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

Despite enormous advances in research in recent decades, the burden of cancer still exists. Additional treatment options for oral squamous cell carcinoma patients exist, but survival outcomes have stagnated in recent years. The development of neo-adjuvant therapies is hindered by the difficulty of identifying cases eligible for window trials. Current prognostic biomarkers are insufficient; thus, a more comprehensive range of prognostic tools is required to facilitate improvements to direly-needed therapies. This thesis aims to develop statistical analysis techniques to analyse Fourier transform infrared (FTIR) spectroscopy data to obtain effective new prognostic tools.

A direct comparison between FTIR data and an existing prognostic biomarker, α-smooth muscle actin (ASMA) was made. Two statistical models were created using each set of variables in addition to a third, hybrid model comprising both sets of variables in combination. A rigorous analysis procedure proved that a logistic regression model utilising FTIR data was capable of predicting prognostic outcomes much more effectively than the ASMA model. The hybrid model also demonstrated good prognostic utility, suggesting that combining a variety of prognostic biomarkers may be an effective strategy for the future.

An optimisation framework was developed to address the lack of consensus on the choice of preprocessing methods currently used. This framework could benefit the wider research community by asserting a standardised sequence of steps for use with FTIR data. The use of the optimisation framework resulted in substantial improvements in classification scores and reaffirmed some conventional wisdom surrounding preprocessing. The adoption of FTIR spectroscopy in a clinical setting could be expedited considerably by this framework by standardising the process of FTIR-based biomarker discovery.

The usage of deep learning based models has grown considerably in medical diagnostics in recent years. A one-dimensional convolutional neural network (CNN) and multilayer perceptron (MLP) model were investigated as prognostic tools. Both models showed promise as effective models; the CNN model showed exceptionally high scores suggesting further research into convolution-based methods could be a fruitful future avenue of research.

Acknowledgements

I would like to thank those colleagues whose guidance has been invaluable to the success of my thesis. Steve, the mentoring you have given me throughout my PhD has helped me to become a better scientist. From my masters project to this thesis, you have been hugely supportive. I want to extend my sincere thanks to you Janet. I will always be grateful for how you went out of your way to support my research and for the many hours you spent helping me annotate samples and proofread my work — I cannot thank you enough. Peter, your kind words, advice, and optimism have been an enormous help over the past four years.

Barney, I doubt I would have made it through the PhD without our friendship. When times were difficult, your sense of humour was a lifeline. To my colleagues Safaa, Caroline, Paul Harrison, David, Phil, Paul Unsworth thank you all so much for your support with anything I ever needed help with; it has been a pleasure to have worked with you all.

To Tara, I cannot thank you enough for your patience and willingness to listen when I needed it most.

To my Mum and Dad, your love and guidance are the only reasons I am where I am today. From buying me my first computer and letting me tinker away with anything I laid my hands on, you were always there to encourage my passions and to help me pursue my dreams — dwi'n caru chi. To my brother Matt, you've always been my biggest supporter. I am proud to have you as my brother and am truly grateful for your ability to see the best in me.

Contents

List of Figures

List of Tables

List of Abbreviations

FTIR	**F**ourier **T**ransform **I**nfra **R**ed
ASMA	**A**lpha **S**mooth **M**uscle **A**ctin
CNN	**C**onvolutional **N**eural **N**etwork
MLP	**M**ulti **L**ayer **P**erceptron
IR	**I**nfra **R**ed
HPV	**H**uman **P**apilloma **V**irus
OSCC	**O**ral **S**quamous **C**ell **C**arcinoma
TMA	**T**issue **M**icro **A**rray
SHO	**S**imple **H**armonic **O**scillator
MCT	**M**ercury **C**admium **T**elluride
QCL	**Q**uantum **C**ascade **L**aser
FEL	**F**ree **E**lectron **L**aser
FPA	**F**ocal **P**lane **A**rray
ML	**M**achine **L**earning
PCA	**P**rincipal **C**omponent **A**nalysis
FFT	**F**ast **F**ourier **T**ransform
EMSC	**E**xtended **M**ultiplicative **S**cattering **C**orrection
LDA	**L**inear **D**iscriminant **A**nalysis
LR	**L**ogistic **R**egression
SVM	**S**uport **V**ector **M**achine
RBF	**R**adial **B**asis **F**unction
SVC	**S**upport **V**ector **C**lassifier
ANN	**A**rtificial **N**eural **N**etwork
CART	**C**lassification **A**nd **R**egression **T**ree
XGB	e**X**treme **G**radient **B**oosting
TN	**T**rue **N**egative
FP	**F**alse**P**ositive
FN	**F**alse **N**egative
PPV	**P**ositive **P**redictive **V**alue
NPV	**N**egative **P**redictive **V**alue
MCC	**M**atthews **C**orrelation **C**oefficient
ROC	**R**eceiver **O**perating **C**haracteristic
AUROC	**A**rea **U**nder the **R**eceiver **O**perating **C**haracteristic
TPR	**T**rue **P**ositive **R**ate
FPR	**F**alse **P**ositive **R**ate
PR	**P**recision **R**ecall
AUPRC	**A**rea **U**nder the **P**recision **R**ecall **C**urve
HNSCC	**H**ead & **N**eck **S**quamou **C**ell **C**arcinoma
ECS	**E**xtra **C**apsular **S**pread
FFPE	**F**ormalin **F**ixed **P**araffin **E**mbedded

GA	**G**enetic **A**lgorithm
ATR	**A**ttenuated **T**otal **R**eflection
RF	**R**andom **F**orest
FFS	**F**orward **F**eature **S**election
GP	**G**aussian **P**rocess
SGD	**S**tochastic **G**radient **D**escent
NIR	**N**ear **I**nfra **R**ed

1 Introduction

Cancer is a leading cause of death, and an obstacle to increasing life expectancy worldwide [1]. Cancer is the cause of nearly 10 million deaths a year globally — up from 8.8 million in 2015 [2]. The incidence of cancer is set to increase globally, reflecting the growth and aging of the human population.

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. This is due to the ineffectiveness of current treatment plans [4, 5]. Neo-adjuvant therapy has the potential to improve prognoses, and aid clinical decision making if applicable cases can be determined at the time of diagnosis.

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 pathologist's 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 for an experienced histopathologist to make an accurate diagnosis [6]. 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 [7, 8] 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 [9, 10], and chemical imaging [11].

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 [12]. Identification of these cases is

paramount to minimising the adverse effects of treatment, and improving patient outcomes. Previous work [13, 14, 15, 16] 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 [17]. 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 test 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 [18].

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 past four years 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 3 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 4 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 5 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. With a surge in usage for applications in medical diagnostics, deep learning is unique in the flexibility of network designs and wide applicability to different sources of data. deep learning proved to be a viable statistical technique for use as a prognostic tool, as evidenced by a thorough analysis routine. The chapter covers the development, optimisation, and evaluation of two differing types of neural network architecture for use as prognostic tools.

Bibliography

[1] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

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

[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] Tseng-Cheng Chen, Chen-Tu Wu, Cheng-Ping Wang, Wan-Lun Hsu, Tsung-Lin Yang, Pei-Jen Lou, Jenq-Yuh Ko, and Yih-Leong Chang. Associations among pretreatment tumor necrosis and the expression of hif-1α and pd-l1 in advanced oral squamous cell carcinoma and the prognostic impact thereof. *Oral Oncology*, 51(11):1004–1010, 2015.

[5] Anthony C Nichols, Pencilla Lang, Eitan Prisman, Eric Berthelet, Eric Tran, Sarah Hamilton, Jonn Wu, Kevin Fung, John R de Almeida, Andrew Bayley, et al. Treatment de-escalation for hpv-associated oropharyngeal squamous cell carcinoma with radiotherapy vs. trans-oral surgery (orator2): study protocol for a randomized phase ii trial. *BMC cancer*, 20(1):1–13, 2020.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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 several key areas covered in this thesis. Additional detail will be given in later chapters where appropriate. Section 2.1 will cover some fundamental aspects of oncology and provide an overview of the field of histology and biomarker discovery. An overview of the sample preparation process and implications for measurements will 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. 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 that share the characteristics of abnormal cellular growth and a tendency to spread into surrounding tissue [1]. The first description of 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 symptoms, 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 Figure 2.1:

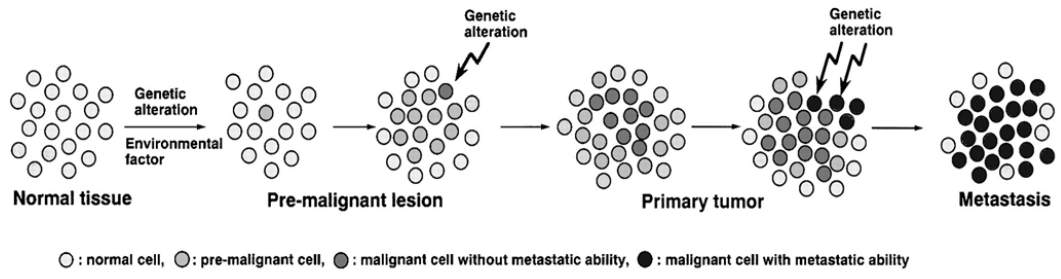


FIGURE 2.1: Malignant Progession. [6]

Malignant progression occurs in 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 [6, 7]:

2.1.2 The Hallmarks of cancer

Capabilities

As normal cells progress towards malignancy, a sequence of hallmarks is typically acquired. A tumour comprises a complex mass of numerous distinct cell types, interacting with each other in varying ways. Alongside these malignant cells are normal cells that contain 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 because 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 To proliferate, cancerous cells must also avoid tumour suppressing signals. Two proteins moderate these tumour suppressant signals: the RB1 *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 [10]. TP53 is sensitive to indicators of stress and damage within tumour cells and can 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 the defective cells. These two proteins form independent and redundant tumour suppressant systems. Therefore, a cell must 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 that a cell undergoes when significant cell stress is detected from within the cell due to DNA damage or signalled from other cells. In the case where apoptosis is triggered in a defective cell, those which have acquired the ability can avoid programmed death [11]. Due to acquired mutations, tumour cells can resist apoptosis by becoming de-sensitised to internal and external signalling. The level of attenuation of apoptosis in tumours has been shown to be severe in tumours that are well-progressed [11].

Biological immortality Malignant cells differ from normal cells in their ability to circumvent the states of senescence (cell ageing) and crisis (cell death) [12]. These processes prevent 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 biologically immune cells is the presence of telomerase which prohibits telomere shortening of chromosomes within the cell [7] and senescence and crisis from stopping uncontrolled proliferation. A large body of evidence suggests that the presence of telomerase allows for the 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 necessary nutrients and oxygen and remove waste products. Angiogenesis is the process of creating this blood supply by activating an ”angiogenic switch” [13]. 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 [14]. These changes allow neoplastic cells to obtain more significant 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 to avoid detection by the immune system and resist interference. It is important to note that several 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 immune 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 developing these changes possess an advantage over neighbouring cells, making reproduction more likely and establishing a cancer lineage. Alongside genetic changes, epigenetic changes such as DNA methylation and histone modifications have been linked to cancer cells [15]. 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 adversely affecting systems that monitor genomic integrity.

Inflammation of surrounding tissue Other than innate 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 provide enzymes that 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 how a tumour will develop is not possible; it depends at least in part on 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.

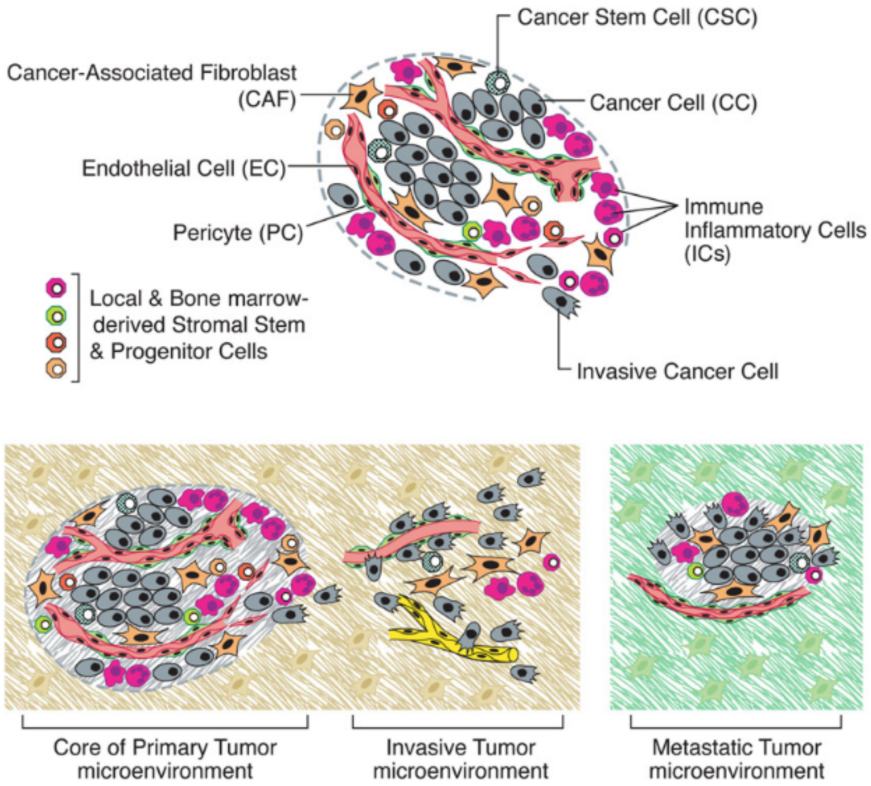


FIGURE 2.2: (Upper) The microenvironment of the tumour showing a complex array of interacting cell types. (Lower) Distinct microenvironments in which neoplastic cells are typically found. These environments develop progressively through the duration of the lineage of a collection of neoplastic cells [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. The hallmarks discussed above will lead to changes in the tumour microenvironment which will lead to chemical changes which can be quantified using FTIR spectroscopy.

2.1.4 Oral Cancer

Whilst the work covered in this thesis seeks to develop methods that apply 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 tissue growth 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 characterised by hardness or fixation [16, 17].

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 smoking and alcohol, other risks factors such as betel quid chewing and various narcotics are associated with the development of oral cancer. The human papilloma virus (HPV) is widely reported as a virus that carries oncogenetic potential. However, results are conflicting as to the true extent of this potential [16]. It has been noted previously that some HPV genomes have been incorporated into oral cancer cells [17], but this has not yet proven to be a useful diagnostic variable when employed as a screening test [17]. Epstein-Barr virus [18] and Hepatitis C Virus (HCV) [19] are also considered to be viruses whose oncogenetic potential is felt 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 most significant clinical need lies. This is due to the ineffectiveness of current treatment plans [20, 21]. Neo-adjuvant therapy has the potential to improve prognoses and aid clinical decision making if applicable cases can be determined at the time of diagnosis.

The influence of individual genetics is also widely recognised as an influencing factor in developing oral squamous cell carcinoma (OSCC). Mice with a genetic predisposition to developing OSCC have been bred — suggesting

a genetic causation. However, the genetics of oral cancer is complex, and a causative genetic link has not been firmly established in humans [17]. 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. This is thought to be due to significant risk factors containing chemicals known to alter DNA — leading to many of the aforementioned hallmarks.

Of particular relevance to the work contained in this thesis is the occurrence of different pathological sites in which OSCC typically develops. The most common site of occurrence of OSCC overall is the lower lip, with the most common site within the mouth being the tongue [17]. Within the oral cavity, OSCC is particularly prevalent in the lower mouth, along the borders of the tongue, the 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 [17].

2.1.5 Molecular Oncology

Molecular oncology is the interdisciplinary approach to cancer treatment that focuses on the effects of tumours at the molecular scale. As an interdisciplinary 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 to improve patient outcomes. However, the impact of molecular oncology is not solely limited to the development of treatments, 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 [22].

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 the limited resources available to dedicate to the pursuit. This issue is exacerbated even further due to heterogeneity present in many tumours, meaning that singular biomarkers are often ineffective on their own. FTIR microscopy is a high-throughput, objective, and relatively inexpensive technology; combined 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; 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 [23, 24]. Biomarkers may be present in any part of the body, including bodily fluids such as blood serum, urine, cerebrospinal fluid; biomarkers may also be found in any tissue situated in the body. Many currently used biomarkers are found in bodily fluids and are a standard diagnostic tool employed by clinicians for many 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 to enhance contrast in desired regions of the image. 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 [24]. 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 Figure 2.3.

Several requirements must be met for a biomarker to translate to a clinical setting. An ideal biomarker achieves the following:

- It is specifically associated with the presence or state of a disease and can 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 biomarker's ability to do so could lead physicians to make decisions on a patient's treatment, which may be useless or detrimental to their well-being.

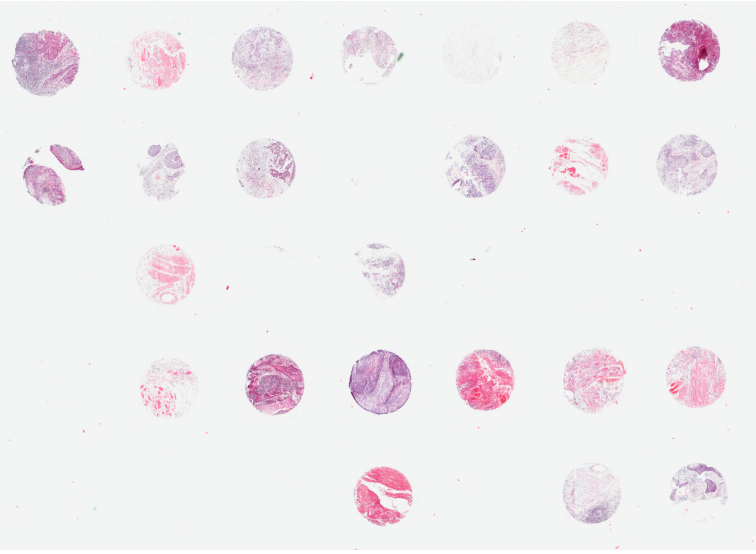


FIGURE 2.3: An example of a TMA showing cores taken from resected tumour tissue arranged in a grid like fashion. Cores have been stained with H&E to show contrast in protein and nucleic acid concentrations.

2.1.7 Histology

Pathology is the branch of medicine that is concerned with the study of disease by studying patient samples (urine, blood, tissue, etc.) to aid or provide diagnosis or prognosis [25]. The main focuses of pathology are evaluating structural and functional changes in patient samples. In the UK NHS, 80-90% of diagnoses performed are based on information gained from laboratory-based medical specialists [25].

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 microscope slides by a highly-skilled histopathologist. A standard method of diagnosis is by examining H&E stained tissue samples using an optical microscope. An example of a H&E stained image is shown in Figure 2.4.

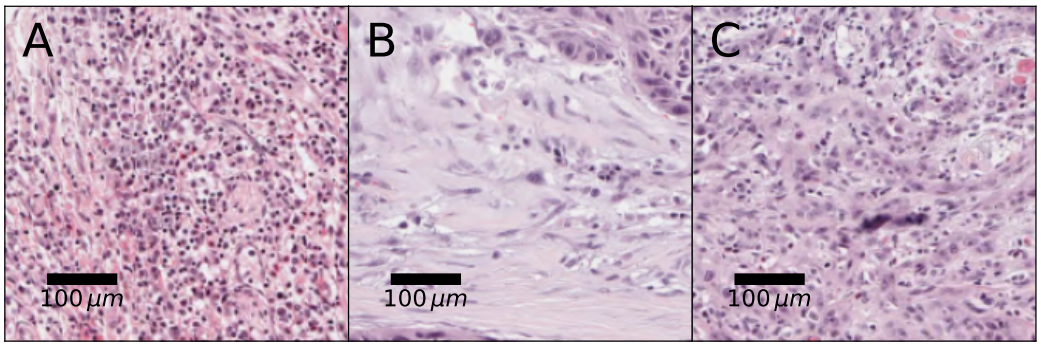


FIGURE 2.4: Examples of H&E stained samples from varied locations within the oral cavity.

Histopathologists typically aim to diagnose potentially malignant tissue according to a standardised classification system [26]. The classification system used to assess oral cancers is based upon the visual interpretation of both architectural features and cytology [26]. Whilst agreement between histologists on the extent of morphological features present within samples is mainly consistent, intra and inter-observer variability continue to hinder this process due to its inherent subjectivity [26, 27, 24]. Any discrepancies in judgement between clinicians could result in misclassification, resulting in over or under treatment for the patient.

The sequence of steps that a tissue sample follows when being prepared for a TMA is outlined below:

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

Biopsy/resection In the process of removing a tissue specimen from a patient's body, various 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 [25].

Paraffin embedding For tissue to be viewed under an optical microscope it must be cut thin enough for light to pass through. Paraffin wax at approximately 58° C is used to permeate the sample [25].

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 [28]; 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 makeup 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[29]. The sample is then stained using the required chemical to produce the desired type of contrast.

The entire slide preparation process can take up to 48 hours [25] 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 rather than formalin-fixation. However, this is a highly skilled process that often results in inferior quality samples and results in a greater

challenge of interpretation. 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 [30].

2.2 Experimental techniques

With the advent of advances in equipment and data analysis, IR chemical imaging has emerged as a solid contender to improve clinical diagnostic capabilities [31, 32, 33, 34].

IR spectroscopy methods come in many forms and modalities with their own respective strengths and weaknesses, but all seek to interrogate the underlying chemical composition of the sample being analysed through the absorption of specific wavelengths of IR light. A brief overview of the physics involved in vibrational spectroscopy shall be given, focusing on the physical mechanisms of absorption and optics. A general overview of the technology underpinning FTIR microscopes shall also be explained.

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 $\vec{B}$, and an electric field $\vec{E}$ oscillating in a synchronised manner whilst propagating through space at velocity **c** in a direction perpendicular to the oscillating fields — as shown in Figure 2.5.

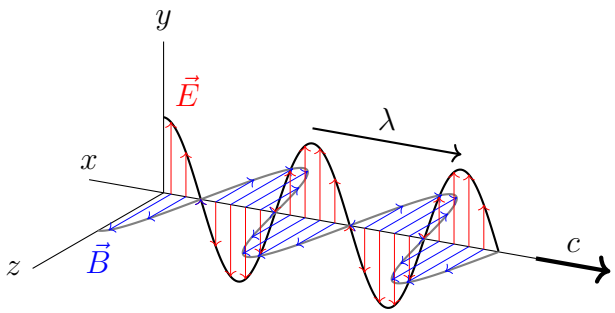


FIGURE 2.5: An electromagnetic wave of wavelength **λ** comprised of a magnetic $\vec{B}$ and an electric field $\vec{E}$ oscillating synchronously whilst propagating in space at velocity **c**

In accordance with Maxwell's equations governing electromagnetic fields Equations (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 Equation (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 [35].

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 crucial 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 Equation (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 to 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 Equation (2.7) and Equation (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. 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 Equation (2.7) becomes:

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

Given Equation (2.12) and Equation (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 an 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 Equation (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 Equation (2.16) to a case with multiple oscillators with their own respective fundamental frequencies and damping factors and combining with Equation (2.7) and Equation (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 Equation (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 Equation (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.

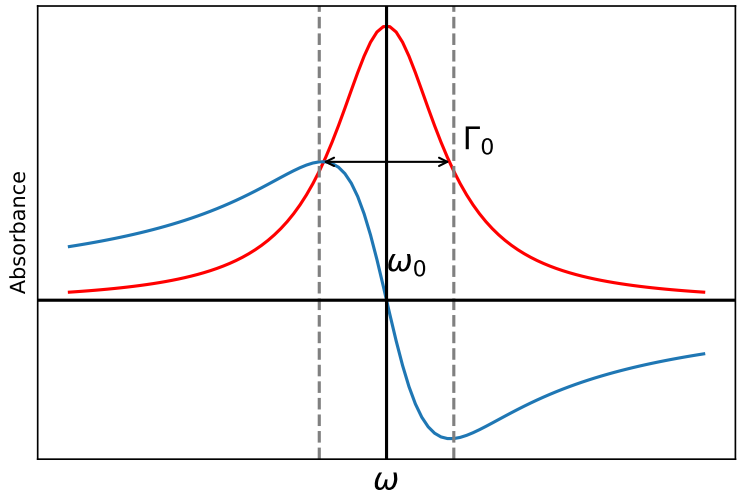


FIGURE 2.6: Complex electric susceptibility of a dielectric as a function of ω. The range of ω values where anomalous dispersion occurs is enclosed in grey dashed lines. This feature repeats at every ω_i.

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, 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 could not 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 fully appreciate the complexity of many phenomena.

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 compared to classical field theories and is therefore much more complex.

Energy quantisation is a requirement to explain many 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. To bring concordance between these

two ideas, a framework that can describe all physical phenomena is needed; this is what quantum mechanics seeks to achieve. 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:

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

This implies that the particle's total probability density is contained within a defined volume. The previous section used the variables ω_i and Γ_i to represent the fundamental frequency and damping coefficient of an oscillating system of charges. To give some insight into the origins of these terms, it is necessary to describe 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 Equation (2.24) take the general form of Equation (2.25). Proof see [36].

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

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:

$$\beta = \frac{2mE}{\hbar^2} \tag{2.27}$$

$$\alpha = \sqrt{\frac{\hbar}{m\omega}} \tag{2.28}$$

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:

$$\alpha^2\beta = 2n + 1 \tag{2.29}$$

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

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

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

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

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

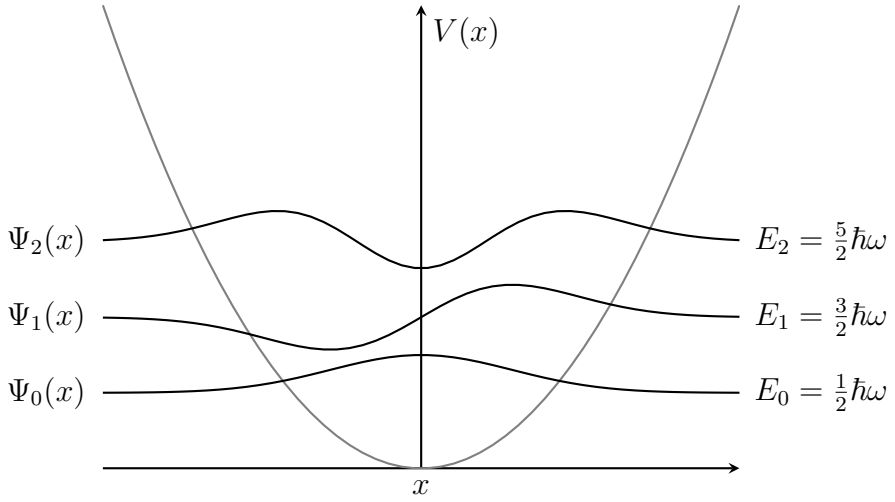


FIGURE 2.7: Potential energy of the quantised harmonic oscillator with first three allowed eigenfunctions and their corresponding energy eigenvalues.

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 that 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. Due to the Pauli exclusion principle, repulsive forces are experienced in regions where $r < r_0$, and attractive forces where $r > r_0$.

