FTIR Spectroscopy for cancer diagnosis. How can glass substrates be used to bring it closer to clinical practice?

Thesis submitted for the degree of
Doctor of Philosophy

Lewis Michael Morgan Dowling

December 2023

Faculty of Medicine and Health Sciences

Keele University

Contents

Acknowledgements

I would like to express great thanks to my supervisor Professor Josep Sulé-Suso. With his expertise and guidance, he has helped support me throughout my PhD research. He has provided me the opportunities to turn my research into publications and present my research at conferences. Without his support as a supervisor this work would not have been possible.

Thank you to Dr Ibraheem Yousef at ALBA synchrotron who has helped supervise my research while at the MIRAS line in ALBA synchrotron. With his support at ALBA synchrotron, I was able to learn how to use the instrumentation and collect data for my research.

I am grateful to Dr Paul Roach for allowing use of the FTIR spectroscopy instrumentation at Loughborough University and his guidance in using the instruments and data collection.

I would like to express my gratitude to Photothermal Spectroscopy Corp and Dr Mustafa Kansiz for their collaborative work with us and allowing use of their mIRage IR microscope to conduct O-PTIR spectroscopy research.

Thank you to Achim Kohler and his group at the Faculty of Science and Technology at NMBU who provided their expertise to help improve my data analysis.

Thank you to my family and Camille Morales who have given me their full support throughout my studies and who have always encouraged me.

I would like to acknowledge Keele University and ALBA synchrotron for funding this research.

Abbreviations

Adenocarcinoma in situ (AIS)

Ammonium chloride potassium (ACK)

Anaplastic lymphoma kinase (ALK)

Area under the curve (AUC)

Attenuated total reflection (ATR)

Body mass index (BMI)

Bovine pituitary extract (BPE)

Breast cancer gene 1 (BRCA1)

Breast cancer gene 2 (BRCA2)

Breast conserving surgery (BCS)

Chronic obstructive pulmonary disease (COPD)

Circulating tumour cells (CTCs)

Circulating tumour DNA (ctDNA)

Circulating tumour RNA (ctRNA)

Computerised tomography (CT)

Core needle biopsy (CNB)

De-novo lipogenesis (DNL)

Ductal carcinoma in situ (DCIS)

Epidermal growth factor receptor (EGFR)

Erythroblastic oncogene B (ERBB2)

European Medicines Agency (EMA)

Extended Multiplicative Signal Correction (EMSC)

Extracellular vesicles (EVs)

Fine needle aspirate cytology (FNAC)

Focal plane array (FPA)

Foetal bovine serum (FBS)

Food and Drug administration (FDA)

Fourier Transform Infrared (FTIR)

Gentamicin sulphate-amphotericin (GA-1000)

Haematoxylin and eosin (H&E)

Hierarchical cluster analysis (HCA)

Hormonal replacement therapy (HRT)

Human epidermal growth factor (hEGF)

Human epidermal growth factor receptor 2 (HER2)

Immunohistochemistry (IHC)

Infrared (IR)

Invasive ductal carcinoma (IDC)

K-nearest neighbours (KNN)

Linear accelerator (LINAC)

Linear discriminant analysis (LDA)

Mammary epithelial cell growth basal medium (MEBM)

Mercury cadmium telluride (MCT)

Minimally invasive carcinoma (MIA)

miRNA (microRNA

mRNA (messenger RNA)

National health service (NHS)

Next generation sequencing (NGS)

Non-small cell lung cancer (NSCLC)

Oestrogen receptor alpha (Er)

Optical path difference (OPD)

Optical photothermal infrared (O-PTIR)

Papanicolaou (Pap)

Squamous cell carcinoma (SqCC)

Tumour educated platelet (TEP)

United Kingdom (UK)

United States of America (USA)

List of figures

List of tables

List of Equations

Abstract

Cancer incidence rates are increasing world-wide including in the UK. An increase in cancer cases puts further pressure on pathology departments that are often already struggling to meet targets to diagnose cancers in a timely manner. Delays in diagnosis will cause the delay of treatment being provided and worse patient outcomes. Current diagnostic methods for cancer rely on cytological/histological staining of biopsies and a diagnosis is made in a subjective manner by a pathologist. These methods are time consuming and require great expertise. New diagnostic methods are needed to help relieve pressures on pathology departments. There is a consensus that vibrational spectroscopy techniques have the potential to be tools that could aid in cancer diagnostics. Despite an increasingly growing body of research demonstrating how vibrational spectroscopy methods could be utilised for clinical diagnostics there has been several barriers to the translation of such methods.

The research in this thesis aims to investigate and demonstrate methodologies to utilise modes of infrared spectroscopy with glass substrates for lung and breast cancer diagnostics. One of the major barriers for the use of infrared spectroscopy in cancer diagnostics is the expense and difficulty of procurement of conventional substrates. This thesis aimed to investigate a methodology to use a glass coverslips substrate for the classification of lung and breast cancer cells using IR spectroscopy. Glass coverslips were used because of their affordability and accessibility, an important consideration for the translation of diagnostic methods.

In vitro cancer cell lines and healthy tissue derived cell lines were used to model this research to test the feasibility of the proposed methods. This research first investigated a sample preparation method for cytology samples to be analysed with FTIR spectroscopy. The

next sections demonstrated the proposed method could be used to classify lung and breast cancer cells in-from non-malignant cells in-vitro using FTIR spectroscopy and a random forest classifier. The methodology was next used to demonstrate how FTIR spectroscopy could be used to identify individual lung cancer cells from leukocytes in mixed samples. This is the first time this has been demonstrated. Finally, related IR spectroscopy technique, O-PTIR spectroscopy, was investigated for how it could be used with glass slides for the classification of lung cancer cells from non-malignant cells. The research in this thesis has demonstrated that glass substrates are viable for the classification of lung and breast cancer cells with high accuracy using sample preparation methods that are commonplace in pathology laboratories for current diagnostic procedures.

Chapter 1: Literature review

The number of cancer cases in the UK are continuing to rise with a 12% incidence rate increase for all cancers in the UK from the 1990s to 2017 (Cancer Research UK, 2018). This ever-increasing incidence of cancer generates a greater workload for pathology departments and an increased turn around for key cancer diagnoses. A delay in diagnoses causes a delay in treatment and possible worsening of patients' condition and survival and an increase in their stress and anxiety. An automated system that could identify abnormal cells in cytology samples for further investigation would be ideal for managing this increased workload. This could reduce the time pathologists would spend looking at samples to deem if they are positive or negative for cancer. Additionally, this kind of system could be used to also help in differentiating types of cancer to further improve diagnosis times.

There is a strong body of research showing that Fourier Transform Infrared (FTIR) microspectroscopy has potential as a technique that could aid pathologists in their work investigating tissue/cytology samples from patients with cancer or suspected cancer. Despite the plethora of work carried out in research settings, FTIR microspectroscopy has yet to be translated to the clinical setting (Finlayson, Rinaldi and Baker, 2019). One of the main drawbacks has been the substrates that samples are placed on for FTIR microspectroscopy (CaF_2, BaF_2, ZnSe) as they are often expensive, costing up £50-60 per slide. This would make a diagnostic system based on FTIR microspectroscopy very expensive with the large number of samples that need to be studied in a clinical setting. The glass slides commonly used in pathology departments as a substrate for cytology samples obscures the fingerprint region

of the spectra because the glass absorbs IR radiation (carbohydrates, proteins, nucleic acids) (Bassan et al., 2014) (Pilling et al., 2017) (Rutter et al., 2018).

Previous work by our group showed that thin soda lime glass coverslips of a thickness of 0.12-0.17 mm could be used as a substrate and that the lipid bands and amide I and II bands were visible (Rutter et al., 2019). These coverslips allow for the study of bands in the fingerprint region down to 1350 cm^{-1} which cannot be seen on regular glass slides. It is believed that, due to the coverslips being thinner than the glass slides commonly used, less IR radiation is absorbed which allows the amide I and II peaks to be viewed. The current research which I am conducting aims to expand on this work. The research will use FTIR microspectroscopy to firstly differentiate between different lung cancer cells (A549, CALU-1), normal counterparts (NL20) and peripheral blood mononuclear cells (PBMC) placed upon the soda lime glass coverslips. This work will be another step towards translating FTIR spectroscopy to a system that could be utilised in a clinical setting. The work will make use of both benchtop spectrometers and synchrotron light based FTIR microspectroscopy.

My main research question is: can soda lime glass coverslips be used as a substrate for FTIR microspectroscopy of lung and breast cancer cells to distinguish them from normal cells? Answering this question could move FTIR microspectroscopy a step closer to being used in a clinical setting as it would show that expensive CaF_2, BaF_2 and ZnSe substrates do not have to be used for transmission FTIR microspectroscopy studies to distinguish between the cells as outlined. This would reduce the cost of the technique and make it easier for pathologists to perform their routine test on the sample such as staining and immunohistochemistry as the coverslips can simply be stuck on to a glass slide to allow for the staining.

Lung cancer

Lung Cancer incidence, burden, and mortality

In the UK there are approximately 47,200 new lung cancer cases diagnosed per year. These lung cancer cases account for 13% of all UK cancer diagnoses a year (Cancer Research UK, 2016). Lung cancer mortality rates have remained high for the last 40 years with approximately 35,600 lung cancer deaths a year in the UK, accounting for 21% of all UK cancer deaths. The situation is similar worldwide with lung cancer accounting for 27% of cancer deaths in the United states of America (USA) in 2015, and 20% within the EU in 2016 (Malhotra *et al.*, 2016). Lung cancer survival rates remain poor with 5 in 100 people diagnosed with lung cancer surviving beyond 10 years in the UK (Cancer Research UK, 2016), and this survival rate is lower still in less developed countries with poorer healthcare. Low survival rates can largely be attributed to the fact that most lung cancers are diagnosed once they are symptomatic in later stages of the disease which makes treatments with a curative intent difficult.

Lung cancer risk factors

Tobacco smoking is the largest cause of all histological types of lung cancer. Lung cancer cases have decreased in the UK by 8% in the last 20 years in most part due to a reduction of the number of smokers and a ban on smoking within enclosed public spaces. The risk among continuous smokers compared to that of non-smokers has been measured in the order of 20-50-fold greater (Malhotra *et al.*, 2016). This risk of lung cancer is reduced in ex-smokers

compared to continuous smokers however excess risk is still increased later in life compared to non-smokers. Passive smoking exposure is also a risk. The excess risk of non-smokers with a spouse who smokes is estimated to be in order of 20-30% (Asomaning *et al.*, 2008).

A family history of cancer has been found to be a significant risk factor for lung cancer. A major susceptibility locus for lung cancer has been mapped to chromosome 6q23-25 (Bailey-Wilson *et al.*, 2004). The fact that a large number of smokers do not develop lung cancer suggests that a genetic predisposition may contribute to the carcinogenesis of lung cancer.

Age is an important risk factor for most types of cancer including lung cancers. 75% of lung cancer cases are diagnosed in people over 65 years of age (Cancer Research UK, 2016). The UK and much of the developed world are facing ageing populations which will contribute to an increase in the incidence of cancer cases.

Other pulmonary conditions could increase the risk of developing lung cancer. It has been suggested that chronic obstructive pulmonary disease (COPD) could increase the risk of developing lung cancer independently of smoking (Turner *et al.*, 2007). However, others have suggested that it is impossible to remove the residual effect of smoking from the potential risk of COPD (Powell *et al.*, 2013). A relative risk of 1.8 (95% CI 1.3-1.3) was reported from a meta-analysis of twenty-two studies on asthma and lung cancer in never smokers (Santillan, Camargo and Colditz, 2003). A population-based case control study in Shanghai investigated the links between tuberculosis (TB) and lung cancer (Zheng *et al.*, 1987). The study found those with a history of TB within the last 20 years had a risk of lung cancer of 2.5-fold and there was a correlation between the location of TB lesions and tumours.

Occupational exposure to carcinogens provides a significant risk for lung cancer. One study estimated that 14.5% of lung cancer cases in the UK can be attributed to exposure to carcinogenic agents from occupations (Rushton *et al.*, 2012). The most significant occupational carcinogenic agents are radon, asbestos, silica, heavy metals and polycyclic aromatic hydrocarbons. All forms of asbestos are carcinogenic to the human lung and is known to cause mesothelioma. Asbestos was widely used in the UK for building insulation prior to its ban in 2000. The use of asbestos is banned in 55 countries but is still an occupational hazard in those countries that still use asbestos. Silica is a material used in pottery making, ceramics and brick making that is also shown to be carcinogenic to the lungs and has demonstrated to have similar effects on the lung to asbestos (Steenland *et al.*, 2001). Radon is a radioactive material that emits ionising α-particles as products of decay, those that work in mining industries are at increased risk of radon exposure.

Lung cancer types

Lung cancer can be categorised into two main histological groups: small cell lung carcinoma (SCLC) that make up 15% of lung cancers and non-small cell carcinomas (NSCLC) which account for 85% of lung cancers (Ryan and Burke, 2017). NSCLC can be further subcategorised as squamous cell carcinoma (SqCC), adenocarcinoma and large cell carcinoma.

Adenocarcinomas are the most common type of lung cancers, comprising 40% of all lung cancer cases. Adenocarcinomas are defined as an epithelial neoplasm with mucin production or pneumocyte immunohistochemical marker expression (Inamura, 2017). Adenocarcinomas

typically form a peripherally located mass with central fibrosis and pleural puckering. As of 2015, the World Health Organisation (WHO) released a new classification that divides adenocarcinoma into adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma or invasive adenocarcinoma based on the extent of invasiveness. AIS is defined as an adenocarcinoma with a lepidic pattern and a diameter <3 cm. If the tumour diameter is >3 cm it is defined as lepidic predominant adenocarcinoma. Minimally invasive carcinoma (MIA) is an adenocarcinoma with a diameter of <3 cm and invasion size of <5 mm. If there is the presence of lymphovascular invasion, pleural invasion or tumour necrosis, it excludes an MIA diagnosis even if both the tumour size and invasion size comply (Ryan and Burke, 2017). Invasive adenocarcinomas are further classified into five differentiation patterns: lepidic, papillary, acinar, micropapillary, and solid adenocarcinoma. Defining the classification of the adenocarcinoma is very important because it will affect the prognosis and treatment plan.

SqCC represents around 25% of lung cancers. In the 2015 WHO classification SqCC was divided into keratinising, non-keratinising and basaloid SqCC. Prior to the 2015 classification, basaloid SqCC was classed as a large cell carcinoma, but it was recategorized based on newly identified SqCC markers from immunohistochemistry (Inamura, 2017).

Neuroendocrine tumours are a new classification for lung tumours established in the 2015 WHO classification. Three subtypes exist within the neuroendocrine classification: SCLC, large cell neuroendocrine carcinoma and carcinoid tumours. Neuroendocrine tumours are very aggressive and are correlated with a long history of smoking (Zappa and Mousa, 2016).

Lung cancer diagnosis

One of the important factors for the prevention of disease progression and successful treatment of the disease is early detection and identifying the cancer before systemic invasion. If a patient is suspected of having lung cancer, they are sent for an evaluation using imaging techniques. The first stage of diagnosis is imaging the chest through x-ray, computerised tomography (CT), and positron emission tomography (PET) (Nasim, Sabath and Eapen, 2019). CT scans are the most used technique for the staging of tumours and to assess the success of treatment. The scans provide information on the tumour size, location and anatomical characteristics. CT scans have limitations detecting metastasis in normal sized lymph nodes and differentiating between tumour adhesion and infiltration. PET is used for the assessment of pulmonary nodules and metastases. CT and PET scans are used in conjunction if it is likely that metastases are present. The lungs can be difficult to image due to their large size and surface area. High resolution CT can identify lung nodules of less than 1 cm in size, but it is not sensitive enough to identify bronchogenic invasive legions (Gohari and Haramati, 2004). Benign granulomas in the lung can mimic early and preinvasive lung cancer leading to false positive results. After imaging, the tumour type and stage must be confirmed through tissue diagnosis. This is done through sputum cytology, lymph node biopsy, fine needle aspiration and video assisted thoracoscopy.

Small biopsy and cytology specimens are the primary methods of diagnosis for lung cancer. Sputum cytology can be used for the early detection of lung cancer. Sputum can be obtained through deep coughing and the application of saline mist (Ammanagi *et al.*, 2012). Sputum cytology is recommended for patients who cannot undergo more invasive methods.

However, the usefulness of sputum cytology is largely dependent on tumour cells' location and the tumour size. Cytology can suffer from low cellularity and non-cancer cells within the suspension (Wardwell and Massion, 2005). Subpleural tumours can be difficult to biopsy and come with a greater risk of pneumothorax, tearing the pleura and laceration of surrounding parenchyma (Gohari and Haramati, 2004).

Biopsy and cytological samples are typically stained with haematoxylin and eosin (H&E). Haematoxylin is a basic dye that stains the cell nucleus purplish blue, and eosin is an acidic dye that stains structures including the extracellular matrix and cytoplasm pink and red. H&E staining provides contrast between structures to allow identification of tumours within a tissue sample. The morphology of the tumour cells in the tissue can be seen from the staining.

Immunohistochemistry (IHC) is an important step of the diagnosis process. IHC is used to discriminate benign from malignant tumours, metastases from primary tumours and SCLC from NSCLC (Bubendorf *et al.*, 2017). In the case of non-squamous NSCLC, IHC is followed by molecular characterisation as the sub-type will affect prognosis and treatment options. Mutations causing increased expression and presence of epidermal growth factor (EGFR) and anaplastic lymphoma kinase (ALK) are detected through molecular characterisation by sequencing and PCR. The results of molecular characterisation will inform the clinician if targeted therapies and/or immunotherapies can be used.

Lung cancer diagnostic pathways still present with many problems including poor resolution, false positives and negatives and a lack of biochemical information. Histological and cytological diagnoses can be subjective in nature and rely on the pathologist. There is a risk

of multiple biopsies needed before diagnosis is made which can cause further suffering to patients.

A major problem with current diagnostic methods for lung cancer diagnosis is the lack of early detection. The current diagnostic methods rely on imaging which can easily miss tumours in early stages when they are smaller in size, and patients are often not imaged until symptoms are apparent in later stages of disease. The requirement of tissue biopsies to make a diagnosis also makes it difficult to diagnose early-stage disease if it is detected because the small tumour can be difficult to access to obtain the biopsy. As there is risk involved from radiation during imaging and the invasive surgery required for biopsies there is currently no screening program for lung cancer in the UK and many other countries. More methods of diagnosis are needed to improve early detection of lung cancer to increase survival.

Lung cancer treatment

The treatment options for lung cancer largely depend on the stage of the cancer and the type. The TNM-based staging system is a commonly used system which describes the anatomical extent of the cancer and its severity (Lemjabbar-Alaoui *et al.*, 2015). The T in the TNM system indicates the size and extent of the primary tumour, N is the extent of involvement of the regional lymph nodes and M is the presence or absence of distant metastatic spread. A number is given to each of these three categories to describe the extent of each. Subsets are combined into stage groupings which indicate the severity of the

cancer. NSCLC has four stages (I-IV) with the lower the stage being the least severe. SCLC has two stages: limited and extensive. Full lung cancer TNM staging is shown in appendix 1.

For the early stages (I-II) of the disease, the primary treatment is surgery if the tumour is resectable. Surgery provides the best long-term survival for the disease at this stage with a five-year survival rate of 60-80% for stage I NSCLC patients after resection of the tumour and 30-50% for stage II. For patients unable to undergo surgery or with unresectable tumours, radiotherapy and chemotherapy are the recommended treatments.

A majority of NSCLC patients are diagnosed in advanced stages (III-IV) of the disease. Stage III NSCLC is a heterogenous disease that varies from a resectable primary tumour with microscopic metastases to the lymph nodes to an unresectable bulky tumour with many nodal locations. The treatment approach is determined by the tumour location and whether it is resectable. Standard treatment for resectable tumours is surgery followed by adjuvant chemotherapy. For unresectable tumours treatment will include a combination of chemotherapy and radiotherapy, and more recently immunotherapy and targeted therapies. Stage IV NSCLC which accounts for 40% of new diagnoses is difficult to treat and has very low survival rate, treatment for grade IV NSCLC is often palliative treatment. The choice of treatment for grade IV NSCLC will depend on many factors including co-morbidities, histology and molecular genetic features of the cancer. Treatment for stage IV NSCLC includes radiotherapy, combination chemotherapy, targeted therapy and/or immunotherapy.

Therapeutic progress in recent years can be mostly attributed to targeted therapies that target the specific molecular genetic mutations of the cancer. Patients who have a cancer that does not have an approved targeted therapy, the first line treatment is platinum-based

doublet therapy with or without bevacizumab. Erlotinib (Tarceva), gefitinib (Iressa) or afatinib (Giotrif) are targeted drugs that act on NSCLC with epidermal growth factor receptor (EGFR) mutations (Hirsch *et al.*, 2017). In 2007 it was discovered that an oncogenic ALK gene rearrangement was present in some NSCLCs (Soda *et al.*, 2007). The ALK rearrangement result from translocations or inversions on chromosome 2 that fuse to regions of exon 20 of the ALK gene. There have been several ALK targeted therapies developed since this discovery including crizotinib, ceritinib, and alectinib. For patients with EGFR or ALK positive NSCLC, targeted therapies form the backbone of treatment. The third target to be approved after EGFR and ALK was ROS1. ROS1 rearranged phenotypes have been described as a distinct molecular phenotype in 1-2% of NSCLC patients. ROS1 rearrangements causes fusion of ROS1 tyrosine kinase domain with partner genes, usually on another chromosome.

Recent focus has been on immunotherapies. One such immunotherapy target is the programmed death ligand 1 (PD-L1) and its receptor programmed death-1 (PD-1). PD-L1 is an inhibitory immune checkpoint molecule. Agents that target PD-L1/PD-1 have shown promising results in NSCLC treatment. Two antibodies that target PD-1; nivolumab and pembrolizumab while another two antibodies target PD-L1; atezolizumab and durvalumab, have been approved by the US Food and Drug Administration (FDA) and European Medicines Agency (EMA) for the treatment of NSCLC. Only about 20% of patients respond to these treatments as a monotherapy so it is important to identify the patients who will benefit most from this therapy (Hirsch *et al.*, 2017).

Breast Cancer

Breast Cancer incidence, burden, and mortality

Breast cancer accounts for 15% of new cancers in the UK and is the most common cancer in women. Breast cancer incidence rates have increased by 18% since the early 1990s (Cancer Research UK, 2018). While breast cancer mortality has decreased by a sixth in the UK in the last decade, it still accounts 7% of all cancer death. Breast cancer mortality rates is projected to continue to fall this next decade by 13%. This can be contributed to improvements in treatment from immunotherapies and the screening program diagnosing many cases in early stages.

Breast cancer risk factors

The most significant risk factor for breast cancer is sex with less than 1% of breast cancers occurring in men. Breast cells in women are vulnerable to increases in oestrogen and progesterone. A change in endogenous hormone levels can increase the risk of breast cancer in premenopausal and postmenopausal women. Men produce an insignificant amount of oestrogen along with less breast tissue so are at less risk of developing breast cancer (Łukasiewicz *et al.*, 2021). Age is another significant risk factor for breast cancers. Most breast cancers are diagnosed after the age of 50. Aging populations in the UK and much of the developed world is a significant cause for the rising incidence of breast cancer in

developed nations. Family history of breast cancer presents as a significant risk factor with approximately 13-19% of patients reporting a first-degree familial relation with breast cancer. Mutations in the breast cancer gene 1 (BRCA1) and breast cancer gene 2 (BRCA2) tumour suppressor genes are involved in about 3% of breast cancers. Mutations of BRCA1 and BRCA2 can be hereditary and are more common in some population groups such as those of Ashkenazi Jewish heritage (Petrucelli, Daly and Pal, 2022).

Low physical activity and a high body mass index (BMI) have been linked to increase risk of breast cancer. Women above 50 years old with a high BMI are at a greater risk of breast cancer than those with a low BMI. Greater BMI has also been associated with more aggressive cancers with greater lymph node metastasis and greater tumour size (Łukasiewicz *et al.*, 2021).

Certain drugs have been shown to increase the risk of breast cancer. The use of hormonal replacement therapy (HRT) longer than 5-7 years can increase breast cancer risk. The prolonged use of HRT can cause overstimulation of oestrogen receptors on breast cells (Williams and Lin, 2013).

Breast cancer types and development

Invasive breast cancer presents in varying behaviour and morphology with WHO distinguishing 18 different histological breast cancer types (Vajpeyi, 2005). Invasive ductal carcinoma (IDC) is the most frequently diagnosed subgroup of breast cancer (75%) and is diagnosed when a tumour fails to be diagnosed into a histological subtype (Łukasiewicz *et*

al., 2021). The remaining 25% of invasive breast cancers are recognised as specific subtypes from their distinctive growth patterns and cytological features. Molecular classification of breast cancer is also important with subtypes based on the expression of receptors that determine treatment.

Luminal breast cancers account for 70% of breast cancers in Western populations. They are oestrogen receptor (ER)-positive tumours. Most luminal breast cancers present as IDC but can be differentiated into invasive lobular, mucinous, invasive cribriform ad invasive micropapillary carcinomas. Molecular classifications of luminal tumours can be typed as luminal A or B subtypes. Luminal A tumours present with ER or progesterone receptor (PR) and have an absence of human epidermal growth factor receptor 2 (HER2). Luminal B tumours have a worse prognosis and present as ER positive, PR negative and HER2 positive.

HER2-enriched breast cancers are characterised by high expression of HER2 and are negative for ER and PR. HER2 cancers grow more aggressively than luminal cancers, but management and treatment of these cancers has improved with immunotherapies against HER2.

Triple negative breast cancers (TNBC) are characterised by being negative for the expression of ER, PR and HER2. BRCA1 mutations are a major contributing factor for the development of a majority if TNBC. TNBC are often aggressive and have a worse prognosis due to there not being available immunotherapies as for other breast cancers (Almansour, 2022).

Claudin-low breast cancers are mostly ER, PR and HER2 negative and are characterised by low expression of genes related to cell adhesion including claudins, occluding and cadherins. Epithelial mesenchymal transition patterns are common in these tumours, and they exhibit stem-like gene expression (Pommier *et al.*, 2020).

Breast cancer diagnosis

The first stage of diagnosing breast cancer is through imaging of the breasts with ultrasound, mammography and/or MRI to test for presence, size, location and number of tumours. If there is suspicion of a tumour, a biopsy sample is taken of the potential tumour.

There are three methods for taking a biopsy of a potential breast tumour, a fine needle aspirate (FNAC), core needle biopsy (CNB) or an open biopsy. FNAC uses a thin needle to take a sample of fluid and cells at the site of the suspicious lesion. FNAC has the benefit of being a fast and cost-effective procedure that has a high safety profile with little complications (Ohashi *et al.*, 2016). However, FNAC can have high false negative rates and can often not provide enough tissue to produce a conclusive diagnosis. CNB involves using a larger needle to take a core biopsy of the tissue at the lesion site. CNB requires the use of local anaesthetic and is a more costly and invasive procedure than FNAC. The use of CNB has become more widely used because of its higher sensitivity, selectivity and accuracy relative to FNAC. But CNB is more expensive, and the processing of the tissue can be time consuming which can lead to a delay in a diagnosis which puts more stress on to patients as they wait for a result. Currently both biopsy methods are used for breast cancer diagnosis with no consensus on which method to use routinely.

Once a biopsy is collected, a pathologist must firstly decide if it is cancerous or not. If the biopsy contains cancerous material, it must be staged and graded. This is done by staining the biopsy and examination under a microscope by a pathologist (Cardoso *et al.*, 2019). Similar to the diagnosis of lung cancer, breast cancer is staged using TNM or a numerical

stage. The earliest stage of breast cancer is ductal carcinoma in situ which is a pre-invasive cancer where the cancer cells have not spread into any of the surrounding breast issue. Early-stage invasive breast cancer has spread into the surrounding breast tissue, but the tumour is small and remains in the breast and has not spread away from the breast tissue. Locally advanced breast cancer is when the cancer has spread from the breast to nearby lymph nodes or to the chest wall. Advanced breast cancer is when the cancer has metastasised to other areas of the body. The grading describes the morphology of the cancer in comparison to normal cells and tissue. Low grade cancers have similar morphology to normal breast cells and are well differentiated with the tissue of ductal cancers forming small tubules and lobular cancer forming cords. Low grade cancers are slow growing and have a good prognosis. In intermediate grade cancer, the cell morphology looks abnormal, and the tissue is moderately differentiated. The cancer is faster growing than early graded cancer and has a poorer prognosis. High grade cancer has very abnormal morphology and are poorly differentiated having few recognisable tissue structures present in normal tissue. High grade cancer grows aggressively and has a poor prognosis. Another important stage of diagnosis is identifying the presence of ER, PR and HER2 through immunohistochemistry (Hammond *et al.*, 2010). The treatments used will be dependent on the presence or lack of these receptors.

In the UK there is a breast cancer screening program for women over the age of 50. Women over 50 are invited every 3 years util the age of 70 to have a breast screening (Breast screening | Breast cancer | Cancer Research UK, 2018). Women are screened using a mammography to search for any lesions that may potentially be cancer. The aim of the screening programme is to find cancer at an early stage, so it is more easily treatable. If the mammography shows a potential abnormality, the patient is invited back for further

imaging. If the area is confirmed to be suspicious, a biopsy is taken for testing. Current evidence suggests the number of deaths by breast cancer is reduced by 1300 a year in the UK because of screening. However, there are risks of false positives, overdiagnosis and overtreatment from screening (Marmot et al., 2012). Many women have biopsies taken which are diagnosed as not being cancerous. This can cause stress and anxiety in patients, and they undergo unnecessary procedures. Overdiagnosis and overtreatment is when a DCIS that may not have ever posed a risk of progressing to an invasive stage is treated. As it is not possible to know if these pre-invasive or slow growing cancers will progress, they must be treated. This leads to treatments such as surgery, radiotherapy and chemotherapy that could have been unnecessary.

Breast cancer treatment

The primary treatment for breast cancer is surgery. There are two major types of surgery for removal of breast cancer, breast conserving surgery (BCS) and mastectomy (Łukasiewicz *et al.*, 2021). BCS is the removal of the cancerous tissue while limiting removal of healthy breast tissue. Mastectomy is the removal of the entire breast. BCS is the preferred method in early stage and lower grade cancers because of less complications and a lower psychological burden (Chung *et al.*, 2015). Mastectomy is used when the tumour is large and would be difficult to remove with BCS. It is also used when the cancer is very aggressive and would likely recur after the removal of the tumour.

Chemotherapies can be used as a neoadjuvant or adjuvant. When selecting chemotherapy, it is important to tailor the therapy to the individual based on the characteristics of the cancer.

Neoadjuvant therapy is used for locally advanced tumours, to reduce the size of larger tumours to aid in BCS and for aggressive tumours where biological therapies (triple-negative) are not effective.

Radiotherapy is typically used after surgery or chemotherapy to ensure complete destruction of cancerous cells and minimise cancer reoccurrence of the cancer (Jonathan Yang and Ho, 2013). Use of radiotherapy is also favourable for the treatment of unrescetable tumours and metastatic tumours to prevent further cancer growth.

Targeted biological therapies have significantly improved the prognosis of HER2 positive breast cancer (Maximiano *et al.*, 2016). The first drug of this kind and still current standard is trastuzumab. Biological therapies use recombinant antibodies to target the overexpressed HER2 on the cancer cells blocking the signalling of HER2. HER2 overexpression promotes the activation of multiple downstream pathways that encourage cancer growth and proliferation. Therefore, by stopping this action the use of biological therapies can also enhance the efficacy of chemotherapy.

Hormonal therapy can be used against ER positive breast cancer. Hormonal therapy reduces oestrogen levels or prevents the cancer cells from being stimulated by oestrogen (Williams and Lin, 2013). Drugs that reduce oestrogen levels include aromatase inhibitors. ER stimulation can be prevented by blocking drugs called selective oestrogen receptor modulators or by ER degradation with selective oestrogen receptor degraders. Hormonal therapy is used to slow down growth and proliferation of the cancer and can be used in conjunction with chemotherapy.

Fourier transform infrared (FTIR) spectroscopy.

FTIR spectroscopy

FTIR is a vibrational spectroscopy technique that can be used to ascertain the chemical structure of a sample (Pallua *et al.*, 2018). FTIR spectroscopy uses infrared (IR) radiation which, when passed through a sample, some radiation is absorbed, some radiation is transmitted, and some is reflected. The absorbed radiation causes vibrations of the covalent bonds in the molecules which can be detected as signal.

Molecular species have different electronegative charges that generate a disproportional charge across molecules that result in dipoles (Cheeseman *et al.*, 2019). The natural vibrations of molecules cause the distance between the negative charge centre of each atomic species in a specific bond to fluctuate, this generates an electric field and is known as resonance frequency. If the frequency of the IR radiation matches the resonance frequency, then IR radiation is absorbed. The absorbed radiation causes greater vibrations of the molecular bonds producing a measurable signal. The signal generates a spectral fingerprint unique to a molecule as the frequency of vibration is unique for each functional group. The modes of vibration in the molecular bonds include bending, scissoring, rocking and symmetric and asymmetric stretching (Figure 1). The IR spectra is plotted as absorbance as function of wavenumber. The number of vibrational modes a molecule has depends on the number of atoms in the molecule and the degrees of freedom in the molecule. In a simple non-linear molecule consisting of three atoms such as a water molecule there are 3 degrees of freedom which consist of the vibrational modes symmetrical stretching, asymmetrical

stretching, and symmetric bending. A non-linear simple molecule of 3 atoms such as CO_2 has 4 degrees of freedom consisting of the vibrational modes symmetrical stretching, asymmetrical stretching, symmetrical bending, and asymmetrical bending. The symmetrical bending mode is inactive for IR spectroscopy, producing no band because there is no change in the dipole moment.

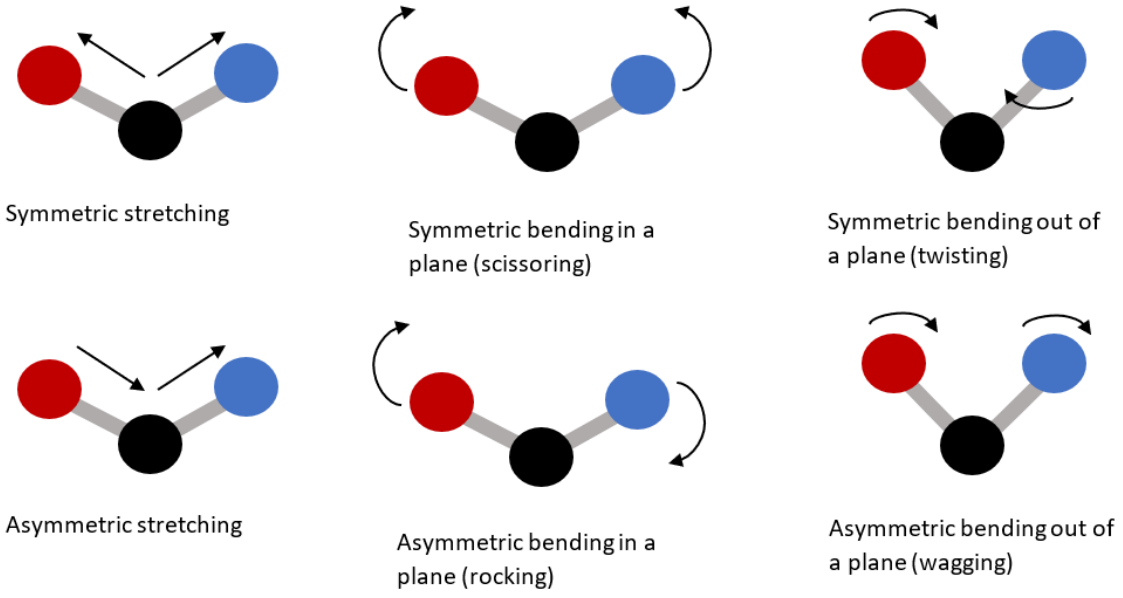

Figure 1 Modes of vibration in molecules from the absorbance of IR radiation.

Most FTIR spectrometers consist of an IR light source, interferometer, sample compartment and a detector. The IR light source generates radiation that passes through the interferometer, through the sample and into the detector. The interferometer as shown in Figure 2, is core to the functioning of FTIR spectrometers. The interferometer consists of two perpendicular mirrors and a beam splitter. The beam splitter splits the IR beam into two beams which are recombined in the interferometer and conducted into the detector. The beam splitter transmits half the light and reflects the other half to produce the two beams. One of the mirrors is stationary while the other moveable. The reflected light and

transmitted light hit the stationary and moving mirror respectively. The beams are reflected by the mirrors and recombined at the beam splitter. If the path travelled by the beams is the same, it is called the zero-path difference. But when the moveable mirror moves away from the beam splitter, a difference in the length of the beams paths is created. The extra difference to the moving mirror is defined as the optical path difference (OPD). The interferogram is a function of time and the values outputted by this function make up the time domain. A Fourier transform is applied to the time domain to obtain a frequency domain which is deconvolved to produce a spectrum.

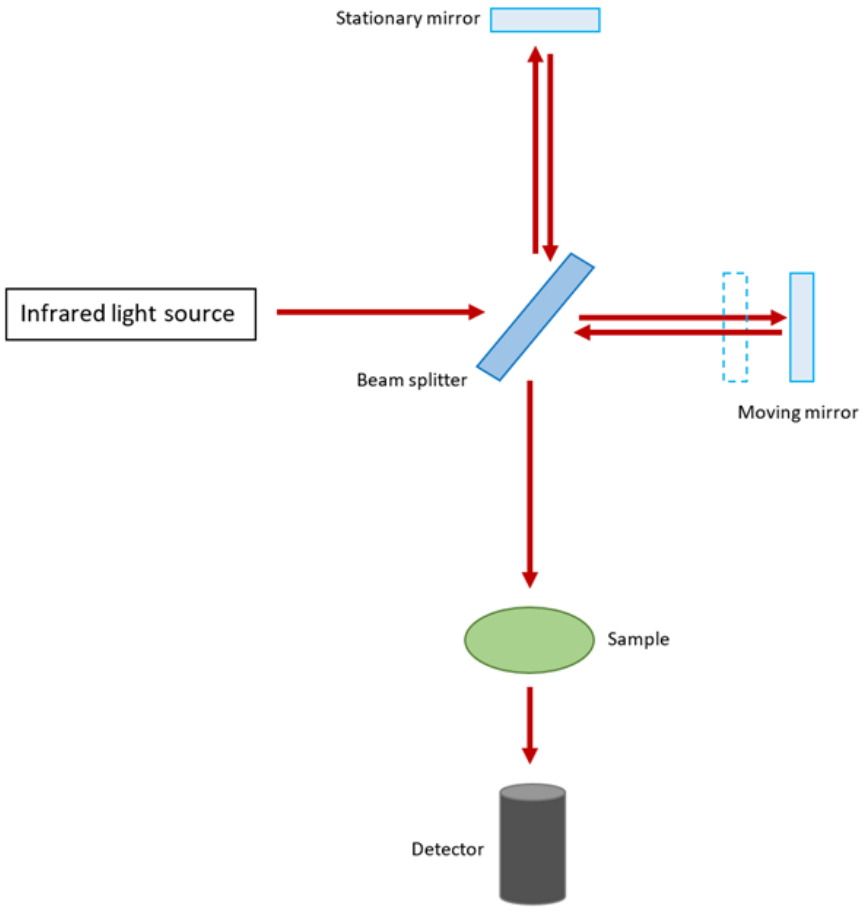

Figure 2 Diagram of a Michelson interferometer.

FTIR spectroscopy IR light sources

The most ubiquitous infrared source in benchtop FTIR instruments is a globar source (Hermes *et al.*, 2018). Globar radiation sources consist of a silicon carbide rod. An electric current is passed through the rod, heating it up to 1300 °K. Due to the high temperatures produced, a cooling system is needed to prevent arcing, thus the globar is surrounded by a water jacket cooled by liquid nitrogen. The radiation source allows the emittance of continuous mid-IR radiation. Globar sources have a large spectral emission range but low spectral intensity due to the principle of black body radiation.

Synchrotron light sources allow a superior signal-to-noise ratio and higher resolution spatial mapping of samples compared to globar sources. Synchrotron sources work on the principle based on the Bremsstrahlung effect. Whereby electrons are accelerated to relativistic speeds while the transverse part of their momentum is modulated by an array of magnets, resulting in the emission of electromagnetic radiation that spans a large range of wavelengths (Hermes *et al.*, 2018). The electrons are accelerated by a linear accelerator (LINAC) to form a stream of electrons. The electron stream is accelerated by a series of particle accelerators in the booster ring until it becomes a stable beam in the storage ring. The rings contain a magnetic lattice used to curve the electrons between straight sections. As electrons follow the curved path, they emit electromagnetic radiation that is siphoned off into beam lines for specific uses such as IR radiation for FTIR spectroscopy. Figure 3 shows the general design of synchrotrons described here.

Synchrotron technology has shown rapid growth in its use since synchrotrons were first used in the 1950s. Synchrotrons are currently in their third generation with fourth generation synchrotrons on the horizon. Third generation synchrotrons have higher resolutions than earlier generations and have many straight sections for insertion devices. The insertion devices are magnetic arrays that bend the electrons at specific curvature radius to produce light at energy levels with specific characteristics for specific purposes (Huang, 2013). There are two main types of insertion devices, wigglers and undulators. Using wigglers, the objective is to apply an intense magnetic field locally to obtain energetic X-rays and repeat the oscillation several times in a longitudinal direction. The light produced at wigglers is a high energy and intense beam. With undulators the light that emerges at each wiggle interferes with light at other wiggles and produces an interference pattern in both the spatial and energy planes. The light generated is spatially very concentrated into a narrow cone and in several specific energies called harmonics. Undulators are used when extremely brilliant light is required and are used for FTIR experiments.

This last decade has seen the development of quantum cascade lasers (QCL) as a reliable light source for IR spectroscopy. A QCL is a heterogenous diode laser where the IR radiation is generated by applying a voltage to the diode (Childs *et al.*, 2015). The heterogenous nature of a QCL allows light to be emitted in a range of wavelengths in the mid and far IR region. The wavelengths emitted by the QCL can be selected by tuning in where the QCL is placed in an external cavity and a grating is tilted. The diode of the QCL is formed from distinct stacked semiconductor layers. QCL are becoming a more popular light source for IR spectroscopy due to their advantages of having excellent signal-to-noise ratios and having a quick acquisition time making them good for applications like chemical imaging. QCL have these advantages as it emits all its photons at approximately the same wavelength. This

allows for the full dynamic range of detectors to be utilised to detect signals at single wavelengths. This differs from traditional FTIR spectroscopy thermal sources where the photons are spread across a broad range of wavelengths. QCL sources By using the full dynamic range of the detector can use uncooled detectors such as deuterated alanine doped Tri-Glycine Sulphate (DTGS) detectors for microspectroscopy and imaging applications. An instrument with a traditional globar source used for FTIR microspectroscopy or FTIR imaging applications requires the use of a mercury cadmium telluride (MCT) detectors to achieve good signal to noise ratio. MCT detectors commonly must be cooled using liquid nitrogen which adds extra cost and must be re-filled during prolonged use which also takes time. The uncooled DGTS detectors can be used at room temperature.

The light source of an instrument is an important consideration when using IR spectroscopy as they all have their advantages and disadvantages. Globar sources are good for obtaining the whole spectra and are widely available and the most affordable source. Synchrotrons provide much higher spatial resolution because of the high brilliance produced which allow for smaller aperture sizes to be used to gain the higher spatial resolution. However, a specialised synchrotron facility is required therefore accessibility can be a problem. QCL provides excellent signal to noise ratio and fast acquisition times but struggle when a broad spectral reading is required. The right light source should be chosen depending on the needs of what is being measured.

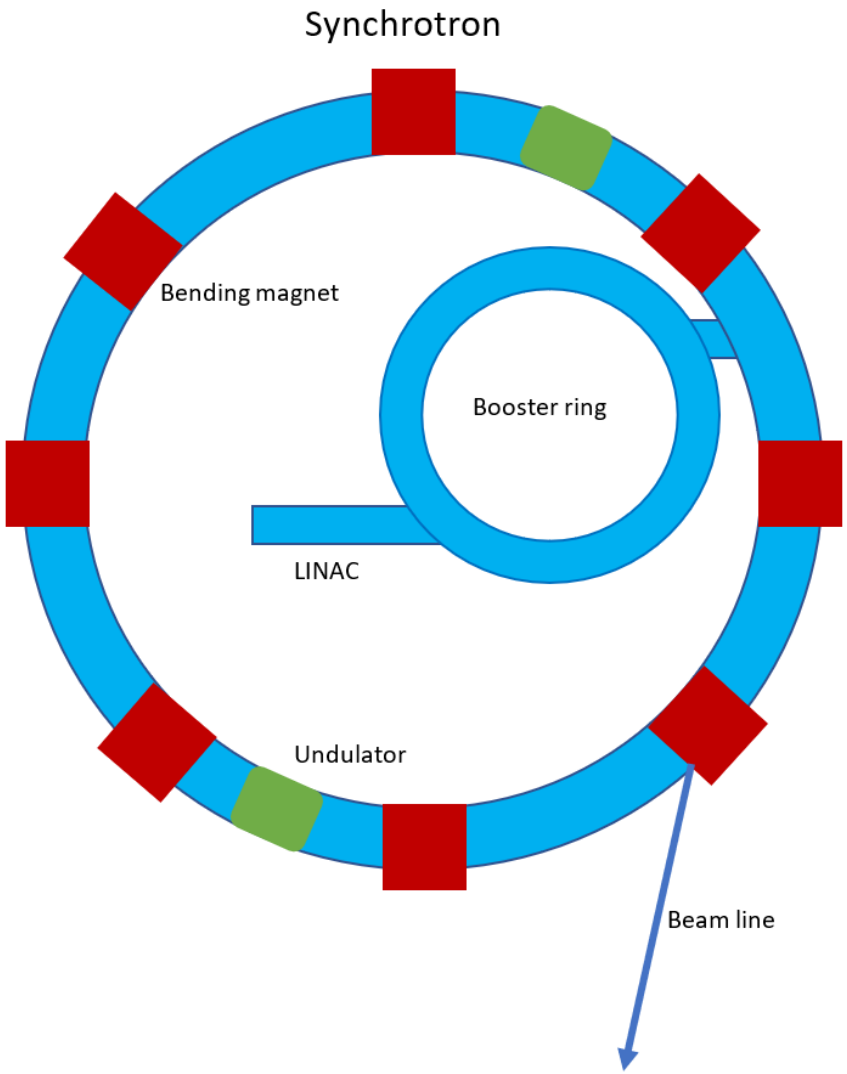

Figure 3 Diagram of the basic layout of a synchrotron.

FTIR spectroscopy analysis of biological materials

FTIR spectroscopy is used to analyse the chemical structure of materials. As such it can be utilised to analyse the biochemical structure of tissues and cells in a non-destructive manner. Biological samples are comprised of several key groups of biomolecules: proteins, lipids, nucleic acids and carbohydrates. Biomolecule will produce an IR spectrum based upon the bonds between the atoms that make up the molecule. A tissue or cell type will produce a

spectrum reflecting its biochemical make up, providing information on what types of biomolecules are present within the sample. The key groups of biomolecules can be identified by spectral bands at certain ranges of wavenumbers with an identifiable band structure (Baker *et al.*, 2009). FTIR analysis of biological samples is conducted in the mid-IR range 400 cm^{-1} – 4000 cm^{-1} (25 μm – 2.5 μm). The bands between 3000-2800 cm^{-1} in the are produced by stretching vibrations of C-H in CH_3 and CH_2 groups in the acyl chains. The main contributor to these band is from the fatty acid chains of lipids also with some contribution from proteins. Amide bands representing the protein fraction of the sample appear at 1700-1310 cm^{-1} in three spectral bands: amide I, II and III (Diem, M. Romeo, *et al.*, 2004). FTIR spectroscopy can provide information about the secondary structure of proteins in a sample, that is stretching vibrations of C=O bond represented by the amide I band and bending vibrations of N-H with stretching of C-H bond in the amide II band. The amide I band appears at 1700-1600 cm^{-1} with an intense peak at 1650 cm^{-1}. Amide II band occurs at 1575-1480 cm^{-1}. At 1301-1229 cm^{-1} the amide III band appears because of intracellular proteins. Like amide I, the amide III band results from vibrations from C-N and N-H bonds. Broad bands are produced by at 3400-3380 cm^{-1} due to the O-H bond stretching vibrations and another band is produced at 2930-2900 cm^{-1} due to CH_2 and C-H stretching. At lower ends of the mid-IR spectrum bands from carbohydrates can be seen at 1200-950 cm^{-1} from C-O, C-C stretching and C-OH bending and bands at 950-700 cm^{-1} from C-OH, C-CH, O-CH and C-H bending vibrations. Bands between 1250-1080 cm^{-1} correspond to vibrations from phosphate containing groups primarily from the backbone of nucleic acids.

Wavenumbers (cm^{-1})	Macromolecules assignment	Bond assignment
4000-3100	Proteins	-OH and -NH stretching mode, amide A band.
3100-2800	Lipids & proteins	-C-H symmetric and asymmetric stretching vibrations of CH_2 and CH_3.
1735	Lipids	Ester C=O stretching.
1695-1615	Proteins	C=O stretching.
1550-1520	Proteins	N-H bending.
1500-920	Carbohydrates & nucleic acids	Phosphate groups, CH_3 bending, C-O stretching.

Table 1 Band allocations for FTIR spectra of biological materials.

There are three main FTIR spectroscopy modalities used for the analysis of biological materials: transmission, attenuated total reflection (ATR) and transflection (Baker *et al.*, 2009). Each mode has inherent advantages and disadvantages, and mode selection is determined by the types of samples being measured. With transmission FTIR spectroscopy the IR radiation is passed through the sample and the transmitted radiation is measured. The spectra obtained from the transmission FTIR will be representative of the bulk of materials.

Using ATR FTIR spectroscopy, the IR beam is directed onto an optically dense crystal that generates an evanescent wave that extends beyond the crystal's surface onto a sample in direct contact with the crystal (Figure 4). The evanescent wave will be attenuated where the sample absorbs IR radiation. The attenuated beam returns to the crystal where it exits at the opposite end of the crystal from where it entered and is directed to the detector. ATR FTIR spectroscopy typically measures 0.5-2 μm deep into a sample due to the intensity of the evanescent wave decaying exponentially with distance from the surface of the ATR crystal. Therefore, ATR FTIR measures the properties of the surface of the material and just below the surface unlike transmission FTIR where the bulk of the material is measured. The

evanescent wave occurs because of total internal reflection resulting from the differing refractive index of the ATR substrate and the sample. When the IR beam hits the surface between the ATR substrate and sample which are characterised by differing refractive indices at a certain angle of incidence the light is totally reflected. This angle of incidence is referred to as the critical angle. Snell's law can be used to calculate the critical angle. Snell's law states that the ratio of two refractive indices is equal to the inverse ratio of the angle of incidence and the angle of refraction. For the special case of no refraction, the angle of incidence becomes the so-called critical angle.

There are three forms of reflectance FTIR spectroscopy; transfelction, diffuse reflectance and specular reflectance. Tranflection has been used in many studies for the analysis of biological materials whereas the two others are more common for analysing materials. In transflection mode, the sample is placed on a reflective substrate and measurements are generated from the IR beam travelling through the sample and being reflected by the substrate through the sample and, as such the bulk of the material will be measured. Another reflectance mode of FTIR spectroscopy is diffuse reflectance where a sample cup is filled with a mixture of a transparent matrix such as KBr and the sample. The IR light is scattered off particles within the sample in all directions and collected with mirrors to be directed to a detector. Diffuse reflectance is commonly used to analyse samples of a particulate nature. The third mode of reflectance is specular reflectance where the IR light is reflected off the sample surface. This method is used for analysing reflective materials.

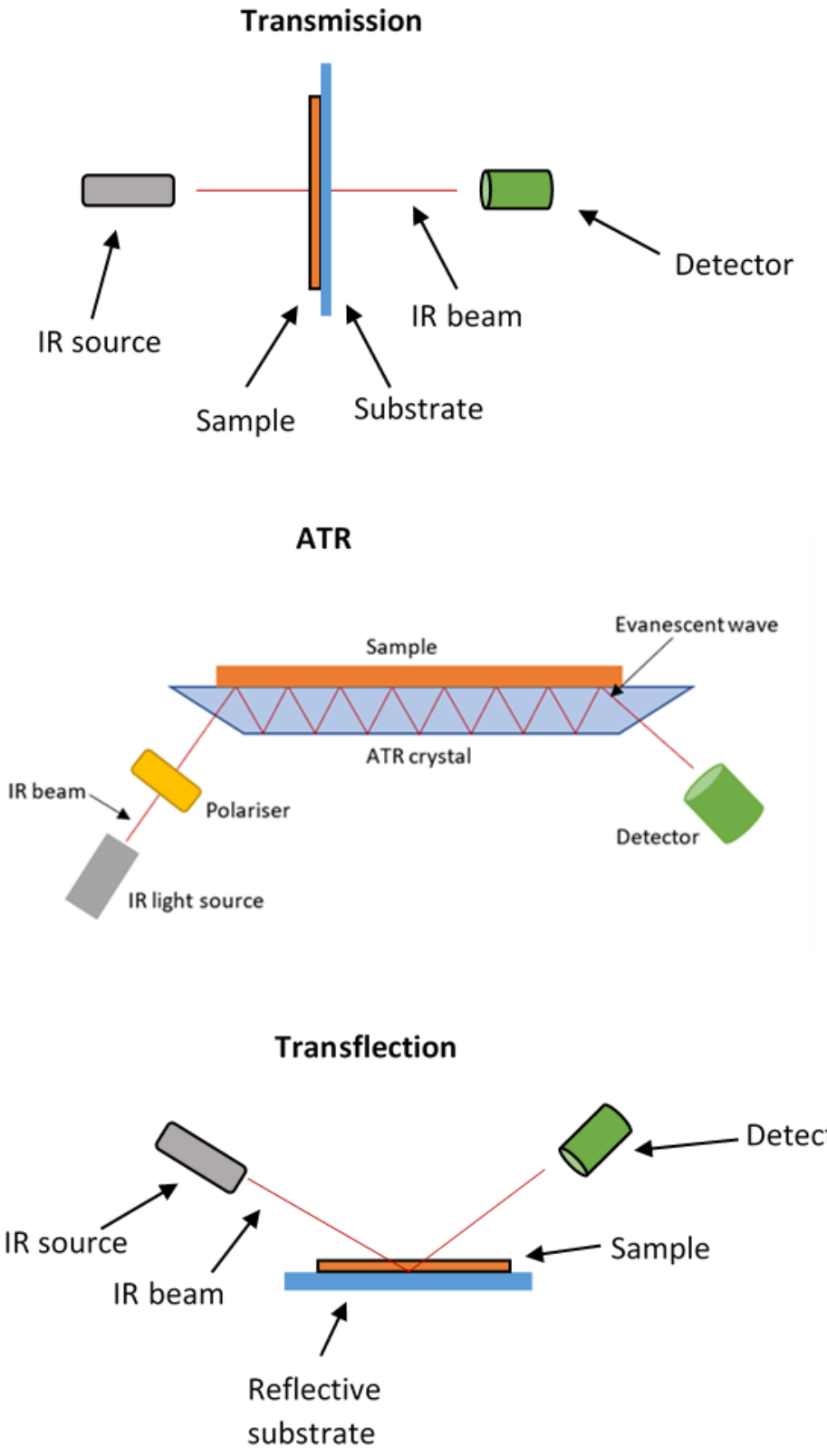

Figure 4 Modes of FTIR spectroscopy commonly used for analysis of biological materials transmission, ATR and transflection.

