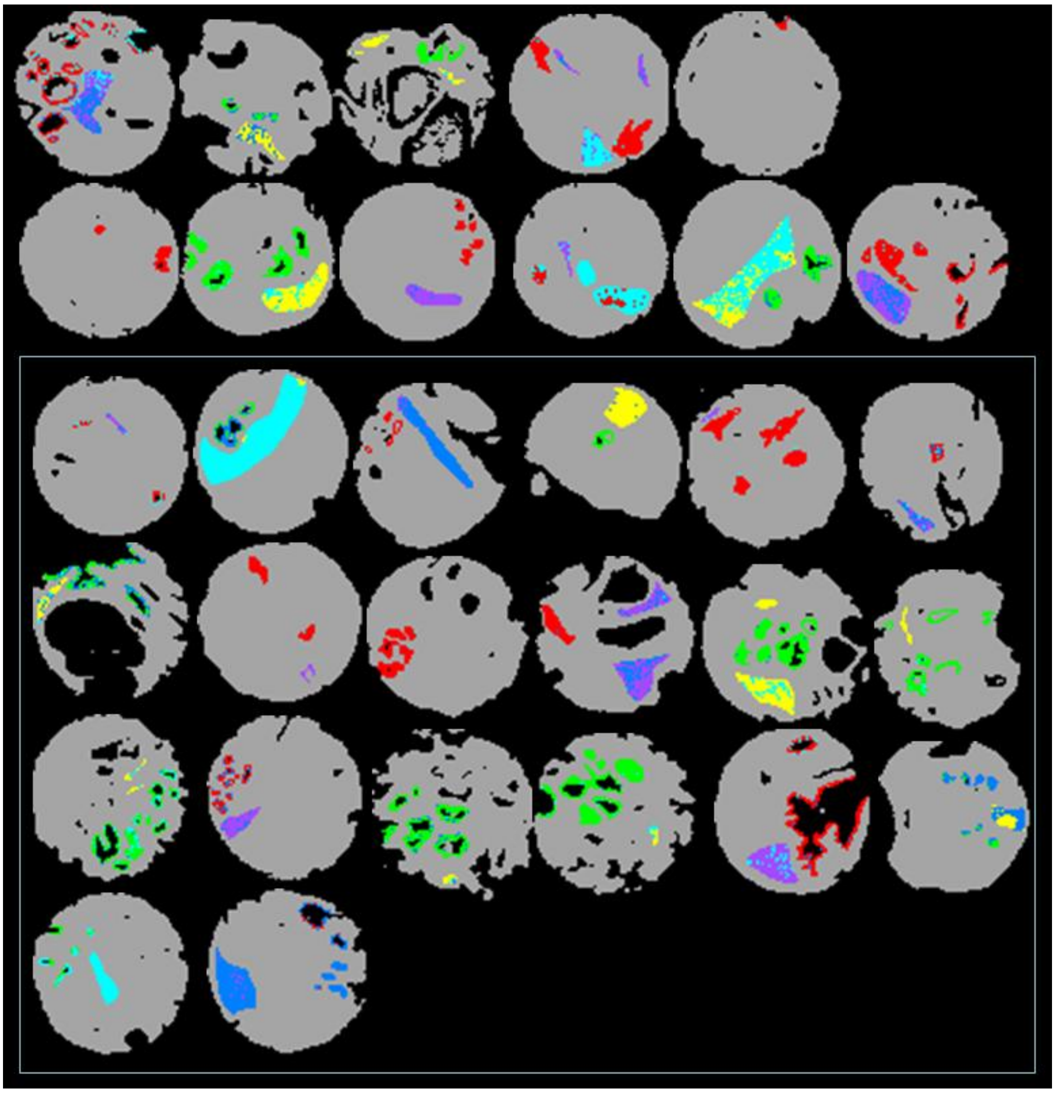


Figure 8.22 Classification results of 31 cores without any threshold using random forest, where dark blue area of each core was made of pixels contained spectra pre-processed before classification and the pale blue area was made of pixel contained spectra were correct, red represents cancerous epithelium annotation, purple stands for cancerous stroma annotation, green shows NAT epithelium annotation and the yellow indicates NAT stroma annotation