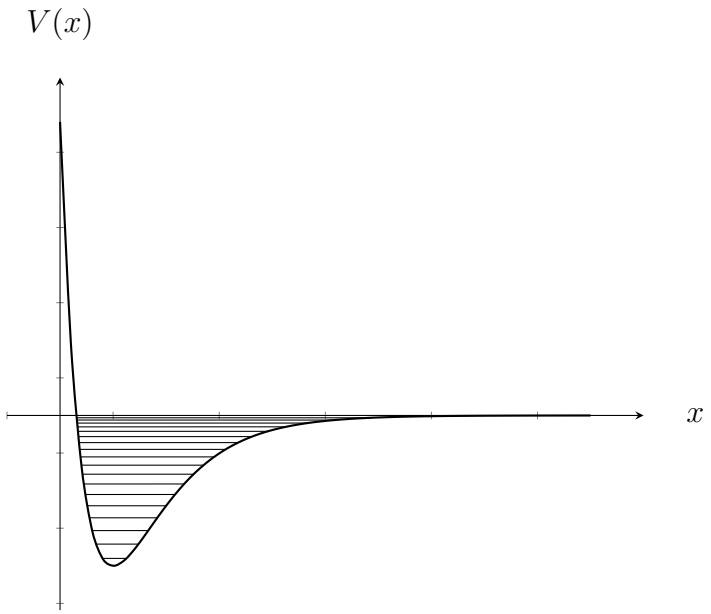


FIGURE 2.8: Potential energy function of the quantised anharmonic oscillator.

In contrast to a harmonic oscillator, the restoring force felt by an object undergoing an 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 [37]. These additional transitions have less energy compared to fundamental changes and give rise to *hot bands*. The intensity of an absorption band is 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 the 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 Figure 2.9.

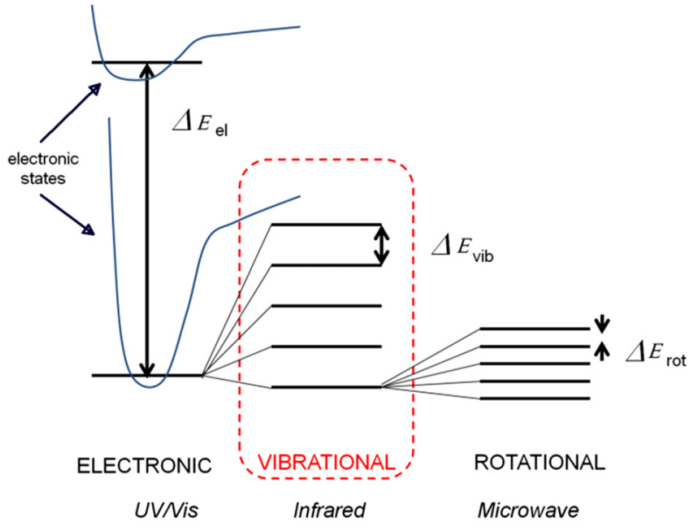


FIGURE 2.9: Energy changes present in molecular spectra. [37]

A molecule with N atoms has 3N degrees of freedom corresponding to translational motion in x,y,z, and rotational motion centred 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 Figure 2.10.

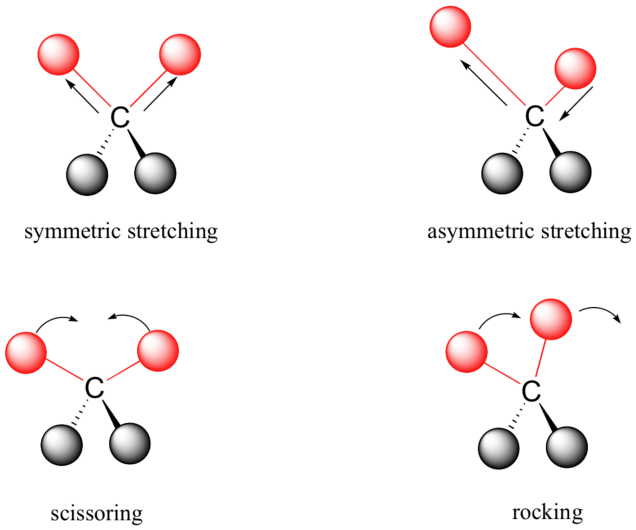


FIGURE 2.10: Energy changes present in molecular spectra. [38]

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 absorption regions in the mid-IR 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 [39]. Spectra are very information-rich; peak positions give information about molecular structures present in the sample, peak intensities yield information about the concentration of such molecules, and peak widths provide information about the sample's chemical state. It is inexpensive, quick, and operators can be trained quickly using modern hardware and software. Some consideration needs to be taken when examining certain samples: water and CO_2 contributions can be a limiting factor when seeking to analyse spectra accurately, but solutions exist to mitigate these effects.

IR Spectroscopy uses the interaction of IR light with matter across several 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 several factors: the wavelength of the IR light, the atoms involved in the molecular bond, and the strength of intermolecular interactions [40]. 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 [41].

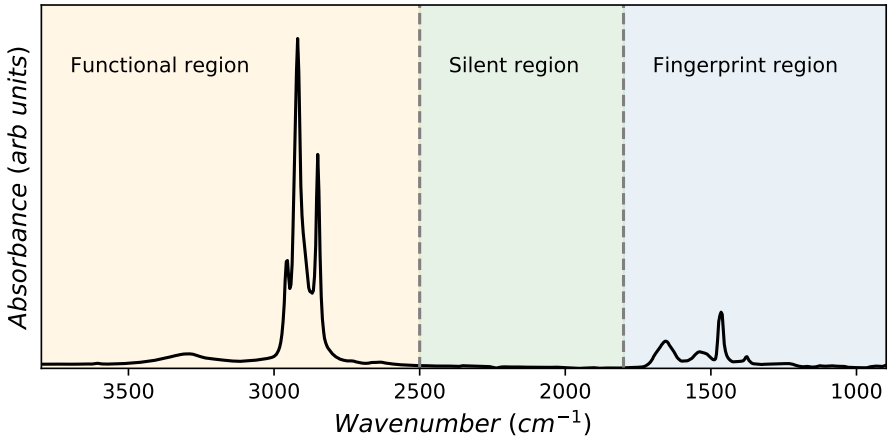


FIGURE 2.11: A typical biological FTIR spectrum example showing three distinct regions of the spectrum: the ”functional region” (3800-2500 cm^{-1}), ”silent region” (2500-1800 cm^{-1}), and ”fingerprint” region (1800-900 cm^{-1})

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 Figure 2.11. A tissue sample or cell is just a collection of molecules in a unique environment that 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 IR 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 sample's absorption 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 [37] 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 a measured μ value must be met with scepticism as a chemical compound will not have a single value associated with it. Scattering effects are indistinguishable from absorption in IR spectra, so the assumption that any "missing" light intensity is purely due to absorbance effects may be false [42].

IR Detectors When measurements are performed in transmission mode using an IR spectrometer, samples are fixed to a substrate that is transparent in the IR, such as CaF_2. 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. The light is then re-collimated by a condenser lens before passing through

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

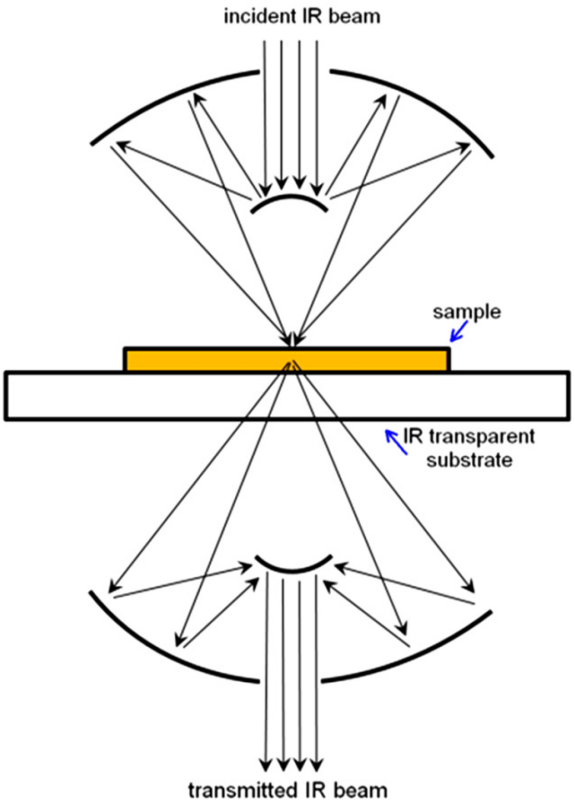


FIGURE 2.12: A Schwarzchild-Cassegrain showing the incident IR beam passing through a series of mirror optics, sample, and a second series of optics before passing through to the detector [37].

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 is then sent as an analogue signal to an accompanying analogue 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 [43]. An MCT

must be cooled to temperatures similar to liquid nitrogen (77K) to minimise the effects of noise, induced by thermally excited current carriers. MCT detectors are far more sensitive than other comparable detectors [43] 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 To observe a wide range of samples, a suitable IR light source is required; three such light sources are the quantum cascade laser (QCL), free-electron laser (FEL), and a globar. Each light source has its 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 [44]. A typical globar comprises a silicon carbide rod heated to around 1000 - 1650 °C, and a variable interference filter. As the globar emits a continuous IR spectrum across a wide range of specific wavelengths, it can be filtered into specific bands of wavelengths or used in conjunction with an interferometer.

A QCL is a semiconductor laser that 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 [37]. Shown in Figure 2.13 is a simplified schematic of a QCL.

Sources can output either a *continuous* IR spectra in the case of a FEL and globar; 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 significantly increased by averaging over a small spectral range instead of scanning over the entire spectrum as in an FTIR [37].

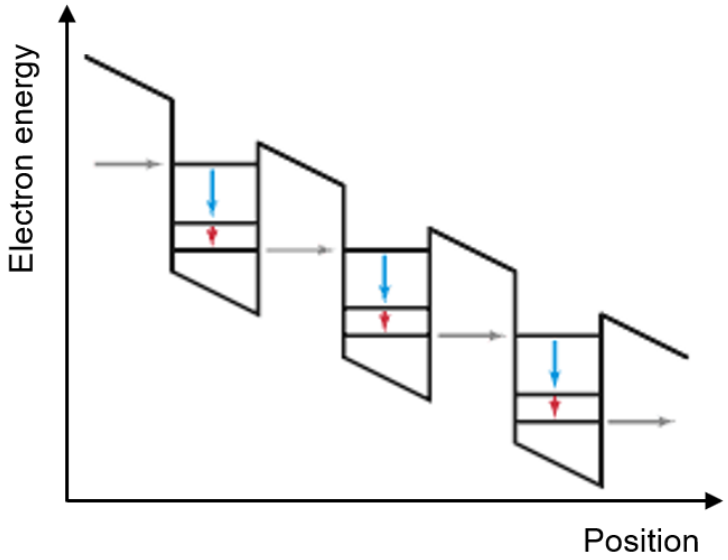


FIGURE 2.13: A simplified schematic of the gain region of a QCL, showing electron energy against the position. The electron is injected at the left-most grey arrow and undergoes a radiative transition (blue arrow); the electron then undergoes a further non-radiative transition (red arrow) before tunnelling to the next quantum well and repeating this process [45].

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 several tissue types such as: brain [46, 47, 48], colon [49], skin [50], liver [51] and many others and is considered to be one of the most popular IR techniques available today [52]. 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 that 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 an instrument that produces an interference pattern by superimposing two beams of light. The

light is incident from the IR source onto a beamsplitter that transmits a portion of the IR light and reflects the other portion onto a fixed mirror. The part of the beam transmitted is incident onto a movable mirror that 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 Figure 2.14.

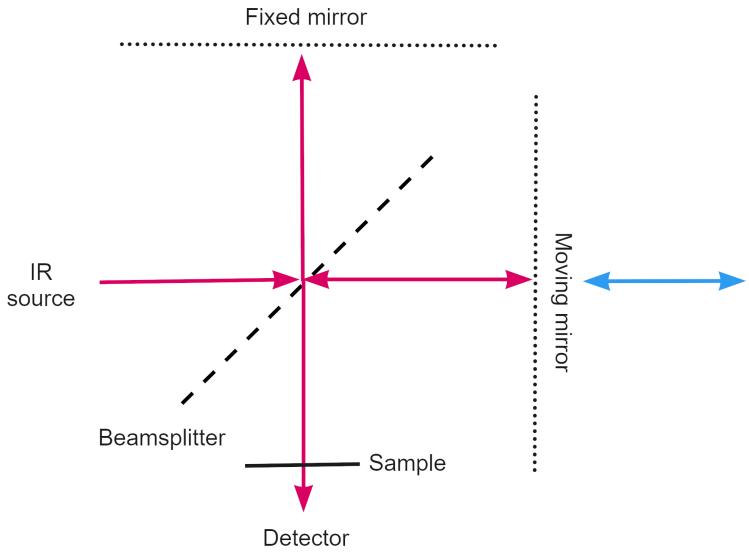


FIGURE 2.14: A Michelson interferometer used in a FTIR spectrometer. [37].

The recombined light interferes constructively if the distance between the beamsplitter and movable mirror is equal. 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 Equation (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 that can be accounted for by including 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 [42]:

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

The resulting interferogram combines 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 Figure 2.15. To obtain a spectrum 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.

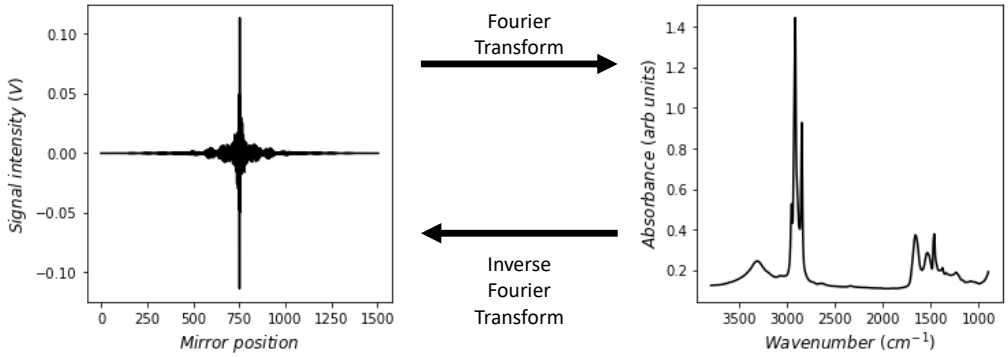


FIGURE 2.15: The conversion of an interferogram to a wavelength dependent transmittance spectrum

Measurement of spatial variation 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 micrometre resolution [30]. Imaging is achieved by directing the light emitted from the sample region onto a focal plane array (FPA) using focusing optics to define the pixel size [37]. 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.

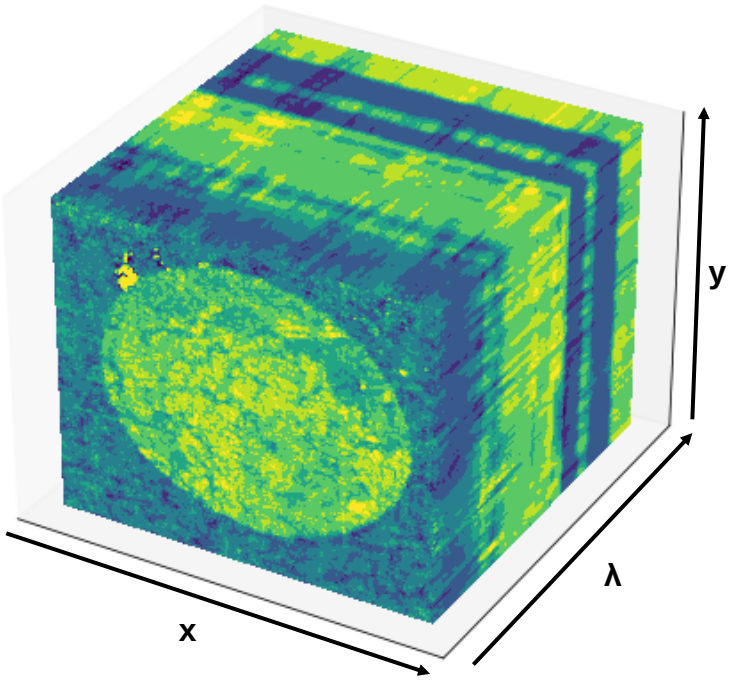


FIGURE 2.16: An FTIR datacube example showing spatial variation in x and y with spectral absorbance varying in λ.

Several factors govern the achievable spatial resolution of an optical technique such as FTIR. 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, a lens with a high NA will be able to resolve objects at smaller scales. However, the resolution of an optical instrument is always subject to the diffraction limit. Due to

dispersive effects associated with glass lenses, mirrors are the typical choice for the optical path of a FTIR microscope.

Diffraction and Resolution 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 source used in acquisition [53]. The diffraction limit is the minimum size of the spot to which a beam of light can be focused using standard lensing elements. The focused spot forms a symmetric pattern of concentric rings called an Airy disk pattern, as shown in Figure 2.17.

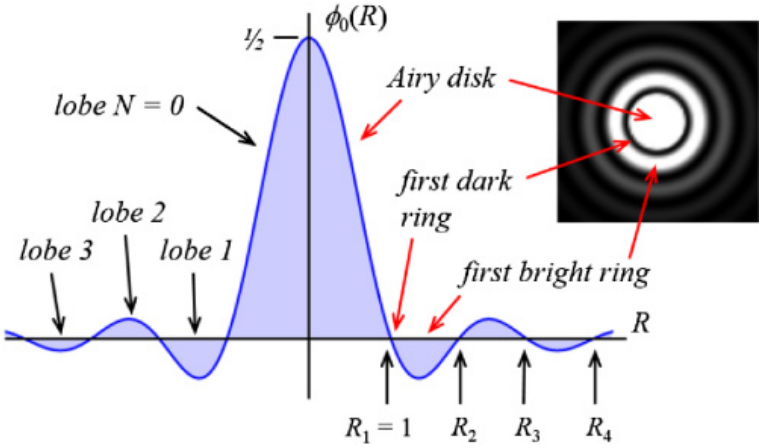


FIGURE 2.17: An Airy disk diffraction pattern showing periodic bright and dark fringes [54].

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 at least a distance away given by Equation (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 over 1. Consequently, this means that with good quality optics, microscopy systems can achieve resolutions of up to $\lambda/2$ [53]. This can, in practice be overcome with good estimates of the point spread function (PSF) of the detector 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 [53].

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 [37, 42]. Combined with modern FPA detectors, it is possible to rapidly and cost-effectively extract large datasets for later analyses.

2.3 Data Analysis

To utilise and gain insight from experimental data collected from FTIR microscopes, the data must be summarised and interpreted. A considerable variety of methods exist which approach the problem from different angles, some are well-established techniques that originate from multivariate statistics, whereas other techniques are categorised as *machine learning* (ML). Effective data analysis is an essential step when developing a clinical diagnostic tool. A high false-positive rate will result in unnecessary procedures, and a high false-negative rate can result in unnecessary deaths.

With the aim of the project being to gain an understanding of the biological systems present in cancer and to develop prognostic tools, it is necessary to convert the *raw data* containing chemical information about each sample into

meaningful insights by quantifying relationships or categorising clusters of data points. The following section starts with a description of a few procedures that are common preprocessing techniques. Then an overview of several statistical methods and machine learning algorithms will be given with a discussion of the process of validating classifier results.

2.3.1 Machine Learning & Statistics

ML is an approach to data analysis that involves learning from example data rather than relying on heuristics. This approach has led to advances in fields such as finance [55], healthcare [56], bioinformatics [57]. ML objectives are generally either *classification* or *regression* problems. The goal of classification is to obtain the function f which maps an input vector $\boldsymbol{X}$ to a discrete output t. The input vector $\boldsymbol{X}$ is the list of variables that are used to describe a data point, e.g. colour, weight, length, and t being the label applied to the data point, e.g. orange, apple, banana. A regression problem seeks to turn $\boldsymbol{X}$ into a numerical output, for example, the number of bathrooms or floor space of a house into the market value of the house. Some overlap exists 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 that can be used and what insights can be gained.

2.3.2 Preprocessing

Analysing 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 several stages; this process is illustrated in Figure 4.1. Preprocessing is a vital step in the analysis workflow, as it has been shown to increase the performance of classification models [58], as well as to improve the validity and interpretability of results.

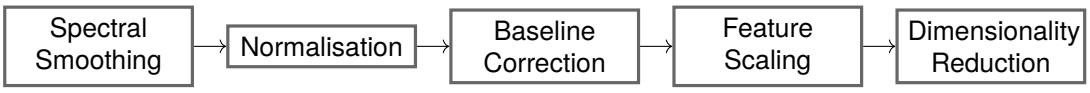


FIGURE 2.18: A typical preprocessing pipeline diagram

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 Equation (2.34) on the thickness of the sample, the absolute absorbance value will vary strongly. This is considered a confounding factor and is typically dealt with by several methods, such as vector normalisation, min-max scaling, or 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 [59], whereby a polynomial is fit to a local moving window of a specified length. Other methods such as 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 spherical morphological structures present in the sample are of comparable size to the incident radiation. This effect is particularly strong in cells, but tissues are also adversely affected to some extent. The impact in embedded tissue samples is mitigated somewhat by the presence of paraffin wax which results in a more homogenous refractive index throughout the sample [60, 61, 62]. The data present in this thesis has been subject to an Extended Multiplicative Scattering Correction (EMSC) algorithm outlined in [63].

Feature Scaling

Feature scaling takes place to remove absolute variable values effectively. This does not detrimentally affect the data as it is only relative values between subgroups of data relevant to discriminatory tasks. This step often helps subsequent classifier steps and is imperative for PCA.

2.3.3 Dimensionality Reduction

When a dataset is highly-dimensional, 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 reducing the amount of information given to the classifier.

Principal Component Analysis (PCA)

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

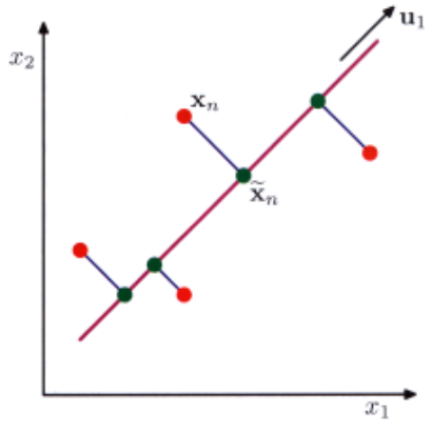


FIGURE 2.19: The principal component shown here by u_1 with the orthogonal projections (shown in green) of the original space data points (in red) projected onto it [64]

This process is performed by obtaining the eigenvectors of the covariance matrix of the data; these eigenvectors then become the principal components [65].

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) differs from PCA in that it uses information about the identity of each data point to linearly transform the basis of the dataset. The main goal of LDA is to map to a space that gives good inter-class separability and avoids overfitting to the data.

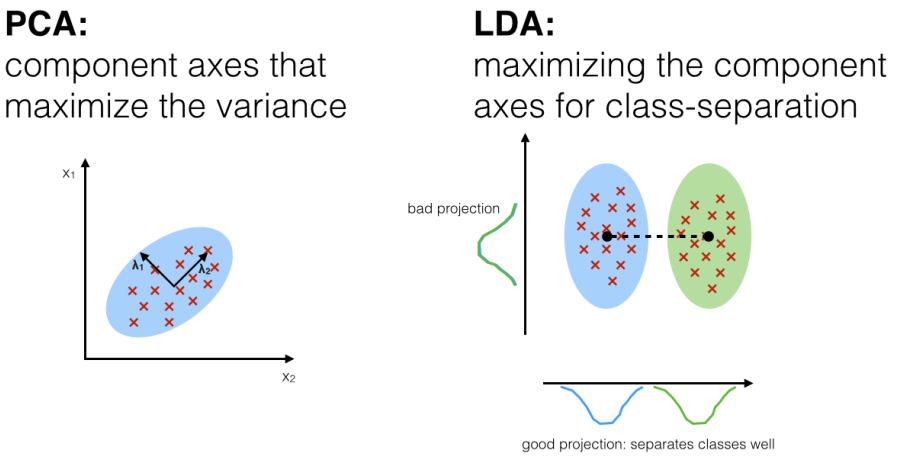


FIGURE 2.20: A comparison of PCA and LDA [66]

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

2.3.4 Machine learning algorithms

This section will briefly overview some machine algorithms used in 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 Figure 2.21.

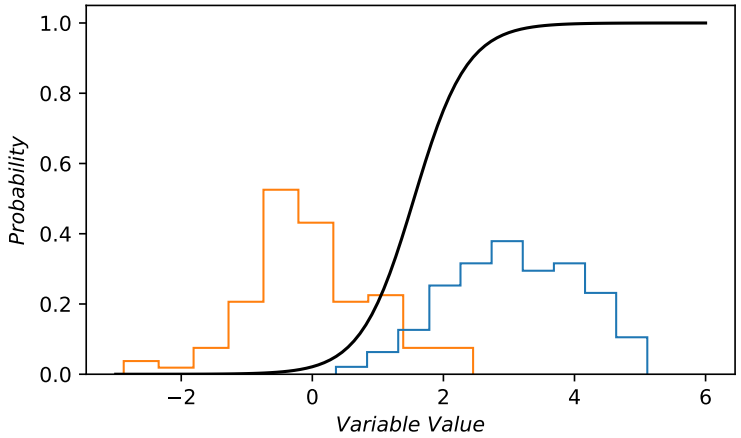


FIGURE 2.21: A univariate logistic regression example showing two generated distributions in orange (Y=0) and blue (Y=1); the fitted logistic function is shown in black with maximum probability predicted in the region spanning the blue histogram (Y=1).

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 [67] for a thorough derivation.

Support Vector Machines

Support vector machines (SVM) have become a commonly used method for classification and regression capable of performing well on complex datasets [68, 64, 69]. SVMs can be separated into two distinct types: linear and non-linear. As its name suggests, a linear SVM forms a linear decision boundary across the input parameter space separating classes.

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

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 that seeks to obtain the optimal separation boundary 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 hyperparameter 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 [69]. 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 that allow classes to become linearly separable. A comparison of a linear and nonlinear SVM is shown in Section 2.3.4.

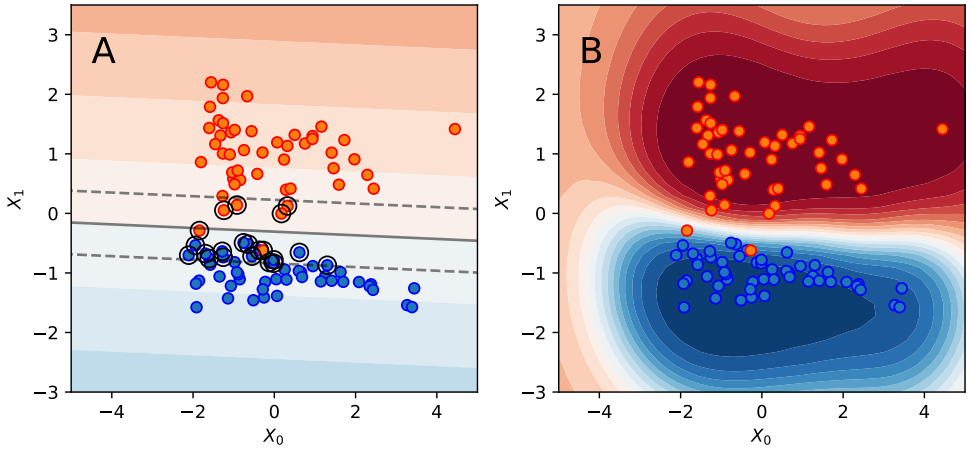


FIGURE 2.22: SVM classification boundaries showing a linear SVM (A), and nonlinear SVM using a RBF kernel function (B). (A) shows the support vector boundary in a solid grey line with support vector points highlighted with a black circle.

Artificial Neural Networks

An artificial neural network (ANN) is a technique loosely based on the functioning of a biological neuron. First introduced in 1943 [70], 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 can receive an input signal, perform some processing, and output the resultant value. This process is summarised in Figure 2.23.

