The University of Liverpool

PhD Thesis

The Development of Machine Learning Methods for Head and Neck Cancer Prognosis

---

Author:

Conor Whitley

Supervisors:

Dr. David Martin

Dr. Steve Barrett

Professor Marta Garcia-Finana

Dr Ruwanthi Kolamunnage-Dona

Abstract

Contents

List of Figures

List of Tables

1 Introduction

In 2015, 90.5 million people worldwide were living with some form of cancer [1]. With around 8.8 million deaths a year, accounting for 15.7% of deaths worldwide attributed to cancer, [1] and 14.1 million new cases occurring each year [2] the incidence of cancer cases is set to increase.

In order to improve the prognoses of patients, the diagnosis must take place at an earlier stage so that effective treatment may be sought, and the progression of the disease limited. The need for an inexpensive, rapid, and accurate diagnosis method is one of the 'holy grails' of cancer detection. In addition to timely diagnosis an important factor in patient outcomes is the choice of treatment. Targeted therapeutic interventions can be utilised to target aggressive cancers where appropriate, however current methods of determining relevant cases have their shortcomings. A false-negative carries the risk of missing a case of metastasis, whereas a false-positive leads to an unnecessary lymph node dissection which can result in disfigurement, pain and other long-term consequences [3]. It has been hypothesised that a molecular fingerprint may exist which characterises patients with more aggressive cases of the disease [3].

The objective of this research project is to develop state-of-the-art classification models for use in the analysis of Fourier Transform Infra-Red (FTIR) spectra. These models will be designed with the purpose of aiding clinical decision makers in predicting the prognoses of head and neck cancers in order to direct

patients to more appropriate treatment. While the quality of the predictions attained by many of the models presented in this thesis are comparable to current biomarkers; these models would not be aimed to replace any current diagnostic methods entirely. This is due to the fact that many biomarkers can be used in unison to attain superior performance overall – this work seeks to augment these processes.

In cases where cancer has been identified in a patient, models to predict accurate prognoses can be used to direct the patient towards more appropriate treatment – potentially improving patient outcomes in the long term. The potential to develop such models using IR spectra as a prognostic biomarker is a relatively unexplored avenue of research and could be of significant value clinically.

In addition to the development of predictive models, the preprocessing steps which are necessary to facilitate reliable predictions will also be heavily covered due to the importance of these steps in the overall performance of the model. The evaluation of the performance of predictive models will also explained in-depth due to its complexity and importance with regards to medical diagnoses. An appropriate statistical methodology will also be given to ensure that any developed predictive tools are able to generalise well to a wider population and become an effective tool in a pathologists arsenal.

The insights gained from the process of developing a discriminatory tool for cancer prognostics can lead valuable insight into the mechanisms of cancer. Due to the ability of vibrational spectroscopy to tap into the underlying chemical moeities of a sample; any statistical model developed using this information can be interrogated, and pertinent information about the differences in chemical signature within pathological groups can be extracted and potentially used for other purposes.

1.1 Significance of the project

The light microscope is the standard instrument used in the examination of histological specimens. When supplemented by various staining agents like hematoxylin and eosin (H&E), areas of tissue rich in various chemical groups are highlighted. Hematoxylin is able to stain cell nuclei a blue/purple colour, and eosin will stain all other tissue structures in varying shades of pink – allowing histopathologists to discriminate between differing tissue types. In the majority of well-progressed cancer cases, this staining is sufficient enough for an experienced histopathologist to make an accurate diagnosis [4]. The information yielded by these techniques for use in diagnosis is strictly morphological, and whilst improving tissue contrast further – these methods barely scratch the surface of the information contained within the tissue samples.

The diagnosis of cancerous tissue through optical microscopy and staining is dogged by varying degrees of inter and intra-observer errors [5, 6] due to the subjective nature of some biomarkers. Even very skilled pathologists may disagree on particular samples which show a cancer in the early stages of dysplasia. However catching a cancer in these early stages is crucial for effective treatment and will result in better prognoses for the patient. The need for an objective analysis procedure has been known for some time, and attempts have been made to implement automated analysis procedures using both optical microscopy of H&E stained specimens [7, 8], and chemical imaging [9].

A key issue facing clinical decision makers is the determination of the optimal course of treatment for a patient dependent upon the progression of the disease. In cases where lower biological aggression is demonstrated, a de-escalation of therapy may be possible [10]. Identification of these cases is paramount to minimising the adverse effects of treatment, and improving patient outcomes. Previous work [11, 12, 13, 14] has hypothesised that tumours

which may be responsive to adjunctive therapeutic treatment may carry a distinct molecular fingerprint; the identification of which would facilitate screening of patients towards appropriate treatment. For approximately 50% of HPV negative head and neck squamous cell carcinoma patients, current treatment plans are ineffective. Neo-adjuvant therapy has the potential to improve prognoses, and aid clinical decision making if applicable cases can be determined in a timely manner.

1.2 Primary research questions

The primary goals of this project are to build upon existing techniques, and push the boundaries of knowledge and capabilities of existing methods. However there are a number of key points to be considered when developing a tool for clinical diagnosis perspective. Clinical diagnostic methods are subject to rigorous testing, and must achieve a number of milestones before transitioning to a clinical setting [15]. The scanning methods and data analysis developed throughout this project must surpass or supplement existing methods in terms of performance, but also pass the requirements expected of a clinical diagnostic test. The following non-exhaustive list must therefore be addressed for any diagnostic test to become clinically validated:

- Can the assay surpass existing diagnostic methods in terms of the relevant performance metrics? (e.g. accuracy, specificity, sensitivity.)
- Can relevant sources of error be identified and addressed?
- Is the test sensitive to the required range for its intended purpose?
- Is the test relatively cheap, and easy to use?

The new insights gained from these scanning techniques are not just suitable for clinical applications, the information contained within the samples will be of

interest to those studying cancer itself or other biological systems. This ”exploratory” perspective is directed more towards areas such as biomarker identification, imaging, and pattern finding [16].

1.3 The Structure of this thesis

This thesis contains a background section covering the necessary information to appreciate the current state-of-the-art research and relevant context to the following sections. The research conducted over the course of the PhD course is spread over three self-contained chapters covering three separate but connected bodies of work. There are many areas of research still requiring investigation when applying vibrational spectroscopy to clinical predictive tasks, the work covered in this thesis seeks to address some of these issues and present a novel approach to solving them.

Chapter **??** covers research into the development of a prognostic tool to risk stratify patients into one of two groups to direct treatment. Prognostic biomarkers are a relatively unexplored area of research in the context of vibrational spectroscopy aided clinical diagnostics. The objective of this chapter is to determine the efficacy of FTIR and α-Smooth Muscle Actin (ASMA) as prognostic variables and evaluate their suitability for a clinical setting.

Chapter **??** covers work undertaken to create an objective method of determining the best combination of preprocessing steps classifier algorithms according to key metrics. This is a widespread issue amongst vibrational spectroscopy and other multivariate classification techniques, as the number of potential configurations is large with the number of parameters associated with each step making the problem even more difficult. The optimisation framework was implemented on a cluster of computers situated within the university in order to

increase the efficiency of the process which has the potential to be implemented on any such system for wider use.

Chapter **??** seeks to demonstrate the possibilities associated with using deep learning. Deep learning is a subset of machine learning involving the use of neural networks with complex architectures for a multitude of purposes. The chapter covers the development, optimisation, and evaluation of two differing types of neural network architecture for use as prognostic tools.

Bibliography

[1] Health Data. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980âĂŞ2015: a systematic analysis for the global burden of disease study 2015., 2015.

[2] World Health Organization. World cancer report. Report, 2014.

[3] Xiaofeng Zhou, Stephane Temam, Myungshin Oh, Nisa Pungpravat, Bau Lin Huang, Li Mao, and David T. Wong. Global expression-based classification of lymph node metastasis and extracapsular spread of oral tongue squamous cell carcinoma. *Neoplasia*, 8(11):925–932, 2006.

[4] G. Orchard and B. Nation. *Histopathology*. Fundamentals of Biomedical Science. OUP Oxford, 2011.

[5] J. B. Lattouf and F. Saad. Gleason score on biopsy: Is it reliable for predicting the final grade on pathology? *BJU International*, 90(7):694–698, 2002.

[6] Daniel C. Paech, Adèle R. Weston, Nick Pavlakis, Anthony Gill, Narayan Rajan, Helen Barraclough, Bronwyn Fitzgerald, and Maximiliano Van Kooten. A systematic review of the interobserver variability for histology in the differentiation between squamous and nonsquamous non-small cell lung cancer. *Journal of Thoracic Oncology*, 6(1):55–63, 2011.

[7] T Araujo, G Aresta, E Castro, J Rouco, P Aguiar, C Eloy, A Polonia, and A Campilho. Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS One*, 12(6):1 – 14, 2017.

[8] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. 2017.

[9] Michael J. Pilling, Alex Henderson, Benjamin Bird, Mick D. Brown, Noel W. Clarke, and Peter Gardner. High-throughput quantum cascade laser

(QCL) spectral histopathology: a practical approach towards clinical translation. *Faraday Discuss.*, 187:135–154, 2016.

[10] Conor P. Barry, Chetan Katre, Elena Papa, James S. Brown, Richard J. Shaw, Fazilet Bekiroglu, Derek Lowe, and Simon N. Rogers. De-escalation of surgery for early oral cancer-is it oncologically safe? *British Journal of Oral and Maxillofacial Surgery*, 51(1):30–36, 2013.

[11] Rebekah K. O'Donnell, Michael Kupferman, S. Jack Wei, Sunil Singhal, Randal Weber, Bert O'Malley, Yi Cheng, Mary Putt, Michael Feldman, Barry Ziober, and Ruth J. Muschel. Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity. *Oncogene*, 24(7):1244–1251, 2005.

[12] Paul Roepman, Lodewyk F.A. Wessels, Nienke Kettelarij, Patrick Kemmeren, Antony J. Miles, Philip Lijnzaad, Marcel G.J. Tilanus, Ronald Koole, Gert Jan Hordijk, Peter C. Van Der Vliet, Marcel J.T. Reinders, Piet J. Slootweg, and Frank C.P. Holstege. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nature Genetics*, 37(2):182–186, 2005.

[13] D. S. Rickman, R. Millon, A. De Reynies, E. Thomas, C. Wasylyk, D. Muller, J. Abecassis, and B. Wasylyk. Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays. *Oncogene*, 27(51):6607–6622, 2008.

[14] Chenguang Zhao, Huiru Zou, Jun Zhang, Jinhui Wang, and Hao Liu. An integrated methylation and gene expression microarray analysis reveals significant prognostic biomarkers in oral squamous cell carcinoma. *Oncology Reports*, 40(5):2637–2647, 2018.

[15] Peter G Murphy. Selection of a suitable assay. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists*, 29 Suppl 1(August):S17–22, 2008.

[16] Júlio Trevisan, Plamen P. Angelov, Paul L. Carmichael, Andrew D. Scott, and Francis L. Martin. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *The Analyst*, 137(14):3202, 2012.

2 The physics approach to cancer diagnostics

The following section is intended to serve as a basic introduction to a number of key areas covered in this thesis. Additional detail will be given in later chapters where appropriate. Section 2.1 will cover some basic aspects of oncology and give an overview of the field of histology and biomarker discovery. An overview of the process of sample preparation and implications for measurements will then be discussed. Section 2.2 will cover topics relating to the experimental aspects of the thesis. The physical phenomena underpinning spectroscopy shall be explained with sections covering electronic hardware and data collection considerations being covered also. Data analysis techniques will be covered briefly in Section 2.3, with a focus on the underlying mechanics and evaluation of classification algorithms.

2.1 Cancer and Histopathology

2.1.1 Cancer

Cancer is the broad term given to a class of diseases which share the characteristics of abnormal cellular growth and a tendency to spread into surrounding tissue [1]. The first description of a breast cancer was recorded by an ancient Egyptian doctor in approximately 3000 BC [2] as a ”bulging tumour of the breast,

a grave disease — with no treatment". The Greek physician Galen noted the crab-like appearance of a solid cross-sectioned tumour and referred to tumours as *Karkinos*. The term Karkinos was later translated into Latin as *Cancer* — from where its modern name originates [2].

Cancer is the second leading cause of death after heart disease globally [3], with around 8.8 million deaths a year — accounting for 15.7% of deaths [4]. As cancer is an entire class of diseases, the specific symptons, causes, and treatments for each type of cancer vary widely. It has become necessary to develop specific treatments and diagnostic tests to account for the varying conditions and circumstances in which cancers are found.

In terms of their cause, different types of cancer are typically divided into one of two types according to their origin: those originating from genetic mutations triggered by environmental factors — amounting to approximately 90-95% [5] and those due to genetic origin accounting for the remaining 5-10% [5].

A cancer typically manifests itself in the form of a tumour or *neoplasm* — a collection of cells which exhibit signs of malignancy. The multi-step process which a cell undergoes when becoming cancerous is known as *Malignant progression*, this process is shown in fig. 2.1:

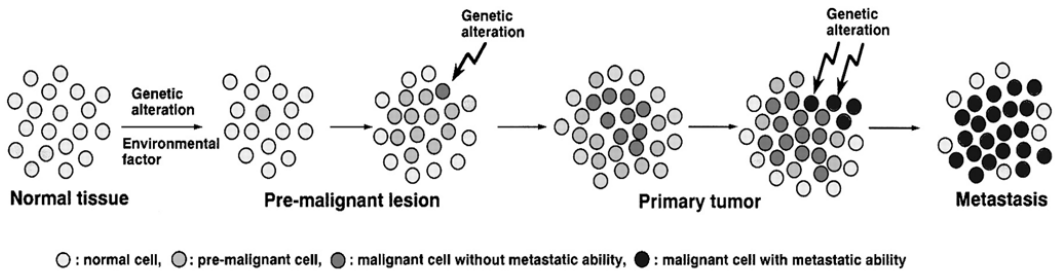

FIGURE 2.1: Malignant Progession. [6]

Malignant progression occurs in a tissue when tumour cells are present — which have undergone a series of genetic mutations, these tumour cells are characterised predominantly by the eight hallmarks of cancer and two enabling characteristics [7]:

2.1.2 The Hallmarks of cancer

Capabilities

As normal cells progress towards malignancy, a sequence of hallmarks are typically acquired. A tumour comprises a complex mass of numerous distinct cell types, interacting with each other in varying ways. Along side these malignant cells are normal cells which comprise a tumour-associated stroma. This tumour-associated stroma is not merely present but plays an active role in facilitating the acquisition of capabilities typical of cancers.

Cell growth and division without proper signalling A fundamental trait of cancer is the uncontrolled proliferation of cells. Normal cells carefully control growth-promoting signals which dictate progression through the cell growth-and-division cycle. It has been observed that cancer cells can acquire the ability to produce growth-factor ligands themselves, or influence associated stromal cells to provide these growth-factor signals [8]. Sources of proliferative signals situated within normal tissues are still not well understood [7]. This issue is further complicated by the fact that growth factor signals dictating cell growth are thought to be modulated temporally and spatially between a cell and its neighbours.

Unabated cell division in the presence of inhibitor signals In order to proliferate, cancerous cells must also avoid tumour suppressing signals. These tumour suppressant signals are moderated by two proteins: the *RB* (Retinoblastoma) protein [9] — which controls whether or not a cell shall proceed through its growth-and-division cycle, and *TP53* — which works similarly to RB but is dependent upon environmental factors within the cell such as levels of oxygenation and glucose. TP53 is sensitive to indicators of stress and damage within

tumour cells and is able to inhibit further cell-cycle progression until conditions return to normal. If conditions reach a point where damage is irreparable, TP53 can trigger cell apoptosis — cleansing defective cells. These two proteins form independent and redundant tumour suppressant systems. A cell must therefore suffer from a defect in the functioning of both systems to be prone to unabated cell division.