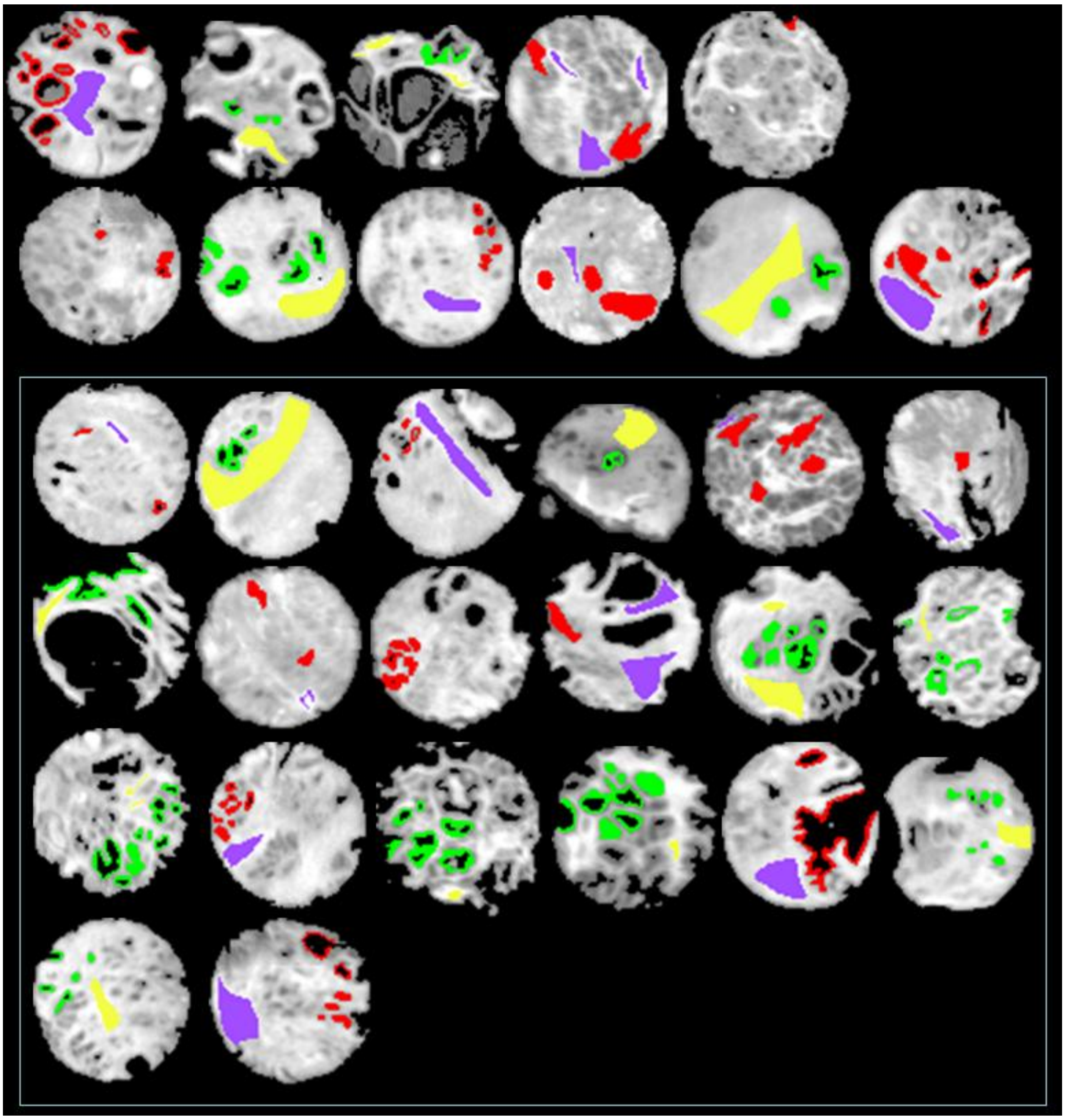


Figure 8.23 Annotations all 31 cores used for the test, where red represents cancerous epithelium annotation, purple stands for cancerous stroma annotation, green shows NAT epithelium annotation and the yellow indicates NAT stroma annotation

The traffic light classification result presenting system was combined with the random forest classification model to further visualise the classification results using the unit of core rather than individual spectrum, which meant, when counting votes and finding the majority party between cancerous and normal classes, the previously obtained classification results of the spectra selected from each core would be considered as new voting elements. For example, in a core, without using any result counting method (mentioned in the next two paragraphs), there were 150 spectra already classified as cancerous and 15 spectra classified as normal. This core would be classified as a cancerous core and coloured in red using the traffic light system.