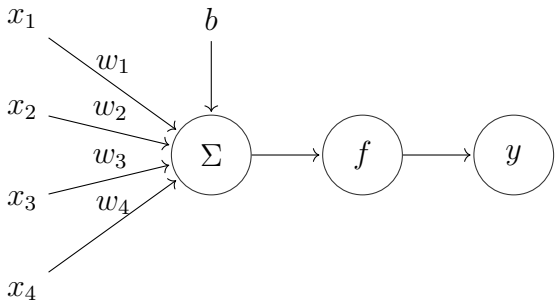


FIGURE 2.23: The perceptron showing input variables x_i multiplied by their respective weights w_i, before being summed over all inputs and added to a bias term b, and then subjected to a nonlinear activation function f

The perceptron equation for several input variables n is given by Equation (2.43)

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

For a perceptron to succeed in its desired application, it must be 'trained'. Training in this sense refers to optimising the weights of a perceptron with respect to the desired metric. Often, metrics such as accuracy, sensitivity, or specificity are used for classification tasks. If the perceptron is used for regression, this metric would indicate the loss/fitness of a proposed function, for example, the means squared error.

The Multilayer Perceptron

A single perceptron is very similar to a single unit of logistic regression and can achieve simple binary classification tasks [68]. 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 Figure 2.24.

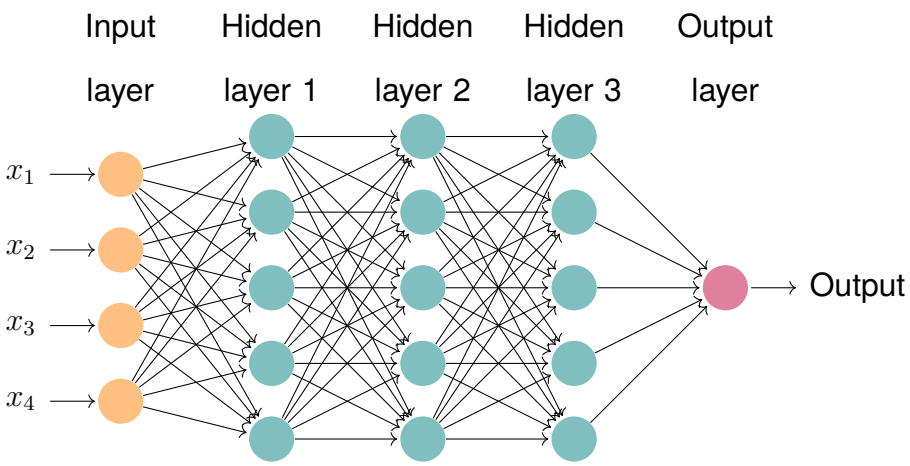


FIGURE 2.24: A multilayer perceptron neural network with an input layer consisting of four input variables $x_0 \dots x_4$, three hidden layers of five nodes each, and a single output layer.

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* [71] is the standard method of determining effective weights for each node; it is made possible because 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 actual 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 choices of activation functions and the number of layers and nodes in a neural network. Additionally, there are many parameters associated with the backpropagation algorithm itself which can be altered. A thorough

description will not be given here but is covered in depth in other sources [68, 64, 67]. An additional process known as hyperparameter optimisation can be performed and will be covered in depth in Chapter 4. Neural networks form the focus of Chapter 5 where an explanation of convolutional neural networks (CNN) shall be given.

Classification and Regression Trees (CART)

A decision tree is an ML algorithm that 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 utilising a set of criteria. Usually, the split is calculated as that which will maximise the entropy gained according to the Gini index [72]. 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.

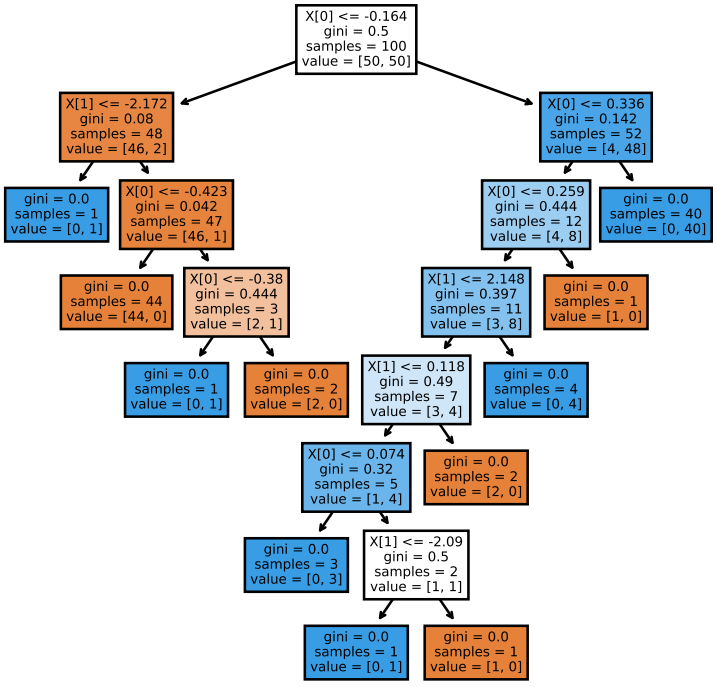


FIGURE 2.25: A typical CART comprising branch nodes shown here by a logical decision operator, and leaf nodes consisting of an output value.

CARTs can be effective classifiers in simple data sets. Still, 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

To mitigate the high variance of CARTs, a technique known as pruning can be employed to reduce the complexity of the tree, and the likelihood of overfitting [73]. Leaf nodes are pruned based on 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 standard method of restricting 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 that regulates the model's complexity.

Bagging and Random Forest

Bagging is a general ML technique that combines many weaker classifiers into an overall more capable 'ensemble' 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 significantly [72]. 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 decides on different features.

Boosting

Boosting, in contrast to bagging, is a technique designed to reduce the bias of a classifier [74]. Bias is a measure of how well a classifier captures the necessary information needed to do its job [64], 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 Equation (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. To optimise a decision tree, boosting is performed. Boosting is achieved by first classifying the set of test points; the misclassified points are then weighted higher so that more focus is placed on classifying them in the next iteration. This process is repeated until the classifier can correctly identify the data set to the required standard. Each

iteration is then weighted according to a learning rate λ and then used to give an overall decision.

Extreme Gradient Boosting (XGBoost)

XGBoost is a highly optimised supervised learning method that builds upon gradient boosted decision trees [75]. It is a highly successful algorithm, and one of the most commonly used ensemble methods used by many data science competition-winning teams [76]. 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 that 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 a crucial 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. However, the performance achieved from this method is not descriptive of the entire data set and wastes 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.

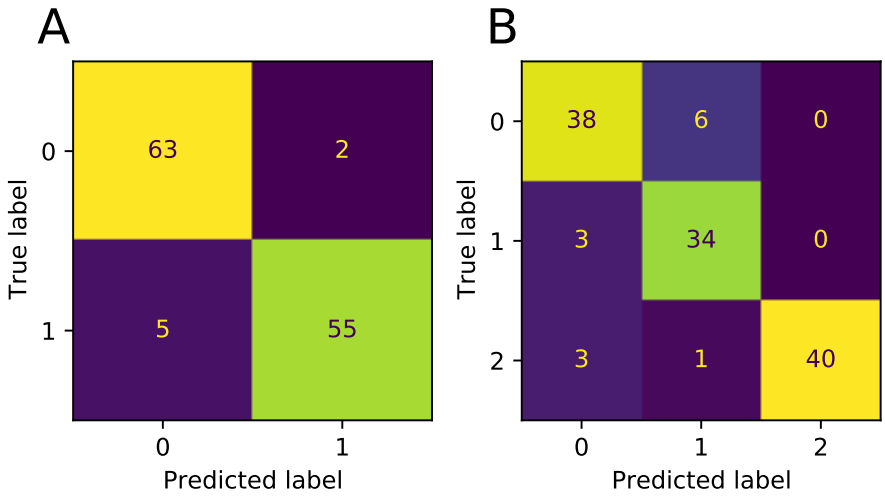


FIGURE 2.26: Binary (A) and multiclass (B) confusion matrices showing classification results.

When evaluating the performance of a classifier, several 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 [57, 77]. 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. Thus, there is a trade-off between the performance of either score – an increase in one score typically involves a decrease in the other. Similar 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 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 dichotomised by taking a threshold over the value.

Receiver Operating Characteristic (ROC) analysis

It is possible to estimate 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 the true positive rate (TPR) (sensitivity) against the false positive rate (FPR) (1 - specificity) for several decision thresholds for a binary classifier; this can be extended to a multi-class problem by utilising a ”one against all” approach.

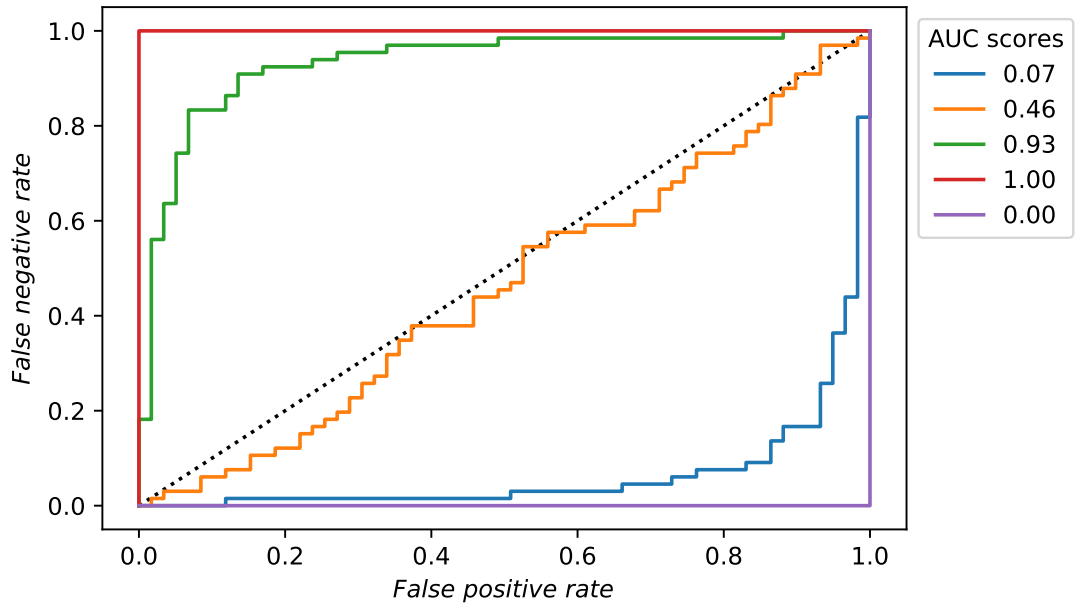


FIGURE 2.27: ROC curves showing a comparison between a number of classifiers with calculated AUROC scores.

A value across the diagonal signifies a classifier that is equivalent to pure chance and effectively useless. A steep curve that approaches the top-left corner is a realistic, well-performing classifier that will have a high AUROC. An AUROC value of 1 signifies a flawless classifier that correctly identifies every instance; more importantly, it does not need to trade-off between specificity and sensitivity; a value of 0.5 is equivalent to a random guess signifying a completely useless classifier. A ROC curve is a collection of TPR and FPR values for

several cut-off points. The curve represents the trade-off between obtaining more true positives and fewer false positives. It allows for a visual evaluation of the classifier's performance over the entire range of thresholds.

Precision Recall analysis

The precision-recall (PR) curve is similar to the ROC 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 several thresholds. When examining a highly imbalanced dataset, a PR curve can avoid overly optimistic estimates of classifier performance [78]. 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.

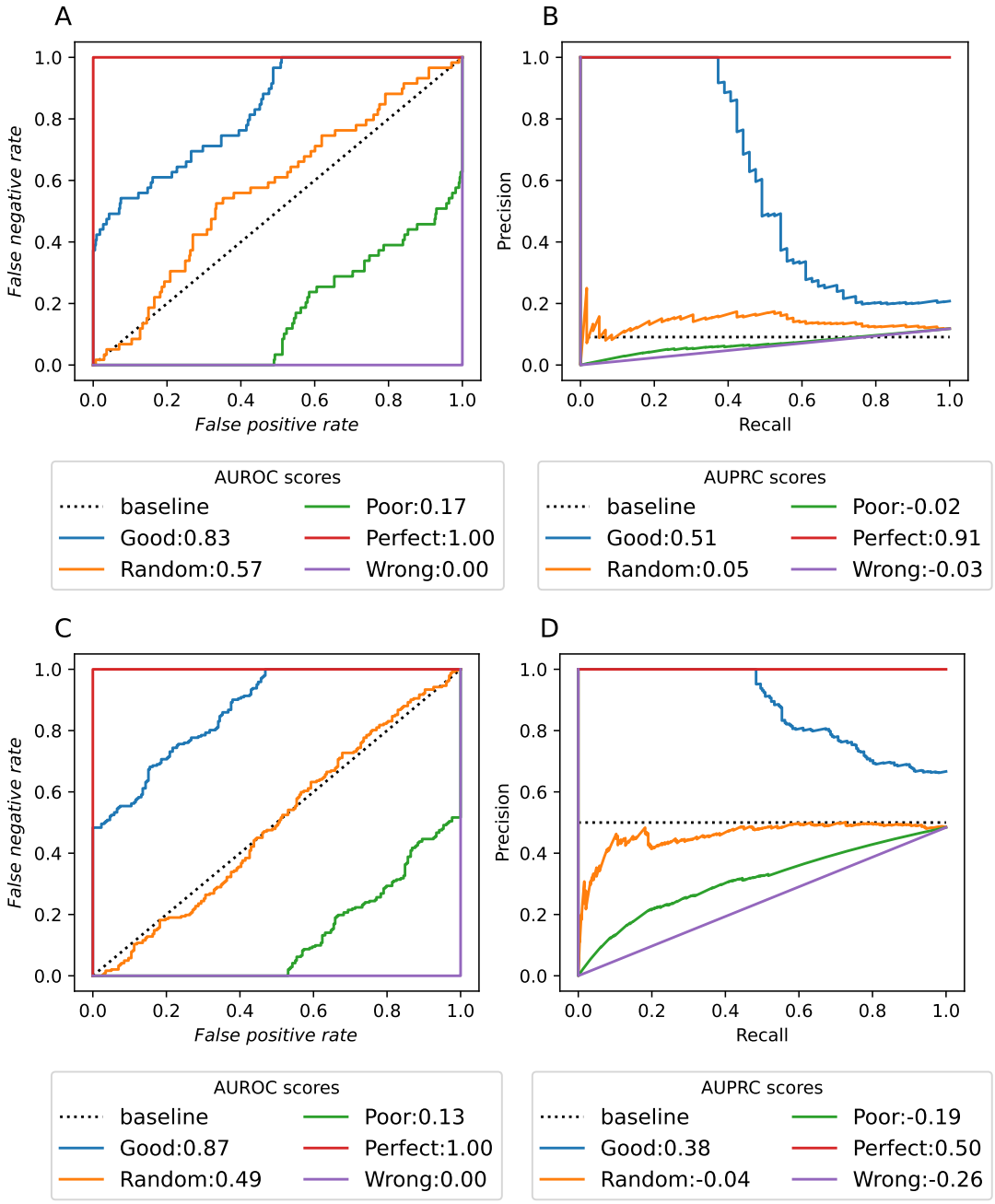


FIGURE 2.28: Classifiers of varying utility evaluated on simulated imbalanced [A,B] and balanced [C,D] datasets.

Figure 2.28[A-B] shows an evaluation of several 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 Figure 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 then subtracted 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, 1980-2015 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] R. J.C. Steele, A. M. Thompson, P. A. Hall, and D. P. Lane. The p53 tumour suppressor gene. *British Journal of Surgery*, 85(11):1460–1467, 1998.

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

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

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] Tseng-Cheng Chen, Chen-Tu Wu, Cheng-Ping Wang, Wan-Lun Hsu, Tsung-Lin Yang, Pei-Jen Lou, Jenq-Yuh Ko, and Yih-Leong Chang. Associations among pretreatment tumor necrosis and the expression of hif-1α and pd-l1 in advanced oral squamous cell carcinoma and the prognostic impact thereof. *Oral Oncology*, 51(11):1004–1010, 2015.

[21] Anthony C Nichols, Pencilla Lang, Eitan Prisman, Eric Berthelet, Eric Tran, Sarah Hamilton, Jonn Wu, Kevin Fung, John R de Almeida, Andrew Bayley, et al. Treatment de-escalation for hpv-associated oropharyngeal squamous cell carcinoma with radiotherapy vs. trans-oral surgery (orator2): study protocol for a randomized phase ii trial. *BMC cancer*, 20(1):1–13, 2020.

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

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

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

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

[26] 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.

[27] 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.

[28] 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.

[29] Safaa Al Jedani, Caroline I. Smith, Philip Gunning, Barnaby G. Ellis, Peter Gardner, Steve D. Barrett, Asterios Triantafyllou, Janet M. Risk, and Peter Weightman. A de-waxing methodology for scanning probe microscopy. *Anal. Methods*, 12:3397–3403, 2020.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

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

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

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

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

[40] 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.

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

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

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

[44] 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.

[45] RP Photonics Consulting. Quantum cascade lasers.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] 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.

[55] 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.

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

[57] 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.

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

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

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

[61] 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.

[62] 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.

[63] 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.

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

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

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

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

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

[69] 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.

[70] 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.

[71] 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.

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

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

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

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

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

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

[78] 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.

3 Prognosis

3.1 Introduction

Head and neck squamous cell carcinomas (HNSCCs) are the sixth most common form of cancer worldwide, amounting to over 500,000 new cases annually. The majority of head and neck cancers are squamous cell carcinomas that originate in the upper aerodigestive epithelium. The development of oropharyngeal squamous cell carcinoma (OPSCC) is known to be linked to carcinogenic HPV, accounting for approximately 51.8% - 71% of cases [1].

OSCC is the 8th most common form of cancer in the UK [2] with a recent increase in incidence reported [3, 4]. In contrast, OSCCs are rarely mediated by HPV [5], and the majority of cases are typically associated with exposure to carcinogens present in tobacco and alcohol [1, 3, 6].

The regions in which head and neck tumours typically develop are anatomically complex and play a vital physiological role in the patient; early diagnosis and selection of appropriate treatment should increase patient survival while maximising the retention of vital organ function.

A key issue facing clinical decision-makers is determining the optimal course of treatment for a patient. In cases where lower biological aggression can be demonstrated, a de-escalation of therapy may be possible [7, 8]. Conversely, early identification of patients with poor prognosis could allow targetting for neo-adjuvant therapy. However, identification of these cases prior to surgery is

not insignificant, even given the acknowledged prognostic importance of extra-nodal extension (ENE) [9]. Physical methods such as magnetic resonance imaging (MRI) and computed tomography (CT) have proven to be of limited benefit [10] despite improvements in analysis techniques utilising DL models [11].

Previous studies [12, 13, 14, 15] have hypothesised that tumours with ENE, which may be responsive to novel therapeutic treatment may carry a distinct molecular fingerprint; the identification of which may allow for screening of patients towards appropriate treatment. Despite small studies identifying putative biomarkers of ENE that might be of use in the analysis of diagnostic biopsy material [16], none of the proposed biomarkers or molecular fingerprints of aggression has progressed into practice. For approximately 50% of HPV negative HNSCC patients, current treatment plans are ineffective [17, 18]. Neo-adjuvant therapy can improve prognoses and aid clinical decision making if applicable cases can be determined at the time of diagnosis.

FTIR microscopy is a well-established technique that has been utilised in a range of biomedical applications in recent years. Due to its ability to access chemical information present within the sample; FTIR microscopy data and accompanied multivariate analysis have been used to diagnose cancer in biofluids [19, 20, 21], surgically resected tissue [22, 23], and cells [24, 25, 26]. FTIR microscopy allows for imaging of sample specimens at thousands of infrared wavelengths simultaneously using a typical spectrometer. This is achieved by using a broadband light source and Michelson interferometer set-up. Individual chemical spectra are then obtained by performing a Fourier transform on the resulting interferogram. Marginal spectral differences in biochemical compounds of interest are typically located in a region known as the fingerprint region (1000cm^{-1}-1800cm^{-1}). It is differences in these absorption bands which contain information that can be utilised to discriminate between samples

of interest.

The use of FTIR in clinical diagnostics is growing quickly, however relatively little work has been aimed towards prognostic biomarkers. The combination of both FTIR microscopy with techniques more familiar to oncologists such as immunohistochemical staining has the potential to improve prognostic predictive capabilities significantly as shown in this work. Previous studies [27, 28, 29, 30, 31, 32, 33] have investigated the viability of a range of prognostic biomarkers for head and neck cancer, with varying degrees of success. Many previously analysed biomarkers are measured on surgically resected tissue — limiting the potential for timely treatment. What is needed are prognostic biomarkers which can be measured in biopsy tissue prior to surgery. The discovery of effective prognostic biomarkers has been difficult and thus far has largely focused on immunohistochemistry techniques. MRI has also been utilised [34, 35, 36] to measure physical attributes such as: tumour thickness, depth of invasion, and the presence of sub-volumes in a non-invasive manner. However, MRI based techniques often quantify biomarkers inaccurately when validated against direct measurements of pathological staging sections [37, 38].

Zawlik et.al [39] investigated FTIR coupled with PCA to investigate the efficacy of chemotherapy in triple-negative breast cancer. They determined that it was possible to monitor changes in the biochemical composition of the tissue in order to monitor the effectiveness of received treatment. Butler et.al [40] have undertaken development of a high-throughput ATR-FTIR based instrument for use in biofluid assays. Their work concluded that it was possible to triage brain cancer utilising FTIR spectroscopy of biofluid samples. Their analysis comprised a large retrospective cohort of 724 patients with a range of brain cancer subtypes and stages. They utilised a binary SVM classifier, and were able to achieve a sensitivity and specificity of 93.2% and 92.8% respectively.

Many other prognostic biomarkers exist [41, 42, 43] but a large proportion

are still in the ”discovery phase” — requiring further study to ascertain prognostic benefit [44]. The use of FTIR within clinical diagnostics likely fits into this category, as many potential barriers facing other potential biomarkers are still present.

This work explores the potential efficacy of FTIR microscopy in combination with a known prognostic biomarker: α-smooth muscle actin (ASMA) expression, as a method of identifying patients with poor prognoses prior to surgery. Previous work [32, 31, 30] has explored the efficacy of ASMA and SERPINE1 [32] as predictive variables for extracapsular spread (ECS) and as prognostic biomarkers for OSCC. ASMA expression is closely associated with the presence of activated fibroblasts, also known as myofibroblasts, in tumour associated stroma. The degree of ASMA expression can be interrogated using appropriate chemical stains and evaluated using an optical microscope.

3.2 Materials and Methods

Tissue preparation

The dataset comprised FTIR spectra taken from diagnostic primary tissue of 29 patients diagnosed with OSCC. The specimens are a subset of those arranged in a previously described tissue microarray (TMA; [32]). Inclusion criteria for this study were: a diagnosis of OSCC; the presence of OSCC in the TMA core; the ability to co-register adjacent H&E stained and FTIR imaged sections; a follow-up period after surgery of at least 24 months; HPV negative. Patients gave written, informed consent and the study was undertaken under ethical approval (Northwest - Liverpool Central REC number EC47.01). All samples were 1mm diameter cores of FFPE tissue arranged into a TMA.

Four adjacent sections of 4μm thickness were taken from the TMA; the first and last sections were stained with H&E and used to assess the presence and location of tumour material in the second and third sections. Specimens were removed from the sample set if no clear area containing predominantly tumour cells was discernable in the H&E stained sections. Samples were also removed from the sample set if the outline of the regions containing tumour cells were markedly different between the first and fourth sections. Images of stained sections were scanned using an Aperio CS2 scanner (Leica Biosystems) and used for IR image annotation. The second and third TMA sections were mounted onto CaF_2 disks for FTIR microspectroscopy.

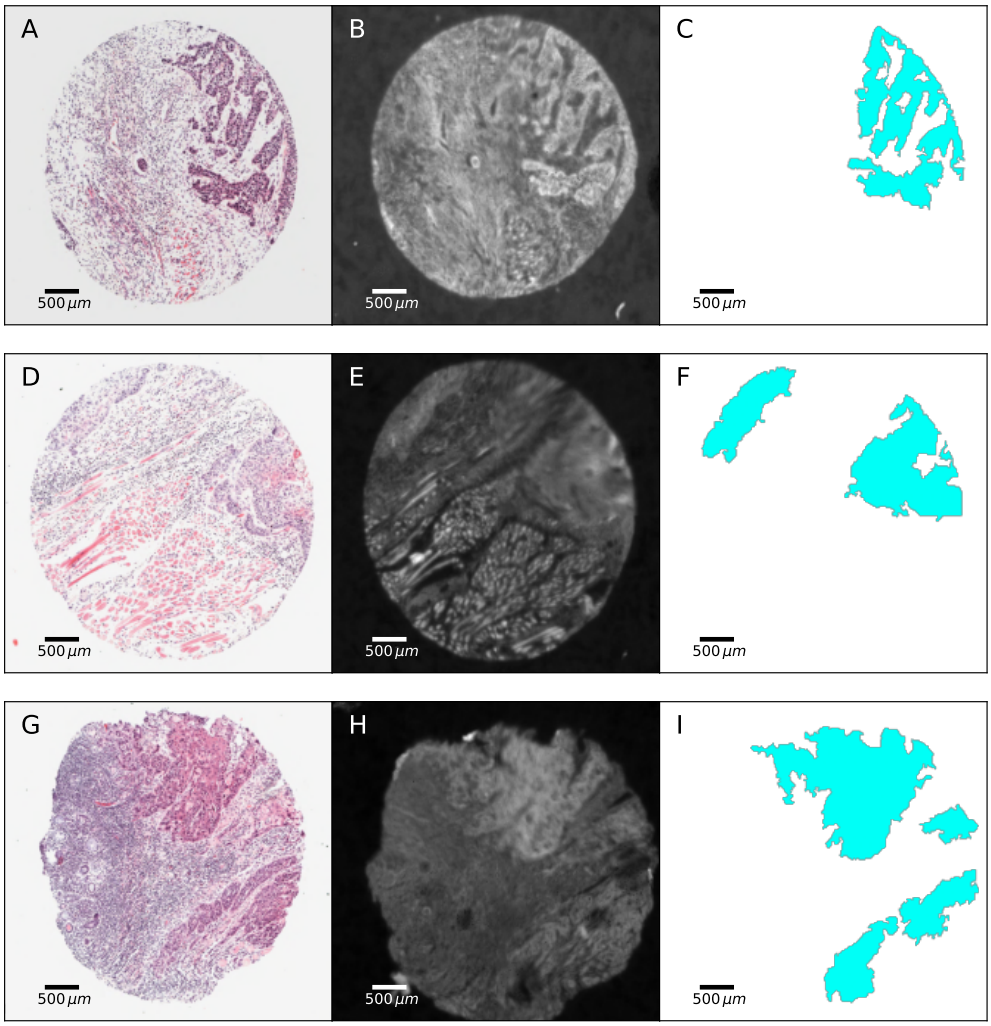


FIGURE 3.1: Annotation of OSCC-containing areas in FTIR images. [A,D,G]: H&E image of a tissue core; [B,E,H]: FTIR image at 1650cm^{-1}; [C,F,I]: Areas from which FTIR data was extracted for analysis

(Figure 3.1) shows three examples of tumour areas selected from H&E sections and used for further analysis. Selections were annotated using GIMP [45] image manipulation software onto IR images at 1650cm^{-1} to ensure spectra were extracted accurately. The obtained mask was then used to compile spectra of tumour regions indexed by patient identity and corresponding metadata.