FTIR micro-spectroscopy and imaging

FTIR microspectroscopy and FTIR spectroscopy imagining are widely used applications of FTIR spectroscopy for the study of biological materials and research into its potential clinical applications. Spectrometers used for FTIR microspectroscopy are attached to a microscope. This allows the cells and tissues to be viewed under a microscope and the IR beam directed at certain cells and areas of tissue. An FTIR microscope can be used for FTIR imaging. Chemical information obtained by FTIR spectroscopy can be combined with the visual topographical images from the microscopy. This allows chemical distribution of biomolecules in cells and tissues to be viewed and false colour images based on the biomolecule content to be produced. With FTIR imaging biological samples can be digitally stained based on biomolecule distribution without altering or destroying the sample itself. FTIR imagining become more practical in the 1990s with the advent of Focal plane array (FPA) detectors. An FPA detector consists of an array of detectors that allow the capture of many spectra over an area at the same time. This makes FTIR imaging much quicker than when using a single element detector. The advancement of QCL IR sources in the last decade is further speeding up acquisition times of spectra for IR imagining.

FTIR spectroscopy as a clinical tool for cancer

FTIR spectroscopy has been investigated as clinical tool for several cancers including breast (Backhaus *et al.*, 2010), colon (Khanmohammadi *et al.*, 2011), prostate (Baker *et al.*, 2009), ovarian (Paraskevaidi *et al.*, 2018), oral (Menzies *et al.*, 2014), bladder (Gok *et al.*, 2016), skin (Kyriakidou *et al.*, 2017) and lung cancers. The first research using FTIR spectroscopy to measure cancer was by Woernley in 1952. Research for FTIR spectroscopy as a clinical tool

for cancer has included its use for detection, diagnosis, response to treatment and follow-up after treatment demonstrating the possibilities for FTIR spectroscopy to be used as a tool in the clinic for the management of cancer from the initial stages of detection and for monitoring the success of treatment or potential relapses.

Effective management and treatment of cancer requires accurate staging and grading of the cancer to best predict the behaviour of the disease, the most suitable treatments and the prognosis of the patient. Most cancers are graded using histology of biopsy samples which are often subjective, possibly leading to misdiagnosis along with being an invasive procedure. FTIR could be used to more accurately stage and grade cancer while being less invasive than current histological methods. One study employed ATR-FTIR spectroscopy coupled with variable selection methods, successive projection algorithm or genetic algorithm combined with linear discriminant analysis (LDA), in order to identify spectral biomarkers in blood plasma or serum samples for the diagnosis and staging of ovarian cancer by histological type and segregation based on age (Lima *et al.*, 2015). In this study 100% sensitivity and selectivity was achieved in the <60 years of age and >60 years of age categories in plasma blood using 42 wavenumbers by GA-LDA. This study demonstrates FTIR spectroscopy can be used to accurately stage ovarian cancer from biomarkers in blood serum and plasma samples thus reducing pain and risk to patients associated with a biopsy.

Surgery is one of the most common treatments for solid tumours. To ensure that the tumour is completely removed during surgery some normal tissue is resected along with the tumour. It is important that resection margins are adequate to avoid under- or over-treatment. Too small of a resection and it may cause reoccurrence of the tumour while too large of a resection will result in prolonged recovery for the patient. Yao, Shi and Zhang used FTIR

spectroscopy coupled with an ATR optical fibre probe to assess the surgical resection margins for colorectal cancers (Yao, Shi and Zhang, 2014). Spectra of the colorectal tumours as well as mucosa 1, 2 and 5 cm from the tumour were measured. The spectra obtained from the tumour and the mucosa 1 cm away were different to the mucosa 2 and 5 cm away with the former site having a decrease in lipids and an increase of proteins and nucleic acids. This technology could be developed to allow real-time assessment of resection margins to avoid unnecessary removal of healthy tissue and ensure the tumour is completely removed thus reducing trauma from unnecessary removal of healthy tissue or from repeat surgeries following relapse.

Monitoring of treatment is essential for the management of cancer and planning personalised medicine. While in recent years treatments for many cancers has improved thanks to targeted therapies and immunotherapies along with better understanding of the molecular genetic causes of cancer, recurrences still often occur. Follow-up post treatment is important for early detection of relapse and secondary tumours, to monitor potential side effects of treatment and to provide psychological and mental health support to the patient. FTIR spectroscopy could be used to detect a relapse or the efficacy of a treatment. Zelig et al used FTIR microspectroscopy to obtain spectra from peripheral blood mononuclear cells (PBMCs) isolated from childhood acute leukaemia patients as well as from patients with a high fever and healthy people as controls (Zelig *et al.*, 2011). Leukaemia was found to be indicated in the spectra by reduced lipids and elevated DNA absorption. These markers diagnostic of leukaemia compared to the controls were used to monitor for biochemical changes in the PBMCs during chemotherapy. This demonstrates how FTIR microspectroscopy could be used as a pre-screening tool and for follow up of treatment for leukaemia.

FTIR spectroscopy as a clinical tool for lung cancer

As discussed in a previous section lung cancer survival rate remains low much in part due to cases being diagnosed in late stages of disease. Therefore, it is of great importance that suspected cases of lung cancer are diagnosed in a timely and efficient manner. New methods to diagnose and monitor lung cancer that are quick and cost effective while reducing the increasing loads on pathology departments due to an ageing population are needed. Disease processes cause a biochemical change in cells and tissues often before morphological changes and symptoms can be seen. FTIR spectroscopy can be used to detect these biochemical changes.

Wang, Wang and Huang were one of the first groups to investigate the use of FTIR spectroscopy as a diagnostic tool for lung cancers (Wang, Wang and Huang, 1997). Their study utilised FTIR spectroscopy in a transflection modality to distinguish normal cells, lung cancer cells and TB cells in pleural effusion. They found that the ratio of peak intensities at 1030 cm^{-1} and 1080 cm^{-1} that represent glycogen and the phosphodiester bonds of nucleic acids respectively, were significantly increased in lung cancer cells compared to normal cells with a higher intensity in peaks in the noted regions. Yano et al analysed human cancerous lung tissue using transmission FTIR microspectroscopy (Yano *et al.*, 2000). The lung tissue sections were mounted to CaF_2 windows which are commonly used as a substrate because of their low change in refractive index and have a transmission over 90% allowing most of the signal to reach the detector. Like Wang et al, Yano et al found an increase in nucleic acids and glycogen of cancerous samples compared to non-cancerous samples. Lewis et al used

ATR-FTIR spectroscopy combined with hierarchical cluster analysis (HCA) and principal component analysis (PCA) to analyse lung sputum cell pellets (Lewis *et al.*, 2010). The sputum was obtained from a range of lung cancer patients including NSCLC and SCLC. Mirroring *Wang et al* and *Yano et al, Lewis et al* found there was increase in glycogen in cancer cells compared to healthy cells indicated by an increase in peak intensity at 1024 cm^{-1} and 1049 cm^{-1}.

The three studies demonstrate the range of sample types that could possibly be used to diagnose or monitor lung cancer using FTIR spectroscopy, and all studies showed that cancerous samples were distinguishable from healthy samples. The sample type used is an important consideration as the protocol including the mode of FTIR spectroscopy and type of detector will be dependent on the type of sample. Therefore, for the clinical use FTIR spectroscopy the type of sample must be carefully considered. Sputum like the one used by *Lewis et al* is an attractive choice due to the minimal invasiveness in obtaining it. Obtaining the sample should ideally be minimally invasive for the patient and is especially important for detection, diagnosis and monitoring of lung cancer where most the patients are elderly and for who invasive procedures provide a significant risk.

FTIR spectroscopy as a clinical tool for breast cancer

There have been several studies investigating how FTIR spectroscopy could be used for the detection and diagnosis of breast cancer. These studies have included the study of a range of materials including tissue, cells and biofluids. Mostaço-Guidolin et al (Mostaço-Guidolin et al, 2010) demonstrated the use of FTIR spectroscopy to characterise the ER+ cell line MCF7 and the ER- cell line SKBr3. They demonstrated differences in the bands 1087 cm−1 (DNA), 1397 cm−1 (CH3), 1543 cm−1 (amide II), 1651 cm−1 (amide I), 2924 cm−1 (fatty acids)

demonstrating FTIR spectroscopy could be useful for distinguishing between ER+ and ER- breast cancers. Around the same time Rehman et al (Rehman et al 2010) demostratd the us of ATR FTIR spectroscopy to identify spectral differences in tissue samples from normal breast, invasive ductal carcinoma, and ductal carcinoma in situ. They found significant differences in the bands between different grades of tumour suggesting FTIR spectroscopy could be helpful for grading breast tumours. More recent research (Tomas et al, 2022) has combined the of FTIR spectroscopy with machine learning to classify breast tissue sections. Normal and malignant breast tissue was measured with ATR FTIR spectroscopy and a neural network was used for classification (Tomas et al, 2022). Souza et al (Souza et al, 2023) demonstrated the use of ATR-spectroscopy to discriminate molecular subtypes of breast cancer from plasma samples. An orthogonal partial least squares discriminant analysis model was used to perform the classification with 100% accuracy to classify luminal A, luminal B, HER2+, triple negative and healthy controls. While excellent classification was achieved, a major drawback is orthogonal partial least squares discriminant analysis does not allow new sample data to be added requiring a new model to be made to add new data. This is not an ideal classifier for translation of the method, other classifiers such as random forest, neural network and support vector, machine algorithms can provide good classification while allowing new data to be input. These studies demonstrate a range of applications for diagnosis and detection of breast cancer. Research in this thesis will investigate how FTIR spectroscopy can be used with glass substrates and preparation methods commonly used in pathology laboratories to demonstrate a methodology for the measurement of cytology samples that is more affordable and less disruptive to current workflows in pathology laboratories.

Challenges of bringing FTIR spectroscopy to the clinic

Many studies have shown promising potential for the use of FTIR spectroscopy as a tool for analysing biological specimens and its use in the screening, diagnosis, management and monitoring of cancer. The challenge is to develop and translate these methods to routine clinical applications.

Most of the studies pertaining FTIR spectroscopy for clinical applications have been performed on small sample sizes. Large scale clinical trials will be needed to prove the efficacy of FTIR spectroscopy as shown in the laboratory and to identify any barriers to implementation. For FTIR spectroscopy to be successful in clinical applications it must fit in with clinical workflows which can only be proved with large scale trials.

Standardisation of sample collection, preparation and storage is needed to achieve experimental reproducibility within individual laboratories and between different laboratories. This would allow for results from different studies to be more easily compared. There is currently no standard preparation method of biological specimens for use in FTIR spectroscopy analysis. Standardised methods of sample collection, preparation and storage will be needed for FTIR spectroscopy to be used in the clinic. More research is needed in this area to find the optimum methods that will produce accurate results that can fit into clinical workflows. As shown in the previous sections, a range of biological specimens have been studied with FTIR spectroscopy including cells, tissues and biofluids. The spectroscopic study of each of these biological specimens has their own challenges.

When analysing biofluids the specific biofluid will influence the sample collection, preparation, and storage. A challenge for the spectroscopic measurement of all biofluids is

their high-water content. Water absorbs light in the mid-IR region which can obscure information on biomarkers present in the sample therefore efforts must be taken to remove water from biofluids. Water's absorption of IR light is what also makes FTIR spectroscopy unsuitable for the measurement of live cells and tissue. Biological materials must be dried and fixed to prevent inference from water. This limits the potential of FTIR spectroscopy for in-vivo measurements.

Mie scattering can render spectra unreliable as features seen in the spectrum are due to the morphology of the structures rather than the biochemistry of the sample. Mie scattering causes a broad sinusoidal oscillation in the baselines of spectra. This results in the distortion of both band position and intensity. The spectra of single cells can also exhibit distortion of band shapes most likely a derivative like distortion on the high wavenumber side of the amide I band (Bassan *et al.*, 2009).

The hurdle for using FTIR spectroscopy to analyse cytological and histological samples in the clinic is implementing it into clinical workflows without causing disruption. The time taken for spectral acquisition is currently limiting FTIR spectroscopy applications in the clinic for the analysis of histological samples. Pathology departments analyse hundreds of tissue samples a day and as such a delay in throughput is detrimental. Single point mode analysis of tissue with single point detectors found in many benchtop spectrometers is intrinsically slow (Finlayson, Rinaldi and Baker, 2019). The emergence of FPA detectors allows for much quicker acquisition times. FPA detectors can collect thousands of spectra concurrently enabling a substantial reduction of acquisition time in comparison to single point detectors. However, the analysis of large tissue sections could still take several hours with a FPA detector. A possible solution to this is the use of tissue microarrays where multiple tissue

cores of submillimetre dimensions are placed onto single slides allowing the analysis of multiple tissue specimens at once (Kwak *et al.*, 2011). These recent advances in FTIR spectroscopy could help to mitigate the challenge of lengthy acquisition times that would disrupt clinical workflows.

Current cytological and histological investigations in pathology labs frequently use stains like haematoxylin and eosin staining (H&E) and Papanicolaou (Pap) stains, that can influence the spectral signatures obtained from FTIR spectroscopy. If FTIR spectroscopy is to be used in conjunction with current pathology methods to further characterise abnormal specimens, the stains could obscure spectral markers. Both H&E and Pap stains have shown to cause a disappearance of peaks in the lipid region at 2850 cm^{-1} and 2920 cm^{-1} in the cell lines CALU-1 and NL20 (Pijanka *et al.*, 2010). This was thought be due to the ethanol used in the staining procedures for both stains. The ethanol also removes phospholipids from cells which removes a peak at 1740 cm^{-1}. H&E staining procedure caused an increase of peak intensity at 1305 cm^{-1} associated with the amide-III band. The malignant CALU-1 cell line was still distinguishable from the non-malignant NL20 cells with FITR spectroscopy after the staining procedures and their alteration to the spectra. The effect of staining on the FTIR spectra must be considered on stained samples if the technique is to be used in the clinic.

The cost of substrates for FITR spectroscopy must be considered especially in public healthcare systems like the National Health Service (NHS). The CaF_2 and BaF_2 slides commonly used for transmission FTIR spectroscopy can cost £60 per slide which would not be economically viable (Finlayson, Rinaldi and Baker, 2019). Glass slides typically used in pathology laboratories for cytology applications are not reliable for FTIR spectroscopy because the glass absorbs IR radiation therefore features of the spectra in the fingerprint

region are lost. Work at Keele University has investigated the used of thin soda-glass slides 0.12-0.17 mm thick (Rutter *et al.*, 2019). Using these slides the lipid, amide I and amide II regions of cells are identifiable and allowed the cell lines CALU-1, K562 and PBMCs to be distinguished from each other. Glass slides are affordable and easy to procure. If glass substrates can be used it would remove the prohibitive costs of CaF_2 and BaF_2 slides.

Instruments from different manufacturers can produce distinct responses and spectral distortions. The differences caused by instrumentation should be considered when comparing results and must be accounted for in possible clinical applications. The environmental variations should be addressed using pre-processing algorithms to compare results from studies using different instruments therefore standardisation in sample preparation, data collection, data pre-processing and analysis are paramount. Currently there is little standardisation the field of clinical vibrational spectroscopy which makes comparison across studies difficult. Lack of standardisation is currently hampering translation of FTIR spectroscopy to the clinic.

Other vibrational spectroscopy techniques

O-PTIR spectroscopy

O-PTIR spectroscopy is a relatively new technique compared to FTIR and Raman spectroscopy. Figure 5 below demonstrates the basic schematic of an O-PTIR spectrometer. O-PTIR spectroscopy combines the modalities of FTIR and Raman using a pump probe system (Kansiz and Prater, 2020). The pump is an IR QCL which is used to excite vibrational

modes in the samples. The pulsed QCL generates a local IR photothermal event. The photothermal response manifests through subtle thermal expansion and refractive index changes. This response is monitored by the optical probe which is typically a 532 nm laser. The changes in probe intensity as a function of IR wavelength tuning of the QCL are demodulated by a lock-in amplifier. The generated spectrum is an IR absorbance spectrum similar to that recorded by transmission FTIR spectroscopy. How the O-PTIR spectroscopy interacts with the substrates that the sample is placed on will differ to transmission FTIR spectroscopy because the IR beam is not travelling through the whole of the substrate. How O-PTIR spectroscopy interacts with alternative substrates such as glass in theory will be different to that of conventional FTIR spectroscopy. Since O-PTIR is a much newer technique there has been little research into how it interacts with different substrates. An advantage of O-PTIR is its improved resolution in comparison to conventional FTIR spectroscopy. Traditionally the resolution of IR spectroscopy is limited by the wavelength of the IR beam to 10-20 μm. The spatial resolution of O-PTIR is limited by the wavelength of the visible beam up to 0.5 μm. Some of the first research using O-PTIR spectroscopy to analyse biological materials was in 2020 to measure collagen from tendons and image amyloid aggregates in neurons (Bakir et al., 2020) (Klementieva *et al.*, 2020). In 2021, O-PTIR was demonstrated to be able to measure the lipid bands and amide I and II bands of live cells in an aqueous environment from cancer cells lines (Spadea et al., 2021). The body of research investigating the potential for O-PTIR spectroscopy for use in cancer diagnostics is still small because of its recency as a technique. Bouzy et al has explored how O-PTIR combined with Raman imaging could be used to investigate the composition of microcalcifications in breast cancer. Microcalcifications are mostly benign, but some can be indication for precancerous lesions.

Better understanding of their composition through vibrational spectroscopy could help to inform which calcifications are indicative of cancer.

In this thesis it was investigated how O-PTIR performs for the classification of NSCLC cells from each other and from non-malignant lung cells on glass substrates which had not been investigated previously. The current limitations of O-PTIR are that the technique has a small user base and there is currently only a small amount of research utilising it. This is a barrier to using the technique for diagnostics because of a limited availability of instruments and people who know how to use the instruments and process and interpret the data.

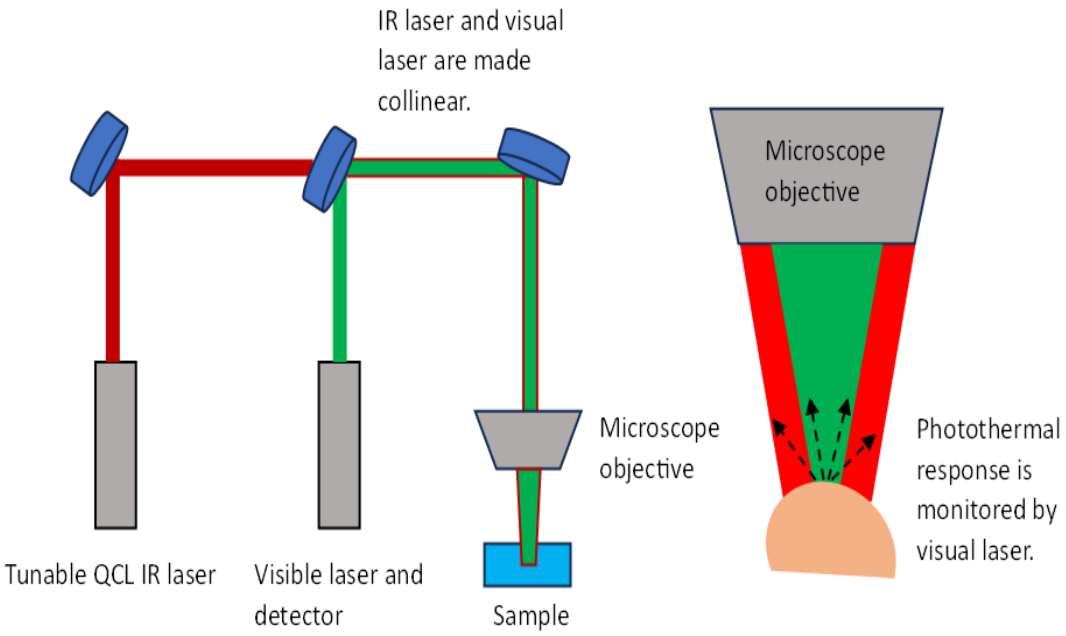

Figure 5 O-PTIR basic schematic.

Raman spectroscopy

Raman spectroscopy is another vibrational spectroscopy technique that has also increasing research into its use for clinical diagnostics. Like FTIR spectroscopy, Raman spectroscopy can be used to gain information on the chemical structure of a sample in a label fee and non-destructive manner. Raman spectroscopy uses a laser wavelength in the visible region of the electromagnetic spectrum to irradiate the sample and cause molecular vibrations. This results in the scattering of light when the sample is irradiated. When the scattering occurs, many of the scattering events is elastic scattering also known as Rayleigh scattering. Elastic scattering is where the energy of the molecule is unchanged after integration with photons. A small portion of the scattering, approximately from 1 in 10 million interactions with photons, inelastic scattering occurs which us also known as Raman scattering. When inelastic scattering occurs there is a transfer of energy between the molecule and the scattered photon. If the molecule is excited to a higher vibrational level from gaining energy from the photon, the photon in turn loses energy and its wavelength increases. This phenomenon is called Stokes Raman scattering. Anti-Stokes Raman scattering is the inverse where the molecule loses energy and the photon gains energy decreasing its wavelength. The majority of molecules are in the ground vibrational level therefore Stokes scatter is more likely to occur. Stokes scatter is most commonly more intense than anti-stokes and for this reason in conventional Raman spectroscopy Stokes Raman scatter is measured in Raman spectroscopy. The energy changes in the modes of scattering are demonstrated in figure 6 below. Each peak on a Raman spectrum corresponds to a specific molecular bond vibration. The intensity of the spectrum is proportional to the concentration of the measured molecules and the scattering cross section of the molecules (Mulvaney and Keating, 2000).

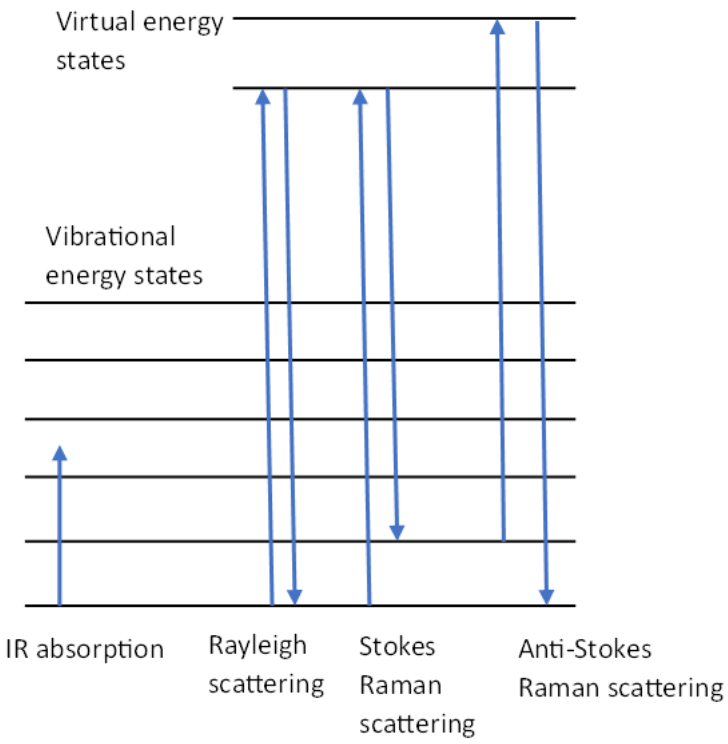

Figure 6 Jablonski diagram showing the energy changes in scattering.

Liquid Biopsies

Liquid biopsies are the use of tumour derived materials in biofluids for the diagnosis, prognosis, and monitoring of cancers. There are various biofluids that could be used for liquid biopsy including blood, urine, saliva, cerebrospinal fluid, sputum, and pleural effusion. There is a range of tumour derived materials that are being researched for use in liquid biopsies including circulating tumour cells (CTCs), extracellular vesicles (EVs), circulating tumour DNA (ctDNA) circulating tumour RNA (ctRNA) and tumour educated cells (Lone *et al.*, 2022). This section will overview the potential of the tumour derived materials found in biofluids and the current limitations that have so far prevented the use of liquid biopsies being utilised in healthcare.

As discussed in earlier sections of this chapter, the current gold standard for the diagnosis of solid cancers like lung and breast cancer requires the use of tissue biopsies from the tumour site. Tissue biopsies are invasive for the patient and are associated with several limitations including patient risk during surgery, cost, sample preparation, and the large amount of expertise required. The use of liquid biopsies would allow for less invasive diagnostic methods that would not require surgical procedures to obtain tumour material. Liquid biopsies could be used for more regular sampling and monitoring of cancer that would not be possible using solid tissue biopsies.

ctDNA is cancer originated DNA that is circulating cell-fee. ctDNA can be short nucleosome associated fragments 80-200 bp in size or much longer fragments >10 kb in size encapsulated in EVs (Ma *et al.*, 2015). There are various mechanisms that introduce ctDNA into circulation including necrosis, apoptosis, cell lysis and release of DNA by the tumour itself. Proof of ctDNA being suitable as a biomarker was demonstrated with the identification of KRAS gene mutations in ctDNA found in the blood of pancreatic cancer patients. Qualitative and quantitative information can both be gained from ctDNA. Qualitive information on tumour mutations can be obtained as shown for KRAS and EGFR mutations. As a quantitative measure, ctDNA can be used to indicate tumour bulk. ctDNA has a short half-life of about 2.5 h, therefore, the amount of ctDNA can give a real time measurement. Currently, to analyse ctDNA, PCR or next generation sequencing (NGS) are used. The biggest limitation for the use of ctDNA is low detection sensitivity in early stages of cancer with ctDNA often accounting for <0.01% of total circulating DNA. The analysis of ctDNA can be time consuming and expensive because ctDNA first needs to be purified from the blood and from other circulating DNA. The reagents for PCR and NGS are expensive and require expertise to perform them.

Circulating tumour cells (CTCs) are cancer cells that have detached from the tumour and entered the peripheral blood circulation (Sundling and Lowe, 2019). They are thought to be a driver of metastasis as CTCs or clusters of CTCs can migrate to other areas of the body through the circulation and multiply to form secondary tumours. CTCs can be generated from primary or secondary tumours from active intravasation or from passive shedding. CTCs can provide a wealth of information on the tumour it originated from as they contain the whole of the DNA and RNA of the cancer cell while also containing the other biomolecules such as lipids, proteins and carbohydrates. CTCs have the potential to be used for diagnosis, prognosis, treatment monitoring and drug discovery. Ex vivo culture of CTCs is valuable for translational research as they can be used to test potential therapies or understand the biology of different mutations. Culture of CTCs could also be used for personalisation of treatment by testing the sensitivity and efficacy of different drugs. It could also be used to monitor the development of drug resistance with cultures grown at different time points. The number of CTCs can be a prognostic indicator with evidence showing a greater number of CTCs being linked to worse outcomes (Matikas *et al.*, 2022). CTC number can also be useful for monitoring treatment, if the number of CTCs is not reducing or rising, it would suggest that the treatment may not be working and the adjustments to the treatment plan might be needed. Like ctDNA, CTCs can be used to identify mutations to genetically type the cancer helping to produce a treatment plan. CTCs have the advantage over ctDNA that it contains the whole of the DNA unlike ctDNA which is often in fragments, many of which may not provide much diagnostic value. While CTCs do contain proteins for transcriptome analysis, the field of single cell protein is premature, and it is difficult to gain sufficient protein material from single cells (Habli *et al.*, 2020). Current limitations for the use of CTCs are the difficulty in isolating them from the blood as they are in small number

compared to the large number of blood cells. This is because most CTCs do not survive for long due to them being cleared by the immune system or being destroyed by the mechanical forces in the blood microenvironment. Isolation techniques for CTCs can be put into two broad groups, separation based upon physical properties or on biological properties. Currently with all the isolation techniques there is contamination from non-CTCs after isolation so identification methods of CTCs must also be developed. CellSearch is the only CTC isolation and detection method that has gained regulatory approval, being approved by the FDA in 2004. Despite CellSearch being available for almost two decades it has had very little adoption by the medical community because of its complexity, difficulty to use and high cost (Andree, van Dalum and Terstappen, 2016). CellSearch utilises antibodies against epithelial markers on CTC, but these markers are often down regulated after EMT, therefore many CTC can be missed.

Extracellular vesicles (EVs) are lipid membrane bound vesicles secreted by cells to mediate intercellular communication. They are secreted by all cell types and can be found in almost all biofluids. EVs are separated into two broad categories, exosomes and microvesicles, based on their content, biogenesis and secretory pathways. EVs contain both DNA and RNA material so like ctDNA and CTCs, they can be used to investigate mutations in the cancer. KRAS and TP53 mutations have been detected in exosomes from the serum of pancreatic cancer patients (Liu *et al.*, 2021a). Again, like ctDNA and CTCs, EVs can provide a quantitative measure of the severity of the disease through the number of EVs originating from the cancer. There is evidence of there being a greater number of exosomes in breast cancer and pancreatic cancer patients. EVs also contain proteins and there is evidence that an increase in EVs showing cancer related membrane proteins could be indicative of the presence of cancer. The current biggest limitation for the use of EVs is the lack of standardisation and

reproducibility in isolation methods. Like the isolation of CTCs, EV isolation uses physical or biological methods (Liu *et al.*, 2021b). EVs can be fragile and current isolation methods can damage and destroy some of the EVs during the process. In the field, there is a lack of standardisation in defining EVs including both the nomenclature and definition of categories of EVs. The current gold standard for EV isolation is the ultracentrifugation. EV isolation is often expensive and obtaining pure yields is difficult with often a compromise between purity and yield.

ctRNA is cell free RNA in circulation originating from cancer cells. On its own, circulating RNA is unstable with a half of about 15 seconds (Lone *et al.*, 2022). The stability of circulating RNA is improved by association proteins, proteolipid complexes and EVs. As with the other tumour derived materials, ctRNA can provide both qualitative and quantitative information. Most classes of RNA have been found in circulation with the most suitable classes of RNA for diagnostics being messenger RNA (mRNA), long non-coding RNA (ncRNA) and micro-RNA (miRNA) (Alba-Bernal *et al.*, 2020). RNA is analysed with PCR techniques such as QRT-PCR or dPCR for single RNA or small panels. Larger panels can be analysed with RNA sequencing. Similar to ctDNA, ctRNA can be used to identify important mutations that relate to treatments and cancer typing such as KRAS and EGFR. Cancer related gene fusions are another potential biomarker of ctRNA. Many lung cancer related gene fusions have been identified in circulating mRNA. The biggest limitation to the use of ctRNA is their instability that makes analysis time sensitive. Careful sample preparation and treatment is needed for the extraction of ctRNAs as lysing agents and anti-clotting agents could damage the RNA.

Tumour educated platelets (TEP) are the most recent material to be considered as a biomarker in liquid biopsies. Platelets are anucleate cells in the blood that are integral for

haemostasis helping to prevent and control bleeding by the formation of thrombotic clots. While platelets do not contain a nucleus and DNA, they do contain RNA and proteins and can transcribe the RNA to form new proteins. Platelets can react to stimuli by releasing RNA and protein signalling complexes packaged in microparticles. Platelet derived microparticles account for >70% of EVs in the peripheral blood. TEPs are platelets that have taken in RNA content from cancer derived microparticles (In 'T Veld and Wurdinger, 2019). TEPs have been shown to have a role in metastasis and drug resistance. CTCs are often found surrounded by TEPs providing protection from the immune system. An advantage of TEPs is their abundance compared to other tumour related materials in circulation, and the isolation of platelets is much easier. Best et al used RNA sequencing to characterise TEPs from RNA profiles in cancer patients from a panel of six different cancers and healthy controls (Best *et al.*, 2015). They distinguished cancer patients from healthy control with 96% accuracy and determined the location of the cancer with 71% accuracy. Currently there are still a lot questions on how TEPs interact with cancer cells and what effects they have on each other, but also, what effect TEPs may have on other cells.

The main limitation for the different tumour related materials found in circulation is their isolation and identification. Isolating the small amounts of these materials from normal blood cells and non-cancer nucleic acids while maintaining high yields and purity is a difficult task. Along with isolation methods, identification methods must be developed that can identify the tumour related materials from other materials as even after isolation there is likely to be contamination of non-cancerous materials. As much of the methods and technology being tested in this field are relatively new, there is currently very little standardisation. More standardisation to approaches will be needed before liquid biopsy platforms can become common place.

FTIR spectroscopy could be a useful tool for the analysis of biofluid for liquid biopsy diagnosis of cancer. A growing number of studies have investigated FTIR spectroscopy modalities for the measurement of a range of biofluids and components within them including blood, urine saliva and pleural fluid. To measure biofluids with FTIR spectroscopy the samples must be dried first because absorbance by water interferes with bands in the spectra therefore some processing is required of the sample before analysis.

There has been considerable interest in the use of blood serum for cancer diagnosis using FTIR spectroscopy. Some of the leading research in this has been by the company Dxcover who developed a disposable silicon slide for measurement of serum samples with ATR-FTIR spectroscopy. In a clinical study Dxcover investigated the diagnosis of brain cancers from serum (Brennan et al, 2021) training their classification algorithm on 724 retrospective patients and testing on a 385-patient cohort. The results of their classification were compared to CT scans. Their results showed good sensitivity and specificity of 81% and 80% respectively. Liquid biopsy has good potential for cancers such as brain cancers where a tissue biopsy cannot easily be obtained without significant risk to the patient. The use of disposable substrates for ATR FTIR spectroscopy such as the substrate produced by discover reduces the risk of sample contamination and damage from placing the sample on a reusable ATR crystal. Another group (Yang et al, 2021) investigated ATR FTIR spectroscopy measurements of serum for lung cancer diagnosis with a cohort of 92 lung cancer patients and 155 healthy people. They dried the serum on a glass slide then removed the dried spot from the slide to place on the ATR crystal. They reported a sensitivity and specificity of 80% and 91.89% using a partial least squares discriminant analysis. These are a couple examples of many groups that are investigating serum analysis with FTIR spectroscopy for cancer diagnosis. For the diagnosis of lung cancer sputum liquid biopsy with FTIR spectroscopy has

also been researched (Lewis et al, 2010) where it was found that there were spectral differences in bands relating to protein and nucleic acid content in the spectra of sputum in lung cancer patients and healthy controls. Other biofluids that have shown evidence of diagnostic potential for cancer include urine for bladder cancer (Ollesch et al, 2014), bile for biliary cancers (Untereiner et al, 2014) and saliva for head and neck cancers (Falamas et al, 2021).

Despite the growing body of research demonstrating the use of FTIR spectroscopy for liquid biopsy cancer diagnostics the use of FTIR spectroscopy to identify CTCs and other tumour related materials is unexplored. FTIR spectroscopy could be used to identify CTCs because the biochemistry of the CTCs is largely different to the surrounding blood cells. FTIR spectroscopy would not rely on labels like current methods of CTC identification. In this thesis the feasibility of using FTIR spectroscopy for CTC identification was investigated.

Objectives

The Objectives of this thesis are:

1. Develop a methodology for the preparation of cells on glass coverslips for FTIR microspectroscopy analysis.
2. Investigate if lung cancer cells can be classified from non-malignant lung cells prepared on glass substrate using FTIR spectroscopy and machine learning.
3. Investigate if different types of lung cancer cells can be classified from each other prepared on glass substrate using FTIR spectroscopy and machine learning.
4. Investigate if breast cancer cells can be classified from non-malignant breast cells prepared on glass substrate using FTIR spectroscopy and machine learning.

5. Investigate if different types of breast cancer cells can be classified from each other prepared on glass substrate using FTIR spectroscopy and machine learning.
6. Investigate if individual lung cancer cells can be identified from blood cells in a mixed sample using FTIR spectra.
7. Investigate the spectra obtained from using O-PTIR spectroscopy to measure cells on a glass slide substrate and if these spectra can be used for classification of the cells.

Chapter 2: Materials and methods

Cell culture methods

Cells

The research conducted or this thesis used these following cell lines.

Breast:

BT-549 is a triple-negative invasive ductal carcinoma line derived from a tumour of a 72-year-old female that had metastasised to 3 of 7 regional lymph nodes. It is representative of an invasive breast cancer.

MCF-7 is an ER+ breast adenocarcinoma derived from a 69-year-old female and is a non-invasive breast cancer cell line.

MCF-10A is a non-tumorigenic breast epithelial cell line that was derived from a 36-year-old female and is used to model non-cancerous breast cells.

These cell lines were kindly gifted by Dr Gianpiero di Leva, Keele University.

Lung:

A549 is a human lung adenocarcinoma cell line. The line was derived from an epithelial lung tumour of a 58-year-old Caucasian male. The A549 cell line was purchased from the European Collection of Cell Cultures, Salisbury United Kingdom (UK).

CALU-1 is a squamous lung carcinoma cell line. The CALU-1 cell line was derived from a 47-year-old Caucasian male with epidermoid cancer in the lung. The cell line was purchased from the European Collection of Cell Cultures, Salisbury UK.

NL20 is a non-malignant cell line consisting of immortalised human bronchial cells derived from a 20-year-old Caucasian female. The NL20 cell line was established through the transfection of the replication-defective SV40 large T plasmid, p129. The line was purchased from the American Collection of Cell Culture. NL20 is used to model non-cancerous lung cells.

Culture conditions

A549, CALU-1 and MCF7 were cultured in Dulbecco's modified eagle's medium (DMEM) with 4.5 g/L glucose and supplemented with 10% foetal bovine serum (FBS), 1% antibiotic (100x), 1% L-glutamine (200 nM), 1% HEPES buffer solution (1M), 1% non-essential amino acids (100x) and 5% sodium pyruvate (100 nM). Cells were seeded into T75 flasks and media was changed every 2-3 days. A549, CALU-1 and MCF7 were passaged by removing culture media and adding 4 ml trypsin then incubated for 5 minutes. After incubation, the trypsin was neutralised by adding 8 ml culture medium. Cells were collected and spun at 1200 rpm for 5 minutes. The supernatant was discarded, and the cell pellet was resuspended in fresh medium. Cell viability was determined by trypan blue exclusion method. Cells were split and seeded into new T75 flasks; this was done every 7 days.

BT-549 was cultured in Roswell Park Memorial Institute (RPMI) media supplemented with 10% FBS, 1% antibiotic (100x), 1% L-glutamine (200 nM), 1% HEPES buffer solution (1 M), 1% non-essential amino acids (100x) and 1% sodium pyruvate (100 nM). BT-549 was grown as an adherent culture. BT-549 cells were passaged by removing culture medium, adding 4 ml trypsin and incubated for 5 minutes. After incubation, the trypsin was neutralised by adding 8 ml culture medium. Cells were collected and spun at 1200 rpm for 5 minutes. The supernatant was discarded, and the cell pellet was resuspended in fresh medium. Cell viability was determined by trypan blue exclusion method. Cells were split and seeded into new T75 flasks; this was done every 7 days.

NL20 cells were cultured in Ham's 12 culture media supplemented with FBS (4%), $NaHCO_3$ (1.5 Gr/L), glucose (2.7 Gr/L), L-Glutamine (2mM, 1%), non-essential amino acids (0.1 mM, 1%), antibiotics (1%), insulin (5 µg/ml), EGF (10ng/ml), hydrocortisone (0.5 ng/ml). Media was changed every 3-4 days. Cells were passed every 5-6 days.

MCF10A cells were cultured in mammary epithelial cell growth basal medium (MEBM) supplemented with the Lonza Singlequot™ kit containing bovine pituitary extract (BPE) 2.00 ml, human epidermal growth factor (hEGF) 0.50 ml, Insulin 0.50 ml, Hydrocortisone 0.50 ml and gentamicin sulphate-amphotericin (GA-1000) 0.50 ml. All cells were grown as an adherent culture in T-75 flasks and incubated at 37 °C at 5% CO_2. Media was changed every 2-3 days and cells were split once a week or before reaching confluency.

The NL20 and MCF10A cells were detached from flasks by incubation with a dissociation solution consisting of 100 ml HBSS, 5.3 ml FBS, 21 mg EDTA. The NL20/MCF10A were incubated 4 minutes with the dissociation solution in the incubator. Culture medium was then used to neutralise the solution. Cells were collected and spun at 1200 rpm for 5

minutes. The supernatant was discarded, and the cell pellet was resuspended in fresh medium. Cell viability was determined by trypan blue exclusion method.

All cells were incubated at 37 °C and 5% CO_2. All cells were routinely tested for presence of mycoplasma.

Survival assays

Cell viability was determined using the trypan blue assay. 0.4% trypan blue solution was added 1:1 to the cell suspension and mixed thoroughly. The stained suspension was loaded on to a haemocytometer. Non-viable cells take up the trypan blue solution and appear blue when viewed under the microscope while viable cells will be unstained.

Sample preparation

Two methods of sample preparation were carried out to place cells on to the slides, cytospin and smear. Cells were collected from the flasks as described in culture conditions above using trypsin or dissociation media. Pelleted cells were resuspended in 0.9% NaCl and cell concentration was brought to 10^6 cells/ml. 20 µl of the cell solution was pipetted into the cytospin funnel. The cells were spun for 1 minute at 900 rpm which deposited them onto the soda lime glass coverslips (24 x 50 mm x 0.13–0.17 mm thickness, GalvOptics, UK) or glass slides (1 mm thickness, ThermoFisher). The cells were deposited in a circular area 1 cm in diameter. The cells were then fixed with 4% buffered paraformaldehyde (PFA) or methanol.

For the smears, 20 µl of a cell concentration of 10^6 cells/ml in 0.9% NaCl were placed on the edges of a GalvOptics coverslip or a glass slide. A second coverslip was placed on the cell solution and used to spread it over the length of the substrate.

Two fixatives were used, 4% PFA in 0.9% NaCl, or methanol. 100 µl of 4% PFA was pipetted on to the samples and incubated for 15 minutes at room temperature. After incubation, excess PFA was removed by washing once with 0.9% NaCl and three times with distilled water. For fixation with methanol, samples were placed in cold methanol for 2 minutes, then allowed to dry. Washings were not carried out with methanol fixation as it is a volatile compound that evaporates from the sample.

To investigate if cancer cells can be classified from blood cells in a mixed sample, samples of doped blood were made. This was to replicate CTCs in blood and to assess if FTIR spectroscopy is viable as a diagnostic tool for CTCs. Blood was doped with CALU-1 or A549 cells at 100,000 lung cancer cells per 1 ml blood. This is a much higher number of cancer cells than would be found circulating in the blood of cancer patients, but it was done so there was plenty of cancer cells in the samples to measure. Once the blood was doped with the cancer cells, the red blood cells were removed from the blood by lysis with an ammonium chloride potassium (ACK) lysing buffer (Thermo Fisher Scientific). 10ml of ACK lysing buffer was added per 1ml of whole blood and was incubated for 5 minutes at room temperature. After the incubation time with ACK lysing buffer the blood was centrifuged for 5 minutes at 300 x g. The supernatant was removed, and the pellet resuspended in 5 ml saline (0.9%). The pellet contained leukocytes and the doped cancer cells. The remaining cells were centrifuged again at 300 x g for 5 minutes at room temperature. The supernatant was removed, and the pellet resuspended using 0.5 ml saline. Samples of these cells were immediately prepared

using a 35 µl of the cell solution in a cytospin at 900 rpm for 1 minute on to glass coverslips. The cells were immediately fixed on the coverslip with 100 µl 4% PFA incubated for 15 minutes at room temperature. After incubation the excess PFA was poured off and to remove any remaining PFA, the slips were washed once with saline and thrice with deionized water.

All cell lines used to produce samples were harvested from flasks after two weeks of culture for each experimental repeat. Cells were harvested during growth phase at around 70% confluency in the flask. Samples were prepared in at least triplicate in three independent experiments for the research in each chapter. Cells were collected by detaching from flask using trypsin or dissociation media and resuspended as described for each cell lie the cell culture section of this chapter. Once counted by trypan blue exclusion method, the cell number needed to produce samples were placed into a 15 ml tube and diluted with saline to the required concentration of 1 x 10^6 cells per ml for the samples of a single cell line.

FTIR microspectroscopy

FTIR spectra in the mid-IR range were obtained at the benchtop using a Thermo Fisher Nicolet iN10(MX) spectrometer at Loughborough University, UK. The spectrometer was fitted with a mercury cadmium telluride (MCT) detector and cooled with liquid nitrogen. Spectra were collected at 4 cm^{-1} resolution with 256 co-added scans per cell using an aperture size of 15 x 15 µm centred on the centre of the cell. The time taken to collect a

spectrum of a cell was 90 seconds. Background measurements were made under the same conditions on an area of the slide without any cells. The instrument was operated in transmission mode.

Figure 7 Nicolet iN10 spectrometer at Loughborough university.

The same Nicolet iN10(MX) spectrometer was used to take larger map measurements for research on classification of lung cancer cells from leukocytes in a doped blood sample. The spectra were collected at 4 cm^{-1} resolution with 256 co-scans per cell using an aperture size of 15 x 15 µm. Spectra were taken in 10 µm steps in both the x and y planes. The instrument was operated in transmission mode.

Measurements at the ALBA synchrotron MIRAS beamline were recorded using a Bruker Vertex 70 spectrometer with a Hyperion 3000 microscope attached. The spectrometer was fitted with an MCT detector cooled with liquid nitrogen. Spectra were collected at 4 cm^{-1} resolution with 256 co-scans per cell using an aperture size of 15 x 15 µm centred on the cell nucleus. The time taken to collect a spectrum of a cell was around 90 seconds. Background measurements were taken after every 10 measurements on a clear section of the glass substrate. The instrument was operated in transmission mode.

The wider aperture of 15 x 15 µm was used to measure spectral information from all regions of the cell including the cytoplasm and cell membrane. The cancer cells used have an average diameter of 20 µm and the leukocytes have an average diameter of 10 µm therefore when taking measurements information is collected from the different cellular structures.

Figure 8 Bruker Vertex 70 spectrometer with Hyperion 3000 microscope at MIRAS beamline in ALBA synchrotron.

Each spectrum measured was from a different individual cell. For the research in each chapter spectra were measured equally from each repeat across the three experiments. This was to ensure that variation within the cell populations and variation withing measurement conditions were accounted for.

O-PTIR spectroscopy

For the research using O-PTIR spectroscopy measurements of A549 and CALU-1 lung cancer cells were taken using a mIRage O-PTIR micro-spectrometer from Photothermal spectroscopy Corp. The spectrometer used a dual range QCL IR pump beam covering the

spectral ranges of 3000-2700 cm^{-1} and 1800-914 cm^{-1}. The QCL operated at 100 KHz pulse rate and 100% power at 2.5% duty cycle. The probe beam was an optical 532 nm laser operated at 28% power. The spectrometer was fitted with a room temperature silicon photodiode detector to record the reflected optical beam intensity. Spectra were collected at a resolution of 6 cm^{-1} with a single scan per replicate spectrum. A single spectrum took approximately 1 second to scan. Spectra were collected in reflection mode, but the output spectra are transmission like IR spectra because of the pump-probe system of O-PTIR spectroscopy. Background spectra were collected off a clean Kevley Low-E slide once per day. The system was purged with dry nitrogen gas to minimise water vapour. For each individual cell measured nine spectra were recorded across the cell and these nine spectra were averaged to produce the spectra for each cell. 50 A549 and 50 CALU-1 cells were measured in this manner. The O-PTIR spectroscopy measurements were obtained by Dr. Mustafa Kansiz at Photothermal Spectroscopy Corp. While it is not used for this study, it is important to note that the O-PTIR instrument is capable of concurrently recording both IR and Raman spectra at the same spatial resolution.

Data pre-processing

Pre-processing is important to reduce uncontrollable variables affecting spectral measurements. This is of greater importance when analysing biological materials that inherently have variation. Environmental conditions such as temperature, humidity and

instrument drift can all have impacts on spectral quality, reproducibility and repeatability. Pre-processing is a vital step in spectral analysis to remove the unwanted variance. Spectral pre-processing also aids in interpretability by both humans and machines when trying to gain information from the spectra. The performance of classifiers can be largely affected by the treatment and processing of the data. The reduction of unwanted variability allows classification models to focus on relevant information in the spectra.

All spectra were cropped to the area of interest to be analysed. The region of the spectra below 1350 cm^{-1} was removed because the glass substrates used absorb the IR radiation interfering with this region. The areas analysed include 3500-2700 cm^{-1} which contains peaks corresponding to CH_2 symmetric and asymmetric stretching of lipid groups and the amide A peak which comes from the N-H stretching of the amide bonds in proteins. The region between 1800-1350 cm^{-1} is an area of the fingerprint region which includes amide I and II which result from the C=O stretching and C-N stretching vibrations of the peptide bonds in proteins, respectively. The spectra collected from using glass substrates therefore provided information on the protein and lipid content of the analysed cells.

A Savitzky-Golay filter was used to remove noise from the spectra. Savitzky-Golay filters are commonly used for the pre-processing of FTIR spectra because they filter out less high frequency noise than some other smoothing filters. It is important that the Savitzky-Golay filter is not overused, and important features are removed. PCA denoising was another pre-processor used to reduce noise in spectra. PCA denoising functions by performing a matrix decomposition of the dataset.

Normalisation is used in pre-processing to scale spectra within a similar range. It is used to remove variation caused by optical path length differences which can be caused by variation

in sample thickness. Two methods of normalisation were used in this research standard normal variate (SNV) and Extended Multiplicative Signal Correction (EMSC). SNV normalisation begins by mean centring the spectra and then divides the mean centred spectra by the standard deviation over the spectral intensities.

EMSC is a model based pre-processing technique. It is a technique used with vibrational spectroscopy data to correct for additive baseline effects, multiplicative scaling effects, and interference effects (*Afseth et al 2012*). In this work it was used to correct for variation in sample thickness and environmental variables such as water vapour and carbon dioxide. The EMSC model can be described by Equation 1 below.

$$Z_{\mathrm{app}}(\tilde{\nu}) = c + bZ_{\mathrm{ref}}(\tilde{\nu}) + d\tilde{\nu} + e\tilde{\nu}^2 + \varepsilon,$$

Equation 1 EMSC model.

Where Z_{app} is a measured spectrum, Z_{ref} is a reference spectrum, b is a multiplicative parameter, c, d, e are constant, linear and quadratic parameters respectively, ε is a residual term, $\tilde{\nu}$ are spectral wavenumbers (*Tafintseva et al 2019*). The reference spectra used were the average spectra of the training datasets used for each classification. The affects of EMSC on spectra are demonstrated below in figure 9 where the baselines of the spectra are shown to be normalised and corrected to allow for a better comparison of the spectra from different cell types.

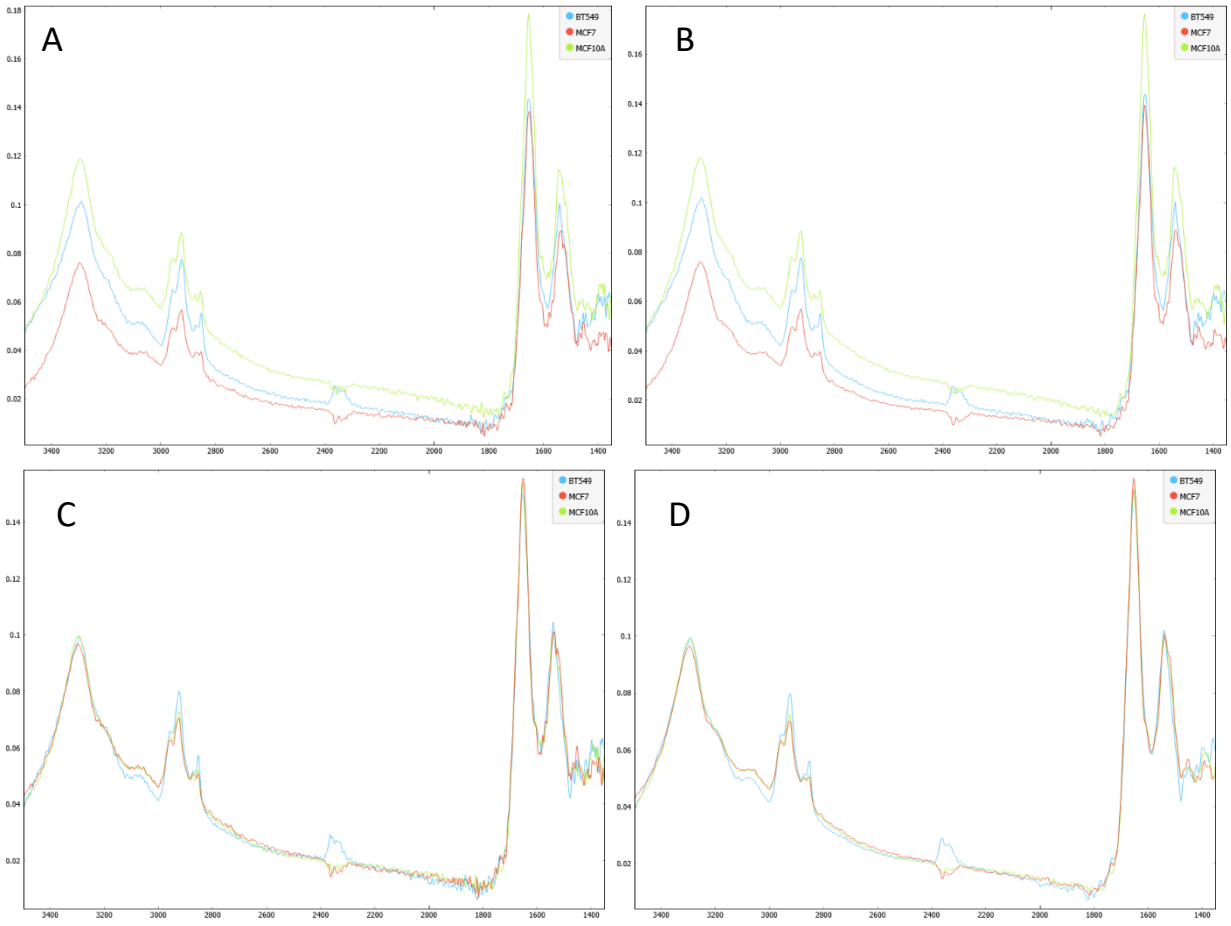

Figure 9 Effect of pre-processing on spectra. A) Raw spectra B) Spectra with denoising applied (PCA denoising and Savitzky-Golay filter) C) Spectra with EMSC applied D) Spectra with de-noising and EMSC applied.