Again, two result counting methods were applied after classification. The first one using spectra from the annotated area of independent test cores. Only the top-scored 10% spectra were kept for each core and among these spectra if one or more than one spectrum was classified as cancerous spectrum and coloured in red. If no spectrum was classified as cancerous ones and the number of diagnosed NAT spectra was larger than half of the total number of selected spectra and coloured in green. In any other cases, the cores would be coloured in yellow. The coloured cores of both other cores from the same patients and cores from different patients in the training group were shown in Figure 8.24. In this plot, cores in the box were cores from different patients while the rest of cores were other cores from the same patients in the training process. The classification accuracy for cores from the same patients was 7 out of 11, which was 64%, while the accuracy for different patients was 9 out of 20, which was 45%. These accuracies were similar to the previously calculated classification accuracies in the unit of spectrum.

The second method, shown in Figure 8.25, using all the spectra from each independent test core, which were first applied to the epithelium and stroma separating model, and only the top 50% scored spectra were kept. These spectra were then sent to the cancer NAT classification model, and only the top 10% scored spectra and their scores were used for class assignments. If the number of diagnosed cancerous spectra was larger than the number of classified NAT spectra, with the condition that the number of cancerous spectra was larger than half of the total number of kept spectra, the core was picked as a cancerous core. If the number of diagnosed cancerous spectra was less than the number of diagnosed stroma spectra and the number of stroma spectra was larger than the total number of kept spectra, the core was named as a NAT core. In any other cases, the core was coloured in yellow which means not determined and needs further classification. For the Random forest classification method, 5 cores were diagnosed as further investigation was required, which means that the model struggles to determine which class these spectra belong. There was no major lead in the classification process. The classification accuracy for other cores from the same patients was 5 out of 11, which was 45%, while the accuracy for cores from different patients was 8 out of 20, which was 40%. These classification accuracies were not ideal but similar to the previous results from individual spectrum classification.

Comparing two traffic light result counting methods, although both methods obtain similar classification results to previous independent tests in the unit of spectra, the first method works slightly better than the second. One possible explanation was that the model could only produce low scores when it comes to the independent test. Although the top-rated spectra were selected, these spectra still had relatively low scores, which mean the classification model was not sure about which class these spectra belonged and possible over-fitting has occurred during model training.

Comparing results obtained from other cores from the same patients as those used in training and cores from different patients, classification accuracy from the same patients was 15% to 20% higher than classification accuracy from different patients, which indicates that either the classification considered common features of the training patients as the features of prostate cancer or the individual differences among people were difficult to overcome with limited number of patients. In order to obtain better classification accuracy, larger training data set and better models were required.

