UNIVERSITY OF LIVERPOOL

DOCTORAL THESIS

Infrared Spectroscopic Techniques & Predictive Modelling Applied to Oral Cancer Diagnostics

Author:
Barnaby George ELLIS

Supervisor:
Prof. Peter WEIGHTMAN

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

SciaScan Research Group
Department of Physics

October 14, 2022

Declaration of Authorship

I, Barnaby George ELLIS, declare that this thesis titled, "Infrared Spectroscopic Techniques & Predictive Modelling Applied to Oral Cancer Diagnostics" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

UNIVERSITY OF LIVERPOOL

Abstract

Faculty of Science & Engineering

Department of Physics

Doctor of Philosophy

Infrared Spectroscopic Techniques & Predictive Modelling Applied to Oral Cancer Diagnostics

by Barnaby George ELLIS

Cancer is a leading cause of mortality and morbidity worldwide. The burden cancer imposes on health services is increasing, owing to an ageing and growing population. Whilst effective treatment is essential, the prevention, diagnosis and risk stratification of early stage cancer is paramount. Subjectivity and ambiguity are a hindrance in the diagnosis of many cancers. Oral cancer is a condition that is often diagnosed at a late stage as asymptotic early stage conditions regularly go undetected.

Vibrational spectroscopy is a family of techniques which allow an objective view of the intrinsic chemistry of a sample. It has shown great promise in the field of cancer diagnosis, but there is still a significant gulf between research efforts and clinical adoption.

The work contained within this thesis describes the utilisation of novel vibrational spectroscopy analytical methods to extract important information from pathological oral tissue. The important information is used to provide further insight into the biochemistry of malignancy, as well as to attempt to predict malignant transformation in early stage cancer.

Acknowledgements

Throughout the course of my PhD I have been lucky enough to work with experts from a variety of fields. Enormous gratitude is due to my project advisor Professor Peter Weightman and the SciaScan research group, with whom I have worked intimately over the past 4 years. Particular thanks go to Conor Whitley, Steven Barrett, Paul Harrison and Caroline Smith.

The biological aspect of my work was predominantly facilitated by Dr Janet Risk, who has been invaluable in her patience and expert advice. Thank you for taking me from a complete novice to a semi-complete novice. Big thanks are also extended to Phillip Gunning, Dr Asterios Triantyflou and Professor Richard Shaw, who aided in the histopathological and clinical aspect of the work.

I am grateful to Professor Peter Gardener and his laboratory team at the University of Manchester for facilitating FTIR experiments. Thanks are also extended to Professor Antonio Cricenti and Dr Marco Luce for expert advice on operating their bespoke IR-SNOM instrument.

And perhaps most of all, the biggest thanks goes to my girlfriend Madeleine, family and housemates for listening to me toil over this thesis for the past year. I can only apologise for making the COVID-19 lockdown yet more unbearable. You're all exceptionally patient.

Contents

Declaration of Authorship i

Abstract ii

Acknowledgements iii

1 Introduction 1

2 Clinical Problem 4

2.1 Oral Cancer 4

2.1.1 Definition and Statistics 4

2.1.2 Diagnosis 6

2.2 Potentially pre-Malignant Oral Epithelial Lesions 10

2.2.1 Clinical Lesions 10

Oral Leukoplakia 10

Oral Erythroplakia 12

2.2.2 Oral Epithelial Dysplasia 12

Clinical Determinants 13

Diagnosis 14

2.2.3 Testing 16

Performance metrics 16

Clinical Adjunctive Tests 19

Biomarkers 21

2.3 Summary and Motivations for Work 21

3 Theoretical Concepts 23

3.1 Introduction 23

3.2 Infrared Spectroscopy and Imaging 23
3.2.1 Fundamentals 23
Light-Matter Interactions 23
Molecular Vibrations 27
3.2.2 Fourier Transform Infrared Spectroscopy 32
Sources 32
Interferometer 33
Detectors 35
3.2.3 Microscopy and Imaging 37
Hyperspectral Imaging 39
3.2.4 Application to Cancer Diagnostics and Characterisation 41
3.3 Analytical Techniques 43
3.3.1 Pre-processing 44
Denoising 44
Baseline Correction 46
Normalisation 47
Feature Extraction 49
3.3.2 Machine Learning 50
Unsupervised Learning 51
Supervised Classifiers 51
Metric Analysis 53
3.4 Scanning Near Field Optical Microscopy (SNOM) 55
3.4.1 History 57
3.4.2 Principles 58

4 Oral Cancer Analysis **62**
4.1 Background 62
4.2 Methods 66
4.2.1 Sample Selection and Preparation 66
4.2.2 FTIR Experiments 69
4.2.3 Data Labelling and Pre-processing 75
4.3 Phase I: Multi-tissue Primary Site Model 79