Data analysis

Once the spectral data was pre-processed, it could then be analysed. The first step of analysing the spectral data was visual inspection of spectra picking out the visually identifiable changes in band intensity, shape and position of the average spectra of the cells

being compared. Average spectra were the average of all the individual spectra collected from individual cells. The spectra collected were balanced across the biological and technical replicates.

Principal component analysis (PCA) was used to visualise spectral datasets and aid in identifying where spectral differences are in the spectra through the loading plots of the PCA. The PCA scores were produced using the Quasar software. PCA is an unsupervised learning method used for dimension reduction of data to reduce the dimensions of datasets with a large number of variables. FTIR spectroscopy data is highly dimensional data with each wavenumber representing a different variable. PCA can be used to transform a dataset reducing the number of variables that still retains most of the information of the original data. PCA reduces the dimensions of data by projecting by geometrically projecting into onto a lower dimension space called principal component (PCs). The first PC contains the most variance and is chosen by minimizing the total distance between data and their projection on the PC. The second PC and subsequent PCs are chosen in similarly but with each PC being uncorrelated with the previous PCs. The projection of PC1 is uncorrelated with projection onto PC2 therefore the PCs are geometrically orthogonal. The maximum number of PCs is the smallest of either the number of features in the dataset or the number of samples. PCs are defined as a linear combination of the original variables in the data. The coefficients are stored in a loading matrix which can be interpreted as a rotation matrix that rotates the data so that the projection with the greatest variance is placed on the first axis. For vibrational spectroscopy studies PCA is often used to help visualize the data with each point on the PCA score plot representing a different spectral reading. This allows the PCA score plot to be used to help visualize differences within different categories and groupings can help to identify how closely related different spectra are.

On some datasets the components from the PCA scores were tested for normality by using a Levene's test of normality. For the normally distributed data, the components were compared with Student's t-test to assess if there were significant difference in the data between different cells. For non-normally distributed data the non-parametric equivalent, the Kruskal-Wallis test was used.

For classification tasks, a machine learning based random forest (RF) classifier was used to test if different cells could be classified from each other using the spectral data collected. The Quasar software was used to perform the classification. RF is an ensemble classifier using many decision trees to come to a consensus. The random nature of a RF comes from the feature selection in each individual decision tree being random. Each decision tree starts with a root node which is the selected feature with the highest information gain, branches then split off from this node until a decision is made. A single decision tree is often prone to overfitting but by using many trees with random subsets of features, a RF avoids overfitting. The most important parameters to be considered are the number of decision trees, the depth of each decision tree, and the maximum number of features considered at each split in a tree. The parameters were adjusted for the different datasets and the classifier was applied based upon which parameters provided the best output for the given dataset.

The RF classifier was used to produce hyperspectral maps for the classification of cancer cells within blood. The RF classifier was used to assign a colour to 10 µm tiles using the spectral data. From this, it was compared to the images of the cells to determine if FTIR spectroscopy could be used to identify single cancer cells within blood. The RF classifier was trained using a training set of known spectra of A549 or CALU-1, leukocytes and background measurements. Before classification, each tile of the map was annotated based on the visual

images. This allowed measurement of how well the RF classifier performed at identifying what areas of the map contained leukocytes, cancer cells or background.

The performance of the classification by the RF classifiers was measured using the area under the curve (AUC), classification accuracy, precision, recall and F1. AUC is the area under the receiver operator curve. The higher the AUC, the better a model is at predicting classes. Classification accuracy is the proportion of correctly identified instances. Precision is the proportion of true positives among the instances classified as positive. Recall (sensitivity) is the proportion of true positives amongst all the positive instances in the data. F1 is the weighted harmonic mean of precision and recall. Precision and recall can be described by Equations 2 and 3 respectively.

$$Precision = \frac{True\ Positives}{True\ Positive + False\ Positives}$$

Equation 2 Precision.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Equation 3 Recall.

Staining

To confirm the identity of cancer cells within samples of blood doped with cancer cells, a Giemsa stain was used. Giemsa stain is a Romanowsky type stain commonly used in pathology laboratories for the routine examination of blood films. It is a differential stain containing a mixture of dyes including azure blue, methylene blue, and eosin dye. Azure blue

and eosin are acidic dyes that stain the cytoplasm and granules in cells. Methylene blue is a basic dye that stains acidic components of the cell including the nucleus. The Giemsa stain was produced from a stock Giemsa solution (Atom Scientific) that contained Giemsa powder, glycerol and methanol. The stock stain was diluted to a working stain with a Gurr buffer which is pH 6.8 phosphate buffer. The buffer was produced using buffer tablets (Thermo Scientific). The working stain was made using a dilution of 1:40 Giemsa stock solution:Gurr buffer. 2-3 drops of the Giemsa stain were applied to the sample covering the whole sample area. The sample was incubated for 45 minutes at room temperature. After incubation, the excess stain was poured of the sample slips. Any remaining excess stain was washed off using the Gurr buffer. Once stained the cancer cells in the doped blood samples could be visually confirmed under a microscope due to the difference in size of the cancer cells when compared to the surrounding leukocytes.

Chapter 3- Optimisation of sample preparation on glass substrates for FTIR microspectroscopy characterisation of lung cancer cell lines.

Introduction

The numbers of cancer cases in the UK are continuing to rise with a 12% incidence rate increase for all cancers in the UK from the 1990s to 2017 (Cancer Research UK, 2017). This ever-increasing incidence of cancer generates a greater workload for pathology departments and an increased turn around for key cancer diagnoses. A delay in diagnoses causes a delay in treatment, a worsening in patients' condition and an increase in their stress and anxiety. A four-week delay to treatment could lead to an increased 6-8% chance of a patient dying. A system that could identify abnormal cells in cytology samples for further investigation more efficiently is needed to manage the massive workload of pathology laboratories. Such a system should reduce the time pathologists would spend looking at samples to deem if they are positive or negative for cancer and get results to patients quicker.

Fourier Transform Infrared (FTIR) microspectroscopy has potential as a technique to aid pathologists in their work investigating tissue/cytology samples from patients with cancer or suspected cancer (Finlayson et al., 2019). FTIR microspectroscopy produces provides information on the biochemistry of the cells which can be used to distinguish normal cells from abnormal cells and different cancers. Despite the plethora of work carried out demonstrating its potential, FTIR microspectroscopy has yet to be translated to the clinical setting (Finlayson et al., 2019). One of the major drawbacks has been the substrates that samples are placed on for transmission FTIR microspectroscopy (CaF_2, BaF_2, ZnSe) as they are often expensive, costing up £50-60 per slide. This would make a diagnostic system based on

FTIR microspectroscopy very expensive considering the large number of samples that need to be produced in a clinical setting. The glass slides commonly used in pathology departments as a substrate for cytology samples obscures the fingerprint region (Rutter et al., 2018) (Pilling et al., 2017) (Bassan et al., 2014) of the IR spectra because the glass absorbs IR radiation. The spectrum is obscured below 2000 cm^{-1} removing information on the protein, nucleic acid and carbohydrate content of the cells. This has caused glass not to be considered an appropriate substrate for use with FTIR spectroscopy.

Previous research has shown soda lime glass coverslips of a thickness of 0.12-0.17 mm could be used as a substrate and retain the higher wavenumber bands that correspond to vibrations from fatty acid chains in lipids (3000-2800 cm^{-1}) and the band amide A (3100-3500 cm^{-1}) corresponding to NH vibrations in proteins. These thinner coverslips allow for the study of bands down to 1350 cm^{-1} which cannot be seen on regular glass slides (Rutter et al., 2019). This allows information on both lipids and proteins to be gained while using a more affordable and accessible substrate. However, the amide III (1350-1200 cm^{-1}) and bands corresponding to carbohydrates and nucleic acids below 1350 cm^{-1} in the fingerprint region are lost. It is believed that due to the coverslips being thinner than the glass slides which are about 1 mm thick, less IR radiation is absorbed by the substrate which allows the amide I and II peaks to be viewed. The research in this chapter aims to expand on this work demonstrating a methodology to use glass coverslips for FTIR spectroscopy analysis of lung cancer cells. The research used FTIR microspectroscopy to differentiate between two different lung cancer cells (A549, CALU-1), placed upon the soda lime glass coverslips. This work aims to take another step towards translating FTIR spectroscopy to a system that could be utilised in a clinical setting.

The closer sample preparation is to methods commonly used in clinical settings for current cytological analysis, the easier an FTIR spectroscopy diagnostic method will be to translate to the pathology laboratories. Diverse cytological preparation methods with different fixation methods are used in pathology laboratories. Two sample preparation techniques used in hospitals worldwide to prepare cytology samples are smears and cytospins. Cytology samples for lung cancer produced from bronchio–alveolar lavage or pleural fluid are usually prepared as cytospins (Nalwa et al., 2018). Fine needle aspirations and blood samples are commonly prepared as smears (Kshatriya & Santwani, 2016). Smears preserve the morphological features of cells but can have uneven distribution of cells and inadequate cellularity. Smear quality is dependent on the practitioner. Cytospins provide good cellularity which can increase diagnostic potential and are less dependent on the practitioner to ensure a good quality sample but some of the morphological features are lost. These two methods were chosen to prepare the samples on to the glass coverslip substrate to identify the best preparation for FTIR spectroscopy analysis. Using widely used preparation methods would make it much more likely that the test could fit within current clinical workflows. Additionally, two fixation methods commonly used for cytology were tested. The two fixation methods of 4% PFA and methanol were chosen. Methanol is a quick and simple method of fixation, but alcohols can affect the lipid content of cells (Brown et al., 2012). PFA can take longer to fix a sample but does not affect the lipid content as much as methanol. The basis of the work in this chapter was to find out which preparation method (cytospin or smear) and fixative (4% PFA or methanol) using glass coverslips as substrates is best for analysis of lung cancer cells in terms of usability with FTIR microspectroscopy and spectral quality.

Aims

1. Determine if cytospins or smears are the better method of preparation for FITR microspectroscopy of cytology samples on glass coverslips.
2. Determine between methanol or 4% paraformaldehyde as the best method of fixation for FTIR microspectroscopy of cytology samples on glass coverslips.
3. Investigate if FTIR microspectroscopy can be used to distinguish between lung cancer cells (A549, CALU-1).

Methods

Cell Culture

Two NSCLC cell lines were used for these experiments, A549 (adenocarcinoma) and CALU-1 (SqCC). Refer to the cell culture section of chapter 2 for culture details.

Sample preparation

A549 and CALU-1 cells were brought to a concentration of 1x 10^6/ml. Cytospin or smear methods were used to apply the cells to Galvoptics glass coverslips. To prepare the cytospin, a glass slide was first placed in the cytospin clip, and the coverslip was then placed on top of the slide, the paper filter and funnel were then placed on top of the coverslip. For the cytospin, 20 µl of the cell suspension was added to the bottom of each funnel and spun at 900 rpm for 1 minute. The cells were immediately fixed with 4% PFA or methanol. For the preparations of smears 20 µl of the cell suspension was pipetted on to the coverslip. Another

coverslip was used to spread the droplet across the coverslip. The smear samples were allowed to air dry before fixation to prevent the cells from being washed off during fixation in 4% PFA or methanol.

To fix the samples with 4% PFA, 100 µl of PFA was applied to the sample and incubated for 15 minutes. After incubation, the samples were washed with 200 µl of 0.9% saline once and 200 µl of water thrice. After the washings, the samples were air dried. To fix with methanol, the samples were submerged in methanol for 2 minutes. After 2 minutes, the samples were removed from the methanol and air dried to allow excess methanol to evaporate off.

All cells used to produce samples were harvested from flasks after two weeks of growth post thawing. Cells were harvested during growth phase at around 70% confluency in the flask. Samples were prepared in at triplicate in three independent experiments.

FTIR microspectroscopy

100 spectra of each cell line for each method of preparation and fixation were collected. Each spectrum was measured from a different individual cell with the measurement centred on the middle of the cell. A Nicolet iN10 benchtop spectrometer was used to collect the spectra. The spectrometer used a globar IR source and a MCT detector cooled with liquid nitrogen. Spectra were collected at 4 cm^{-1} resolution with 256 co-added scans per cell using an aperture size of 15 x 15 µm centred on the centre of the cell. Measurements were taken equally from across samples from the three independent experiments.

Pre-processing and data analysis

The spectra were cropped to the areas of interest, 3100-2700 cm^{-1} and 1800-1350 cm^{-1}. 3100-2700 cm^{-1} contains bands mainly from the vibrations of C-H groups in the hydrocarbon chains of lipids. 1800-1350 cm^{-1} contains the amide I band at 1695-1615 cm^{-1} resulting from the stretching of C=O in the amide bonds of proteins and the amide II band at 1550-1520 cm^{-1} resulting from the bending of N-H bonds in the amide bond. A Savizky-Golay filter with a window size of 15 and polynomial of 2 was used to de-noise the spectra. SNV was used to normalise the spectra. Normalisation is applied to remove baseline defects than can be caused by variations in sample thickness. The average spectra for each preparation condition were generated by averaging the 100 spectra collected of each cell line for each condition.

Data analysis was carried out in the Unscrambler X software. PCA plots of the spectra were generated to allow for comparison of the spectra of the two cell lines and preparation conditions. A Levene's test of normality was used to test if the data was normally distributed. A Kruskall-Wallis test was used for the non-normally distributed data and a Students' t-test for the normally distributed data. Both were performed on the PC data to test if the two cell lines and the preparation conditions were significantly different.

Results

The cytospin produced samples with a uniform distribution of cells with the cells grouped in one circular area on the coverslip. The smear produced a non-uniform distribution of cells

across a large area of the coverslip. The cells on a smear were spread across the coverslip in small groups. These patterns of distribution made collecting 100 spectra from different cells take much longer for the smear samples than the cytospin samples. Furthermore, smears must be dried prior to fixing to minimise loss of cells from the coverslip during fixation and washing. The cytospins can be fixed as soon as they are prepared because the cells are strongly attached to the coverslip. This drying step in the preparation of smears can cause biochemical changes in the cells which will affect the FITR spectra of cells.

Methanol was tested as a fixative because it only needs 2 minutes for fixation and no washing steps because the methanol evaporates due to its volatility. Compared to PFA which can take up to 20 minutes with incubation and washing. The time saved using methanol fixation is advantageous if many samples must be produced. The methanol fixation however stripped lipids from the cells shown in the spectra of figures 10 and 12 by a decrease in the size of the peaks and a flattening in their shape at 2850 cm^{-1} and 2920 cm^{-1}.

Figures 11 and 13 show the amide I and II bands in the region 1800-1350 cm^{-1} for A549 and CALU-1 cell lines prepared using smear and cytospin respectively. The spectra demonstrate that amide I and II can be identified in both cell lines on the glass coverslips when fixed with PFA or methanol. Both fixative agents retain these bands to provide information on the protein content of cells. However, a difference in the shape of the bands can be seen in the cytospins and smears. This difference is most noticeable in the CALU-1 spectra where the amide I and amide II bands have a flatter peak than the in cells prepared by smear compared to the cytospin preparation. This infers a change in the protein content caused by the preparation method.

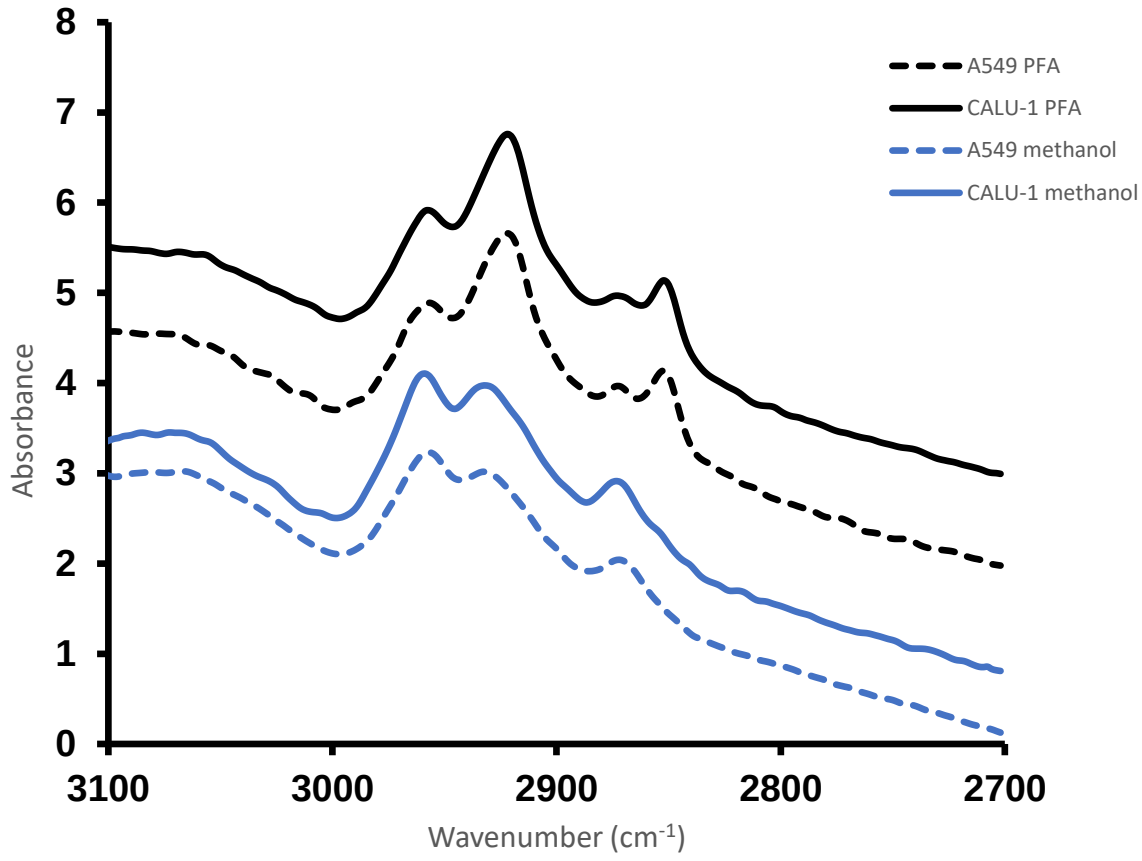

Figure 10 Average spectra from 100 cells of A549 and CALU-1 cells for the region 3100-2700 cm^{-1} prepared on glass coverslips as a smear. Cells were fixed with 4% PFA or methanol. Spectra offset for clarity.

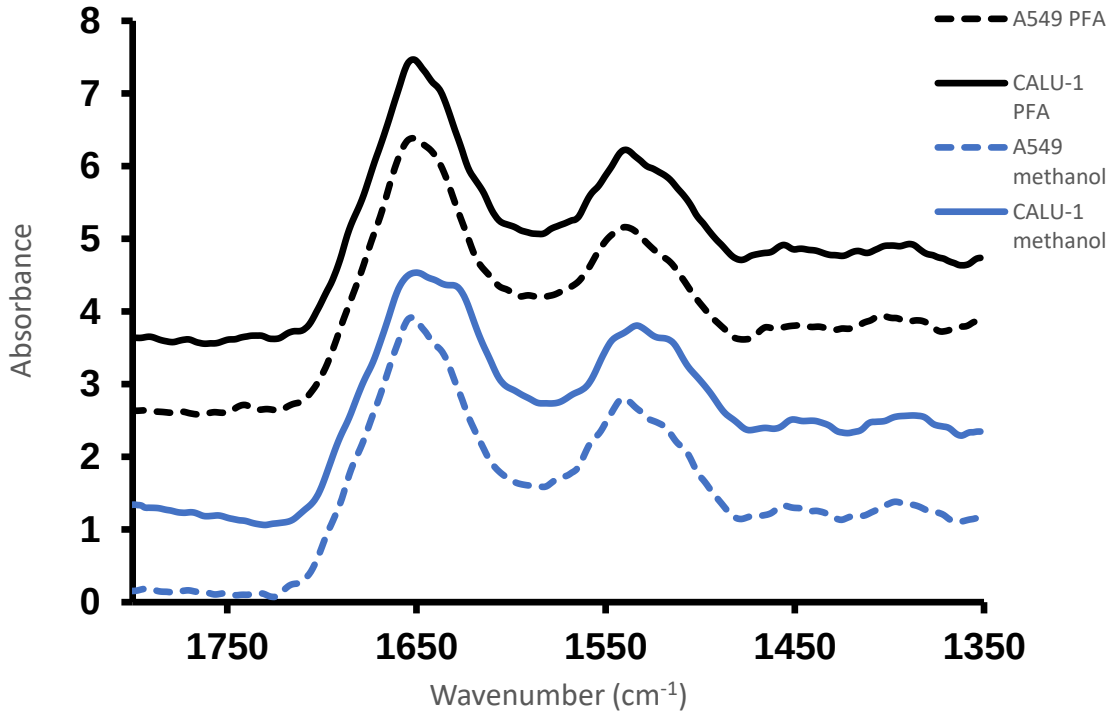

Figure 11 Average spectra from 100 cells of A549 and CALU-1 cells for the region 1800-1350 cm^{-1} prepared on glass coverslips as a smear. Cells were fixed with 4% PFA or methanol. Spectra offset for clarity

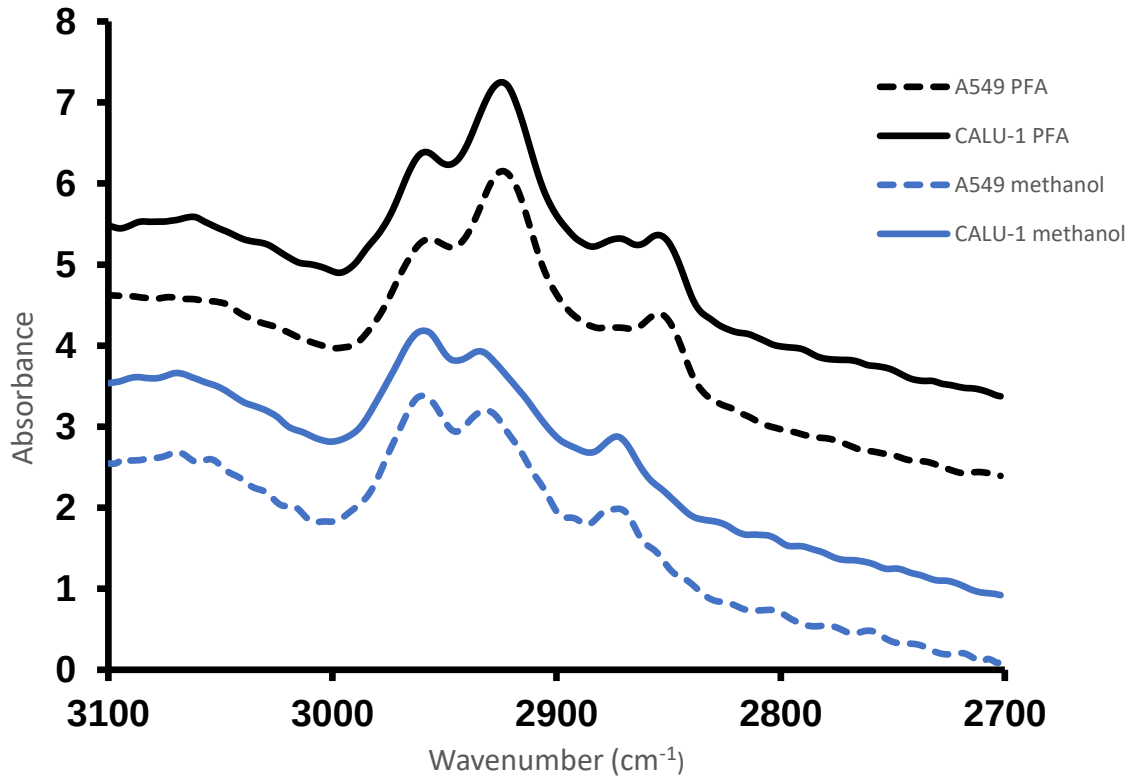

Figure 12 Average spectra from 100 cells of A549 and CALU-1 cells for the region 3100-2700 cm^{-1} prepared on glass coverslips as a cytospin. Cells were fixed with 4% PFA or methanol. Spectra offset for clarity

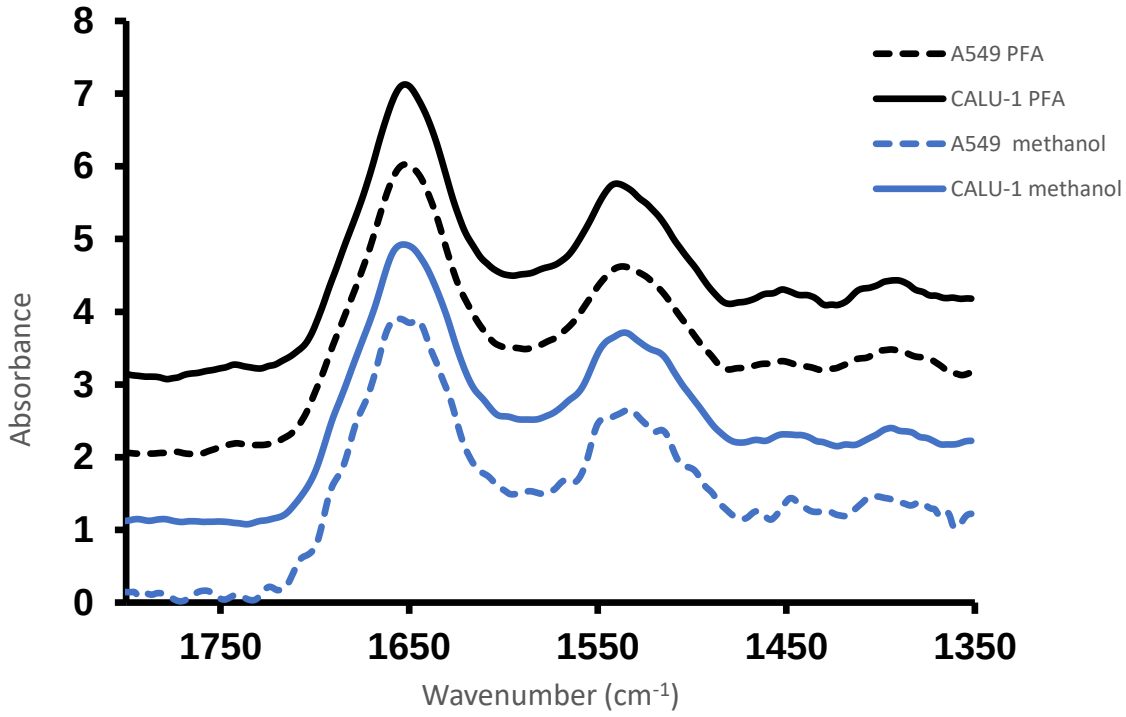

Figure 13 Average spectra from 100 cells of A549 and CALU-1 cells for the region 1800-1350 cm^{-1} prepared on glass coverslips as a cytospin. Cells were fixed with 4% PFA or methanol. Spectra offset for clarity.

The next step was to analyse the data with PCA to assess the differences between spectra of different preparation conditions and between A549 and CALU-1. Figures 14 and 16 below shows the PCA scores and loadings for the 3100-2700 cm^{-1} region of A549 and CALU-1 prepared by cytospin and fixed with PFA and methanol on glass coverslips. The PCA plot shows that there is a clear grouping and separation of cells prepared with the different fixatives thus a difference in the spectra of cells fixed with methanol or PFA. As expected from the mean spectra there is a clear difference from the fixative of choice on the lipid region (Figures 12 and 14) of the cells. Despite the loss of lipid content with methanol there was separation of A549 and CALU-1 on the PCA scores for smear and cytospin preparation

for both fixation methods. It is not identifiable if the separation in the lipid region with methanol fixation is from actual difference in the lipid content of the cells or differences in what lipids the methanol removed from the cells. Separation between the cells from the fixative used can also be seen for both preparation methods in the region 1800-1350 cm^{-1} (Figures 15 and 17). Separation of A549 and CALU-1 was clearer in the cytospin PCA score of 1800-1350 cm^{-1} than for the smear. Statistical analysis of a PCA comparing cytospin versus smear sample preparation showed statistically significant differences between these two types of samples regardless of type of fixation or cell line for PC1 for both regions (Table 2).

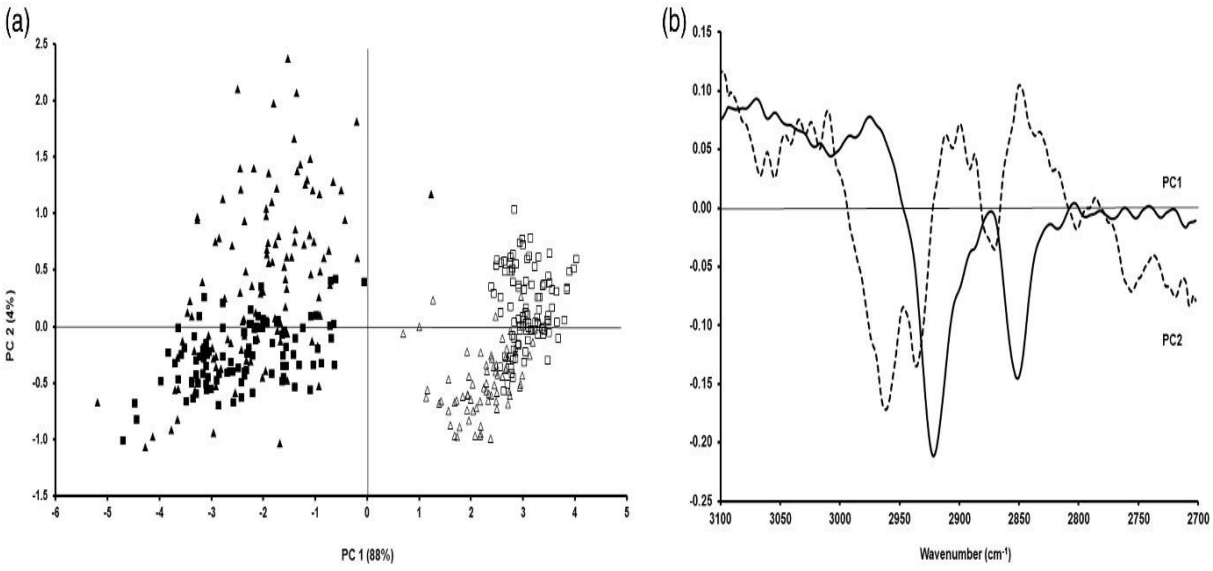

Figure 14 PCA score (a) for A549 (triangles) and CALU-1 (squares) cells prepared using cytospin and fixing them with methanol (open triangles and open squares) or PFA (filled triangles and filled squares) for the 3100-2700 cm^{-1} region and the corresponding PC loadings (b).

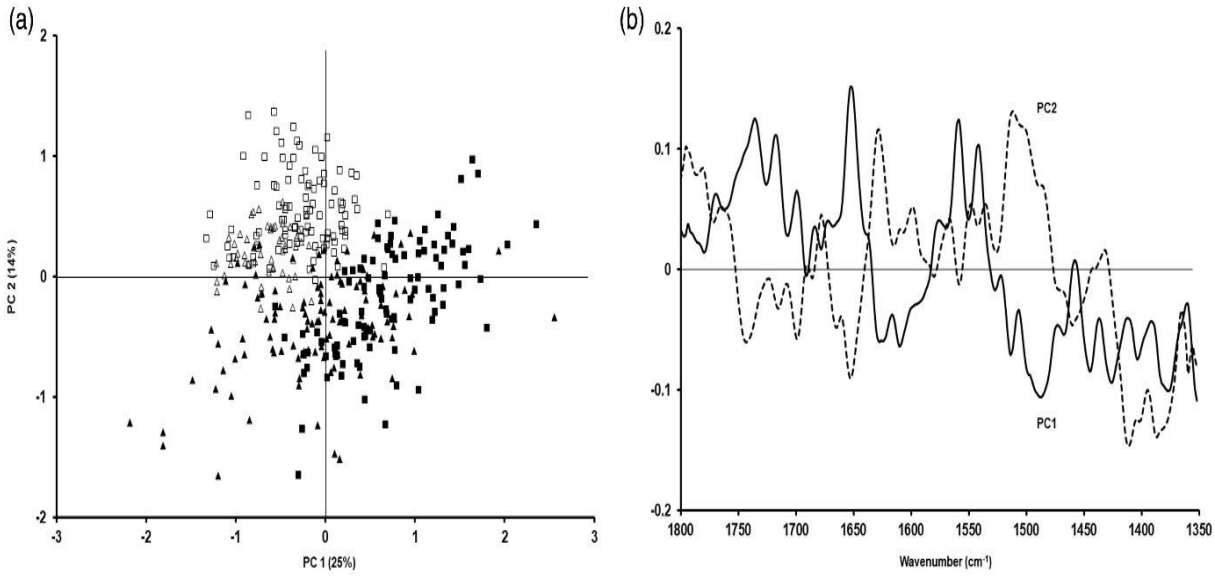

Figure 15 PCA score (a) for A549 (triangles) and CALU-1 (squares) cells prepared using cytospin and fixing them with methanol (open triangles and open squares) or PFA (filled triangles and filled squares) for the 1800-1350 cm^{-1} region and the corresponding PC loadings (b).

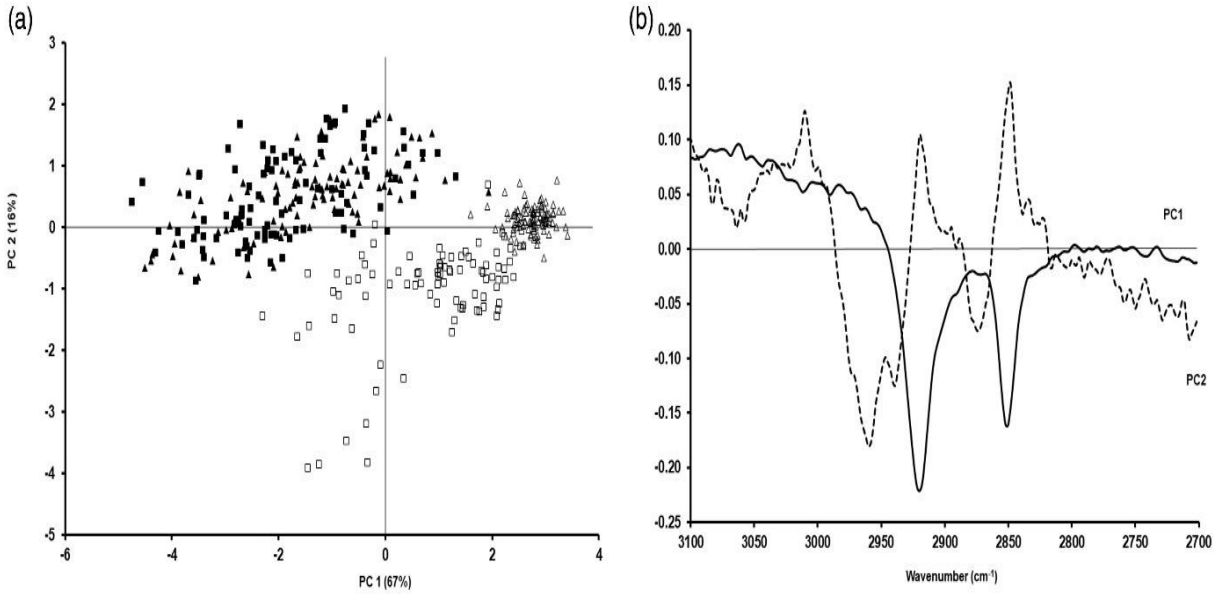

Figure 16 PCA score (a) for A549 (triangles) and CALU-1 (squares) cells prepared using smear and fixing them with methanol (open triangles and open squares) or PFA (filled triangles and filled squares) for the 3100-2700 cm^{-1} region and the corresponding PC loadings (b).

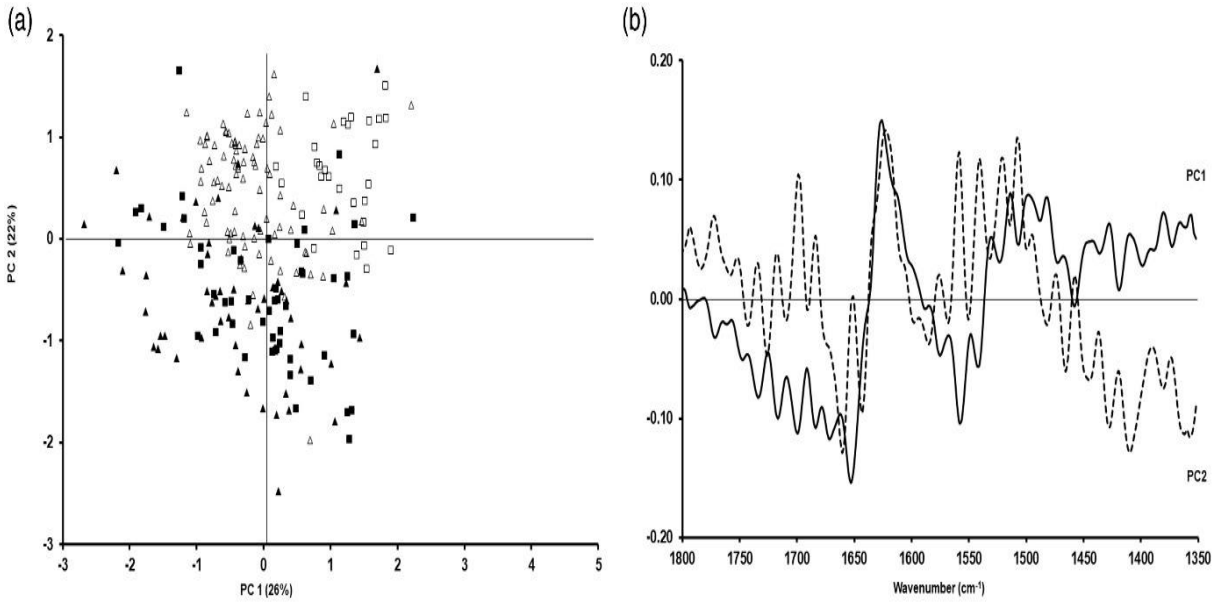

Figure 17 PCA score (a) for A549 (triangles) and CALU-1 (squares) cells prepared using cytospin and fixing them with methanol (open triangles and open squares) or PFA (filled triangles and filled squares) for the 1800-1350 cm^{-1} region and the corresponding PC loadings (b).

	Lipid region		1800 cm^{-1} to 1350 cm^{-1} region	
	PC1	**PC2**	**PC1**	**PC2**
Formalin				
A549	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**
CALU-1	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	$p = 0.395$
Methanol				
A549	***p* < 0.001**	***p* = 0.021**	***p* < 0.001**	$p = 0.071$
CALU-1	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**

Table 2 Statistical significance between different types of sample preparation (cytospin versus smear) based on types of fixative (PFA, methanol) and cell type (A549, CALU-1). Statistically significant values in bold.

The statistical analysis of the PCAs in Figures 14 to 17 is shown in Table 3 comparing fixation with PFA or methanol. Statistically significant differences can be seen between these two types of fixative regardless of cell line and whether samples have been prepared as cytospins or smears. In order to assess whether the differences were not just due to sample preparation but also to biochemical differences between these two cell lines, statistical analysis from the principal components comparing A549 cell line (lung adenocarcinoma) and CALU-1 (lung epidermoid carcinoma) was carried out. Again, the statistical analysis in Table 4 is based on the PCAs as shown in Figures 14 to 17. Statistically significant differences between these two types of cell lines in both regions for PC1. These differences are present regardless of the preparation or fixation method. For PC2, the differences were statistically significant for samples prepared as cytospins and fixed either with PFA or methanol (Table

3). This is the first time that different types of lung cancer cell types have been separated with FTIR microspectroscopy using glass coverslips as substrates.

	Lipid region		**1800 cm^{-1} to 1350 cm^{-1} region**	
	PC1	**PC2**	**PC1**	**PC2**
Cytospin				
A549	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**
CALU-1	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**
Smear				
A549	***p* < 0.001**	***p* < 0.001**	*p* = 0.125	***p* < 0.001**
CALU-1	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	**p < 0.001**

Table 3 Statistical significance between different types of fixative (PFA versus methanol) based on sample preparation (cytospin, smear) and cell type (A549, CALU-1). Statistically significant values in bold.

	Lipid region		**1800 cm^{-1} to 1350 cm^{-1} region**	
	PC1	**PC2**	**PC1**	**PC2**
Cytospin				
Formalin	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	***p* = 0.001**
Methanol	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**
Smear				
Formalin	***p* = 0.038**	*p* = 0.162	***p* = 0.008**	*p* = 0.298
Methanol	***p* < 0.001**	***p* < 0.001**	***p* < 0.001**	*p* = 0.099

Table 4 Statistical significance between the different cell types (A549 versus CALU-1) based on sample preparation (cytospin, smear) and fixative (PFA, methanol). Statistically significant values in bold.

Discussion

For new methods and techniques to be introduced into clinical laboratories, it is important that these methods do not disrupt current workflows and standard procedures used. For FTIR spectroscopy to be successfully used in the clinical laboratory, sample preparation is the crucial first step. Sample preparation should be standard across laboratories and fit with current workflows. Ideally, preparation methods for FTIR spectroscopy should use techniques already established in pathology laboratories. The sample preparation must reach a balance though of not altering the biochemistry of cell and tissue samples while fitting in with clinical practices.

Two preparation methods commonly used in histopathology laboratories are smears and cytospins (Strimpakos et al., 2014). The smear requires no special equipment to prepare therefore it can be performed at anytime and anywhere because it does not require an instrument like the cytospin. This could be an important factor to consider for countries with less developed health infrastructure where smaller hospitals may not be able equipped with the cytospin machine or only have a small number. However, the fact that samples prepared with a smear require time to dry before applying the fixative to avoid washing off the cells from the substrate may alter the biochemistry of the cell and adds time to the preparation. Air drying could cause delocalisation of biomolecules as a result of large surface tension forces associated with the water air interface (Baker et al., 2009). The cytospin method of preparation is more consistent in quality as it uses a cytospin machine to apply the cells to the substrate whereas the smear's quality is dependent on the technician. This makes the

cytospin more reliable and repeatable across samples. The cytospin method maintains a more consistent quality in the biochemistry and spectral quality as well as the cellularity of the samples. The samples prepared using the cytospin are quicker to analyse with FTIR microspectroscopy as the cells are concentrated into a single area on the substrate. The cells on a smear are spread across the substrate so time is taken to find the small, spread groups of cells. I found this to be particularly time consuming when analysing 100 cells for each sample. The smears took significantly longer to collect 100 spectra. The cytospin preparation would allow for multiple cells to be analysed quickly and easily saving time and allowing for more confidence in the result because more cells can be measured.

Fixation of cells is a vital step of preparation of cell samples for FTIR spectroscopy analysis. Unfixed cells have a high-water content, the water absorbs IR radiation and obscures information of biomolecules in the cells making analysis difficult. Removal of cells from growth media and air drying can change the osmotic pressure within the cells resulting in shrinking or swelling of cells. Swelling can result in membrane rupture and the leaching of intracellular components. Also, the drying of living cells can initiate autolytic processes. Lysosomes release enzymes that cause denaturing of proteins and dephosphorylation of mononucleotides, phospholipids and proteins. The effects of autolysis from inappropriate preparation of cellular samples will obscure the information that can be gained from cells with FTIR spectroscopy analysis. Fixation quenches the autolysis process and minimises the leaching of biomolecules. Therefore, finding a fixation process that retains the biochemical information of the cells while not disrupting workflows is of great importance for FTIR spectroscopy to be translated to clinical use.

4% PFA and methanol were tested because they are two fixative agents commonly used in pathology laboratories. Fixation of samples with methanol is simple and quick requiring the placement of the sample in methanol for two minutes. PFA fixation is much more time-consuming requiring 15-20 minutes for fixation then washes with saline and water. However, as methanol is an alcohol it removes some of the lipids from the cells as demonstrated in Figures 10 and 12 which demonstrated decreased intensities and removal of band features in the lipid regions of both A549 and CALU-1 cells in both the cytospin and smear preparation methods. Whereas the lipid band features were still present after fixation with PFA. Fixation with PFA was decided to be the preferred method of fixation of the two methods tested as it retains more information from the lipid content of the cells. Both the methanol and the PFA maintained the band features of the amide I and II. As such if the interest of a study or clinical evaluation is in the amide bands methanol fixation has the advantage of being faster than PFA. Otherwise, PFA would be a more suitable fixative for FTIR spectroscopy studies as it retains more of the cell's biochemical features. This is especially important if using a glass substrate where information of nucleic acids and carbohydrates is already lost. Research by Meade et al compared live cells to three fixative methods for Raman spectroscopy (Meade et al., 2010), a technique like FITR spectroscopy that measures the biochemical content. The study looked at PFA and methanol as fixatives as well as Carnoy's fixative (60% absolute ethanol, 30% chloroform, 10% glacial acetic acid). They found that all three fixative agents affected the vibrational modes of lipid, protein, nucleic acid and carbohydrate moieties compared to live cells. But found that PFA produced the spectrum most like that of live cells, agreeing with my work that PFA would be the more appropriate fixative for vibrational spectroscopy of cells. They also found that methanol and Carnoy's fixative also altered the nucleic acid spectral contributions significantly even though

they were fixatives recommended for nucleic acid study. This is further evidence for caution to be taken if methanol is to be used as a fixative in vibrational spectroscopic analysis.

This research utilised lung cancer cell lines to demonstrate the feasibility of the sample preparation methods to collect FTIR spectra of NSCLC samples prepared on a glass coverslip substrate. The sample preparation methodology proposed in this chapter is to create a method to allow measurement of cytology samples with FTIR microspectroscopy that will cause minimal disruption to clinical workflows while also being accessible. For the translation of this methodology cytological samples would have to be obtained from the lungs. Currently there are a few techniques of lung cancer cytology, but they can often produce a false negative diagnosis due to low cell number and difficulty distinguishing the morphology. This is where FTIR microspectroscopy could be used to help identify lung cancer cells when the morphology is not sufficient as it provides biochemical information while allowing staining for morphological analysis. While cytological techniques are currently less accurate for lung cancer diagnosis than tissue biopsies, they are however less invasive and carry less risk to a patient. Techniques used for cytological evaluation of lung cancer include induced sputum, thoracentesis, bronchioalveolar lavage, bronchial brushing, bronchial washing and fine needle aspiration. Wang et al previously demonstrated the use of FTIR spectroscopy for the analysis of lung cancer cytology from pleural fluid. That study demonstrated that there could be a difference found in the fingerprint region of the spectra from normal lung cells and the spectra from lung cancer cells. Despite this research showing that FTIR spectroscopy could be a useful diagnostic tool for lung cancer cytology over 25 years ago and a growing body of research, there has been little translation of FTIR spectroscopy from research settings to clinical diagnostics. One of the reasons for this is that much of the research uses methods that could be disruptive to current diagnostic workflows.

The methodology proposed in this chapter aims to address this problem by utilising sample preparation methods and materials that are used across pathology labs already to prepare samples for diagnosis.

Conclusions

The preparation of cytology samples on glass coverslips using a cytospin and PFA fixation is what I would advocate for FTIR spectroscopy analysis in a clinical laboratory setting. Ideally this sample preparation could form part of a methodology for the use of FTIR spectroscopy for cytological cancer diagnosis. Such a system would first identify abnormal samples, separating them from samples deemed non-pathological based on the biochemical properties of the cells. The preparation method proposed would allow for many cells to be measured easily and quickly while maintaining the integrity of the protein and lipid cellular content. The glass substrates would make use of FTIR spectroscopy on large scales for diagnostics affordable and accessible. Additionally, the use of glass with the proposed sample preparation would allow further testing with the standard cytological staining and immunohistochemistry tests. The FTIR spectroscopy experiments carried out in the rest of this thesis use the cytospin and PFA preparation proposed in this chapter.

Chapter 4: FTIR Spectroscopy Combined with Machine Learning classification of lung cancer cells from non-malignant lung cells on a glass substrate.

Introduction

For FTIR spectroscopy to become commonplace in clinical diagnosis of cancer it must be affordable and minimally disruptive to current practice. The previous chapter outlined a sample preparation method to use FTIR spectroscopy with glass coverslips for the analysis of lung cancer cells. As discussed in the previous chapter, the cytospin was selected because it produced reproducible samples with good cellularity that allowed for quick analysis of many cells while maintaining good spectral quality. Fixation with 4% PFA maintained the spectral information because it retained protein and lipid content, as such, was used to prepare samples for this research. The ideal use of FTIR spectroscopy in a clinical setting would be in an automated system which would allow for efficient objective diagnosis with need for little hands-on time for the pathologists. The system would need to be able to separate lung cancer from non-cancerous samples. The time saved by triaging the samples would help to reduce the large caseload pathologists have to lead to more timely diagnosis and treatment. This would require the use of machine learning classifiers to classify cancerous samples from non-cancerous samples using the information on the biochemical content of the cells contained in the spectra.

As shown in chapter 3 using a glass coverslip substrate, FTIR spectroscopy can be used to distinguish two different types of NSCLC cells. The next step was to investigate if lung cancer

cells can be classified from non-malignant lung cells with FTIR microspectroscopy. To see how well the cancer cells could be classified from non-malignant cells, spectra collected from two lung cancer lines (A549, CALU-1) and a non-malignant lung line (NL20) were fed into a random forest (RF) classification model, a form of machine learning. Machine learning is a subset of artificial intelligence (AI) that uses algorithms to build a model based upon sample data commonly referred to as training data (Kotsiantis *et al.*, 2014) . These models can be used to make predictions or decisions without a programme explicitly instructing the decision. Machine learning in recent years has been utilised with spectroscopy data including FTIR spectroscopy to classify and categorise samples based upon trained classification models. Machine learning models are valuable tools that could potentially be used within an automated FTIR spectroscopy system for the diagnosis of lung cancer. Once trained a model could decide on an unknown sample and help to classify it as a normal sample or an abnormal cancerous sample.

If reliable classification of cancerous lung cells from lung cells from a non-cancerous tissue can be achieved using spectral data collected from samples prepared on glass substrates, it will be a step closer to demonstrating how such a methodology could be utilised to examine cytology samples.

Aims

1. To investigate if using FTIR spectroscopy lung cancer cells can be classified from non-malignant lung cells using a glass coverslip substrate with the proposed sample preparation method.

2. To investigate if using FTIR spectroscopy two different NSCLC cells can be classified from each other with a RF classifier using a glass coverslip substrate with the proposed sample preparation method.

Methods

Cell culture

Three cell line were used for the experiments in this chapter, NL20, A549 and CALU-1. NL20 is a non-malignant lung cell line derived from non-cancerous tissue while A549 and CALU-1 are NSCLC cell lines. Refer to the cell culture section in the chapter 2 for a detailed methodology of the culture conditions.

Sample preparation

Cells were collected from the flasks by trypsinisation and centrifugation. Cells were resuspended in 0.9% normal saline and brought to a concentration of 1 million cells per 1 ml. Cells were applied to the glass coverslips by cytospin ran at 900 rpm for 1 minute using 20 µl of the cell solution. The samples were immediately fixed using 4% paraformaldehyde (PFA) and incubated for 15 minutes. After fixation excess PFA was washed off with one wash of 0.9% normal saline and three washes with deionised water. Samples of the three cell lines were made in three different experiments and four samples coverslips of each cell line was produced per experiment.

FTIR spectroscopy

The samples were measured using transmission FTIR spectroscopy on a Bruker Vertex spectrometer with a synchrotron light source attached to a Hyperion 3000 microscope. Measurements were taken from the centre of each cell. An aperture size of 15x15 µm was used and 256 co-added scans of each cell were taken. The background was measured on a clear section of the glass slip without any cells. A background measurement was taken before each cell measurement. 150 cells of each cell line were measured with 50 cells from each of the experiments measured from cells across the four samples from each experiment. Each spectrum recorded was from a different individual cell.

Pre-processing and data analysis

The first step of pre-processing the spectra was to remove the portion of spectra obscured by the glass <1350 cm^{-1}. Spectra were cropped to the regions 3500-1350 cm^{-1}, 3500-2700 cm^{-1} and 1800-1350 cm^{-1}. Noise was removed from the spectra with PCA denoising set to 10 principal components and a Savitzky-Golay filter with a window size of 5 and a polynomial of 2. Extended multiplicative signal correction (EMSC) was applied to the spectra for baseline correction and normalisation. EMSC also aids in the removal of scattering effects from the spectra, inference from variation in sample thickness, water vapour and carbon dioxide. The 150 spectra recorded of each cell line were averaged to produce the average spectra. The second derivative spectra were generated by applying a second derivative to the spectra in

the Savitzky-Golay processor. The PCA denoising was increased to 12 principal components and window size of the Savitzky-Golay filter increased to 17 to reduce the noise introduced from the second derivative. All the pre-processing steps were performed using the software, Quasar.

Classification was performed using a random forest classifier. The spectra were split randomly 70:30 into a training set and testing set. The RF classifier contained 200 decision trees, the square root of the number of attributes was set for the number of attributes split at each node and no pruning was applied to the tree size. The results of the classification were assessed by the classification accuracy, F1, precision and recall and through the confusion matrices. The data analysis was performed using Quasar software.

Results

As demonstrated in the previous chapters and previous studies the spectra produced from using a glass coverslip substrate produces useable spectral data up to 1350 cm^{-1} which includes peaks providing information on the lipids and proteins of the measured cells. Figure 18 shows the average spectra of A549, CALU-1 and NL20 (3500-1350 cm^{-1}). From inspection of the average spectra there were differences in the absorbance, peak shape and position. Both A549 and CALU-1 spectra had a higher absorbance in the amide I and amide II peaks than the NL20 (Figure 20). The position of the amide I peak of A549 and CALU-1 were shifted to 1553 cm^{-1} whereas the amide I peak of NL20 was at 1551 cm^{-1}. The A549 and CALU-1 spectra also had a higher absorbance in the peaks at 2922 cm^{-1} and 2852 cm^{-1} these peaks are from the CH_2 symmetrical and asymmetrical stretching modes mostly from CH_2 groups in the lipid fatty acid chains (Figure 19). The 2922 cm^{-1} peak was shifted compared to the peak

for NL20 where it was positioned at 2925 cm^{-1}. The differences in these peaks showed that there are biochemical differences in the lung cancer derived cells (A549 and CALU-1) and NL20 derived from normal lung tissue.