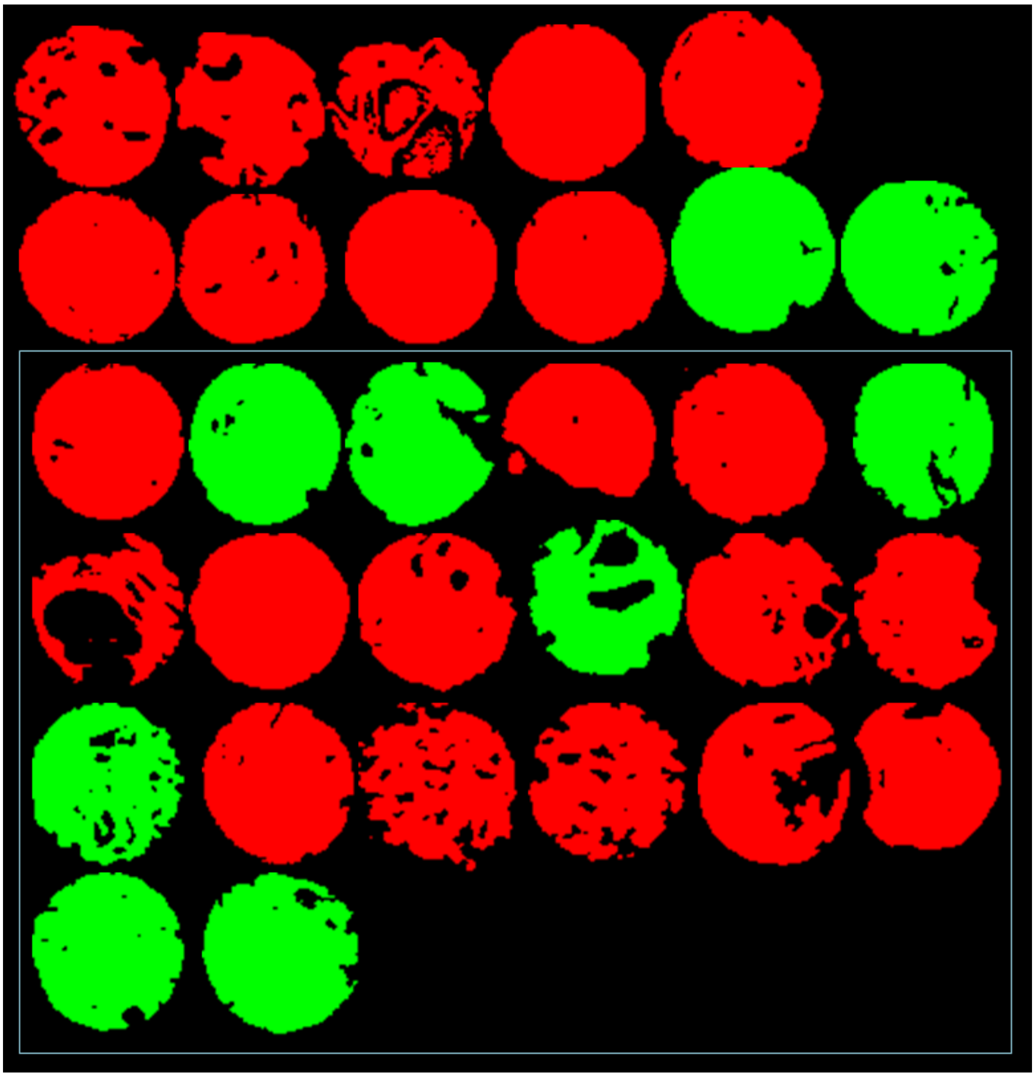


Figure 8.24 classification results presented in the traffic light system with the integration method one using random forest, where red stands for cancer prediction and green shows NAT prediction