4.3.1 Parameter selection and Model Training 79
4.3.2 Results 82
4.3.3 Discussion 85
4.4 Phase II & III: Binary Metastasis in Lymph Node Analysis 89
4.4.1 Spectral Biomarker Extraction 89
4.4.2 IR-SNOM Instrument 91
4.4.3 Experimental 100
Sample Preparation and Model Validation 102
SNOM Experiments 103
Image Processing 106
4.4.4 Results 107
4.4.5 Discussion 110
4.5 Conclusion 113

5 Pipeline Optimisation Framework 115
5.1 Introduction 115
5.1.1 Background and Motives 115
5.1.2 Bayesian Hyperparameter Search 118
5.2 Framework Components 124
5.2.1 Steps and Methods 124
5.2.2 Pipeline Construction 128
5.2.3 Job Dispatching and Execution 128
5.3 Testing 131
5.3.1 Methods 131
5.3.2 Results 132
5.3.3 Discussion 137
5.4 Conclusion 141

6 Dysplasia Transformation Analysis 142
6.1 Background 142
6.2 Experimental 145
6.2.1 Sample Selection & Preparation 145
6.2.2 FTIR Experiments 146

6.2.3 Labelling & Quality Checks 146

6.2.4 Pre-processing 149

6.3 Results 150

6.3.1 All Patients Model 150

6.3.2 Low Grade Patients Model 157

6.3.3 High Grade Patients Model 160

6.4 Discussion 164

6.4.1 All Patients Model 165

6.4.2 Low Grade Patients Model 168

6.4.3 High Grade Patients Model 174

6.4.4 Spectral Biomarker Analysis 180

6.5 Conclusion 185

7 Future and Outlook **187**

7.1 Multi-Centre Studies 189

7.2 Different Sample Domain 189

7.3 Deep Learning 190

Bibliography **193**

List of Figures

2.1 Schematic of stratified squamous epithelium. 5

2.2 Possible pathways induced by tobacco smoking [14] 6

2.3 Kaplan-Meier survival curves for different stages of oral cancer [18]. . . 8

2.4 Microscopic image of H&E stained tissue section from dysplastic epithelium. Individual nuclei are discernible from the surrounding connective and epithelial tissue. Scale = 200 μ m. 10

2.5 Oral leukoplakia of the lateral tongue. OL is characterised by a flat, homogeneous white appearance. 11

2.6 Examples of dysplasia grading 15

2.7 Confusion matrix of a diagnostic test 17

2.8 Simple depiction of a diagnostic test and the resulting ROC curve 18

2.9 Velscope and Toluidine blue can locate abnormal dysplastic tissue *in vivo*. (a) Photograph of lateral tongue ROI with leukoplakia. (b) Photograph of ROI illuminated with Velscope. (c) Photograph of ROI stained with toloudine blue. (d) Microscopic image of HE stained tissue confirming the presence of dysplasia. 20

3.1 Basic depiction of an EM wave [63] 24

3.2 Behaviour of EM radiation at a boundary between two media with refractive indexes $n_1(\overline{v})$ and $n_2(\overline{v})$. 26

3.3 Stretching vibrations of a simple linear diatomic molecule. r_0 is the equilibrium inter-atomic separation, $r_{1,2}$ are the inter-atomic separations at points of maximum extension and compression respectively. . . 28

3.4 Harmonic oscillator potential energy (V) as a function of inter-atomic separation (r). Red lines indicate discrete energy levels allowed in a quantum harmonic oscillator 30

3.5 Anharmonic oscillator potential energy (V) as a function of inter-atomic separation (r). 31

3.6 The amide I (a) and amide II (b) vibrational modes. The amide I mode ($\approx$ 1650 cm^{-1}) is a stretching vibration centred on the C=O bond, and the amide II mode ($\approx$ 1550 cm^{-1}) is a bending vibration of the hydrogen atom. 31

3.7 A schematic of a Michelson Interferometer. 34

3.8 Semiconductor detector schematic. When a photon with an energy greater than bandgap of the semiconductor is incident on the detector, an electron is excited to the conduction band, leaving a hole in the valence band. Movement of these charge carriers leads to a measurable electric current. 37

3.9 Basic schematic of an FTIR microscope in transmission mode. 38

3.10 Depiction of hyperspectral datacube resulting from a 4×4 FPA. Each slice in the data is a unique array of values corresponding to the absorbance at a particular wavenumber. 40

3.11 Typical biological spectrum and corresponding peak assignments. Type of vibrations signified by a v (stretching), δ (bending), s (symmetric) and as (asymmetric). Extracted with permission from [74]. Colour of peak label indicates group of biomolecules each spectral biomarker is related to (lipid: blue, protein: red, nucleic acid: green, carbohydrate: yellow.) . 42

3.12 Basic depiction of SG smoothing. A linear polynomial is fitted to a window of 21 points. The window is propagated through the data to acquire a smoothed signal. 44

3.13 Example of raw (a) and Mie scattering corrected (b) spectra of prostate tissue. There is a prominent undulating baseline in (a), with peak shifts in the amide region. Reproduced with permission from [85]. 48