The second derivative spectra in Figures 21-23 resolved further differences between the cell lines. NL20 had a lower absorbance at the peak positions of 2865 cm^{-1}, 2903 cm^{-1} and 1674 cm^{-1} and a higher absorbance at 2987 cm^{-1} and 2889 cm^{-1} than A549 and CALU-1. These differences in intensity infer biochemical differences in the lipids and proteins of the cancer cells from the NL20 cells.

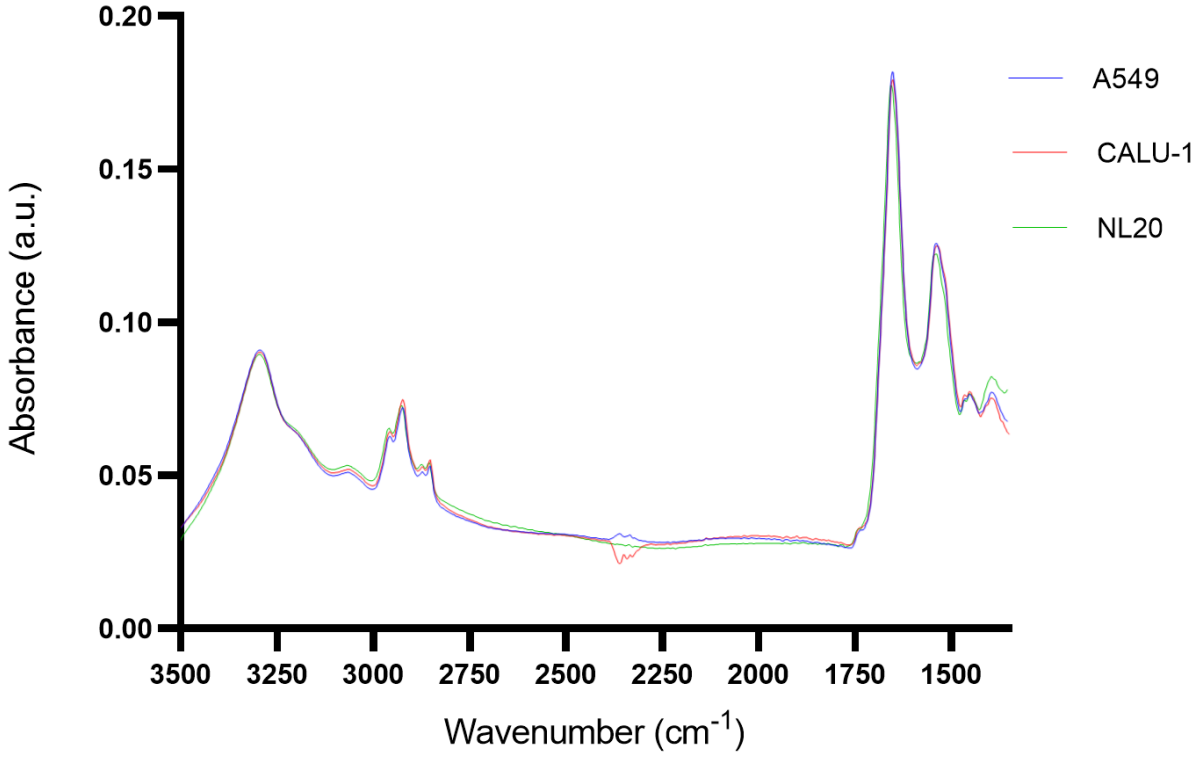

Figure 18 Average spectra from 150 cells of each cell line A549, CALU-1 and NL20 in the region 3500-1350 cm^{-1}. Each of the spectra contributing to the average spectra was from a different individual cell.

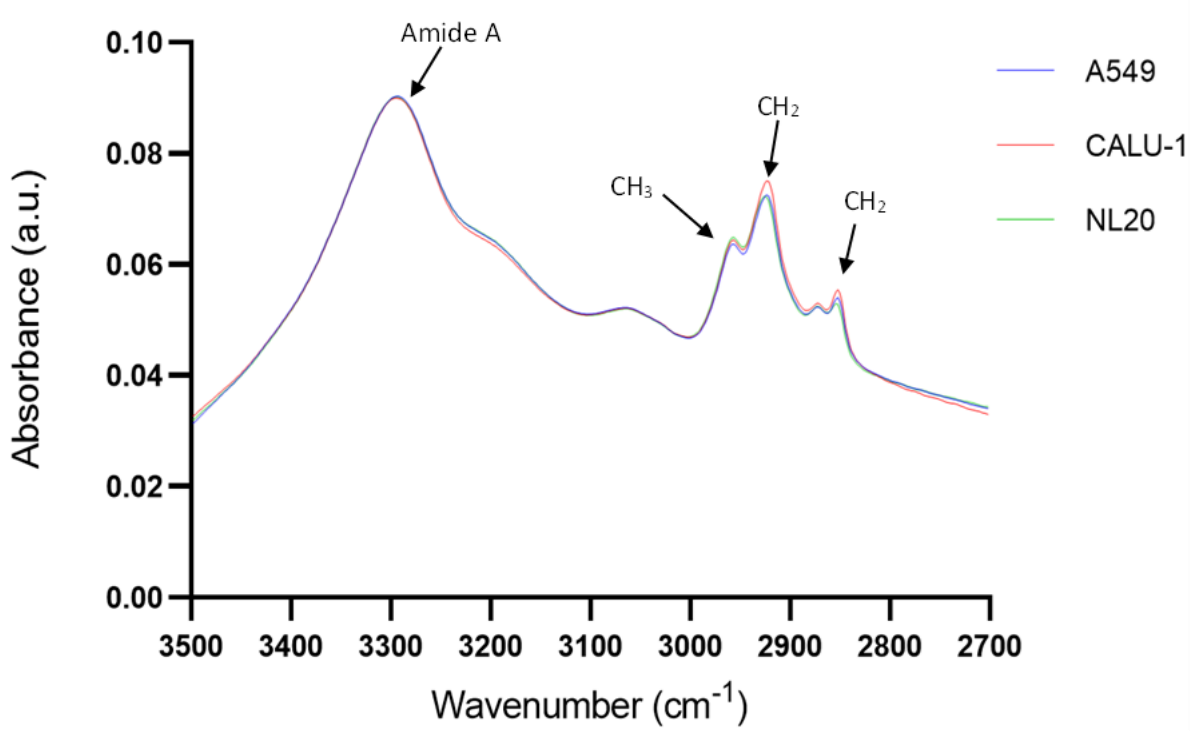

Figure 19 Average spectra from 150 cells of each cell line A549, CALU-1 and NL20 in the region 3500-2700 cm^{-1}. Each of the spectra contributing to the average spectra was from a different individual cell. This region of the spectra contains the amide A band, CH_3 symmetrical stretching and CH_2 symmetrical and asymmetrical stretching bands.

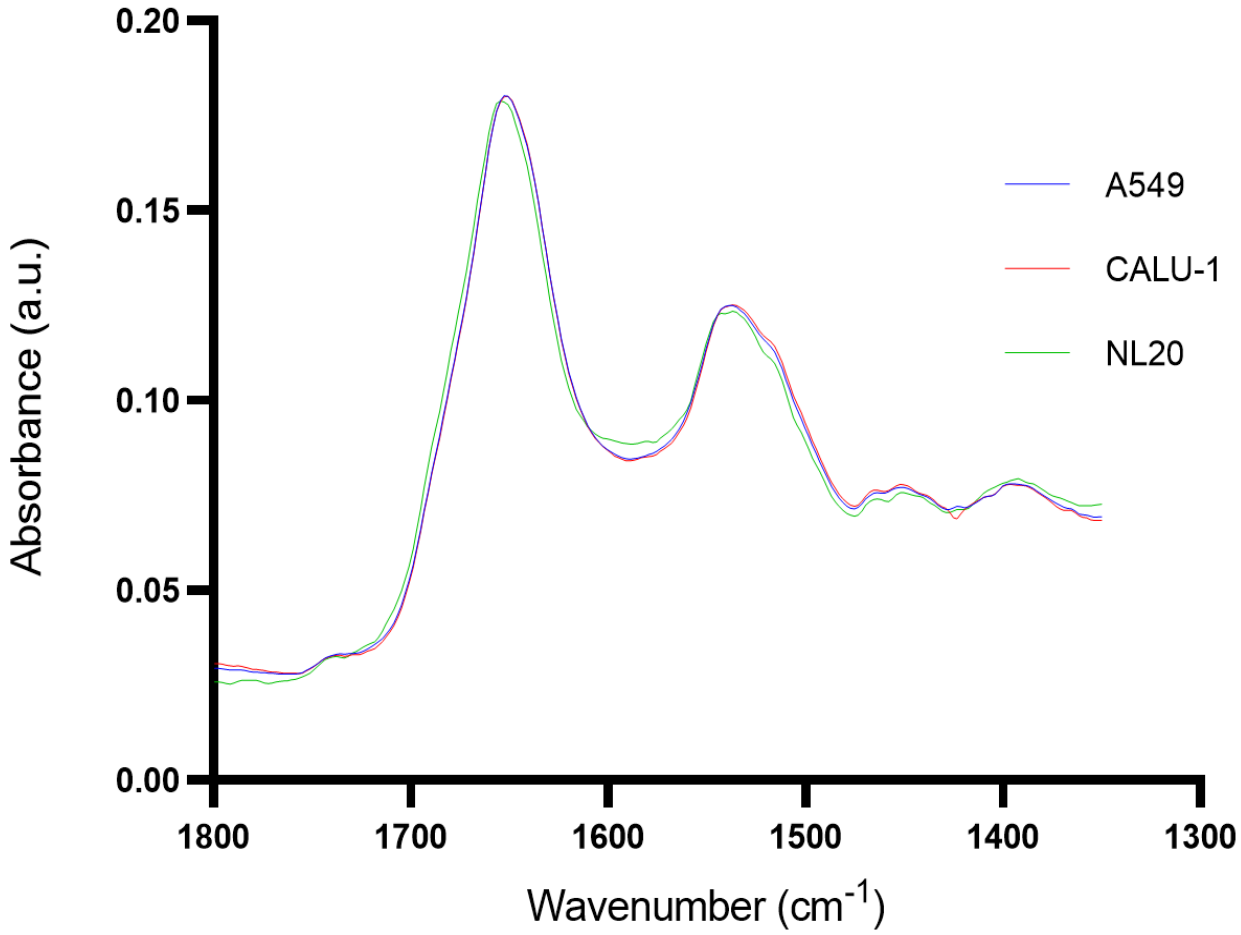

Figure 20 Average spectra from 150 cells of A549, CALU-1 and NL20 in the region 1800-1350 cm^{-1}.

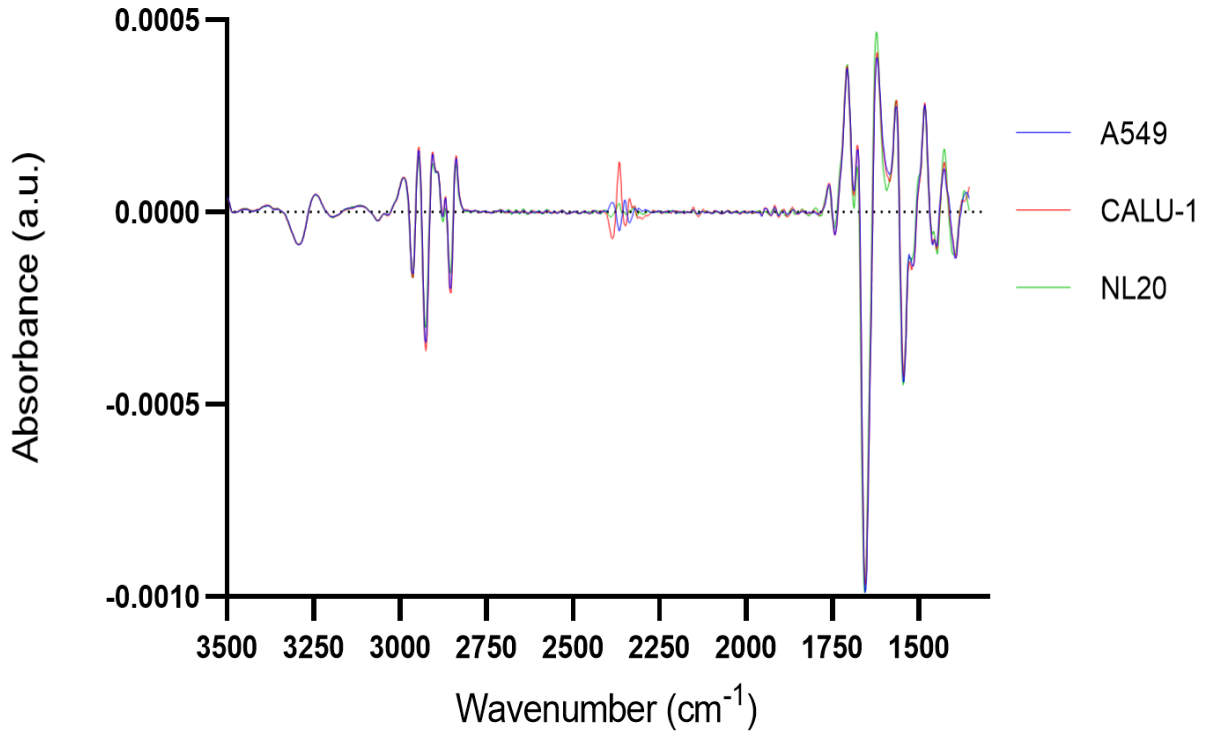

Figure 21 Average 2nd derivative from 150 spectra of A549, CALU-1 and NL20 in the region 3500-1350 cm^{-1}.

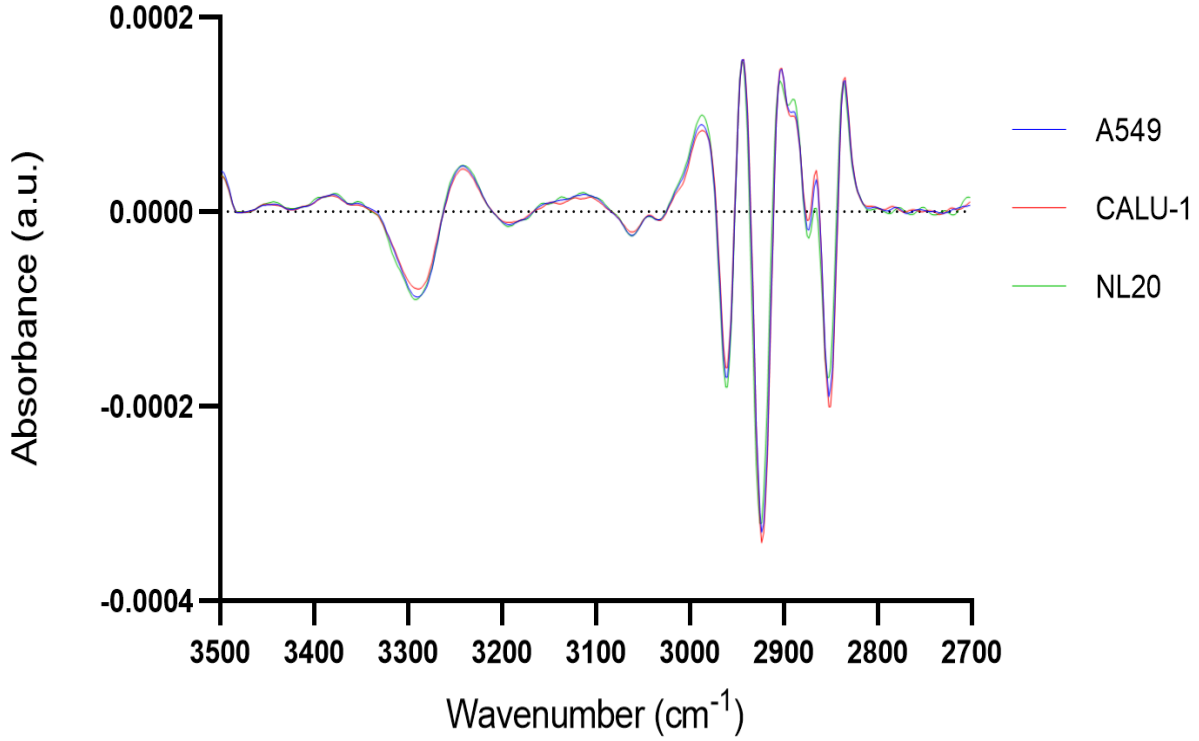

Figure 22 Average 2nd derivative spectra from 150 spectra of A549, CALU-1 and NL20 in the region 3500-2700 cm^{-1}.

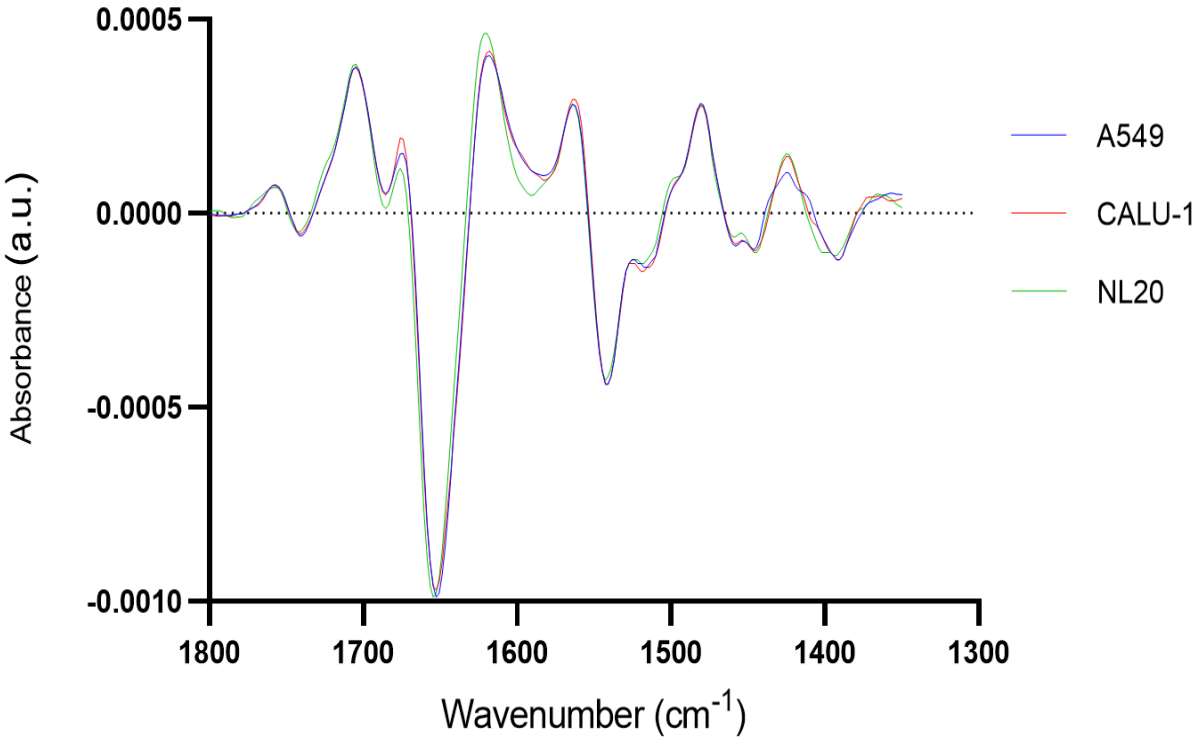

Figure 23 Average 2nd derivative spectra from 150 spectra of A549, CALU-1 and NL20 in the region 1800-1350 cm^{-1}.

To assess if FTIR spectroscopy using a glass substrate with could be utilised as a diagnostic tool for lung cancer, a RF classifier was used to test if the spectral data could be used to classify the lung cancer cells from NL20. Table 5 below shows the results of the RF classification of the cells using the spectra performed of all three cell lines together. Tables 6-8 show the classification of cells in a pairwise manner. For all groups the 2nd derivative spectra produced a better classification than the normal spectra improving on accuracy, precision and recall. The 3500-2700 cm^{-1} region provided a better classification than the 1800-1350 cm^{-1} region for the three cell lines together and the pairs apart from the pair of A549 and NL20, where the two regions performed equally. Using the larger region 3500-1350 cm^{-1} improved the classification over the smaller spectral regions for the classification

of all three cells together and for the A549 and NL20 pair but not the other two pairs for the classification of CALU-1 from A549 or NL20. Overall classification using the region 3500-2700 cm^{-1} performed the most consistently of the three regions tested.

Spectral region (cm^{-1})	AUC	Classification accuracy	F1	Precision	Recall
3500-1350	0.859	0.718	0.714	0.723	0.718
3500-1350 2nd derivative	0.960	0.885	0.884	0.888	0.885
3500-2700	0.899	0.724	0.721	0.724	0.724
3500-2700 2nd derivative	0.955	0.851	0.852	0.856	0.851
1800-1350	0.839	0.684	0.680	0.684	0.684
1800-1350 2nd derivative	0.888	0.718	0.716	0.718	0.718

Table 5 Random forest classification results of A549, CALU-1 and NL20 spectra.

Spectral region (cm^{-1})	AUC	Classification accuracy	F1	Precision	Recall
3500-1350	0.914	0.825	0.823	0.827	0.825
3500-1350 2nd derivative	0.994	0.933	0.933	0.941	0.933
3500-2700	0.923	0.817	0.814	0.823	0.817
3500-2700 2nd derivative	0.969	0.917	0.916	0.923	0.917
1800-1350	0.918	0.842	0.841	0.842	0.842
1800-1350 2nd derivative	0.970	0.917	0.917	0.917	0.917

Table 6 Random forest classification results of A549 and NL20 spectra.

Spectral region (cm^{-1})	AUC	Classification accuracy	F1	Precision	Recall
3500-1350	0.909	0.846	0.845	0.850	0.846
3500-1350 2nd derivative	0.972	0.942	0.942	0.943	0.942
3500-2700	0.995	0.952	0.952	0.953	0.952
3500-2700 2nd derivative	0.994	0.962	0.962	0.962	0.962
1800-1350	0.886	0.827	0.827	0.828	0.827
1800-1350 2nd derivative	0.938	0.817	0.817	0.817	0.817

Table 7 Random forest classification results of CALU-1 and NL20 spectra.

Spectral region (cm^{-1})	AUC	Classification accuracy	F1	Precision	Recall
3500-1350	0.792	0.707	0.704	0.707	0.707
3500-1350 2nd derivative	0.962	0.894	0.894	0.896	0.894
3500-2700	0.844	0.732	0.730	0.731	0.732
3500-2700 2nd derivative	0.962	0.902	0.902	0.904	0.902
1800-1350	0.723	0.618	0.615	0.615	0.618
1800-1350 2nd derivative	0.832	0.748	0.745	0.750	0.748

Table 8 Random forest classification results of A549 and CALU-1 spectra.

Below Figure 24 shows the confusion matrices for the RF classification of the three cell lines together (Table 5) using the 2nd derivative spectra. The confusion matrix for classification using the region 3500-1350 cm^{-1}, shows that the classifier performed well for the classification of A549 and NL20 classifying ≥90% of the both cells in the test set correctly. Where the classifier struggled using this region was in the classification of CALU-1 in which it misclassified 19.6% of CALU-1 as A549 and 3.9% as NL20. The classification using the region 3500-2700 cm^{-1} performed worse than using 3500-1350 m^{-1} for the classification of A549 and NL20 but still performed well correctly classifying 87.7% and 86% of A549 and NL20 cells respectively. However, the region 3500-2700 cm^{-1} performed best for the classification of

CALU-1 correctly classifying 82.4% of the cells with no misclassifications as NL20 and 17.6% misclassified as A549. The region 1800-1350 cm^{-1} performed worse than the other regions for the classification of all three cells. The classification of CALU-1 was also the worst performing as with the other regions, only classifying 64.7% of CALU-1 correctly. The confusion matrices demonstrate that where the classifier was producing the most misclassifications across the three regions was wrongly classifying CALU-1 as A549.

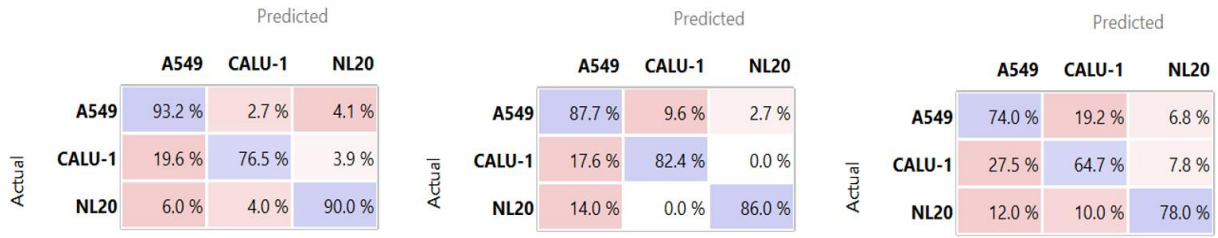

Figure 24 Confusion matrices of RF classification of A549, CALU-1 and NL20 using 2nd derivative FTIR spectra. Left: spectral region 3500-1350 cm^{-1}, middle: spectral region 3500-2700 cm^{-1}, right: spectral region 1800-1350 cm^{-1}.

Figure 25 below, shows the confusion matrices for the classification of A549 and NL20 using the three regions of the 2nd derivative spectra. All three selected regions of the spectra provided good classifications of A549 and NL20. Using the regions 3500-1350 cm^{-1} and 3500-2700 cm^{-1} classified A549 cells with higher accuracy than using 1800-1350 cm^{-1}. However, 1800-1350 cm^{-1} classified NL20 with more accuracy. Figure 24 shows the confusion matrices for the classification of the pair CALU-1 and NL20. As was also demonstrated with the classification of all three cells (Figure 24), the region 3500-2700 cm^{-1} produced the most accurate classification of CALU-1 and NL20 from and the worst performance was when using the region 1800-1350 cm^{-1}. Using 3500-1350 cm^{-1} the classification of CALU-1 performed as

well as with 3500-2700 cm^{-1} but misclassified more NL20 cells. Figure 25 shows the confusion matrices of the classification for the pair A549 and CALU-1. Again, the classifier struggled to more to classify CALU-1 than A549 using all three regions.

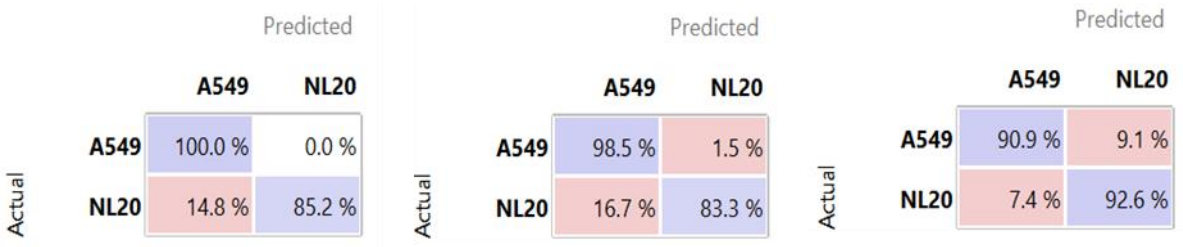

Figure 25 Confusion matrices of RF classification of A549 and NL20 using 2nd derivative FTIR spectra. Left: spectral region 3500-1350 cm^{-1}, middle: spectral region 3500-2700 cm^{-1}, right: spectral region 1800-1350 cm^{-1}.

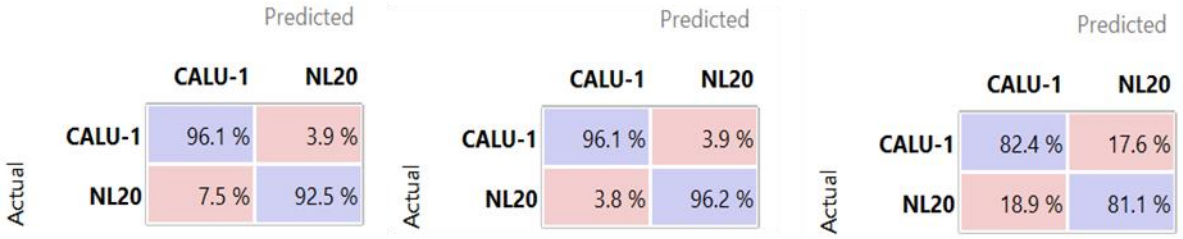

Figure 26 Confusion matrices of RF classification of CALU-1 and NL20 using 2nd derivative FTIR spectra. Left: spectral region 3500-1350 cm^{-1}, middle: spectral region 3500-2700 cm^{-1}, right: spectral region 1800-1350 cm^{-1}.

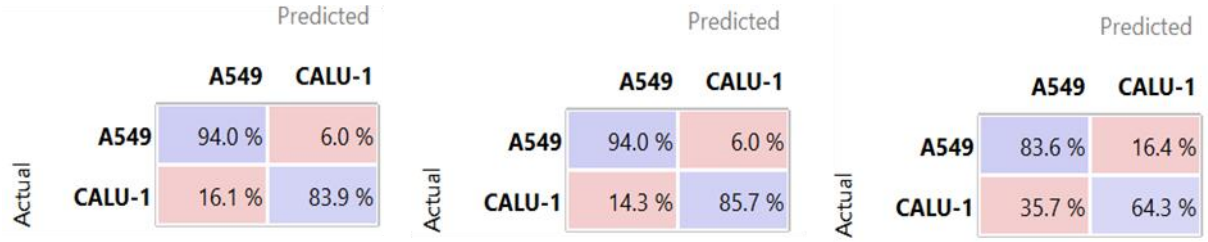

Figure 27 Confusion matrices of RF classification of A549 and CALU-1 using 2nd derivative FTIR spectra. Left: spectral region 3500-1350 cm^{-1}, middle: spectral region 3500-2700 cm^{-1}, right: spectral region 1800-1350 cm^{-1}.

Discussion

For FTIR spectroscopy to make the transition from research laboratories to clinical pathology laboratories for cancer diagnostics the substrates used must be made more cost-effective. The current commonly used substrates such as calcium fluoride and barium fluoride slides cost up to £50-60 per single slide which would make the use of FTIR spectroscopy for diagnostics prohibitively expensive. Alternative substrates must be investigated and assessed for FTIR spectroscopy to make the transition. Glass could be an alternative substrate that would reduce the cost of using FTIR spectroscopy while also being a commonplace material available in every pathology laboratory that pathologists and biomedical scientists are accustomed to working with. While there would be a compromise with using glass from the loss of some spectral information, there should be enough information retained on the lipids and proteins of the cells in the samples to classify cancer cells from normal cells.

The mean spectra showed a higher absorbance in the bands at 2920-2850 cm^{-1} in the cancer cells than the normal tissue derived NL20. The peaks at 2920 cm^{-1} and 2850 cm^{-1}

corresponds to the stretching of CH_2 groups in the methylene chains of cell membrane lipids. The peak at 2870 cm^{-1} corresponds to the stretching of CH_3 groups mostly within lipid fatty acid chains. A possible reason for the increase in intensity of the lipid related bands in the cancer cells spectra could be due to an increase in synthesis of lipids with long aliphatic chains. In cancer cells, there is a change in lipid metabolism to increased de-novo lipogenesis (DNL) of lipids while normal cells receive most of their lipids from circulating lipids (Merino Salvador et al., 2017). Many cancers adopt an alternative metabolic pathway for fatty acid synthesis. Cancer cells will rely on glutamate or acetate as substrates for fatty acid synthesis. One of the main pathways for fatty acid synthesis for cancer is through isocitrate dehydrogenase-1 dependent pathway in which the reductive carboxylation of glutamine derived α-ketoglutarate is used to synthesise acetyl-CoA (Beloribi-Djefaflia S, Vasseur S, 2016). The increased DNL is often used in cancer to fuel membrane biogenesis which contributes to cell proliferation and maintenance of a malignant phenotype. Polyunsaturated fatty acid synthesis is limited by DNL in mammalian cells which therefore results in more saturated or mono-unsaturated lipids in cancer cells. The saturated lipids pack more densely than unsaturated lipids thus changing the membrane characteristics of cancer cells. There is evidence of NSCLC differentially expressing 91 phospholipid species than in normal tissues (Marien et al., 2015). FTIR spectroscopy detects overall biochemical differences in the cells and these large-scale changes in lipid content will result in the spectral differences seen in the cancer cells.

While using glass coverslips cuts off the spectra at 1350 cm^{-1} which removes information from nucleic acids and carbohydrates in the fingerprint region, the amide I and II bands can still be viewed unlike when thicker glass slides. The amide I and II bands provide information on the protein content of the cells. Both cancer cells had a higher absorbance in the amide I

and II bands than NL20. The higher absorbance infers the cancer cells had a higher protein content than NL20. The amide I region is a sensitive region arising from stretching vibrations in C=O. Amide II arises from stretching vibrations in the C-N bond and bending vibration of the N-H bond in amide bond. Proteins are the main effector machinery in cells and there is a large change in what proteins are produced, the amounts of proteins and the structure of proteins in cancer cells compared to healthy cells resulting in spectral changes in IR spectra.

The metrics from the RF classification demonstrate that is possible to classify adenocarcinoma and SqCC lung cancer cells from non-malignant cells using FTIR spectroscopy data with good classification accuracy, precision, and recall. Using the 2nd derivative spectra further boosted the performance of the classifier. Spectral differences are more pronounced in the 2nd derivative spectra which is likely the cause of the stronger classification using 2nd derivative spectra. These spectral differences reflect the biochemical differences in the proteins and lipids of the cells. Also demonstrated was that classification of the two types of NSLC cells can be achieved. The ability to classify different types of cancer further enhances the utility FTIR spectroscopy would have in cancer diagnostics. FTIR spectroscopy could not only be used to triage normal samples from abnormal samples but also help to inform the decision making of the pathologist when typing a cancer. This would be particularly useful in cases where the morphology from a biopsy is unclear causing difficulty in coming to a diagnosis.

As was shown in the confusion matrices (Figure 24) the classifier performed well in the classification of the cancerous cells and the non-malignant NL20 from each other. Using the 3500-1350 cm^{-1} region which performed best for classifying all three cells together less than 10% of the cancer cells were misclassified as NL20 and only 10% of NL20 were misclassified

as cancer cells. This demonstrated that the methodology performs well in separating cancerous samples from non-cancerous samples. Where the classifier struggled was classifying the different cancerous cells from each other when classifying the three cell lines together. However, when the classification of the cancer cells was done as a pair there was a more accurate classification. This suggests that the methodology should first be used to separate cancer from non-cancer and then in a separate step could be used to aid in the typing of NSLC.

A549 and CALU-1 are representative of two different types of NSCLC, adenocarcinoma and SqCC respectively. With advances in personalised therapy for lung cancer knowing the type of lung cancer has become more important (Wang, Herbst and Boshoff, 2021). In the past these two cancer types have been largely treated similarly but now with more understanding of the differing biological signatures (Relli et al., 2019) there are clear clinical implications in terms of treatment and prognosis (Kawase et al., 2012). Therefore, the typing of NSCLC is becoming a more important diagnostic step because the personalised treatment plan will differ due to distinct gene expression and signalling pathways (Tian, 2017). For example, there adenocarcinoma has been seen to have better survival rates for the use of gemcitabine-platinum and taxane-platinum regimens together (Weiss et al., 2007). Whereas SqCC treatment could benefit more from cisplatin plus etoposide treatment. These factors further demonstrate how a FTIR spectroscopy platform that can classify types of NSLC cancers as demonstrated by the classification of A549 and CALU-1 could be of benefit. Adenocarcinoma and SqCC combined account for 85% of lung cancer cases. A majority of typing of lung cancers would be between these two types. If FTIR spectroscopy could quicken the typing stage, the diagnostic process would be improved for both clinician and patients.

The difference in classification performance from using the different spectral regions demonstrated how when using FTIR spectra for the classification of cells it is important to consider which regions of the spectra are being used. Which regions of the spectra are being used for the classification should influence the sample preparation. The region 3500-2700 cm^{-1} provided the best overall performance. The 4% PFA fixative retains the lipids in the cells unlike the methanol fixation in chapter 3. If methanol or other alcohol-based fixatives are used the lipid content of the cells can be lost. The sample preparation methodology used allowed for the use of the bands from the lipid content to produce a good classification. The cytospin as mentioned in chapter 3, again produced good cellularity to allow for efficient measurements of cells and consistent sample quality.

Ultimately the ideal use of FTIR spectroscopy for the clinical diagnosis of cancer would be in an automated system that can separate cancer from non-cancer in a robust manner with a sensitivity and specificity that is superior to current diagnostic methods. It was demonstrated that the whole fingerprint region is not necessary for a robust classification. The smaller regions of the spectra used can provide a robust classification. The lipid and amide A region of the spectra provided the best classification overall across classification of different pairs and the group of cell lines. If lung cancer can be classified from non-cancer only using the 3500-2700 cm^{-1} region of the spectra, standard 1 mm glass slides could be a viable substrate for diagnostic purposes because the spectra above 2000 cm^{-1} is not obscured by the thicker glass (Rutter et al., 2018). Using glass slides would reduce the disruption caused by implementing FTIR spectroscopy diagnostics into current workflows in pathology laboratories because glass slides are currently used for histological diagnostic techniques. While the use of the lipid bands for classification causes no problems for unstained samples it would not be applicable to stained samples as staining removes lipid

content from cells due to the alcohols in the dyes. Cells would have to be measured prior to any staining for the lipid content to be used for classification. However, if coverslips are used as a substrate, stained samples could be classified using the amide I and II bands. While the amide region provided a worse classification than the lipid bands, it still provided a good classification when A549 or CALU-1 was classified from NL20 as pairs. Furthermore, after spectroscopy measurements are taken the coverslip can be mounted to a glass slide for easier handling if staining of the sample is required to allowing the methodology to be easily fit into current diagnostic methodology.

Conclusions

NSCLC can be classified from non-malignant lung cells using a RF classifier with FTIR spectroscopy data. This was possible using glass coverslips as a substrate. The spectral region 3500-2700 cm^{-1} provided a better overall classification of A549, CALU-1 and NL20 from each other than the region 1800-1350 cm^{-1}. Use of the 2^{nd} derivative spectra further improved the performance of the classifier. Using the 2^{nd} derivative of both regions of 3500-2700 cm^{-1} and 1800-1350 cm^{-1} gave the best classification of all cell lines together and the A549 and NL20 as a pair. For classification of the pairs CALU-1 and NL20, and A549 and Calu-1 the 2^{nd} derivative of the region 3500-2700 cm^{-1} gave the best classification. This research demonstrated that the sample preparation methodology proposed in chapter 3 using a cytospin and PFA fixation with a glass coverslip substrate is applicable for classification tasks of lung cancer cells using FTIR spectroscopy.

Chapter 5: Classification of breast cancer cells from non-cancer breast cells on a glass substrate using FTIR microspectroscopy with machine learning.

Introduction

In the previous chapters, I have investigated the use of FTIR spectroscopy with glass substrates to classify lung cancer cells from non-malignant lung cells using the sample preparation methodology outlined in chapter 3. This chapter will continue this research by investigating if the same methodology of FTIR microspectroscopy with a RF classifier can classify breast cancer cells and non-cancerous breast cells using a benchtop spectrometer. It is important to assess if the methodology works for other solid cancers to fully assess its viability. This work was performed using a Thermo Nicolet iN10 benchtop spectrometer with a globar light source. It was important to show how the methods perform using a benchtop spectrometer as this is what will be available to most laboratories. Breast cancer was chosen to be investigated as it is the most common cancer in the UK accounting for 15% of recorded cancer cases (Breast cancer statistics, Cancer Research UK, 2017). Being able to reduce the time taken for diagnosis of breast cancer would greatly help relieve pressure on pathology laboratories and improve patient outcomes through quicker diagnoses.

Current diagnosis of breast cancers relies upon the use of imaging modalities and histological imaging of biopsy samples (Cardoso *et al.*, 2019). The presence of breast cancers is initially detected using imaging modalities, mainly mammography. A mammography is the application of a low dose of x-rays to image the position and size of a cancer which can give

indication to stage and invasiveness of the tumour. When a potential tumour is found, a biopsy sample is taken. The biopsy sample is prepared for histological diagnosis by fixation, cutting and staining. Histological methods of diagnosis require a trained pathologist to assess if the biopsy is indicative of cancer by comparing the cell morphology and tissue differentiation to normal cells and tissue. This method of diagnosis can be subjective and often require multiple pathologists if a case is difficult to distinguish. The use of FTIR spectroscopy would help to provide on objective measurement of whether the sample is cancerous or not and rely less on subjective assessments (Su and Lee, 2020). This proposed system would use FTIR Spectroscopy on the biopsy sample after mammography and before histological analysis. Being able to separate what is cancer and not cancer with FTIR spectroscopy would reduce the number of histological investigations on needed saving time for pathologists and allow for more focus upon the cancerous samples and get to a diagnosis quicker. Once the presence of the cancer has been confirmed the next steps of diagnosis can be carried out including staging and determining the type of breast cancer.

In the UK there is a breast cancer screening offered to women between the ages of 50 and 70 years of age using mammography to screen for the presence of cancer (Breast screening for Breast cancer, Cancer Research UK, 2018). Screening helps to identify breast cancer in the early stages which is vital for providing a patient the best treatment and chances of survival. However, with screening there are many cases of false positives and overdiagnosis where non-cancerous and benign legions might be identified as cancer (Marmot et al., 2012). FTIR spectroscopy could be used to filter out these cases where screening has caught non-cancerous or benign legions that do not require any further action at the current time. It was important, that the proposed methodology was tested if it could classify a non-invasive

cancer from an invasive cancer because it would allow pathologists to further allocate time and resources to the most urgent cases where the cancer has become invasive.

Any newly proposed diagnostic methods for breast cancer should work in conjunction with current methods to identify the key markers on the cancer because they are critical for deciding the best treatment plan. There are three main targets in the molecular testing of breast cancer, the oestrogen receptor alpha (ER), the progesterone receptor (PR) and epidermal growth factor 2 (ERBB2/HER2) (Hammond *et al.*, 2010). Approximately 70% of invasive breast cancers express ER. Expression of ERα and PR is closely linked, and PR expression is a marker of ERα signalling. HER2 is overexpressed in in approximately 20% of invasive breast cancers. Patients with HER2+ cancer can benefit from immunotherapy such as trastuzumab and pertuzumab and treatment from small molecule tyrosine kinase inhibitors (Maximiano *et al.*, 2016). Approximately 15% of breast cancers are triple negative meaning they do not express any three of these molecular markers. Triple negative breast cancer patients have a high rate of relapse in the first 3 to 5 years post treatment. These key markers are identified through immunohistochemistry. FTIR spectroscopy used with glass substrates can be fit into current workflows of immunohistology for marker identification. The label free and non-destructive nature maintains sample integrity for further analysis such as immunohistochemistry and the glass substrate allow microscopy to be performed.

The staging of breast cancers is an important step with the cancer being more easily treated in early stages, demonstrated by only 25% of patients surviving more than 5 years after diagnosis of stage 4 cancer while 98% of patients survive more than 5 years after diagnosis of stage 1 cancer (Breast cancer statistics, Cancer Research UK, 2017). The stage of the cancer is vital for determining the appropriate treatment plan. Therefore, it was important

to test if the proposed methodology using FTIR spectroscopy can aid in the staging of breast cancers by being able to classify between an invasive and a non-invasive breast cancer.

In this study two cancer breast cell lines and a normal breast cell line prepared on glass substrates were measured with FTIR spectroscopy. A RF classifier was used to test if the cells could be classified from each other using the FTIR spectra. The three cell lines represent different stages of breast cancer. MCF10A is derived from non-cancerous breast epithelium, MCF7 is a non-invasive ductal carcinoma line and BT549 is an invasive ductal carcinoma line. After demonstrating that this methodology can be used for classifying lung cancer from non-malignant lung cells and different types of NSCLC, it was important to test if the methodology can be used with other types of solid tumours and if it can be used to aid in distinguishing non-invasive cancers from invasive cancers.

Aims

1. Assess how well breast cancer cells can be classified from non-cancer derived breast cells on a glass substrate using benchtop FTIR spectroscopy and a RF classifier.
2. Investigate if non-invasive breast cancer cells can be classified from invasive breast cancer cells on a glass substrate using benchtop FTIR spectroscopy and a RF classifier.
3. Investigate which region of the spectra provides the best classification of the cells.

Methods

Cells

The breast cancer cell lines BT549 and MCF7 and the normal breast cell line MCF10A were used for the experiment. BT549 is derived from an invasive ductal carcinoma. MCF7 is derived from a metastatic breast adenocarcinoma. MCF10A is derived from non-cancerous breast epithelial cells. For a detailed methodology on the culture of the cells refer to the relevant section in chapter 2.

Sample preparation

Cells were collected from the flasks and resuspended in 0.9% normal saline and brought to a concentration of $1x10^6$ per 1 ml. Cells were applied to the glass coverslips by cytospin ran at 900 rpm for 1 minute using 20 µl of the cell solution. The samples were immediately fixed using 4% PFA and incubated for 15 minutes. After fixation, excess PFA was washed off with one wash of 0.9% normal saline and three washes with deionised water. Samples were prepared in three individual experiments with four samples of each cell line being produced each experiment. The cells from all three cell lines were collected after two weeks of culture.

FTIR spectroscopy

The samples were measured using transmission FTIR spectroscopy on a Thermo Fisher Nicolet iN10(mx) spectrometer with a globar light source. An aperture size of 15 x 15 µm was used and 256 co-added scans were taken per spectrum. The background was taken of clear section of the glass coverslip without any cells before each cell measurement. Measurements were taken from the centre of each cell. 100 spectra from each cell line were recorded with each spectrum being from a different individual cell. The measurements were taken equally across the three experiments and across the replicates from each experiment.

Pre-processing and data analysis

The spectra were pre-processed by first cropping the spectra removing the region below 1350 cm^{-1} which was obscured by the glass. The spectra were cropped to three regions for analysis 3500-1350 cm^{-1}, 3500-2700 cm^{-1} and 1800-1350 cm^{-1}. PCA denoising using 10 components and a Savizky-Golay filter with a window size of 5 and polynomial of 2 was used to reduce noise in the spectra. EMSC using the average spectra of the cells as a reference was used to normalise the spectra and correct the baseline for any defects caused by variation in sample thicknes. Average spectra were produced from 100 spectra of individual cells from each cell line.

2^{nd} derivative spectra were produced by adding a 2^{nd} derivative with the Savitzky-Golay filter. The PCA denoising was changed to 12 components and the window size of the Savitzky-Golay filter changed to 21 to remove noise added by the 2^{nd} derivative.

The spectra were split 70:30 into a training and testing set respectively. A RF classifier was used to test if the three cell lines could be classified to assess how well invasive breast cancer, non-invasive breast cancer and non-cancerous breast cells could be classified from each other using FTIR spectroscopy data measured using a glass coverslip substrate.

Results

The aim of this investigation was to test if breast cancer cells can be classified from benign breast cells and non-invasive cancer cells from invasive cancer cells using FTIR spectroscopy with a glass substrate. Visual inspection of the average spectra shown in Figures 28-30 of the three cell lines showed that there are biochemical differences between the cells. There was a difference between the three cell lines in the absorbance and shape of lipid bands and amide A (3500-2700 cm^{-1}) and the amide I and II bands (1800-1350 cm^{-1}). The invasive BT549 cells had more prominent peaks in the lipid bands which infers a greater amount of lipids in the cells compared to MCF7 and MCF10A. All three cell lines had a different absorbance at the amide I peak with BT549 having the lowest absorption and MCF7 having the highest. The shape of the amide II band for BT549 spectrum had a different shape to the other two cell lines spectra with the BT549 having a sharper point at the peak and having a less pronounced shoulder at 1521 cm^{-1}. The amide II band of MCF7 was shifted two wavenumbers from that of BT549 and MCF10A. The BT549 spectrum the highest absorbance in amide II, followed by MCF7 and MCF10A had the lowest absorbance. The absorbance of BT549 is lower in the amide A band than MCF7 and MCF10A and the peak is shifted four wavenumbers. These differences in the amide peaks infer differences across the three cells in their protein content.

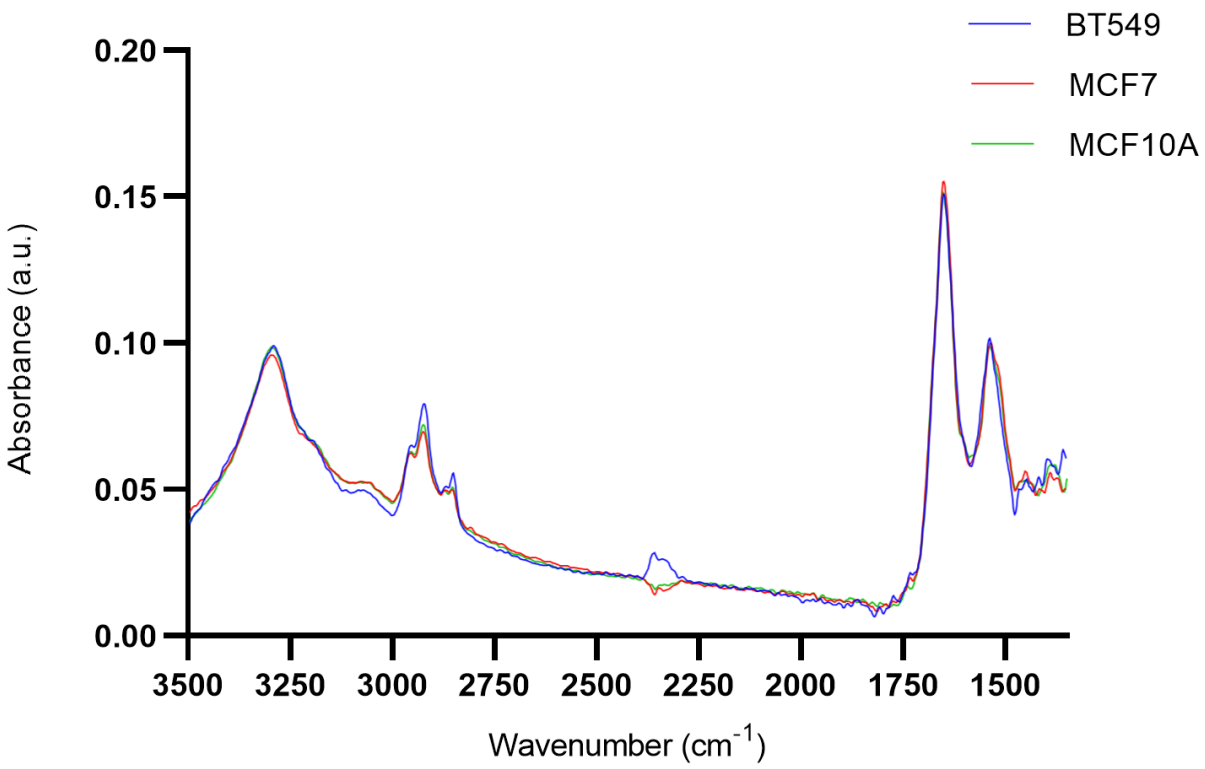

Figure 28 Average FTIR spectra from 100 spectra of BT549, MCF7 and MCF10A in the region 3500-1350 cm^{-1}.

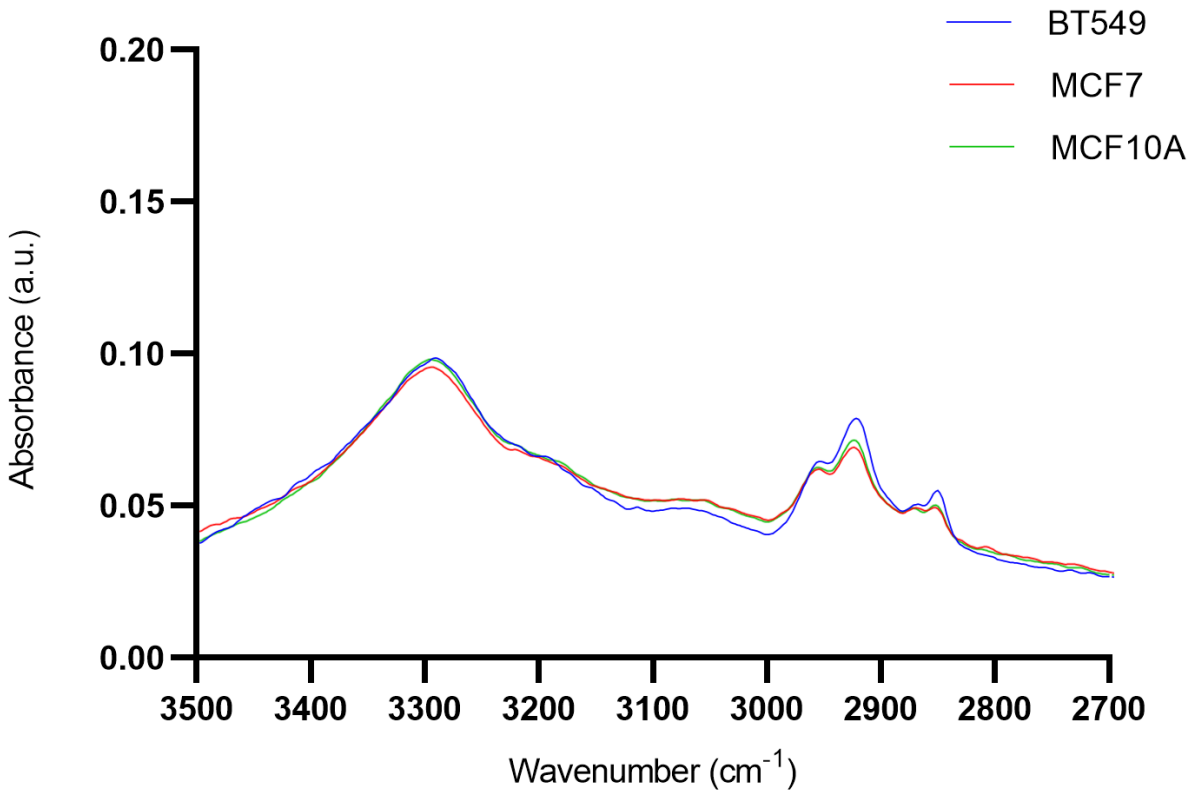

Figure 29 Average FTIR spectra from 100 spectra of BT549, MCF7 and MCF10A in the region 3500-2700 cm^{-1}.