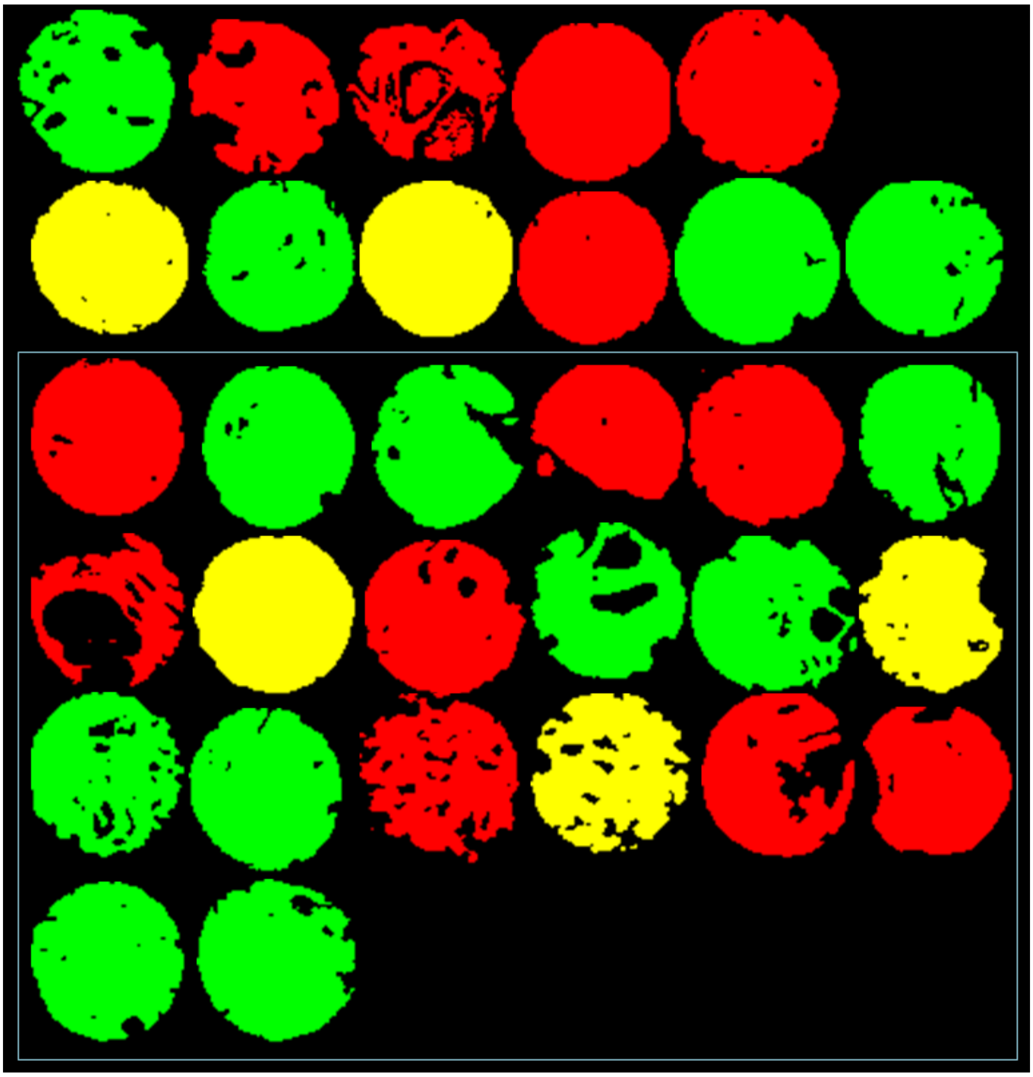


Figure 8.25 classification results presented in the traffic light system with the integration method two using random forest, where red stands for cancer prediction, yellow indicates uncertain case, and green shows NAT prediction

8.3.2.3. Section conclusion

Classification results in terms of separating cancerous and NAT spectra were poor and certainly would not be considered good enough for clinical application. AdaBoost has both relatively low classification accuracy on both the training validation and the independent test, while random forest succeeds in the training validation but fails badly in the independent test due to possible over-fitting. Although thresholds were applied and only half top-scored spectra were kept, still the classification accuracies were poor and do not improve very much. It means that the features found by the models could just be similarities among patients in the training but not the general features of prostate cancer. It could be that the differences between patients were larger than the differences between cancer and non-cancerous tissue making the prostate cancer features harder to identify.

Despite the classification results, AdaBoost seems to have more stability in the training and independent test.

Comparing the classification results obtained on glass with the work conducted previously by other groups, with infrared transparent substrates, not only the infrared spectral information could be used for cancer diagnosis, but also correlated with Gleason grade and tumour stage. Using three-band Gleason criteria, a sensitivity over 70% and specificity over 81% were reported in 2006 [20], and overall sensitivities and specificities of 92.3 and 98.9% were presented in 2008 [6]. Using glass as substrate still has large potentials, as the classification results have large improvement space comparing with works on CaF_2 and BaF_2.

In term of human individual differences, it has large influences on classification results. Regardless of which classification methods were applied, tests on the cores from the same patients with the training group did work 10-20% better than those 20 cores from completely different patients. It means a part of the spectral patterns obtained by models used as classification features were actually not prostate cancer-related information. It could be some biological similarities among the training patients group.

Another observation was that the double-sized independent patient group worked much poorer than the smaller independent group. One possible reason for it could be that the smaller group of patients has more similarities with the training group compared with the larger independent patient group. The classification accuracy observed previously could be based on the similarities between training and the smaller independent test group. With another group, these similarities were not detected or reduced, which lead to the poor classifications as observed.