FTIR Microspectroscopy

FTIR measurements of TMA cores were taken at room temperature using a Varian Cary 670-FTIR spectrometer with an attached Varian Cary 620-FTIR microscope produced by Varian (now Agilent Technologies, Santa Clara CA, USA); with a liquid nitrogen-cooled 128$\times$128 pixel MCT focal plane array with an effective field of view for each pixel of 5.5μm. The sample stage was enclosed in a perspex box and pumped with dry air until a humidity of 1% was achieved to mitigate the effects of water contributions on measured IR spectra. Images were acquired at a resolution of 6cm^{-1} over a spectral range of 990cm^{-1} to 3800cm^{-1} using a co-addition of 128 scans. The attenuator and integration time of the FPA were chosen to maximise the signal-to-noise ratio. Background scans were acquired using a blank CaF_2 disk situated within the perspex box before each session of measurements; data was extracted from raw output using MATLAB methods from ChiToolBox [46].

Data Preprocessing & Analysis

The selection of tissue areas to include in the analysis was undertaken by a consultant oral pathologist (AT), who identified regions containing high proportions of tumour cells on the H&E images. These were subsequently co-registered with IR images at 1650cm^{-1} (the amide-I peak) from the same tissue core to extract IR data for analysis.

To correct for atmospheric scattering, extracted spectra were pre-processed using an open-source extended multiplicative scattering correction (EMSC) code [47]. An unsupervised quality control check of all data to eliminate anomalous spectra through the use of the multivariate Hotelling's T^2 statistic [48, 49]; spectra determined to have a T^2 value lying outside the 95th percentile were deemed to be anomalous and were omitted from further analysis.

The following preprocessing steps were carried out on the dataset before a final classification step was performed using a LR classifier. The absorbance contribution attributable to paraffin situated in the range (1340cm^{-1} - 1460cm^{-1}) was removed and all spectra were limited to the fingerprint region (1000cm^{-1}-1800cm^{-1}). Vector normalisation was used to account for sample thickness; wavenumber absorbance features were mean-centred, and variance scaled to one; before a final PCA step to reduce the dimensionality of the dataset. Seven principal components were taken to assist convergence when fitting the LR classifier. A large L2 regularisation term (1×10^5) was applied to the objective function when fitting the LR model to mitigate the potential for overfitting.

Scientific Python packages [50, 51, 52] were used to implement classification models and survival analysis. The classification power of an FTIR spectrum as a prognostic biomarker was estimated using the AUROC and AUPRC metrics. To obtain an estimate of the variability of the classification power of FTIR, bootstrap out-of-bag sampling was utilised as follows. A training data set was constructed by drawing 80% of patients in the total dataset without replacement. The remaining 20% was used as the ”out of bag” test set on which fitted models were evaluated and statistics calculated. This process was repeated 100 times, ensuring that no two sample sets were identical. When fitting the LR model, data points were inversely weighted to compensate for the differing number of acquired spectra per patient and by risk group to mitigate the imbalanced nature of the dataset. Predictions of the risk group from the LR model are presented as a list of probabilities for each risk group; the final prediction scores for each patient are the median probability predicted for that patient.

Calculated statistics included: AUROC, Matthew's correlation coefficient (MCC), specificity, sensitivity, PPV, and NPV — all statistics were calculated using appropriate weightings to compensate for class/patient imbalance and to ensure a classification statistics were not skewed. Prognostic efficacy was

investigated using Kaplan-Meier survival analysis, a Cox proportional hazards regression, and a log-rank test.

Prediction of patient outcomes

Rather than utilising arbitrary outcome cutoffs to identify the degree of risk, the cohort was stratified into ”high” and ”low” risk categories. The choice of the risk group for each patient was determined through an optimisation routine which maximised the log-rank statistic with respect to the groupings of patients solely using outcome data. The optimisation procedure was performed using a genetic algorithm (GA) based approach utilising the distributed evolutionary algorithms for python (DEAP) library [53]. The ”individuals” involved in the GA routine are vectors comprising the identity of the risk group of each patient in the analysis. The ”fitness” of an individual set is the log-rank statistic — calculated using the patient risk groups specified by that individual.

Plotting the maximum individual fitness of the generation against the generation number, shows a plateauing of the log-rank statistic at around 40 (Figure 3.2 [A]). The resulting risk groups show a clear distinction in clinical outcomes (Figure 3.2 [B,C]).

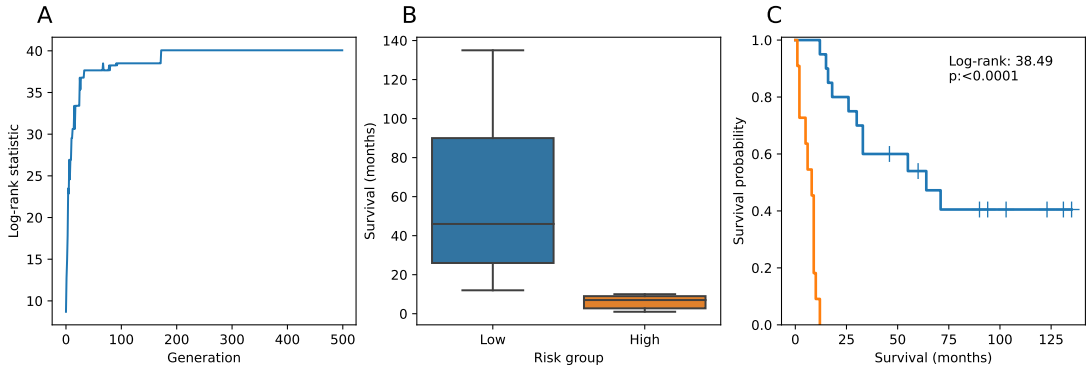


FIGURE 3.2: Stratification of patients into high and low-risk. A: Maximum log-rank statistic vs GA generation, plateauing around 40. B: Whisker box plots of survival duration in each risk group. C: Kaplan-Meier Survival curves showing optimal risk stratification of the patient cohort; also shown are the log-rank statistic and corresponding p-value for the optimal groupings.

3.3 Results

The inclusion criteria designated in Materials and Methods gave a sample set of 29 patients from the original 102 previously described [32]. Of these 29 patients, 19 remained alive 12 months after surgery, while 14 remained alive at two years. The cohort was representative of this original patient set except that the male: female ratio was inverted (Table 3.1) but now corresponds more closely with larger datasets (Table 3.1; final column). In agreement with the published cohort, the sample set was enriched for cases with ENE (i.e. poor prognosis) in comparison to the general HNSCC population [54].

TABLE 3.1: Characteristics of the sample cohort

	All (N=29)	Outcome at 12 months		Outcome at 24 months		Original cohort [32]	Larger, local cohort [54]
		Dead	Alive	Dead	Alive	N=102	N=489
Age (years)							
Mean	60	70.4	56.3	64.5	57.2	60	
Range	29-85	59-85	29-72	48-85	29-68	29-89	
Median	61	70.5	56.5	64	60		
α-SMA							
High/Intermediate	26 (94)	10	16	14	12	60 (64)	ND
Low	3 (6)	0	3	1	2	33 (36)	ND
Gender							
F	7 (24)†	0	7	2	5	57 (56)^{a}	187 (38)^{b}
M	22 (76)	10	12	13	9	45 (45)	302 (62)
T Stage*							
1	1 (3)	0	1	0	1	8 (8)	123 (25)^{b}
2	14 (48)	4	10	7	8	57 (54)	175 (35)
3	2 (7)	0	2	2	0	12 (11)	47 (10)
4	8 (28)	4	4	3	5	20 (19)	144 (30)
4a	4 (14)	2	2	2	2	9 (8)	
N Stage$^{\#}$							
0	7 (24)	2	4	1	5	38 (37)^{b}	314 (64)^{c}
1	7 (24)	2	4	2	4	18 (17)	64 (14)
2a	1 (3)	1	0	1	0		101(20)
2b	16 (55)	4	9	8	5	45 (44)	
2c	3 (10)	1	2	3	0		
Pathological Site							
Floor of mouth	8 (28)	2	6	4	4	35 (34)^{b}	162 (33)
Other	12 (41)	5	7	5	7	24 (24)	183 (36)
Tongue	9 (31)	3	6	6	3	37 (36)	144 (30)

a: $p<0.005$

b: p=NS

c: $p<0.00001$

*: T stage 1+2 v 3+4

†: numbers in parenthesis are percentages

#: N stage 0 v 1 v 2

FTIR and ASMA data from the reduced sample set were evaluated as prognostic indicators of death within one year of surgery, both separately and together (Figure 3.3). A total of 168,460 FTIR spectra were obtained from the 19 patients who survived beyond one year, and 96,402 spectra were obtained from 10 patients who died within 12 months.

Evaluation of individual spectra

Using prediction scores derived from individual spectra only, ROC and PR analysis were performed (Figure 3.3). The ASMA model shows poor performance across all thresholds for both analyses; the FTIR model is slightly better and is a reasonable predictor in its own right. However, the combined model shows superior ROC and PR curves overall. The baseline shown in Figure 3.3[B] corresponds to the prevalence of spectra labelled as high-risk.

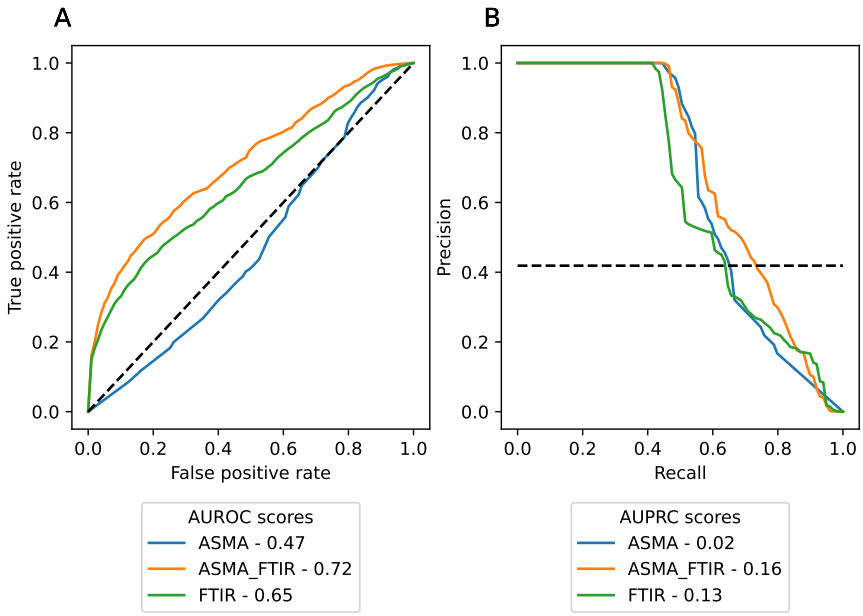


FIGURE 3.3: Median ROC and PR curves shown in solid lines; dashed lines represent baselines scores associated with random chance. AUROC and AUPRC scores are shown for each set of prognostic indicators.

(Figure 3.4) shows distributions of classification statistics calculated for each set of prognostic variables. As shown in (Figure 3.3[A,B]) AUROC and AUPRC scores of the FTIR models show promising scores but also a small spread in comparison to the ASMA model. Many other statistics show the combined model as the most effective, with the exception of the sensitivity score.

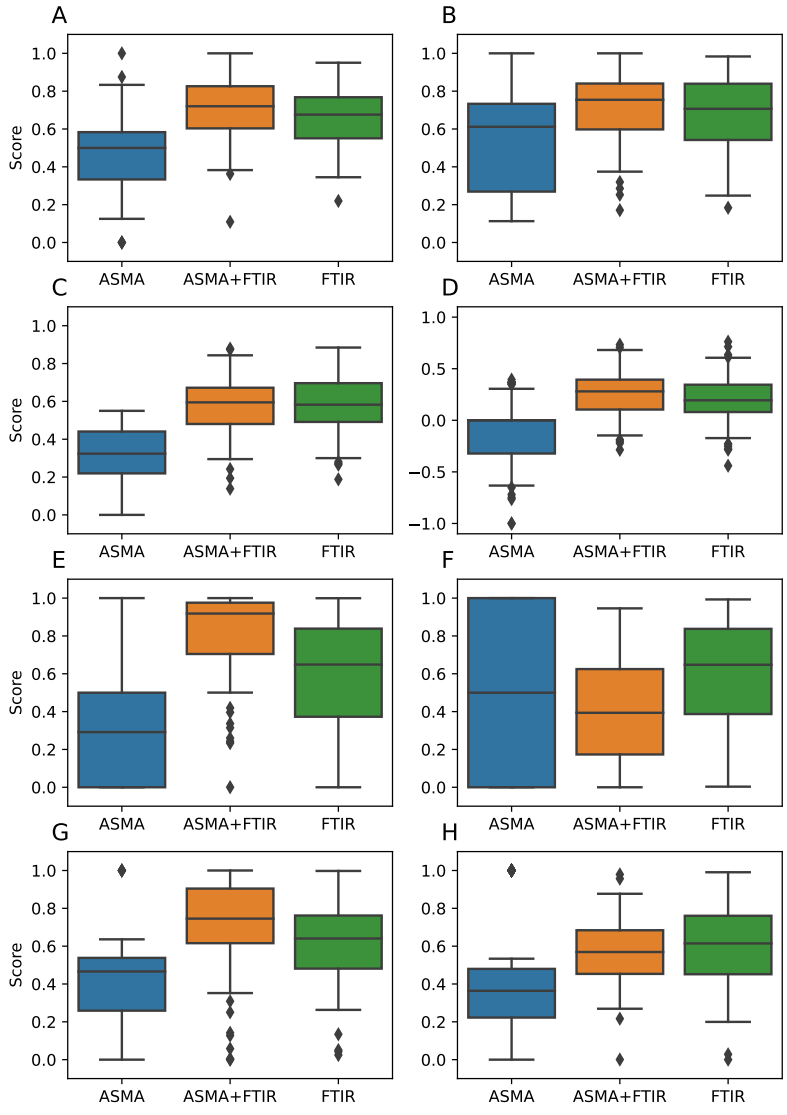


FIGURE 3.4: Whisker boxplots of classification statistics calculated across all data subsets. AUROC (A); AUPRC (B); F1 (C); MCC (D); specificity (E); sensitivity (F); PPV (G); NPV (H). Boxes show the median, 25^{th}, and 75^{th} percentiles; whiskers extend to points that lie within 1.5 inter quartile ranges of the lower and upper quartiles; points lying outside this range are shown as individual diamonds.

Evaluation on a patient basis

To obtain a prediction of the risk group for each patient, final prediction scores were taken as the median probability predicted across all spectra for any given patient.

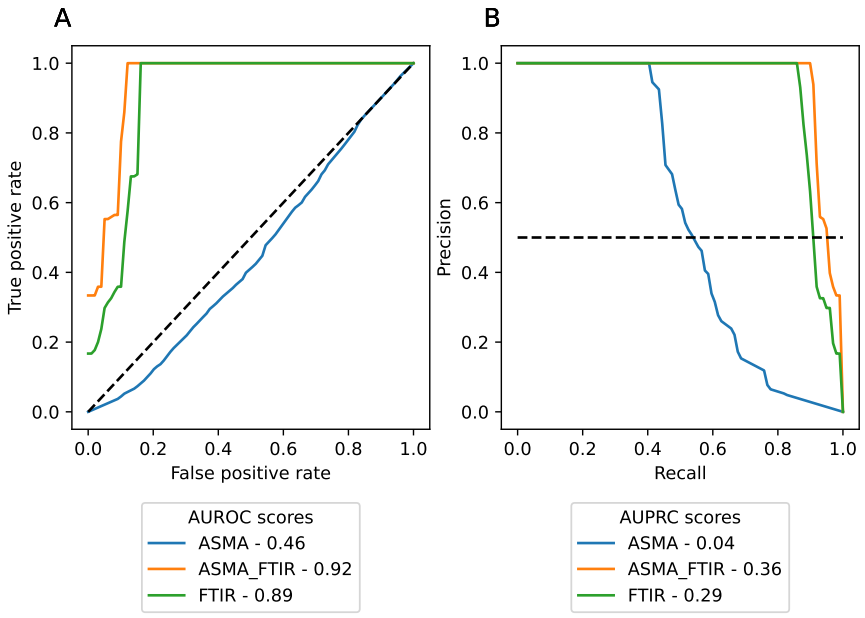


FIGURE 3.5: Median ROC and PR curves shown in solid lines; dashed lines represent baselines scores associated with random chance. AUROC and AUPRC scores are shown for each set of prognostic indicators.

The median AUROC obtained from FTIR alone was 0.89 (Figure 3.6 [A,C]); incorporation of the ASMA data into this analysis, increased the AUROC slightly to 0.92; while ASMA alone achieved a significantly poorer score of 0.46. Precision and recall scores both remained high for the FTIR and combined ASMA/FTIR models across a range of decision thresholds — indicating that both models can balance both statistics effectively, and that imbalance in the dataset was not detrimental (Figure 3.6 [B]).

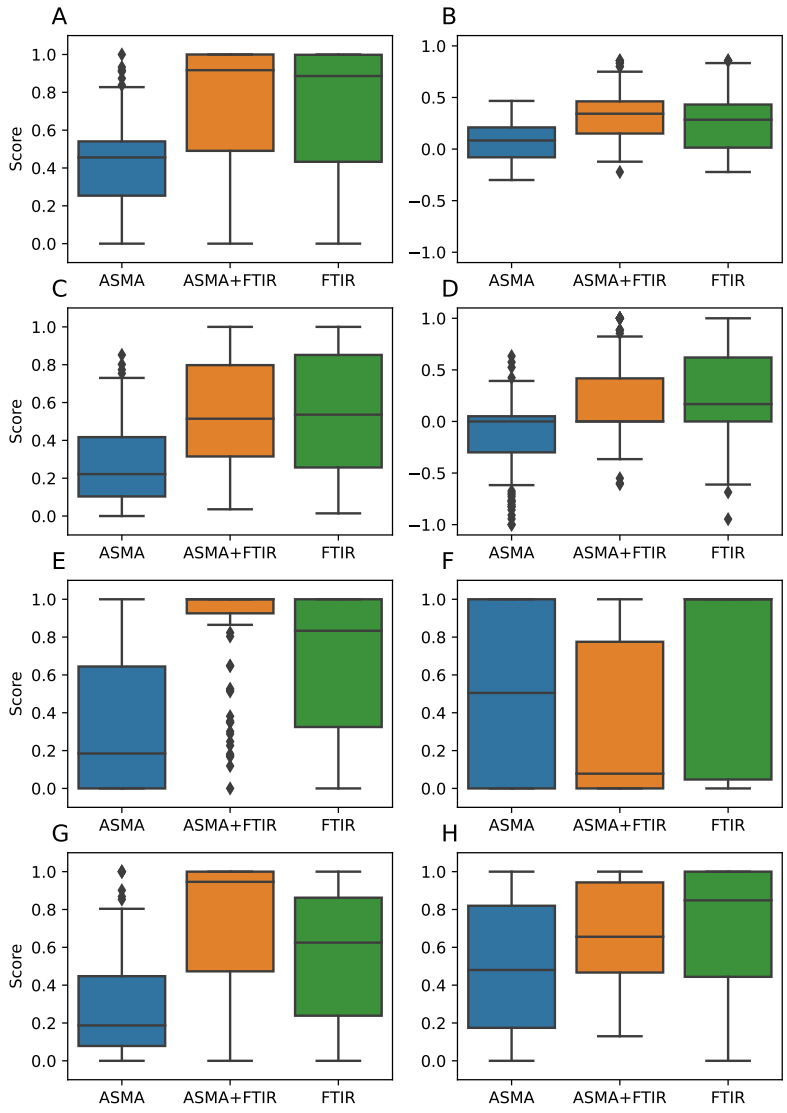


FIGURE 3.6: Whisker boxplots of classification statistics calculated across all data subsets. AUROC (A); AUPRC (B); F1 (C); MCC (D); specificity (E); sensitivity (F); PPV (G); NPV (H). Boxes show the median, 25^{th}, and 75^{th} percentiles; whiskers extend to points that lie within 1.5 inter quartile ranges of the lower and upper quartiles; points lying outside this range are shown as individual diamonds.

Furthermore, additional classification statistics produced comparable conclusions, showing that the combined model was a good predictor of poor outcome (Figure 3.6 [A-H]).

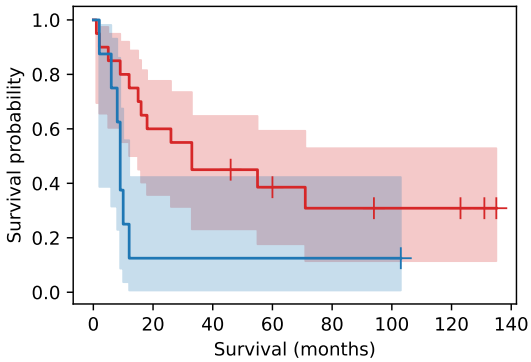
TABLE 3.2: Median classification statistics. Classification thresholds (Table 3.2) used to convert probabilities to binary decisions were determined to be those that maximised the log-rank statistic.

Variables	AUROC	AUPRC	F1	MCC	Spec	Sens	PPV	NPV	Thresh
ASMA	0.46	0.41	0.22	0.00	0.18	0.50	0.19	0.48	0.48
ASMA+FTIR	0.92	0.85	0.51	0.00	1.00	0.08	0.95	0.66	0.69
FTIR	0.89	0.79	0.54	0.17	0.83	1.00	0.62	0.85	0.34

A comparison of the median scores of these statistics shows that ASMA alone is a poor predictive variable for this dataset with low scores in all metrics (Table 3.2), while the FTIR alone scores are consistently high. The combined model shows high specificity and low sensitivity, with a high PPV and moderate NPV indicating numerous false negatives.

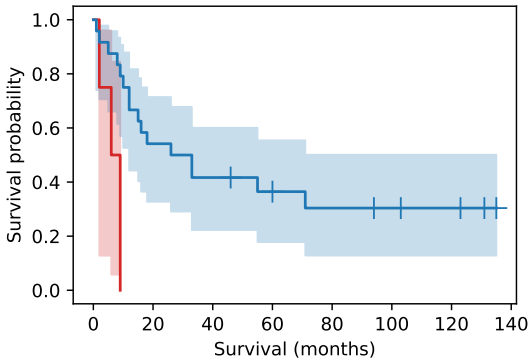
The classification threshold for each model was used to assign patients to high or low-risk groups and survival analyses were undertaken. As expected, the use of ASMA alone did not show good separation of the high and low-risk groups as shown in Kaplan-Meier curves (Figure 3.7a) and, indeed, the risk group predictions were inverted to that expected. The use of FTIR alone to predict risk showed significant separation of the groups (Figure 3.7c; p=0.01), while the combined model produced good separation with a highly significant p value (Figure 3.7b; $p<0.005$).

(A) ASMA



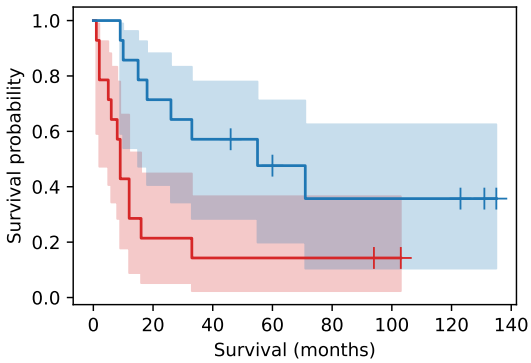
Risk		0	20	40	60	80	100	120
High	At risk	21	13	9	6	5	4	3
	censored	0	0	2	1	0	1	1
Low	At risk	8	2	2	1	1	1	-
	censored	0	0	0	1	1	1	-

(B) ASMA + FTIR



Risk		0	20	40	60	80	100	120
High	At risk	4	-	-	-	-	-	-
	censored	0	-	-	-	-	-	-
Low	At risk	25	14	10	7	6	4	3
	censored	0	0	2	1	0	1	1

(C) FTIR



Risk		0	20	40	60	80	100	120
High	At risk	14	4	3	3	2	1	-
	censored	0	0	0	0	1	1	-
Low	At risk	15	11	8	5	4	3	3
	censored	0	0	2	1	0	1	1

FIGURE 3.7: Kaplan-Meier survival curves for each risk group according input variables. Low-risk:blue, high-risk:red. Confidence intervals are computed using the exponential Greenwood method [55].

A univariate Cox proportional hazard model was fit to prediction scores obtained from the LR model to assess the prognostic utility of the prediction score before conversion to a binary decision. Both models using FTIR data have significantly higher hazard ratios than the pure ASMA model suggesting the LR

classifier is able to stratify risk groups effectively. However, the 95% confidence intervals for all models spanned a wide range.

TABLE 3.3: Cox proportional hazards model fit statistics

Variables	Coefficients	z	p	Hazard ratio
ASMA	-0.03	-0.02	0.98	0.97 (0.06-15.08)
ASMA+FTIR	1.84	2.12	0.03	6.29 (1.14-34.59)
FTIR	2.02	2.07	0.04	7.52 (1.12-50.62)

(Figures 3.8a to 3.8c) show predicted survival curves for five simulated patients. Patients' prediction scores vary between 0 (definitely low-risk) and 1 (definitely high-risk), showing how LR predictions translate to expected survival outcomes. As expected (Figure 3.8a) showed very little stratification between simulated predictions; this is consistent with a hazard ratio of $\sim$1 as shown in (Table 3.3). The combined model shows a much better level of stratification by risk than ASMA alone, with distinct survival curves clearly visible. The FTIR model also shows distinct survival curves consistent with a large hazard ratio. The addition of ASMA data appears to be detrimental to the overall prognostic utility of the combined model.

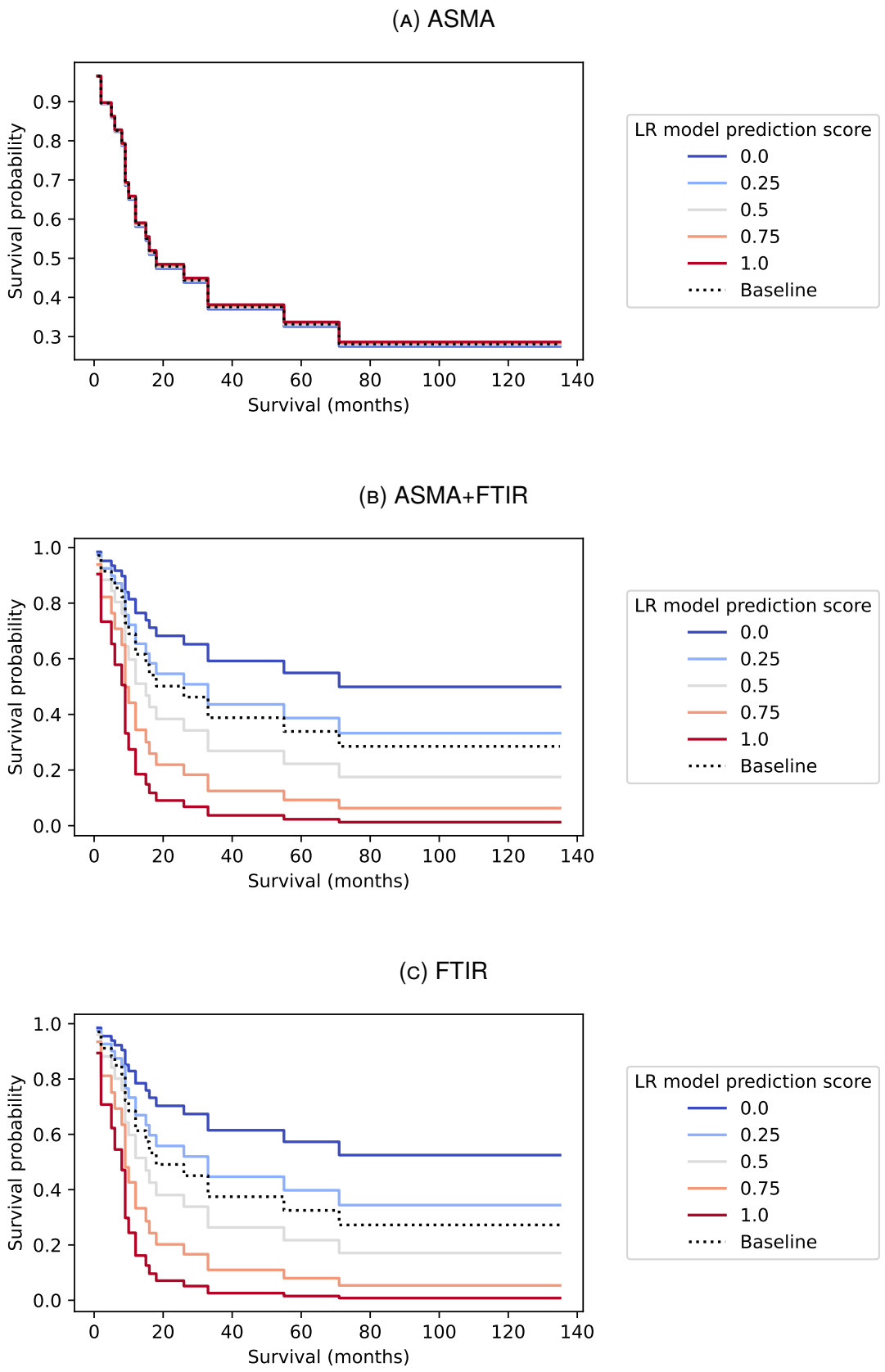


FIGURE 3.8: Predicted survival curves for five simulated patients varying between 0 and 1. A dotted line represents the hazard baseline.