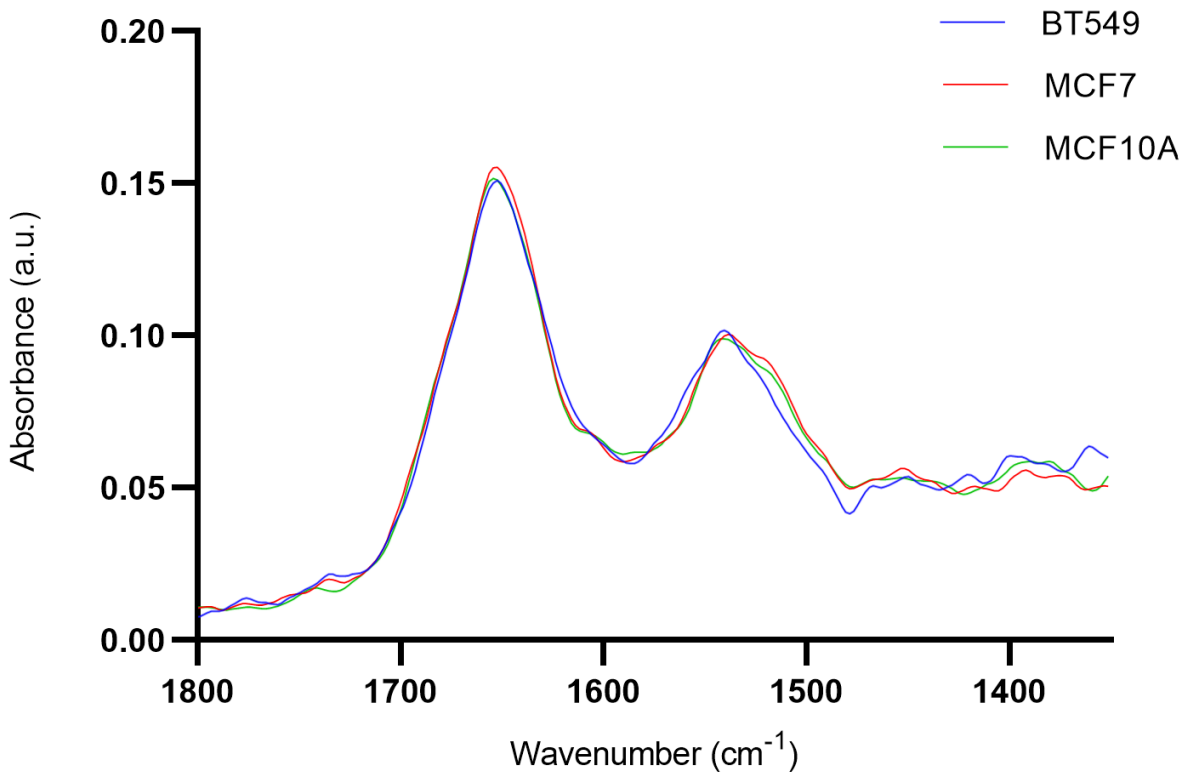

Figure 30 Average FTIR spectra from 100 spectra of BT549, MCF7 and MCF10A in the region 1800-1350 cm^{-1}.

A RF classifier was chosen to test the classification because RF works well with data that has many features and multiple classes such as this spectral dataset where each wavenumber is its own feature. Table 9 shows the performance of the RF classification. Classifications were done using the three d spectral regions to test if there is an optimum region of the spectrum to use for classifications. The regions used were the whole spectrum (3500-1350 cm^{-1}), the amide I and II peaks together (1800-1350 cm^{-1}) and the lipid peaks and amide A together (3500-2700 cm^{-1}). The region3500-1350 cm^{-1} and 3500-2700 cm^{-1} both performed similarly well with the classifier both with an F1 score of 0.901. The region 1800-1350 cm^{-1} also performed well with a F1 score of 0.862 albeit lower than the other two regions tested.

Region (cm^{-1})	AUC	Classification accuracy	F1	Precision	Recall
3500-1350	0.970	0.901	0.901	0.901	0.901
1800-1350	0.952	0.864	0.862	0.863	0.864
3500-2700	0.987	0.901	0.901	0.906	0.901

Table 9 Random forest classification result of BT549, MCF7 and MCF10A spectra.

Table 10 lists the 10 features with the most importance for the RF classifier in each spectral region used. The feature importance informs which features in the data were most predictive of the target variable. FTIR spectra have a large number of features with each wavenumber being a feature therefore narrowing which wavenumbers are the most useful for classifying the cells. The ten highest ranked features in the regions 3500-1350 cm^{-1} and 3500-2700 cm^{-1} were within the amide A band. This is likely why the classification performed similarly using those two regions of the spectra because the RF gave the most importance to features in the amide A band. As can be seen in Figures 28 and 29 there are differences in absorbance in the amide A band between the three cell lines especially with the less prominent peak of MCF7. The ten most important features for the RF classification using 1800-1350 cm^{-1} were mostly within the amide II band between 1479-1484 cm^{-1}. As noted above there were noticeable differences in the shape, absorbance, and position in the average spectra of the three cell lines in amide II.

3500-1350 cm^{-1}	3500-2700 cm^{-1}	1800-1350 cm^{-1}
3122	3322	1482
3230	3320	1477
3066	3417	1450
3137	3442	1475
3060	3303	1481
3085	3050	1484
3232	3305	1799
3276	3048	1479
3120	3411	1793
3126	3311	1587

Table 10 Ten most important features for the random forest classification.

Figures 31-33 are the confusion matrices of the RF classification using the three selected spectral regions. The confusion matrices show the percentage of the cells correctly and incorrectly classified. The RF classifier performed best for the classification of BT549 where 95.7% of BT549 cells were correctly classified with all three regions used. The remaining 4.3% of the BT549 cells were misclassified as MCF10A when the regions 3500-1350 cm^{-1} and 3500-2700 cm^{-1} were used and using 1800-1350 cm^{-1} BT549 was misclassified as MCF7. The

correct classification of MCF7 and MCF10A was less consistent. The region 3500-2700 cm^{-1} provided the best classification of MCF7 with 93.1% of cells correctly classified and the region 1800-1350 provided the worst with 75.9% of cells correctly classified. 86.2% of MCF10A cells were correctly classified using 3500-2700 cm^{-1} and 1800-1350 cm^{-1} and 79.3% were correctly classified using 3500-2700 cm^{-1}. The misclassifications of MCF10A using 3500-1350 cm^{-1} and 1800-1350 cm^{-1} (3.4%) were of BT549 and the remaining misclassifications were MCF7. While all the misclassifications of MCF10A were MCF7 using 3500-2700 cm^{-1}.

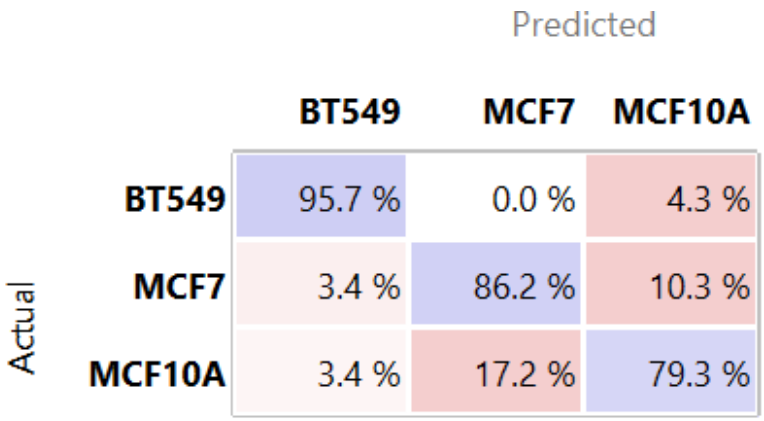

Figure 31 Confusion matrix of random forest classification of BT549, MCF7 and BT549 using FTIR spectra in the region 3500-1350 cm^{-1}.

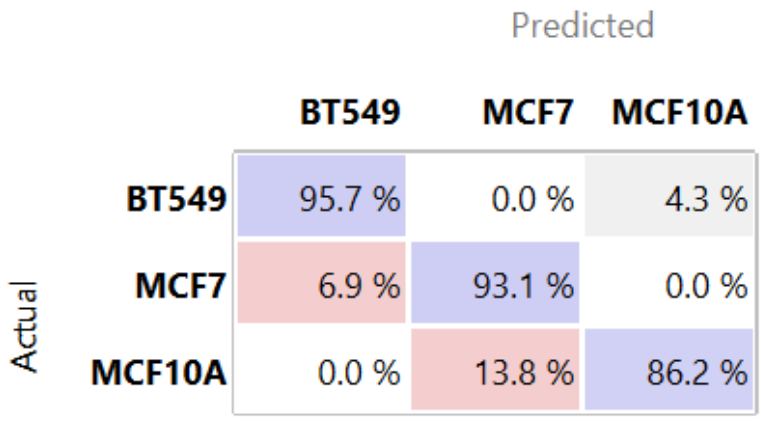

Figure 32 Confusion matrix of random forest classification of BT549, MCF7 and BT549 using FTIR spectra in the region 3500-2700 cm^{-1}.

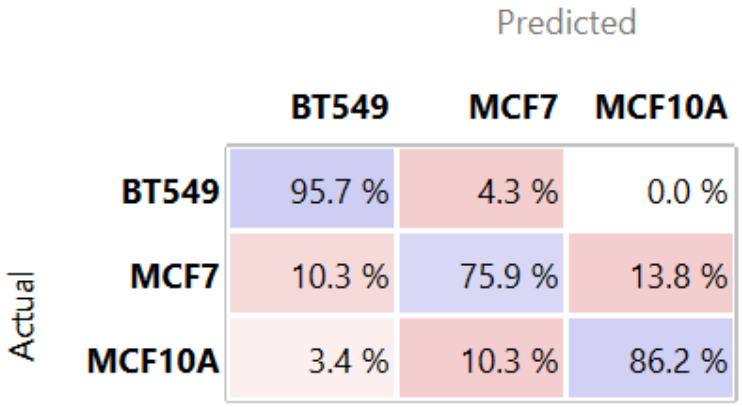

Figure 33 Confusion matrix of random forest classification of BT549, MCF7 and BT549 using FTIR spectra in the region 1800-1350 cm^{-1}.

Using the 2nd derivative spectra (Figures 34-36) improved the classification after increasing the components in the PCA denoising to 12 and the window size of the Savitzky-Golay filter to 21 to account for the noise introduced by the 2nd derivative. The classification metrics using the 2nd derivative spectra are shown in Table 11 below. The 2nd derivative spectra can resolve differences in the bands that are not clear in the normal spectra. There were

differences in all the major bands shape, position and absorbance as shown in the average 2nd derivative spectra of BT549, MCF7 and MCF10A. In the region 3500-2700 cm^{-1} there is a difference in the absorbance of all three cells at the peak 3236 cm^{-1}. BT549 has a considerably lower absorbance and less pronounced peak at this position. The 3236 cm^{-1} corresponds to the amide A peak from the N-H stretching vibration in the peptide bonds of proteins. The MCF7 has a higher absorbance showing a more pronounced peak than the other cells at 2988 cm^{-1}. The 2988 cm^{-1} peak is not prominent in the normal spectra this demonstrates how the 2nd derivative can reveal more differences within the spectra. The 2900 cm^{-1} is also prominent in the 2nd derivative spectra but not the normal spectra. The BT549 cells had the highest absorbance at this peak and was shifted to 2898 cm^{-1}. The MCF7 had the lowest absorbance, and the peak was split. MCF10A had a higher absorbance than MCF7 but lower than BT549. The shoulder at 2872 cm^{-1} is much more prominent in the BT549 spectrum than the MCF7 and MCF10A spectra. At 2832 cm^{-1} BT549 again has a higher absorbance than MCF7 and MCF10A. These peaks at 2900-2832 cm^{-1} are from CH_2 stretching vibrations of lipids and proteins. In the 1800-1350 cm^{-1} there is a large shift of MCF7 at 1713 cm^{-1} where BT549 and MCF10A have the peak positioned at 1703 cm^{-1}. The peak forms part of the amide I peak and a shift in peak position could suggest a change in protein secondary structures. At 1620 cm^{-1} differences were in this peak that is also part of amide I. The BT549 peak is shifted 1618 cm^{-1} and has a lower absorbance and the MCF10A absorbance is lower than the MCF7. The differences in the amide I peaks suggest different overall protein compositions in the cells. Spectral differences could be seen in the peaks from 1585-1485 cm^{-1} however there is a considerable amount of noise introduced from the 2nd derivative, so it is difficult to interpret if the spectral differences were from biochemical differences in the cells or caused by noise.

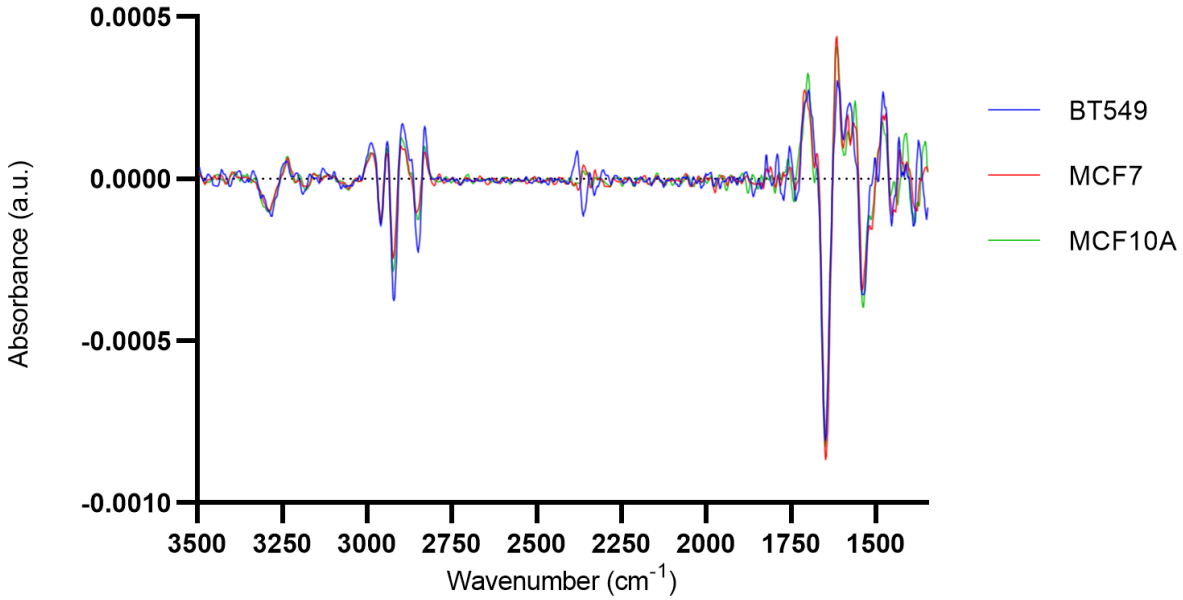

Figure 34 Average 2nd derivative FTIR spectra from 100 of BT549, MCF7 and MCF10A in the region 3500-1350 cm^{-1}.

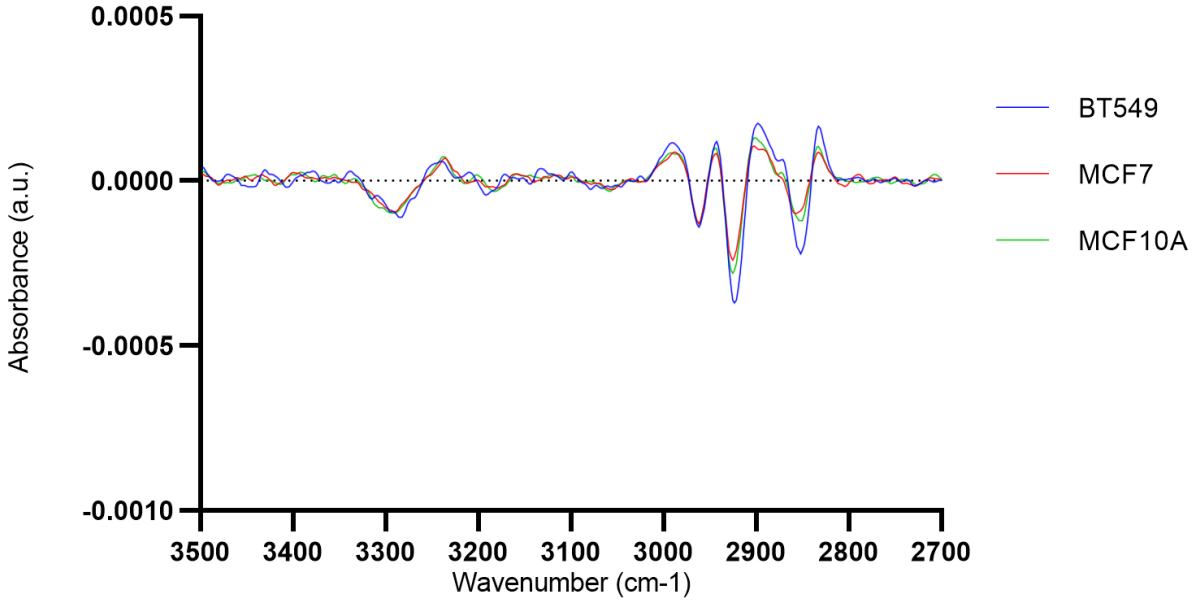

Figure 35 Average 2nd derivative FTIR spectra from 100 spectra of BT549, MCF7 and MCF10A in the region 3500-2700 cm^{-1}.

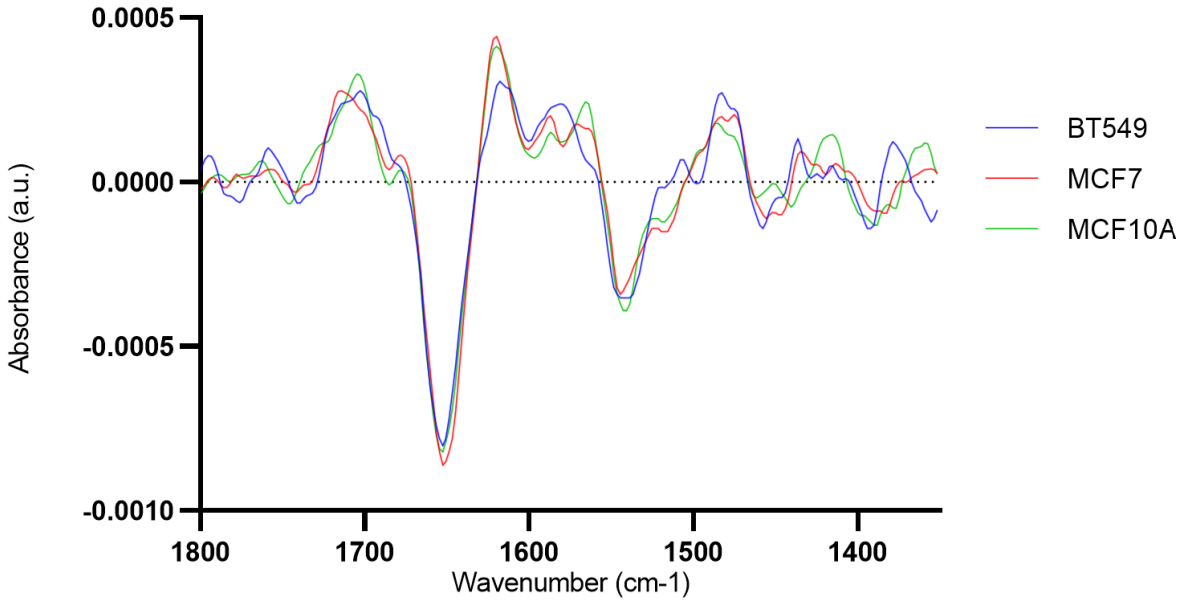

Figure 36 Average 2nd derivative FTIR spectra from 100 spectra of BT549, MCF7 and MCF10A in the region 1800-1350 cm^{-1}.

Table 11 and Figures 37-39 below show the results of the random forest classification of BT549, MCF7 and MCF10A using the 2nd derivative spectra. There was an overall improvement in the classification performance shown by an improvement in all the metrics in Table 11 in comparison to the classification with the normal spectra in Table 9. 3500-1350 cm^{-1} gave the best overall classification followed by 3500-2700 cm^{-1} and 1800-1350 cm^{-1}. The classification of MCF7 and MCF10A was improved using the 2nd derivative spectra over the normal spectra using the region 3500-1350 cm^{-1}. The classification of MCF7 had the biggest improvement in the region 3500-1350 cm^{-1} with 93.1% of the cells correctly classified compared to 86.2% using the normal spectra 96.6% and 79.3% of MCF7 were correctly identified using 3500-2700 cm^{-1} and 1800-1350 cm^{-1}. The classification of MCF10A with 2nd derivative spectra improved for 3500-2700 cm^{-1} and 1800-1350 cm^{-1} at 96.6% correctly classified but did not improve for 3500-2700 cm^{-1}. All the misclassifications of MCF10 were

classified as MCF7 when the 2nd derivative spectra were used. 95.7% of BT549 cells correctly classified using the regions 3500-2700 cm^{-1} and 1800-1350 cm^{-1}, this was the same amount as the normal spectra. The classification of BT549 worsened using the 2nd derivative of 3500-1350 cm^{-1} with 91.3% of the cells correctly classified. All the BT549 misclassifications using 2nd derivative spectra were attributed to MCF7.

Region (cm^{-1})	AUC	Classification accuracy	F1	Precision	Recall
3500-1350	0.982	0.938	0.939	0.940	0.938
1800-1350	0.970	0.901	0.899	0.903	0.901
3500-2700	0.990	0.926	0.926	0.932	0.926

Table 11 Random forest classification results for classification of BT549, MCF7 and MCF10A using 2nd derivative FTIR spectra.

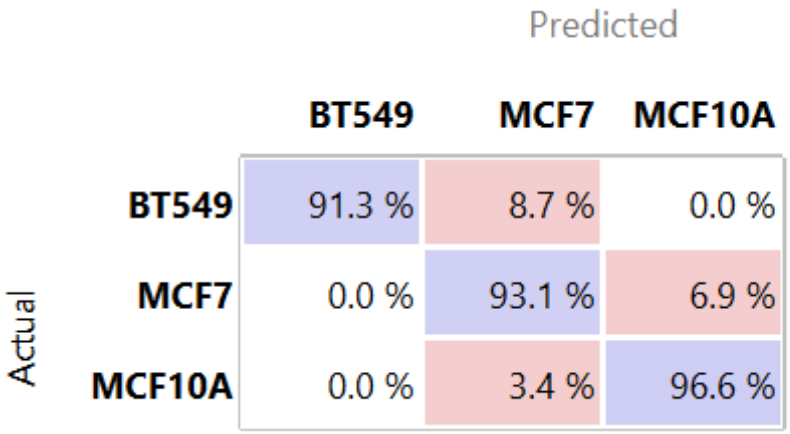

Figure 37 Confusion matrix of random forest classification of BT549, MCF7 and MCF10A using 2nd derivative FTIR spectra in the region 3500-1350 cm^{-1}.

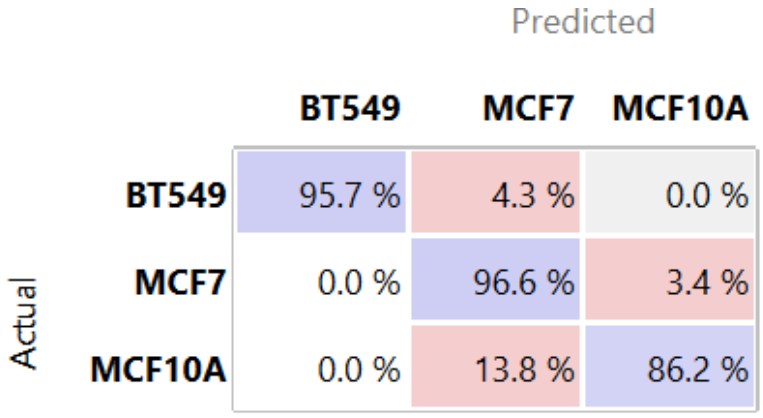

Figure 38 Confusion matrix of random forest classification of BT549, MCF7 and MCF10A using 2nd derivative FTIR spectra in the region 3500-2700 cm^{-1}.

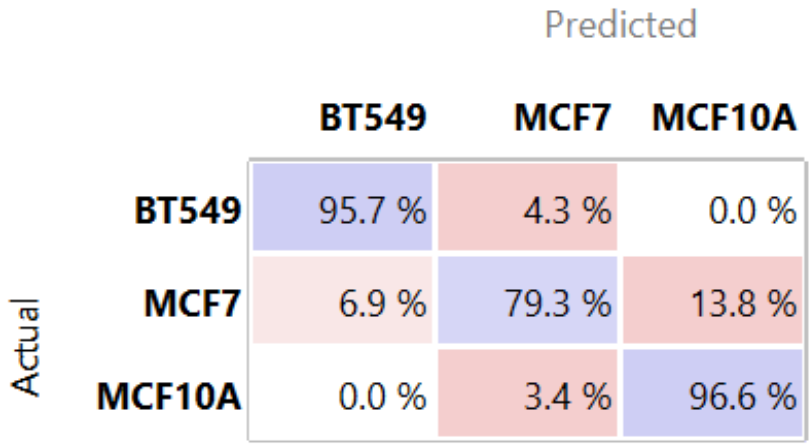

Figure 39 Confusion matrix of random forest classification of BT549, MCF7 and MCF10A using 2nd derivative FTIR spectra in the region 1800-1350 cm^{-1}.

Discussion

The previous chapter demonstrated that two different NSCLC cells could be classified from healthy tissue derived NL20 cells using a RF classifier and FTIR spectroscopy. This chapter aimed to demonstrate the same methods could be used for the classification of other solid tumours in this case breast cancer. While the classification of A549 and CALU-1 from NL20

used a synchrotron-based spectrometer it was important to demonstrate in this chapter that a benchtop spectrometer can also provide high quality spectra for the classification of cancer cells from non-cancer cells. While the spectra measured using the benchtop spectrometer had more noise, they were still of good quality with clearly defined bands. The bands within the average spectra of BT549, MCF7 and MCF10A demonstrated spectral differences that indicated differences in both the proteins and lipids of the cells. This is expected as there are significant changes in the protein and lipid content of breast cancer from normal cells and from non-invasive to invasive cancer.

All selected spectral regions 1800-1350 cm^{-1}, 3500-2700 cm^{-1} and 3500-1350 cm^{-1}, provided good classification of BT549, MCF7 and MCF10A with classification accuracy above 85%. Although, the 3500-2700 cm^{-1} region provided the best classification overall. This region was also the best performing region for the classification of the lung cancer cells and NL20 followed by 3500-1350 cm^{-1} and 1800-1350 cm^{-1} when using the normal spectra. The classification of lung cancer cells from non-malignant lung cells and classification of breast cancer cells from non-malignant breast cells both performed best when using the higher wavenumber region 3500-2700 cm^{-1}. This region of the spectra contains information on both the lipids and protein content of the cells which could be why it results in a better classification than the 1800-1350 cm^{-1} where the bands are from vibrations in just the proteins. 3500-2700 cm^{-1} giving the best classification for both breast and lung cancer means the same methodology can be used for the two different cancers. Having a shared methodology that is used across different cancers would allow for an easier translation of FTIR spectroscopy to clinical diagnostics because it would simplify the process. Glass as a substrate is viable for both breast and lung cancer because 3500-2700 cm^{-1} is not affected using glass.

As with the lung cancer cell classification, the 2nd derivative spectra gave a better performance for the breast cancer classification than the spectra with no derivatives applied. The 2nd derivative spectra resolve broad bands into individual bands and resolves lower frequency areas of the spectra into higher frequency peaks. The bands resolved by the 2nd derivative all had differences in the three breast cell lines. These differences likely resulted in the better classification than the raw spectra where the differences in the bands are less pronounced. Using the 2nd derivative spectra, the region 3500-1350 cm^{-1} provided the best classification, correctly classifying >90% of BT549, MCF7 and MCF10A. For the classification of the lung cells, this region also gave the best classification when classifying the three cell lines together. None of the invasive breast cancer cells (BT549) and only 6.9% of the non-invasive breast cancer cells (MCF7) were misclassified as MCF10A. Only 3.4% of MCF10A were misclassified as non-invasive cancer cells. The results of this classification demonstrate that this methodology with the 2nd derivative spectra has the potential for separating cancerous samples from non-cancerous samples for both non-invasive and invasive breast cancer. Also demonstrated is the methodology can be used to separate the non-invasive cancers from invasive cancers. The methodology demonstrated the feasibility to be further developed to be used in a clinical setting. To further develop this methodology would require the use of primary cells from patients and testing on larger datasets. The methodology has shown that it could be developed to help triage cancer cases from non-cancerous cases and then further triage invasive cancer from non-invasive cancer. Triage into these categories would allow the most urgent invasive cases to be escalated further to the next steps in the diagnostic pathway for staging and typing to allow treatment plans to be made.

The increased noise introduced by using a 2nd derivative must be considered when applying it during pre-processing. The parameters of denoising methods must be increased in the pre-processing phase to produce a readable spectrum. This can complicate the preparation of the data because denoising must be balanced so the unwanted noise is removed but the important features are not removed. The spectra of the breast cell lines were collected using a benchtop spectrometer which has a lower resolution than the synchrotron light source used to collect the lung cell lines' spectra in chapter 4. The lower resolution of the benchtop spectrometer is noticeable in the noisier 2nd derivative spectra of the breast cell lines in comparison to the 2nd derivative spectra of the lung cell lines. Some of the differences seen in the 2nd derivative spectra of the breast could be because of noise and not from the biochemical differences in the cells. For this reason, when using a benchtop spectrometer, the noise introduced into spectra with a derivative applied could hinder classification if there is a lot of noise in the raw spectra. In such a case not using the 2nd derivative could be the better choice as it still provides a good classification while being more easily interpretable with less ambiguity of what is causing the differences seen in the spectral bands. In a situation where the raw spectra collected contain a considerable amount of noise using the 2nd derivative spectra is not advisable because the classifier could mistake the noise for features. The spectra acquired for this research were of high quality and the 2nd derivative spectra could still be used.

Using FTIR spectroscopy in the manner I have proposed to separate cancerous samples from non-cancerous samples the method must have a low rate of false positives as demonstrated by the good precision and recall from the classifications. A classification with many false positives would not provide much benefit as a new diagnostic technique because it would not aid the pathologists by separating the cancerous and non-cancerous samples. False

positives are especially a concern for breast cancer diagnosis because of the screening program that will produce more false positives and overdiagnoses. Time and resources are put into further testing of falsely diagnosed samples, taking it away from true positive cases of cancer. It is estimated for each breast cancer death prevented by screening, three cases are over diagnosed (Marmot et al., 2012). In England in the year 2016-2017, 70,000 women received a false positive screening result. Overdiagnosis can cause undue stress for patients while awaiting the results of diagnostic testing and in some cases results in unnecessary further testing and treatments. FTIR spectroscopy would be most beneficial at the stage of breast cancer diagnosis after there is a possible cancerous body found during screening, but it is uncertain if it is cancer. The FTIR spectroscopy diagnostic methods combined with the RF classifier could be used to aid in identification of the cancerous samples from non-cancerous samples and reduce the number of false positive to reduce overdiagnoses. The use of the glass substrates and preparation methods already used in pathology laboratories allows FITR spectroscopy to be used in conjunction with current diagnostic techniques. FTIR spectroscopy diagnostic techniques should not replace the current diagnostic methods because morphological features and the presence of ER, PR and HER2 are still vital for making a full diagnosis. An FTIR spectroscopy measurement can be taken prior to the cytological or immunohistochemistry staining of samples. In cases where the pathologist is unsure of the identity of a sample, analysis with spectroscopy and machine learning classification can provide more information to allow decisions to be less subjective.

FTIR spectroscopy on glass substrates could be particularly useful for the helping to diagnose breast cancer from fine needle aspirate cytology (FNAC). FNAC uses a very thin needle connected to a vacuumed syringe to aspirate a small amount of tissue and cells from a suspicious lesion. The diagnosis from a FNAC can be difficult particularly for borderline

breast lesions with the sensitivity and specificity values varying by a large amount often because of the small amount of cell and tissue retrieved. Typing and grading the tumour can be difficult from histopathological diagnostic techniques and diagnostic accuracy depends on the experience of the pathologist. A sometimes-high rate of false negatives from sampling errors or interpretation errors has put the value of FNAC in doubt among some clinicians (Mitra and Dey, 2016). Despite this FNAC is still used to obtain biopsies because of its efficiency, affordability, and safety profile. FNAC does not require the use of local anaesthetic or radiological assistance therefore harm to the patient is minimal. Core needle biopsy (CNB) uses a larger needle to remove a core of tissue from a suspicious site in the breast. CNB has become the preferred method among many clinicians because it provides higher diagnostic accuracy as a larger amount of tissue is provided giving a better picture of the site when using current diagnostic methods (Mitra and Dey, 2016). Interpretation of these biopsies is easier resulting in fewer false negatives. However, CNB is more invasive for the patient using a larger needle and removes more tissue. Local anaesthetic and radiological assistance are required for CNB procedures unlike FNAC. This is more painful for the patient, takes longer and is more expensive than FNAC. Another risk with the use of CNB with smaller lesions is that it could break up the lesions which makes further sampling and excision difficult. The use of FTIR spectroscopy could help to obtain improved diagnostic accuracy from FNAC biopsies and provide more diagnostic value. The initial diagnosis for the presence of cancer can be made from the biochemical properties of the cells in the sample and rely less on the interpretation of morphology. More information could be gained from FNAC in cases where the morphology is poor and difficult to interpret. Improving the diagnostic value of FNAC would be beneficial to patients because it is less invasive and has less risk than CNB. Many women tested for breast cancer after screening do not have cancer

and many undergo unneeded procedures to obtain biopsies. The screening process would be improved if the invasiveness of the procedure can be reduced while diagnostic value is maintained.

Research by Lasalvia et al also compared the FTIR spectra of MCF7 and MCF10A on glass substrates. They found that the average spectrum of MCF7 had a lower absorbance than MCF10A in the lipid bands (Lasalvia, Capozzi and Perna, 2021). I found a similar result in the average spectra of MCF7 and MCF10A shown in Figure 8 with a lower absorbance in the lipid bands of MCF7 than MCF10A at 2926 cm^{-1}. The FTIR spectrum is influenced by the pre-processing used which creates difficulty in direct comparisons between studies where different pre-processing is used. Lasalvia et al used SNV as a normalisation step while I used EMSC. Both are acceptable methods of normalising FTIR spectra, I selected EMSC for this study because it produced a better classification of the cells than SNV. The wide range of pre-processing used in clinical FTIR spectroscopy studies makes it difficult to directly compare different studies on the same subject. There is currently no agreed upon standard or optimal pre-processing method in the field of clinical FTIR spectroscopy. This is another hurdle in the translation of FTIR spectroscopy to a clinical application. If FTIR spectroscopy is to be used for diagnostics there must be a standard method of processing the spectroscopy data collected across pathology laboratories and hospitals. There are many variables to account for when applying pre-processing steps including choosing which pre-processing steps to apply, what order to apply the steps and which method to choose for each step. Each of these variables will affect the spectra and classification. The number of variables involved and there being no one correct method of pre-processing makes coming to a consensus on the optimum pre-processing methods a difficult task. Coming to a consensus

on how and how much data should be processed is a discussion that needs to take place within the field.

For FTIR spectroscopy to be used reliably in an automated manner for clinical diagnosis a large sample size of spectra must be collected for each type of cancer to get the best representation of the spectra of the cancer, find the best areas of the spectra for classification and train the classification algorithms to be precise and reliable. This is currently a hurdle in the field of diagnostic FTIR spectroscopy as there are few large-scale studies of patient samples. There will need to be large scale collaborations between academic researchers, healthcare providers and industry to conduct large scale data collection before FTIR spectroscopy could be used reliably in an automated manner for diagnosis of cancer. For this to happen there also needs to be more consensus in the field on the protocols used including how samples are prepared, the FTIR spectroscopy modality (transmission, ATR, transflection) used, substrates and pre-processing of spectra and data analysis.

Conclusions

The research in this chapter has demonstrated that the methodology can also be used for the classification of breast cancer cells from non-cancerous breast cells. Also demonstrated was that invasive and non-invasive breast cancer cells can be classified from each other. Like the classification of lung cancer cells, the 2nd derivative spectra produced a better classification than the spectra with no derivatives applied. The 3500-1350 cm^{-1} and 3500-2700 cm^{-1} produced better classifications for both lung and breast cancer cells than the region 1800-1350 cm^{-1}. It was demonstrated that FTIR spectroscopy with cells placed on a

substrate of glass coverslips could feasibly be used for the classification of breast cancer cells to aid in diagnosis.

Chapter 6: The use of FTIR spectroscopy to identify individual cancer cells from leukocytes in mixed samples.

Introduction

There has long been an interest in the use of liquid biopsies for cancer diagnosis and prognosis. Analysis of tissue biopsies is the current gold standard for diagnosis of solid cancers including lung cancer. Retrieval of tissue biopsies is an invasive method usually requiring a surgical procedure that can cause pain and distress to patients. Reliance on tissue biopsies prevents repeated sampling of the cancer and a poor-quality biopsy can make diagnosis difficult. Liquid biopsies utilise biofluids and their components for diagnostics. Liquid biopsies are less invasive and allow for more frequent testing and monitoring with less pain and distress to patients. Blood and its components have had the largest amount of research devoted to it for liquid biopsy diagnosis because it is easily accessible and contains a wealth of information. Blood contains several tumour related materials that could be used as biomarkers including cell-free DNA (cfDNA), circulating tumour DNA (ctDNA), extracellular vesicles (EVs), mRNA (messenger RNA), miRNA (microRNA), circulating tumour cells (CTCs) and tumour educated platelets (Lone et al., 2022).

The focus of this research was how FTIR spectroscopy could be used for the identification of CTCs. CTCs are cells that have detached from primary or secondary tumours and travelled away from the site of the tumour in the blood (Yang et al., 2019). The migration and seeding of CTCs and CTC clusters is thought to cause the growth of metastases. The identification of CTCs in peripheral blood could be a valuable tool for the diagnosis and prognosis of cancers in a non-invasive manner. The current difficulty in the use of CTCs is their low number in the

blood due to most CTCs perishing from mechanical stresses and attacks from the immune system. The number of individual CTCs in blood number in range from 1 to >50 CTCs per 7.5 ml of blood (Syrigos et al., 2018). Clusters of CTCs known as circulating tumour micro emboli contain at least 2 CTCs with some containing up to 50 CTCs. CTCs are often found associated with other cells such as leukocytes, cancer associated fibroblasts, endothelial cells, and platelets. Currently, only one method of CTC identification called CellSearch has gained FDA approval. The CellSearch has been available since 2004 but has had little use and acceptance in the clinical field for cancer diagnostics due to being complex, difficult to use and expensive (Andree et al., 2016).

With the current difficulty in identification of CTCs several methodologies have been developed and proposed. These methods can be broadly separated into two categories: 1. Label dependent where CTCs are identified based on expression of surface antigens. 2. Label independent where CTCs are identified based on physical properties such as size, density, deformability or dielectric properties (Sundling & Lowe, 2019). Methods of positive CTC selection using surface antigens as is used for CellSearch, have the disadvantage that selected antigens may not be expressed on the CTCs as antigen expression can vary greatly between not just different cancers but also within tumour subtypes within a tumour. A selection of different antibodies may be required for CTC identification which can quickly become complex and expensive. Matters are further complicated by the epithelial-mesenchymal-transition (EMT). EMT is dedifferentiation process in which epithelial cells gain mesenchymal traits that confer stem like properties to aid in migration (Kalluri & Weinberg, 2009). The loss of epithelial markers makes using antibodies against epithelial antigens often unsuccessful for CTC identification. There is still a lack of a robust methodology for CTC identification in peripheral blood. An ideal method of CTC identification would be able to

identify CTCs of all types of tumours, be fast, not too complex, and have a high throughput while minimally disturbing current cancer management pathways.

FTIR spectroscopy could be a technique that has the qualities to form part of a methodology for CTC identification. It can identify cells by their biochemical properties and as the previous chapters have demonstrated and multiple other studies, it has potential as a diagnostic tool for cancer. As it is label free it does not depend on antibody antigen interactions like the CellSearch which makes FTIR spectroscopy less complex, expensive and it does not require the same level of expertise to use. As the biochemical properties of CTCs are very different to the surrounding blood cells, it should facilitate the recognition of CTCs in blood.

This research investigated if individual lung cancer cells can be distinguished from leukocytes on a glass substrate using FTIR spectroscopy. This was investigated using whole peripheral blood doped with lung cancer cells. This was a feasibility study to test how FTIR spectroscopy could be used for CTC identification. To the best of my knowledge, this was the first time FTIR spectroscopy was tested for use as method of CTC identification in blood samples placed on glass substrates. There has been little research around using vibrational spectroscopy for the identification of CTCs in peripheral blood.

Aims

1. Assess if A549 and CALU-1 lung cancer cells can be distinguished from leukocytes on glass substrates using FTIR spectroscopy.
2. Using RF with spectral data, assess if A549 and CALU-1 cells can be identified in samples of mixed cancer cell and leukocytes populations on glass substrates.

Materials and Methods

Cells

The following cell lines were used for this research: A549 (European Collection of Cell Cultures – ECACC) lung adenocarcinoma, CALU-1 (ECACC) lung squamous cell carcinoma. For detailed cell culture methodology please refer to the relevant section of chapter 2.

Human peripheral blood was obtained through venepuncture of healthy volunteers. The research had ethical approval by the Keele University FMHS Faculty Research Ethics Committee (MH-210190). 4 ml of blood was taken per volunteer and collected in EDTA containing tubes. The blood was taken immediately to the laboratory to be processed into samples.

Sample preparation

Cancer cells were removed from culture flasks as described above. Following centrifugation at 1200 rpm for 5 minutes, cells were resuspended in 0.9% saline. 50,000 cancer cells were pipetted into 1 ml of whole blood. A higher concentration of cancer cells than would be found physiologically was used to obtain samples with enough individual cancer cells to allow to collect enough data for a robust training and testing dataset for testing of the methodology and classification of the cancer cells from blood cells, and at the same time, not to obtain groups of cancer cells together in the final samples as the aim was to study individual cancer cells. The methodology was adapted to study single cancer cells in blood as the clinical application would to identify individual cancer cells. After the cancer cells were

added to the whole blood, red blood cells were removed by incubating the blood with Ammonium-Chloride-Potassium (ACK) lysing buffer (Thermo Fisher Scientific) for 5 minutes at room temperature at a concentration of 10 ml ACK buffer per 1 ml of blood. The blood was then centrifuged at 300 x g for 5 minutes at room temperature. The supernatant was disposed removing most of the red blood cells. The pellet containing leukocytes and cancer cells was resuspended with 5 ml of cold 0.9% saline. The remaining cells were centrifuged again at 300 x g for 5 minutes. The supernatant was removed, and the pellet was resuspended in 0.5 ml of 0.9% saline. The resuspended mixture of leukocytes and cancer cells were immediately used to prepare samples on glass coverslips with a cytospin.

The red cell depleted doped blood was deposited on glass coverslips using a cytospin at 900 rpm for 1 minute. The deposited cells were immediately fixed by incubating at room temperature for 15 minutes with 100 µl of 4% PFA. After fixation excess PFA was poured of the slips into a disposal container and then washed once with 0.9% saline and thrice with dH_20 to ensure all the PFA was removed. Samples were air dried to remove excess moisture for each cell line, 3 independent experiments were prepared, and for each independent experiment, 6 samples were prepared. Thus, 18 samples were prepared for each cell line. Each independent experiment corresponded to cells at different passage number.

FTIR Spectroscopy

FTIR spectra were obtained using a Thermo Nicolet iN10(MX) spectrometer. Developing this methodology into a clinical application will entail mapping areas of blood samples containing suspected CTCs. Thus, IR spectra of individual cancer cells were obtained by mapping an area

containing individual cancer cells or leukocytes. The maps were collected using an aperture of 15 x 15 µm. Spectra were measured with a 10 µm step size in the X axis and Y axis. This method would ensure that at least one 15x 15 µm aperture size spectrum will include only cancer cell. The size of each individual cancer cell for both cell lines is 20-30 µm diameter after the cytospin. These cancer cell lines were chosen to allow visual identification of the cancer cells for the purposes of the experiment. Individual spectra of cancer cells and leukocytes were also collected to build a training dataset. Spectra were collected at 4 cm^{-1} resolution, with 256 co-added scans. Background measurements were obtained under the same conditions from areas of coverslip without a biological sample.

Staining

To confirm the identity of the cancer cells in the prepared samples, a Giemsa stain was used. Giemsa stain is a differential stain containing a mixture of azure blue, methylene blue and eosin dye. Pathology laboratories commonly use Giemsa staining for blood work such as leukaemia and malaria diagnosis. A staining solution was prepared by diluting a stock Giemsa solution (Atom Scientific) (methanol <25%, glycerol <25%, ethylene glycol <25% and Giemsa powder) at a ratio of 1:40 with a Gurr buffer (Giemsa solution: Gurr buffer) (Thermo Fisher Scientific). The Gurr buffer was produced by adding the Gurr buffer tablet to 100 ml of distilled water as per the manufacturer instructions to produce a pH 6.8 phosphate buffer. The sample was covered with the Giemsa solution and incubated at room temperature for 45 minutes. The excess staining solution was then poured off the sample slips and remaining

excess stain was washed off with the buffer. Samples were air dried. The stained samples were imaged with microscopy to confirm the identity of the measured cancer cells.

Pre-processing and data analysis

Spectra were cropped to the regions 3500-1350 cm^{-1}, 3500-2700 cm^{-1} and 1800-1350 cm^{-1}. The spectra were denoised with PCA denoising with 10 components and a Savitzky-Golay filter with a window size of 5 and polynomial of 2. EMSC was used to normalise the spectra and remove any baseline defects. The average spectra of the training set were used as a reference for the EMSC. The pre-processing was carried out using the Quasar software.

The spectra in the maps were annotated as A549/CALU-1, leukocytes, or background. Annotations were done using the stained and non-stained images of the mapped area. The larger size of the A549 and CALU-1 cells allowed them to be visually identified. Larger cancer cell lines were chosen for this study so it could be assessed if the cancer cells are being correctly identified within the maps by the classifier.

A RF classifier was used to classify each 10 µm tile based on spectral data. The RF classifier contained 200 decision trees. A further increase in the number of trees did not significantly improve the classification. The classifier was trained using a training data set consisting of spectra from A549/CALU-1, leukocytes, and background measurements. The number of spectra of cancer cells and leukocytes in the training set was balanced to 300 of each. The maps that comprised the test data were different maps used for training.

Each tile in the maps were coloured by the classifier based on the probability of the tile containing spectrum from A549/CALU-1. The tiles were coloured on a colour scale from yellow being of high probability (>0.9) to dark blue being low probability (<0.2), the colour scale is shown in the Figure 40 below. Output from the RF classifier including the AUC, CA, F1, precision and recall which was measured against if the spectra were correctly classified against the annotations.

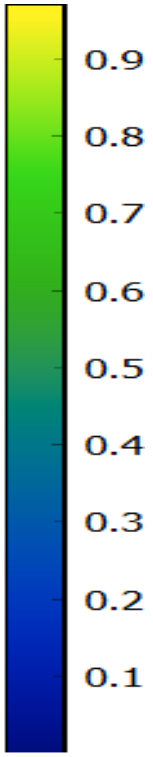

Figure 40 Colour scale for random forest classification of maps.

Results

The first step was to obtain FTIR spectra from individual cancer cells in both cell lines and clusters of leukocytes. The spectra were obtained from the sample prior to staining. Despite the lack of staining, cancer cells were easy to identify due to their larger size in comparison

to leukocytes. All samples were stained with a Giemsa stain to confirm the identity of the cancer cells. Figures 41 and 42 demonstrates the larger morphology of the cancer cells in comparison to leukocytes and darker purple colour from the stain. All the cells thought to be cancer cells prior to staining were confirmed as cancer cells following saining. Figures 43-47 shows the average FTIR spectra of the leukocytes compared to A549 and CALU-1. The amide I, amide II and amide A bands were less intense in the leukocyte spectrum than the A549 and CALU-1 spectrum which infers greater protein content in the cancer cells. The amide I of the leukocyte spectrum is also shifted to 1650 cm-^{1} compared to A549 and CALU-1 positioned at 1653 cm^{-1}. In the region 2800-3000 cm^{-1} stretching vibrations of from CH_2 and CH_3 groups in lipids were of a higher intensity in the leukocyte spectrum in the bands at 2922 cm^{-1} and 2851 cm^{-1} while at 2956 cm^{-1} has a lower intensity. Both the A549 and CALU-1 average spectra show similar differences to the average leukocyte spectrum.

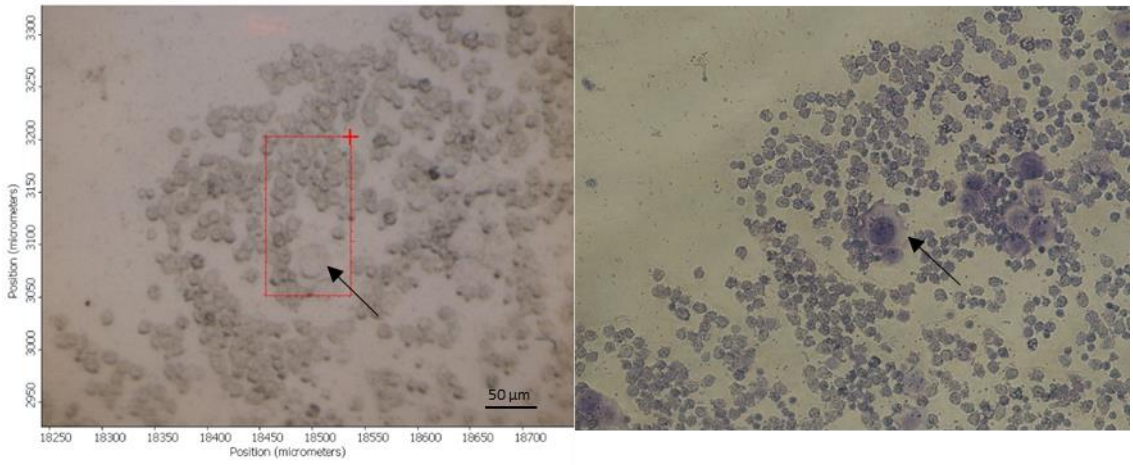

Figure 41 Image of a stained and unstained mapped area containing A549 cells and leukocytes. The arrows point to an A549 cell. The A549 cells can be identified from their larger size and deep purple colour in the stained image.

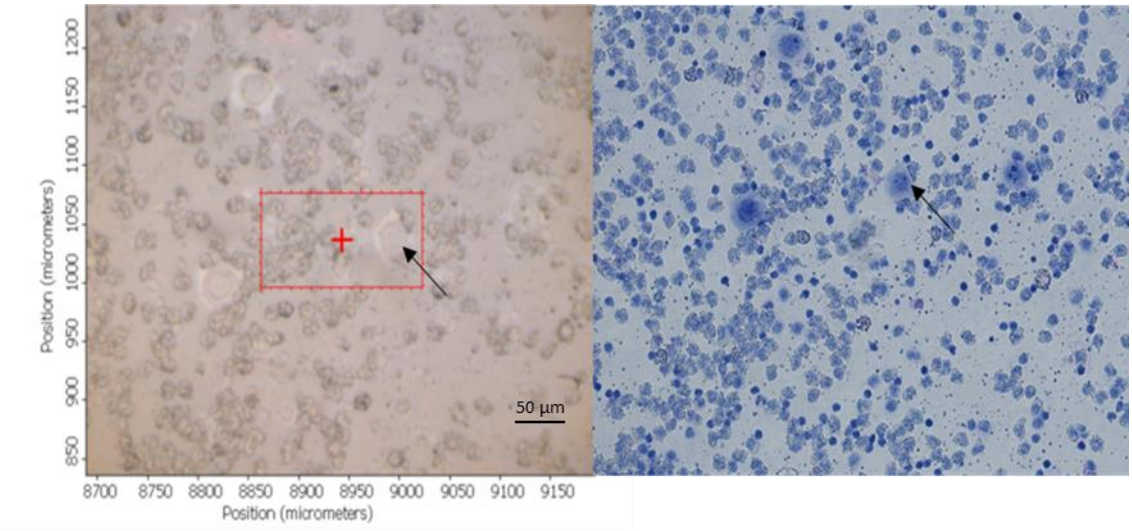

Figure 42 Image of a stained and unstained mapped area containing CALU-1 cells and leukocytes. The arrows point to a CALU-1 cell. The CALU-1 cells can be identified from their larger size and deep purple colour in the stained image.

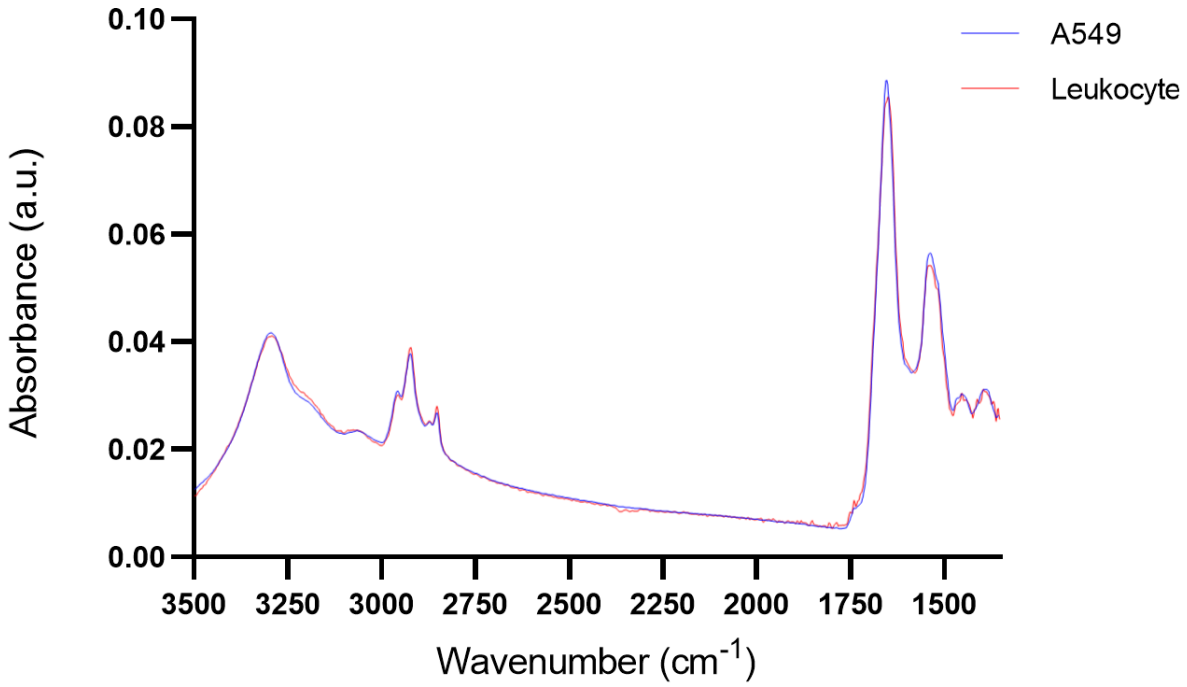

Figure 43 Average spectra of A549 cells and leukocytes in the region 3500-1350 cm^{-1}.