Comparing performances of these two classification methods taking core as classification unit, AdaBoost seems to have a slightly higher classification accuracy regardless which result counting methods were adopted. One possible reason was that AdaBoost weights frequently appeared features, which was considered to be more helpful, while random forest treats each feature equally regardless its importance.

However, for both classification methods, the independent test results presented previously was not good. This could be caused by a small number of feature input (determined by the nature of glass substrates) and other interruption information from spectra, which could be the mounting material used to fix coverslip on the slide.

8.4. Conclusion and future works

To sum up, spectral data analysis, using glass as a substrate for infrared measurements still has great potential for pre-screening in for cancer diagnosis. The digital histology is promising as both AdaBoost and random forest have very high accuracy in for separating epithelium and stroma cells. Using glass substrates combined with FTIR technology can augment and potentially replace some pathological staining processes to save the number of piece of precious biopsy tissue required from patients for multiple stains and time of the pathologists. More cell type classification can be focused in the future to further validate the advantages of technology.

In terms of cancer diagnosis, neither AdaBoost nor random forest gives satisfying results in the unit of individual spectrum and core. Classification result presentation methods and further results processing can provide different angle to further understand the classification outputs. One of the possible causes for the relatively low classification accuracy of both machine learning models, in this case, can be lack of data, since comparing with other infrared transparent substrates the nature of glass slides limited the wavenumber ranges measured. Lost information from 2800 to 1000 cm^{-1} could contain important biological features which will help with cancer diagnosis. Another reason could be that the training patient group size is too small, so both models can not deal with human individual differences well. Cancer in each patient can have slight differences based on individual body conditions. The general cancer features are hidden patterns which are different to find. However, similarities, which are not cancer related, among training patients are easier to find when a small group of training patients are applied. These similarities are selected as classification feature patterns for cancerous and NAT spectra, which can directly lead to failure in independent tests. With a large increase in the number of patients involved in model training, the non-cancer related features will have less influence on the classification results, as they will be more and more difficult to find. The size of independent test is also important. A larger group size provides more reliable validation for models.

For future researches on glass with unstained tissue, further research should be conducted on the mounting material of glass slides. Since these materials were not designed for infrared measurements, it can cause a potential problem on the spectra.

Other more advanced classification methods can be conducted to further dig the subtle correlation between prostate cancer and spectra collected with a limited range of wavenumber. Deeping learning algorithm, for example, artificial neural network, can be applied, which provides additional angle to the data collected. Instead of numerical spectral data, chemical images can be applied to build the graphical neural network. Combination of images and spectra, the new approach has great potential to improve the current classification accuracies and move the clinical transition of infrared technology forward.

8.5. Reference

[1] About prostate cancer | Prostate cancer | Cancer Research UK, Cancer Res. UK. (2018). https://www.cancerresearchuk.org/about-cancer/prostate-cancer/about (accessed May 3, 2019).

[2] P. Bassan, J. Mellor, J. Shapiro, K.J. Williams, M.P. Lisanti, P. Gardner, Transmission FT-IR chemical imaging on glass substrates: Applications in infrared spectral histopathology, Anal. Chem. (2014). doi:10.1021/ac403412n.

[3] J.B. Lattouf, F. Saad, Gleason score on biopsy: Is it reliable for predicting the final grade on pathology?, BJU Int. (2002). doi:10.1046/j.1464-410X.2002.02990.x.

[4] M.J. Baker, E. Gazi, M.D. Brown, J.H. Shanks, N.W. Clarke, P. Gardner, Investigating FTIR based histopathology for the diagnosis of prostate cancer, J. Biophotonics. (2009). doi:10.1002/jbio.200810062.

[5] L.F.S. Siqueira, K.M.G. Lima, A decade (2004 - 2014) of FTIR prostate cancer spectroscopy studies: An overview of recent advancements, TrAC - Trends Anal. Chem. (2016). doi:10.1016/j.trac.2016.05.028.

[6] M.J. Baker, E. Gazi, M.D. Brown, J.H. Shanks, P. Gardner, N.W. Clarke, FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer, Br. J. Cancer. (2008). doi:10.1038/sj.bjc.6604753.