Avoidance of apoptosis Apoptosis is the highly-regulated process which a cell undergoes either when significant cell stress is detected from within the cell due to DNA damage, or when signalled to do so from other cells. In the case where apoptosis is triggered in a defective cell, those which have acquired the ability can avoid programmed death [10]. Tumour cells can resist apoptosis by becoming de-sensitised to internal and external signalling due to acquired mutations. The level of attenuation of apoptosis in tumours has been shown to be severe in tumours which are well-progressed [10].

Biological immortality Malignant cells differ from normal cells in their ability to circumvent the states of senesence (cell aging), and crisis (cell death) [11]. These processes avoid the uncontrolled proliferation of cells in the body, avoiding a ”hoarding” of nutrients by these cells and the potential to adversely affect surrounding tissue. A characteristic of cells which are biologically immune is the presence of telomerase which prohibits telomere shortening of chromosomes within the cell [7] and senesence and crisis from stopping uncontrolled proliferation. A large body of evidence suggests that the presence of telomerase allows for unlimited proliferation of cells — facilitating the growth of macroscopic tumours.

Construction of blood vessel networks (angiogenesis) As is the case for normal tissue, an adequate blood supply is required to provide the necessary

nutrients and oxygen, and remove waste products. Angiogensis is the process of creating this blood supply by activating an ”angiogenic switch” [12]. This is typically temporary for healthy adults, however this switch is permanently activated and continuously promotes the growth of vasculature to help support neoplastic tissue [7].

Invasion of surrounding tissue and metastases Typically occurring in later stages in the progression of cancer: metastases of malignant cells to neighbouring tissue sites and other organs of the body. These invasive and metastatic malignant cells are characterised typically by a change in shape and a reduction in E-cadherin — a key molecule in cell-to-cell adhesion [7].

Deregulation of metabolism A change in the metabolic processes favoured by malignant cells has been observed in many types of cancer [13]. These changes allow neoplastic cells to obtain larger amounts of energy to fuel cellular growth and division. Typically normal cells respire, converting glucose to ATP. However cancer cells have been observed to *reprogram* their glucose metabolism, resulting in a conversion to a state termed ”aerobic glycolysis” [7].

Evasion of the immune system For a tumour to grow it must have the capacity avoid detection by the immune system and to resist attacks from it. It is important to note that a number of cancers are induced by viruses, which a compromised immune system may struggle to eradicate. However, only $\sim$20% of tumours are virus-induced, the remaining $\sim$80% are thus able to overcome interference from the immmune system.

Enabling Characteristics

Genomic instability and mutation For a tumour to become established, the characteristics listed above must be acquired through a series of genetic changes. This happens gradually as individual cells acquiring these changes possess an advantage over neighbouring cells, making reproduction more likely and establishing a cancer lineage. Along side genetic changes, epigenetic changes such as DNA methylation and histone modifications [14]. In normal tissue mutation rates are usually low, however tumour cells can increase the level of mutation by increasing the sensitivity to mutagenic agents and by adverseley affecting systems which monitor genomic integrity.

Inflammation of surrounding tissue Other than innate the characteristics of neoplastic tissue, inflammation of tissue caused by varying degrees of an immune response can often have a counterproductive effect of promoting tumour growth. Inflammation can exacerbate and aid certain acquired characteristics by supplying molecules useful to the tumour to its local vicinity. An immune response can supply enzymes which modify the extracellular matrix allowing invasion, angiogenesis, and metastasis [7].

2.1.3 The tumour microenvironment

The environments in which neoplastic cells develop vary widely and will change over time. An exact prediction of exactly how a tumour will develop is not possible; it is dependent at least in part to the structural environment in which it resides and the interaction with other bodily systems. Figure 2.2 depicts typical tumour microenvironments in which neoplastic cells and their associated normal cells can be found.

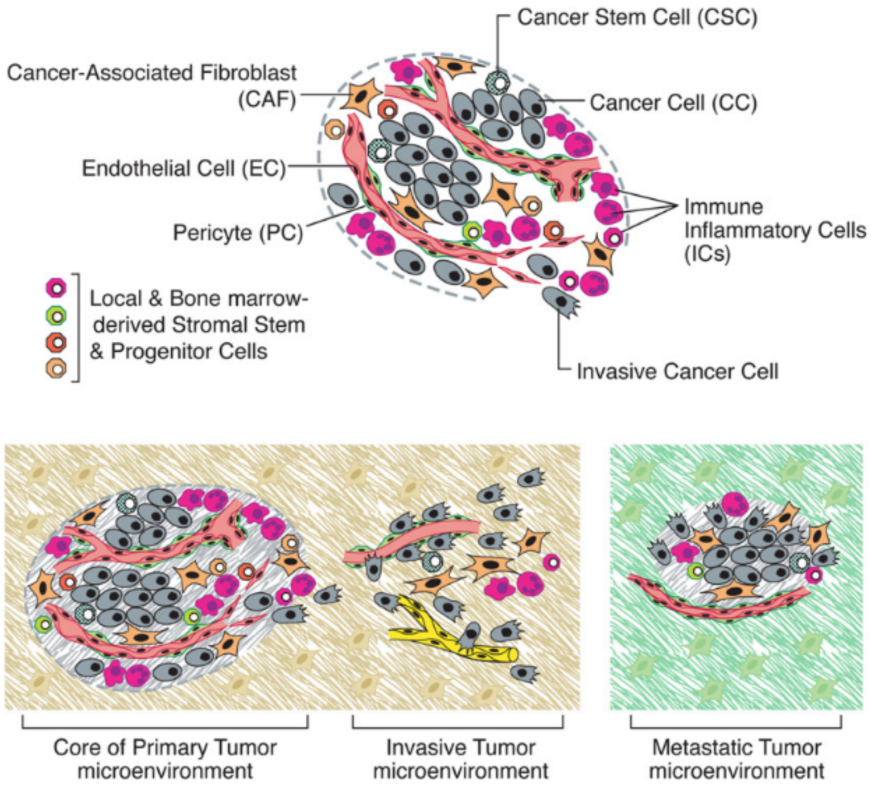

FIGURE 2.2: (Upper) The microenvironment of the tumour showing a complex array of interacting cell types. (Lower) Distinct microenvironments in which neoplastic cells are typically found. These environments develop progressively through the duration of the lineage of a collection of neoplastic cells [7].

The objective of the work in this thesis is effectively to observe variables associated with the tumour microenvironment. The variables in question vary according to the technique being utilised and are known informally as '-omics' e.g. genomics, proteomics, metabolomics etc.

2.1.4 Oral Cancer

Whilst the work covered in this thesis seeks to develop methods which are applicable to a range of diagnostic applications, the primary focus has been the development of diagnostic and prognostic tools for the treatment of oral cancer.

Oral cancer is characterised by the growth of tissue in various regions of the oral cavity, pharyngeal regions, and salivary glands. Oral cancer usually presents as an ulcer with fissuring or raised exophytic margins. It may also present as a lump, as a red lesion (erythro- plakia), as a white or mixed white and red lesion, as a non-healing extraction socket or as a cervical lymph node enlargement, characterized by hardness or fixation [15, 16].

Risk factors associated with oral cancer in the western world include tobacco and alcohol consumption; with 75% of all cases of oral cancer being associated with tobacco. In addition to tobacco and alcohol other risks factors such as betel quid chewing and various narcotics are associated with the development of oral cancer. Viral infections are also commonly associated with an increased chance of developing oral cancer. Human papilloma virus (HPV) is widely reported as a virus which carries oncogenetic potential, however results are conflicting as to the true extent of this potential [15]. It has been noted previously that some HPV genomes have been incorporated into oral cancer cells [16], but this has not yet proven to be a useful diagnostic variable when employed as a screening test [16]. Epstein-Barr virus [17] and Hepatitis C Virus (HCV) [18] are also considered to be viruses which oncogenetic potential through the influence of oncoproteins, these are however not known to follow oral cancer with a high incidence. The work covered in this thesis seeks to identify prognostic biomarkers in non virus-induced cancers as this is where the greatest clinical need lies.

The influence of an individuals genetics is also widely recognised as an influencing factor in the development of OSCC. Mice with a genetic prepdisposition to developing OSCC have been bred — suggesting a genetic causation. However the genetics of oral cancer are complex and a causative genetic link has not been firmly established in humans [16]. Some cancer predisposition

syndromes such as Li Fraumeni syndrome, and Fanconi anaemia have an increased prevalence of oral cancer, suggesting that p53 and DNA repair processes are important in this disease. This is thought to be due to major risk factors containing chemicals known to alter DNA.

Of particular relevance to the work contained in this thesis is the occurence of different pathological sites in which OSCC typically develops. The most common site of occurence of OSCC overall is the lower lip, with the most common site within the mouth being the tongue [16]. Within the oral cavity OSCC is particularly prevalent in the lower mouth, along the borders of the tongue, floor of the mouth, and adjoining areas. Whilst only comprising 20% of the area of the oral cavity, approximately 70% of oral cancers are known to occur in these regions.

2.1.5 Molecular Oncology

Molecular oncology is the interdisciplinary approach to cancer treatment which focuses on the effects of tumours at the molecular scale. As an interdiscplinary field molecular oncology frequently overlaps with chemistry and cytology and may be able to offer some level of insight into FTIR spectra. The ultimate goal of molecular oncology is to develop targeted therapies which can improve patient outcomes. However, the impact of molecular oncology is not solely limited to the development of therapies, as a large amount of effort is directed towards the prevention of cancer and the development of *molecular imaging* methods which may allow for the detection and study of malignant cells in situ [19].

The methods presented in this thesis may allow for such molecular imaging through the examination and FTIR microscopy images. Due to the ability of IR spectroscopy to access the chemical information contained in a sample, and in combination with imaging microscopes, FTIR microscopy could form the basis

for such a technology. A key limiting factor in the development of molecular imaging and cancer treatment is not the lack of target molecules, but in the limited resources available to dedicate to the pursuit. FTIR microscopy is a high-throughput, objective, and relatively inexpensive technology; in combination with vast data sets and an ever growing range of statistical techniques, it may be possible to expedite this process significantly.

2.1.6 Biomarker Discovery

The primary objective of a diagnostic or prognostic tool is to infer the presence or state of a disease; in order to accomplish this an indicator variable known as a *biomarker* is employed. A biomarker may come in many forms and can be considered any chemical, physical, or biological variable; the measurement of a biomarker can be molecular, cellular, biochemical or physiological in nature [20, 21]. Biomarkers may be present in any part of the body including bodily fluids such as blood serum, urine, cerebro-spinal fluid; biomarkers may also be found in any type of tissue situated in the body. A large proportion of currently used biomarkers are found in bodily fluids and are a common diagnostic tool employed by clinicians in a multitude of purposes.

Tissue biomarkers are those typically examined post biopsy after undergoing a series of steps to enable them to be viewed under an optical microscope. These methods are often supplemented using immunohistochemical stains in order to bring out contrast in desired regions of the image. In order to validate biomarkers for the clinic, a large volume of data is typically required to ensure that a biomarker generalises well to a larger patient cohort and is not solely a feature of a subset of patient data [21]. A Tissue Micro Array (TMA) is a collection of samples often taken from hundreds of biopsies using a needle punch biopsy arranged in a grid-like fashion, this is demonstrated in fig. 2.3.

In order for a biomarker to translate to a clinical setting, a number of requirements must be met. An ideal biomarker achieves the following:

- It is specifically associated with the presence or state of a disease, and is able to differentiate between similar physiological conditions.
- Standard biological sources can be used to observe the biomarker e.g. bodily fluids, tissue.
- The measurement of the biomarker must ideally be: quick, simple, accurate, and inexpensive.
- The biomarker is comparable to a measurable and standardised baseline reference.

Biomarkers must be discerned using statistically robust methods and offer a benefit to clinicians which justifies the cost of implementing the test. Any failings in a biomarkers ability to do so could lead to physicians to make decisions on a patients treatment which may be useless or detrimental to their wellbeing.

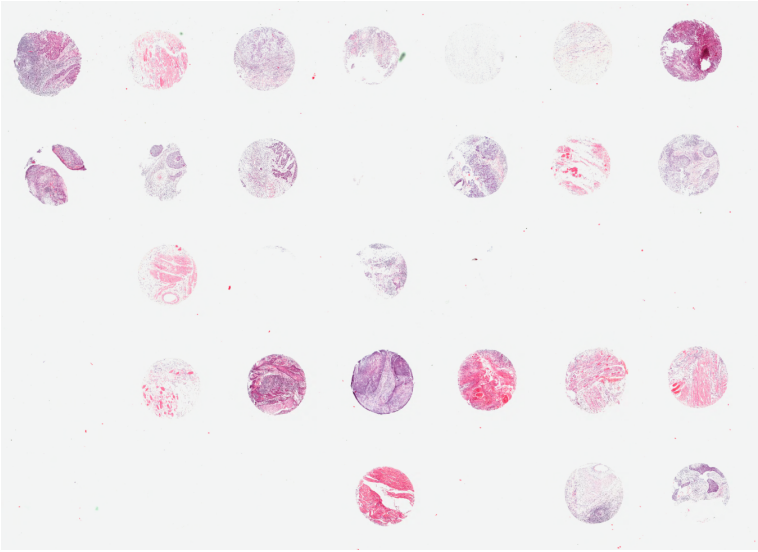

FIGURE 2.3: An example of a TMA showing cores taken from resected tumour tissue arranged in a grid like fashion. Cores have been stained with H&E to show contrast in protein and nucleic acid concentrations.

2.1.7 Histology

Pathology is the branch of medicine which is concerned with the study of disease, by study of patient samples (urine,blood,tissue etc.) to aid or provide diagnosis or prognosis [22]. The main focuses of pathology are the evaluation of structural and functional changes in patient samples. In the UK NHS, 80âĂŤ90% of diagnoses performed are based on information gained from laboratory-based medical specialists [22].

Histopathology is the microscopic study of patient tissue samples, and is primarily concerned with diseases like cancer, infection, and inflammation. In contrast to pathology, histopathology is based on the visual inspection of samples – a subjective process relying on the interpretation of morphological information present in stained microscopes slides by of a highly skilled histopathologists. A common method of diagnosis is through the examination of H&E stained tissue samples using an optical microscope. An example of a H&E stained image is shown in fig. 2.4.

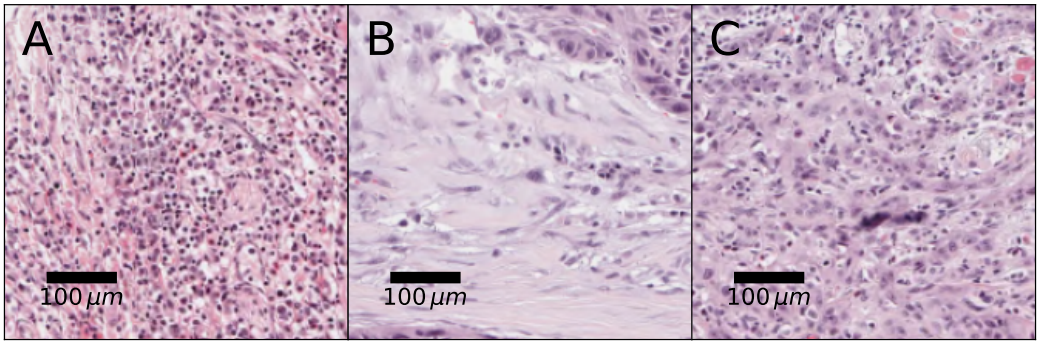

FIGURE 2.4: Examples of H&E stained samples from varied locations within the oral cavity.

Histopathologists typically aim to diagnose potentially malignant tissue according to a standardised classification system [23]. The classification system used to assess oral cancers is based upon the visual interpretation of both architecural features and cytology [23].