3.14 Typical workflow of supervised ML experiments. The optimisation of hyperparameters is often an iterative process which should be isolated from the test data to prevent bias leaking through to evaluation. 53

3.15 Simple depiction of the formulation and assessment of a metric. 54

3.16 A sub-micron imaging system conceptualised by Edward Synge. A small aperture in an opaque screen is used as a nano-source, with a sample situated in the near field zone of the source. 58

3.17 Depiction of aperture SNOM operating in collection mode. A plane wave is incident on a sample. The high frequency components in the diffracted light decay away rapidly with distance from the surface. A limited object can be used to convert the evanescent field into propagating waves which are directed towards a detector through an optical fiber. 60

4.1 Pallua *et al* [107], fig. 7. (a) HE stained sample from TMA, regions of interest corresponding to cancer, hornification and desmoplastic stroma identified by histopathologist. (b) - (d) Peak area maps for 3 different wavenumber ranges: 1385 cm^{-1} - 1380 cm^{-1} (glycoproteins), 1225 cm^{-1} - 1220 cm^{-1} (nucleic acids), 1085 cm^{-1} - 1080 cm^{-1} (nucleic acids). (e) HCA pseudo-colour image. (d) KMC pseudo-colour image. 65

4.2 (a) HE of entire tissue section from OSCC patient, with regions of interest (ROI) defined by histopathologist. (b) ROI 1: Invasive OSCC with surrounding stroma tissue. (c) ROI 2: Normal epithelium with adjacent stroma and skeletal muscle tissue. 68

4.3 Labelled photograph of apparatus used for FTIR-MS experiments. . . . 70

4.4 Schematic of the FTIR system. More detailed figures for the optical system are in Fig. 3.9 and Fig. 3.7. 70

4.5 Representations of the illumination distribution on the FPA. The 2D intensity map is shown in (a), whilst the intensity as a function of flattened position is shown in (b). A tight, uniform intensity distribution is desired, represented by the compact, shallow curve in (b) 72

4.6 Plots of different apodisation functions (a-c) and respective Fourier transforms (d-f) . . . 73

4.7 Mean and standard deviation of an example background transmission spectrum obtained from CaF_2 housed in a purged chamber. . . . 74

4.8 Procedure for finding and selecting ROIs using the visible mosaic function in the resolutions pro software. (a) is the large visible mosaic, (b) is the cutout selected for IR imaging, informed by the H&E counterpart in (c). . . . 75

4.9 Representative LM images of H&E tissue containing the eight identified tissue types. . . . 77

4.10 Mean spectra and standard deviation for (a) OSCC, (b) NE, (c) CS, (d) NS, (e) MT, (f) LYM, (g) SM, (h) BL. (i) stack of the average spectrum from each tissue for comparison. . . . 78

4.11 Plots of the mean (with error-bars) of the AUC (a), sensitivity (b), specificity (c) for each of the tissues within each different permutation of hyperparameters. The horizontal axis labels denote the hyperparameter permutation in the order [***down-sampling factor, optimising function***]. . 81

4.12 Importance plot for multi-class study. Tissue types from top to bottom: OSCC, NE, CS, NS, LYM, SM, BL. . . . 83

4.13 Distributions of the best metric for OSCC (a) and LYM (b). Positive class is circled in black. . . . 84

4.14 Backlabelled images. (a) and (b) were used to extract a portion of spectra used for model construction, (c) was not used for analysis. (d-f) are LM HE counterparts to (a-c) respectively for comparison. . . . 85

4.15 Tissue mean spectra in the region 1500 - 1700 cm^{-1}. Best metric for OSCC has been drawn with two dotted lines. . . . 88

4.16 Distributions of absorbance at (a) 1252 cm^{-1} and (b) 1252 cm^{-1}, which both form the best metric $\delta_{1252,1285}$, the distribution of which is shown in (c). . . . 90

4.17 Labelled photographs of the SNOM instrument used in this work. (b) is a photograph of the oscilloscope unit that is used to monitor multiple signals from the QCL, MCT and lock-in; and the lock-in amplifier which is used to isolate a periodic signal from noise with a continuous frequency spectrum. (c) is a zoomed in view of the SNOM head/stage and microscope system. 92

4.18 a) Schematic diagram of the driving and sensing bimorph which are driven into oscillation by an external voltage. Extracted with permission from James Ingham's PhD thesis [119]. (b) Shows how the contraction and expansion of either side of the driving bimorph results in oscillation. 93

4.19 QCL emission spectrum. 95

4.20 Optical path taken by the IR beam from the QCL to the optical fibre tip. 96

4.21 Path taken by probe relative to the sample surface. For the forward scan (green), the position of the sample is stepped to the right until the end of the line, at which point it is stepped all the way back to the left for the backwards scan (purple). At each point (red dot), the signal is recorded so that an image can be formed. At the end of each line, the stage steps up so the line beneath can be obtained. 99