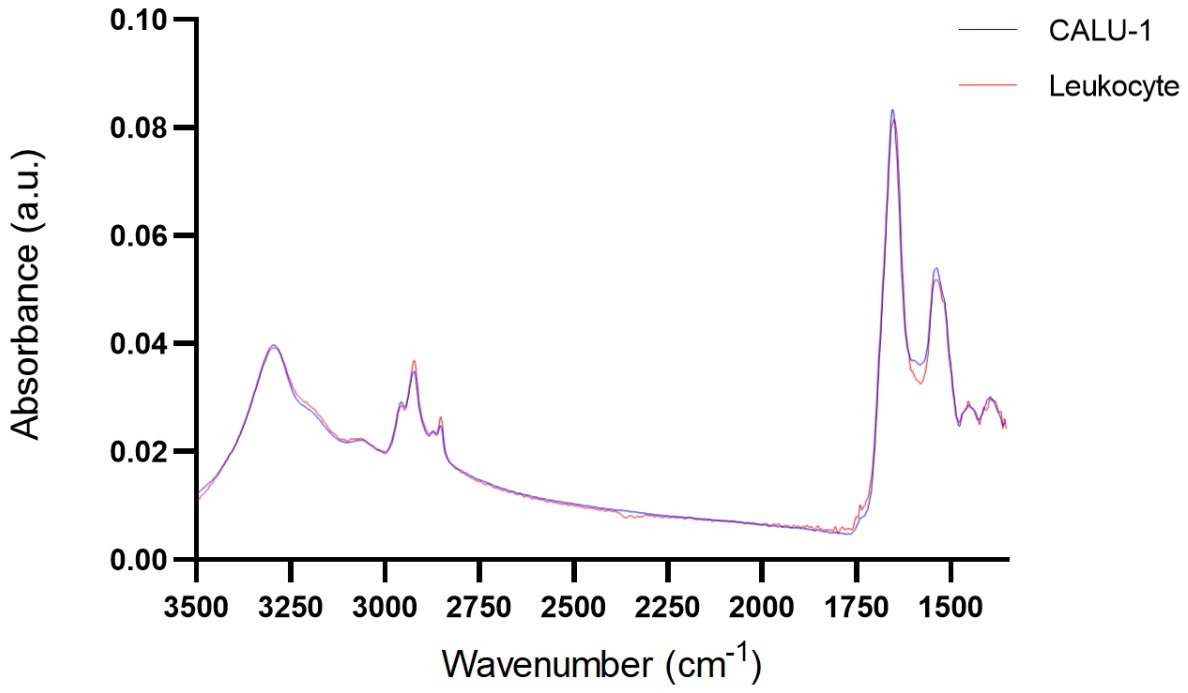

Figure 44 Average spectra of CALU-1 cells and leukocytes in the region 3500-1350 cm^{-1}.

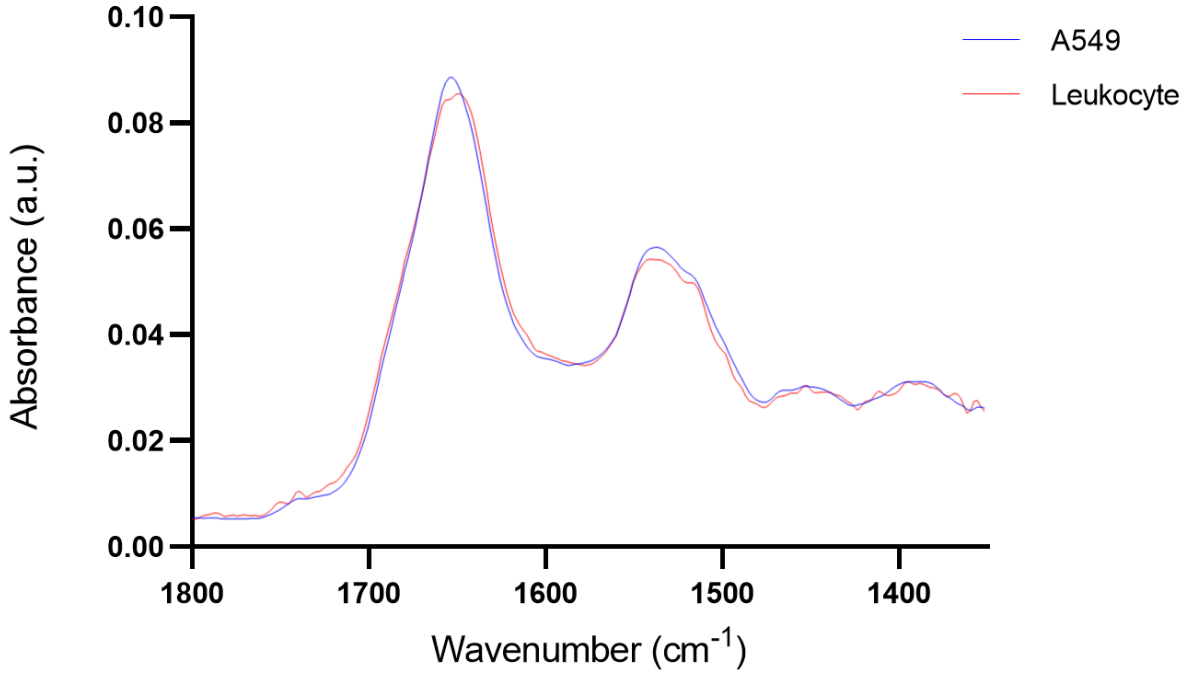

Figure 45 Average spectra of A549 cells and leukocytes I the region 1800-1350 cm^{-1}.

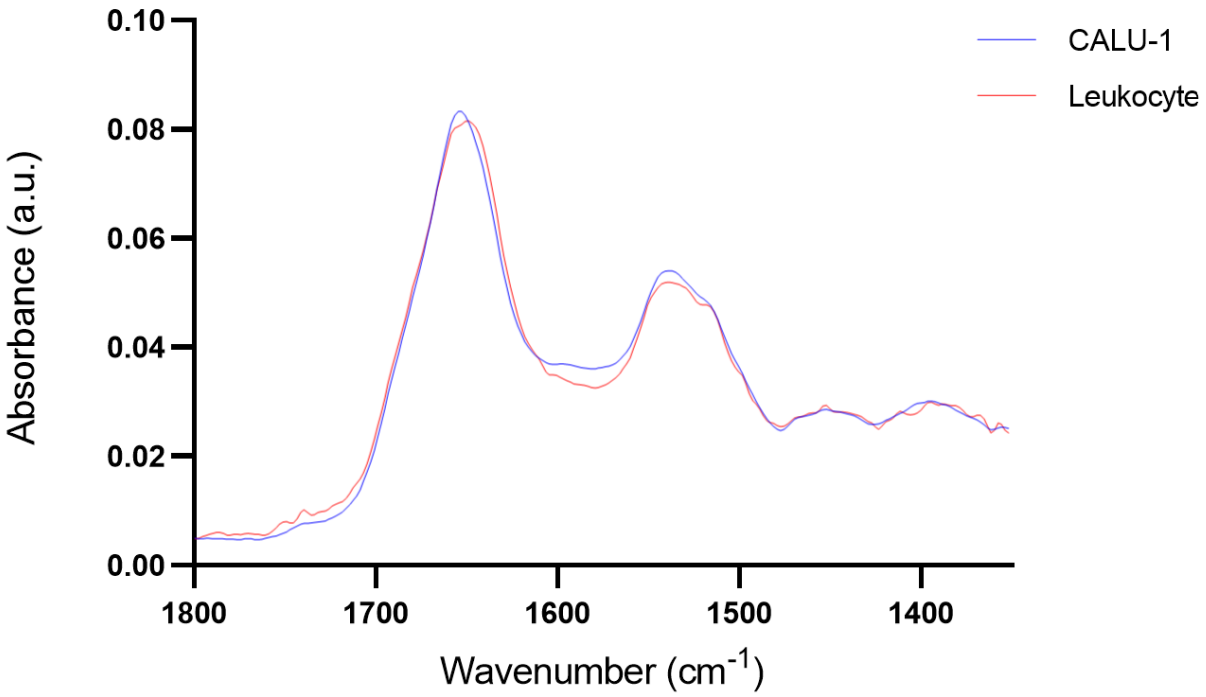

Figure 46 Average spectra of CALU-1 cells and leukocytes in the region 1800-1350 cm^{-1}.

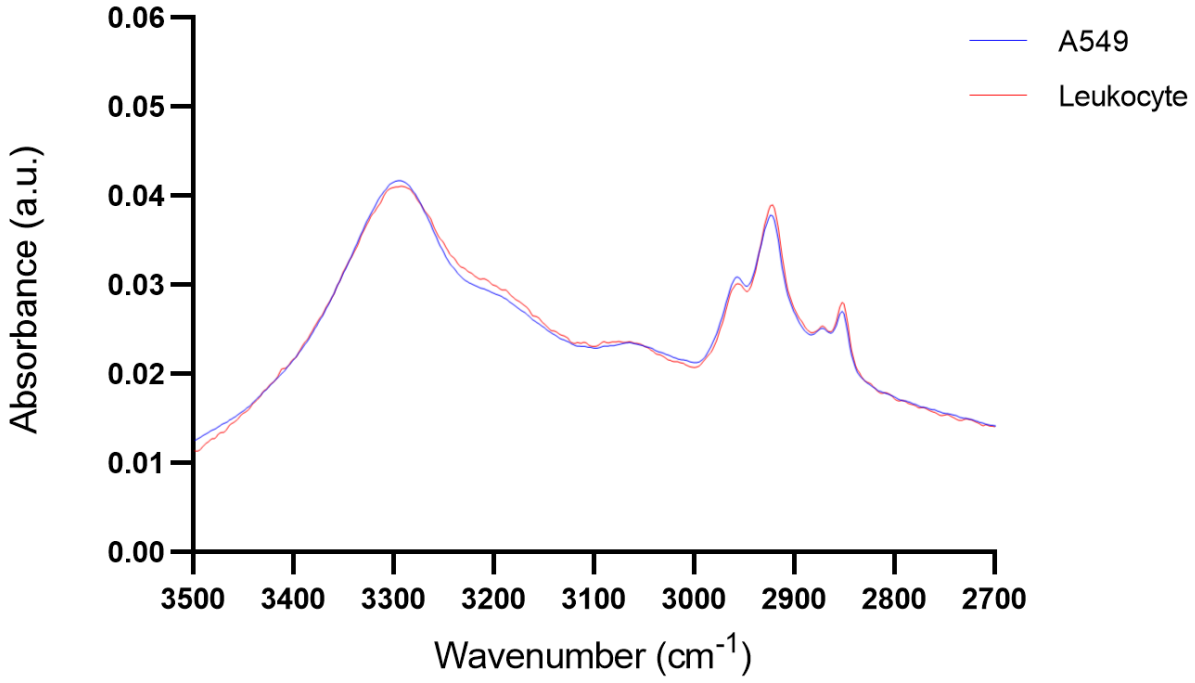

Figure 47 Average spectra of A549 cells and leukocytes in the region 3500-2700 cm^{-1}.

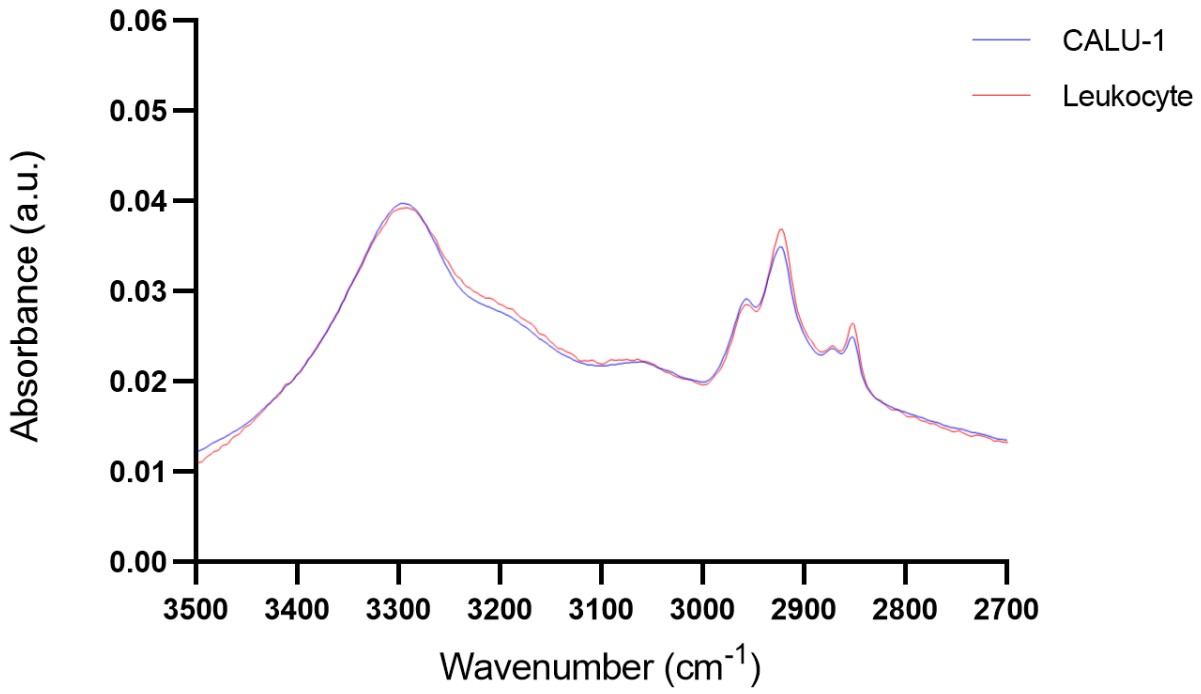

Figure 48 Average spectra of CALU-1 cells and leukocytes in the region 3500-2700 cm^{-1}.

The application of FTIR spectroscopy for cancer cell identification in liquid biopsies will entail using a machine learning to classify cells by their FTIR spectra which show the biochemical differences between cells. A RF classifier was chosen for this research because it handles data with many features well and is less prone to overfitting. 8 maps containing A549 (Figures 49-56) and 8 containing CALU-1 were classified (Figures 57-64). In all 16 maps, areas containing cancer cells were identified by the RF classifier. The tiles on the maps coloured yellow and green were given a high probability of the spectra being from a cancer cell while blue tiles had a low probability. Comparing the coloured maps to the visual images, the areas coloured as likely containing cancer correlated to the location of the cancer cells. Using the spectral region 3500-2700 cm^{-1} the maps were coloured more accurately with less tiles falsely coloured to contain spectra from cancer cells. Using the regions 3500-1350 cm^{-1} and

1800-1350 cm^{-1}, the classifier also identified the cancer cells but had more tiles misclassified as likely containing cancer cells in locations with no cancer cells. Comparing the misclassified tiles to images of the maps, the tiles misclassified as cancer not near a cancer cell largely correspond to areas where there are large numbers of leukocytes clumped together. The misclassifications of tiles in the areas close to the cancer cells were because of the aperture used. The tiles were 10 μm and the aperture size was 15 μm therefore tiles adjacent to the cancer cells contained spectral measurements from the cancer cells.

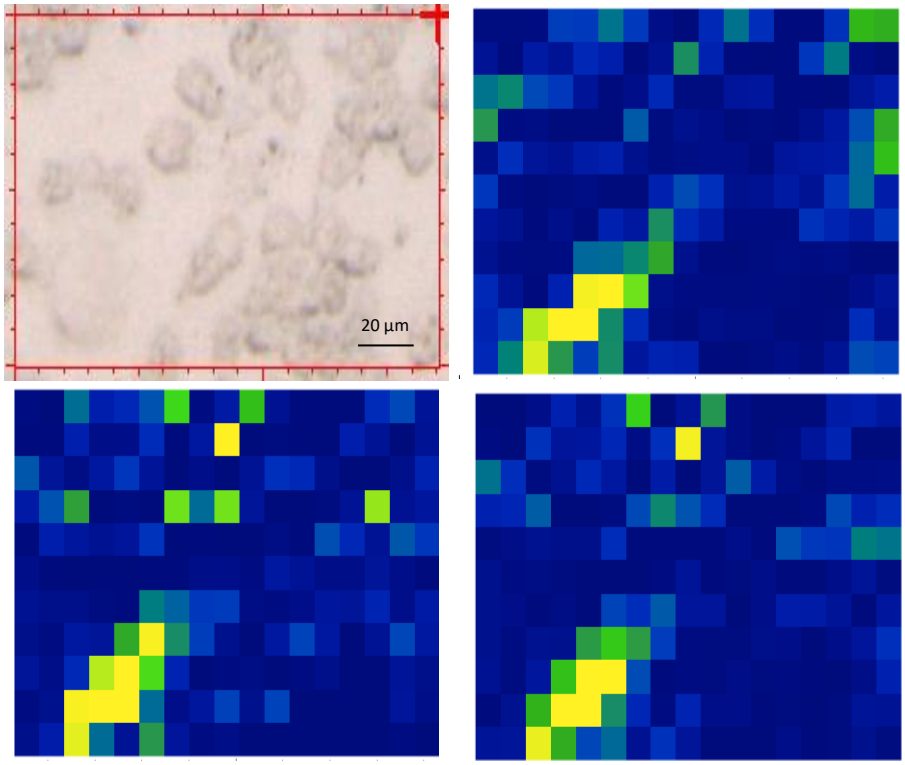

Figure 49 False colour maps coloured by RF classifier based on probability of A549 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

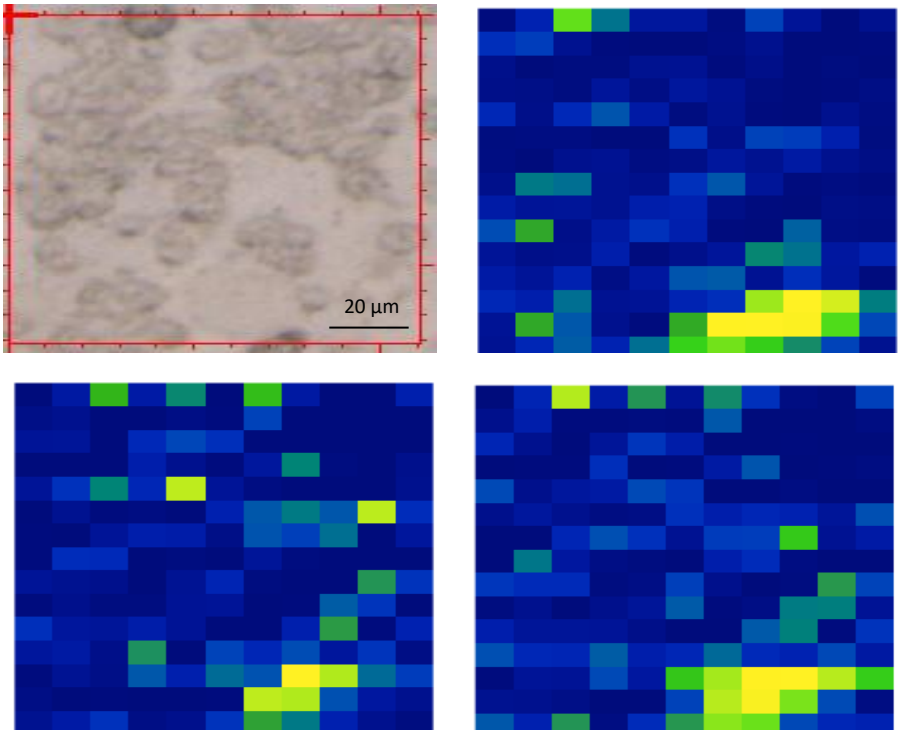

Figure 50 False colour maps coloured by RF classifier based on probability of A549 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

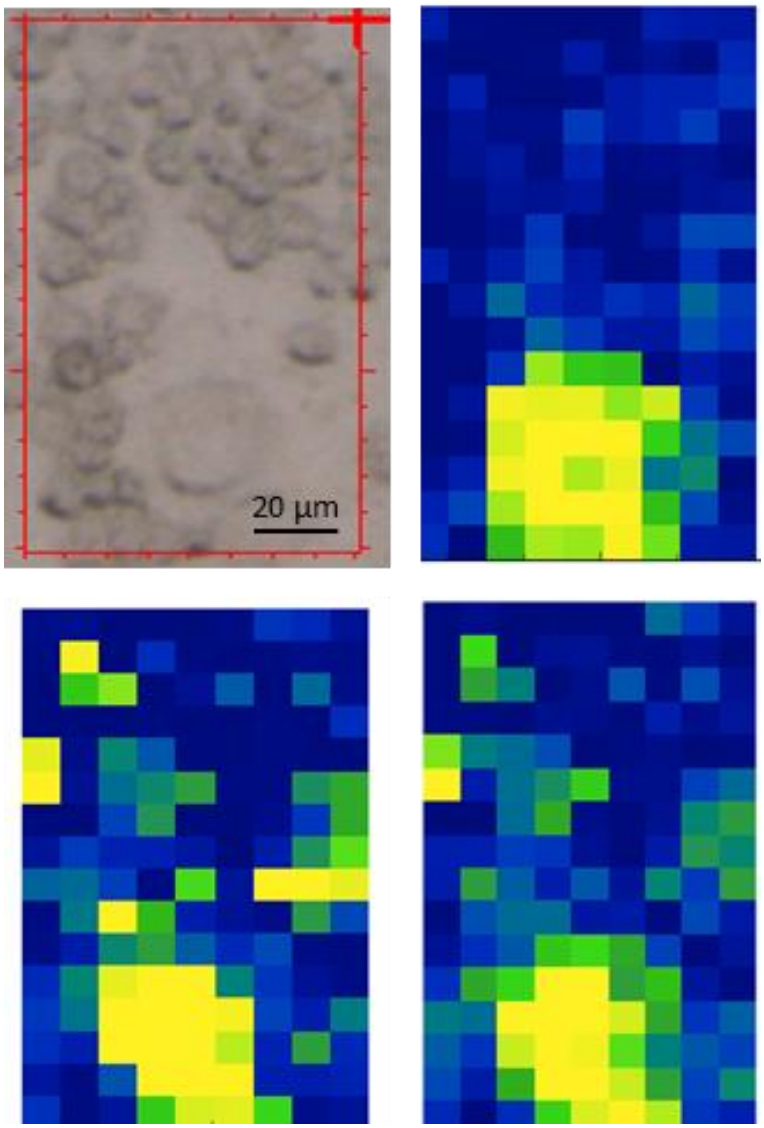

Figure 51 False colour maps coloured by RF classifier based on probability of A549 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

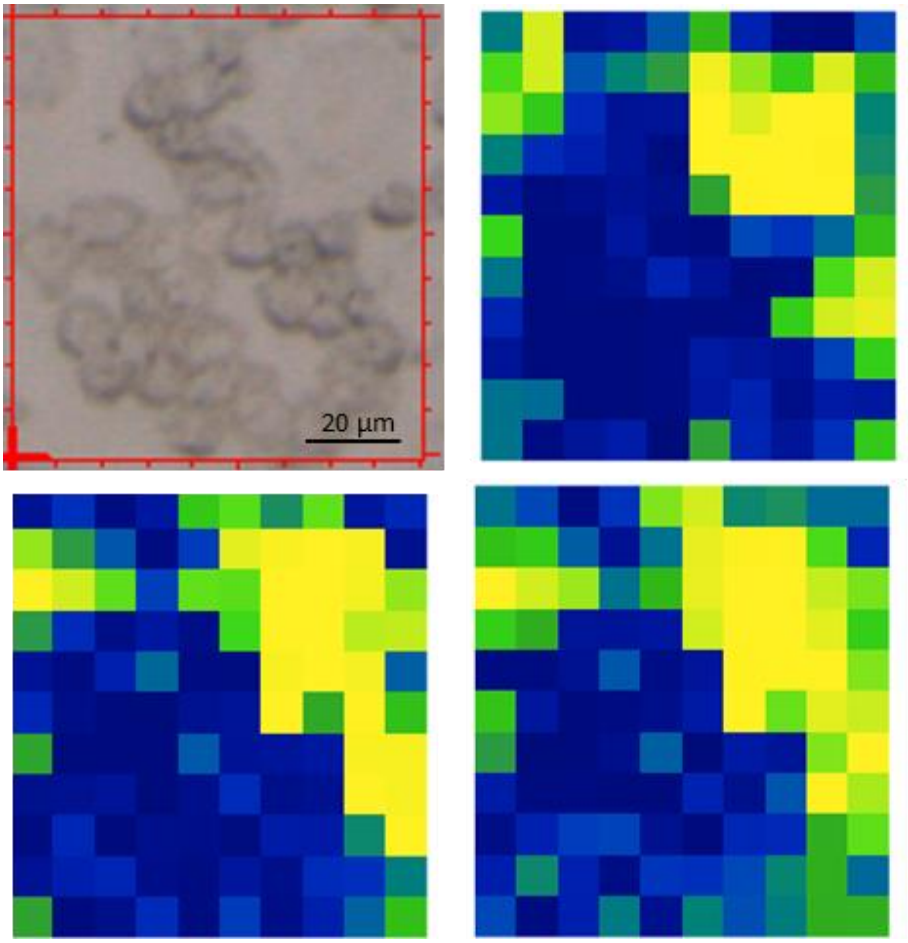

Figure 52 False colour maps coloured by RF classifier based on probability of A549 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

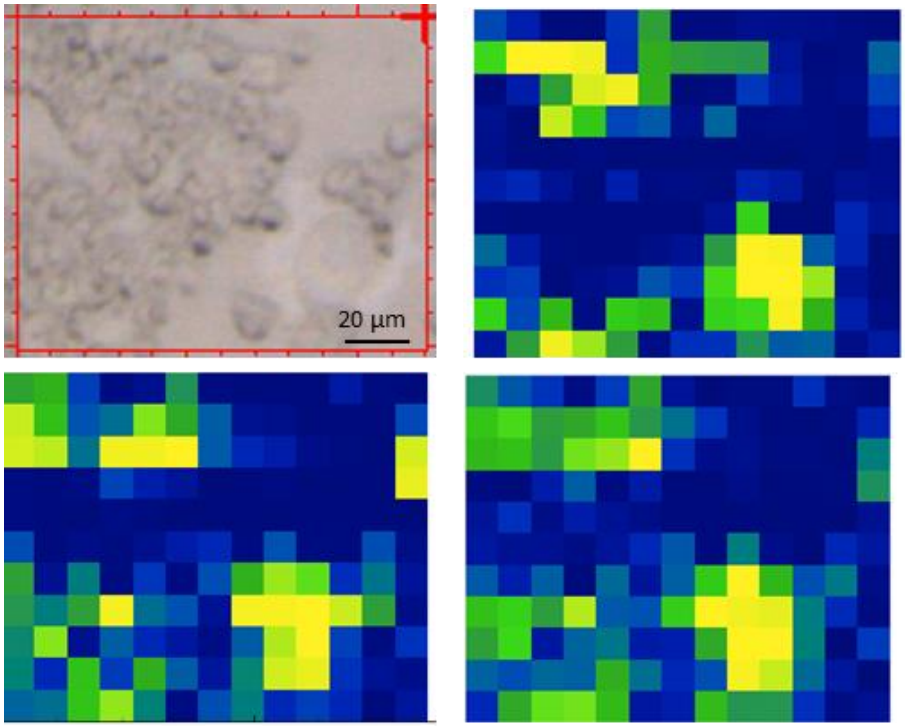

Figure 53 False colour maps coloured by RF classifier based on probability of A549 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

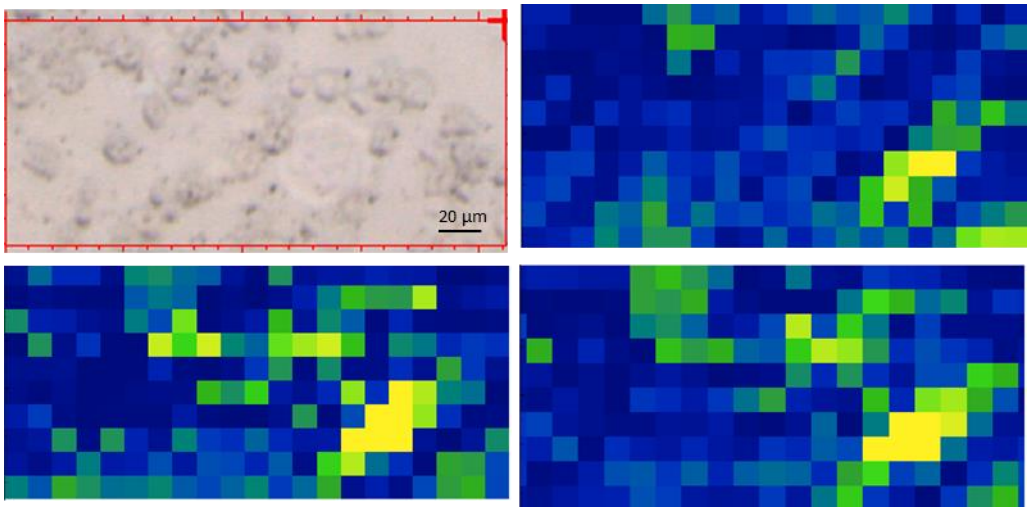

Figure 54 False colour maps coloured by RF classifier based on probability of A549 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

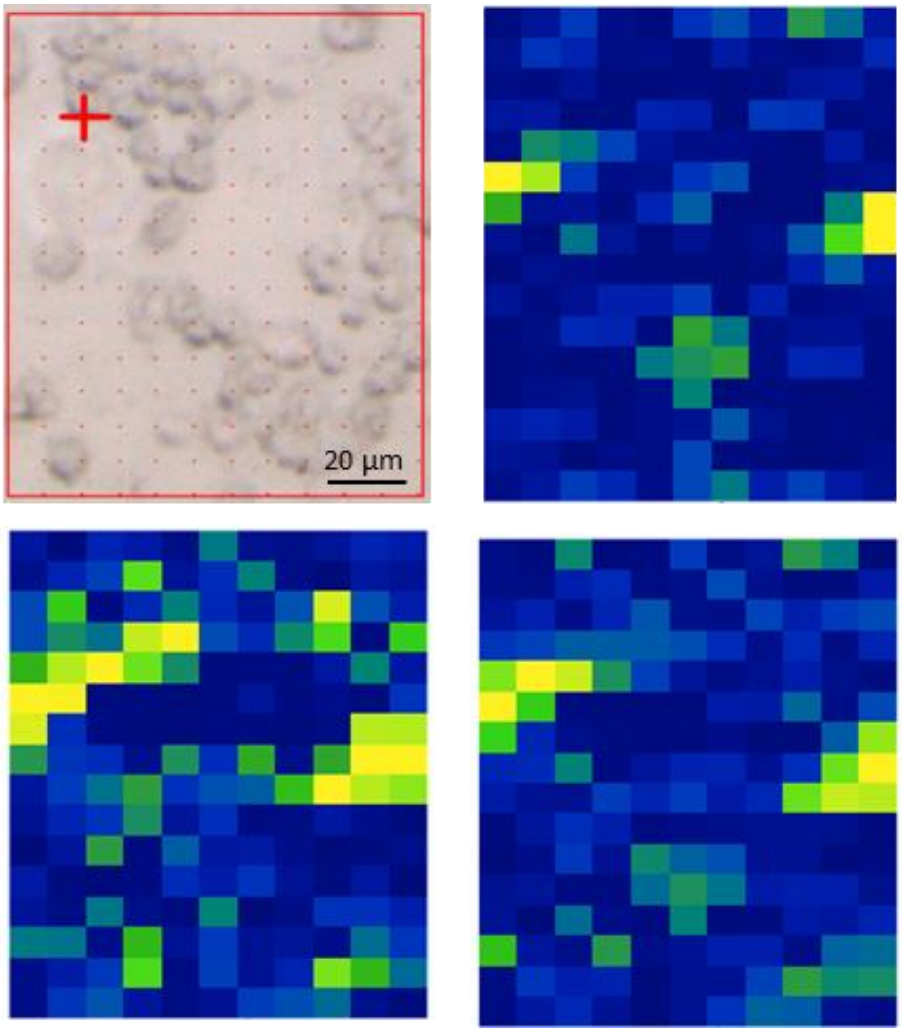

Figure 55 False colour maps coloured by RF classifier based on probability of A549 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

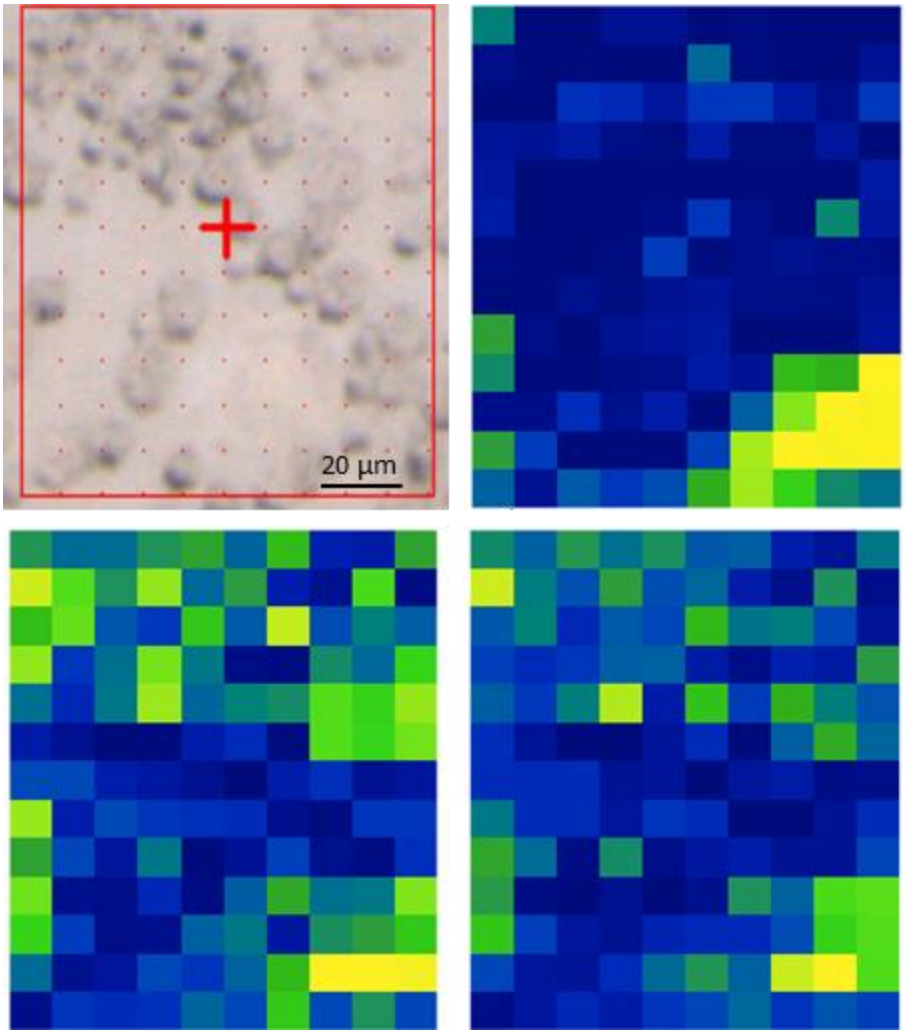

Figure 56 False colour maps coloured by RF classifier based on probability of A549 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

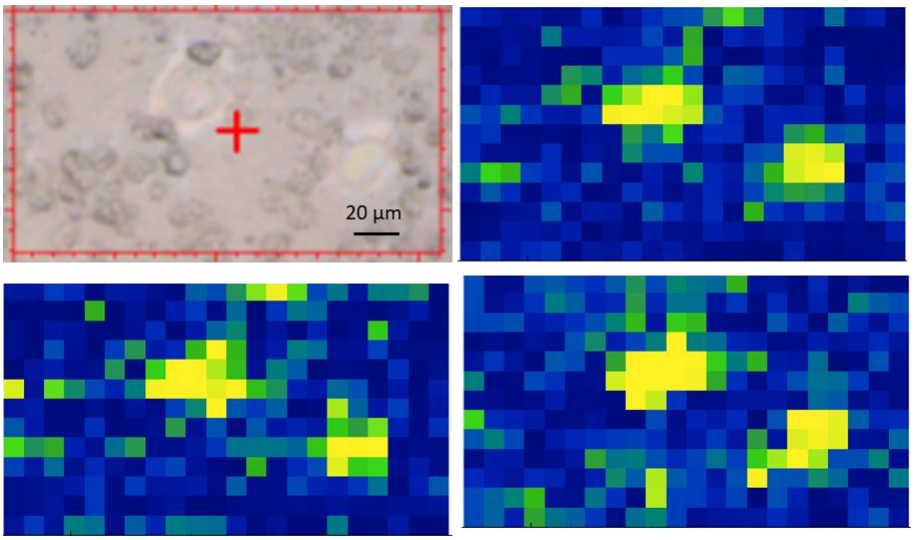

Figure 57 False colour maps coloured by RF classifier based on probability of CALU-1 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

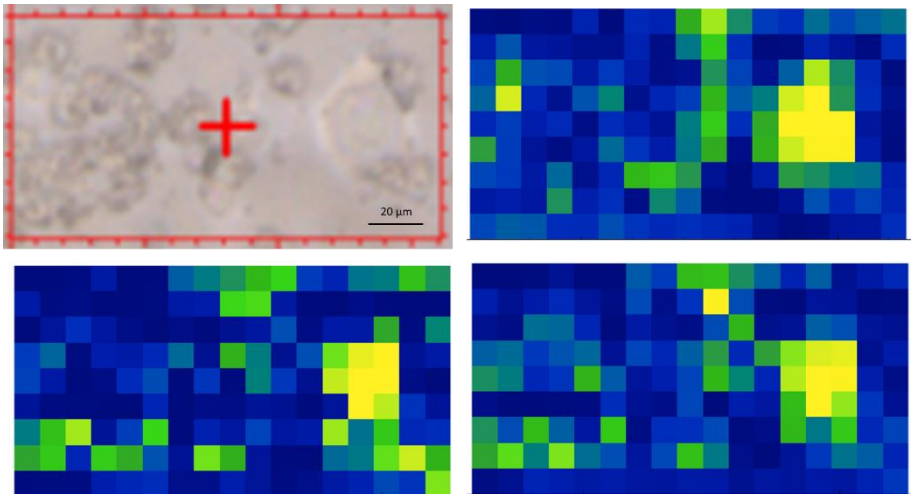

Figure 58 False colour maps coloured by RF classifier based on probability of CALU-1 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

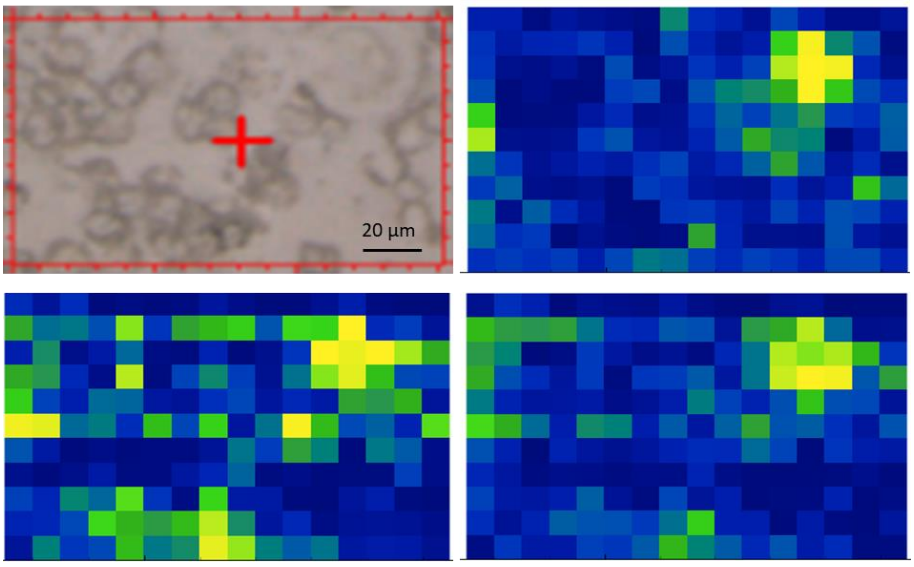

Figure 59 False colour maps coloured by RF classifier based on probability of CALU-1 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured

using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

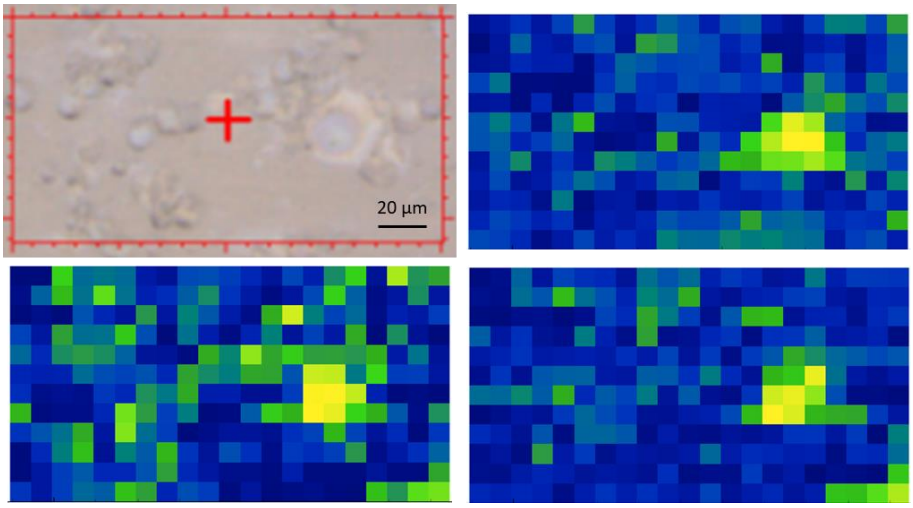

Figure 60 False colour maps coloured by RF classifier based on probability of CALU-1 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

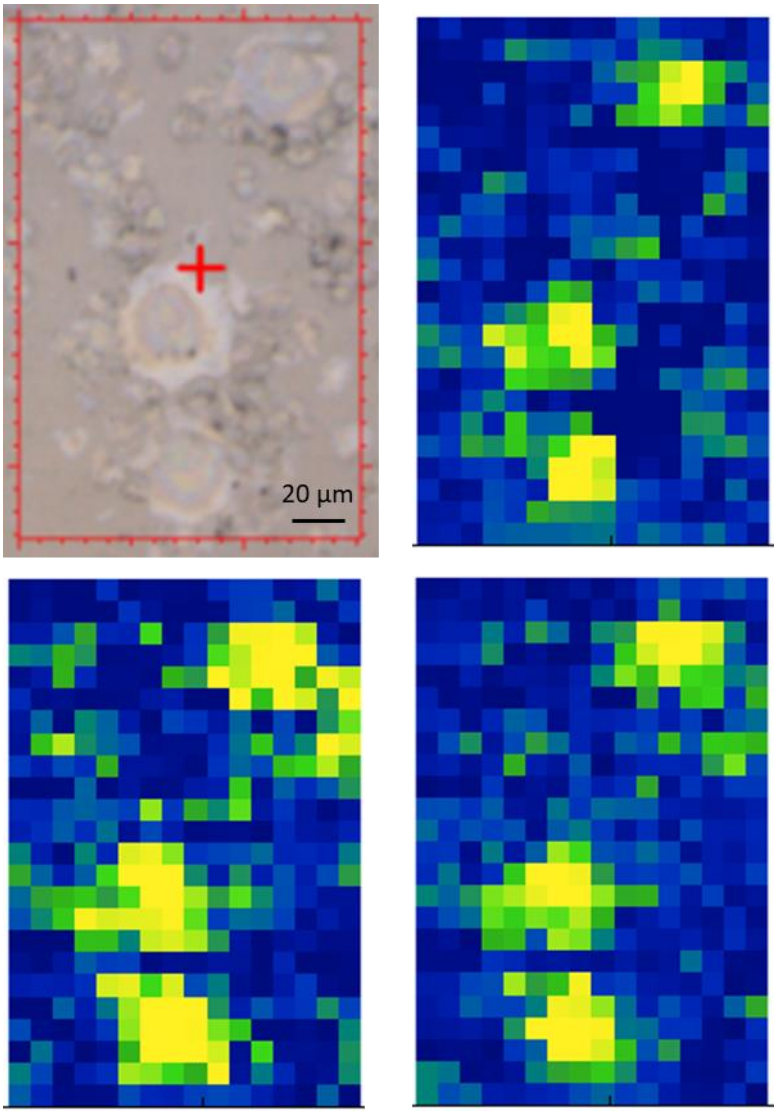

Figure 61 False colour maps coloured by RF classifier based on probability of CALU-1 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

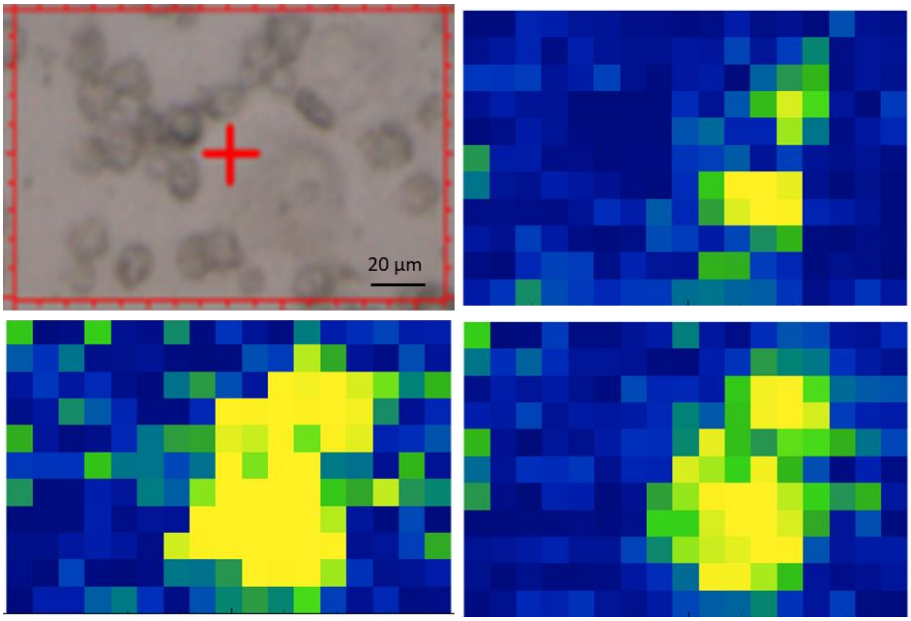

Figure 62 False colour maps coloured by RF classifier based on probability of CALU-1 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

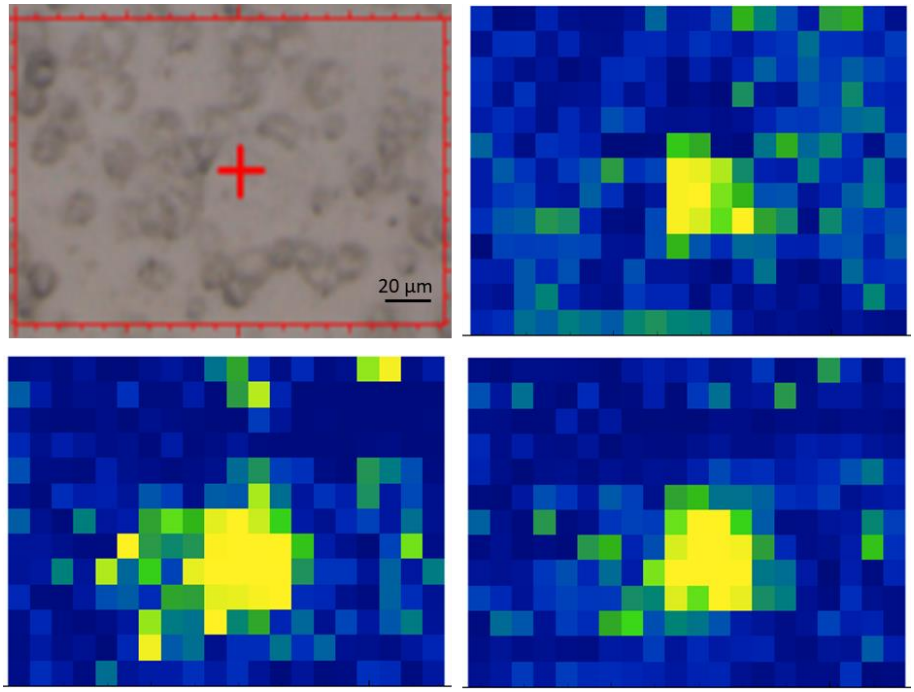

Figure 63 False colour maps coloured by RF classifier based on probability of CALU-1 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

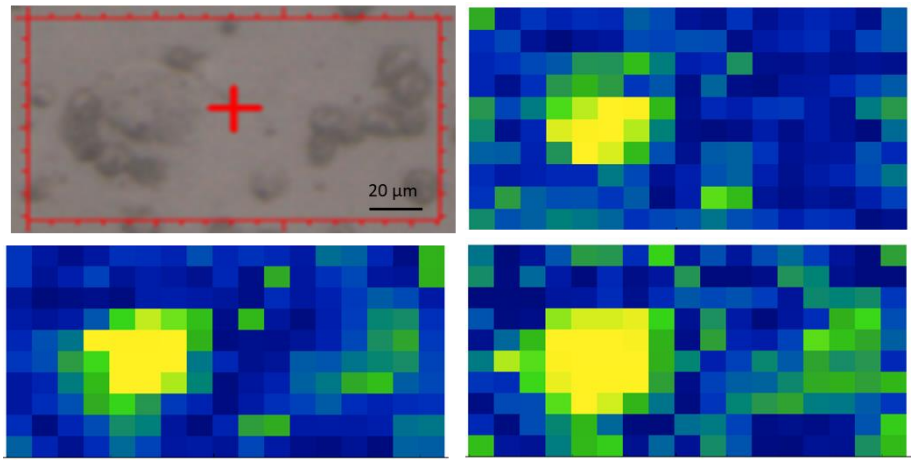

Figure 64 False colour maps coloured by RF classifier based on probability of CALU-1 cell in each tile. Top left: Microscope image of mapped area. Top right: Hyperspectral map coloured

using FTIR spectra region 3500-1750 cm^{-1}. Bottom left: Hyperspectral map coloured using FTIR spectra region 1800-1350 cm^{-1}. Bottom right: Hyperspectral map coloured using FTIR spectra region 3500-1350 cm^{-1}.

Table 12 below provides the average classification results from the maps using the best classification result from each map. These classification results were calculated from the classification of each tile in the maps by the RF classifier compared to the annotations of each tile. Each tile was annotated based on the visual. The maps containing A549 cells had a stronger average classification result than the CALU-1. As mentioned above, the misclassification of areas having cancer cells were often in areas densely populated by leukocytes or close to the edge of the cancer cells. Many misclassifications were from leukocytes as background and vice versa. These misclassifications often occurred at the edge of leucokcytes where the spectral signal was weak, so it was misclassified as background or where there were no cells present but there was a spectral signal from nearby cells due to the aperture size used. Overall, the results indicated that FTIR spectroscopy used with a glass substrate is a capable tool for identification of individual lung cancer cells in blood.

Cell line	AUC	Classification accuracy	F1	Precision	Recall
A549	0.885	0.774	0.788	0.820	0.768
CALU-1	0.897	0.727	0.735	0.804	0.718

Table 12 Average classification results from classification of cancer cell and leukocyte maps using the best classification for each map.

Discussion

Over recent years the management of cancer has been moving towards a more personalised treatment system (Hoeben et al., 2021). The appearance of new immunotherapies, targeted therapies and better tumour subtyping through genetic testing is tailoring the treatment of cancer to each individual patient and tumour. Liquid biopsies have the potential to enhance personalised cancer management by allowing more frequent testing and monitoring of the disease in a non-invasive manner. In the case of CTC identification, it could provide more diagnoses at early stages of the disease and characterisation of the cancer with less need for invasive surgeries (Yang et al., 2019). CTC dissemination is thought to start in early stages of cancer progression and if these cells can be identified it could allow diagnosis of cancer earlier than would be detected with imaging and biopsies of the tumour. Early diagnosis of lung cancer is currently difficult with <30% of cases being diagnosed in stage I or II of the disease. This is one of the key reasons why lung cancer survival remains low with only around 16% of patients surviving their disease for 5 years (Cancer Research UK, 2018). Another advantage of liquid biopsies is that they allow more frequent monitoring of the cancer than is possible from biopsies of the tumour. A liquid biopsy from blood only requires a blood and not an operation. A difficulty of current biopsy methods is a poor-quality biopsy can make diagnosis difficult and another biopsy is not always possible because the tumour is difficult to access. The continued presence of CTCs post-treatment could indicate that the cancer has not been fully eradicated. The biggest challenge in using CTC for diagnostics thus far is their scarcity which makes their isolation and identification difficult. The only CTC

identification method currently approved for clinical use is CellSearch which identifies CTCs through epithelial markers. Despite the CellSearch being available since 2004 there is no widespread use of CTC cancer diagnosis. New methods of CTC identification that are needed that are simple and inexpensive to use CTCs for cancer diagnostics. A reason it has not been widely used is that it is unable to detect CTCs that express EpCAM which is further complicated by the EMT downregulating epithelial marker expression. For CTC identification to become more widely used for cancer diagnostics, new methods are needed that are ideally label free, accessible, and affordable. FTIR spectroscopy is a technique yet to be investigated in the literature for its potential to aid in CTC identification. This research is some of the first to test the feasibility of using FTIR spectroscopy to identify lung cancer cells in blood. As discussed previously in this thesis FTIR spectroscopy is an attractive option because of its label-free and non-destructive nature while being relatively simple to use.

It is clear the biochemical profile of CTCs in both their epithelial and mesenchymal phase will be very different to that of blood cells. It can be hypothesised that exploiting these biochemical differences would lead to systems that could identify CTCs. Most research in the area of FTIR spectroscopy for cancer diagnostics has been aimed at differentiating cancerous cells or tissues from their non-malignant counterparts. In some instances, the malignant cells or tissues can be quite like their non-malignant counterparts. Shown in previous chapters of this thesis is the classification of different types of lung and breast cancer cells from each other. Whereas there is very little similarities between blood cells and CTCs. Most liquid biopsy research thus far utilising FTIR spectroscopy has focused on measuring the biofluid itself for chemical changes such as the serum, urine, or saliva. There is currently no literature at the time of writing as far I am aware of the use of FTIR spectroscopy for CTC identification.