Whilst agreement between histologists on the extent of morphological features present within samples is largely consistent, intra and inter-observer variability continues to hinder this process due to its inherent subjectivity [23, 24, 21].

Slide preparation

The process under which a tissue goes through from biopsy to the microscope follows a few key steps:

Biopsy/resection The process of removing a tissue specimen from a patient's body, a wide variety of methods are used in practice depending on the area to be examined.

Fixation A crucial step in the process where the tissue sample is preserved in a fixative to prevent decomposition. The standard practice is to use neutral buffered formalin [22].

Paraffin embedding In order for a tissue to be viewed under a microscope it must be cut thin enough for light to pass through. Paraffin wax at approximately 58° C is used to permeate the sample [22].

Microtomy The fixed sample is then sliced precisely to a few micrometres using a *microtome*. A ribbon of paraffin embedded tissue is then extracted and floated in a water bath to prevent creases in the sample [25], this ribbon is then fixed to either a glass slide or a Calcium Fluoride disk for use in IR spectroscopy or other methods – due to its opacity in the IR.

Staining and mounting At this point the procedure can be halted and the sample can be used unstained in IR spectroscopy and other techniques which do not require histological staining. Further progression in the process rids the tissue sample of paraffin and any lipids present, this will alter the chemical make up of the sample which may have implications for subsequent analysis. If the sample is to be de-waxed it is subjected to a sequence of xylene and alcohol

washes. The sample is then stained using the required chemical to bring out the desired type of contrast.

The entire slide preparation process can take up to 48 hours [22] and is unsuitable for intraoperative diagnosis which requires a report within the time that the patient is under general anaesthetic. This may be overcome by freezing the tissue after biopsy, however this is a highly skilled process which results in inferior quality samples and greater difficulty for histopathologists. In order for a diagnostic process relying on samples prepared in this way to be used within this time frame, they must be able to perform the diagnosis under cryogenic conditions. This has significant implications for techniques reliant upon IR spectroscopy as water has strong absorbance in the spectral regions typically used for diagnosis [26].

2.2 Experimental techniques

With the advent of advances in equipment and data analysis, IR chemical imaging has emerged as a strong contender to improve clinical diagnostic capabilities [27, 28, 29, 30].

IR spectroscopy methods come in many forms and modalities with their own respective strengths and weaknesses, but all seek to observe the underlying chemical composition of the sample being analysed through the absorption of specific wavelengths of infra-red light. A brief overview of the physics involved in vibrational spectroscopy shall be given with a focus on

The operating characteristics, advantages, and disadvantages of FTIR will be discussed in the following chapter.

2.2.1 Electromagnetic Radiation

The Classical Perspective

In classical electromagnetism, electromagnetic waves comprise a magnetic field **B**, and an electric field **E** oscillating in a synchronised manner whilst propagating through space at velocity **c** in a direction perpendicular to the oscillating fields — as shown in fig. 2.5.

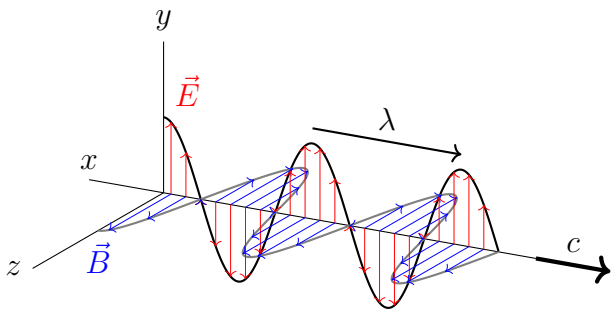

FIGURE 2.5: An electromagnetic wave of wavelength **λ** comprised of a magnetic $\vec{B}$ and an electric field $\vec{E}$ oscillating synchronously whilst propagating in space at velocity **c**

In accordance with Maxwell's equations governing electromagnetic fields eqs. (2.2) and (2.4), a change in the electric field of a wave invokes a change in the magnetic field of a wave and vice versa. This phenomenon implies that neither type of wave can exist in isolation.

$$\vec{\nabla} \cdot \vec{E} = 0 \qquad (2.1) \qquad \vec{\nabla} \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad (2.2)$$

$$\vec{\nabla} \cdot \vec{B} = 0 \qquad (2.3) \qquad \vec{\nabla} \times \vec{B} = \mu_0 \epsilon_0 \frac{\partial \vec{E}}{\partial t} \qquad (2.4)$$

$\vec{E}$ = Electric field vector (NC^{-1}) $\vec{B}$ = Magnetic field vector (T)

ϵ_0 = Permittivity of free space (Fm^{-1}) μ_0 = Permeability of free space (NA^{-2})

The amplitudes of both fields vary temporally according to the frequency of the wave; the greater the frequency of this oscillation, the greater the energy of the wave. The frequency and wavelength of the wave are related by eq. (2.5)

$$f = c/\lambda \tag{2.5}$$

f = Frequency of oscillation of the electromagnetic wave in Hz (s^{-1})

c = The velocity of light ($3 \times 10^8\ ms^{-1}$)

λ = Wavelength of the electromagnetic wave (m)

A common convention in spectroscopy is to describe the energy of an electromagnetic wave in terms of its *wavenumber*. The wavenumber is simply the inverse of the wavelength $\nu = 1/\lambda$ and is measured in m^{-1} but commonly stated as cm^{-1} in IR spectroscopy [31].

The physical process underpinning spectroscopy in general is *absorption*. Absorption occurs as a result of the dispersive effects of dielectric media in which the dynamics of the situation become considerably more complicated. Due to the interaction of the electric field with charges within the dielectric media, it is important to consider the implications this has on the electric and magnetic fields situated within the media.

The charges within a dielectric media become spatially separated when interacting with an electric field — an effect known as *polarisation*. The extent of this polarisation for an atomic system of two equal charges is given by eq. (2.6).

$$p = e\Delta x \tag{2.6}$$

p = Electric dipole moment (Cm)

q = Electrical charge (C)

x = Displacement (m).

When considering larger systems of charges: p is multiplied by the number of charges per unit volume N to give the electric polarisation of the dielectric P. The displacement x is proportional to the polarisation P, and is proportional the strength and direction of the electric field $\vec{E}$, thus the relation can be stated as:

$$\vec{P} = \epsilon_0 \chi_e \vec{E} \tag{2.7}$$

$\vec{P}$ = Polarisation per unit volume (Cm^{-2})

χ_e = Electrical susceptibility

In order to quantify the total electrical field strength in any given position and moment, a new quantity $\vec{D}$ is introduced:

$$\vec{D} = \epsilon_0 \vec{E} + \vec{P} \tag{2.8}$$

$\vec{D}$ = Displacement per unit volume (Cm^{-2}),

Combining eq. (2.7) and eq. (2.8) gives:

$$\vec{D} = \epsilon_0 \vec{E} + \epsilon_0 \chi_e \vec{E} = (1 + \chi_e)\epsilon_0 \vec{E} \tag{2.9}$$

From which the following relations can be derived:

$$\vec{D} = \epsilon_r \epsilon_0 \vec{E} \tag{2.10}$$

$$\epsilon_r = 1 + \chi_e \tag{2.11}$$

These equations relate the electric field strength $\vec{E}$ to the overall electric displacement $\vec{D}$ by taking into consideration the effect of the polarisation induced by $\vec{E}$.

In the case of an electromagnetic wave interacting with matter, the situation becomes even more complex due to the time-varying electric field associated

with the wave. In order to account for the oscillating nature of the wave, it is necessary to consider the inertia of the charges present in the media. Given an oscillating electric field:

$$\vec{E} = \vec{E_0} e^{j\omega t} \tag{2.12}$$

It is necessary to allow the electric susceptibility χ_e to take on a complex form to account for the phase difference implied by the lag in $\vec{P}$ compared to $\vec{E}$. Thus eq. (2.7) becomes:

$$\vec{P} = \epsilon_0(\chi_{e1} - j\chi_{e2})\vec{E} \tag{2.13}$$

Given eq. (2.12) and eq. (2.13), the polarisation $\vec{P}$ of the matter is now shown to be dependent upon the frequency of the oscillating electric field in which it is situated. The interaction of the electromagnetic wave with matter leads to a oscillating force driving the displacement of charges in the material. This is governed by the *Lorentz* force given by:

$$\vec{F} = e(\vec{E} + \vec{v} \times \vec{B}) \tag{2.14}$$

In the far-field regime where the distance from a source to a point is greater than 2λ, the wave can be approximated as a plane wave, therefore the amplitude of the electric field strength in comparison to its associated magnetic field is $E \approx \frac{B}{c}$. Within molecules, charged particle velocities are $v \ll c$ therefore eq. (2.14) can be approximated to:

$$\vec{F} = e\vec{E} = e\vec{E_0} e^{j\omega t} \tag{2.15}$$

Assuming a simple harmonic oscillator (SHO) model for charges in the media; the following equation of motion can be derived:

$$x = \frac{1}{(\omega_0^2 - \omega^2) + j\omega\Gamma} \frac{q}{m} \vec{E}_0 e^{j\omega t} \tag{2.16}$$

Γ = Velocity dependent damping factor

ω_0 = Fundamental oscillation frequency

Generalising eq. (2.16) to a case with multiple oscillators with their own respective fundamental frequencies and damping factors and combining with eq. (2.7) and eq. (2.13).

$$\vec{P} = \sum_i \frac{N_i q^2/m}{(\omega_i^2 - \omega^2) + j\omega\Gamma_i} \vec{E}_0 e^{j\omega t} = \chi_e \epsilon_0 \vec{E}_0 e^{j\omega t} \tag{2.17}$$

Extracting χ_e and combining with eq. (2.11) we find that the complex form of the electric susceptibility can be separated into its constituent components:

$$\chi_{e1} = \frac{q^2}{m\epsilon_0} \sum_i \frac{N_i(\omega_i^2 - \omega^2)}{(\omega_i^2 - \omega^2)^2 + \omega^2\Gamma_i^2} \tag{2.18}$$

$$\chi_{e2} = \frac{q^2}{m\epsilon_0} \sum_i \frac{N_i\omega_i\Gamma_i}{(\omega_i^2 - \omega^2)^2 + \omega^2\Gamma_i^2} \tag{2.19}$$

From where the refractive index of the material can be deduced as shown by eq. (2.20).

$$n = \sqrt{1 + \chi_{e1} - j\chi_{e2}} \tag{2.20}$$

As the frequency of the incident electromagnetic radiation ω approaches resonant frequencies of the charges within the media ω_i, the susceptibility becomes

complex and anomalous dispersion occurs in a region close to ω_i bounded by Γ_0.

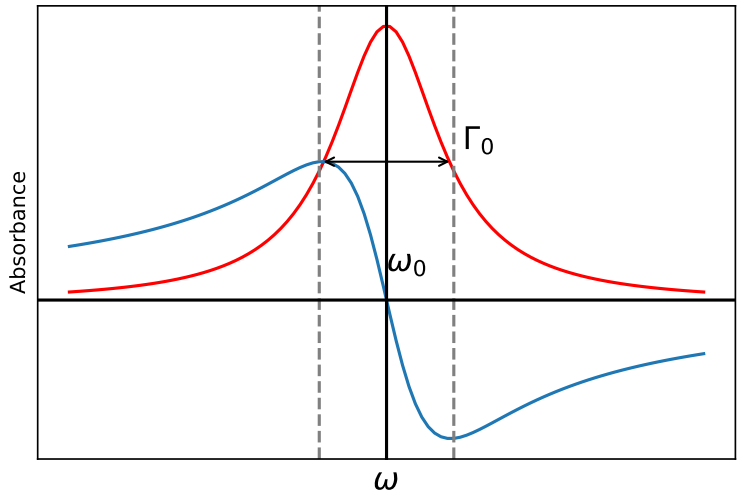

FIGURE 2.6: Complex electric susceptibility of a dielectric as a function of ω. The range of ω values where anomalous dispersion occurs is enclosed in grey dashed lines. This feature repeats at every ω_i.

The parameters ω_i and Γ_i are material specific and have multiple values for each region of resonance, this gives rise to a characteristic spectrum associated with a material where absorbance occurs at each resonance peak. These values of ω_i and Γ_i might be energy transitions that are either electronic, vibrational, or rotational in nature but all involve a change in the dipole moment of the system in question. The origin of these ω_i and Γ_i values have so far been overlooked, however in order to explain these terms in reasonable detail it is necessary to look to quantum mechanics.

The Quantum Perspective

Classical electromagnetism is able to explain a wide range of phenomena associated with waves and particles, and often serves as a useful approximation. However, phenomena such as the photoelectric effect and black body radiation, were unable to be truly explained using classical physics — thus quantum

theories of waves and particles were developed to understand these phenomena. Quantum mechanics forms the basis of all current understanding of the Universe at all scales and is necessary to appreciate the complexity of many phenomena fully.

All equations in the previous section are classical equations, and thus take no account of quantum effects. It is now understood that when particles interact they do so through the exchange of discrete *quanta* of energy, the inclusion of quantum effects into classical field theories gives rise to *quantum* field theories. The quantisation of a field theory leads to the appearance of many new features in comparison to classical field theories and are therefore much more complex.

Energy quantisation is a requirement to explain a number of phenomena. Max Planck assumed that the energy carried by an electromagnetic wave of frequency ω can only exist in quantised amounts corresponding to:

$$E = \hbar\omega \tag{2.21}$$

$\hbar$ = Reduced Planck's constant $(1.05 \times 10^{-34} Js)$

This prompts the realisation that interactions between matter and radiation are not a continuous process, but are instead mediated by an exchange of discrete amounts of energy. However, some phenomena such as reflection, refraction, and wave interference can *only* be understood by considering electromagnetic radiation to consist of waves. In order to bring concordance between these two ideas, a framework which is able to describe all physical phenomena is needed, this is what quantum mechanics seeks to achieve. In order to explain both types of observation, a quantum entity is assigned a wave-function $\Psi(r, t)$. A wave-function may be a real or complex valued function, but a key stipulation is that it has the property:

$$\int |\Psi(r,t)|^2 \, d^3r = \int \Psi(r,t)\Psi(r,t)^* \, d^3r = 1 \tag{2.22}$$

Implying that the total probability density of the particle is contained within a defined volume. In the previous section, the variables ω_i and Γ_i were used to represent the fundamental frequency, and damping coefficient of an oscillating system of charges. In order to give some insight into the origins of these terms it is necessary to represent this system of charges using quantised energy levels. Given the classical representation of a simple harmonic oscillator as the sum of its constituent energy contributions:

$$V(x) = \frac{1}{2}kx^2 = \frac{1}{2}m\omega^2x^2 \tag{2.23}$$

Its equivalent Schrödinger equation is:

$$\left(\frac{-\hbar}{2m}\frac{d^2x}{dx^2} + \frac{1}{2}m\omega^2x^2\right)\Psi(x) = E\Psi(x) \tag{2.24}$$

Due to the nature of a harmonic oscillator being in a quantum mechanically bound state, eigenfunctions of eq. (2.24) take the general form of eq. (2.25). Proof see [32].

$$\Psi(x) = e^{\frac{-x^2}{2\alpha^2}}(a_0 + a_1x + a_2x^2...) \tag{2.25}$$

And satisfy Hermite's equation:

$$\frac{\partial\Psi(x)}{\partial x^2} + \beta\Psi - \left(\frac{x^2}{\alpha^4}\right)\Psi = 0 \tag{2.26}$$

Where:

$$\beta = \frac{2mE}{\hbar^2} \qquad (2.27) \qquad\qquad \alpha = \sqrt{\frac{\hbar}{m\omega}} \qquad (2.28)$$

Due to the nature of a bound oscillator state, wave-functions cannot diverge as $x \to \infty$ and must be quantised. Solutions meeting this requirement satisfy:

$$\alpha^2\beta = 2n + 1 \qquad (2.29)$$

Therefore, the first three allowed wave-functions meeting this requirement are:

$$\Psi_0(x) = c_0 e^{\frac{-x^2}{2\alpha^2}} \qquad (2.30)$$

$$\Psi_1(x) = c_1 \left(\frac{x}{\alpha}\right) e^{\frac{-x^2}{2\alpha^2}} \qquad (2.31)$$

$$\Psi_2(x) = c_2 \left(\frac{2x^2}{\alpha^2 - 1}\right) e^{\frac{-x^2}{2\alpha^2}} \qquad (2.32)$$

Figure 2.7 depicts the first three oscillator eigenfunctions of a quantised harmonic oscillator.