4.22 Size of the features of the hysteresis calibration grid. 99

4.23 Forwards (a) and backward (b) topography scans of the calibration grid. 100

4.24 Curves of observed position as a function of true position for horizontal (a) and vertical (b) orientations. 101

4.25 Corrected forward (a) and backward (b) scans. 101

4.26 H&E (a,c) and IHC (b,d) images of the cores extracted for study. Black arrows indicate periphery of tumour, white arrow indicates keratin pearl within tumour. The green boxes indicate the region of interest (ROI) that will be targeted. 104

4.27 Core 1 absorbance images at (a) 1252 cm^{-1} and (b) 1285 cm^{-1}. The ratio between the two images is shown in (c), corresponding the optimum metric emerging from section 4.4.1. Black and white arrows used to indicate the same features as in Fig. 4.26b and d. Each FTIR image is plotted with a colour table covering the 5th to 95th percentiles of the image intensity range. 105

4.28 (a) H&E stained image, (b) IHC image stained for pan-cytokeratins (dark brown), (c) topography, IR SNOM images at (d) 1751 cm^{1}, (e) 1650 cm^{1}, (f) 1369 cm^{1}, (g) 1285 cm^{1}, (h) 1252 cm^{1} and (i) ratio of 1252 cm^{1}/1285 cm^{1} [i.e. (h)/(g)]. All images are 300 μm × 300 μm. Each SNOM IR image is plotted with a colour table covering the 5th to 95th percentiles of the image intensity range. Image (a) was obtained from a section adjacent to that used to obtain image (b), which was in turn adjacent to that used to obtain images (c) to (i). 108

4.29 H&E stained image of region enclosed by green square in Fig. 4.28a (left) and line profiles (right) taken through the core at the white line showing (a) topography, (b) 1751 cm^{-1}, (c) 1650 cm^{-1}, (d) 1369 cm^{-1}, (e) 1285 cm^{-1}, (f) 1252 cm^{-1} and (g) ratio of 1252 cm^{-1}/1285 cm^{-1} [i.e. (f)/(e)]. H&E image (left) was obtained from a section adjacent to that used to obtain SNOM line profiles. Each line profile has been normalised to its min/max values. 109

5.1 Comparison between a grid search and random search for HP optimisation. The two methods are used on the same problem, for which two different optimal combinations are depicted by the green and yellow stars. The grid search is configured to search over 9 user-defined positions in the space, whereas the random search is configured to randomly sample combinations from the space. 119

5.2 2D function $f(\theta_1, \theta_2)$. Implementing random or grid search to find the maximum to this function would be inefficient. 120

5.3 Finding the optimum of a 2D function using BO over 20 iterations. 1st, 10th and 20th iteration shown for brevity. Each panel shows the true function (a), mean function (b), standard deviation (c) and expected improvement (d). 123

5.4 Summaries of the transformer and estimator `scikit-learn` base classes. 125

5.5 Example of a simple decision tree with two root nodes and three leaf nodes for a three class classifier. 127

5.6 Structure of the pipeline search dictionary. All combinations of methods (taking one from each step) are generated, combining all HPs from the selected methods into one domain. The combined HP vector is subject to BO. 129

5.7 Process of dispatching n pipelines and a set of mutual inputs to each job in the condor framework. 130

5.8 Flowchart of overall optimisation process 132

5.9 Mean scores of pipelines for various metrics [(a): AUC, (b): MCC, (c): sensitivity, (b): specificity] , ranked by AUC. Red circle in (a) marks pipeline which applies no pre-processing and logistic regression. 133

5.10 Mean confusion matrix and ROC curve shown with standard errors for best (a,b) and second best (c,d) pipelines as shown in table 5.1. 134

5.11 GP hyperparameter surfaces showing mean function in red and standard deviations in blue averaged across 50 sample iterations. 135

5.12 Histograms of optimum hyperparameters over the 50 train-test splits. . 136

5.13 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.137

5.14 Frequency each method either enhances (green) or diminishes (red) relative to the median score (AUC = 0.48). Steps are (a) smoothing, (b) baseline, (c) normalisation, (d) scaling, (e) feature-extraction, (f) classifier.139

6.1 Workflow for the labelling protocol. Pre-processed image (a) is subject to two tiers of KCA (b-c), with the epithelial and dysplastic regions guided and confirmed by histopathologist informed by H&E image (d). The dysplastic layer in this image were defined as the brown cluster. . . 148