3.4 Discussion

It has previously been established that the presence of ENE is a marker of poor prognosis in oral cancer. However, the presence of ENE cannot be firmly established until after pathological inspection of surgically removed nodal material. This prevents the use of neo-adjuvant treatment in those patients with poor prognosis, for whom current treatment regimens do not work. It has been shown that the chemical 'fingerprint' of the primary tumour, readily observed through the use of FTIR microscopy is able to identify individuals whose disease will progress following treatment.

Much of the existing literature concerning prognostic indicators utilises disease-specific survival and overall survival as indicators of patient prognoses. Stratification of these measures into groups according to risk is typically decided according to a number of years e.g. one year, two years. The decision to use a cut off threshold of a discrete number of years is somewhat arbitrary, thus, it would be desirable to determine a threshold that was decided objectively. A GA approach was used to stratify patients in this cohort into either a low or high-risk group, where the log-rank statistic calculated from a log-rank test was maximised — indicating the best separation. This threshold was determined to be 11 months and was used to dichotomize patients into risk groups.

Results presented here suggest that FTIR spectra can be used to stratify patients into useful clinical risk groups. Models utilising a combination of FTIR and ASMA data showed some diminished prognostic utility — suggesting that the addition of ASMA data was detrimental. Many other statistics show the combined model as the most effective, with the exception of the sensitivity score; indicating a high proportion of false negatives and thus suggesting that low-risk patients are often misclassified and would be undertreated in a clinical setting. The FTIR model scores are generally high, indicating that FTIR spectra

can serve as a prognostically useful biomarker. A slightly lower PPV however suggests the FTIR model incurs more false positives — indicating that patients with better prognoses may be given inappropriate treatment.

A significant improvement was observed when taking the median prediction score for each patient. This is potentially a reflection of the fact that the molecular fingerprint for poor prognosis varies in magnitude across the measured tumour section — genetic heterogeneity within OSCC lesions has been noted previously [56]. The aggregation of scores across an entire tumour section may cause a 'regularising' effect, thus mitigating the effects of overfitting to specific chemical fingerprints. This regularisation effect may be dependent upon the size of the measured tumour section present within the core. Data subsets containing patients with relatively few measured spectra may experience lower scores as a result; possibly explaining the large degree of variation within sample scores to some extent.

Survival analysis showed that groups allocated by the classifier had significantly different outcomes. The combined model (Figure 3.7b) is able to closely replicate the ideal survival curves in (Figure 3.2[C]). Model predictions using FTIR data showed a marked improvement over the pure ASMA model. Hazard ratios of 6.29 and 7.5 for the ASMA+FTIR and FTIR model respectively show that the prognoses of patients allocated to the high-risk group are significantly poorer. 95% confidence intervals for hazard ratios span a large range (FTIR: 1.12-50.62; ASMA+FTIR: 1.14-34.59). This is indicative of a high degree of heterogeneity in the sample cohort and would justify further exploration with a larger cohort in the future.

A relatively small sample set was a key issue facing this study due to the difficulty in acquiring and imaging large numbers of samples. Despite attempts to determine the feasibility of FTIR as a prognostic tool through multiple sampling of the dataset; a larger study would be required in the future to estimate wider

clinical utility. A large degree of variation was observed across some classification statistics, potentially signifying a large degree of biological heterogeneity in the dataset. A potential cause for this could be the effect of inherent molecular heterogeneity of the tumour microenvironment, or perhaps varying extents of lymphocyte infiltration present in specimens. The difficulty of annotating samples is also likely to introduce noise into the dataset; alongside inconsistencies in sample preparation and measurement procedure.

3.5 Conclusion

It would be of considerable benefit to be able to direct patients with poor prognoses towards appropriate treatment. Currently, it is not ethically possible to select patients for neo-adjuvant treatment in window trials; patients who would not benefit from such treatment would incur unnecessary adverse effects and additional health risks. It is currently only possible to determine patient prognoses after post-surgical nodal biopsies have taken place. The work presented here would allow for this crucial window of opportunity to be seized and to enable the development of new treatment methods to take place.

The use of FTIR in a clinical setting is still in its infancy, however, the work covered here shows that it has the potential to be of significant benefit as a prognostic tool. The addition of ASMA information was shown to be beneficial in certain cases, and demonstrates that additional information from other modalities could lead to the creation of a novel and informative prognostic tool. FTIR spectroscopy has been shown to be capable of detecting a molecular fingerprint associated with poor prognosis; combining FTIR spectra with additional measurement modalities from the same patient would further test this hypothesis. As multiple adjacent slices of the same biopsy sample can be acquired and divided across multiple imaging modalities; the respective strengths

could be compounded whilst simultaneously mitigating the weaknesses of each modality with minimal disruption to current clinical routines.

Bibliography

[1] Matt Lechner, Jacklyn Liu, Liam Masterson, and Tim R. Fenton. HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management. *Nature Reviews Clinical Oncology*, 0123456789, 2022.

[2] Head and neck cancers statistics, Sep 2021.

[3] Mauricio E. Gamez, Ryan Kraus, Michael L. Hinni, Eric J. Moore, Daniel J. Ma, Stephen J. Ko, Jean Claude M. Rwigema, Lisa A. McGee, Michele Y. Halyard, Matthew R. Buras, Robert L. Foote, and Samir H. Patel. Treatment outcomes of squamous cell carcinoma of the oral cavity in young adults. *Oral Oncology*, 87(August):43–48, 2018.

[4] Torbjörn Ramqvist and Tina Dalianis. An epidemic of oropharyngeal squamous cell carcinoma (OSCC) due to human papillomavirus (HPV) infection and aspects of treatment and prevention. *Anticancer research*, 31(5):1515–9, 2011.

[5] Victor Lopes, Paul Murray, Hazel Williams, Ciaran Woodman, John Watkinson, and Max Robinson. Squamous cell carcinoma of the oral cavity rarely harbours oncogenic human papillomavirus. *Oral Oncology*, 47(8):698–701, 2011.

[6] Athanassios Argiris, Neck Cancer Program, Michalis V Karamouzis, Neck Cancer Program, David Raben, Robert L Ferris, Neck Cancer Program, Cancer Immunology, and Immunoprevention Program. Head and neck cancer Athanassios. *Lancet*, 371(9625):1695–1709, 2008.

[7] 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.

[8] Ari J. Rosenberg and Everett E. Vokes. Optimizing Treatment De-Escalation in Head and Neck Cancer: Current and Future Perspectives. *The Oncologist*, 26(1):40–48, 2021.

[9] Mahul B. Amin, Frederick L. Greene, Stephen B. Edge, Carolyn C. Compton, Jeffrey E. Gershenwald, Robert K. Brookland, Laura Meyer, Donna M. Gress, David R. Byrd, and David P. Winchester. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more âĂIJpersonalizedâĂİ approach to cancer staging. *CA: A Cancer Journal for Clinicians*, 67(2):93–99, 2017.

[10] Sang Ik Park, Jeffrey P. Guenette, Chong Hyun Suh, Glenn J. Hanna, Sae Rom Chung, Jung Hwan Baek, Jeong Hyun Lee, and Young Jun Choi. The diagnostic performance of CT and MRI for detecting extranodal extension in patients with head and neck squamous cell carcinoma: a systematic review and diagnostic meta-analysis. *European Radiology*, 31(4):2048–2061, 2021.

[11] Benjamin H. Kann, Daniel F. Hicks, Sam Payabvash, Amit Mahajan, Justin Du, Vishal Gupta, Henry S. Park, James B. Yu, Wendell G. Yarbrough, Barbara A. Burtness, Zain A. Husain, and Sanjay Aneja. Multi-institutional validation of deep learning for pretreatment identification of extranodal extension in head and neck squamous cell carcinoma. *Journal of Clinical Oncology*, 38(12):1304–1311, 2020.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Diana Graizel, Ayelet Zlotogorski-Hurvitz, Igor Tsesis, Eyal Rosen, Ron Kedem, and Marilena Vered. Oral cancer-associated fibroblasts predict poor survival: Systematic review and meta-analysis. *Oral Diseases*, 26(4):733–744, 2020.

[17] Tseng-Cheng Chen, Chen-Tu Wu, Cheng-Ping Wang, Wan-Lun Hsu, Tsung-Lin Yang, Pei-Jen Lou, Jenq-Yuh Ko, and Yih-Leong Chang. Associations among pretreatment tumor necrosis and the expression of hif-1α and pd-l1 in advanced oral squamous cell carcinoma and the prognostic impact thereof. *Oral Oncology*, 51(11):1004–1010, 2015.

[18] Anthony C Nichols, Pencilla Lang, Eitan Prisman, Eric Berthelet, Eric Tran, Sarah Hamilton, Jonn Wu, Kevin Fung, John R de Almeida, Andrew Bayley, et al. Treatment de-escalation for hpv-associated oropharyngeal squamous cell carcinoma with radiotherapy vs. trans-oral surgery (orator2): study protocol for a randomized phase ii trial. *BMC cancer*, 20(1):1–13, 2020.

[19] Alexandra Sala, David J. Anderson, Paul M. Brennan, Holly J. Butler, James M. Cameron, Michael D. Jenkinson, Christopher Rinaldi, Ashton G. Theakstone, and Matthew J. Baker. Biofluid diagnostics by FTIR spectroscopy: A platform technology for cancer detection. *Cancer Letters*, 477(December 2019):122–130, 2020.

[20] Vera E. Sitnikova, Mariia A. Kotkova, Tatiana N. Nosenko, Tatiana N. Kotkova, Daria M. Martynova, and Mayya V. Uspenskaya. Breast cancer detection by ATR-FTIR spectroscopy of blood serum and multivariate data-analysis. *Talanta*, 214(October 2019):120857, 2020.

[21] Daniela Lazaro-Pacheco, Abeer Shaaban, Gouri Baldwin, Nicholas Akinwale Titiloye, Shazza Rehman, and Ihtesham ur Rehman. Deciphering the structural and chemical composition of breast cancer using FTIR spectroscopy. *Applied Spectroscopy Reviews*, 0(0):1–15, 2020.

[22] 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.

[23] Sebastian Berisha, Mahsa Lotfollahi, Jahandar Jahanipour, Ilker Gurcan, Michael Walsh, Rohit Bhargava, Hien Van Nguyen, and David Mayerich. Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks. *Analyst*, 144(5):1642–1653, 2019.

[24] J. Ingham, M.J. Pilling, D.S. Martin, C.I. Smith, B.G. Ellis, C.A. Whitley, M.R.F. Siggel-King, P. Harrison, T. Craig, A. Varro, D.M. Pritchard, A. Varga, P. Gardner, P. Weightman, and S. Barrett. A novel FTIR analysis method for rapid high-confidence discrimination of esophageal cancer. *Infrared Physics and Technology*, 102, 2019.

[25] Donna E. Maziak, Minh T. Do, Farid M. Shamji, Sudhir R. Sundaresan, D. Garth Perkins, and Patrick T.T. Wong. Fourier-transform infrared spectroscopic study of characteristic molecular structure in cancer cells of esophagus: An exploratory study, 2007.

[26] A. L. M. Batista de Carvalho, M. Pilling, P. Gardner, J. Doherty, G. Cinque, K. Wehbe, C. Kelley, L. A. E. Batista de Carvalho, and M. P. M. Marques. Chemotherapeutic response to cisplatin-like drugs in human breast cancer cells probed by vibrational microspectroscopy. *Faraday Discuss.*, 187:273–298, 2016.

[27] Richard J. Shaw, Derek Lowe, Julia A. Woolgar, James S. Brown, E. David Vaughan, Christopher Evans, Huw Lewis-Jones, Rebecca Hanlon, Gillian L. Hall, and Simon N. Rogers. Extracapsular spread in oral squamous cell carcinoma. *Head Neck*, 36(10):NA–NA, 2009.

[28] Maxime Mermod, Genrich Tolstonog, Christian Simon, and Yan Monnier. Extracapsular spread in head and neck squamous cell carcinoma: A systematic review and meta-analysis. *Oral Oncology*, 62:60–71, 2016.

[29] David Chin, Glen M. Boyle, Rebecca M. Williams, Kaltin Ferguson, Nirmala Pandeya, Julie Pedley, Catherine M. Campbell, David R. Theile, Peter G. Parsons, and William B. Coman. Novel markers for poor prognosis in head and neck cancer. *International Journal of Cancer*, 113(5):789–797, 2005.

[30] M. G. Kellermann, L. M. Sobral, S. D. Da Silva, K. G. Zecchin, E. Graner,

M. A. Lopes, I. Nishimoto, L. P. Kowalski, and R. D. Coletta. Myofibroblasts in the stroma of oral squamous cell carcinoma are associated with poor prognosis [3]. *Histopathology*, 51(6):849–853, 2007.

[31] Marilena Vered, Alex Dobriyan, Dan Dayan, Ran Yahalom, Yoav P. Talmi, Lev Bedrin, Iris Barshack, and Shlomo Taicher. Tumor-host histopathologic variables, stromal myofibroblasts and risk score, are significantly associated with recurrent disease in tongue cancer. *Cancer Science*, 101(1):274–280, 2010.

[32] J Dhanda, A Triantafyllou, T Liloglou, H Kalirai, B Lloyd, R Hanlon, R J Shaw, and D R Sibson. SERPINE1 and SMA expression at the invasive front predict extracapsular spread and survival in oral squamous cell carcinoma. (August):2114–2121, 2014.

[33] Chun Ta Liao, Li Yu Lee, Chuen Hsueh, Chien Yu Lin, Kang Hsing Fan, Hung Ming Wang, Chia Hsun Hsieh, Shu Hang Ng, Chih Hung Lin, Chung Kan Tsao, Chung Jan Kang, Tuan Jen Fang, Shiang Fu Huang, Kai Ping Chang, Lan Yan Yang, and Tzu Chen Yen. Pathological risk factors stratification in pN3b oral cavity squamous cell carcinoma: Focus on the number of positive nodes and extranodal extension. *Oral Oncology*, 86(September):188–194, 2018.

[34] Christine T. Lwin, Rebecca Hanlon, Derek Lowe, James S. Brown, Julia A. Woolgar, Asterios Triantafyllou, Simon N. Rogers, Fazilet Bekiroglu, Huw Lewis-Jones, Hulya Wieshmann, and Richard J. Shaw. Accuracy of MRI in prediction of tumour thickness and nodal stage in oral squamous cell carcinoma. *Oral Oncology*, 48(2):149–154, 2012.

[35] Peng Wang, Aron Popovtzer, Avraham Eisbruch, and Yue Cao. An approach to identify, from DCE MRI, significant subvolumes of tumors related to outcomes in advanced head-and-neck cancer. *Medical Physics*,

39(8):5277–5285, 2012.

[36] Tobias Waech, Shila Pazahr, Vittoria Guarda, Niels J. Rupp, Martina A. Broglie, and Grégoire B. Morand. Measurement variations of MRI and CT in the assessment of tumor depth of invasion in oral cancer: A retrospective study. *European Journal of Radiology*, 135, 2021.

[37] Jenny K. Hoang, Jyotsna Vanka, Benjamin J. Ludwig, and Christine M. Glastonbury. Evaluation of cervical lymph nodes in head and neck cancer with CT and MRI: Tips, traps, and a systematic approach. *American Journal of Roentgenology*, 200(1):17–25, 2013.

[38] Ming Hui Mao, Shu Wang, Zhi en Feng, Jin Zhong Li, Hua Li, Li zheng Qin, and Zheng xue Han. Accuracy of magnetic resonance imaging in evaluating the depth of invasion of tongue cancer. A prospective cohort study. *Oral Oncology*, 91(October 2018):79–84, 2019.

[39] Izabela Zawlik, Ewa Kaznowska, Jozef Cebulski, Magdalena Kolodziej, Joanna Depciuch, Jitraporn Vongsvivut, and Marian Cholewa. FPA-FTIR Microspectroscopy for Monitoring Chemotherapy Efficacy in Triple-Negative Breast Cancer. *Scientific Reports*, 6:1–8, 2016.

[40] Holly J. Butler, Paul M. Brennan, James M. Cameron, Duncan Finlayson, Mark G. Hegarty, Michael D. Jenkinson, David S. Palmer, Benjamin R. Smith, and Matthew J. Baker. Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. *Nature Communications*, 10(1):1–9, 2019.

[41] J Zapała, J Czopek, A Lazar, and R Tomaszewska. Proliferative index activity in oral squamous cell carcinoma : indication for postoperative radiotherapy? pages 1189–1194, 2014.

[42] Thomas Scholzen and Johannes Gerdes. The Ki-67 Protein : From the Known and. 322(August 1999):311–322, 2000.

[43] Neelam G Shah, Trupti l Trivedi, Rajen A Tankshali, Jignesh V Goswami, Dhaval H Jetly, Shilin N Shukla, Pankaj M Shah, and Ramtej J Verma. Prognostic significance of molecular markers in oral squamous cell carcinoma : a multivariate analysis. (December):1544–1556, 2009.

[44] César Rivera, Ana Karina Oliveira, Rute Alves Pereira Costa, Tatiane De Rossi, and Adriana Franco Paes Leme. Prognostic biomarkers in oral squamous cell carcinoma: A systematic review. *Oral Oncology*, 72:38–47, 2017.

[45] The GIMP Development Team. Gimp.

[46] Alex Henderson. Chitoolbox, 2022.

[47] 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.

[48] Pitard F.F. An Introduction to the Theory of Sampling: An Essential Part of Total Quality Management. *Comprehensive Chemometrics*, pages 1–16, 2009.

[49] Wolfgang Karl Härdle and Léopold Simar. *Applied multivariate statistical analysis, fourth edition*. 2015.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[51] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

[52] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[53] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175, jul 2012.

[54] Simon N. Rogers, James S. Brown, Julia A. Woolgar, Derek Lowe, Patrick Magennis, Richard J. Shaw, David Sutton, Douglas Errington, and David Vaughan. Survival following primary surgery for oral cancer. *Oral Oncology*, 45(3):201–211, 2009.

[55] S Sawyer. The Greenwood and Exponential Greenwood Confidence Intervals in Survival Analysis. *Health (San Francisco)*, (i):2–5, 2003.

[56] Alhadi Almangush, Ilkka Heikkinen, Antti A. Mäkitie, Ricardo D. Coletta, Esa Läärä, Ilmo Leivo, and Tuula Salo. Prognostic biomarkers for oral tongue squamous cell carcinoma: A systematic review and meta-analysis. *British Journal of Cancer*, 117(6):856–866, 2017.

4 FTIR Preprocessing Pipeline Optimisation

4.1 Introduction

In recent decades, vibrational spectroscopy has emerged as a useful analysis method applicable to a range of specimens, from food and pharmaceutical samples to security and medical applications. Vibrational spectroscopy techniques interrogate the molecular structure of a specimen by utilising infrared radiation, either directly (infrared absorption spectroscopy) or indirectly (Raman scattering spectroscopy). The absorption of discrete photon energies by molecular vibrational modes within the sample gives rise to characteristic spectral profiles, which can be associated with the inherent chemistry of the sample under interrogation.

4.1.1 Optimisation

Despite the undeniable promise of the field, it is somewhat hampered by the lack of consensus on precisely how to preprocess the data and subsequently analyse it. The number of available preprocessing methods and predictive models is vast; each method typically has one or more parameters associated with it, exacerbating the problem further.

Analysis of spectroscopic data is typically a multi-step procedure, starting with preprocessing, and ultimately the employment of pattern recognition and machine learning tools to classify the data into distinct groups. Preprocessing is a vital step in the analysis process of IR data, as it has been shown to generally increase performance of classification models [1] allowing them to generalise more effectively to larger clinical cohorts, and to increase the interpretability of results.

FTIR and Raman data sets are highly dimensional. In the case of Raman mapping and FTIR imaging, sample numbers are very large, with data sets comprising large numbers of patients typically having tens of thousands of spectra. A further motivating factor known as the 'no free lunch' (NFL) theorem states that 'any two optimisation algorithms are equivalent when their performance is averaged across all possible problems' [2]. This was stated generally for optimisation problems; due to the close relationship between optimisation and machine learning methods, the same theorem applies [3]. The implications of the NFL theorem would require a thorough search of multiple machine learning models to ensure some degree of certainty that the task is possible.

An optimisation protocol that can efficiently search across a highly dimensional space would be of considerable benefit to users of multivariate analysis techniques. One approach [4] utilised a genetic algorithm (GA) to perform an optimisation search procedure to determine more effective processing pipelines. A single 'individual' in this framework was represented as a processing pipeline; many generations of preprocessing pipelines were allowed to 'evolve' in a manner analogous to Darwinian evolution. The 'fittest' individuals which influence subsequent generations were chosen to be those with the lowest prediction error, hence directing the process towards optimal values. The optimised pipeline resulted in a 16% reduction in the model error compared with

the raw data model. Similar work [5] utilised a GA approach to assist with feature selection. The authors were interested in using Curie-point pyrolysis mass spectrometry to differentiate between bacterial spore samples. The highly dimensional nature of these datasets makes interpretation prohibitively difficult; the GA approach was used to select a subset of these features for further analysis. A trial-and-error approach was more recently reported [6] which trialled every permutation of preprocessing steps within a defined search space on an ATR-FTIR biofluid dataset comprising patients with varying types of brain cancer. The authors utilised RF and SVM classifiers. However, the hyperparameters of the final classifiers were not optimised in this approach. A brute force approach will cover every possible combination, but the number of combinations grows extremely quickly and is generally infeasible as a strategy.

This work proposes a novel approach to objectively optimise an effective preprocessing and classification pipeline. We perform a Bayesian hyperparameter search on several candidate pipelines using a parallel computing approach — more efficiently searching all possible solutions. The focus of this paper will be on FTIR imaging data, but the process is in theory generalisable to *any* pipeline type inference problem.

4.2 Theoretical

4.2.1 Preprocessing

Preprocessing of FTIR data can be broken into several discrete steps, with each step designated to mitigate unwanted spectral aberrations and measurement artefacts. Using modern object-oriented programming languages, preprocessing step can be encapsulated as a transformer object — an abstract representation of a preprocessing step. A transformer will typically take a data set as

input, perform the transformation associated with it, and then output the transformed data. A sequence of transformers with or without a final estimator can be visualised as a pipeline. A pipeline consists of a sequence of transformers that take data as an input and pass the data through each transformer sequentially until a final result is obtained. Such a preprocessing sequence is briefly set out below.

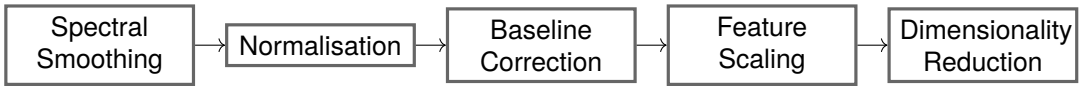


FIGURE 4.1: Preprocessing pipeline flowchart

Spectral smoothing - *Accounts 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 [7], whereby a polynomial is fit to a local moving window of a specified length. Other methods such as PCA de-noising and FFT filtering are also commonly applied.

Normalisation - *Accounts for unimportant changes in absolute absorbances*. This important step accounts for absolute differences in absorption according to the Beer-Lambert law [8] — which is to be expected with a biologically diverse data-set from multiple samples. This step helps mitigate the effect of sample thickness on the analysis — a potential confounding factor.

Baseline correction - *Accounts for variations* . Spectra are often superimposed on a non-linear baseline due to background scattering interference effects. Methods such as rubberband correction and spectral differentiation are often applied to compensate for these effects.

Feature scaling - *Ensures variables all lie within the same range*. A scaling step is sometimes applied to scale the absorbance for each wavenumber to a common range. This can often increase the performance of some classifiers and is sometimes a required preprocessing step for dimensionality reduction algorithms.

Dimensionality reduction - *A reduction in the number of predictive variables*. Reducing the dimensionality of the data to a much lower space often facilitates more robust classifier results and can mitigate issues arising from correlated variables. Feature selection/extraction techniques such as PCA [9] and forward feature selection (FFS) [10] are often used to reduce the number of variables while retaining maximum important information.

In addition to numerous preprocessing methods, each may have intrinsic parameters, referred to from here on as *hyperparameters*, which must also be determined. Exhaustively trialling every possible combination of preprocessing transformations to a dataset is often prohibitive. Whilst this would yield the true optimal combination of methods, it would be at a high computational cost.

Whilst it is possible to exhaustively trial all preprocessing transformations to find the true optimal combination of methods, this may come at a prohibitive computational cost. For instance, consider a case where there are five preprocessing steps, each containing five methods to select from. Each of these methods has a hyperparameter space that spans twenty possible values. The total number of pipelines to construct and evaluate is equal to 10^{10}, rendering this brute-force approach prohibitive when faced with a large, multidimensional search space.

There is a demand for an optimisation protocol that can efficiently and intelligently search across a high dimensional space. One approach [4] utilised a GA by which generations of preprocessing sequences were allowed to ‘evolve’ in a

manner analogous to Darwinian evolution. The optimised pipeline resulted in a 16% reduction in the model error compared with the raw data model. GA does not scale well with complexity as each generation is dependent on the previous. Therefore improvements to the runtime using parallelisation are not possible. A trial-and-error approach was more recently reported [6] which trialled every permutation of preprocessing steps within a defined search space on an ATR-FTIR biofluid dataset of brain cancer patients before classification using either random forest (RF) or RF/GA fed support vector machines (SVM). This brute force approach is comprehensive but can also lead to prohibitive runtimes. Furthermore, the hyperparameters of the final estimators were not optimised in this approach. This work proposes a novel approach to objectively optimise a preprocessing and classification pipeline. We combine parallel processing with Bayesian hyperparameter search to search a large parameter space efficiently. An FTIR-imaging dataset comprising several patients and over 100,000 spectra is used to test the framework on an existing problem.