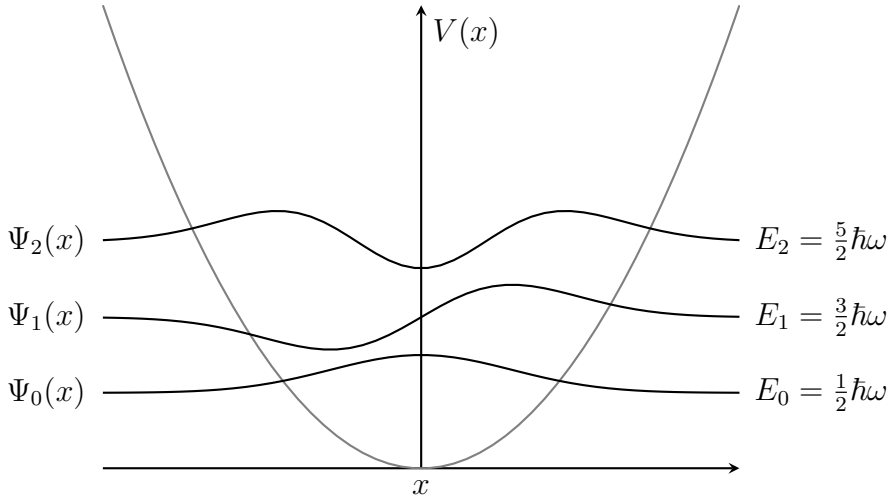

FIGURE 2.7: Potential energy of the quantised harmonic oscillator with first three allowed eigenfunctions and their corresponding energy eigenvalues.

When a photon is incident upon a chemical sample, an electron is promoted to an excited state if the photon energy $E = \hbar\omega$ is equal to ΔE_{ij}. These principles apply to the vibrational energy transitions within molecules which spectroscopy techniques seek to observe. In reality the harmonic oscillator model is inaccurate except for regions at the bottom of the potential energy curve, instead the potential energy function of an atom follows that of an anharmonic oscillator. The potential energy between two atoms $V(r)$ as a function of the separation r reaches a minimum at r_0. This is due to repulsive nuclear forces experienced in regions where $r < r_0$ and attractive forces between electrons and the nucleus in regions where $r > r_0$.

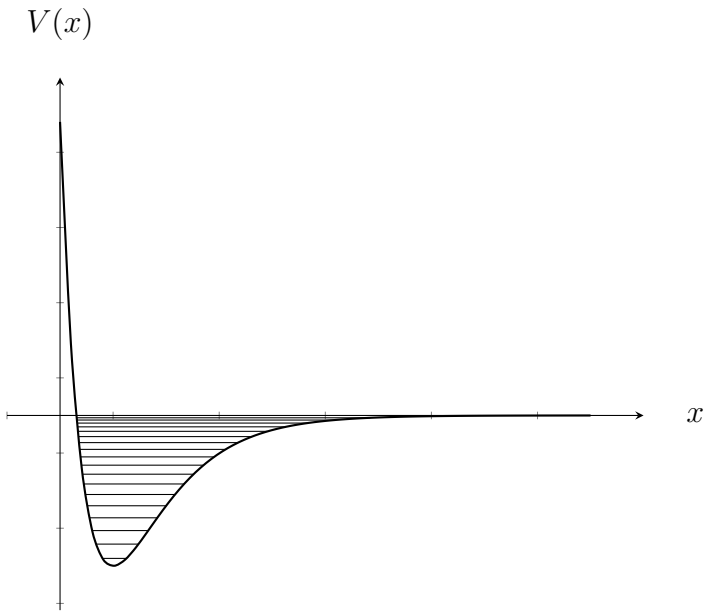

FIGURE 2.8: Potential energy function of the quantised anharmonic oscillator.

In contrast to a harmonic oscillator, the restoring force felt by an object undergoing anharmonic oscillation is non-linear and dependent upon the displacement from the equilibrium position. An IR spectrum contains multiple peaks corresponding to many different energy transitions, many more transitions are allowed due to the anharmonicity of the potential function; energy changes where $\Delta n > 1$ are allowed — these transitions are known as overtone bands [33]. These additional transitions have less energy compared to fundamental changes and give rise to *hot bands*. The intensity of an absorption band are proportional to the change in molecular dipole moment, therefore larger changes in the molecular dipole moment give rise to larger absorption peaks. In gas phase spectroscopy IR spectra show distinct peaks due to rotational energy transitions being more readily resolved.

Molecular vibrations vary from simple coupling of diatomic molecules to a much more complicated situation involving many atoms. Vibrational energy changes are much smaller than electronic energy level changes, as shown in fig. 2.9.

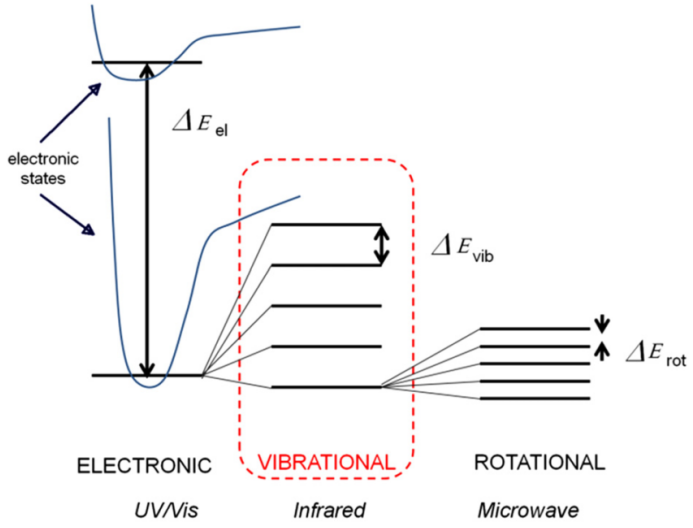

FIGURE 2.9: Energy changes present in molecular spectra. [33]

A molecule with N atoms has 3N degrees of freedom corresponding to translational motion in x,y,z, and rotational motion centered about the x,y,z axes. The remaining 3N-6 degrees of freedom correspond to vibrational modes involving harmonic displacement of atoms from their equilibrium positions. An illustration of vibrational modes in a CH_2 group is given in fig. 2.10.

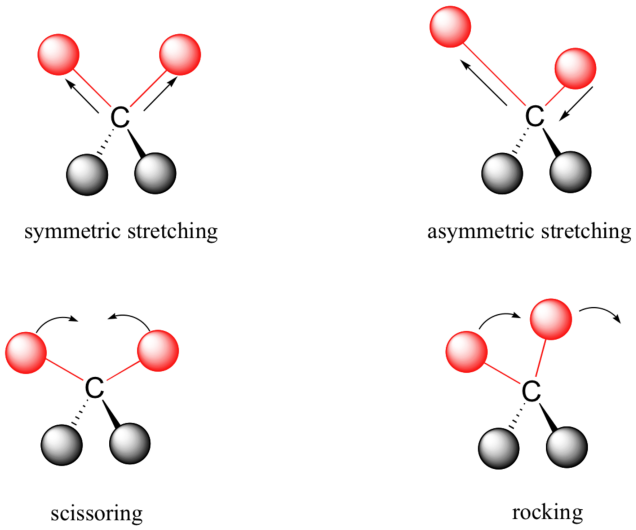

FIGURE 2.10: Energy changes present in molecular spectra. [34]

2.2.2 IR spectroscopy

IR spectroscopy is a well known technique that has grown in complexity and variety over the past few decades. IR spectroscopy is almost universal in its applicability due to many molecules having strong regions of absorption in the mid-infrared region. Samples in any physical state can be examined (with some preparation), and many different types of samples such as polymers, powders, and organic and inorganic compounds can have their IR spectra measured [35]. Spectra are very information rich; with peak positions giving information about molecular structures present in the sample, peak intensities yielding information about the concentration of such molecules, and peak widths providing information about the chemical state of the sample. It is inexpensive, quick, and operators can be trained quickly using modern hardware and software. Some consideration needs to be taken however when examining certain samples: water, and CO_2 contributions can be a limiting factor when seeking to examine spectra accurately, but solutions exist to mitigate these effects.

IR Spectroscopy uses the interaction of IR light with matter across a number of wavelengths to produce an absorption (or transmittance) spectrum, this absorption spectrum arises from the vibrational interactions of the IR light with the molecular bonds present in the sample. The absorption is dependent upon a number of factors: the wavelength of the IR light, the atoms involved in the molecular bond, and the strength of intermolecular interactions [36]. This interaction typically occurs in the mid and far-IR spectral region, where molecular vibration frequency and incident light frequency are approximately equal, and when a change in molecular dipole moment occurs [37].

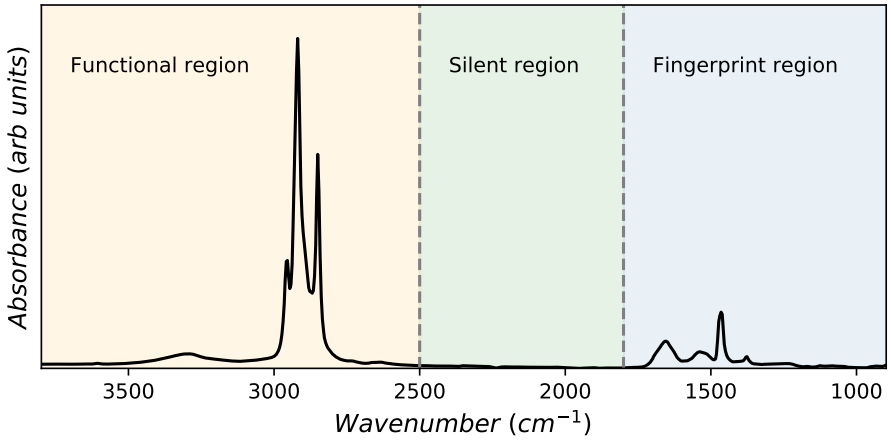

FIGURE 2.11: A typical biological FTIR spectrum example showing three distinct regions of the spectrum: the ”functional region” (3800-2500 cm^{-1}), ”silent region” (2500-1800 cm^{-1}), and ”fingerprint” region (1800-900 cm^{-1})

The interactions between the constituent atoms of a molecule and the incident light results in a unique IR spectrum for the molecule — an example spectrum is shown in fig. 2.11. A tissue sample or cell is just a collection of molecules in a unique environment which will also display a unique IR spectrum. This can be used to characterise entire sections of tissue, or cell phenotypes based on the collective contributions of the constituent molecules.

A common mode of operation for Infra-red spectroscopy is *transmission* mode. A beam of IR light is incident upon the sample where a portion of the light is absorbed according to the vibrational modes of the molecules present in the sample; the amount of absorbance is in proportion to the concentration of the molecules present, according to the Beer-Lambert law.

$$I = I_0 e^{-\mu x} \tag{2.33}$$

I = The attenuated intensity,

I_0 = Intensity of the incident IR beam,

μ = Absorption coefficient of the attenuating material,

x = The thickness of the attenuating material.

An IR spectrum measures the absorption of the sample as a function of the incident photon energy. The measured absorbance is calculated using the following:

$$A = log_{10}\frac{I_0}{I} \tag{2.34}$$

A = Absorbance,

This technique is generally used for thin samples in the region of 1-20 μm [33] where the Beer-Lambert law is valid. If samples thicker than 20μm are used the relationship begins to break down. The Beer-Lambert law allows for the determination of molecular concentrations for many applications, however it is not without its flaws. Particularly dense samples will not absorb linearly and assumptions about the origins of an measured μ value must be met with some skepticism as a chemical compound will not have a single value associated with it. Scattering effects are indistinguisable from absorption in IR spectra so the assumption that any "missing" light intensity is purely due to absorbance effects may be false [38].

IR Detectors When measurements are performened in transmission mode using an IR spectrometer, samples are fixed to a substrate which is transparent in the IR such as CaF_2. In order to accurately capture the spectrum of a sample with minimal absorbance elsewhere mirrors are utilised instead of conventional glass optics. A Schwarzchild-Cassegrain objective is used to focus the incoming light onto the sample from above, at which point the light passes through the sample and is re-collimated by a condenser lens before passing through

subsequent mirrors to the detector. A schematic of this process is shown in fig. 2.12.

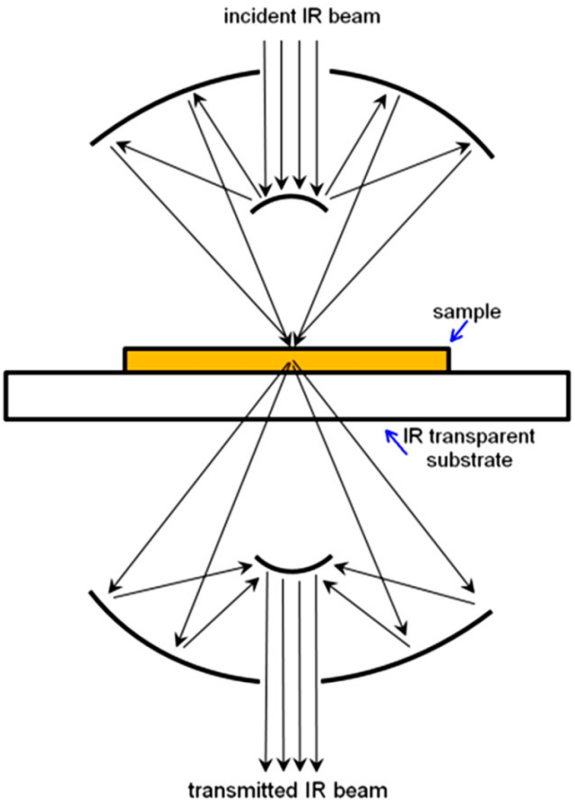

FIGURE 2.12: A Schwarzchild-Cassegrain showing the incident IR beam passing through a series of mirror optics, sample, and a second series of optics before passing through to the detector [33].

In order to quantify the intensity of IR light for further analysis, the signal must first be converted into electron pulses, digitised, and measured by a computer. The conversion of IR light to electron pulses is typically performed by a mercury cadmium telluride (MCT) detector. An MCT is a semiconductor compound with a bandgap tuned to the desired wavelength range through the addition of cadmium. When an IR photon strikes the MCT an electron within the valence band of the detector is promoted to the conduction band where it then sent as an analog signal to an accompanying analog to digital converter. In practice this happens for large numbers of photons and thus electron pulses are observed which are proportional to the intensity of the photon flux [39]. An MCT must

be cooled to temperatures similar to liquid nitrogen (77K) in order to minimise the effects of noise induced by thermally excited current carriers. MCT detectors are far more sensitive than other comparable detectors [39] and convert IR photons to electron pulses much more quickly, this has led to MCT detectors becoming the detector type of choice for many scientific applications.