6.2 Example of a quality control test for data from a single image. Black crosses in (a) and red pixels in (b) indicate data points which are deemed outliers by the test. (b) IR image at 1650 cm^{-1}. Green dots indicate non-outliers. Scale bar = 200 µm. 149
6.3 The method and HP search spaces used to optimise each model. 151
6.4 Mean spectra when the data is grouped by (a) outcome and (b) grade. . 152
6.5 A matrix depicting the patient pairings from the two outcome groups. . 152
6.6 Mean MCC (a), AUC (b), sensitivity (c) and specificity (d) for pipelines (sorted in order of decreasing mean MCC). All patients incorporated into study. 153
6.7 Mean and median ROC curves over the 224 individual splits, which are shown in paler colours. 155
6.8 Number of times patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies. 156
6.9 Mean MCC (a), AUC (b), sensitivity (c) and specificity (d) for pipelines (sorted in order of decreasing mean MCC). Low grade patients study. . . 158
6.10 Mean and median ROC curves over the 42 individual splits (shown in paler colours) for low grade patients. 159
6.11 (a) Mislabelling rates for each transforming (a) and non-transforming (b) patient. 160
6.12 Number of times low-grade patients patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies. 161
6.13 Mean MCC (a), AUC (b), sensitivity (c) and specificity (d) for pipelines (sorted in order of decreasing mean MCC). High grade patients. 162
6.14 Mean and median ROC curves over the 70 individual splits (shown in paler colours) for high grade patients. 163
6.15 Mislabelling rates for each transforming (a) and non-transforming (b) patient. 164

6.16 Number of times high-grade patients patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies. 165

6.17 Number of times patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies, shown for different thresholds. 167

6.18 Number of times high-grade patients patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies, shown for different thresholds. 169

6.19 The spatial arrangement of test spectra super-imposed on the absorbance map at 1650 cm^{-1}, for each of the models where 12194 (T) appears in test set, of which there are 7 (a-g). The color encodes the transformation probability for each spectrum. Hard and soft voting probabilities are also shown to inform the patient outcome decision. Scale bar in (a) = 200 µm. 170

6.20 H&E stained sample of ROI shown in Fig. 6.19. White arrows are pointing the basal layer of the epithelium, which contains regions found confusing by the model. 171

6.21 The spatial arrangement of test spectra super-imposed on the absorbance map at 1650 cm^{-1}, for each of the models where 12182 (T) appears in test set, of which there are 6 (a-f). The color encodes the transformation probability for each spectrum. Hard and soft voting probabilities are also shown to inform the patient outcome decision. 172

6.22 H&E stained sample of ROI shown in Fig. 6.21. The white arrows are pointing toward the rete ridges of the basal layer, which contains regions which the model finds confusing. 173

6.23 Number of times low grade patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies, shown for different thresholds. 174

6.24 The spatial arrangement of test spectra super-imposed on the absorbance map at 1650 cm^{-1}, for each of the models where 12329 (NT) appears in test set, of which there are 10 (a-j). The color encodes the transformation probability for each spectrum. Hard and soft voting probabilities are also shown to inform the patient outcome decision. Scale bar in (a) = 200 µm. 176

6.25 H&E stained sample of ROI shown in Fig. 6.24 177

6.26 The spatial arrangement of test spectra super-imposed on the absorbance map at 1650 cm^{1}, for each of the models where 12127 (T) appears in test set, of which there are 7 (a-g). The colour encodes the transformation probability for each spectrum. Hard and soft voting probabilities are also shown to inform the patient outcome decision. Scale bar in (a) = 200 μm 178

6.27 H&E stained sample of ROI shown in Fig. 6.26 179

6.28 Number of times patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies, shown for different thresholds. 179

6.29 (a) Data points (spectra) transformed into predictions by the trained sigmoid function (see Eq. (3.24)) for one of the train-test splits. Those with a sigmoid output of > 0 are predicted as the positive class. (b) The ROC curve for the transformed points shown in (a). 181

6.30 (a) Logistic regression model for a single train-test split, for which the ROC curve is shown in (b) 181

6.31 Combined weight plot for PCA-LR model for low-grade patients. 182

6.32 Decision boundary (top) and combined weight plot for PCA-LCA model (bottom) for high-grade patients. The stars delineate the wavenumbers with the highest weight amplitude, listed in table 6.5 183