4.2.2 Bayesian hyperparameter Search

To efficiently search for optimal pipeline configurations, it was necessary to perform a Bayesian hyperparameter search due to the computational expense of evaluating each pipeline. An open-source python library *scikit-optimize* [11] was used to leverage a Gaussian process (GP) regression over the parameter space associated with each job.

A GP is an efficient method of searching for maxima or minima over a complicated function. The GP model is utilised to approximate the loss function with limited data. The loss function is the score obtained when a proposed set of hyperparameters is used. The search for an optimal set of hyperparameters is summarised in Equation (4.1).

$$\boldsymbol{\theta}^* = \underset{\theta}{\operatorname{argmin}}\, f(\boldsymbol{\theta}) \tag{4.1}$$

Where $\boldsymbol{\theta}^*$ is the optimal hyperparameter configuration, and $f(\boldsymbol{\theta})$ corresponds to the process of training and evaluation of the pipeline in question using the hyperparameter vector $\boldsymbol{\theta}$. This framework is well-suited to the problem, which can be described as a sequential model-based optimisation task. The loss function is estimated sequentially, using previous evaluations to determine the optimal hyperparameter configuration to evaluate next. To evaluate the next point, which will result in the most significant improvement in the loss function given all previous evaluations and current estimates of the space; the expected improvement criterion is used [12]:

$$EI(\boldsymbol{\theta}) = (\mu(\boldsymbol{\theta}) - f(\widehat{\boldsymbol{\theta}}))\Phi(Z) + \sigma(\boldsymbol{\theta})\phi(Z) \tag{4.2}$$

$$Z = \frac{\mu(\boldsymbol{\theta}) - f(\widehat{\boldsymbol{\theta}})}{\sigma(\boldsymbol{\theta})} \tag{4.3}$$

Where $\Phi(z)$, and $\phi(z)$, are the cumulative density function and probability density function of a multivariate normal distribution.

4.2.3 Gaussian Processes

A GP regression seeks to estimate a distribution over an infinite set of candidate functions over a noisy loss function [12]. A GP frames the problem in such a way that each point in the optimisation space is considered to be a dimension in a multivariate Gaussian, described by a mean function Equation (4.4) and covariance matrix Equation (4.5):

$$m(x) = \mathbb{E}[f(x)] \tag{4.4}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{4.5}$$

Where a GP is defined as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{4.6}$$

The mean function approximates the underlying hyperspace, which we wish to estimate. Whereas the covariance function quantifies the relationship between the points in this space. The covariance function $k(x, x')$ can be represented by several different functions and can be used to instil prior knowledge of the relationship between data points. A commonly used covariance function is the *Matern* kernel:

$$k_{Matern}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{v} K_{\nu} \left(\frac{\sqrt{2\nu} r}{l} \right) \tag{4.7}$$

Where

$$r = x - x' \tag{4.8}$$

Where K_{ν} is a modified Bessel function of order ν and l is the parameter governing the length of the relationship between data points. The Matern kernel can cope with a noisy loss function and is the default kernel for the GP optimisation in skopt's BayesSearchCV. A GP is generally considered to be a non-parametric method. However, it is conceptually helpful to consider a GP as having an *infinite* number of parameters. This is due to the fact that a GP

instead seeks to estimate the posterior distribution over an *infinite* number of potential functions. This can be contrasted with a typical random variable as instead of drawing a scalar value from the corresponding distribution, a *function* is drawn instead. The function is drawn from a Gaussian distribution of mean Equation (4.4) and covariance Equation (4.5).

Toy example

To demonstrate the sequential optimisation procedure described above and reenforce the intuition behind a GP, a *toy* example has been constructed, a visual representation is shown in Figure 4.2. Shown in plot (a) is the true underlying function to be approximated. Plots (b) and (c) show the mean and standard deviation at each point respectively. Plot (d) shows the expected improvement at each point in the space given by Equation (4.2).

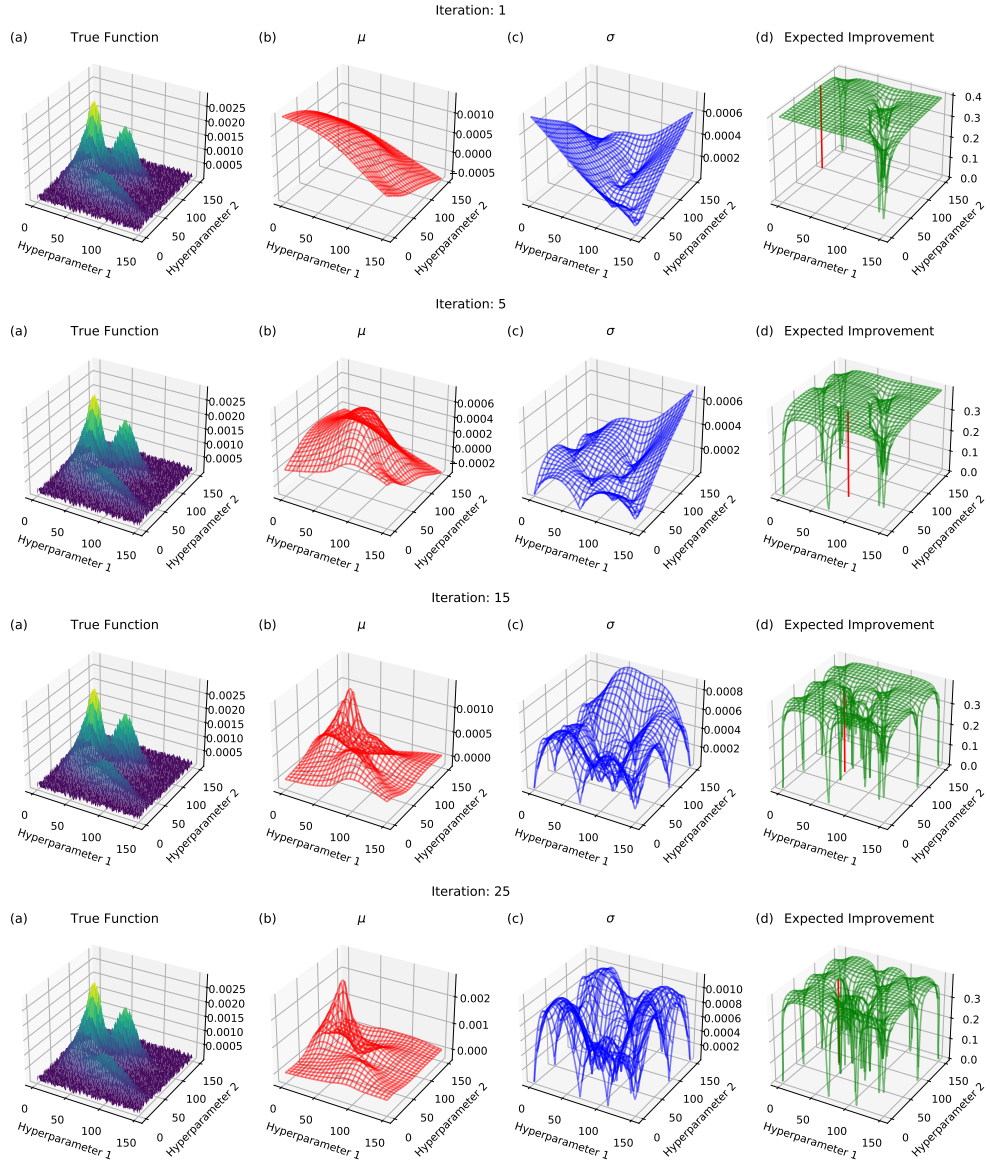


FIGURE 4.2: Bayesian hyperparameter search using a GP

The optimiser is initialised with 5 random points to assist with the convergence of the algorithm. These initial points are then used to fit a GP regression at the first step. The GP regression (Figure 4.3) then predicts the mean function (b) and standard deviation (c) at each point in the search space given. The expected improvement is then calculated using the observed data at that step. The maximum expected improvement is then taken as the point which will next be sampled. This next point is then taken with all previous data, and the process is repeated for the desired number of iterations. The hyperparameter

configuration chosen is then taken to be that which resulted in the most optimal score from the sequence of previous evaluations.

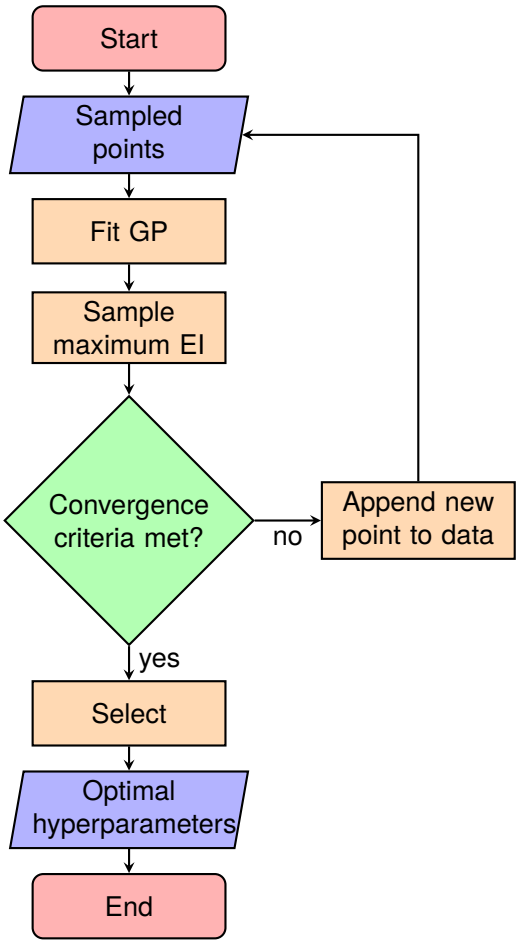


FIGURE 4.3: GP hyperparameter optimisation flowchart showing the overall process.

The number of iterations can be dictated by several criteria such as the convergence of the loss score, a preferred number of iterations, or a set amount of time. The convergence criteria are generally subject to the constraints of the available computing resources, but if resources are plentiful, convergence criteria are usually used.

Figure 4.4 shows a comparison between the mean function represented in Figure 4.2 (a), and the true function we wish to approximate (b).

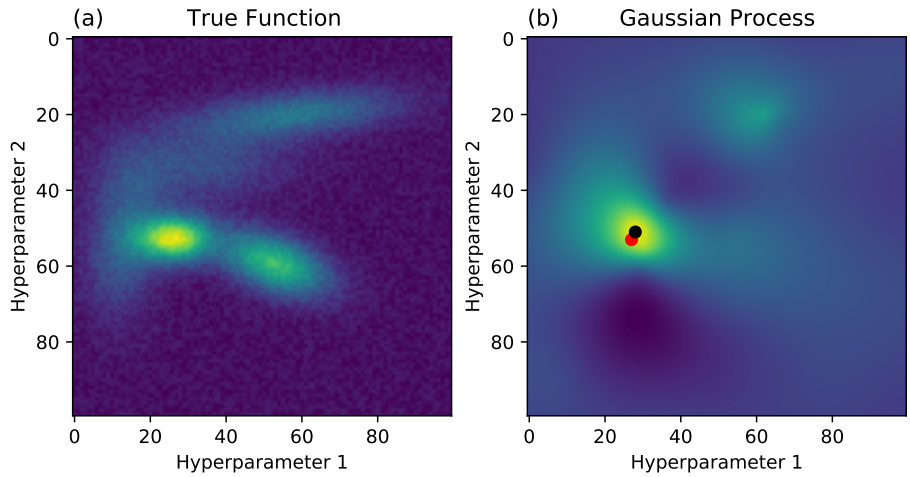


FIGURE 4.4: A comparison of true function to GP regression approximation. The optima predicted by the GP regression (red dot) is very close to that of the true function (black dot).

Shown in red and black dots are the global optimum of the true function, and the maximum of the GP mean function. Whilst a GP can approximate a high-dimensional loss surface well, it only serves to *direct* the search process. The optimal hyperparameter set chosen is selected as the set which was *known* to yield a low loss score through evaluation. This avoids using the somewhat speculative maximum given by the GP mean function, which can be particularly important when considering very high dimensional search spaces where the number of points sampled is relatively low compared to the entire volume of the space.

4.3 Methods

To utilise the framework, first the methods and hyperparameter search spaces they wish to trial must be input. Also defined in this stage is the order in which the steps are applied, with an estimator always occupying the final step. At

this point, it is instructive to define the validation procedure to use in the scoring stage. Protocols such as k-fold cross-validation and leave-one-out cross-validation are preferential over a single train-test split to mitigate the risk of overfitting. A completely independent set of data is put aside to test the final, optimised model. Each possible combination of methods is generated and distributed to a network of computers. The initial release of the framework is optimised for those utilising the well established HTCondor service [13]. HTCondor is an open-source high throughput package that enables the user to distribute parallelisable computationally expensive tasks (jobs) to a pool of idle computers on a local network, a method known as 'cycle-scavenging'. A flowchart summary of the process is shown in Figure 4.5.

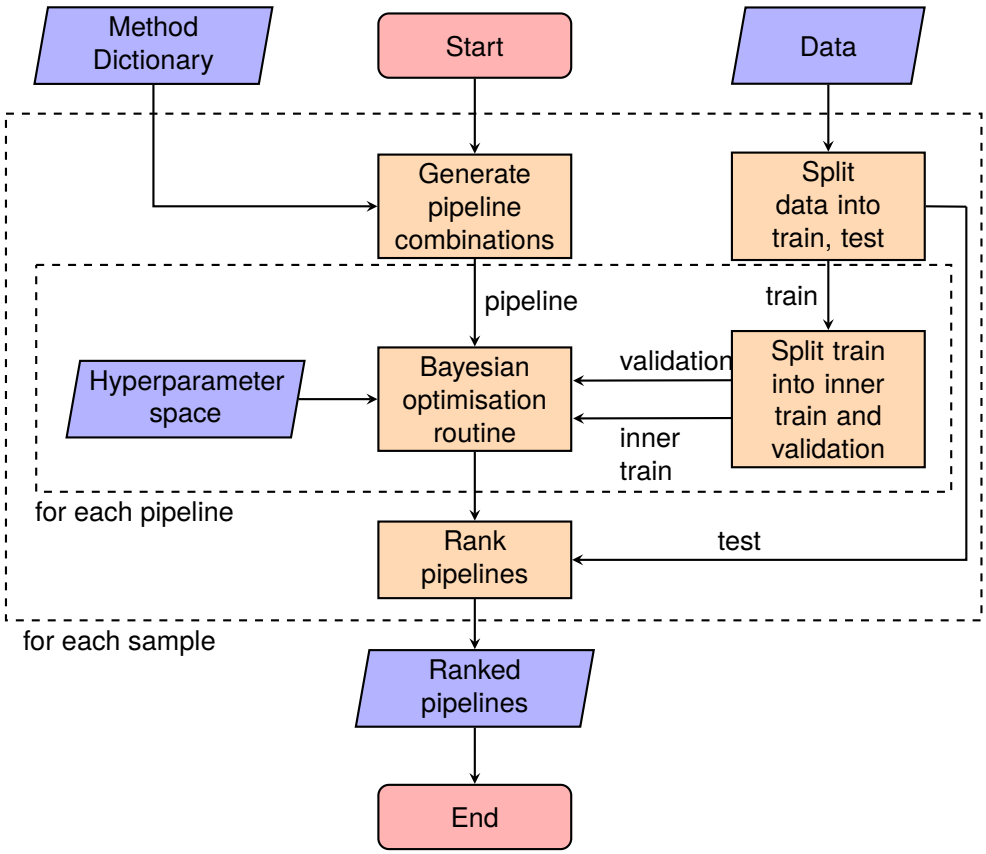


FIGURE 4.5: Flowchart of overall optimisation process

The number of jobs is directly related to the number of methods n belonging to each step i by the following:

$$n_{pipelines} = \prod_{i=0}^{n_{parameters}} n_i \tag{4.9}$$

Each distributed job is a unique combination of preprocessing transformers and final estimator. Contained within the job configuration is the set of search spaces associated with the hyperparameters, which can be initialised as one of three different types:

- **Categorical** — hyperparameter values can be any from an unordered list.
- **Continuous space** — hyperparameters are drawn from a defined probability distribution.
- **Integer space** — hyperparameters are discretised values from a given range.

Different sampling distributions can be specified when viable hyperparameter values span a wide range. Specifying sampling distributions for hyperparameters allows the user to provide prior knowledge of feasible values; or avoided entirely by specifying a uniform distribution.

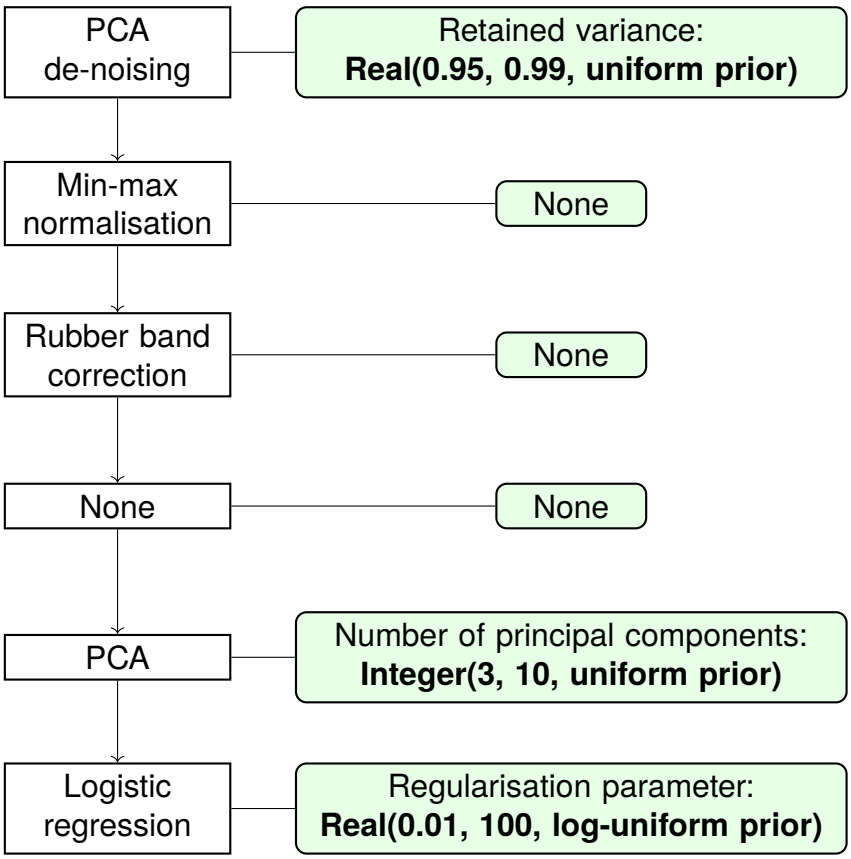


FIGURE 4.6: An example pipeline showing each step with associated hyperparameter search arguments (green).

As an example, consider a processing pipeline consisting of PCA-de-noising, followed by min-max normalisation, PCA, and ending with a logistic regression classifier. Three steps have hyperparameters that require tuning: the explained variance of the PCA de-noising method, the number of principal components retained by the PCA transform, and the regularisation parameter in logistic regression. A reasonable range of values to optimise over for the explained variance would be 0.95-0.99 (95-99%) to eliminate low variance noise; the space should be initialised as **Real(0.95, 0.99, uniform prior)**. The regularisation parameter value of the logistic regression classifier is inversely proportional to the regularisation strength — smaller values correspond to a stronger regularisation effect and mitigate the potential for overfitting. Optimal values exist on a much wider domain, nominally taking up values between 10^{-2} and 10^{2}, therefore the search space argument is initialised as **Real(0.01, 100, log-uniform**

prior) sampling from a log-uniform distribution.

The initialised job now begins the Bayesian search regime, through which the hyperparameter space is sequentially sampled, updating the loss function at each iteration and informing the subsequent search choices taken as the maximum expected improvement in the loss score Equation (4.2). The number of iterations, loss function, and validation protocol, are all pre-defined parameters which can be selected based on the size of the search space and type of optimisation problem. In the case of a classification task with a large parameter space, a large number of iterations will likely be needed. Once a set of hyperparameters has been found the fine tuned pipeline is validated on an unseen dataset to prevent information leakage and thus overly optimistic results. The completed jobs are then aggregated and ranked according to the mean validation AUROC score.

4.4 Results

To test the framework, an evaluation was performed on the same FTIR dataset described in 3. The objective was to obtain the optimised pipeline with the best mean performance across several train-test splits. The task was to predict whether a patient would live beyond or less than one year of the most recent review date.

Using a test set of preprocessing methods 576 unique pipelines were evaluated using the optimisation routine. Approximately one third of patients were held out for final model evaluation, with the remaining patients used for model training and optimisation. The loss function was the aggregated mean AUROC score of three patient-stratified folds; the optimiser iteration limit was set to 50.

The jobs were distributed to the HTCondor framework at the University of Liverpool, UK. Each of the 1900 PCs on the network is equipped with an Intel Core i3 (quad-core) processor running at 3.3 GHz, 8 GB RAM and 120 GB storage. Completed jobs were extracted from HTCondor and the results for each permutation across the 50 train-test splits were aggregated to compare average results. Pipelines were ranked according to the mean AUROC score across the 50 iterations; classification scores for these pipelines are shown in Figure 4.7.

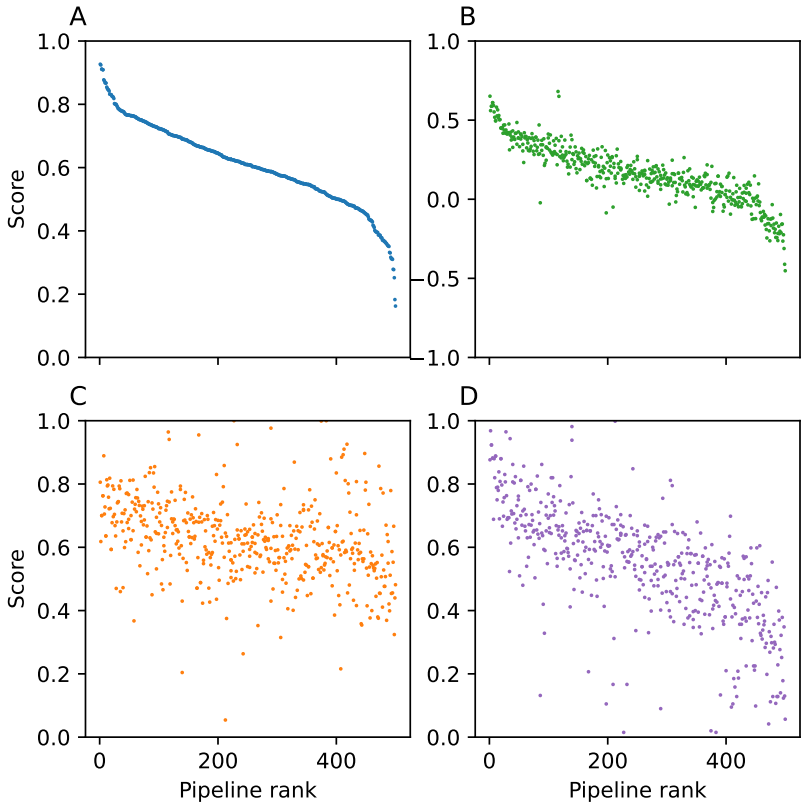


FIGURE 4.7: Classification statistics for the top 50 pipelines ranked according to AUROC score; AUROC (A), MCC (B), Specificity (C), Sensitivity (D).

The top 5 ranked pipelines discerned from the optimisation procedure are summarised in Table 4.1 and Table 4.2.

TABLE 4.1: Best performing pipelines with optimal processing steps and number of parameters n_θ.

Rank	Spectral smoothing	Baseline correction	Normalisation	Feature scaling	Feature extraction	Classifier	n_θ
1	N/A	N/A	Amide I	Robust	N/A	LR	1
2	SG	N/A	Min-Max	Standard	PCA	LR	3
3	N/A	N/A	Vector	Robust	N/A	LR	1
4	N/A	N/A	Amide I	Robust	N/A	LR	1
5	N/A	RB	Vector	Standard	PCA	LR	2

TABLE 4.2: Top ranking pipeline classification scores as decimals.

Rank	TN	FP	FN	TP	AUROC
1	0.51	0.16	0.11	0.22	0.63 ± 0.02
2	0.47	0.20	0.12	0.21	0.62 ± 0.02
3	0.44	0.22	0.10	0.24	0.61 ± 0.02
4	0.45	0.22	0.09	0.25	0.61 ± 0.02
5	0.42	0.25	0.11	0.22	0.61 ± 0.02

Table 4.1 summarises the specific methods used in each of the top 5 ranked pipelines.

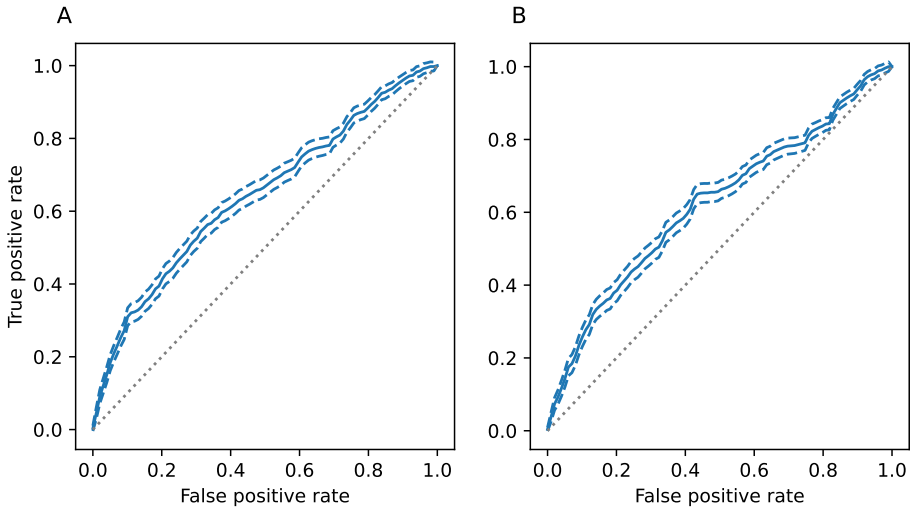


FIGURE 4.8: ROC curves shown with standard errors for best (A) and second-best (B) pipelines.

Figure 4.9 shows the loss functions sampled during the hyperparameter search for pipeline two. The loss surface is shows a strong dependence upon the fraction of components used in the feature extraction step. This contrasts

with the relatively low dependence upon the regularisation parameter associated with the logistic regression classifier. This is likely due to the fact that both parameters play a regularising role in the inference procedure so as to avoid overfitting. If both steps were to have parameters indicating a high regularisation effect, this would be detrimental to the classification performance so feature extraction seems to be preferred.