[7] P. Bassan, J. Lee, A. Sachdeva, J. Pissardini, K.M. Dorling, J.S. Fletcher, A. Henderson, P. Gardner, The inherent problem of transflection-mode infrared spectroscopic microscopy and the ramifications for biomedical single point and imaging applications, Analyst. 138 (2013) 144–157. doi:10.1039/C2AN36090J.

[8] M.J. Pilling, P. Bassan, P. Gardner, Comparison of transmission and transflectance mode FTIR imaging of biological tissue, Analyst. (2015). doi:10.1039/c4an01975j.

[9] C. Hughes, J. Iqbal-Wahid, M. Brown, J.H. Shanks, A. Eustace, H. Denley, P.J. Hoskin, C. West, N.W. Clarke, P. Gardner, FTIR microspectroscopy of selected rare diverse sub-variants of carcinoma of the urinary bladder, J. Biophotonics. (2013). doi:10.1002/jbio.201200126.

[10] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, Analyst. 142 (2017) 1258–1268. doi:10.1039/c6an02224c.

[11] Sunil R. Lakhani; Ian O. Ellis;, S.J.P.H.T. Schnitt, V.M.J. van De, WHO Classification of Tumours of the breast, 2012. doi:10.1017/CBO9781107415324.004.

[12] A. Cutler, D.R. Cutler, J.R. Stevens, Random Forests, (2012). doi:10.1007/978-

1-4419-9326-7.

[13] B. Schölkopf, Z. Luo, V. Vovk, Empirical inference: Festschrift in honor of Vladimir N. Vapnik, Empir. Inference Festschrift Honor Vladimir N. Vapnik. (2013) 1–287. doi:10.1007/978-3-642-41136-6.

[14] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, RUSBoost : A Hybrid Approach to Alleviating Class Imbalance, 40 (2010) 185–197.

[15] F.N. Pounder, K. Reddy, R. Bhargava, Development of a practical spatial-spectral analysis protocol for breast histopathology using Fourier transform infrared spectroscopic imaging, Faraday Discuss. 187 (2016) 43–68. doi:10.1039/C5FD00199D.

[16] MathWorks, Fitcensemble, MathWorks Doc. (2019). https://uk.mathworks.com/help/stats/fitcensemble.html (accessed July 27, 2019).

[17] A.C.S. Talari, M.A.G. Martinez, Z. Movasaghi, S. Rehman, I.U. Rehman, Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues, Appl. Spectrosc. Rev. 52 (2017) 456–506. doi:10.1080/05704928.2016.1230863.

[18] S. Kasraeian, D.C. Allison, E.R. Ahlmann, A.N. Fedenko, L.R. Menendez, A comparison of fine-needle aspiration, core biopsy, and surgical biopsy in the diagnosis of extremity soft tissue masses, in: Clin. Orthop. Relat. Res., 2010. doi:10.1007/s11999-010-1401-x.

[19] A.B. Barqawi, R. Turcanu, E.J. Gamito, M.S. Lucia, C.I. O’Donnell, E.D. Crawford, D.D. La Rosa, F.G. La Rosa, The value of second-opinion pathology diagnoses on prostate biopsies from patients referred for management of prostate cancer, Int. J. Clin. Exp. Pathol. (2011).

[20] E. Gazi, M. Baker, J. Dwyer, N.P. Lockyer, P. Gardner, J.H. Shanks, R.S. Reeve, C.A. Hart, N.W. Clarke, M.D. Brown, A Correlation of FTIR Spectra Derived from Prostate Cancer Biopsies with Gleason Grade and Tumour Stage, Eur. Urol. (2006). doi:10.1016/j.eururo.2006.03.031.

Chapter 9

Conclusion and Future Works

To sum up, this thesis is mainly focused on two types of tissues, breast and prostate tissues.

For breast tissue cores, digital histology and cancer diagnosis, for tissue on both CaF_2 and glass substrates were discussed. Generally, for both digital histology prediction and cancer diagnosis, the classification accuracies were higher on the CaF_2 slides compared with the glass slides.