List of Tables

2.1 TNM staging of OSCC 7

2.2 WHO architectural and cytological criteria used for grading OED. 14

4.1 Patients selected for imaging. 69

4.2 Identified tissue types and their respective ROIs and number of labelled spectra. 77

4.3 MA.v4 hyperparameters 79

4.4 Results on hold-out test data 82

4.5 Best metric for each tissue type in the multi-class study. 84

4.6 Results for MT vs LYM binary model 90

4.7 Wavenumbers selected for SNOM study 101

4.8 Wavenumber specific experimental parameters. 106

5.1 Top ranking pipelines. 134

6.1 Top ranking pipelines for all dysplasia patients model. 154

6.2 Top ranking pipelines for low grade dysplasia patients model. 157

6.3 Top ranking pipelines for high grade dysplasia patients model. 163

6.4 Worst pipelines for high grade dysplasia patients model. 175

6.5 Important wavenumbers for outcome prediction 184

List of Abbreviations

AFM	Atomic Force Microscopy		
AI	Artificial Intelligence	**API**	Application Programming Interface
BL	Basal Layer	**BO**	Bayesian Optimisation
CA	Cluster Analysis	**CAF**	Cancer Associated Fibroblast
CDF	Cumulative Density Function	**CS**	Cancer Stroma
EI	Expected Improvement	**EM**	Electromagnetic
EMSC	Extended Multiplicative Signal Correction	**FTIR**	Fourier Transform Infrared
FTIR-MS	FTIR Microspectroscopy	**FFPE**	Formalin Fixed Paraffin Embedded
FT	Fourier Transform	**FFS**	Forward Feature Selection
GP	Gaussian Process	**H&E**	Haemotoxylin & Eosi
HP	Hyperparameter	**HSI**	Hyperspectral Imaging
IHC	Immunohistochemistry	**IR**	Infrared
LDA	Linear Discriminant Analysis	**LM**	Light Microscopy
LYM	Lymphoid Tissue	**MA**	Metric Analysis
MCC	Matthew's Correlation Coefficient	**MCT**	Mercury Cadmium Telluride
MI	Michelson Interferometer	**ML**	Machine Learning
MT	Metastatic Tumour	**MTR**	Malignant Transformation Rate

NA	Numerical Aperture	**NE**	Normal Epithelium/Maturation Layer
NPV	Positive Predictive Value	**NS**	Normal Stroma
OE	Oral Erythroplakia	**OED**	Oral Epithelial Dysplasia
OL	Oral Leukoplakia	**OPMD**	Oral Pre-Malignant Disease
OSCC	Oral Squamous Cell Carcinoma	**PCA**	Principal Component Analysis
PDF	Probability Density Function	**PMMA**	Polymethyl Methacrylate
PPOEL	Potentially pre-Malignant Oral Epithelial Lesion	**PPV**	Positive Predictive Value
QCL	Quantum Cascade Laser	**RF**	Random Forest
ROC	Receiver Operating Characteristic	**ROI**	Region of Interest
SG	Savitzy Golay	**SM**	Skeletal Muscle
SNR	Signal to Noise Ratio	**SNOM**	Scanning Near Field Optical Microscopy
SPM	Scanning Probe Microscopy	**SVD**	Singular Value Decomposition
STM	Scanning Tunnelling Microscopy	**WHO**	World Health Organisation
T	Transforming OED	**TB**	Toluidine Blue
TMA	Tissue Micro-Array	**TNM**	Tumour Node Metastasis
TPMD	Tetramethylpentadecane	**XGBoost**	Extreme Gradient Boosting

Chapter 1

Introduction

The burden of cancer is a crisis for health systems worldwide. Developed and developing nations unanimously regard cancer as a leading cause of death, with the World Health Organisation (WHO) estimating in 2019 that cancer ranks as the first or second leading cause of mortality in 112 out of 183 countries [1]. The severity of the crisis is rapidly increasing, largely owing to ageing populations, socioeconomic factors and increased exposure to high risk carcinogens such as tobacco, alcohol and processed foods.

Cancer can be broadly defined as a biological malfunction that leads to the abnormal and uncontrolled proliferation of cells. The cancer cells often accumulate to form solid tumours, and many are able to spread (metastasise) to distant regions of the body via the circulatory and lymphatic systems. Breast, lung, prostate, colorectal and stomach cancer are amongst the most prevalent cancers [2], with lung cancer having by far the highest mortality rate

Recently, the global cancer observatory (GLOBOCAN) estimated global incidence and mortality rates for all cancers at 19.3 million and 10.0 million per year respectively [3]. This alarmingly high mortality rate results from a myriad of interconnected factors, particularly deficiencies in health services, diagnostic difficulties and the scarcity of expensive treatment, particularly in poorer countries.

Early diagnosis is of particular importance and focus in cancer research. The high mortality associated with many cancers can be mainly attributed to the difficulties in diagnosing early stage cancer. Many cancers may not present symptoms until a late stage, at which point the chances of survival are significantly diminished. A sobering example of this is the mortality rate of lung cancer, which suffers an abysmal 5-year survival rate of merely 21.7% [4]. However, when diagnosed as early stage IA1 small cell cancer, the 5-year survival rate dramatically increases to 92% [5]. Clearly, improvements in early diagnostics are pivotal in the fight against cancer.

Vibrational spectroscopy is a well established field which has recently been experiencing a surge in applications, owing primarily to advances in technology, analytical techniques and understanding. Vibrational spectroscopic techniques, such as Fourier transform infrared (FTIR) spectroscopy and Raman spectroscopy, all share the common defining feature that they are able to objectively and non-destructively probe the chemical properties of a sample. This property has led to widespread analytical applications in food sciences [6], pharmaceuticals [7] and biomedical sciences [8].