Light sources In order to observe a wide range of samples a suitable source of IR light is required; three such sources of light are the Quantum Cascade Laser (QCL), Free-Electron Laser (FEL), and a globar. Each source of light has its own respective advantages and disadvantages in terms of: spectral output, source stability, intensity, and cost. An IR-FEL requires a large supporting facility and can be extremely expensive to operate, FELs also suffer from source stability issues and are highly unsuitable for a clinical environment [40]. A typical globar is comprisesed of a silicon carbide rod heated to temperatures of around 1000 - 1650 °C, and a variable interference filter. As the globar emits a continuous IR spectrum across a wide range specific wavelengths it can be filtered into specific bands of wavelengths or used in conjunction with an interferometer.

A QCL is a semiconductor laser which utilises epitaxially grown quantum wells containing electrons in lasing states within a sequence of quantum wells. QCLs allow for a spectrally narrow-band beam when used in conjunction with narrow-band mid-IR reflectance filters [33]. Shown in fig. 2.13 is a simplified schematic of a QCL.

Sources can output either a *continuous* IR spectra in the case of a FEL and glowbar; or a *discrete* spectrum as in the case of a QCL. A QCL offers a distinct advantage when used with a compatible imaging system; as discrete wavebands are scanned sequentially, the signal to noise ratio can be greatly

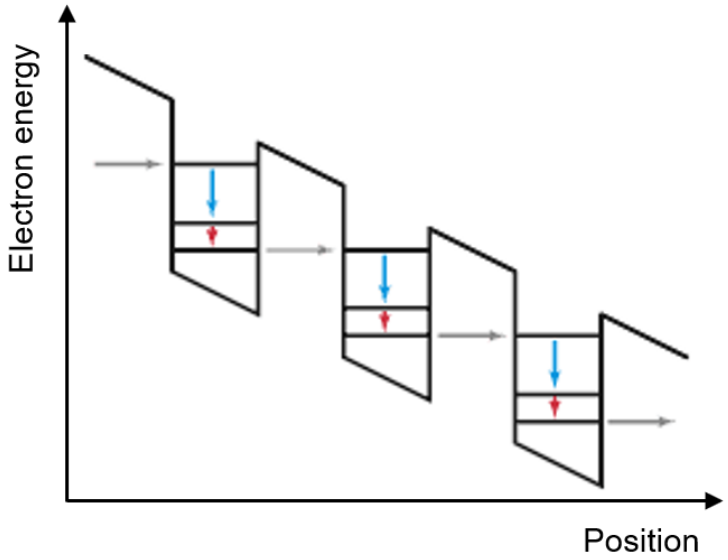

FIGURE 2.13: A simplified schematic of the gain region of a QCL, showing electron energy against position. The electron is injected at the left-most grey arrow and undergoes a radiative transition (blue arrow), the electron then undergoes a further non-radiative transition (red arrow) before tunnelling to the next quantum well and repeating this process [41].

increased by averaging over a small spectral range instead of scanning over the entire spectrum as in an FTIR [33].

2.2.3 Fourier-transform infrared spectroscopy (FTIR)

The application of FTIR to biological samples is relatively novel, with a range of potential applications across biomedical sciences. FTIR has been used to investigate the development of cancer in a number of tissue types such as: brain [42, 43, 44], colon [45], skin [46], liver [47] and many others and is considered to be one of the most popular IR techniques available today [48]. It provides a way to assay the chemical structure of a sample in a non-destructive manner. FTIR has proven to be a rapid and cost-effective technique which requires minimal preparation, and could potentially be used to help alleviate the subjectivity present in histopathological diagnosis.

Operating Principle A typical set-up for an FTIR spectrometer is a Michelson interferometer and a detector. A Michelson interferometer is a instrument which produces an interference pattern by superimposing two beams of light. The light is incident from the IR source and onto a beamsplitter, the beamsplitter transmits a portion of the IR light and reflects the other portion onto a fixed mirror. The portion of the beam which is transmitted is incident onto a movable mirror which reflects back to the beam splitter, re-combines with the other portion of light, and proceeds to the IR detector. The two beams undergo superposition and create an interference pattern. A schematic of a Michelson interferometer is shown in fig. 2.14.

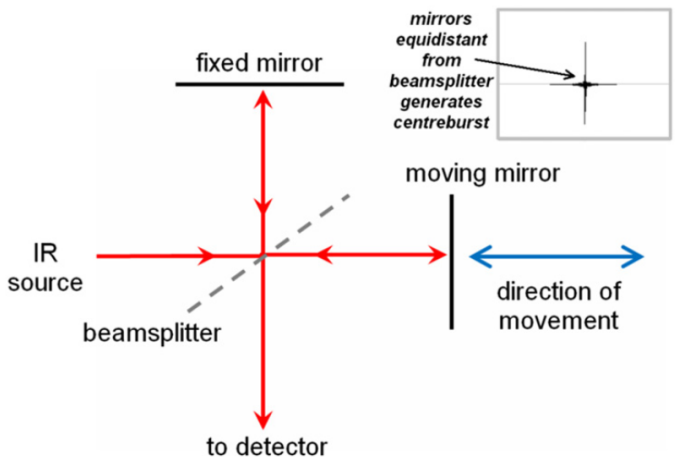

FIGURE 2.14: A Michelson interferometer used in a FTIR spectrometer. [33].

If the distance between the beamsplitter and movable mirror is equal then the re-combined light interferes constructively. This condition repeats for every integer multiple of a wavelength, destructive interference occurs every half wavelength for a given wavelength of light. The intensity $I^{'}$ of the beam at the detector is given by eq. (2.35).

$$I^{'}(\delta) = 0.5I(v_0)\left(1 + \cos(2\pi)\frac{\delta}{\lambda}\right) \tag{2.35}$$

Where the retardation is given by:

$$\delta = n\lambda \tag{2.36}$$

The beamsplitter and detector each have wavelength dependent efficiencies which can be accounted for with the inclusion of a correction factor $H(v_0)$. Including the wavenumber dependent responsivity of the detector $G(v_0)$ $(V \cdot W^{-1})$ therefore gives the measured intensity in volts [38]:

$$S(\delta) = 0.5I(v_0)H(v_0)G(v_0)\left(1 + \cos(2\pi)\frac{\delta}{\lambda}\right) \tag{2.37}$$

The resulting interferogram is a combination of the intensities of each wavelength of light as the mirror is moved. This interferogram is then transformed to a frequency domain spectrum using a Fourier transform as shown in fig. 2.15. In order to obtain a spectrum which is characteristic of the sample, a background measurement is taken in the absence of the sample, the background spectrum is then subtracted to obtain the sample spectrum.

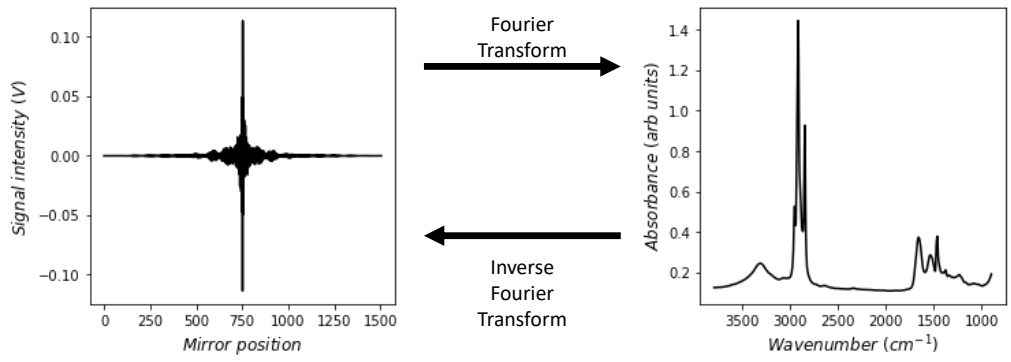

FIGURE 2.15: The conversion of an interferogram to a wavelength dependent transmittance spectrum

Measurement of spatial variation In order to obtain spatially varying spectra across the sample of interest either imaging or mapping can be performed. Mapping is done by collecting the absorbance spectra at each position in the desired area of the sample, this area can be changed through the use of piezoelectric motors to give micrometer resolution [26]. Imaging is achieved by directing the light emitted from sample region onto a Focal Plane Array (FPA) using focusing optics to define the pixel size [33]. Mapping the spectra of a sample area in this way creates a *data cube* with each spatial pixel corresponding to a measured spectrum at that point.

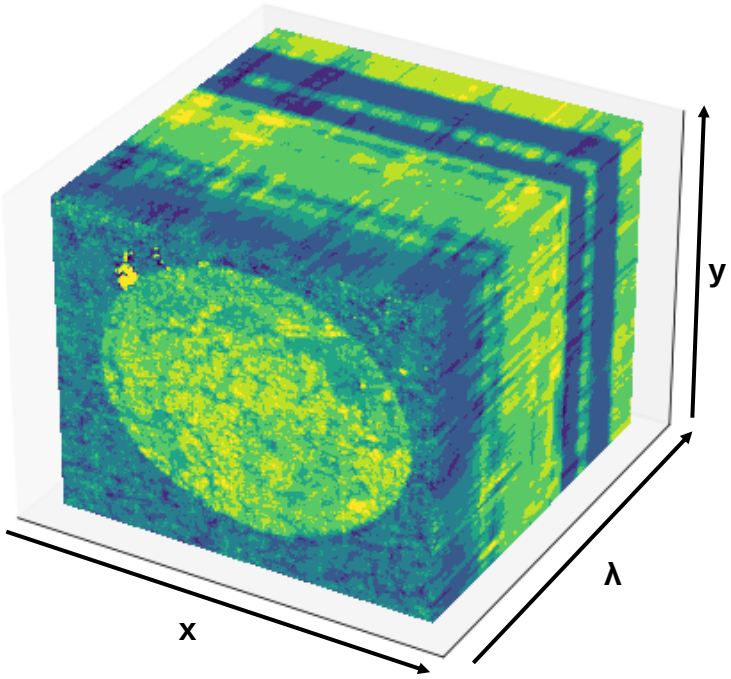

FIGURE 2.16: An FTIR datacube example showing spatial variation in x and y with spectral absorbance varying in λ.

The achievable spatial resolution of an optical technique such as FTIR is governed by a number of factors. The first is the magnifying lens used to focus the light onto the sample; if a powerful lens is used, the FPA will image a smaller area due to its decreased field of view, and so each pixel represents a smaller

area. Second is the Numerical Aperture (NA) of the lens, which is representative of the ability of a lens to collect light over a range of angles, therefore lens with a high NA will be able to resolve objects at smaller scales. The resolution of an optical instrument however is always subject to the diffraction limit. It should be noted that due to dispersive effects associated with glass lenses, mirrors are the typical choice for the optical path of a FTIR microscope.

Diffraction and Resolution In order to image objects at smaller scales using optical techniques, the resolution of the imaging technique must surpass the diffraction limit, which is limited to roughly half the wavelength of the light sourced used in acquisition [49]. The diffraction limit is the minimum size of the spot which a beam of light can be focused to using normal lensing elements. The focused spot forms a symmetric pattern of concentric rings called an Airy disk pattern as shown in fig. 2.17.

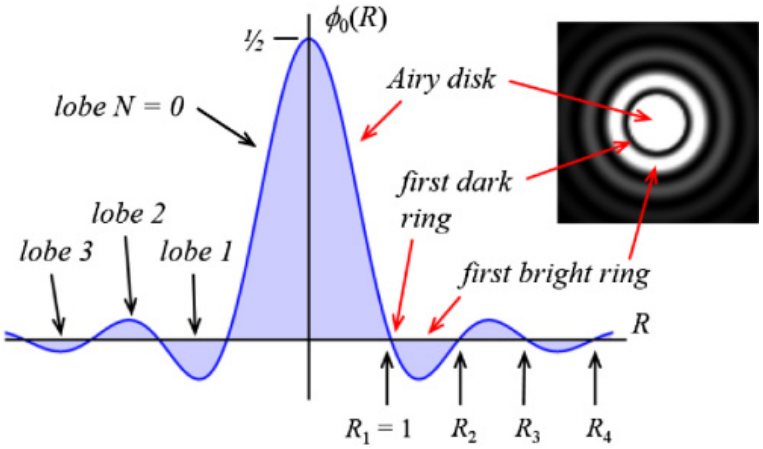

FIGURE 2.17: An Airy disk diffraction pattern showing periodic bright and dark fringes [50].

In order for two objects which are close by to be individually resolvable, they must obey the *Rayleigh criterion* which states that the two objects must be a at least a distance away given by eq. (2.38), the distance between the two first bright fringes of the Airy disk produced by each object.

$$d = \frac{0.61\lambda_0}{\sin(\theta)} \tag{2.38}$$

d = Distance between the centre of the Airy disk and the first minimum intensity

λ_0 = Wavelength of the light in vacuum

θ = Light convergence angle

With the improvement in NA of many microscopy systems available today, it is possible to achieve values in excess of 1. Consequently this means that with good quality optics microscopy systems can achieve resolutions of up to $\lambda/2$ [49]. This can in practice be overcome with good estimates of the Point Spread Function (PSF) of the detector and in conjunction with a high signal-to-noise ratio, however this theoretical limit is generally never achieved due to limitations of experimental conditions and aberration effects in optical instruments [49].

FTIR has become a commonplace instrument for the analytical chemist over the past few decades owing to its robust reproducible spectra, low cost, and versatility of use [33, 38]. In combination with modern FPA detectors it is possible to extract large datasets rapidly and cost effectively for use in later analyses.

2.3 Data Analysis

In order to gain further insight into the data collected from the previously mentioned scanning methods the data must be summarised and interpreted. A huge variety of methods exist which approach the problem from different angles, some are well-established techniques which have their origin in multivariate statistics; whereas other techniques are categorised as *machine learning* (ML).

Effective data analysis is an extremely important step when developing a clinical assay procedure as a high false positive rate will result in a large number

of unnecessary procedures, and a high false negative rate can result in unnecessary deaths.

With the aim of the project being to gain understanding of the biological systems present in cancer and to develop methods of diagnosis, it is necessary to convert the *raw data* containing spectral and spatial information about each sample into meaningful information by quantifying relationships or categorising clusters of data points.

The following section starts off with a description of a few procedures which are common data preprocessing techniques in the field of data science. Then an overview of a number of statistical techniques and machine learning methods will be given with a focus on

Then *XGBoost* the main machine learning method utilised in the analysis of data in this project will be described from basic principles to the full algorithm. Finally an overview of the process of validating classifier methods will be given.

2.3.1 Machine Learning & Statistics

ML is an approach to data analysis which involves learning from example data rather than through the use of heuristics. This approach has led to advances in fields such as finance [51], healthcare [52], bioinformatics [53]. ML objectives are generally either *classification* or *regression* problems. The goal of a classification is to obtain the function f which maps an input vector $\boldsymbol{X}$ to a discrete output $\boldsymbol{Y}$. The input vector $\boldsymbol{X}$ is the list of variables which is used to describe a data point e.g. colour, weight, length, and $\boldsymbol{Y}$ being the label applied to the data point e.g. orange, apple, banana. A regression problem is one that seeks to turn $\boldsymbol{X}$ into a *continuous* output, for example the number of bathrooms or floor space of a house into the market value. There does exist some overlap between these types of problems but they are generally evaluated differently.

These problems can be further separated into *supervised* and *unsupervised* learning problems; supervised learning is when each data point has an associated label so that the algorithm can learn more directly, unsupervised learning is therefore learning in the absence of labels. These differences have implications for the types of algorithms which can be used and to what performance that can be achieved.

2.3.2 Preprocessing

Analysis of spectroscopic data is typically a multi-step procedure, starting with a sequence of preprocessing steps before classification or regression. This process naturally follows a ”pipeline” like procedure where data is passed sequentially through a number of steps, this process is illustrated in fig. 2.18. Preprocessing is a vital step in the analysis workflow, as it has been shown to generally increase performance of classification models [54], as well as to increase the validity and interpretability of results.