For breast tissues on CaF_2 slides, not only were spectra of epithelium cells involved in model training and construction process but also the stroma region of tissues was used as training inputs for classification models. Although the stroma spectra did not provide predictions with higher classification accuracies than epithelium cells, the advantage is that it is available when epithelial spectra were lacking in the best tissue biopsy. Using stroma spectra as training samples could avoid the problem of lack of model training information. With larger training data, the model can potentially provide reliable and robust predictions on unknown samples. In terms of classification methods, two machine learning algorithms, random forest and AdaBoost were applied to the same datasets. In the training validation tests, both methods worked approximately equally well on the same samples. However, in the independent tests, AdaBoost worked better than the random forest. The differences could be caused by the nature of AdaBoost algorithm, weighting features which favour the classification.

For breast tissue on glass, with reasonable training data sampling, pre-processing steps and classification methods, epithelium spectra obtained an averaged 81.3% and 83.2% classification accuracies for cancer and NAT classes respectively. This is higher than the cancer diagnosis rates of the pathologist in 2010, 75.4%[1]. This further supports, together with the literature[2], the idea that the H&E stained tissue on glass slides could be used for tissue cancer diagnosis. This is a significant step forward for the translation of FTIR cancer diagnosis to clinical adoption. There was also a limitation of this study, in that only grade II breast cancer and NAT cores were used for model construction. In order to construct a more robust breast cancer diagnosis model on glass slides, the training set should include all type of breast cancer in the future.

In addition, the question of whether caveolin-1 staining can be used as diagnostic markers for breast cancer was also addressed in this thesis. Even though the random forest models could not find clear differences between caveolin-1 stained stroma and the unstained stroma, using spectra extracted from caveolin-1 stained stroma tissue as training inputs, the constructed model could successfully separate cancerous and normal cores in the independent test. This showed that caveolin-1 did have potential as a cancer diagnosis stain marker. However, with the limitation of caveolin-1 stain, only a very small number of spectra were used in the analysis. For more reliable reflection of the correlation between caveolin-1 and cancer prediction, more stained sample should be used in the future.

In terms of studies on prostate tissue, analysis has been conducted on the possibility of applying digital histology and cancer diagnosis on unstained prostate tissues. For digital histology, two machine learning methods were applied to classify epithelium and stroma spectra. The classification accuracies were 90% and 85% using both methods. This showed the potential of application of digital histology using unstained tissue on glass slides, which will have the potential to significantly benefit patients as multiple stains can be conducted using data from a single tissue slide.

In term of cancer diagnosis, neither classification method gave satisfying results. Two possible reasons could be found to explain the failure. Firstly, the low accuracies on independent test dataset were due to lack of training features in model construction. Comparing with other infrared transparent substrates the nature of glass slides limited the wavenumber ranges of measuring, which caused the limitation of feature selection. The second one was that unknown glue was found in the sample spectra, which interfered with the tissue spectra and influenced the classification results. Two digital glue removal methods, least-square fitting and EMSC were applied to remove the influence of glue. However, both methods failed to completely removal glue from the spectra of mixture. Two possible reasons could be raised. The first one was that the assumed used glue was similar but not the same with the glue on the slides. Secondly, the content of glue could be unevenly distributed, which caused variations in the spectra. In order to conduct more research using unstained tissues on glass slides, further studies concerning the glue used on the slides are important.

For future research on cancer diagnosis of human tissue on glass, other more complicated and advanced classification algorithms can be applied, for example, deep learning based on the artificial neural network. Artificial neural network based on digital stained chemical images of individual cores is potentially great classification tool in term of cancer diagnosis, as it combines the chemical information of samples extracted by infrared spectra, which can be presented by different chemical stains of cancer biomarkers on generated chemical images of samples, and the appearance of cells in the sample tissue, which is used by pathologists as standard to make diagnosis decisions.

9.1. Reference

[1] S. Kasraeian, D.C. Allison, E.R. Ahlmann, A.N. Fedenko, L.R. Menendez, A comparison of fine-needle aspiration, core biopsy, and surgical biopsy in the diagnosis of extremity soft tissue masses, in: Clin. Orthop. Relat. Res., 2010. doi:10.1007/s11999-010-1401-x.

[2] M.J. Pilling, A. Henderson, J.H. Shanks, M.D. Brown, N.W. Clarke, P. Gardner, Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation, Analyst. 142 (2017) 1258–1268. doi:10.1039/c6an02224c.