Of particular interest is the exploitation of vibrational spectroscopic techniques to investigate and diagnose cancer. Combined with microscopy and sophisticated analysis, it has been shown to be an effective and promising objective diagnostic and exploratory tool in tissue (histopathological), cell (cytological), biofluid, and in-vivo applications. FTIR microspectroscopy (FTIR-MS) is a unification of microscopy and spectroscopy which enables the relatively rapid acquisition of spatially resolved infrared spectra, revealing the intrinsic chemical distribution of a sample at microscopic length scales of the order of 10^{-5} m.

Machine learning (ML) is a branch of artificial intelligence (AI) which can be generally defined as algorithms which 'learn' rules and patterns from sample training data, with the intention of applying the learned rules and patterns to unseen data in order to categorise them. Supervised learning is a subset of ML algorithms which build discriminatory models based on prior knowledge of data labels. Unsupervised learning, on the other hand, infers sub-groups based on the training data alone [9].

This thesis is based around the application of infrared imaging and novel analytical techniques to the characterisation of oral tissue. Firstly, FTIR-MS and scanning near-field optical microscopy are exploited as a tool to investigate the differences between different pathologies associated with oral squamous cell carcinoma (OSCC). Following this, a novel framework which optimises an analysis pipeline is described with real world examples. Finally, FTIR-MS and the analysis framework are deployed to predict the malignant potential of oral epithelial lesions.

Chapter 2

Clinical Problem

2.1 Oral Cancer

2.1.1 Definition and Statistics

The incidence rate of oral cancer in the UK has increased by 30% in the past two decades, a trend which is expected to continue [10]. Oral cancer refers to malignancy within the mouth, occurring at sites such as the lips, gums, tongue, inner lining of the cheeks (buccal), floor of the mouth and hard palate. Whilst there are several diseases that can be characterised as oral cancer, oral squamous cell carcinoma (OSCC) is by far the most common, accounting for over 90% of cases [11]. As for all cancers, OSCC progression stems from DNA mutation, leading to an accumulation of abnormalities which ultimately result in the uncontrolled malignant spread of cancer [12].

Mucous membrane (mucosa) is the term given to the protective tissue that surrounds organs and orifices such as the mouth. As its name suggests, mucosa is characterised by it's ability to secrete mucous, a thick fluid which acts as a first line of defence against pathogens. Oral mucosa is the lining inside the oral cavity, consisting of an epithelial layer (stratified squamous epithelium) and underlying connective tissue. The stratified squamous epithelium is characterised by superficial horizontally flattened (squamous) cells stacked atop several layers of more regularly shaped cells, acting as a protective layer to the deeper lying tissue. A basic schematic of a stratified squamous epithelium is shown in Fig. 2.1.

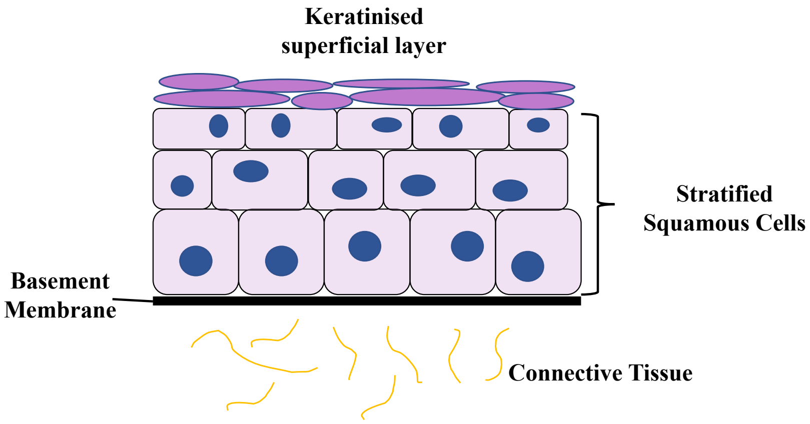


FIGURE 2.1: Schematic of stratified squamous epithelium.

It is within the squamous cells of the epithelium that OSCC originates. OSCC's development is a multi-step process by which the accumulation of genetic mutations within the nuclei of the epithelial cells leads to their atypia and uncontrolled proliferation, which can eventually result in invasive and metastatic tumour growth [13]. Exposure to carcinogens such as tobacco, alcohol and betel nut have been shown to increase the risk of malignancy [14], with factors such as age and gender also posing risk, yet these may be entangled with socio-economic factors.

The oral cavity is immediately prone to insult from carcinogens such as tobacco and alcohol. Additional to this is the fact that the distribution of carcinogenic insult is essentially uniform across the oral cavity. For instance, tobacco smoke disperses to take the shape of the oral cavity once inhaled, similarly for alcoholic beverages. As a result, relatively large regions of exposed tissue experience a uniform increase in malignant risk. This theory, coined 'field cancerization', forms the basis of the current understanding of the mechanisms that lead to the occurrence of multiple macroscopically distant lesions in the oral cavity.