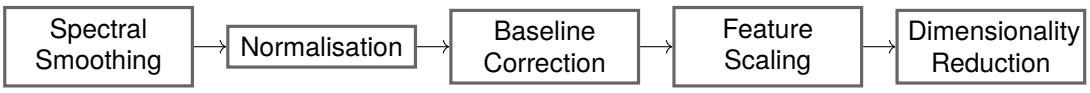

FIGURE 2.18: A typical preprocessing pipeline diagram

An outline of each step in the preprocessing sequence is set out below.

Normalisation

Spectral normalisation takes place to account for the variable thickness of samples. Due to the dependence of eq. (2.34) on the thickness of the sample, the absolute absorbance value will vary strongly. This is considered to be a confounding factor and is typically dealt with by a number of methods for example

vector normalisation, min-max scaling, or through the use of spectral differentiation.

Spectral Smoothing

Spectral smoothing methods seek to account for high frequency noise in the data. This unwanted noise may have instrumental, environmental, or sample origins. There are several associated methods, including the commonly used Savitzy-Golay [55], whereby a polynomial is fit to a local moving window of specified length. Other methods such as principal component analysis (PCA) de-noising and fast Fourier transform (FFT) filtering are also commonly applied.

Baseline Correction

In transmission IR spectroscopy, the incident light beam will experience a degree of *Mie* scattering, modifying the observed IR spectrum. The original chemical spectrum sits atop an induced non-linear baseline caused by the wavelength dependent scattering of the incident light. Mie scattering occurs if there are spherical morphological structures present in the sample which are of comparable size to the incident radiation. This effect is particularly strong in cells but tissues are also adversely effected to some extent. The effect in embedded tissue samples is mitigated somewhat by the presence of paraffin wax which results in a more homogenous refractive index throughout the sample [56, 57, 58]. The data present in this thesis has been subject to an Extended Multiplicative Scattering Correction (EMSC) algorithm outlined in [59].

Feature Scaling

Feature scaling takes place to effectively remove absolute variable values. This does not detrimentally effect the data as it is only relative values between sub-groups of data which are of relevance for discriminatory tasks. This step often helps subsequent classifier steps and is imperative for PCA.

2.3.3 Dimensionality Reduction

When a dataset contains a large number of features which are used to describe a data point it can become computationally expensive to process. The goal of dimensionality reduction (DR) is to decrease the number of components of the feature vector x to reduce computation time and/or expense, DR can also allow higher dimensional data sets to be visualised in lower dimensions. DR can also be advantageous for classifier performance as it can play a role in regularising the classifier due to the reduction in information given to the classifier.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a DR technique which seeks to re-orient the axes representing a dataset so that the axes are those which maximise the variance. This is to reduce the complexity arising from any linear dependence between the components of the feature vector and disregard redundant information. The data is mapped to a subspace which maximises the variance of the orthogonal projections of the data points as shown in fig. 2.19.

This process is performed by obtaining the eigenvectors of the covariance matrix of the data, these eigenvectors then become the principal components [61]. Using a PCA however does come with the disadvantage that the data can be

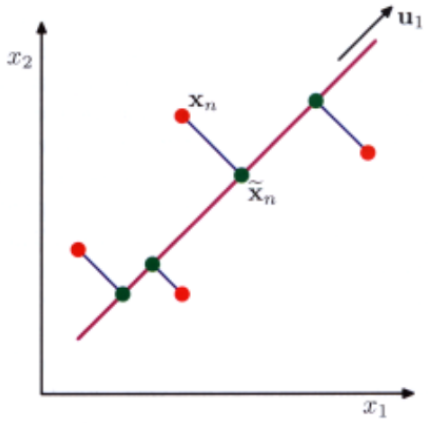

FIGURE 2.19: The principal component shown here by u_1 with the orthogonal projections (shown in green) of the original space data points (in red) projected onto it [60]

difficult to interpret. When information about which features were the most important in discrimination is needed, some information has been disregarded in the process of mapping to the new subspace, and a degree of mixing goes on between each of the components in the original space.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) differs from PCA in that uses information about the identity of each data point to remap the data. The main goal of LDA is map to a space which gives good inter-class separability and avoid over fitting to the data.

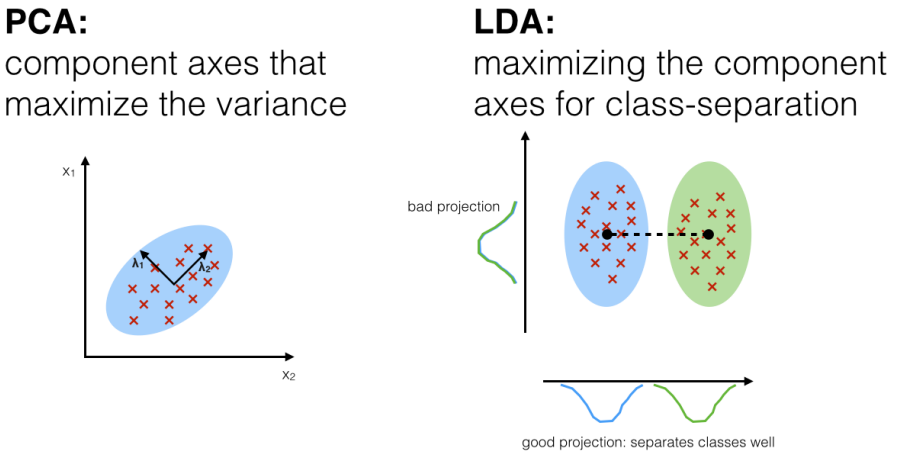

FIGURE 2.20: A comparison of PCA and LDA [62]

Another goal of LDA is to minimise intra-class variance to avoid scattering the data across the subspace. LDA can allow for the visualisation of the underlying structure of the data, and can aid in visualising the relationships between groups of data.

2.3.4 Machine learning algorithms

This section will give a brief overview of some machine algorithms used in the work compromising this thesis. Neural networks will be covered in more detail in later chapters where they are utilised as the primary focus.

Logistic Regression

Logistic regression (LR) is a relatively simple classification method with a basis in classical statistics. It is closely related to linear regression but with an additional logistic function step used to convert input vectors into a usable probability estimate. LR is based on the following equation:

$$Pr(Y = 1) = \frac{1}{1 + e^{-z}} \tag{2.39}$$

Where

$$z = \beta_0 + \beta_1^T X_1 + \beta_2^T X_2 ... = \sum_{i=0}^{n} \beta_i^T X_i \tag{2.40}$$

This is illustrated in a univariate case for a two-class problem in fig. 2.21.

Optimal values for coefficient vector β_i are determined through maximum likelihood estimation. In practice this is done by iteratively maximising the log-likelihood with respect to β using Newton's method or otherwise. See [63] for a thorough derivation.

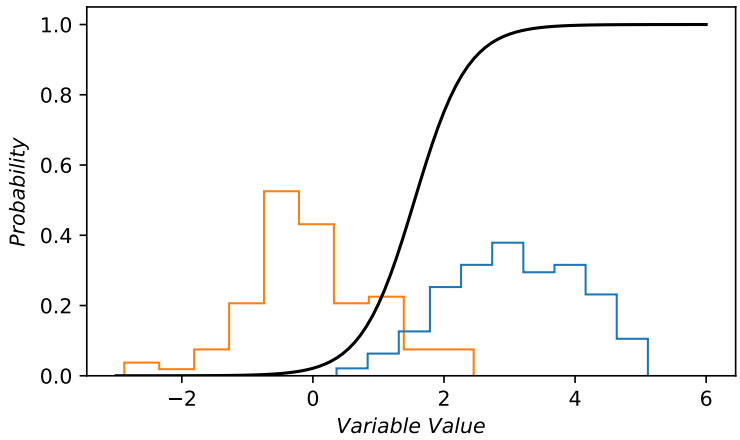

FIGURE 2.21: A univariate logistic regression example showing two generated distributions in orange (Y=0) and blue (Y=1); the fitted logistic function is shown in black with maximum probability predicted in the region spanning the blue histogram (Y=1).

Support Vector Machines

Support vector machines (SVM) have become a commonly used method for classification and regression capable of performing well on complex datasets [64, 60, 65]. SVMs can be separated into two distinct types: linear, and non-linear. A linear SVM as its name suggests forms a linear decision boundary across the input parameter space separating classes.

$$\hat{y} = \begin{cases} 0 \text{ if } \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + b < 0 \\ 1 \text{ if } \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + b \geq 0 \end{cases} \tag{2.41}$$

w = Gradient associated with the linear decision boundary

b = Constant offset of the decision boundary

Values for w and b are determined through an optimisation procedure which seeks obtain the optimal boundary of separation between classes. This is accomplished by minimising an objective function which also allows for some misclassification through a *slack variable* ζ. The objective function is given by

$$\underset{w,b,\zeta}{\text{minimise}} \ \frac{1}{2}w^T w + C\sum_{i=0}^{n}\zeta_i \tag{2.42}$$

C = A tunable hyperparamter allowing for a level of misclassification

When a dataset is not linearly separable it is still possible to utilise an SVM as a classification method through the use of a *kernel*. A kernel takes the original n-dimensional parameter space of the dataset and transforms it into a new m-dimensional space called a *feature space*, where m $>$ n [65]. A kernel can take many forms such as a linear, polynomial, or radial basis function (RBF) kernel; these kernels perform calculations based on the original data to derive more features which can allow for a linear separation between classes. A comparison of a linear and nonlinear SVM is shown in section 2.3.4.

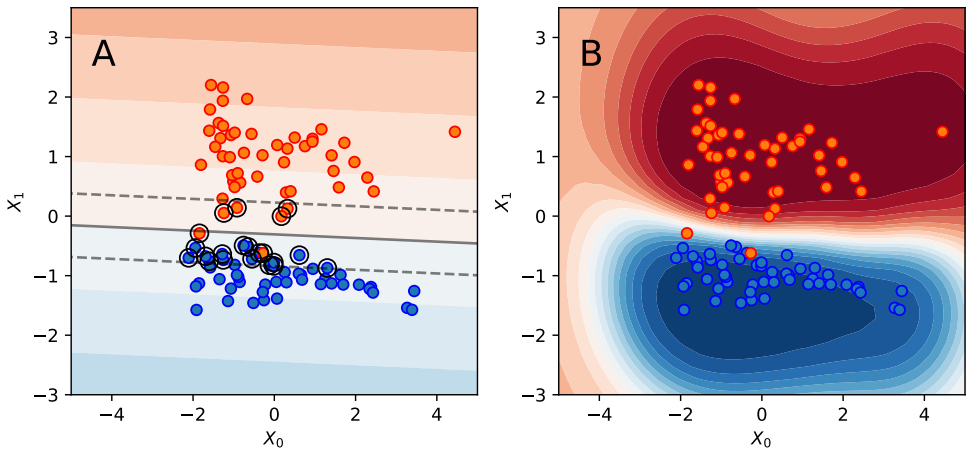

FIGURE 2.22: SVM classification boundaries showing a linear SVM (A), and nonlinear SVM using a RBF kernel function (B). (A) shows the support vector boundary in a solid grey line with support vector points highlighted with a black circle.

Artificial Neural Networks

An artificial neural network (ANN) is a technique loosely based on the functioning of a biological neuron. First introduced in 1943 [66], ANNs are a simplified computational model of how a biological neuron might work in an animal brain.

ANNs vary widely in complexity and structure, with the addition of specially designed layers and functions ANNs can accomplish increasingly complex tasks such as: natural language processing, computer vision, and time series prediction. Like its biological analogue, a neuron is able to receive an input signal, perform some processing, and output the resultant value. This process is summarised in fig. 2.23.

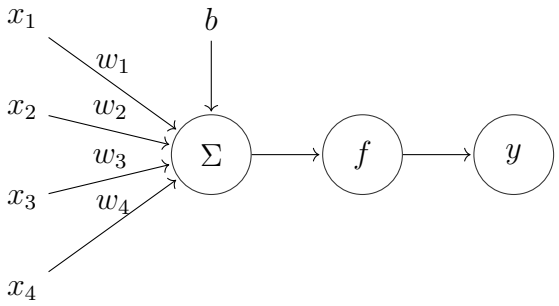

FIGURE 2.23: The perceptron showing input variables x_i multiplied by their respective weights w_i, before being summed over all inputs and added to a bias term b, and then subjected to a nonlinear activation function f

The perceptron equation for a number of input variables n is given by eq. (2.43)

$$y = f\left(\sum_{i=0}^{n} x_i \cdot w_i + b\right) \tag{2.43}$$

In order for a perceptron to succeed in its desired application it must be 'trained'. Training in this sense refers to an optimisation of the weights of a perceptron with respect to a desired metric – often metrics such as accuracy, sensitivity, or specificity are used for classification tasks. If the perceptron is to be used for regression this metric would be an indicator of the loss/fitness of a proposed function.

The Multilayer Perceptron

A single perceptron is very similar to a single unit of logistic regression and can achieve simple binary classification tasks [64], however most classification tasks are substantially more complicated .When the output of a perceptron is used as the input of another perceptron ANNs can model complex non-linear relationships; such structures are known as 'deep neural networks'. The term *Deep learning* is associated with the research and development of these complex models. A typical multi layer perceptron network is illustrated in fig. 2.24.

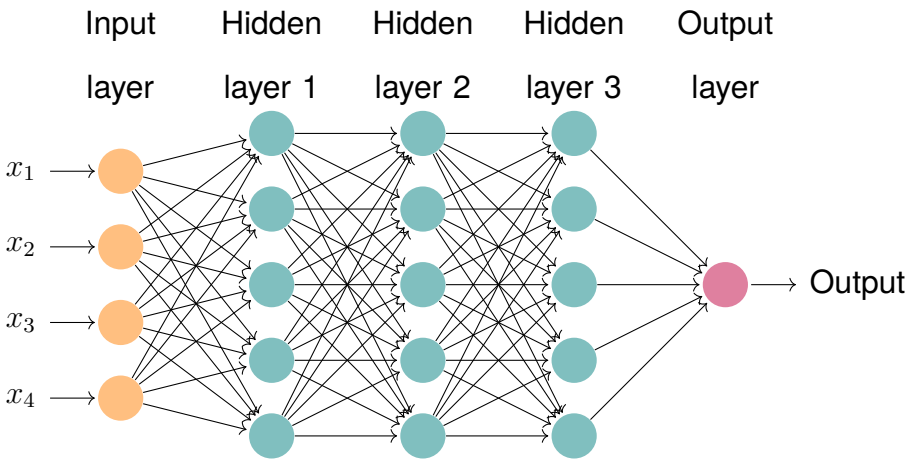

FIGURE 2.24: A multilayer perceptron neural network with an input layer consisting of four input variables $x_0 \dots x_4$, three hidden layers of five nodes each, and a single output layer.

The use of successive layers of nodes allows an ANN to model significantly more complex relationships. Optimal choices for network weights must be determined through an optimisation procedure. *Backpropagation* [67] is the standard method of determining effective weights for each node, it is made possible due to the fact that all weights and bias terms can be related to the error of the network through a series of *gradients* – a chain of partial differential equations. A *forward pass* is used to calculate the output of a network given an input, this output is compared to the true output value and an error is computed. With

each weight and bias in the network being related to this error, the network adjusts each parameter accordingly in a *backward pass*. This process continues with batches of samples from a dataset until the network error converges. The 'trained' network can now be used to perform predictions on unlabelled data samples.

There are many options for the choice of activation functions and the number of layers and nodes in a neural network, additionally there are many parameters associated with the back propagation algorithm itself which can be altered. A thorough description will not be given here but is covered in depth in other sources [64, 60, 63]. An additional process known as hyperparameter optimisation can be performed and will be covered in depth in **??**. Neural networks form the focus of **??** where an explanation of convolutional neural networks (CNN) shall be given.