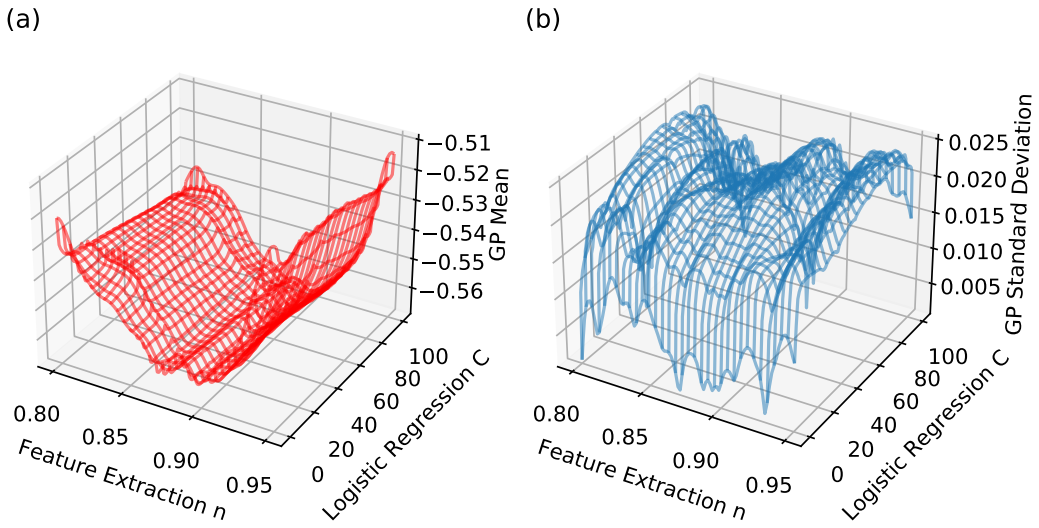


FIGURE 4.9: GP hyperparameter surfaces showing mean function in red and standard deviations in blue averaged across 50 sample iterations.

Optimised pipelines from each of the 50 train test splits each have a unique set of hyperparameters; these parameters are ”tuned” to specific training and validation data sets during the training phase. To determine a more generally applicable set of parameters the mode of the distribution of values was taken. Figure 4.10 shows histograms of each of the selected hyperparameters for the top two pipelines over 50 samples.

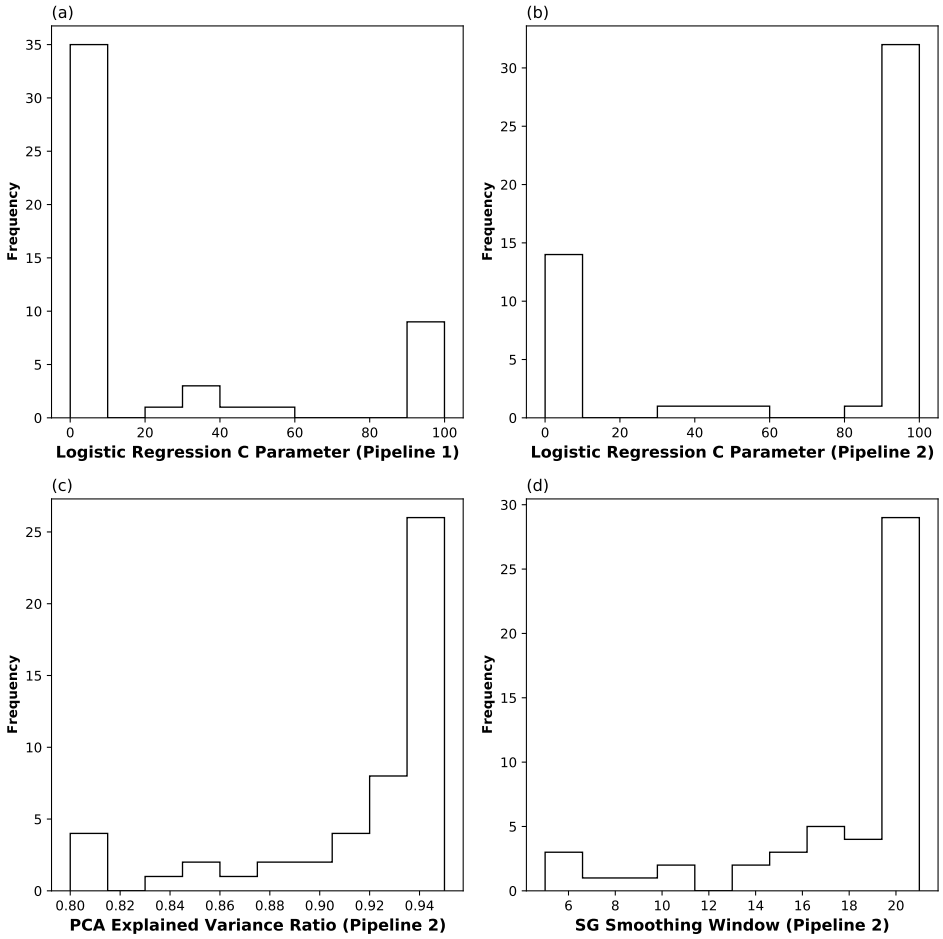


FIGURE 4.10: Histograms of optimum hyperparameters over the 50 train-test splits.

To acquire a complete measure of the performance of the optimised pipelines, the model is sequentially tested 50 times by randomly drawing train-test splits without replacement. Modal values from Figure 4.10 are used as final model hyperparameter values. Aggregated statistics for pipelines 1 and 2 are shown in Figure 4.11.

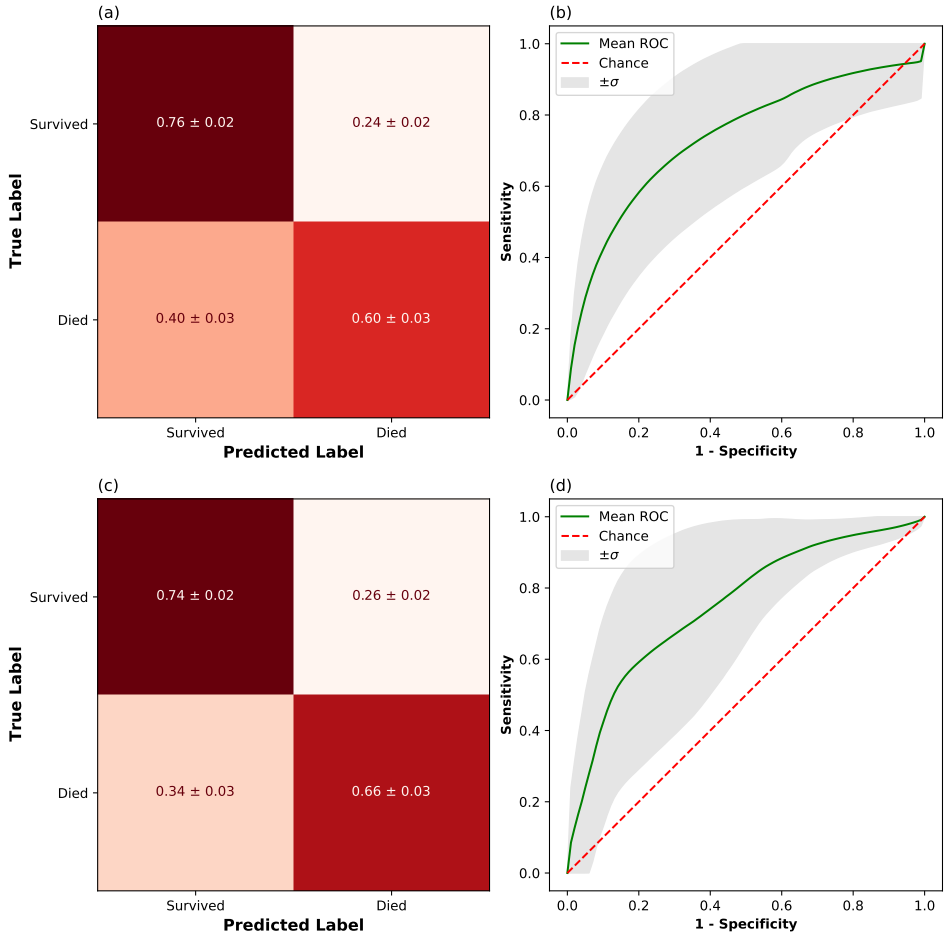


FIGURE 4.11: Mean confusion matrix and ROC curve shown with standard errors for best (a,b) and second best (c,d) pipelines trained and tested on full dataset.

Figure 4.11 shows a significant increase in both sensitivity and specificity when the optimised pipelines and hyperparameters are deployed on the full dataset. A lower specificity translates to more false positives indicating the model struggles to identify patients with a poor prognosis.

4.5 Discussion

classification scores vary widely but generally exhibit good classification scores well above random chance; shown in Figure 4.7. The more holistic measures

of AUROC and MCC follow a similar trend, whereas the relatively noisy sensitivity and specificity traces imply that there is often a trade off between the two metrics, where high sensitivity often leads to low specificity. Ranking metrics by AUROC or MCC favours pipelines with balanced sensitivity and specificity scores. The trace in Figure 4.7(A) shows a number of small, relatively high scoring pipelines, before gradually decreasing towards an AUROC of around 0.4, and an MCC of 0.0. The steep increase in AUROC and MCC scores towards the higher end imply the expense of the optimisation procedure is justified.

Table 4.1 indicates that the optimal classifier for this dataset is logistic regression, with various choices of preprocessing options preceding this step. Normalisation and scaling are never bypassed, suggesting this is an imperative step. Two instances in the top five classifiers utilise PCA to reduce dimensionality, suggesting this step is not such an important for this dataset paired with logistic regression. Similarly, spectral smoothing by Savitzy-Golay filtering appears in the second pipeline, but is absent for the top ranking and remaining pipelines in the top five.

In order to investigate the effects of different methods on the performance of the pipeline, the frequency that a certain method either enhances or diminishes performances relative to a reference can be informative. Here, the reference score is the median score of all pipelines in the analysis.

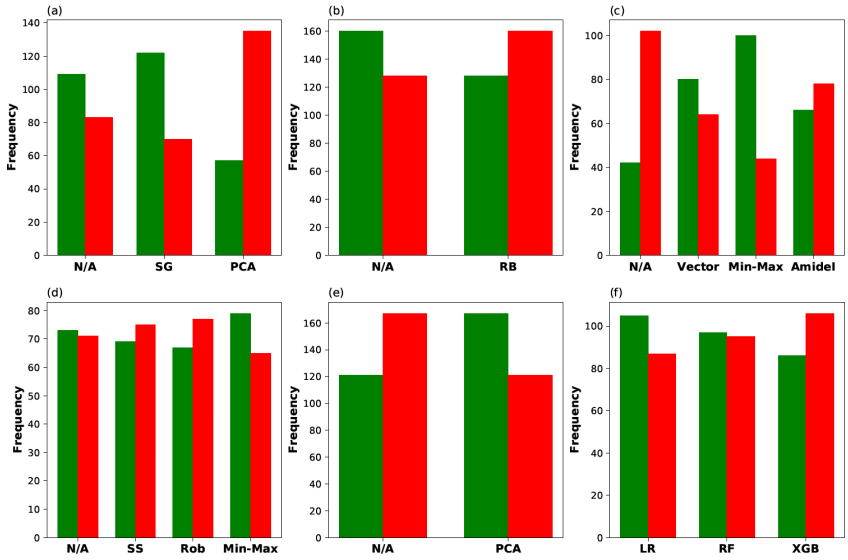


FIGURE 4.12: Frequency each method either enhances (green) or diminishes (red) relative to the median score (AUROC = 0.48). Steps are (a) smoothing, (b) baseline, (c) normalisation, (d) scaling, (e) feature-extraction, (f) classifier.

Figure 4.12 shows some interesting insights of the effects of various methods. The choice of smoothing method has a significant effect; the majority of pipelines that utilise PCA de-noising perform worse than the median, whilst Savitzy-Golay smoothing predominantly increases scores. It could be argued that baseline correction has an insignificant effect, perhaps slightly detrimental, this could be attributed to the data already being subject to a previous scatter correction prior to the analysis, negating the requirement to perform a baseline correction. Normalisation is a step that can not be bypassed, an expected result as spectra originate from different samples, each with differences in sample thickness. Mitigating the effects of sample thickness has a positive impact on classification scores.

It appears that min-max normalisation occurs most frequently in the higher-performing pipelines. Scaling of the data appears to have a significant effect on the performance of the pipeline, but the choice of scaling does not seem to be hugely important. It should be reiterated that the top five pipelines in Table 4.1

employ a scaling method to the data in the pipeline, suggesting that re-scaling each wavenumber variable is beneficial for logistic regression. It would also appear that application of PCA to decompose the data prior to classification is slightly more beneficial than not. As previously stated, logistic regression emerges as a favourable classifier to the tree based random forest and gradient boosted classifiers, implying that a simpler, linear based model is preferred to complexity, perhaps as complex models are more prone to overfitting and have a much larger hyperparamater space to optimise. In fact, the dramatic drop off at approximately AUROC = 0.40 is the result of pipelines with an XGBoost classifier, which has a large hyperparameter space that requires fine tuning. It may be the case that more iterations within the Bayesian hyperparameter search would produce more favourable results for the tree based models such as RF and XGBoost, but this would increase the time taken for the optimisation to execute.

The histograms in Figure 4.10 show distributions of hyperparameters for each of the two top performing pipelines. Interestingly, it reveals that the logistic regression regularisation value in pipeline one Figure 4.10(A) converges to a much lower value (~ 0.01) than for pipeline two, where it appears to converge towards 100. Pipeline two applies smoothing and feature extraction in addition to normalisation and scaling, which themselves have a regularisation effect on the subsequently fitted models. This may be the reason as to why the ultimate C parameter of logistic regression needn't be as low as 0.01 for pipeline two.

Taking the modal hyperparameter selections from Figure 4.10 and training and testing on all available data (using the same patients for each of the 50 train-test splits) enhances the scores significantly, as shown by the mean confusion matrices and ROC curves in Figure 4.11. There is a 14% increase in mean specificity for pipeline one and a 3% increase in mean sensitivity. Pipeline two exhibits an 11% increase in specificity and 9% increase in sensitivity. This

would suggest that the strategy of sampling equally small subsets of data from each patient for the purposes of efficiency and stratification is sound, and translates well to a more realistic scenario where all the available data from different patients would be used instead.

4.6 Conclusion

The work presented here demonstrates a versatile framework capable of determining a near-optimal data preprocessing and classification pipeline. This optimisation framework has been employed on a real inference problem and has successfully demonstrated that this process can be performed objectively and without specific prior knowledge of optimal parameters. The framework's performance has been tested across a range of sample datasets. It has shown that effective configurations can be determined through a rigorous analysis, as proven by validation on held-out data. The choices of preprocessing methods resulting in pipelines with the highest ranks seem to follow conventional logic — normalisation is necessary, Savitzky-Golay smoothing is beneficial, PCA is advantageous depending on the choice of the classifier. Valuable insights have been gained from the procedure, showing that some preprocessing methods are particularly beneficial compared to others.

To gain further knowledge of effective preprocessing methods, it would be useful to perform the optimisation procedure in a broader variety of datasets. This could yield insights into which preprocessing steps are effective or given classes of inference problems.

This framework could be utilised by other researchers to perform a similar process for a given problem and set of preprocessing steps. It is by no means limited to FTIR spectroscopy and could be extended to other inference problems with minimal adjustment.

Bibliography

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

[2] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[3] David H. Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390, 10 1996.

[4] Roger M. Jarvis and Royston Goodacre. Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics*, 21(7):860–868, 2005.

[5] Elon Correa and Royston Goodacre. A genetic algorithm-Bayesian network approach for the analysis of metabolomics and spectroscopic data: Application to the rapid identification of Bacillus spores and classification of Bacillus species. *BMC Bioinformatics*, 12, 2011.

[6] Holly J Butler, Benjamin R Smith, Robby Fritzsch, Pretheepan Radhakrishnan, David Palmer, and Matthew J Baker. Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy. *The Analyst*, 2018.

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

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

[9] R O Duda, P E Hart, and D G Stork. Pattern classiïňĄcation. *New York: John Wiley, Section*, 2001.

[10] Isabelle Guyon and Andre Elisseeff. Feature Extraction, Foundations and Applications: An introduction to feature extraction. *Studies in Fuzziness and Soft Computing*, 2006.

[11] Tim Head, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi. scikit-optimize/scikit-optimize. 9 2020.

[12] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian process for machine learning*. The MIT Press, 2006.

[13] Douglas Thain, Todd Tannenbaum, and Miron Livny. Distributed computing in practice: The Condor experience, 2005.

5 Deep Learning Prognostic Tools

5.1 Introduction

In recent years research a subcategory of machine learning called 'deep learning' has rapidly emerged to become the state-of-the-art in field of artificial intelligence. Deep learning is able to achieve superior performance where previous approaches have fallen short, and has facilitated the development of many useful applications [1]. Deep learning is a broad term covering a large swathe of statistical models all comprising a multi-layered ANN of some form. The recent surge in interest in ANNs has been driven by a number of factors. In the last twenty years computing power, storage, and the availability of data have increased dramatically [2]. This has facilitated a shift in the way that artificial intelligence systems are created; modern algorithms have been developed which learn from data very efficiently, rather than being explicitly programmed by a human to accomplish a given task. Due to the highly parameteric nature of deep ANNs, large quantities of data are required to obtain optimal model configurations, this data-focused approach is responsible in large part for the recent success of ANNs [2].

To evaluate the potential of CNNs as a prognostic model for oral cancer, a one-dimensional CNN was created to establish if any improvement was gained by adding spectral information. The network follows a similar structure to a 2D CNN; the objective is that the network would extract higher level features

from raw absorbance values — potentially correlating to levels of known biochemicals. A crucial factor in the transmission of FTIR microscopy into a clinical setting is its ability as a technique to be universally adopted. To do this factors including sample preparation, measurement environment, and measurement technique must be as uniform as possible. Preprocessing data is an attempt to mitigate the effects of potential inconsistencies in measurement practice but the process is never perfect. It would be desirable for any analysis method to be able to obviate the need for preprocessing all together by being robust to invariances introduced by experimental practice alone. A CNN is able to manage this to some extent due to it expressing translational invariance — attributable to the convolution [3], and pooling layers [4, 5]. By effectively scanning the entire spectrum and learning how to recognise patterns spanning multiple wavenumbers, the model is robust to slight alterations in wavelength-dependent absorbance.

Convolutional neural networks

As outlined in 2.3.4 an ANN consists of a number of layers, each of which comprise a number of nodes. Nodes in an ANN are a representation of a relatively simple equation known as the perceptron equation (Figure 2.23); which are combined in complex ways and used as a highly parametric model of a particular inference problem. A neural network trained using labelled data will optimise free parameters within the network to minimise a loss function constructed for the problem at hand. In particular, neural networks have allowed for advances in applications where data contains temporal or spatial information. Convolutional neural networks (CNNs) are a type of network containing specialised layers capable of extracting spatial information from data. This is accomplished with the use of a kernel which is convolved over the input data. A kernel comprises a number of parameters which are refined during a training

phase to capture the most relevant spatial information. The first few layers of a CNN are utilised as a method of feature extraction; values of the convolution kernels are refined progressively to extract useful spatial features in the data; these features are then fed into a standard MLP network for further feature extraction, and later classification or regression.

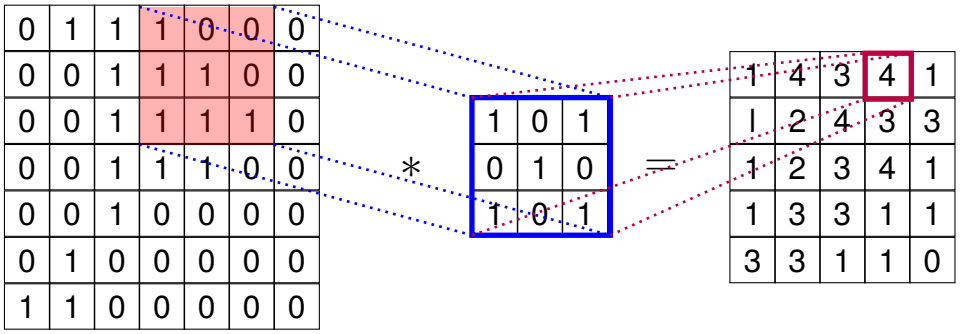


FIGURE 5.1: A 2D convolution layer showing a simple (3×3) kernel K convolved over an input image I of size (m×n). The resulting convolution $I * K$ is effectively a spatial map of where in I most closely resembles K.

This convolution operation can be expressed formally as:

$$(I * K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, j - n) \tag{5.1}$$

The kernel K is moved across the dimensions of the data where it is multiplied by the values in the kernel. A mapping of the similarity of the data to the kernel at that point is obtained then aggregated, typically a maximum or mean is taken and the values are pooled and assigned to an output. The behaviour of a kernel layer is dictated by its size, stride — the number of elements the kernel shifts per iteration, and dilation — the mapping of kernel elements to non contiguous elements of the data. Like other layer types convolutional layers can be altered by many parameters but will not be discussed further here. Depending on the dimensionality of the data the kernel can be combined with other kernels

to find correlations between them; allowing colour information in 2D images to be used. A bias term is added to the pooled value and subjected to an activation function like in a normal perceptron layer. The next layer is obtained by convolving over the preceding layer using another kernel, this continues for a number of layers. The input is flattened into a one-dimensional vector where it is passed into a normal MLP where it reaches a softmax layer output. This softmax layer Equation (5.2) outputs scores which sum to one and can be loosely interpreted as probabilistic predictions of a given class.

$$softmax_k(x) = \frac{e^{W_k^T x}}{\sum_{i=1}^{n} e^{W_i^T x}} \tag{5.2}$$

Intermediate steps between these layers can be introduced to assist with regularisation such as dropout layers, batch normalisation layers [6], and many others. Dropout layers [7] are often included in ANN architectures as they provide a strong regularisation effect during the training phase. A simplified representation of a CNN showing the sequential nature of the described layers is shown in Figure 5.2. Dropout layers are typically implemented in a similar way to regular hidden layers, however they differ in that when any forward pass occurs in the training stage a node may become inactive, preventing any change in the weights of connected nodes for that particular pass. Each dropout layer typically has a probability associated with it which dictates the chance of becoming inactive. Dropout works effectively by discouraging weights from converging towards similar values — encouraging redundancy in the network structure.

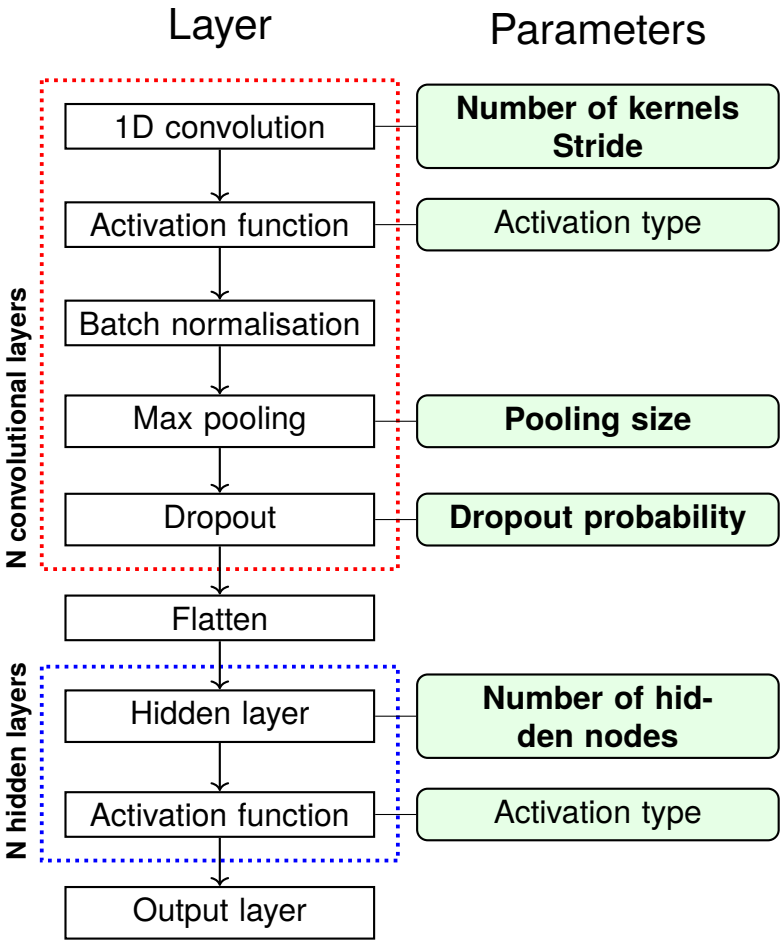


FIGURE 5.2: An typical convolutional network example pipeline; parameters associated with each step are shown in green; extra parameters are shown in bold.

Batch normalisation layers were included in the CNN model to increase the speed and efficiency of the training procedure. Batch normalisation layers work by normalising the distribution of values flowing from one set of nodes to the next, by scaling to a given range. This helps to prevent issues associated with exploding/vanishing gradients whereby update values for nodes increase or decrease rapidly to the detriment of the training process.

Furthermore, many parameters are associated with aspects of the training procedure of the network itself. The objective of the optimisation procedure used to train neural networks is to maximise or minimise a score with respect to the network parameters. Number of optimisation techniques are used but

stochastic gradient descent (SGD) was the chosen method for both ANNs described here. SGD works by computing this objective score on a randomly selected subset of the training data. This is beneficial for large datasets where computing on the entire set may be computationally infeasible. When an update occurs a coefficient called the *learning rate* is used to dictate the weighting of the new value. Another parameter used to influence the optimisation strategy is the *weight decay*; a coefficient used to alter the effect of the gradient value on the objective function. A thorough description of SGD and its associated parameters is available from [8]. Another commonly used technique to improve convergence during the training stage is to initialise neural network weights. Weights were intialised using the method described by Kaiming [9]; initialising layer weights has been shown to decrease convergence times and improve the stability of the optimisation procedure.

5.2 Materials and Methods

To evaluate the benefit of utilising convolution layers to analyse spectra, a comparison of a CNN model was made with a MLP. Both ANN models were constructed, trained, and evaluated using an open-source python libary PyTorch [10]. Additional packages [11, 12, 13] were leveraged to implement the evaluation procedure along side other common machine learning operations. The same procedure as discussed in Chapter 3 was followed; out of bag sampling was utilised to obtain distributions of classification values, inverse weighting was used to mitigate the effects of dataset imbalance.

The dataset comprised FTIR spectra taken from primary tumour sites of 29 patients with a diagnosis of OSCC. Inclusion criteria for this study were as previously described in Chapter 3: a diagnosis of OSCC; the presence of OSCC in

the TMA core; the ability to co-register adjacent H&E stained and FTIR imaged sections; a follow-up period after surgery of at least 24 months; HPV negative.

Images were acquired at a resolution of 6cm^{-1} over a spectral range of 990cm^{-1} to 3800cm^{-1} using a co-addition of 128 scans. Attenuator and integration time of the focal plane array (FPA) were chosen to gain the maximum signal-to-noise ratio. Background scans were acquired using a blank CaF_2 disk situated within the perspex box before each session of measurements.

The preprocessing steps required for each type of network are slightly different. Given that the convolutional layers in the CNN model are used to extract features from multiple wavenumbers simultaneously, the only preprocessing step is to normalise the data. For the MLP model vector normalisation was used in order to account for sample thickness; wavenumber absorbance features were mean-centered; and variance scaled to one; before a final PCA step to reduce dimensionality of the dataset.

5.2.1 Optimisation of network structure

The sheer number of tunable parameters associated with ANNs necessitates a hyperparameter search similar to that described in Chapter 4. The open source optimisation framework Optuna [14] was used to determine and optimal network structure and associated hyperparameters. The hyperparameters in bold in Figure 5.2 were chosen for optimisation for the CNN network; in addition the learning rate and weight decay parameter were included as optimal values are task-dependent and have a large impact on the training efficiency of ANNs [15, 16]. The median AUROC value was calculated across a five-fold cross validation of data subsets to determine the general suitability of the network configuration. Fifty sequential trials were chosen to allow sufficient exploration of the parameter space.

Convolutional network

A summary of the configuration determined by the procedure is given in Table 5.1

TABLE 5.1: Optimal convolutional neural network hyperparameters and values.