The average spectra of lung cancer cells and leukocytes demonstrated the biochemical differences that make the basis of using FTIR spectroscopy for CTC identification. CTCs originate from solid tumours and often from epithelial cells which have a different function to the blood cells, therefore, the phenotype and biochemistry will be markedly different. The spectra indicated differences in both the proteins and lipid content of the cancer cells and the leukocytes. These differences were shown by the absorbances across all the bands, differing band position (amide I) and shape (amide II). With these biochemical differences cancer cells could be identified from blood cells without the need for specific cell markers. The EMT transition provides difficulties in using epithelial markers for CTC identification because their expression is often reduced (Andree et al., 2016). This is drawback of currently approved CTC identification method that label CTCs with a fluorescent tag conjugated to antibodies to attach to EpCAM on the CTCs. Therefore, to identify CTCs by surface markers multiple markers must be targeted which becomes a complex process to ensure no off-target attachment. Selecting the markers to target is difficult for the initial diagnosis because it is not known which surface antigens the CTC express and can cause the CTCs going undetected if marker selection is flawed.

This study demonstrates FTIR spectroscopy can be used to identify lung cancer cells from leukocytes in a mixed sample placed upon a glass substrate. This was possible using both the region 1800-1350 cm^{-1} with the amide I and amide II bands and the 3500-2700 cm^{-1} region containing bands from the CH_2 stretching in lipids and amide A. The 3500-2700 cm^{-1} region provided a better identification of the cancer cells in most of the maps for A549 and CALU-1 with less tiles coloured wrongly by the RF classifier as containing cancer cells. In all 16 maps the lung cancer cells could be identified including multiple cancer cells in some maps. The average precision for the classification of the maps was >80% for both A549 and CALU-1. The

high precision demonstrates the classifier was able to identify a majority of the tiles that contained a cancer cell. Only a small number of maps were tested and only with two NSCLC cell lines. Measurements and training of the classifier for each type of cancer would need to be before the methodology can be applied to different cancers. The false colour maps utilising different spectral regions can be produced simultaneously to provide a compiled view of classifications from using different regions. This is helpful because different types of cancer may be classified better using a different spectral region than the cells measured for this research.

Many of the tiles misclassified as cancer were in areas densely populated by leukocytes. If this methodology was to be developed further, the cancer cells would first have to be isolated due to their low number compared to blood cells. This would also help reduce the misclassifications from densely packed areas of leukocytes. There have been many different methods of CTC isolation suggested which can be broadly separated into biological methods and physical methods. Biological methods rely on the use of antibodies binding to antigens on the cells (Bankó et al., 2019). These antibodies can be conjugated to a magnetic bead. Beads attached to cells are retained when a magnetic field is applied while the other cells are allowed to flow through into a collection receptacle. If using biological isolation with FTIR spectroscopy for CTC identification, negative selection by depletion of blood cells may be the best option. Positive selection as discussed can be difficult due to the EMT and tumour subpopulations not containing the target antigen. While positive selection can produce a purer yield of CTCs many can be lost. The lost CTCs could be of diagnostic relevance for typing the tumour and deciding a treatment plan. More blood cells will be present after negative selection but as demonstrated, the cancer cells can be identified in samples surrounded by leukocytes because of the biochemical and morphological differences.

Leukocytes can be depleted by targeting CD45 which is not present on CTCs from solid tumours. The depletion of the leukocytes would help to reduce the misclassifications' as densely populated areas are reduced. Leukocyte dense areas are misclassified because of the higher absorbance from multiple cells causing the classifier to mistake the area of containing cancer cells which have a higher average absorbance than leukocytes. Physical methods separate cells based on size, density, and deformability. The use microfluidic devices that are fabricated to capture CTCs based on their physical properties differing from blood cells is an area with growing research (Tan et al 2009) (Payne et al 2021). The main limitation of physical isolation methods is they cannot isolate CTCs with similar physical properties to the blood cells such as small cell lung cancer or small cell prostate cancer. Generally physical isolation methods lack specificity and have poor purity. Future work to develop CTC identification with FTIR spectroscopy will have to test it with isolation methods.

The use of the glass substrate allows for further characterisation of the cancer cells after identification with FTIR spectroscopy. The current standard pathology techniques such as staining, or immunohistochemistry can be used because the FTIR spectroscopy is label free and non-destructive. This was demonstrated by the Giemsa stain of the samples to confirm the identity of the cancer cells. Further characterisation of the cells is important to confirm tumour type and subtypes and the presence of surface receptors important for selecting the best treatment plan.

This is some of the first research to investigate how FTIR spectroscopy could be used for CTC identification. There is a larger body of research for the use of FTIR spectroscopy for other liquid biopsy methods, to diagnose cancer using the blood serum (Baker et al., 2022). These methods do not use any specific tumour material like CTCs but use the chemical changes in

the serum itself. The advantage of using FTIR spectroscopy with liquid biopsies of blood is that both the CTCs and serum could be analysed using the same instrument and blood sample. Liquid biopsy analysis with FTIR spectroscopy could be used together to provide the clinicians with a wider range of information without the need for invasive surgeries. While the study of serum requires less sample preparation than CTC analysis, it does not allow for further analysis of tumour related materials. It is also not fully known how or what changes different cancers make to the serum and the biological changes involved that cause the spectral differences in the serum of cancer patients and healthy patients and what changes different cancers cause. Other diseases could also cause changes to the biochemistry of the serum and in result the FTIR spectra changing compared to serum from a healthy patient. A majority of cancer patients are elderly with co-morbidities which could contribute to changes in serum spectra. A combined approach using both serum measurements and CTC measurements could make use of the benefits of both. The serum measurements could provide a more global overview and fast initial information on the presence of cancer while the CTC analysis would provide more specific information on the nature of the cancer. As shown in previous chapters different types of cancer can be classified from the spectra of cancer cells. The identification of the CTCs would allow the further analysis with immunohistochemistry and genetic profiling to fully characterise the cancer.

While there is no literature for the use of FTIR spectroscopy for CTC identification there has been research on the use of Raman spectroscopy, another vibrational technique that can also measure biochemical differences in the cells. Kaminska et al used surfaced enhanced Raman spectroscopy (SERS) to distinguish HeLa cell and a prostate cancer cell line from leukocytes (Kamińska et al., 2019). They used PCA scores to demonstrate how the cells could be distinguished using the spectroscopy but did not employ any methods of classification.

For diagnostic use of spectroscopy for CTC identification classification will have to be used with classifiers such as RF. They combined the SERS with an isolation method using a microfabricated membrane to filter the larger cancer cells from the leukocytes. Physical isolation methods can be used to isolate many types of cancer cells from leukocytes because of the larger size of cancer cells. Size based isolation would not be applicable to small cell carcinomas where the cancer cells are not larger than the leukocytes. Wu et al also used SERS to identify HeLa cells and the liver cancer line HepG2 (Wu et al., 2015). They used a gold nanoparticle bio-probe conjugated to a folic acid ligand that is recognised by the cancer cells because of overexpressed folate receptor alpha. The nanoparticles enhance the signal from the attached cancer cells with the signal increasing with the number of cancer cells. This use of SERS shows its potential for prognostic use because CTC numbers have been shown to have prognostic value. However, the method can only be used for cancer cells that over express folate receptor alpha and does not prepare and isolate cells for further analysis. This other research demonstrates how vibrational spectroscopy can be utilised in differing ways for CTC detection. These methods do not allow as easily further analysis of the CTCs because the substrates used to achieve the surface enhancement such as the microfabricated membrane and conjugated gold nanoparticles are not widely available and simple to manufacture substrates which could make adoption of the SERS techniques difficult on a widespread scale across large healthcare systems.

Conclusions

This work was a proof of concept to demonstrate that FTIR imaging can be used to identify lung cancer cells of lung adenocarcinoma and squamous cell carcinoma from leukocytes. The

model shows how individual cancer cells can be identified from leukocytes with a RF classifier based on the spectral data measuring biochemical differences in the cells. This was shown to be possible using a glass substrate. The choice of substrate allowed for Giemsa staining of the samples as would be used for current diagnostic methods for blood samples. Now it has been shown to be possible to identify cancer cells from blood on glass substrates with FTIR spectroscopy, future stages of this research will look to test methods to improve the classification of CTCs using FTIR spectra by isolating the cancer cells. This research is a first step in creating a methodology for the use of FTIR spectroscopy to detect CTCs.

Chapter 7: Optical Photothermal Infrared Spectroscopy to study lung cancer on glass substrates.

Introduction

In previous chapters I demonstrated how FTIR spectroscopy can be used to classify lung cancer and breast cancer cells from non-malignant lung and breast cells on glass substrates. In this chapter, I investigated another IR spectroscopy technique similar to FTIR spectroscopy called Optical photothermal infrared (O-PTIR) spectroscopy. O-PTIR spectroscopy is a recently developed technique that combines the functionalities of traditional FTIR and Raman spectroscopy. It is important to be aware of new developments in IR spectroscopy technology and how they could be utilised to aid in cancer diagnostics. This chapter investigates how O-PTIR combined with machine learning performs to classify lung cancer cells from non-malignant lung cells placed on glass slides.

O-PTIR spectroscopy uses a pump probe setup where the pump is a tuneable pulsed IR light source provided by a quantum cascade laser (QCL) and the probe is a short wavelength optical laser (Kansiz and Prater, 2020). The QCL is directed onto the sample and tuned to wavelengths across its range that corresponds to vibrational modes of the sample. As the IR radiation is absorbed local modulated heating of the sample occurs which causes subtle thermal expansion and refractive index changes. This is called the photothermal response. The optical probe and the QCL laser are made collinear, and the optical probe monitors the photothermal response. The changes in reflected optical probe beam intensity is monitored and demodulated through a lock-in amplifier to generate an IR absorbance spectrum. This technique enables submicron far-filed IR spectroscopy in reflection mode while generating

transmission like spectral quality. An advantage of O-PTIR spectroscopy in comparison to conventional FTIR spectroscopy is the higher spatial resolution of O-PTIR spectroscopy. Spatial resolution is determined by the Rayleigh criterion, spatial resolution = 0.61 wavelengths/numerical aperture of microscope objective. The higher spatial resolution of O-PTIR is because the spatial resolution is determined by the shorter wavelength optical beam and not the IR beam from conventional FTIR spectrometers (Paulus *et al.*, 2021). As O-PTIR spectroscopy produces an IR spectrum through the pump-probe system, how it interacts with the substrate is different to transmission FTIR spectroscopy. Therefore, I was interested to investigate how the glass substrate affects the spectra produced from O-PTIR spectroscopy.

In the previous chapters glass coverslips were used as a substrate for classification of cancer cells with FTIR spectroscopy. The coverslips could be used to collect good quality spectra up to 1350 cm^{-1} (Dowling *et al.*, 2020) (Rutter *et al.*, 2019). However, coverslips are fragile and can be difficult to handle. There is the risk of ruining the samples through breakage of the coverslips especially if multiple analyses (spectroscopy, staining) are being conducted. Glass slides (1mm thick) are easier to handle and much more difficult to break than coverslips but with standard FTIR spectroscopy only allow spectral measurements up to 2000 cm^{-1} (Pilling *et al.*, 2017). O-PTIR interacts differently with the substrate because the IR beam is not travelling through the whole of the sample and substrate as is the case for transmission FTIR spectroscopy. O-PTIR might offer the possibility to obtain good quality spectra to a lower wavenumber than 2000 cm^{-1} using glass slide substrates.

In this study O-PTIR was used to study the non-tumorigenic lung cell line NL20 and the two lung cancer cell lines A549 and CALU-1 placed on standard 1 mm thick microscope slides.

The cells were prepared as cytospins on the slides as was done for the FTIR spectroscopy experiments. The study aimed to assess whether O-PTIR spectroscopy could identify spectral differences between the three cell lines using the glass slides as a substrate. Combined with the use of O-PTIR was the machine learning method RF to classify the cells using the spectral data. To the best of my knowledge this study was the first to investigate the feasibility of O-PTIR as a diagnostic tool for lung cancers using a glass substrate. Due to its use of an optical probe how the technique interacts with glass substrates is different than FTIR spectroscopy therefore it was important to assess how O-PTIR spectroscopy interacts with a glass substrate and how much information can be gained from the spectra.

Aims

1. To assess how much spectral information can be gained from cells placed on a glass substrate using O-PTIR spectroscopy.
2. To assess if O-PTIR spectroscopy with a RF classifier could be used classify the normal lung tissue derived NL20 cell line from cancer cell lines A549 and CALU-1 prepared on a glass slide substrate.

Methods

Cells

For the research conduced in this chapter the cell line NL20, A549 and CALU-1 were. NL20 is derived from non-cancerous lung tissue while A549 and CALU-1 are derived from NSCLC. For detailed methodology on the culture of the cells refer to the relevant section in chapter 2.

Sample preparation

Cells were removed from the flasks as described above and resuspended in 0.9% saline at a concentration of 1 x 10^6 cells per 1 ml. 20 µl of the cell solution was applied to the glass slides by cytospin running at 900 rpm for 1 minute. The cells were immediately fixed by pipetting 100 µl of 4% PFA on to the sample area and incubating for 15 minutes. After the incubation period the excess PFA was poured off and the samples were washed once with saline and thrice with deionised water.

O-PTIR Spectroscopy

A mIRage O-PTIR micro-spectrometer from Photothermal Spectroscopy Corp. was used for this study. The IR pump beam was a dual range QCL covering the wavenumber ranges 3000-2700 cm^{-1} and 1800-914 cm^{-1}. The QCL operated at 100 KHz pulse rate and 100% power at 2.5% duty cycle. The optical probe beam was a 532 nm laser operated at 28% power. The spectrometer was fitted with a room temperature silicon photodiode detector to record the reflected optical beam intensity. Spectra were collected at a spectral resolution of 6 cm^{-1}

with a single scan per replicate spectra. Single spectra took 1 second to scan. Spectra were collected in reflection mode, but the output spectra are transmission like IR spectra because of the pump probe system. Background QCL spectra were collected once per day off a clean Kevley Low-E slide. The system was purged with dry nitrogen gas to minimise the inference from water vapour. Nine spectra were recorded for each individual cell measured to produce an average spectrum for each cell. 50 cells from each cell line were measured. Measurements were centred on the cell centre to gain spectral information from the whole of the cell. The spectra were obtained by Dr M. Kansiz at Photothermal Spectroscopy Corp.

Pre-processing and data analysis

A Savitzky-Golay filter was applied to the spectra to reduce noise with a window size of 11 and a polynomial of 2. EMSC was used to normalise the spectra and remove baseline defects caused by changes in sample thickness. The average spectra were used as a reference for the EMSC. The average spectra for A549, CALU-1 and NL20 were produced by averaging the spectra average spectra of each measured cell from 50 cells of each cell line. The spectra were cropped to the regions 3000-2800 cm^{-1}, 1780-900 cm^{-1}, 1780-1300 cm^{-1} and the combined spectral regions of 1780-900 cm^{-1} & 3000-2800 cm^{-1} and 1780-1300 cm^{-1} & 3000-2800 cm^{-1}. 2nd derivative spectra were generated by applying a second derivative in the Savitzky-Golay filter.

The cells were classified using the spectral data with a RF classifier containing 200 trees. 66% of the data was split into a training set and the remaining data was used as a test set using random sampling. The results of the classification were assessed by the AUC, CA, precision, recall and the confusion matrices.

Results

The O-PTIR spectroscopy was first assessed for how much spectral information can be gained from cells placed on a standard glass slide (1 mm thickness). Figure 65 shows the mean O-PTIR spectra in the region 1800-900 cm^{-1} for A549, CALU-1 and NL20 from 50 cells of each line. For all three cell lines the spectra showed information up to 900 cm^{-1}. But below 1350 cm^{-1} individual bands cannot be seen and is dominated by a single large band. As O-PTIR is a newer method of vibrational spectroscopy and little work has been done with using a glass substrate it is unclear how it interreacted with the glass to form the large single band below 1350 cm^{-1}. The region 1800-1350 cm^{-1} contained the amide I, II and III bands that arise from vibrations of the functional groups in the amide bonds of proteins. Figure 66 shows the mean spectra 3000-2800 cm^{-1}. This region contained peaks corresponding to CH_2 symmetric and asymmetric stretching vibrations mostly from the fatty acid chains in lipids. O-PTIR uses a QCL which measures spectra quickly in high quality in narrow ranges but cannot measure a broad spectrum covering the whole range at once like conventional FTIR spectroscopy. A dual chip QCL was used to allow both regions to be measured. Spectral differences can be seen across the three spectra in both regions measured. CALU-1 had a much higher absorbance in the amide I and amide III bands than A549 and NL20. In the amide II band A549 had the highest absorbance and had a higher absorbance in amide I and amide III than NL20. A549 had a higher absorbance than the other cells in the 2930 cm^{-1} and 2860 cm^{-1} bands and lower absorbance at 2960 cm^{-1}. In the 2nd derivative spectra in Figures 65 and 66 further spectral differences were resolved. There were spectral differences in most of the 2nd derivative bands. The 2nd derivative spectra at 1300-900 cm^{-1} had a lot of noise causing it to

be uninterpretable. The noise is caused by the glass substrate absorbing a portion of the IR radiation.

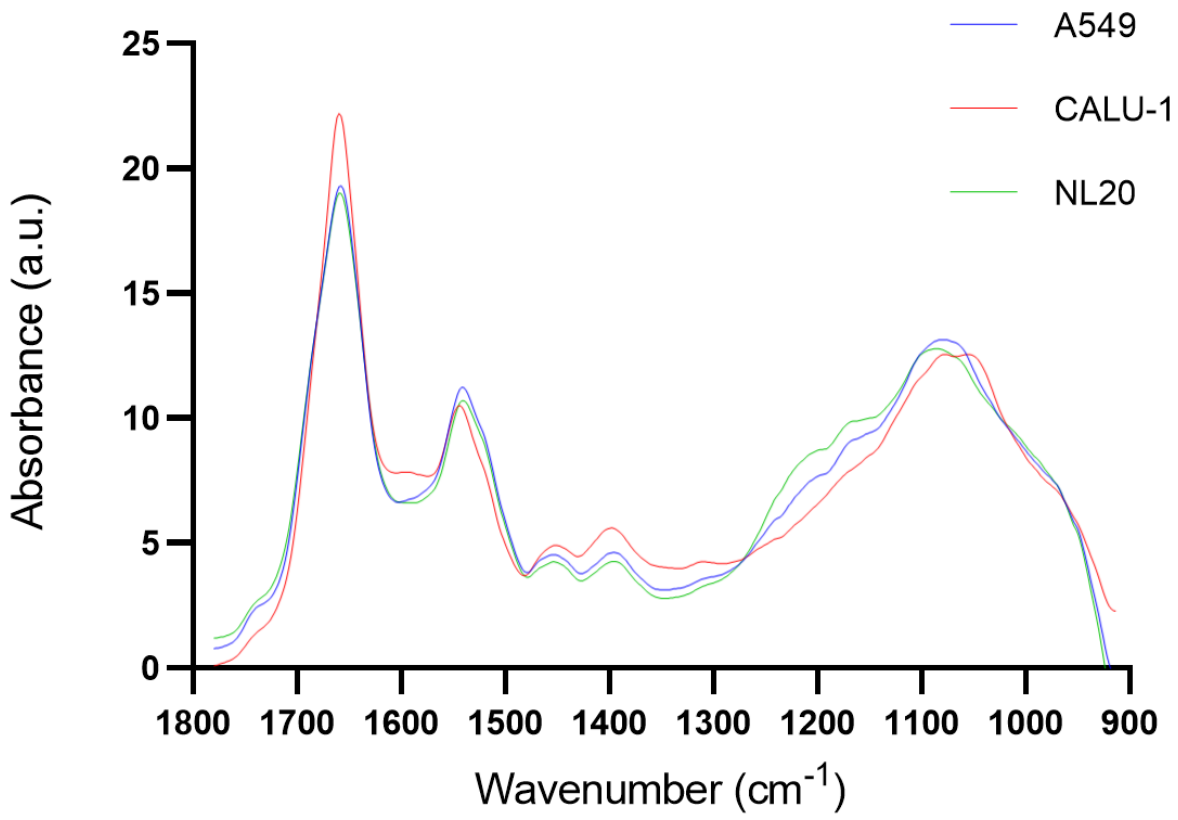

Figure 65 Average spectra in region 1780-900 cm^{-1} of A549, CALU-1 and NL20 from 50 cells of each cell line.

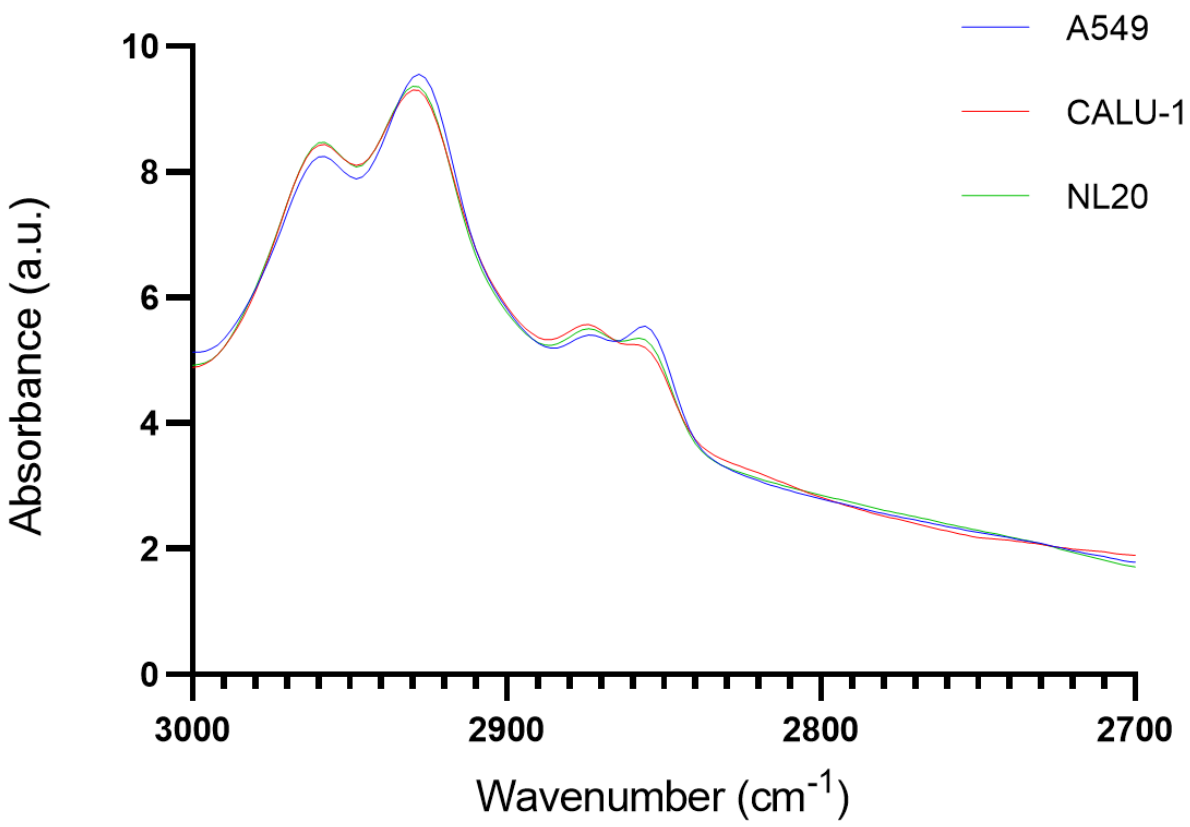

Figure 66 Average spectra in region 3000-2800 cm^{-1} of A549, CALU-1 and NL20 from 50 cells of each cell line.

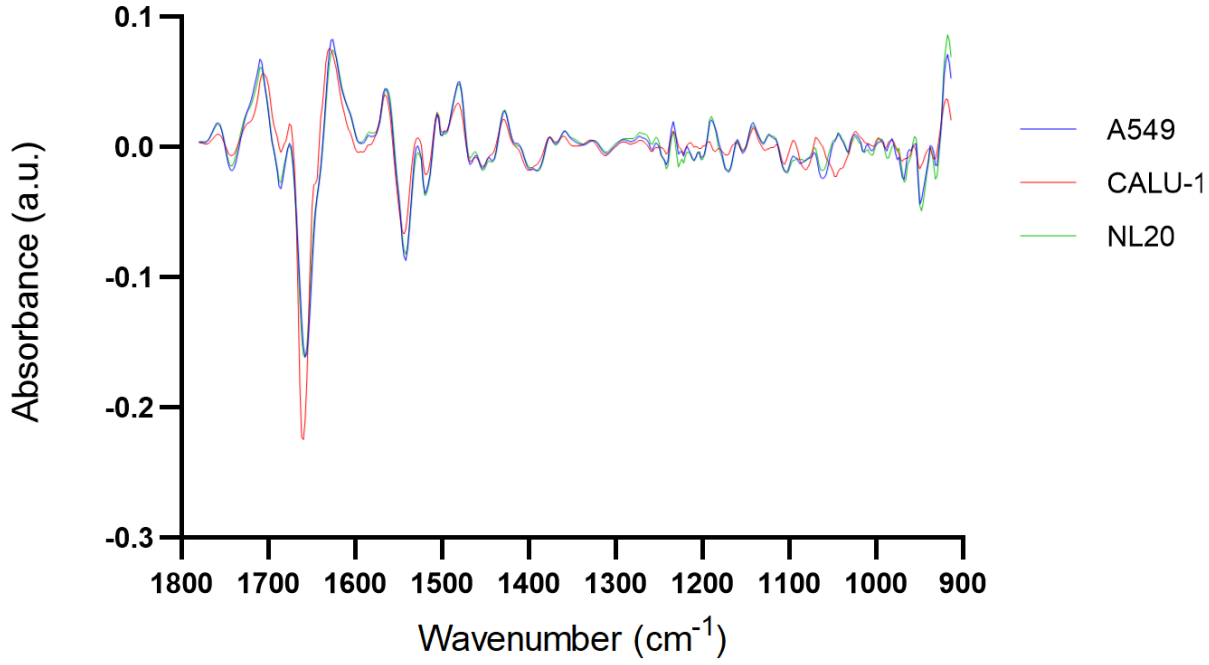

Figure 67 Average 2nd derivative spectra in region 1780-900 cm-1 of A549, CALU-1 and NL20 from 50 cells of each cell line.

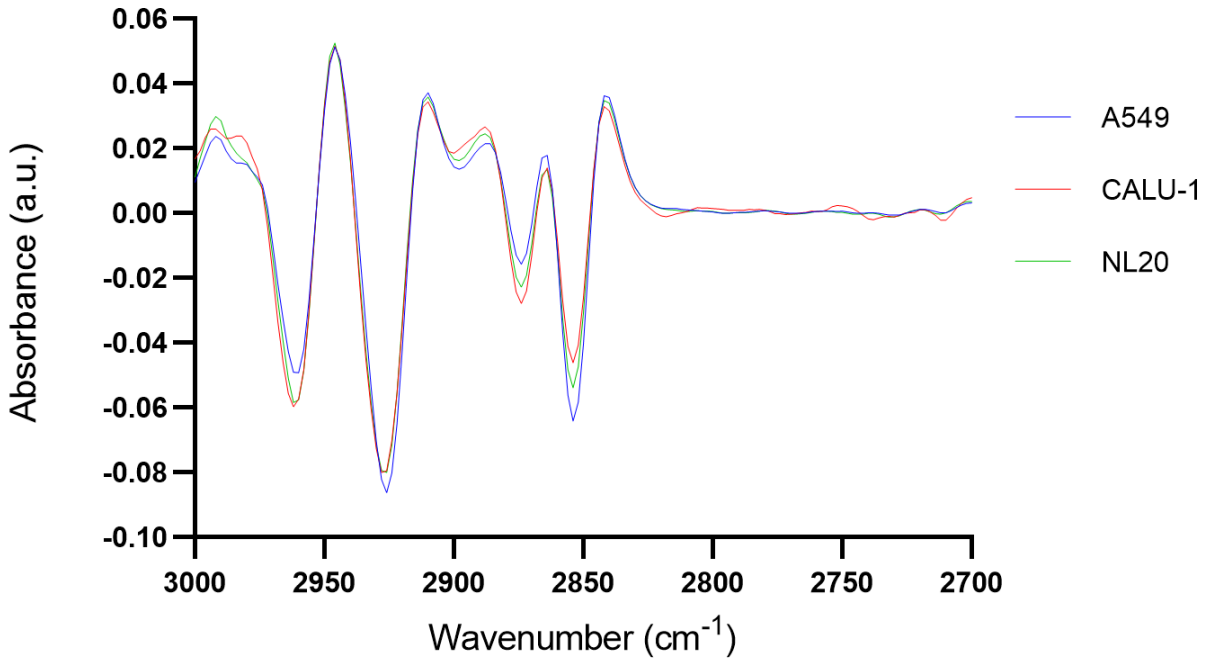

Figure 68 Average 2nd derivative spectra in region 3000-2800 cm-1 of A549, CALU-1 and NL20 from 50 cells of each cell line.

Figures 69-71 shows PCA score plots of the spectra from the cells for all three cell lines. The plot used the regions of the spectra 3000-2800 cm^{-1}, 1780-1300 cm^{-1} and the two regions combined. The CALU-1 cells had a clear separation from the A549 and NL20 cells in all three PCA. There was overlap between A549 and NL20 there was some separation of the two cells line that in the PCA for 3000-2800 cm^{-1}. The loading plots for the PCA showed strong positive features at 2926 and 2854 cm^{-1} which arise from CH and CH_2 stretching vibrations mostly from lipids fatty acid chains.

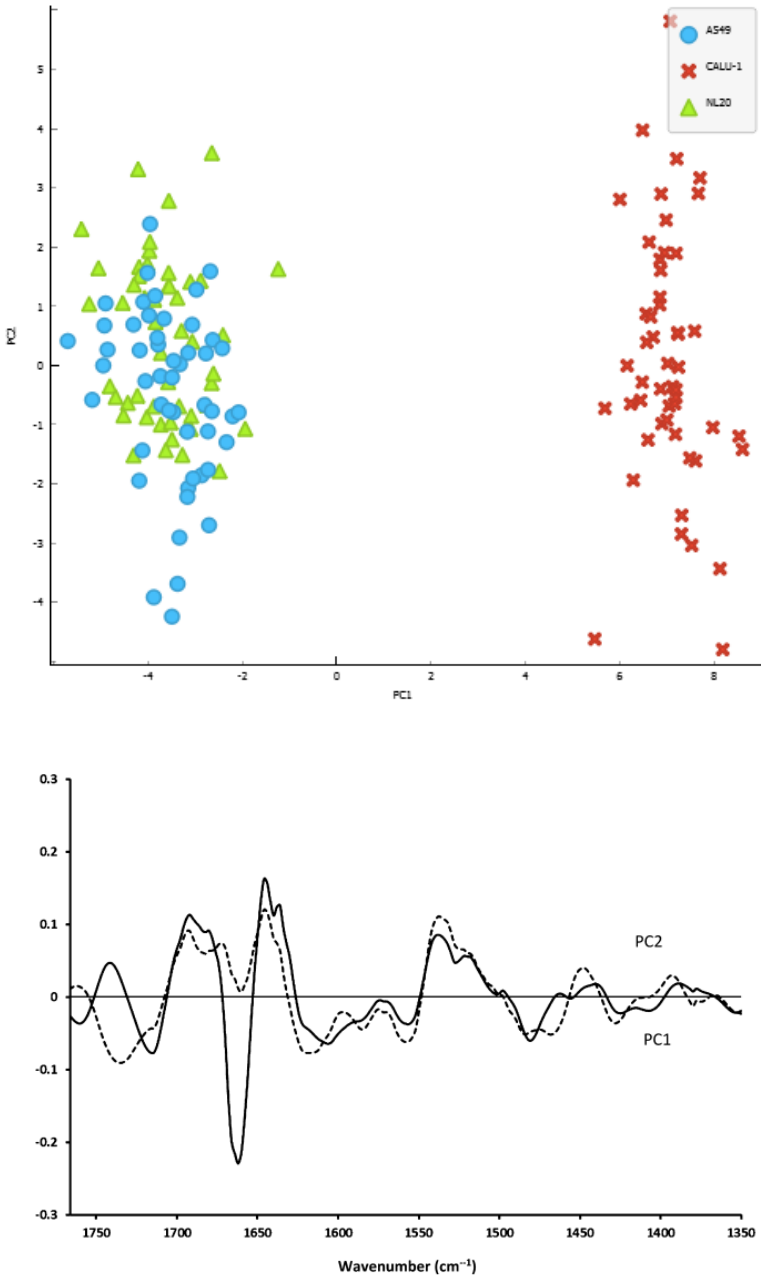

Figure 69 PCA score of A549, CALU-1 and NL20 spectra in region 1780-1300 cm^{-1}. PC1 = 80%, PC2 = 10%.

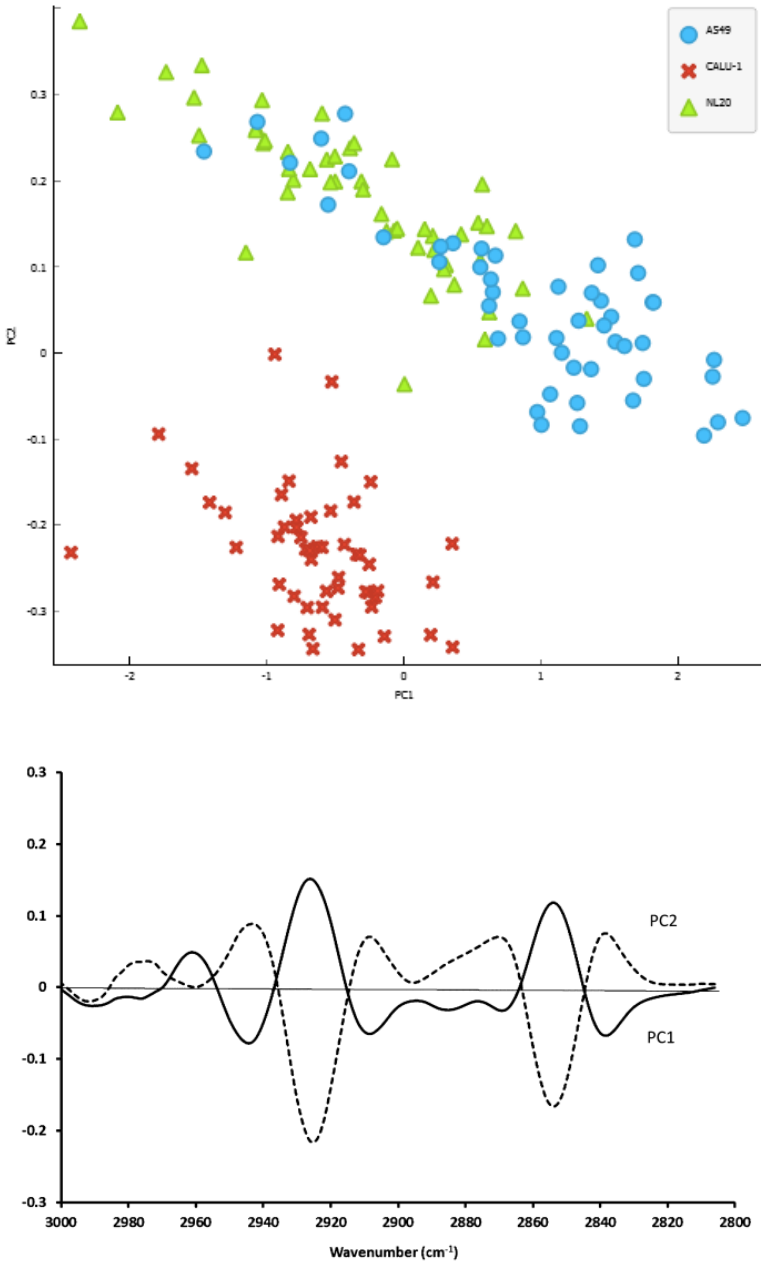

Figure 70 Top: PCA score of A549, CALU-1 and NL20 spectra in region 3000-2700 cm^{-1}. PC1 = 83%, PC2 = 13%. Bottom: loading plot of PC1 and PC2.

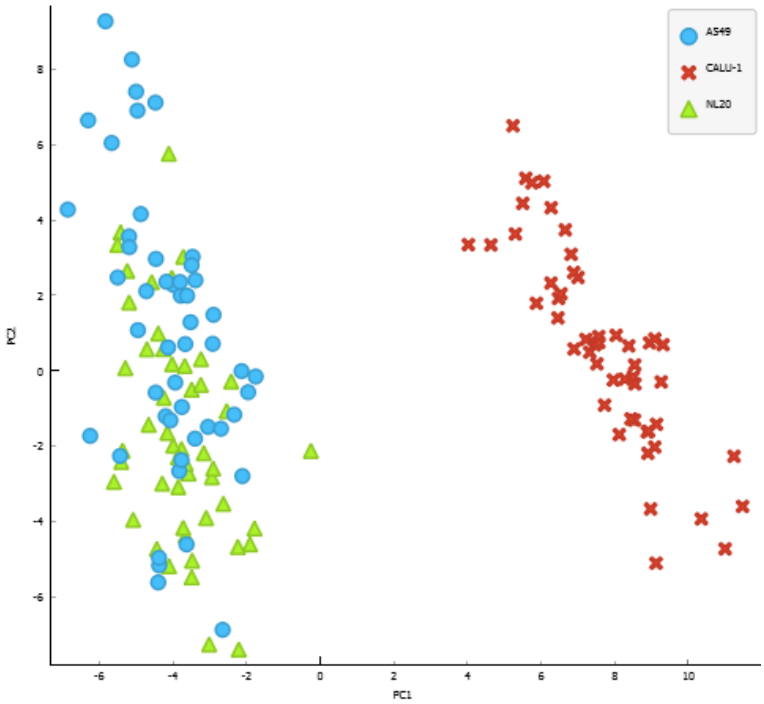

Figure 71 Top: PCA score of A549, CALU-1 and NL20 spectra in region 3000-2700 cm^{-1} and 1780-1300 cm1- combined. PC1 = 63%, PC2 = 19%.

A RF classifier was used to classify the three cell lines using IR spectra. Classifications were performed using the spectral region 1780-1300 cm^{-1} and 3000-2700 cm^{-1} individually and combined. Classifications were also tested including the band at 1300-900 cm^{-1}. Table 13 below shows the classification metrics for the RF classifications with the different spectral regions. From these metrics it showed that the inclusion of the 1300-900 cm^{-1} band did not improve the classification. The classification using the region 1780-1300 cm^{-1} provided the best classification with an F1 score of 0.962. The 3000-2700 cm^{-1} region and combined regions also performed well with F1 scores of 0.940 and 0.948 respectively. RF classification using the 2nd derivative spectra was also tested. The classification metrics using the 2nd derivative spectra are shown in Table 14. The use of the 2nd derivative spectra improved the classification using the region 3000-2700 cm^{-1} and the combined regions with F1 scores of

0.955 and 0.963. The classification metrics using the 2nd derivative of the region 1780-1300 cm^{-1} did not improve from the use of the normal spectra with all the metrics remaining the same. The confusion matrices below (Figures 72 and 73) show the percentages of how the cells were classified and misclassified. The classifications with the raw spectra and 2nd derivative spectra using the region 1780-1300 cm^{-1} and regions combined classified 100% of CALU-1 cells correctly. Using the 3000-2700 cm^{-1} region of the normal spectra, 97.8% of CALU-1 was correctly classified, the small number of misclassifications were attributed to 1.5% A549 and 0.6% NL20. Using the 2nd derivative spectra 0.2% of CALU-1 was misclassified as A549. All instances where A549 was misclassified, it was as NL20 and all the misclassifications of NL20 was as A549. This corresponds to the PCA score plot where there is overlap between the A549 and NL20 clusters and the CALU-1 cluster is separated. The confusion matrices of the classifications including the 1300-900 cm^{-1} band (Figure 73) showed that its inclusion worsened the classification of A549 and NL20 cells.

Spectral region (cm^{-1})	AUC	Classification accuracy	F1	Precision	Recall
3000-2700	0.990	0.940	0.940	0.941	0.940
1780-900	0.984	0.918	0.918	0.919	0.918
1780-1300	0.997	0.962	0.962	0.962	0.962
1780-900 & 3000-2700	0.985	0.920	0.919	0.919	0.920
1780-1300 & 3000-2700	0.992	0.952	0.952	0.953	0.952

Table 13 Classification results of RF classification of A549, CALU-1 AND NL20 using O-PTIR spectra.

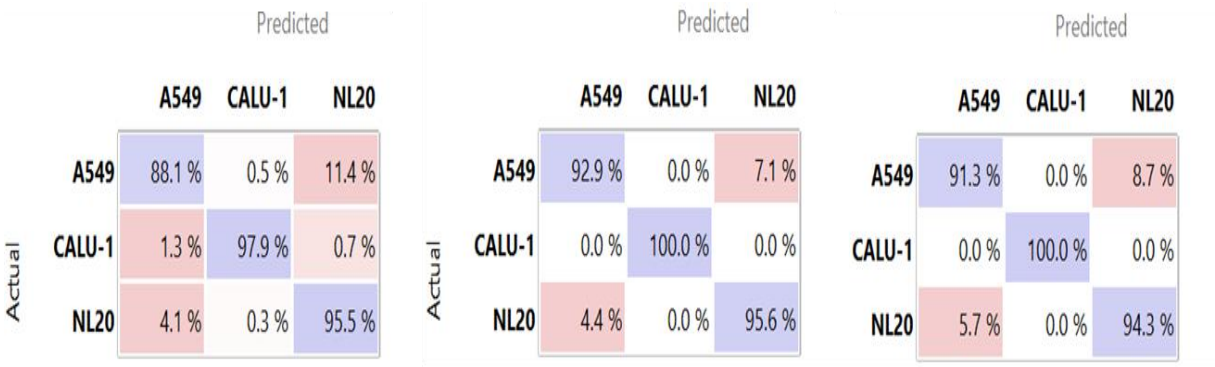

Figure 72 Confusion matrices for RF classification of A549, CALU-1 and NL20 using O-PTIR spectra. Left: 3000-2700 cm^{-1}, middle: 1780-1300 cm^{-1}, right: 3000-2700 cm^{-1} and 1780-1300 cm^{-1} combined.

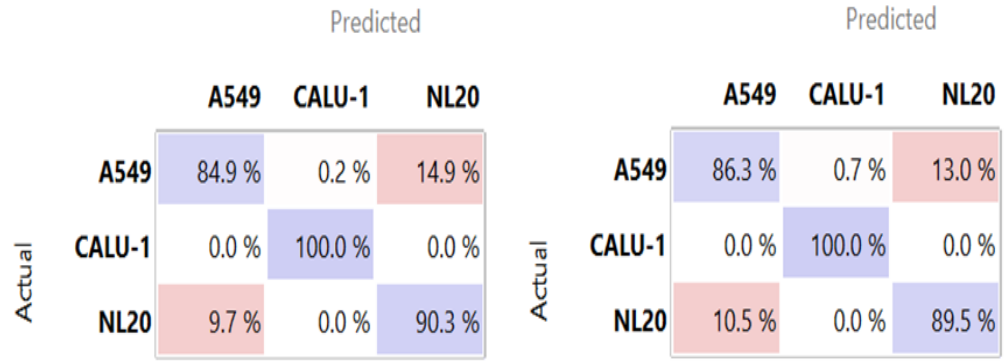

Figure 73 Confusion matrices for RF classification of A549, CALU-1 and NL20 using O-PTIR spectra. Right: 1780-900 cm^{-1}, left: 1780-900 & 3000-2800 cm^{-1}.

Spectral region (cm^{-1})	AUC	Classification accuracy	F1	Precision	Recall
1780-1300	0.996	0.962	0.962	0.962	0.962
3000-2700	0.993	0.954	0.954	0.955	0.954
3000-2700 & 1780-1300	0.997	0.968	0.968	0.968	0.968

Table 14 Classification results of RF classification of A549, CALU-1 AND NL20 using 2nd derivative O-PTIR spectra.

Figure 74 below shows the confusion matrices from the classification of A549, CALU-1 and NL20 using the 2nd derivative O-PTIR spectra. The confusion matrix from the classification using 3000-2700 cm^{-1} showed that using the 2nd derivative improved the classification of all three cell lines. For the classification with the combined region using the 2nd derivative spectra there was an improvement on the classification of both A549 and NL20.

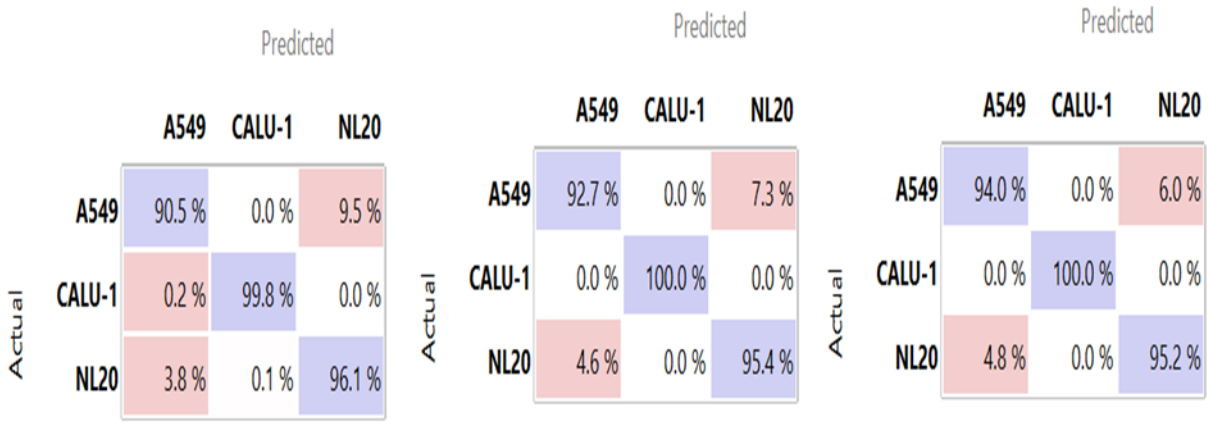

Figure 74 Confusion matrices for RF classification of A549, CALU-1 and NL20 using 2nd derivative O-PTIR spectra. Left: 3000-2700 cm-1, middle: 1780-1300 cm-1, right: 3000-2700 cm-1 and 1780-1300 cm-1 combined.

Table 15 shows the ten features given the most importance by the RF classifier for the regions 3000-2700 cm^{-1}, 1780-1300 cm^{-1} and the two combined. For the 3000-2700 cm^{-1} there are features across all the bands within the ten features. Similarly in the region 1780-1300 cm^{-1} there are features across the three bands amide I, II and III in the ten most important features. When the regions combined were used, features were included from both regions but with a majority of the features from the amide bands. This demonstrated that the RF classifier utilised features from across the spectra to perform the classifications. Table 16 show the features given the most importance by the classifier when the 2nd derivative spectra were used. The classifier gave importance to features across different bands in the 2nd derivative spectra as with the normal spectra. For the classification with the combined spectral regions, the majority of the ten features were also in the amide bands.

3000-2700 cm^{-1}	1780-1300 cm^{-1}	3000-2700 cm^{-1} & 1780-1300 cm^{-1}
2940	1720	2844
2982	1628	2842
2864	1722	1628
2938	1576	1492
2832	1572	2846
2834	1626	1506
2866	1410	1594
2826	1622	1424
2944	1574	1502
2942	1364	1382

Table 15 Ten features given most importance by RF classifier used for classification of A549, CALU-1 and NL20 using O-PTIR spectra.

3000-2700 cm^{-1}	1780-1300 cm^{-1}	3000-2700 cm^{-1} & 1780-1300 cm^{-1}
2800	1578	2936
2804	1302	1302
2802	1594	1352
2916	1718	1590
2864	1582	1436
2818	1560	1602
2936	1442	2862
2918	1436	1696
2820	1724	1582
2830	1588	1438

Table 16 Ten features given most importance by RF classifier used for classification of A549, CALU-1 and NL20 using 2nd derivative O-PTIR spectra.

IR spectra were measured using three different instruments across the chapters in this thesis including a benchtop globar source spectrometer, a synchrotron source spectrometer and an O-PTIR spectrometer. Below figures 75 and 76 show 50 spectra of A549 de-noising pre-processing measured using each of the three instruments to demonstrate the similarities and differences between the spectra produced by the three instruments. The bands in both regions are clear and defined across the three instruments and as demonstrated across the thesis can be used for classification of cancer cell lines from healthy tissue derived cell lines

and the cancer cell lines from each other. The SNR varied across the three instruments. The spectra measured with the globar source showed the most noise, followed by the synchrotron source and the O-PTIR with a QCL IR source showed the least noise in the spectra measured.

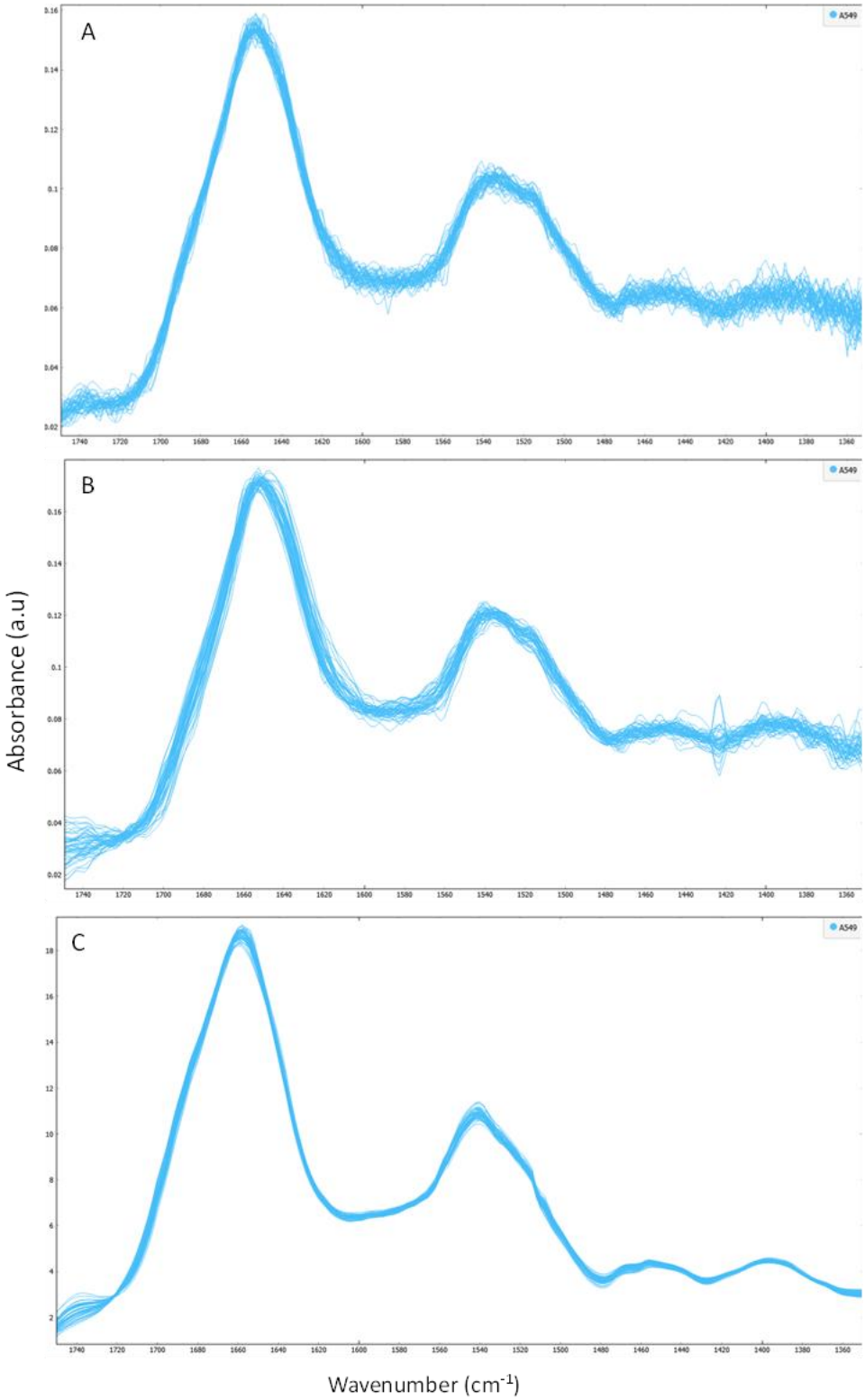

Figure 75 A549 IR spectra in the region 1350-1750 cm^{-1}. A) 50 spectra from benchtop a spectrometer with a globar IR source. B) 50 spectra from a spectrometer with a synchrotron IR source. C) 50 spectra from an O-PTIR spectrometer with a QCL IR source.