Classification and Regression Trees (CART)

A decision tree is a type of ML algorithm which constructs a tree-like structure consisting of branch nodes and leaf nodes. They can be used to perform classification or regression tasks by splitting the data set at each branch node using a set of criteria, usually the split is calculated as that which will maximise the entropy gained according to the Gini index [68]. The splitting generally continues until either: each leaf node leaves a single class, a maximum tree depth is reached, or a given performance metric has been achieved.

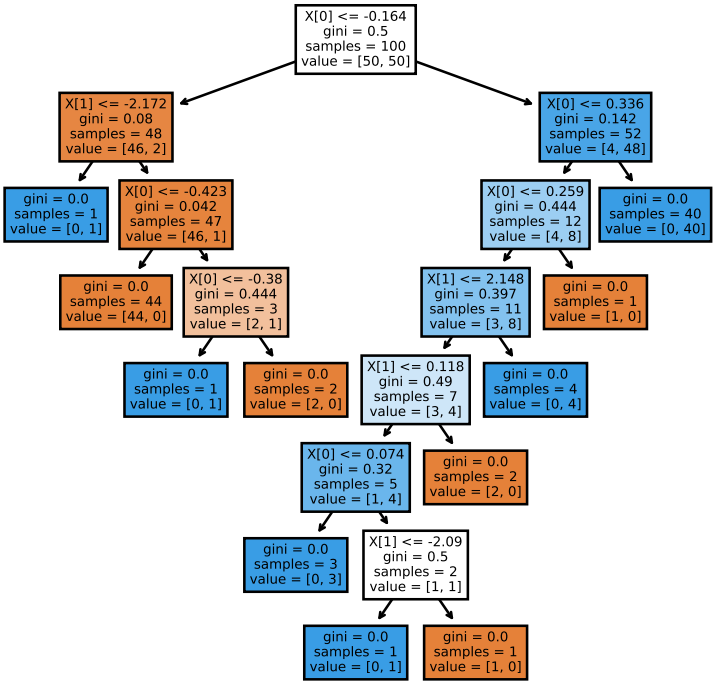

FIGURE 2.25: A typical CART comprising branch nodes shown here by a logical decision operator, and leaf nodes consisting of an output value.

CARTs can be effective classifiers in simple data sets but they suffer from overfitting due to their high variance – a tendency to be sensitive to small changes in the data set and not generalise well. When the performance of a standard decision tree is cross-validated with other data they tend to fall short.

Regularisation and pruning

In order to mitigate the high variance of CARTs, a technique known as pruning can be employed to reduce the complexity of the tree, and likelihood of overfitting [69]. Leaf nodes are pruned on the basis of the misclassification error given by Eq.2.44:

$$E(t) = 1 - [max(p(i|t))] \tag{2.44}$$

Where:

E = Classification error

t = A given tree structure

i = A given decision or class

Essentially if the split does not result in any improvement or is deemed redundant: it is pruned. Regularisation is the act of limiting the complexity of a predicting model in any sense, limiting the depth of the tree is another common method of limiting complexity in CARTS which directs the model towards a more general solution to the problem. This is typically achieved through the optimisation of an objective function:

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{2.45}$$

Where:

obj = The objective function to be minimised or maximised

θ = The set of parameters used in the model

L = The error term associated with the model in the training stage

Ω = The regularisation term which regulates the complexity of the model.

Bagging and Random Forest

Bagging is a general ML technique which seeks to combine the performance of many weaker classifiers into an overall more capable macro-classifier. This is achieved by training each sub classifier on a subset of the samples in the data set. The decision of each sub classifier is factored into the overall decision by either voting or taking an average of the output. This reduces the variance of the classifier and can improve the performance greatly [68]. A random forest classifier seeks to expand upon this further by decreasing the correlation between the decisions of each sub-classifier. This is achieved by limiting the number of

features given to each tree so that each tree effectively makes a decision on different features.

Boosting

Boosting in contrast to bagging and Random Forest is a technique designed to reduce the bias of a classifier [70]. Bias is a measure of how well a classifier captures the necessary information needed to do its job [60], if bias is high it can cause the classifier to miss important information and lead to underfitting. To avoid underfitting, a classifier is optimised using using eq. (2.45). In the case of a CART this would be the structure of the tree itself; however due to the heuristic nature of a decision tree, optimisation can become difficult. In order to optimise a decision tree, boosting is performed. Boosting is achieved by firstly classifying the set of test points; the points which are misclassified are then weighted higher so that more focused is placed on classifying them in the next iteration. This process is repeated until the classifier is able to correctly identify the data set to the required standard. Each iteration in the process is then weighted according to a learning rate λ and then used to give an overall decision.

Extreme Gradient Boosting (XGBoost)

XGBoost is highly optimised supervised learning method which builds upon gradient boosted decision trees [71]. It is a highly successful algorithm and one of the most commonly used ensemble methods used by many data science competition winning teams [72]. It is an *ensemble* classifier, a method which is characterised by a meta-classifier, which uses the input of many sub classifiers known as *weak learners*. The weak learners in this case are decision tree

classifiers which have been enhanced through the use of two ensemble techniques: *bagging* and *boosting*. Another built in feature known as a *regularised learning objective* penalises an overly complex model allowing the classifier to generalise more effectively.

2.3.5 Evaluation of Classifier Performance

Evaluating a predictive classifier is an extremely important step in the process of development. The standard format for training a classifier is to have a training and testing set – one to fit a classifier to, and one to evaluate the performance of the fitted classifier on. The performance achieved from this set up however is not descriptive of the entire data set and is actually wasting some of the data. To overcome this a technique known as *cross-validation* can be used. Cross-validation works simply by selecting a different training and testing set multiple times and then aggregating the results in the desired way.

To gain a true indication of the performance of a classifier it is not sufficient to look at the accuracy; in the case of a binary classifier a confusion matrix is often employed to see exactly what predictions have been made. Shown in table 2.1 are terms used to refer to types of classification result:

TABLE 2.1: Statistical classification terms derived from a confusion matrix.

Statistic	Symbol	Description
Positives	P	The number of positive cases
Negatives	N	The number of negative cases
True Positives	TP	Cases correctly predicted as positive
True Negatives	TN	Cases correctly predicted as negative
False Positives	FP	Cases incorrectly predicted as positive
False Negatives	FN	Cases incorrectly predicted as negative

A confusion matrix is a way of visualising predicted and actual values obtained from a classifier. It allows for a greater level of insight when diagnosing the cause of classification errors.

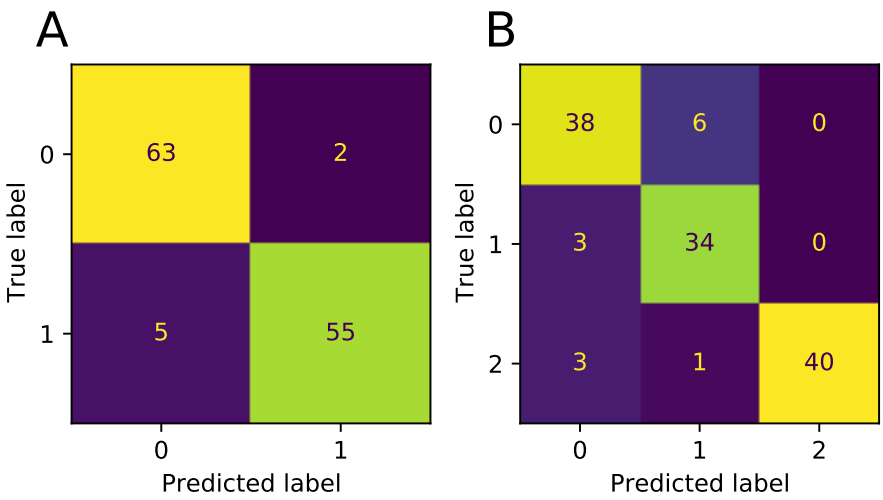

FIGURE 2.26: Binary (A) and multiclass (B) confusion matrices showing classification results.

When evaluating the performance of a classifier a number of metrics can be derived from those shown in table 2.1. The use of these statistics is common in clinical sciences and bioinformatics, and is the common language in which the performance of diagnostics tests are communicated [53, 73]. In a diagnostic test, sensitivity is a measure of the ability of a test to identify true positives; equivalently the specificity of a diagnostic test is a measure of how well a test can identify true negatives. Sensitivity and specificity are inextricably linked and thus there is a trade off between the performance of either scores – an increase in one score typically involves a decrease in the other. A similar pair of metrics are the positive predictive value (PPV) and negative predictive value (NPV). These metrics measure the ratio of true positives/negatives to the total number of positives/negatives. Finally, the Matthews correlation coefficient (MCC) is a more holistic measure of the performance of a diagnostic test and considers all prediction outcomes. A summary of a these metrics is given below.

TABLE 2.2: Classification statistics used in the evaluation of predictive models.

Statistic	
Accuracy	$\frac{(TP+TN)}{P+N}$
Matthews correlation coefficient	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Specificity	$\frac{TN}{TN+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Positive predictive value	$\frac{TP}{TP+FP}$
Negative predictive value	$\frac{TN}{TN+FN}$

The majority of classifier algorithms output an estimate as a continuous value. In the case of logistic regression this estimate can be directly interpreted as a probability, in other cases this estimate is derived in a non-probabilistic way and must be used with caution. These continuous values must be turned into a boolean prediction by taking a threshold over the value.

Receiver Operating Characteristic (ROC) analysis

It is possible to gain an estimate of the classification power of a diagnostic test irrespective of the threshold by calculating a receiver operating characteristic (ROC) curve. A ROC curve can be employed to calculate the area under the receiver operating characteristic curve (AUROC) statistic, a widely accepted measure of classifier performance. A ROC curve is a plot of true positive rate (TPR) (sensitivity) against the false positive rate (FPR) (1 - specificity) for a number of decision thresholds for a binary classifier; this can be extended to a multi-class problem by utilising a ”one against all” approach.

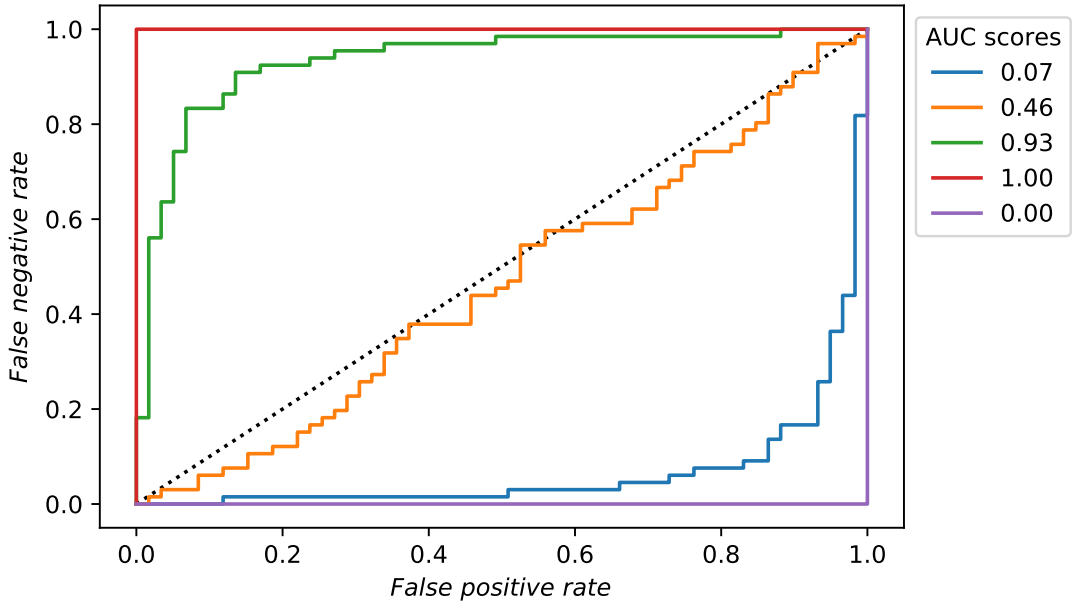

FIGURE 2.27: ROC curves showing a comparison between a number of classifiers with calculated AUROC scores.

A value across the diagonal signifies a classifier which is equivalent to pure chance and effectively useless. A steep curve that approaches the top-left corner is a realistic, well-performing classifier which will have a high AUROC. An AUROC value of 1 signifies a flawless classifier which correctly identifies every instance, more importantly it does not need to make a trade off between specificity and sensitivity; a value of 0.5 is equivalent to a random guess signifying a completely useless classifier. A classifier with a value of less than 0.5 is actually useful as it consistently predicts incorrectly; as this statistic applies to a binary classifier the prediction can be inverted to give a useful result. A ROC curve is a collection of TPR and FPR values for a number of cut-off points. The curve represents the trade off between obtaining more true positives, and fewer false positives and so allows for a visual evaluation of the classifier performance over the entire range of thresholds.

Precision Recall analysis

Similar to the ROC curve is the precision-recall (PR) curve; The precision is equivalent to the PPV, and recall is equivalent to the sensitivity score. The PR curve is formed in the same way as a ROC curve by calculating the precision and recall at a number of thresholds. When examining a highly imbalanced dataset a PR curve can avoid overly optimistic estimates of classifier performance [74]. The PR curve is particularly useful as it considers the precision/PPV which is itself dependent upon the prevalence of positive cases in the data set. The area under the precision recall curve (AUPRC) can be utilised as a summary statistic similar to the AUROC score.

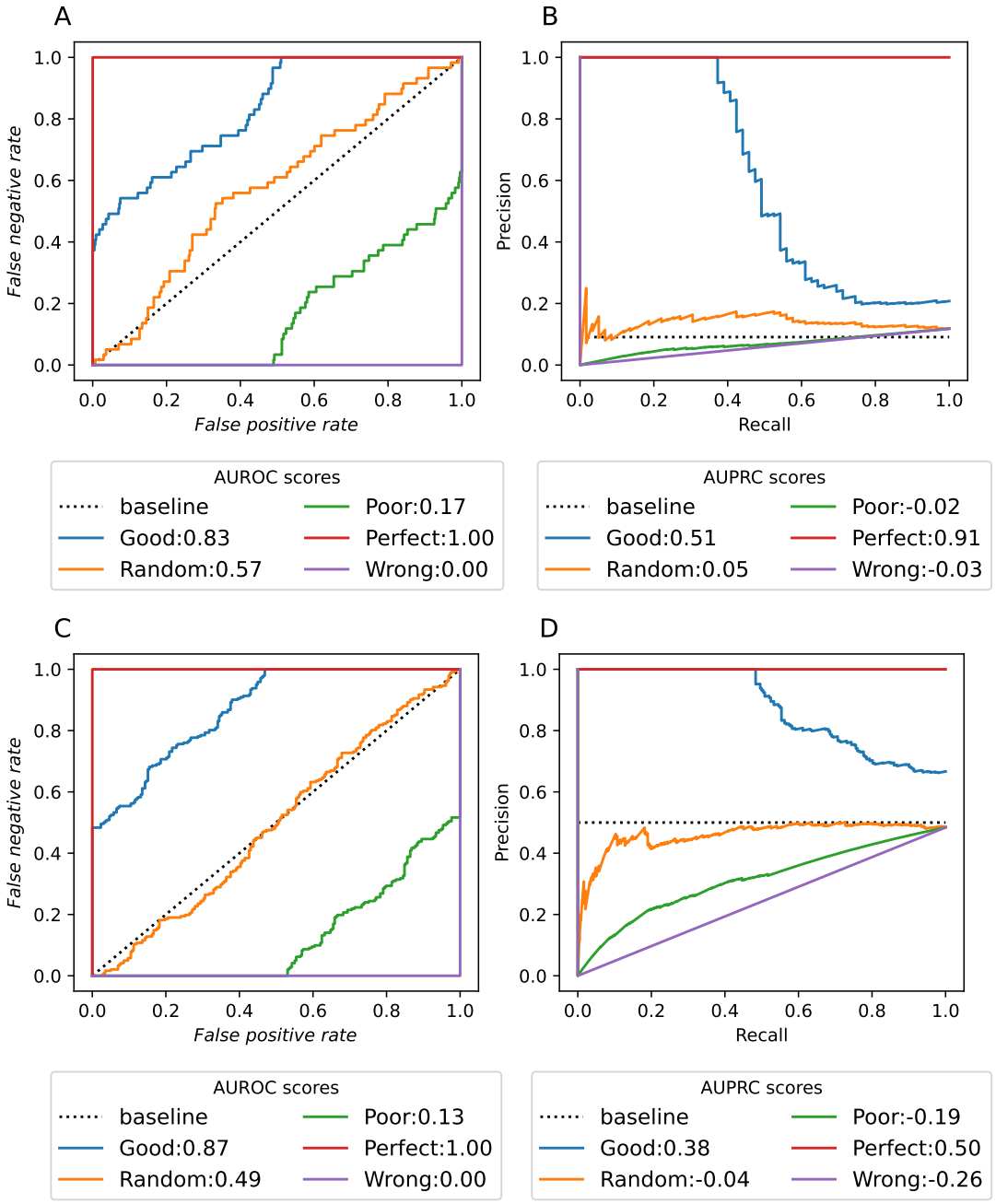

FIGURE 2.28: Classifiers of varying utility evaluated on simulated imbalanced [A,B] and balanced [C,D] datasets.