Parameter name	Value
N convolutional layers	5
N kernels in convolution layer 1	96
N kernels in convolution layer 2	128
N kernels in convolution layer 3	32
N kernels in convolution layer 4	80
N kernels in convolution layer 5	128
Maxpool 1 size	3
Maxpool 2 size	3
Maxpool 3 size	7
Maxpool 4 size	5
Maxpool 5 size	3
N fully connected nodes	80
Dropout probability	0.45
Learning rate	8×10^{-5}
Optimum AUROC value	0.84

A simplified diagram of the CNN network Figure 5.3 configuration determined by the optimisation procedure is shown in Figure 5.3. The CNN network contains five convolutional and maxpooling layers of varying sizes. Deeper network designs are typically better at extracting high-level structural information in data [17].

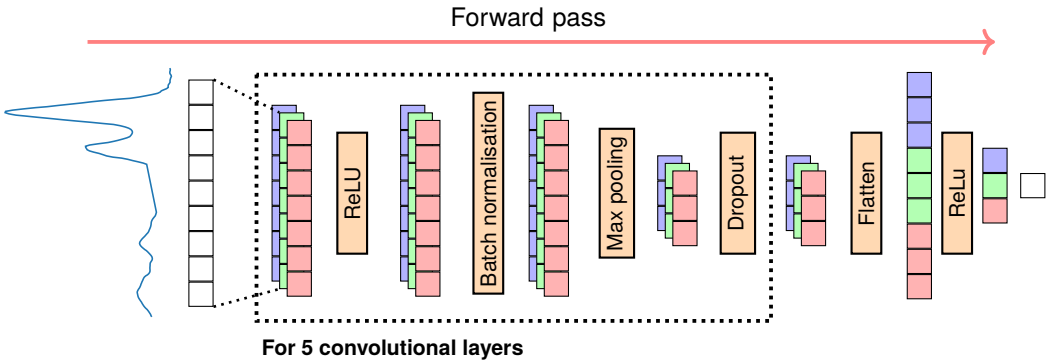


FIGURE 5.3: A simplified schematic of the optimal one-dimensional CNN architecture. The shape of the data as it passes through each layer is represented by vectors; the colour of each vector represents a different kernel. Intermediate layer activations, regularisation steps etc are represented by orange boxes. The final element represents the probability of a poor prognosis for that spectrum.

Multilayer perceptron network

A summary of the optimal network configuration for the MLP determined by the optimisation procedure is given in Table 5.2. A simplified representation of the optimal MLP network is shown in Figure 5.4. The configuration is a relatively shallow MLP with two layers. Both layers have a strong regularisation effect applied by dropout layers with dropout probailities $\sim$0.3. The first layer has 169 nodes followed by 10 in the second layer; a potential explanation for this is that there are fewer higher level features needed in the second layer to achieve a a good level of discrimination between risk groups.

TABLE 5.2: Optimal MLP network parameters.

Parameter name	Value
N hidden layers	2
N nodes in hidden layer 1	169
N nodes in hidden layer 2	10
Dropout layer 1 probability	0.28
Dropout layer 2 probability	0.29
Learning rate	2×10^{-5}
Weight decay	1.1×10^{-3}
Optimum AUROC value	0.77

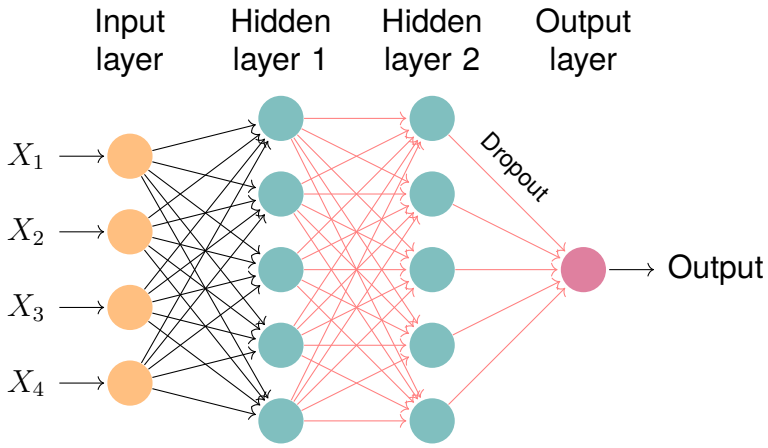


FIGURE 5.4: A multilayer perceptron neural network with an input layer consisting of four input variables $x_0 \dots x_4$, two hidden layers of five nodes each, and a single output layer. Dropout layers are represented by red arrows between network layers.

5.3 Results

Much of the existing literature concerning prognositic indicators utilises disease specific survival and overall survival as indicators of patient prognoses. Stratification of these measures into groups according to risk is typically on a set number of years e.g. one year, two years. The decision to use a cut off threshold of a discrete number of years is somewhat arbitrary, thus, it would be desirable to

determine a threshold that was decided objectively. A GA approach discussed in Chapter 3 was used to stratify patients into either a low or high-risk group. This threshold was determined to be 11 months and was used to dichotomise patients into risk groups which served as prediction objectives for the CNN and MLP models.

As dicussed in detail in Chapter 3, the objective was to predict which risk group a patient falls into, risk groups were determined by a GA optimisation routine seeking to achieve the maximum prognostic information. Predictions of risk groups for each patient were taken as the median probability predicted across all spectra for any given patient. The threshold used to dichotomise probabilistic predictions was set by maximising the MCC score (Table 2.2). The MCC score considers all possible prediction outcomes and is a well-rounded measure of performance for discrete predictions.

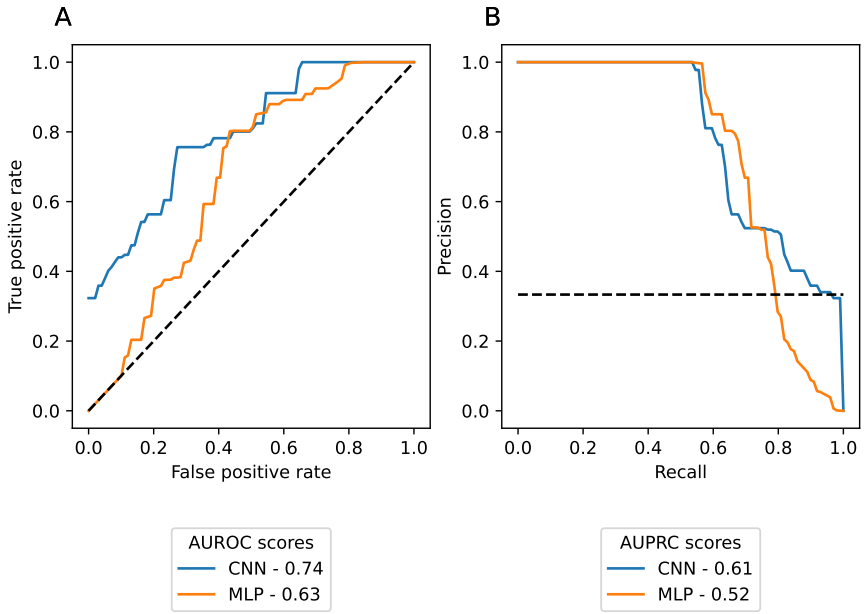


FIGURE 5.5: Median ROC and PR curves shown in solid lines; dashed lines represent baselines scores associated with random chance. AUROC and AUPRC scores are shown for each set of prognostic indicators.

Median ROC and PR curves (Figure 5.5) indicate that both models show

some utility as classifiers. The AUPRC scores for both classifiers are both significantly above the baseline score — Indicating that both models can balance both precision and recall simultaneously, and that imbalance in the dataset was not detrimental to classification scores (Figure 5.5[B]). The AUROC score is modest for the MLP model at 0.63, the CNN model performs better across all classification thresholds with a score of 0.74.

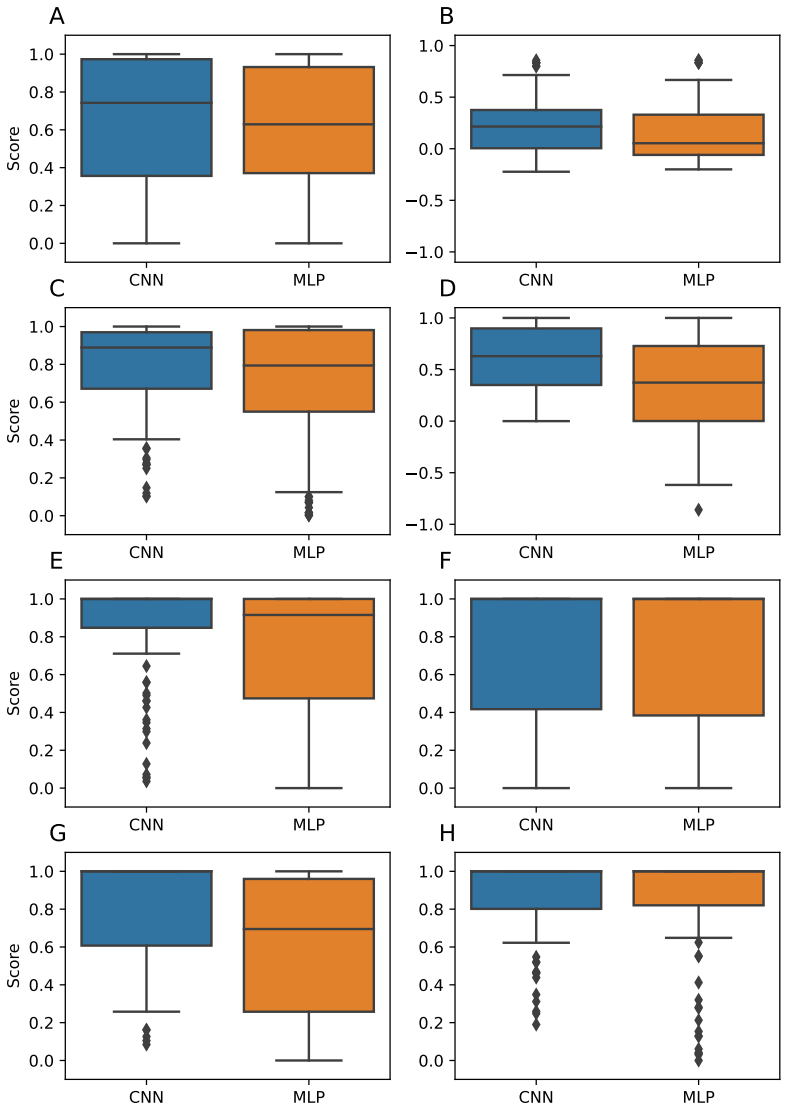


FIGURE 5.6: Whisker boxplots of classification statistics calculated across all data subsets. AUROC (A); AUPRC (B); F1 (C); MCC (D); specificity (E); sensitivity (F); PPV (G); NPV (H). Boxes show the median, 25^{th}, and 75^{th} percentiles; whiskers extend to points that lie within 1.5 inter quartile ranges of the lower and upper quartiles; points lying outside this range are shown as individual diamonds.

Score distributions Figure 5.6[A-H] for both models are generally good. The CNN model is superior for all statistics and generally varies less than the MLP model despite the low threshold value. The low threshold value for the CNN model indicates that the distribution of prediction scores spans a low range.

TABLE 5.3: Median classification statistics. Classification thresholds (Table 5.3) used to dichotomise prediction probabilities were determined to be those that maximised the MCC score.

Variables	AUROC	AUPRC	F1	MCC	Spec	Sens	PPV	NPV	Thresh
CNN	0.74	0.22	0.89	0.63	1.00	1.0	1.0	1.0	0.01
MLP	0.63	0.05	0.79	0.37	0.92	1.0	0.7	1.0	0.58

Kaplan-Meier curves were plotted for groups dichotomised by the threshold in Table 5.3 and show a clear distinction between groups determined by the CNN classifier Figure 5.7; a significant p-value was obtained by performing a log-rank test using the predicted groups. shows effectively no ability to discriminate between risk groups, which is in agreement with an insignificant p-value of 0.4.

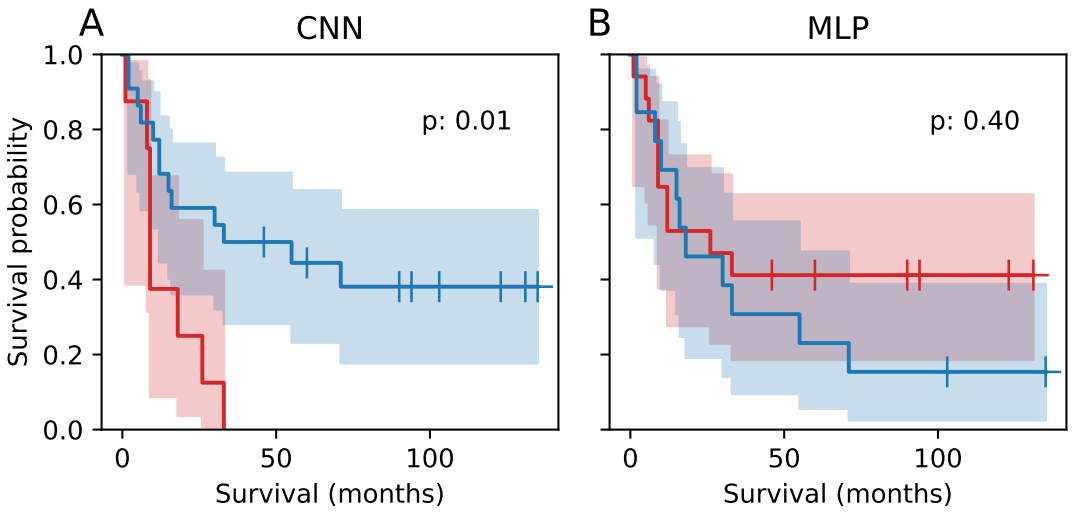


FIGURE 5.7: Kaplan-Meier survival curves of predicted risk groups for CNN [A] and MLP [B] classifiers.

5.4 Discussion

Deep learning methods are a popular choice for a variety of classification tasks in the field of medical diagnostics. Due to the increasing availability of data and computing resources. Deep learning methods coupled with vibrational spectroscopy data sets are rapidly finding use in cancer diagnostics [18]. The development of CNNs has enabled researchers to utilise structural information present in data to quantify spatially and temporally dependent features. FTIR spectra possess such structural information due to the overlap of absorbance bands associated with various chemical moieties present in a sample. The use of multiple convolution layers facilitates the extraction of high-level structural information present in data, determining patterns from multiple wavenumbers simultaneously in a similar way to what might be done by a human searching for known peaks in chemical spectra. Even slight shifts in absorption values induced by inconsistent measurement technique could result in the misclassification of spectra, negatively affecting the prospect of FTIR being widely adopted as a clinical tool.

The work described here is an attempt to establish the viability of CNNs coupled with FTIR spectroscopy as prognostic prediction tools. The objective is to correctly predict the prognosis of a patient from spectra measured from primary tumour sections taken from a TMA array. An initial optimisation procedure was performed to determine a suitable network architecture; the Bayesian optimisation procedure explored a large parameter space seeking to maximise the median AUROC of a five-fold cross validation routine. Using the optimal network configurations, both ANNs were evaluated using a sampling without replacement bootstrap strategy to obtain distributions of classification scores.

ROC and PR analyses were used to estimate the general utility of each ANN configuration. AUROC scores of 0.74 and 0.63 for the CNN and MLP networks

respectively show that both classifiers have some utility across all classification thresholds. AUPRC scores of 0.61 (CNN) and 0.52 (MLP) show that both models display some utility despite some imbalance in the dataset. Classification statistics Figure 5.6[A-G] are generally very strong for both models; a median MCC score of 0.63 for the CNN model indicates a very strong classifier, with the MLP achieving a median score of 0.37. The optimal threshold used to dichotomise the CNN model output was particularly low, this however was not detrimental to the classification performance of the model.

Survival curves for groups predicted by the CNN model Figure 5.7[A] show good separation between risk groups; a significant p-value of 0.01 given by a the log-rank test corroborates this. Interestingly despite the good overall classification performance of the MLP model, the threshold which maximised the classification scores resulted in poor stratification of risk groups. This is likely due to the fact that a high score for the log-rank statistic can be obtained by allocating only a small number of patients to the high-risk group. Conversely, to obtain strong classification scores, as many patients as possible must be allocated to the correct risk group.

MLPs have also been utilised with vibrational spectroscopy methods to detect breast cancer using ATR-FTIR data [19] and used three ANNs in a 10-fold cross validation scheme to discriminate between FTIR spectra collected from 78 malignant and 88 benign breast tumours. The authors found that ANNs had superior performance across many classification statistics in comparison to LR, RF, and LDA classifiers; with the ANN models achieving AUROC scores ~ 0.9. The ANN models however did not perform as well as a SVM classifier on the same data. The ANN models used most closely resembled the MLP classifier discussed here, however the authors did not include additional regularisation layers such as dropout layers.

A second study [20] has leveraged 2D CNNs to dicriminate between different grades of breast cancer from a cohort of 96 patients. The study followed a similar analysis routine to that discussed in Chapter 3, where a series of subsets of the data were drawn without replacement to obtain distributions of classification scores. It was found that the addition of spatial information from convolutional layers improved the model performance considerably over pixel-level predictions of a large range of models including: SVM, random forest, and an RBF kernel. Adding spatial information increased the overall accuracy of the predictions by $\sim$20%; the specificity and sensitivity increased considerably with one class improving by $\sim$60%.

A one-dimensional neural network was used to classify FTIR, Raman, and near infrared (NIR) data derived from a variety of food samples [3]. The ANN developed by the authors was a shallow CNN, which improved on overall accuracy scores of other classifier methods on preprocessed data from 62% to 86%; and from 89% to 96% on raw data. CNNs were used with success on ATR-FTIR data in forensics to detect synthetic cannabinoids [21]. The authors found that CNNs were capable of identify synthetic cannabinoids, achieving 98.7% accuracy and an F1 score of 98.5% — meeting the standards of a forensic screening system. CNNs applied to vibrational spectroscopy have enjoyed further success in forensic applications where they employed to identify amphetamines with an accuracy of over 90% [22].

A rigorous analysis procedure was performed to obtain distributions of classification scores, and to gain an insight into the general applicability of each model to unseen data. The relatively small sample set was a key issue facing this study due to the expense of acquiring and imaging large numbers of samples. A large degree of variation was observed across some classification statistics, potentially indicating a large degree of biological heterogeneity in the

dataset — a known characteristic of OSCC primary tumours [23, 24]. A potential cause for this could be the effect of inherent molecular heterogeneity of the tumour microenvironment [25]; or perhaps varying extents of lymphocyte infiltration present in specimens. The difficulty of annotating samples is also likely to introduce noise into the dataset; alongside inconsistencies in sample preparation and the measurement procedure and environment.

5.5 Conclusion

For FTIR spectroscopy to transition into widespread clinical use as a prognostic tool, measurement errors must be mitigated wherever possible. CNNs have the potential to mitigate some of these effects due to the usage of convolution operations as a means of extracting useful features from FTIR spectra. This work has shown that through the use of an optimisation procedure it was possible to use CNNs to correctly classify FTIR spectra derived from primary tumour sites into useful risk groups. The CNN model evaluated here showed superior classification performance over a comparable MLP network architecture when evaluated using a number of conventional metrics. A thorough out of bag bootstrap procedure was used to obtain distributions of classification scores to estimate the variability of these scores. In comparison to the LR models developed in Chapter 3 the CNN model compared particularly well in a few key statistics. Whilst the FTIR-based LR models enjoyed higher AUROC scores, the dichotomised predictions fell short in comparison to the CNN model. The CNN model achieved a MCC score of 0.63 — signifying a very few misclassifications of any type.

The usage of these models could facilitate the ethical selection of patients for neo-adjuvant treatment in clinical window-trials whilst minimising overtreatment. This could be a crucial first step to improving the range of treatment

options available to patients with OSCC.

Bibliography

[1] Grigorios Tsagkatakis, Anastasia Aidini, Konstantina Fotiadou, Michalis Giannopoulos, Anastasia Pentari, and Panagiotis Tsakalides. Survey of deep-learning approaches for remote sensing observation enhancement. *Sensors (Switzerland)*, 19(18):1–39, 2019.

[2] Michael Haenlein and Andreas Kaplan. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 61(4):5–14, 2019.

[3] Jacopo Acquarelli, Twan van Laarhoven, Jan Gerretzen, Thanh N. Tran, Lutgarde M.C. Buydens, and Elena Marchiori. Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica Chimica Acta*, 954:22–31, 2017.

[4] Eric Kauderer-abrams. Quantifying Translation-Invariance in Convolutional Neural Networks.

[5] Osman Semih Kayhan and Jan C Van Gemert. On Translation Invariance in CNNs : Convolutional Layers can Exploit Absolute Spatial Location. (class 2):14274–14285.

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[7] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[8] George E. Dahl, Tara N. Sainath, and Geoffrey E. Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout.

ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, (2010):8609–8613, 2013.

[9] Kaiming He. Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification.

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Cameron Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

[13] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and

SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[14] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902, 2019.

[15] Yanzhao Wu, Ling Liu, Juhyun Bae, Ka Ho Chow, Arun Iyengar, Calton Pu, Wenqi Wei, Lei Yu, and Qi Zhang. Demystifying Learning Rate Policies for High Accuracy Training of Deep Neural Networks. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pages 1971–1980, 2019.

[16] Thomas M. Breuel. The effects of hyperparameters on SGD training of neural networks. *CoRR*, abs/1508.02788, 2015.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[18] Rasheed Omobolaji Alabi, Omar Youssef, Matti Pirinen, Mohammed Elmusrati, Antti A. MÃďkitie, Ilmo Leivo, and Alhadi Almangush. Machine learning in oral squamous cell carcinoma: Current status, clinical concerns and prospects for futureâĂŤa systematic review. *Artificial Intelligence in Medicine*, 115:102060, 2021.

[19] Rock Christian Tomas, Anthony Jay Sayat, Andrea Nicole Atienza, Jannah Lianne Danganan, Ma Rollene Ramos, Allan Fellizar, Kin Notarte Israel, Lara Mae Angeles, Ruth Bangaoil, Abegail Santillan, and Pia Marie Albano. Detection of breast cancer by ATR-FTIR spectroscopy using artificial neural networks. *PLoS ONE*, 17(1 January):1–24, 2022.

[20] Sebastian Berisha, Mahsa Lotfollahi, Jahandar Jahanipour, Ilker Gurcan, Michael Walsh, Rohit Bhargava, Hien Van Nguyen, and David Mayerich. Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks. *Analyst*, 144(5):1642–1653, 2019.

[21] Catalina Mercedes Burlacu, Steluta Gosav, Bianca Andreea Burlacu, and Mirela Praisler. Convolutional Neural Network D etecting S ynthetic C annabinoids. pages 24–27, 2021.

[22] Catalin Negoita, Mirela Praisler, and Iulia-Florentina Darie. Automatic identification of hallucinogenic amphetamines based on their ATR-FTIR spectra processed with Convolutional Neural Networks. *MATEC Web of Conferences*, 342:05003, 2021.

[23] Sang Ik Park, Jeffrey P. Guenette, Chong Hyun Suh, Glenn J. Hanna, Sae Rom Chung, Jung Hwan Baek, Jeong Hyun Lee, and Young Jun Choi. The diagnostic performance of CT and MRI for detecting extranodal extension in patients with head and neck squamous cell carcinoma: a systematic review and diagnostic meta-analysis. *European Radiology*, 31(4):2048–2061, 2021.

[24] Elham Alsahafi, Katheryn Begg, Ivano Amelio, Nina Raulf, Philippe Lucarelli, Thomas Sauter, and Mahvash Tavassoli. Clinical update on head and neck cancer: molecular biology and ongoing challenges. *Cell Death and Disease*, 10(8), 2019.

[25] Patrick K. Ha, Steven S. Chang, Chad A. Glazer, Joseph A. Califano, and David Sidransky. Molecular techniques and genetic alterations in head and neck cancer. *Oral Oncology*, 45(4):335–339, 2009. Oral Cancer Management. Pitfalls and Solutions.

6 Conclusions and Future Work

The objective of this thesis is to investigate the viability of machine learning methods applied to FTIR spectroscopy for use as prognostic tools in head and neck cancer. Despite phenomenal progress in the development of treatment options for OSCC over the past few decades, there is still work to be done; 5-year survival rates for patients with OSCC still remain around 50%. Additional treatment methods exist in the form of pre-operative neoadjuvant therapy, these treatment options show promise for improving clinical outcomes for patients with OSCC. However, research progress into neoadjuvant therapies is hindered by the difficulty of identifying patients eligible for clinical window trials. Due to the increased risk of adverse effects and additional health risks associated with hormone treatment and chemotherapy, it is unethical to accept patients who do not require additional treatment onto a trial. The difficulty lies in the fact that using pre-existing prognostic biomarkers it is only possible to determine a patient's prognosis after post-surgical nodal resection has taken place. The work undertaken in this thesis has shown that using sophisticated statistical methods to analyse FTIR spectroscopy data it may be possible to predict the prognosis of a patient from diagnostic biopsy tissue.

A list of criteria was set out in chapter 1 which summarises the requirements of any diagnostic method aiming to be widely adopted in a clinical setting. The performance of a diagnostic or prognostic test must be demonstrably superior to or at least be able to supplement existing methods; the method must also be sensitive to its intended range. Both of these criteria have been established

to be true of FTIR spectroscopy in chapters 3 and 5; a rigorous analysis procedure was performed which demonstrated that developed statistical models were able to stratify patients by risk status. Survival analysis was performed to ascertain the prognostic utility of FTIR spectra; results showed that risk groups had clearly distinguished survival curves. Chapter 3 contains a comparison to a known prognostic biomarker: ASMA, which was shown to be substantially less effective than FTIR in the data set used in the study. The two prognostic biomarkers were then combined to create a hybrid model which demonstrated superior performance in some cases over the two individual models. This suggests that combining prognostic biomarkers from a variety of assay methods may be an effective way to develop capable prognostic tools for use in a clinical setting. The current 'gold standard' for prognostic biomarkers are MRI and CT; these methods show reasonable performance as prognostic tools but further work is needed. The biological heterogeneity associated with OSCC has proven to be a difficult obstacle to overcome. The introduction of FTIR microscopy into clinical practice has great potential as a prognostic tool; however, the combination of a wide variety of prognostic biomarkers is likely a necessary next step to truly improve patient outcomes.

Another important criterion is the identification of sources of error, whether or not these sources of error can be addressed is crucial for the prospects for any diagnostic test. Chapter 4 covers the development of a distributed preprocessing optimisation framework for use with FTIR spectra; the work was undertaken in collaboration with another PhD student Barney Ellis. The optimisation framework was designed to provide a means of determining an effective preprocessing and classification pipeline for a given prediction objective. It is crucial to explore the parameter space of a pipeline as the classification performance of a pipeline is highly dependent upon the choice of configuration. The framework proved to be very successful, the top-performing pipeline configurations showed substnatially better scores with the difference between the best

and worst pipelines being $>$70%. The configurations of the top-ranked pipelines tended to be simpler but re-enforced some conventional wisdom about preprocessing FTIR data. The framework was developed on a widely adopted parallel processing framework, and could be easily adapted for other multivariate tasks with similar constraints. The framework will be made available for public use.

An evaluation of deep learning based FTIR analysis in Chapter 5 included a comparison between two different network architectures which were determined through a network optimisation process]. Both models showed some promise as prognostic tools, with the CNN model performing particularly well in comparison to all models developed in this thesis. This is an encouraging result as CNNs are more robust to measurement inconsistencies which would likely be a hindrance to the wider adoption of FTIR as a clinical tool.

To take the work covered here forward for further development; larger collaborative efforts should be made to improve the diversity and quality of data available for model training. CNNs should be explored more thoroughly as they have the potential to alleviate measurement issues, increasing the likelihood of wider adoption of the technique.