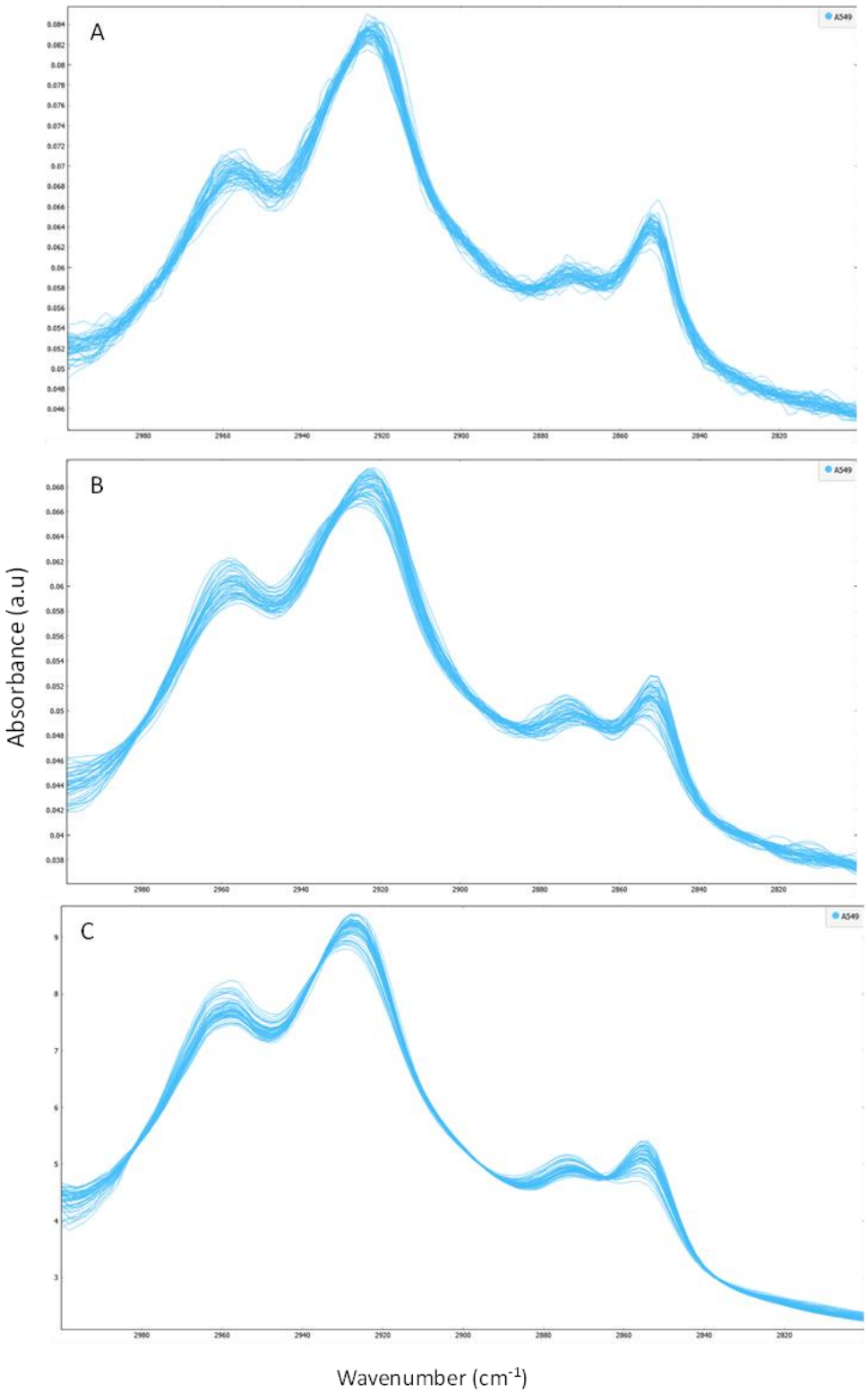

Figure 76 A549 IR spectra in the region 2700-3000 cm^{-1}. A) 50 spectra from benchtop a spectrometer with a globar IR source. B) 50 spectra from a spectrometer with a synchrotron IR source. C) 50 spectra from an O-PTIR spectrometer with a QCL IR source.

Discussion

It was important to assess the bounds of the newer technology of O-PTIR spectroscopy for measuring cells on glass slides because there has been little research with how it interacts with substrates including glass. O-PTIR spectroscopy allowed for spectral information to be gained up 900 cm^{-1} but individual bands could be only resolved up to 1350 cm^{-1}. This allows more of the spectra to be viewed than conventional FTIR spectroscopy when using a glass slide of 1 mm thickness which has a cut off from 2000 cm^{-1} in the spectra for what can be resolved (Rutter *et al.*, 2018). Interestingly it is the same cut off point for O-PTIR spectroscopy when FTIR spectroscopy is used with thinner glass coverslips. O-PTIR spectroscopy enables the use of thicker more durable glass slides to gain the same information as FTIR spectroscopy when a glass coverslip is used. The glass slides are easier to handle because of their durability compared to fragile coverslips and equipment such as microscope stages are made to fit glass slides. The individual bands could not be seen in the fingerprint region between 1400-900 cm^{-1} but there was a broad band that varied in shape in all three cell lines. This region in conventional FTIR spectroscopy with a glass substrate is unresolvable noise because of absorption of IR radiation by the glass. It is uncertain if this single band has any value regarding the biochemical profile of the cells. In an IR spectrum on a non-glass substrate this region would contain smaller bands that occur from vibrations in the functional groups of nucleic acids and carbohydrates (Diem, Melissa Romeo, *et al.*, 2004). The differences in cell thickness could be a reason for the variability seen in the band. If the glass contribution could be subtracted from the spectra more information could be gained from the fingerprint region, this could be an area for future research. Other studies that have analysed tissue sections of > 5 µm in thickness on glass substrates with O-PTIR

spectroscopy (Bakir et al., 2020). The spectra they obtained did not have glass contributions because the light source did not penetrate through the whole of the tissue to the substrate. This demonstrates sample thickness is an important factor to consider when using O-PTIR spectroscopy with glass substrates. The insensitivity to glass O-PTIR spectroscopy has when compared to FTIR spectroscopy is enabled by the reflection geometry of the technique. The IR beam hits the sample where it is attenuated and generates a photothermal effect which is detected by the visible beam probe. The probe detects the IR photothermal effect in reflection mode therefore it is not transmitted through the whole glass substrate as is the case for transmission FTIR spectroscopy so there is less interference from the glass. The ability to gain spectral information on both lipids and proteins while using standard glass slides gives O-PTIR promising potential for use in cancer diagnostics. While the spectra were collected in a reflection mode with the visible light probe, the resultant spectra are more like spectra from transmission FTIR spectroscopy without the scattering artefacts and distortions that are common with transflection FTIR spectroscopy.

The spectra produced from O-PTIR spectroscopy were of high-quality spectra with little noise. The spectrometer used is a benchtop spectrometer. The high resolution and short measurement time was obtainable from the benchtop instrument due to the use of a QCL IR source which has submicron resolution. To achieve high resolution from a benchtop instrument is an important factor to consider for the translation of vibrational spectroscopy techniques to clinical diagnostics because it is benchtop spectrometers that will be used. An important consideration when using QCL IR sources is they do not measure a broad range and cannot capture the whole spectra as other light sources. The use of a dual chip QCL allowed collection of both the higher and lower wavenumber regions presented in the spectra. However, spectral measurements above 3000 cm^{-1} which contain the amide A band

were not captured. Despite missing the amide A band an excellent classification of the cells was achieved with high accuracy, precision and recall.

The RF classifier performed well using the O-PTIR spectra to classify A549, CALU-1 and NL20. The band in the fingerprint region of 1300-900 cm^{-1} did not improve the classification results. This suggests that this band does not offer useful information on the biochemical properties of the cells despite the differences in the band between the cell lines. Therefore, I think that this band is not used for the classification of cells when using O-PTIR spectroscopy with a glass substrate. The region that provided the best classification using the raw spectra was 1780-1300 cm^{-1} which contains bands that correspond to vibrations in the amide bonds of proteins (Diem, Melissa Romeo, *et al.*, 2004). The protein content of cancer will differ greatly from non-cancerous cells and that of different cancers because the mutations that result in cancers will change the protein expression. The lipid bands of the region 3000-2700 cm^{-1} and the two regions combined also provided a good classification of the cells, but the amide bands classification produced the best performance. Combining the bands improved the classification from just using the lipid bands but was not superior to the amide bands alone.

The loading plot for the PCs shown in Figure 70 demonstrated strong positive peaks at 2926 cm^{-1} and 2854 cm^{-1} in the lipid region representing symmetrical stretching and asymmetrical stretching respectively in CH_2 (Baker *et al.*, 2014). A difference in lipid region between the same cell lines around 2850 cm^{-1} was also shown in spectra acquired using FTIR spectroscopy on the same cell lines in chapter 4. Both techniques picked up that there were differences in the lipid content of the three cell lines. The CALU-1 and A549 cells had a higher absorbance in the amide I band than NL20 which was also the case in the FTIR spectra of the same cell lines but the band for CALU-1 had a larger difference in absorbance to the other cell lines in

the O-PTIR spectra. A549 has a higher absorbance in both O-PTIR and FTIR spectra than NL20 but CALU-1 has a lower absorbance than NL20 in the O-PTIR spectra and a higher absorbance in the FTIR spectra. These results may indicate that the two techniques can have different spectral signals. The classification using FTIR spectra (chapter 4) struggled to classify CALU-1 from A549 but using O-PTIR spectra improved the CALU-1 classification considerably while also improving the classification of NL20. The classification of A549 provided a similar percentage of correctly identified cells. Using the 2nd derivative spectra improved the classification performance when using the region 3000-2700 cm^{-1} and the combined regions. This is consistent with the classification of cells in chapter 4 and 5 where the 2nd derivative of the FTIR spectra also improved the performance. While using the 2nd derivative of the 3000-2700 cm^{-1} improved the performance, the region 1780-1300 cm^{-1} still provided the strongest performance. However, using the 2nd derivative of the regions combined improved the classification over only using 1780-1300 cm^{-1}. The improvements in the classification from using the 2nd derivative were only small improving the classification of the individual cell lines by <5%. As the classification performance was already high it would be difficult to improve classification much further. The difference in performance between the regions used was also small only differing by small percentages. The classification using the amide bands of the O-PTIR spectra provided a better classification of A549, CALU-1 and NL20 than the FTIR spectra using the same bands.

While Raman measurements were not taken for this experiment the Mirage O-PTIR instrument can measure simultaneous Raman and IR spectra (Spadea et al., 2021). Having access to both types of spectra could be useful as they provide different but complementary biochemical information. For this experiment however the focus was on the quality of IR spectra available from cytological samples on glass and its comparison to FTIR spectra from

traditional sources as shown in other chapters. The IR spectra provided have been shown to produce a good classification themselves without the need for additional features from Raman spectroscopy. The use of two types of spectroscopies would have the drawback of making the analysis more complex requiring more data processing needed for two different types of spectra. For clinical translation the extra complexity needed for the data processing and analysis could be a hinderance as it would increase the time taken to analyse the sample and would require knowledge of both techniques. The IR spectra alone provided excellent classification of the cells and therefore in this case the use of Raman was not needed.

Figures 75 and 76 showed spectra from the three IR sources used across this thesis including globar, synchrotron and QCL. The spectra from the globar and synchtotron sources produced more noise than the spectra from the QCL source. The QCL was also much faster in measuring spectra taking seconds to measure a spectrum while the other two sources took between 60-90 seconds to record a spectrum. The high SNR and speed of measurement is an advantage of QCL sources. The speed of measurement would especially be an advantage for clinical translation allowing more samples to be measured which could help the pathologist in triaging which patients samples need further investigation. The current disadvantages of QCL source spectrometers is that they are currently more expensive to buy because it is a newer technology and it has to be considered that they can only measure in a limited range of wavenumbers.

One of the current problems for the translation of O-PTIR is that there is only one available instrument in production for the technique which is the Mirage O-PTIR spectrometer. This currently poses the difficulty of there being a lack of capacity for wide scale use of the technique. As O-PTIR has only been recently developed there is a wide scope of research to

find and develop the best ways of utilising this technology. From this experiment it has been shown that it is possible to be used with glass substrates for classification of lung cancer cells from non-malignant lung cells. O-PTIR spectroscopy can use thicker glass than transmission FTIR spectroscopy to obtain IR spectra in the same regions.

Conclusions

Using O-PTIR spectroscopy allows high quality IR spectra of cells to be measured on glass slides down to 1300 cm^{-1} because it has less interaction with the substrate. With O-PTIR spectroscopy the lung cancer cells could be classified from each other and NL20 with a high classification accuracy, precision and recall that was higher than with FTIR spectroscopy. The classification of the SqCC CALU-1 had the largest improvement using O-PTIR spectroscopy. Currently the biggest drawback of O-PTIR spectroscopy is that there is a small userbase with only a small number of instruments available.

Chapter 8: Discussion and future work.

There is demand in cancer diagnostics for diagnostic tools that can complement and support current diagnostic techniques to improve the management of cancer. The increasing incidence of cancer in developed nations is putting further pressure on pathology departments. In the NHS, there is a target that 93% of patients should have an appointment with a cancer specialist two weeks after GP referral. In 2022 48% of NHS trusts in England failed to meet this target every month (NHS Statistics, Provider-Based Cancer Waiting Times for December 2022 – 23, 2023). This crisis in cancer management within the English NHS demonstrates the need for diagnostic methods that improve turnaround time to meet the two-week referral target to ensure patients are treated in a timely manner. The two cancers focused on for this research were lung and breast cancer. Lung cancer is largely diagnosed at later stages of disease which is linked to poor survival. Novel diagnostic techniques are needed for earlier lung cancer diagnoses to improve treatment options and survival. Breast cancer diagnosis has a problem of overdiagnosis from the screening program and accounts for a large number of cancer cases being diagnosed. Novel tools that help to diagnose cancer in an objective manner to accurately identify cancerous samples from non-cancerous samples would help to improve management of cancer cases. This thesis aimed to investigate methodologies for the use of IR spectroscopy with glass substrates for the classification of lung and breast cancers. IR spectroscopy allows classification of cancer and non-cancerous cells from their biochemical makeup. If IR spectroscopy techniques can be translated to clinical use, they could be a useful tool for cancer management. The research in this thesis aimed to bring research in the area of IR spectroscopy for cancer diagnostics by

showing how glass substrates can be used which would reduce the cost of the method and allow it to fit more easily within current practices in pathology laboratories.

The first experimental chapter (chapter 3) focused on finding a methodology to prepare samples of cells on glass substrates. If FTIR spectroscopy is to be adopted for cancer diagnostics within clinical settings it must be minimally disruptive. The aim of the research was to test smear and cytospin as preparation methods for FTIR spectroscopy analysis of cells on glass coverslips. Also evaluated were two methods of fixation methanol or 4% PFA. These preparation and fixation methods are all commonly used to prepare cytological samples for current diagnostics therefore their use for FTIR spectroscopy analysis would have negligible disruption on current workflows.

Both preparation methods could be used to produce samples quickly and both produce high quality spectra. The cytospin method was found to allow for measurement of the cells more easily because the cells are deposited in a smaller area than the smear which spreads the cells across a large area. The cells being concentrated in a smaller area when prepared as a cytospin meant there was less time finding cells on the coverslip and reduced measurement time. The quality of a smear is more variable on the experience of the practitioner therefore cytospins provide more consistent quality of samples that are easily reproducible. The cytospin required a centrifuge to produce the samples while the smear did not need any specialised equipment. For pathology laboratories in hospitals in developed areas of the world this is not a problem because they will already be well equipped with the equipment needed to produce cytospins. However, in areas that are remote and have less developed health infrastructure, the smear could still be used if the equipment for cytospin preparation is not available.

4% PFA was found to be the preferred method for FTIR spectroscopy analysis on the glass substrates because it maintained the integrity of the lipid and protein components of the cells. The lipids and proteins are the only biochemical groups which can be measured using a glass substrate because most of the fingerprint region is lost from absorption of IR radiation by the glass. Therefore, it is important to reduce any further loss of information. The methanol fixation while fixing the sample faster removed lipid content from cells because it is an alcohol. The conclusion of this research was that cytospin with 4% PFA fixation was the most appropriate method to prepare cells on glass for FTIR spectroscopy measurement and analysis.

Future work would evaluate the proposed preparation method at a larger scale within pathology laboratories. It cannot be truly known how the method fits into current workflows until it is tested with a large number of samples in the intended environment. Ideally it should cause minimal disruptions because the methods are currently for sample preparation for other diagnostic techniques.

Chapter 4 aimed to investigate if the methodology for sample preparation developed in chapter 3 could be used for the classification of NSCLC cells from non-malignant lung cells. The adenocarcinoma cell line A549, SqCC cell line CALU-1 and the non-malignant lung cell line NL20 were classified from FTIR spectra using a RF classifier. The RF classifier was selected because it handles data with many features well and is less prone to overfitting. Previously, there has been little research using FTIR spectroscopy with samples prepared on glass coverslip substrates to classify single lung cancer cells from non-malignant lung cells.

The results showed that A549 and CALU-1 could be classified with high accuracy, precision, and recall from NL20. The region 3500-2700 cm^{-1} containing the amide A band and lipid

bands was demonstrated to provide a better classification than the region 1800-1350 cm^{-1}. The classification was further improved by using the 2nd derivative spectra. The methodology also allowed classification of A549 and CALU-1 from each other. These results show that the methodology has potential to separate cancerous samples from non-cancerous samples and help to inform the typing of NSCLC.

This work used cell lines to test the feasibility of the proposed methodology. Future work will test the methods with non-cancerous lung cells and lung cancer cells from patients. It would have to expand the research to test different types and subtypes of lung cancer to assess how well the methodology performs with different types of lung cancer. It will also be important in future work to test the methodology with different stages of lung cancer to assess if it can be used for the diagnosis of early stages of lung cancer. Lung cancer has poor survival in later stages, if FTIR spectroscopy could be used in earlier diagnosis of lung cancer it would improve the management of lung cancer.

Chapter 5 aimed to expand the methodology to assess if it could be applied to multiple cancers by testing it with breast cancer. The invasive ductal carcinoma cell line BT549, non-invasive ductal carcinoma line MCF7 and healthy breast tissue derived line MCF10A were used for this research. The results demonstrated that the methodology could be used for the classification of breast cancer cells from non-malignant breast cells. Also demonstrated was the classification of invasive breast cancer cells from non-invasive breast cancer cells. The region 3500-2700 cm^{-1} provided a better classification than 1800-1350 cm^{-1} of the breast cancer cells like was also demonstrated for lung cancer cells. The spectra used in chapter 5 were collected using a benchtop spectrometer with a globar IR source while the spectra used in chapter 4 were collected using a spectrometer with a synchrotron IR source. It was

demonstrated that the methodology is applicable with both synchrotron and benchtop spectrometers.

To expand this research in the future as with the lung cancer, breast cancer cells and breast cells from patients would have to be measured using the methodology. The research showed the methodology can be applied to multiple types of cancer therefore future work would continue to test the methodology with other solid cancers.

Chapter 6 aimed to assess the feasibility for a methodology of using FTIR spectroscopy with glass substrates for CTC identification. To the best of my knowledge there has been no research published previously investigating FTIR spectroscopy for CTC detection. Blood was doped with the lung cancer cell lines A549 or CALU-1. Areas of the samples containing cancer cells and leukocytes were measured with FTIR spectroscopy. A RF classifier was trained using spectra from A549 or CALU-1 cells and leukocytes. The maps were colored by the RF classier based on the probability of each tile to contain spectra from a lung cancer cell. The colored maps demonstrate that the A549 and CALU-1 cells could be accurately identified from the leukocytes. The identity of the cancer cells was further confirmed by a Giemsa stain which demonstrated how the FTIR spectroscopy can be used in conjunction with current cytological techniques.

This research has demonstrated the feasibility of identifying cancer cells in blood. Future work must test it with CTCs from patients. In theory this method should work with actual CTCs because they, like the cancer cells from the cell lines, will have a hugely different biochemistry from blood cells. Another major step for future work will be testing the methodology with isolation techniques. In order to generate enough areas containing cancer cells in the samples to produce enough measurements for a robust training dataset many

more cancer cells were added to the blood than would be found in the blood of a patient. There are very few CTCs in blood relative to the blood cells with numbers estimated being between 1 and 50 per 7.5 ml, therefore, an isolation step will be needed to enrich the CTCs. There are many different isolation methods that have been researched, therefore multiple methods would have to be tested to find which isolation method would work best in conjunction with the FTIR spectroscopy identification. Future research should also investigate measuring spectra using an instrument fitted with an FPA detector to improve the measurement speed. The use of single point measurements make obtaining the maps time consuming as each spectrum took around 90 seconds to collect. An FPA detector would allow imagining of the whole sample area produced from a cytospin but with the MCT detector it was limited to measuring small area because of the time it takes to collect the measurements.

Chapter 7 investigated the use of O-PTIR for the classification of A549 and CALU-1 cells from NL20 on 1 mm thick glass slides. The first aim of the chapter was to assess what spectral measurements of cells could be gained using O-PTIR with a glass slide substrate. It was found that with O-PTIR spectroscopy, spectra of cells can be recorded up to 1350 cm^{-1} on glass slides. When using conventional FTIR spectroscopy with glass slides the spectra is cut off at 2000 cm^{-1}. This is an advantage for O-PTIR spectroscopy because of the ease of handling the glass slides in comparison to the coverslips. The second aim was to investigate if the O-PTIR spectra were useful for the classification of A549, CALU-1 and NL20. The RF classifier using the O-PTIR spectra classified the cells with high accuracy and had better classification of the cells than with the FTIR spectra used in chapter 4 for the classification of the same cell lines. The biggest current disadvantage of O-PTIR is that there is a small userbase of the technique because only recently it has been developed. There are a limited availability of instruments

and people who know how to use them, which would be a problem if every pathology laboratory required use of such an instrument.

Future work expanding on the research from chapter 7 would use O-PTIR spectroscopy to analyse cytology samples from patients adapting the methodology shown here. This methodology can be expanded to measure other types of cancer. Another area of future work is to see if the glass contribution can be subtracted from the fingerprint region 1300-900 cm^{-1} and whether this would provide biochemical information in the resulting bands on nucleic acids and carbohydrates.

All the research conducted in this thesis used cell lines which is a limitation of this research. Cell lines do not have the heterogeneity that cells obtained from patients would. They can be variation in the same type of tumour across patients and across stages of the disease. Even within a single tumour there can be intra-tumour heterogeneity. The research in this thesis was to demonstrate the feasibility of the proposed methods. For translation of the methods the future stages of research will need to use cancer cells obtained from patients from a range of subtypes and stages of the disease. A difficulty in progressing the research to patient samples will be collecting enough samples of the different types and stages of the cancers to generate robust training data. It may be possible to use cell lines as training data for initial identification of cancer cells in cytology samples, but it would not suffice for specific classification of type and stage.

To conclude, this thesis has addressed the aims set out in each chapter. The research demonstrated the feasibility of preparing cytology samples on glass using a cytospin and PFA fixation for measurement by FTIR spectroscopy. The proposed methodology should be minimally disruptive to current pathology laboratory workflows using the suggested

preparation methods, and a substrate that is widely available and affordable. The spectra collected from the methods were of high quality and allowed classification of cancer cells from non-malignant cells with high accuracy using a RF classifier. Additionally, the methods have shown potential in classifying different types of lung and breast cancers from each other. Also presented was a novel methodology for the identification of cancer cells in blood with FTIR microspectroscopy that showed the feasibility of FTIR spectroscopy as a tool for CTC analysis. The proposed methodologies can be used in conjunction with current diagnostic methods because of the label-free non-destructive nature of FTIR spectroscopy combined with the use of a glass substrate.

Appendices

Appendix 1. TNM classification of lung cancer

The TNM staging system is used to define the extent of the cancer and provide a prognosis to guide treatment. There are three components to TNM staging: the extent and features of the primary tumour (T), lymph node involvement (N) and extent of metastasis (M).

Primary tumour (T)	
Category	Descriptor
Tx	Tumour that is proven histopathologically but cannot be assessed using imagining modalities.
T0	No evidence of a primary tumour
Tis	Carcinoma in situ
T1	Size: <3 cm, local invasion: none, location: in or distal to the lobar bronchus.
T2	Any of the following: Size: >3 cm but <5 cm, local invasion:

	visceral pleura, airway location: invasion of the main bronchus or presence of atelectasis or obstructive.
T3	Any of the following: Size: >5 cm but <7 cm, local invasion: direct invasion of the chest wall, parietal pleura, phrenic nerve, parietal pericardium. Separate tumour nodule(s) in the same lobe of the primary tumour.
T4	Any of the following: Size: >7 cm, airway location: invasion of the carina or trachea, local invasion: diaphragm, mediastinum, heart, great vessels, recurrent laryngeal nerve, oesophagus or vertebral body. Separate tumour nodule(s) in an ipsilateral lobe of the primary tumour.

Lymph nodes (N)	
Category	Definition
Nx	Regional lymph nodes cannot be evaluated.
N0	No regional lymph nodes involved.
N1	Involvement of ipsilateral parabronchial and/or ipsilateral hila lymph nodes.
N2	Involvement of the ipsilateral mediastinal and/or subcarinal lymph nodes.
N3	Involvement of any following lymph nodes: contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene or supraclavicular nodes.

Distant metastasis (M)	
Category	Definition
M0	No distant metastasis.
M1	Presence of distant metastasis. Subdivisions: M1a: separate tumour nodule(s) in a contralateral lobe to that of primary tumours. M1b: single extra-thoracic metastasis.

	M1c: multiple extra-thoracic metastasis to one or more organs.

Stage group	
Stage	TNM categories
Occult carcinoma	(TxN0M0)
Stage 0	(TisN0M0)
Stage IA1	(T1aN0M0) (T1(mi)N0M0)
Stage IA2	(T1bN0M0)
Stage IA3	(T1cN0M0)
Stage IB	(T2aN0M0)
Stage IIA	(T2bN0M0)
Stage IIB	(T (1–2)N1M0) (T3N0M0)
Stage IIIA	(T(1–2)N2M0) (T3N1M0) (T4N(0–1)M0)
Stage IIIB	(T(1–2)N3M0) (T(3–4)N2M0)
Stage IIIC	(T(3–4)N3M0)
Stage IVA	(Any T, Any N, M1a,b)
Stage IVB	(Any T, Any N, M1c)

Appendix 2. TNM staging of breast cancer.

Primary tumour (T)	
Category	Descriptor
Tx	The primary tumour cannot be evaluated.
T0	No evidence of a primary tumour.
Tis	Carcinoma in situ: DCIS or Paget's
T1	Size: ≤2 cm 4 substages based on tumour size: T1mi: ≤ 1mm T1a: >1 mm but ≤5 mm T1b: >5mm but ≤1 cm T1c: >1 cm but ≤2 cm
T2	Size: >2 cm but <5 cm
T3	Size: >5 cm
T4	Any of the following: T4a: tumour has grown into the chest wall. T4b: tumour has grown into the skin. T4c: tumour has grown into the chest wall and skin. T4d: inflammatory breast cancer.

Lymph nodes (N)	
Category	Definition
Nx	Regional lymph nodes cannot be evaluated.
N0	No regional lymph nodes involved or areas of cancer smaller than 0.2 mm in lymph nodes.
N1	Presence of cancer in 1-3 axillary lymph nodes and/or mammary lymph nodes. Cancer in the lymph node is >0.2 mm but ≤2 mm.
N2	Presence of cancer in 4-9 axillary lymph nodes or it has spread to internal mammary lymph nodes but not the axillary lymph nodes.
N3	Presence of cancer in ≥10 axillary lymph nodes and/or presence under the clavicle or collarbone. Cancer may have also spread to the internal mammary lymph nodes.

Distant metastasis (M)	
Category	Definition
MX	Distant spread cannot be evaluated.
M0	No distant metastasis.
M1	Presence of distant metastasis in another part of the body.

Appendix 3. Publications

First author

Dowling, Lewis M., et al. “Optimization of Sample Preparation Using Glass Slides for Spectral Pathology.” Applied Spectroscopy, vol. 75, no. 3, SAGE Publications Inc., Mar. 2021, pp. 343–50, doi:10.1177/0003702820945748.

Contributing author

Kansiz, Mustafa, et al. “Optical Photothermal Infrared Microspectroscopy Discriminates for the First Time Different Types of Lung Cells on Histopathology Glass Slides.” Analytical Chemistry, vol. 93, no. 32, American Chemical Society, Aug. 2021, pp. 11081–88, doi:10.1021/ACS.ANALCHEM.1C00309.

Xie, B., Njoroge, W., Dowling, L. M., Sulé-Suso, J., Cinque, G., & Yang, Y. (2022). Detection of lipid efflux from foam cell models using a label-free infrared method. The Analyst, 147(23). https://doi.org/10.1039/D2AN01041K

References

Alba-Bernal, A. *et al.* (2020) 'Challenges and achievements of liquid biopsy technologies employed in early breast cancer', *EBioMedicine*, 62. Available at: https://doi.org/10.1016/j.ebiom.2020.103100.

Almansour, N.M. (2022) 'Triple-Negative Breast Cancer: A Brief Review About Epidemiology, Risk Factors, Signaling Pathways, Treatment and Role of Artificial Intelligence', *Frontiers in Molecular Biosciences*, 9, p. 32. Available at: https://doi.org/10.3389/FMOLB.2022.836417/BIBTEX.

Ammanagi, A.S. *et al.* (2012) 'Sputum cytology in suspected cases of carcinoma of lung (Sputum cytology a poor man's bronchoscopy!)', *Lung India*, 29(1), pp. 19–23. Available at: https://doi.org/10.4103/0970-2113.92356.

Andree, K.C., van Dalum, G. and Terstappen, L.W.M.M. (2016) 'Challenges in circulating tumor cell detection by the CellSearch system', *Molecular Oncology*, 10(3), p. 395. Available at: https://doi.org/10.1016/J.MOLONC.2015.12.002.

Asomaning, K. *et al.* (2008) 'Second hand smoke, age of exposure and lung cancer risk', *Lung cancer (Amsterdam, Netherlands)*, 61(1), p. 13. Available at: https://doi.org/10.1016/J.LUNGCAN.2007.11.013.

Backhaus, J. *et al.* (2010) 'Diagnosis of breast cancer with infrared spectroscopy from serum samples', *Vibrational Spectroscopy*, 52(2), pp. 173–177. Available at: https://doi.org/10.1016/j.vibspec.2010.01.013.

Bailey-Wilson, J.E. *et al.* (2004) 'A major lung cancer susceptibility locus maps to chromosome 6q23-25', *American Journal of Human Genetics*, 75(3), pp. 460–474. Available at: https://doi.org/10.1086/423857.

Baker, M.J. *et al.* (2009) 'Investigating FTIR based histopathology for the diagnosis of prostate cancer', *Journal of Biophotonics*, 2(1–2), pp. 104–113. Available at: https://doi.org/10.1002/jbio.200810062.

Baker, M.J. *et al.* (2014) 'Using Fourier transform IR spectroscopy to analyze biological materials.', *Nature protocols*, 9(8), pp. 1771–91. Available at: https://doi.org/10.1038/nprot.2014.110.

Bakir, G. *et al.* (2020a) 'Orientation Matters: Polarization Dependent IR Spectroscopy of Collagen from Intact Tendon Down to the Single Fibril Level', *Molecules 2020, Vol. 25, Page 4295*, 25(18), p. 4295. Available at: https://doi.org/10.3390/MOLECULES25184295.

Bakir, G. *et al.* (2020b) 'Orientation Matters: Polarization Dependent IR Spectroscopy of Collagen from Intact Tendon Down to the Single Fibril Level', *Molecules 2020, Vol. 25, Page 4295*, 25(18), p. 4295. Available at: https://doi.org/10.3390/MOLECULES25184295.

Bassan, P. *et al.* (2009) 'Resonant Mie scattering in infrared spectroscopy of biological materials - Understanding the "dispersion artefact"', *Analyst*, 134(8), pp. 1586–1593. Available at: https://doi.org/10.1039/b904808a.

Best, M.G. *et al.* (2015) 'RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics', *Cancer Cell*, 28(5), pp. 666–676. Available at: https://doi.org/10.1016/j.ccell.2015.09.018.

Breast cancer statistics | Cancer Research UK (no date). Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer (Accessed: 20 February 2023).

Breast screening | Breast cancer | Cancer Research UK (no date). Available at: https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/screening/breast-screening (Accessed: 21 February 2023).

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324/METRICS.

Brennan, P. M., Butler, H. J., Christie, L., Hegarty, M. G., Jenkinson, M. D., Keerie, C., Norrie, J., O'brien, R., Palmer, D. S., Smith, B. R., & Baker, M. J. (n.d.). Early diagnosis of brain tumours using a novel spectroscopic liquid biopsy. https://doi.org/10.1093/braincomms/fcab056

Bubendorf, L. *et al.* (2017) 'Nonsmall cell lung carcinoma: Diagnostic difficulties in small biopsies and cytological specimens', *European Respiratory Review*, 26(144). Available at: https://doi.org/10.1183/16000617.0007-2017.

Cancer Research UK (no date) *Lung cancer statistics | Cancer Research UK*. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer#heading-Two (Accessed: 7 October 2019).

Cardoso, F. *et al.* (2019) 'Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†', *Annals of oncology : official journal of the European Society for Medical Oncology*, 30(8), pp. 1194–1220. Available at: https://doi.org/10.1093/ANNONC/MDZ173.

Cheeseman, S. *et al.* (2019) 'Applications of Synchrotron-Source IR Spectroscopy for the Investigation of Insect Wings', in *Synchrotron Radiation - Useful and Interesting Applications*. IntechOpen. Available at: https://doi.org/10.5772/intechopen.84591.

Childs, D.T.D. *et al.* (2015) 'Sensitivity Advantage of QCL Tunable-Laser Mid-Infrared Spectroscopy Over FTIR Spectroscopy', *http://dx.doi.org/10.1080/05704928.2015.1075208*, 50(10), pp. 822–839. Available at: https://doi.org/10.1080/05704928.2015.1075208.

Chung, A. *et al.* (2015) 'Impact of Consensus Guidelines by the Society of Surgical Oncology and the American Society for Radiation Oncology on Margins for Breast-Conserving Surgery in Stages 1 and 2 Invasive Breast Cancer', *Annals of Surgical Oncology*, 22(3), pp. 422–427. Available at: https://doi.org/10.1245/S10434-015-4829-0/METRICS.

Diem, M., Romeo, M., *et al.* (2004) 'A decade of vibrational micro-spectroscopy of human cells and tissue (1994-2004)', in *Analyst*. NIH Public Access, pp. 880–885. Available at: https://doi.org/10.1039/b408952a.

Diem, M., Romeo, Melissa, *et al.* (2004) 'Comparison of Fourier transform infrared (FTIR) spectra of individual cells acquired using synchrotron and conventional sources'. Available at: https://doi.org/10.1016/j.infrared.2004.01.013.

Dowling, L. *et al.* (2020) 'Optimization of Sample Preparation Using Glass Slides for Spectral Pathology.', *Applied spectroscopy*, p. 3702820945748. Available at: https://doi.org/10.1177/0003702820945748.

Falamas, A., Faur, C. I., Ciupe, S., Chirila, M., Rotaru, H., Hedesiu, M., & Cinta Pinzaru, S. (2021). Rapid and noninvasive diagnosis of oral and oropharyngeal cancer based on micro-

Raman and FT-IR spectra of saliva. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 252, 119477. https://doi.org/10.1016/J.SAA.2021.119477

Ferreira, I. C. C., Aguiar, E. M. G., Silva, A. T. F., Santos, L. L. D., Cardoso-Sousa, L., Araújo, T. G., Santos, D. W., Goulart, L. R., Sabino-Silva, R., Maia, Y. C. P., & Li, C. J. (2020). Attenuated Total Reflection-Fourier Transform Infrared (ATR-FTIR) Spectroscopy Analysis of Saliva for Breast Cancer Diagnosis. Journal of Oncology, 2020. https://doi.org/10.1155/2020/4343590

Finlayson, D., Rinaldi, C. and Baker, M.J. (2019) 'Is Infrared Spectroscopy Ready for the Clinic?', *Analytical Chemistry*, 91(19), pp. 12117–12128. Available at: https://doi.org/10.1021/acs.analchem.9b02280.

Gohari, A. and Haramati, L.B. (2004) 'Complications of CT scan-guided lung biopsy: Lesion size and depth matter', *Chest*. American College of Chest Physicians, pp. 666–668. Available at: https://doi.org/10.1378/chest.126.3.666.

Gok, S. *et al.* (2016) 'Bladder cancer diagnosis from bladder wash by Fourier transform infrared spectroscopy as a novel test for tumor recurrence', *Journal of biophotonics*, 9(9), pp. 967–975. Available at: https://doi.org/10.1002/jbio.201500322.

Habli, Z. *et al.* (2020) 'Circulating Tumor Cell Detection Technologies and Clinical Utility: Challenges and Opportunities', *Cancers*, 12(7), pp. 1–30. Available at: https://doi.org/10.3390/CANCERS12071930.

Hammond, M.E.H. *et al.* (2010) 'American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer', *Journal of clinical oncology : official journal of the*

American Society of Clinical Oncology, 28(16), pp. 2784–2795. Available at: https://doi.org/10.1200/JCO.2009.25.6529.

Hermes, M. *et al.* (2018) 'Journal of Optics TOPICAL REVIEW • OPEN ACCESS Mid-IR hyperspectral imaging for label-free histopathology and cytology'. Available at: https://doi.org/10.1088/2040-8986/aaa36b.

Hirsch, F.R. *et al.* (2017) 'Lung cancer: current therapies and new targeted treatments', *The Lancet*. Lancet Publishing Group, pp. 299–311. Available at: https://doi.org/10.1016/S0140-6736(16)30958-8.

Huang, Z. (2013) *BRIGHTNESS AND COHERENCE OF SYNCHROTRON RADIATION AND FELs **.

Inamura, K. (2017) 'Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification', *Frontiers in Oncology*, 7, p. 193. Available at: https://doi.org/10.3389/FONC.2017.00193.

Jonathan Yang, T. and Ho, A.Y. (2013) 'Radiation therapy in the management of breast cancer', *The Surgical clinics of North America*, 93(2), pp. 455–471. Available at: https://doi.org/10.1016/J.SUC.2013.01.002.

Kansiz, M. and Prater, C.B. (2020) 'Super resolution correlative far-field submicron simultaneous IR and raman microscopy: a new paradigm in vibrational spectroscopy', *https://doi.org/10.1117/12.2565489*, 11252, p. 112520E. Available at: https://doi.org/10.1117/12.2565489.

Khanmohammadi, M. *et al.* (2011) 'Application of linear discriminant analysis and attenuated total reflectance fourier transform infrared microspectroscopy for diagnosis of colon cancer',

Pathology and Oncology Research, 17(2), pp. 435–441. Available at: https://doi.org/10.1007/s12253-010-9326-y.

Klementieva, O. *et al.* (2020) 'Super-Resolution Infrared Imaging of Polymorphic Amyloid Aggregates Directly in Neurons', *Advanced Science*, 7(6), p. 1903004. Available at: https://doi.org/10.1002/ADVS.201903004.

Kotsiantis, S. *et al.* (2014) 'Machine learning: A review of classification and combining techniques Machine Learning and Data Mining View project Metaheurestic Optimization in Machine Learning View project Machine learning: a review of classification and combining techniques'. Available at: https://doi.org/10.1007/s10462-007-9052-3.

Kumar, S., Srinivasan, A., & Nikolajeff, F. (2018). Role of Infrared Spectroscopy and Imaging in Cancer Diagnosis. Current Medicinal Chemistry, 25(9), 1055–1072. https://doi.org/10.2174/0929867324666170523121314

Kwak, J.T. *et al.* (2011) 'Multimodal microscopy for automated histologic analysis of prostate cancer', *BMC Cancer*, 11. Available at: https://doi.org/10.1186/1471-2407-11-62.

Kyriakidou, M. *et al.* (2017) 'FT-IR spectroscopy study in early diagnosis of skin cancer', *In Vivo*, 31(6), pp. 1131–1137. Available at: https://doi.org/10.21873/invivo.11179.

Lasalvia, M., Capozzi, V. and Perna, G. (2021) 'Discrimination of Different Breast Cell Lines on Glass Substrate by Means of Fourier Transform Infrared Spectroscopy', *Sensors 2021, Vol. 21, Page 6992*, 21(21), p. 6992. Available at: https://doi.org/10.3390/S21216992.

Lazaro-Pacheco, D., Shaaban, A., Baldwin, G., Titiloye, N. A., Rehman, S., & Rehman, I. ur. (2020). Deciphering the structural and chemical composition of breast cancer using FTIR spectroscopy. Https://Doi.Org/10.1080/05704928.2020.1843471, 57(3), 234–248.

Lemjabbar-Alaoui, H. *et al.* (2015) 'Lung cancer: Biology and treatment options', *Biochimica et Biophysica Acta - Reviews on Cancer*. Elsevier B.V., pp. 189–210. Available at: https://doi.org/10.1016/j.bbcan.2015.08.002.

Lewis, P.D. *et al.* (2010) 'Evaluation of FTIR Spectroscopy as a diagnostic tool for lung cancer using sputum', *BMC Cancer*, 10. Available at: https://doi.org/10.1186/1471-2407-10-640.

Lima, K.M.G. *et al.* (2015) 'Segregation of ovarian cancer stage exploiting spectral biomarkers derived from blood plasma or serum analysis: ATR-FTIR spectroscopy coupled with variable selection methods', *Biotechnology Progress*, 31(3), pp. 832–839. Available at: https://doi.org/10.1002/btpr.2084.

Liu, J. *et al.* (2021) 'Extracellular Vesicles in Liquid Biopsies: Potential for Disease Diagnosis', *BioMed Research International*, 2021. Available at: https://doi.org/10.1155/2021/6611244.

Liu, J., Cheng, H., Lv, X., Zhang, Z., Zheng, X., Wu, G., Tang, J., Ma, X., & Yue, X. (2020). Use of FT-IR spectroscopy combined with SVM as a screening tool to identify invasive ductal carcinoma in breast cancer. https://doi.org/10.1016/j.ijleo.2020.164225.

Lone, S.N. *et al.* (2022) 'Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments', *Molecular Cancer 2022 21:1*, 21(1), pp. 1–22. Available at: https://doi.org/10.1186/S12943-022-01543-7.

Łukasiewicz, S. *et al.* (2021) 'Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review', *Cancers*, 13(17). Available at: https://doi.org/10.3390/CANCERS13174287.

Ma, M. *et al.* (2015) '"Liquid biopsy"—ctDNA detection with great potential and challenges', *Annals of Translational Medicine*, 3(16), p. 235. Available at: https://doi.org/10.3978/J.ISSN.2305-5839.2015.09.29.

Malhotra, J. *et al.* (2016) 'Risk factors for lung cancer worldwide', *European Respiratory Journal*, 48(3), pp. 889–902. Available at: https://doi.org/10.1183/13993003.00359-2016.

Marmot, M. *et al.* (2012a) 'The benefits and harms of breast cancer screening: An independent review', *The Lancet*, 380(9855), pp. 1778–1786. Available at: https://doi.org/10.1016/S0140-6736(12)61611-0.

Marmot, M. *et al.* (2012b) 'The benefits and harms of breast cancer screening: an independent review', *Lancet (London, England)*, 380(9855), pp. 1778–1786. Available at: https://doi.org/10.1016/S0140-6736(12)61611-0.

Matikas, A. *et al.* (2022) 'Detection of circulating tumour cells before and following adjuvant chemotherapy and long-term prognosis of early breast cancer', *British Journal of Cancer 2022 126:11*, 126(11), pp. 1563–1569. Available at: https://doi.org/10.1038/s41416-022-01699-5.

Maximiano, S. *et al.* (2016) 'Trastuzumab in the Treatment of Breast Cancer', *BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy*, 30(2), pp. 75–86. Available at: https://doi.org/10.1007/S40259-016-0162-9.

Menzies, G.E. *et al.* (2014) 'Fourier transform infrared for noninvasive optical diagnosis of oral, oropharyngeal, and laryngeal cancer', *Translational Research*, 163(1), pp. 19–26. Available at: https://doi.org/10.1016/j.trsl.2013.09.006.

Mitra, S. and Dey, P. (2016) 'Fine-needle aspiration and core biopsy in the diagnosis of breast lesions: A comparison and review of the literature', *CytoJournal*, 13(1). Available at: https://doi.org/10.4103/1742-6413.189637.

Mostaço-Guidolin, L. B., Murakami, L. S., Batistuti, M. R., Nomizo, A., & Bachmann, L. (2010). Molecular and chemical characterization by Fourier transform infrared spectroscopy of human breast cancer cells with estrogen receptor expressed and not expressed. Journal of Spectroscopy, 24(5), 501–510. https://doi.org/10.3233/SPE-2010-0466.

Mulvaney, S.P. and Keating, C.D. (2000) 'Raman Spectroscopy'. Available at: https://doi.org/10.1021/a10000155.

Nasim, F., Sabath, B.F. and Eapen, G.A. (2019) 'Lung Cancer', *Medical Clinics of North America*. W.B. Saunders, pp. 463–473. Available at: https://doi.org/10.1016/j.mcna.2018.12.006.

Ohashi, R. *et al.* (2016) 'Diagnostic value of fine needle aspiration and core needle biopsy in special types of breast cancer', *Breast cancer (Tokyo, Japan)*, 23(4), pp. 675–683. Available at: https://doi.org/10.1007/S12282-015-0624-9.

Ollesch, J., Heinze, M., Heise, H. M., Behrens, T., Brüning, T., & Gerwert, K. (2014). It's in your blood: spectral biomarker candidates for urinary bladder cancer from automated FTIR spectroscopy. Journal of Biophotonics, 7(3–4), 210–221. https://doi.org/10.1002/JBIO.201300163.

Pallua, J.D. *et al.* (2018) 'Clinical infrared microscopic imaging: An overview', *Pathology - Research and Practice*, 214(10), pp. 1532–1538. Available at: https://doi.org/10.1016/J.PRP.2018.08.026.

Paraskevaidi, M. *et al.* (2018) 'Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer', *Analyst*, 143(13), pp. 3156–3163. Available at: https://doi.org/10.1039/c8an00027a.

Paulus, A. *et al.* (2021) 'Amyloid structural changes studied by infrared microspectroscopy in bigenic cellular models of alzheimer's disease', *International Journal of Molecular Sciences*, 22(7). Available at: https://doi.org/10.3390/IJMS22073430/S1.

Pereira de Souza, N. M., Machado, B. H., Padoin, L. V., Prá, D., Fay, A. P., Corbellini, V. A., & Rieger, A. (2023). Rapid and low-cost liquid biopsy with ATR-FTIR spectroscopy to discriminate the molecular subtypes of breast cancer. Talanta, 254, 123858. https://doi.org/10.1016/J.TALANTA.2022.123858.

Petrucelli, N., Daly, M.B. and Pal, T. (2022) 'BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer', *GeneReviews®* [Preprint]. Available at: https://www.ncbi.nlm.nih.gov/books/NBK1247/ (Accessed: 22 March 2023).

Pijanka, J. *et al.* (2010) 'Synchrotron-based FTIR spectra of stained single cells. Towards a clinical application in pathology', *Laboratory Investigation*, 90(5), pp. 797–807. Available at: https://doi.org/10.1038/labinvest.2010.8.

Pilling, M.J. *et al.* (2017) 'Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation', *Analyst*, 142(8), pp. 1258–1268. Available at: https://doi.org/10.1039/c6an02224c.

Pommier, R.M. *et al.* (2020) 'Comprehensive characterization of claudin-low breast tumors reflects the impact of the cell-of-origin on cancer evolution', *Nature Communications 2020 11:1*, 11(1), pp. 1–12. Available at: https://doi.org/10.1038/s41467-020-17249-7.

Powell, H.A. *et al.* (2013) 'Chronic obstructive pulmonary disease and risk of lung cancer: The importance of smoking and timing of diagnosis', *Journal of Thoracic Oncology*, 8(1), pp. 6–11. Available at: https://doi.org/10.1097/JTO.0b013e318274a7dc.

Rushton, L. *et al.* (2012) 'Occupational cancer burden in Great Britain STRUCTURE OF THE SUPPLEMENT', *British Journal of Cancer*, 107, pp. 3–7. Available at: https://doi.org/10.1038/bjc.2012.112.

Rutter, A. V. *et al.* (2018) 'Fourier transform infrared spectra of cells on glass coverslips. A further step in spectral pathology', *The Analyst*, 143(23), pp. 5711–5717. Available at: https://doi.org/10.1039/C8AN01634H.

Rutter, A. V *et al.* (2019) 'Identification of a Glass Substrate to Study Cells Using Fourier Transform Infrared Spectroscopy: Are We Closer to Spectral Pathology?', *Applied spectroscopy*, p. 3702819875828. Available at: https://doi.org/10.1177/0003702819875828.

Ryan, C. and Burke, L. (2017) 'Pathology of lung tumours', *Surgery (United Kingdom)*. Elsevier Ltd, pp. 234–242. Available at: https://doi.org/10.1016/j.mpsur.2017.02.002.

Santillan, A.A., Camargo, C.A. and Colditz, G.A. (2003) 'A meta-analysis of asthma and risk of lung cancer (United States).', *Cancer causes & control : CCC*, 14(4), pp. 327–34. Available at: https://doi.org/10.1023/a:1023982402137.

Sitnikova, V. E., Kotkova, M. A., Nosenko, T. N., Kotkova, T. N., Martynova, D. M., & Uspenskaya, M. V. (2020). Breast cancer detection by ATR-FTIR spectroscopy of blood serum and multivariate data-analysis. Talanta, 214, 120857. https://doi.org/10.1016/J.TALANTA.2020.120857.

Soda, M. *et al.* (2007) 'Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer', *Nature*, 448(7153), pp. 561–566. Available at: https://doi.org/10.1038/nature05945.

Spadea, A. *et al.* (2021) 'Analysis of Fixed and Live Single Cells Using Optical Photothermal Infrared with Concomitant Raman Spectroscopy', *Cite This: Anal. Chem*, 93, p. 3950. Available at: https://doi.org/10.1021/acs.analchem.0c04846.

Statistics » Provider-based Cancer Waiting Times for December 2022 – 23 (Provisional) (no date). Available at: https://www.england.nhs.uk/statistics/statistical-work-areas/cancer-waiting-times/monthly-prov-cwt/2022-23-monthly-provider-cancer-waiting-times-statistics/provider-based-cancer-waiting-times-for-december-2022-23-provisional/ (Accessed: 6 March 2023).

Steenland, K. *et al.* (2001) 'Pooled exposure-response analyses and risk assessment for lung cancer in 10 cohorts of silica-exposed workers: An IARC multicentre study', *Cancer Causes and Control*, 12(9), pp. 773–784. Available at: https://doi.org/10.1023/A:1012214102061.

Su, K.-Y. and Lee, W.-L. (2020) 'Fourier Transform Infrared Spectroscopy as a Cancer Screening and Diagnostic Tool: A Review and Prospects', *Cancers*, 12(1), p. 115. Available at: https://doi.org/10.3390/cancers12010115.

Sundling, K.E. and Lowe, A.C. (2019) 'Circulating Tumor Cells: Overview and Opportunities in Cytology', *Advances in Anatomic Pathology*, 26(1), pp. 56–63. Available at: https://doi.org/10.1097/PAP.0000000000000217.

Tomas, R. C., Sayat, A. J., Atienza, A. N., Danganan, J. L., Ramos, M. R., Fellizar, A., Israel, K. N., Angeles, L. M., Bangaoil, R., Santillan, A., & Albano, P. M. (2022). Detection of breast

cancer by ATR-FTIR spectroscopy using artificial neural networks. PLoS ONE, 17(1). https://doi.org/10.1371/JOURNAL.PONE.0262489.

Turner, M.C. *et al.* (2007) 'Chronic obstructive pulmonary disease is associated with lung cancer mortality in a prospective study of never smokers', *American Journal of Respiratory and Critical Care Medicine*, 176(3), pp. 285–290. Available at: https://doi.org/10.1164/rccm.200612-1792OC.

Vajpeyi, R. (2005) 'WHO Classification of Tumours: Pathology and Genetics of Tumours of the Breast and Female Genital Organs', *Journal of Clinical Pathology*, 58(6), p. 671. Available at: /pmc/articles/PMC1770678/ (Accessed: 20 February 2023).

In 'T Veld, S.G.J.G. and Wurdinger, T. (2019) 'Tumor-educated platelets', *Blood*, 133(22), pp. 2359–2364. Available at: https://doi.org/10.1182/BLOOD-2018-12-852830.

Wang, H.P., Wang, H.C. and Huang, Y.J. (1997) 'Microscopic FTIR studies of lung cancer cells in pleural fluid', *Science of the Total Environment*, 204(3), pp. 283–287. Available at: https://doi.org/10.1016/S0048-9697(97)00180-0.

Wardwell, N.R. and Massion, P.P. (2005) 'Novel strategies for the early detection and prevention of lung cancer', *Seminars in Oncology*, 32(3 SOPEC. ISS), pp. 259–268. Available at: https://doi.org/10.1053/j.seminoncol.2005.02.009.

Williams, C. and Lin, C.Y. (2013) 'Oestrogen receptors in breast cancer: basic mechanisms and clinical implications', *Ecancermedicalscience*, 7(1). Available at: https://doi.org/10.3332/ECANCER.2013.370.

Yano, K. *et al.* (2000) 'Direct measurement of human lung cancerous and noncancerous tissues by Fourier transform infrared microscopy: Can an infrared microscope be used as a

clinical tool?', *Analytical Biochemistry*, 287(2), pp. 218–225. Available at: https://doi.org/10.1006/abio.2000.4872.

Yang, X., Ou, Q., Qian, K., Yang, J., Bai, Z., Yang, W., Shi, Y., & Liu, G. (2021). Diagnosis of Lung Cancer by ATR-FTIR Spectroscopy and Chemometrics. Frontiers in Oncology, 11. https://doi.org/10.3389/FONC.2021.753791.

Yao, H., Shi, X. and Zhang, Y. (2014) 'The use of FTIR-ATR spectrometry for evaluation of surgical resection margin in colorectal cancer: A pilot study of 56 samples', *Journal of Spectroscopy*, 2014. Available at: https://doi.org/10.1155/2014/213890.

Zappa, C. and Mousa, S.A. (2016) 'Non-small cell lung cancer: current treatment and future advances.', *Translational lung cancer research*, 5(3), pp. 288–300. Available at: https://doi.org/10.21037/tlcr.2016.06.07.

Zelig, U. *et al.* (2011) 'Pre-screening and follow-up of childhood acute leukemia using biochemical infrared analysis of peripheral blood mononuclear cells', *Biochimica et Biophysica Acta - General Subjects*, 1810(9), pp. 827–835. Available at: https://doi.org/10.1016/j.bbagen.2011.06.010.

Zheng, W. *et al.* (1987) 'Lung cancer and prior tuberculosis infection in Shanghai', *British Journal of Cancer*, 56(4), pp. 501–504. Available at: https://doi.org/10.1038/bjc.1987.233.

Annex: Letter of ethical approval

Keele University FMHS Faculty Research Ethics Committee
health.ethics@keele.ac.uk
g.p.j.moss@keele.ac.uk

9th November 2021

Dear Prof Sule-Suso

REC Project Reference:	MH-210190
Type of Application	Main application

Keele University's Faculty of Medicine and Health Sciences Research Ethics Committee (FMHS FREC) reviewed the above project application.

Final Opinion
Thank you for summarising the amendments in a detailed but extremely clear manner. The FMHS FREC can now recommend that this study receives a **Favourable Ethical Opinion**.

Conditions / recommendations:
There are no **conditions** attached to this application. There are, however, standard reporting requirements to consider, below:

Reporting requirements
The University's standard operating procedures give detailed guidance on reporting requirements for studies with a favourable opinion including:

- Notifying the relevant FREC of substantial amendments to an approved study
- Notifying the relevant FREC of issues which may have an impact upon ethical opinion of the study
- Progress reports
- Notifying the relevant FREC of the end of the study

Documents reviewed
The documents reviewed were:

Document	Version	Date
All documents submitted with MH-210190 including revisions		

Yours sincerely,

Dr Gary Moss

Chair