Figure 2.28[A-B] shows an evaluation of a number of classifiers of varying utility on a simulated imbalanced dataset, where the ratio of the positive to negative class is 1:10. The baseline score in fig. 2.28[B] is set at 0.09 to reflect the imbalance and is calculated by:

$$baseline = \frac{P}{(P+N)} \tag{2.46}$$

Figure 2.28[C-D] shows a simulated dataset of equal class distribution where the baseline is set at 0.5.

This baseline is the substracted from the calculated AUPRC to give the final scores where negative scores indicate a poor classifier. Figure 2.28[A,C] show ROC curves for these predictions but give no real insight into the consequences of class imbalance. Incorrect conclusions could be drawn from Figure 2.28[B] if the baseline was not adjusted for class prevalence and set at 0.5. A consequence of this baseline adjustment is that when comparing biomarkers evaluated on cohorts with differing class distributions the prevalence of the dataset must be taken into account.

Bibliography

[1] National Cancer Institute. What is cancer?, Feb 2015.

[2] Hajdu Steven I. A note from history: Landmarks in history of cancer, part 1. *Cancer*, 117(5):1097–1102.

[3] Freddie Bray and Bjørn Møller. Predicting the future burden of cancer. *Nature Reviews Cancer*, 6(1):63–74, 2006.

[4] Health Data. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980âĂŞ2015: a systematic analysis for the global burden of disease study 2015., 2015.

[5] Preetha Anand, Ajaikumar B. Kunnumakara, Chitra Sundaram, Kuzhuvelil B. Harikumar, Sheeja T. Tharakan, Oiki S. Lai, Bokyung Sung, and Bharat B. Aggarwal. Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical Research*, 25(9):2097–2116, 2008.

[6] Jun Yokota. Tumor progression and metastasis. *Carcinogenesis*, 21(3):497–503, 2000.

[7] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646 – 674, 2011.

[8] Nikki Cheng, Anna Chytil, Yu Shyr, Alison Joly, and Harold L Moses. TGF-β signaling deficient fibroblasts enhance Hepatocyte Growth Factor signaling in mammary carcinoma cells to promote scattering and invasion. *Molecular cancer research : MCR*, 6(10):1521–1533, 2008.

[9] Deborah L. Burkhart and Julien Sage. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nature Reviews Cancer*, 8(9):671–682, 2008.

[10] J. M. Adams and S. Cory. The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene*, 26(9):1324–1337, 2007.

[11] Maria A. Blasco. Telomeres and human disease: Ageing, cancer and beyond. *Nature Reviews Genetics*, 6(8):611–622, 2005.

[12] Baeriswyl Vanessa and Christofori Gerhard. The angiogenic switch in carcinogenesis. *Semin Cancer Biol*, 19(5):329–37, 2009.

[13] Khalid O. Alfarouk, Daniel Verduzco, Cyril Rauch, Abdel Khalig Muddathir, Adil H.H. Bashir, Gamal O. Elhassan, Muntaser E. Ibrahim, Julian David Polo Orozco, Rosa Angela Cardone, Stephan J. Reshkin, and Salvador Harguindey. Glycolysis, tumor metabolism, cancer growth and dissemination. A new pH-based etiopathogenic perspective and therapeutic approach to an old cancer question. *Oncoscience*, 1(12):777–802, 2014.

[14] María Berdasco and Manel Esteller. Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry. *Developmental Cell*, 19(5):698–711, 2010.

[15] Anastasios K Markopoulos. Current Aspects on Oral Squamous Cell Carcinoma. *The Open Dentistry Journal*, 6(1):126–130, 2012.

[16] A Cawson, R and W Odell, E. *Oral Pathology and Oral Medicine*. Cawson2008, 8th editio edition, 2008.

[17] Jamshid Jalouli, Salah O. Ibrahim, Ravi Mehrotra, Miranda M. Jalouli, Dipak Sapkota, Per-Anders Larsson, and Jan-M. Hirsch. Prevalence of viral (hpv, ebv, hsv) infections in oral submucous fibrosis and oral cancer from india. *Acta Oto-Laryngologica*, 130(11):1306–1311, 2010.

[18] M. A. Gonzalez-Moles, J. Gutierrez, M. J. Rodriguez, I. Ruiz-Avila, and A. Rodriguez-Archilla. Epstein-Barr virus latent membrane protein-1 (LMP-1) expression in oral squamous cell carcinoma. *Laryngoscope*, 112(3):482–487, 2002.

[19] Miguel H. Bronchud, MaryAnn Foote, Giuseppe Giaccone, Olufunmilayo I. Olopade, and Paul Workman, editors. *Principles of Molecular Oncology*. Humana Press, Totowa, NJ, nov 2004.

[20] World Health Orgnization. Biomarkers and risk assessment: concepts and principles. Environmental Health Criteria 155. *Environmental Health Criteria*, (155):82, 1993.

[21] Kewal K. Jain. *The handbook of biomarkers*. 2010.

[22] G. Orchard and B. Nation. *Histopathology*. Fundamentals of Biomedical Science. OUP Oxford, 2011.

[23] S Warnakulasuriya, J Reibel, J Bouquot, and E Dabelsteen. Oral epithelial dysplasia classification systems: Predictive value, utility, weaknesses and scope for improvement. *Journal of Oral Pathology and Medicine*, 37(3):127–133, 2008.

[24] Daniel C. Paech, Adèle R. Weston, Nick Pavlakis, Anthony Gill, Narayan Rajan, Helen Barraclough, Bronwyn Fitzgerald, and Maximiliano Van Kooten. A systematic review of the interobserver variability for histology in the differentiation between squamous and nonsquamous non-small cell lung cancer. *Journal of Thoracic Oncology*, 6(1):55–63, 2011.

[25] Hani A Alturkistani, Faris M Tashkandi, and Zuhair M Mohammedsaleh. Histological Stains: A Literature Review and Case Study. *Global Journal of Health Science*, 8(3):72, 2015.

[26] Michael Pilling and Peter Gardner. Fundamental developments in infrared spectroscopic imaging for biomedical applications. *Chem. Soc. Rev.*, 45(7):1935–1957, 2016.

[27] Michael J. Pilling, Alex Henderson, Benjamin Bird, Mick D. Brown, Noel W. Clarke, and Peter Gardner. High-throughput quantum cascade laser

(QCL) spectral histopathology: a practical approach towards clinical translation. *Faraday Discuss.*, 187:135–154, 2016.

[28] Júlio Trevisan, Plamen P. Angelov, Paul L. Carmichael, Andrew D. Scott, and Francis L. Martin. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *The Analyst*, 137(14):3202, 2012.

[29] H. Fabian, P. Lasch, M. Boese, and W. Haensch. Infrared microspectroscopic imaging of benign breast tumor tissue sections. *Journal of Molecular Structure*, 661-662(1-3):411–417, 2003.

[30] Matthew J. Baker, Hugh J. Byrne, John Chalmers, Peter Gardner, Royston Goodacre, Alex Henderson, Sergei G. Kazarian, Francis L. Martin, Julian Moger, Nick Stone, and Josep Sulé-Suso. Clinical applications of infrared and Raman spectroscopy: state of play and future challenges. *The Analyst*, (1985):1735–1757, 2018.

[31] Matthew J. Baker, Caryn S. Hughes, and Katherine A. Hollywood. *Biophotonics: Vibrational spectroscopic diagnostics*. 2016.

[32] F. Mandl. *Quantum mechanics*, volume 23. John Wiley & Sons, Manchester, 1st edition, 1992.

[33] Matthew J. Baker, Caryn S. Hughes, and Katherine A. Hollywood. *Biophotonics: Vibrational spectroscopic diagnostics*. 2016.

[34] Tim Soderberg. 4.3: Infrared spectroscopy, Jul 2020.

[35] Brian C. Smith. *Fundamentals of fourier transform infrared spectroscopy, second edition*. 2011.

[36] M. J. Baker, E. Gazi, M. D. Brown, J. H. Shanks, P. Gardner, and N. W. Clarke. FTIR-based spectroscopic analysis in the identification of clinically

aggressive prostate cancer. *British Journal of Cancer*, 99(11):1859–1866, 2008.

[37] Andreas Barth. Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta - Bioenergetics*, 1767(9):1073–1101, 2007.

[38] Peter R. Griffiths and James A. De Haseth. Fourier transform Raman spectrometry. *Chemical Analysis*, 171:375–393, 2007.

[39] Anand Subramanian and Luis Rodriguez-Saona. *Fourier Transform Infrared (FTIR) Spectroscopy*, volume volume. Elsevier Inc., 1 edition, 2009.

[40] A. D. Smith, M. R F Siggel-King, G. M. Holder, A. Cricenti, M. Luce, P. Harrison, D. S. Martin, M. Surman, T. Craig, S. D. Barrett, A. Wolski, D. J. Dunning, N. R. Thompson, Y. Saveliev, D. M. Pritchard, A. Varro, S. Chattopadhyay, and P. Weightman. Near-field optical microscopy with an infra-red free electron laser applied to cancer diagnosis. *Applied Physics Letters*, 102(5):1–5, 2013.

[41] RP Photonics Consulting. Quantum cascade lasers.

[42] Michael J. Walsh, Maneesh N. Singh, Helen F. Stringfellow, Hubert M. Pollock, Azzedine Hammiche, Olaug Grude, Nigel J. Fullwood, Mark A. Pitt, Pierre L. Martin-Hirsch, and Francis L. Martin. FTIR microspectroscopy coupled with two-class discrimination segregates markers responsible for inter- and intra-category variance in exfoliative cervical cytology. *Biomarker Insights*, 2008(3):179–189, 2008.

[43] Christoph Krafft, Matthias Kirsch, Claudia Beleites, Gabriele Schackert, and Reiner Salzer. Methodology for fiber-optic Raman mapping and FTIR imaging of metastases in mouse brains. *Analytical and Bioanalytical Chemistry*, 389(4):1133–1142, 2007.

[44] C. Beleites, G. Steiner, M. G. Sowa, R. Baumgartner, S. Sobottka, G. Schackert, and R. Salzer. Classification of human gliomas by infrared

imaging spectroscopy and chemometric image processing. *Vibrational Spectroscopy*, 38(1-2):143–149, 2005.

[45] Qingbo Li, Can Hao, Xue Kang, Jialin Zhang, Xuejun Sun, Wenbo Wang, and Haishan Zeng. Colorectal Cancer and Colitis Diagnosis Using Fourier Transform Infrared Spectroscopy and an Improved K-Nearest-Neighbour Classifier. *Sensors*, 17(12):2739, 2017.

[46] Elodie Ly, Olivier Piot, Anne Durlach, Philippe Bernard, and Michel Manfait. Differential diagnosis of cutaneous carcinomas by infrared spectral micro-imaging combined with pattern recognition. *The Analyst*, 134(6):1208, 2009.

[47] M Diem, L Chiriboga, and H Yee. Infrared spectroscopy of human cells and tissue. VIII. Strategies for analysis of infrared tissue mapping data and applications to liver tissue. *Biopolymers*, 57(5):282–290, 2000.

[48] Kevin Yeh, Seth Kenkel, Jui Nung Liu, and Rohit Bhargava. Fast infrared chemical imaging with a quantum cascade laser. *Analytical Chemistry*, 87(1):485–493, 2015.

[49] Robert C. Dunn. Near-Field Scanning Optical Microscopy. *Chemical Reviews*, 99(10):2891–2928, 1999.

[50] I. Gris-Sánchez, D. Van Ras, and T. A. Birks. The Airy fiber: an optical fiber that guides light diffracted by a circular aperture. *Optica*, 3(3):270, 2016.

[51] T.B. Trafalis and H. Ince. Support vector machine for regression and applications to financial forecasting. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, (x):348–353 vol.6, 2000.

[52] Igor Kononenko. Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

[53] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[54] Peter Lasch. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometrics and Intelligent Laboratory Systems*, 117(August 2012):100–114, 2012.

[55] Abraham Savitzky and Marcel J.E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 1964.

[56] Rong Wang and Yong Wang. Fourier transform infrared spectroscopy in oral cancer diagnosis. *International Journal of Molecular Sciences*, 22(3):1–21, 2021.

[57] Tomasz P. Wrobel, Danuta Liberda, Paulina Koziol, Czeslawa Paluszkiewicz, and Wojciech M. Kwiatek. Comparison of the new Mie Extinction Extended Multiplicative Scattering Correction and Resonant Mie Extended Multiplicative Scattering Correction in transmission infrared tissue image scattering correction. *Infrared Physics and Technology*, 107(March):103291, 2020.

[58] Johanne H. Solheim, Evgeniy Gunko, Dennis Petersen, Frederik Großerüschkamp, Klaus Gerwert, and Achim Kohler. An open-source code for Mie extinction extended multiplicative signal correction for infrared microscopy spectra of cells and tissues. *Journal of Biophotonics*, 12(8):1–14, 2019.

[59] A. Köhler, J. Sulé-Suso, G. D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. G. Van Pittius, G. Parkes, and H. Martens. Estimating and correcting Mie scattering in synchrotron-based microscopic fourier transform infrared spectra by extended multiplicative signal correction. *Applied Spectroscopy*, 62(3):259–266, 2008.

[60] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 53. 2013.

[61] I T Jolliffe. Principal Component Analysis, Second Edition. *Encyclopedia of Statistics in Behavioral Science*, 30(3):487, 2002.

[62] Sebastian Raschka. Linear discriminant analysis - bit by bit, aug 2014.

[63] James Franklin. *The elements of statistical learning: data mining, inference and prediction*, volume 27. 2005.

[64] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn , Keras , and*. O'Reilly Media.

[65] Shan Suthaharan. *Machine Learning Models and Algorithms for Big Data Classification*, volume 36 of *Integrated Series in Information Systems*. Springer US, Boston, MA, jun 2016.

[66] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, dec 1943.

[67] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Internal Representations by Error Propagation. *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*, pages 399–421, 2013.

[68] Wei-yin Loh. Encyclopedia of Statistics in Quality and Reliability. pages 315–323, 2008.

[69] David Hutchison and John C Mitchell. *Lecture Notes in Computer Science*. 2011.

[70] Robert E Schapire. The boosting approach to machine learning: an overview. *Nonlinear Estimation and Classification*, 171:149–171, 2003.

[71] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[72] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. 2016.

[73] David Powers. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):1–24, 2011.

[74] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):1–21, 2015.