There are multiple epigenetic pathways associated with tobacco smoking, depicted in Fig. 2.2. The term 'epigenetic' implies influence of non-genetic factors on the intrinsic genetic expression of cells. The *TP53* gene codes for the p53 protein which regulates the cell cycle by steering cell growth arrest and apoptosis (cell death). p53 damage leads to a significant reduction in tumour suppression, leading to it being dubbed 'the guardian of the genome' [15]. *GLUT-1* is part of a family of proteins which mediate the influx of glucose into the cell, and thus the metabolic capability of the cell leading to growth advantage. The deregulation of GLUT-1 has been strongly correlated with grade of tumour [16], [17].

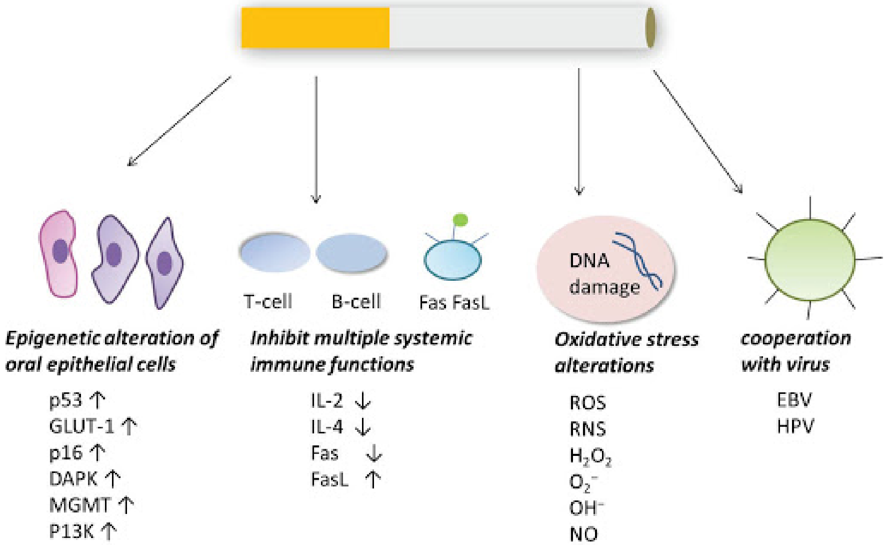


FIGURE 2.2: Possible pathways induced by tobacco smoking [14]

2.1.2 Diagnosis

Risk stratification and management of OSCC is influenced by numerous factors, primarily the stage and grade of tumours. Categorisation into stages is carried out using criteria reflecting the extent of tumour growth. The most commonly used system for staging is the *American Joint Committee on Cancer* (AJCC) Tumour, Node, Metastasis (**TNM**) system [18], which is based on the following:

- **Tumour**: Determine size of primary site tumour, assess spread to nearby tissue in oral cavity.
- **Node**: Check for lymph node spread.
- **Metastasis**: Check for secondary tumours in distant organs.

Each of the three categories are assessed and assigned a stage. Table 2.1 summarises the criteria associated with each category. The three categories carry significant prognostic value, with the extent of the nodal (N) stage highly correlated with poor outcomes [19].

TABLE 2.1: TNM staging of OSCC

Tumour	**Node**	**Metastasis**
Tis- Carcinoma in situ	**N0**- No cancer in lymph nodes	**MO**- No spread to other parts of body.
T1- $x_T < 2\,\text{cm}$, $d < 5\,\text{mm}$	**N1**- Cancer present in one lymph node same side of neck as primary tumour. $node < 3\,\text{cm}$.	**M1**- Cancer spread to other parts of body
T2- $2\,\text{cm} < x_T < 4\,\text{cm}$, $5\,\text{mm} < d < 10\,\text{mm}$	**N2**- *(a)* Cancer present in one lymph node same side of neck as primary tumour. $3\,\text{cm} < x_N < 6\,\text{cm}$. *(b)* More than one lymph node same side as neck as primary tumour contain cancer cells, $x_N < 6\,\text{cm}$. *(c)* Cancer present in lymph node other/both side(s) of neck to primary tumour. $x_N < 6\,\text{cm}$	
T3- $x_T > 4\,\text{cm}$, $d > 10\,\text{mm}$	**N3**- *(a)* Cancer containing lymph node, $x_N >$ 6cm. *(b)* Any number of lymph nodes contains cancer, spread into surrounding tissue.	
T4- *(a)* Grown into surrounding structures (sinuses, skin, jaw), *(b)* Grown beyond surrounding structures (skull, neck)		

x_T: Tumour size
x_N: Lymph node size
d: Depth

Each of the three stages is then used to stratify a patient into an overall stage group, enumerated from stage 0 to stage IV, with stage IV indicating the poorest prognosis. Each of the strata contains a near uniform survival across all patients. The prognostic value of the staging system can be demonstrated through a Kaplan-Meier survival curve, which is a non-parametetric representation of the survival of a population over a defined time period. The Kaplan-Meier curves in Fig. 2.3 clearly illustrates the correlation between stage and survival.

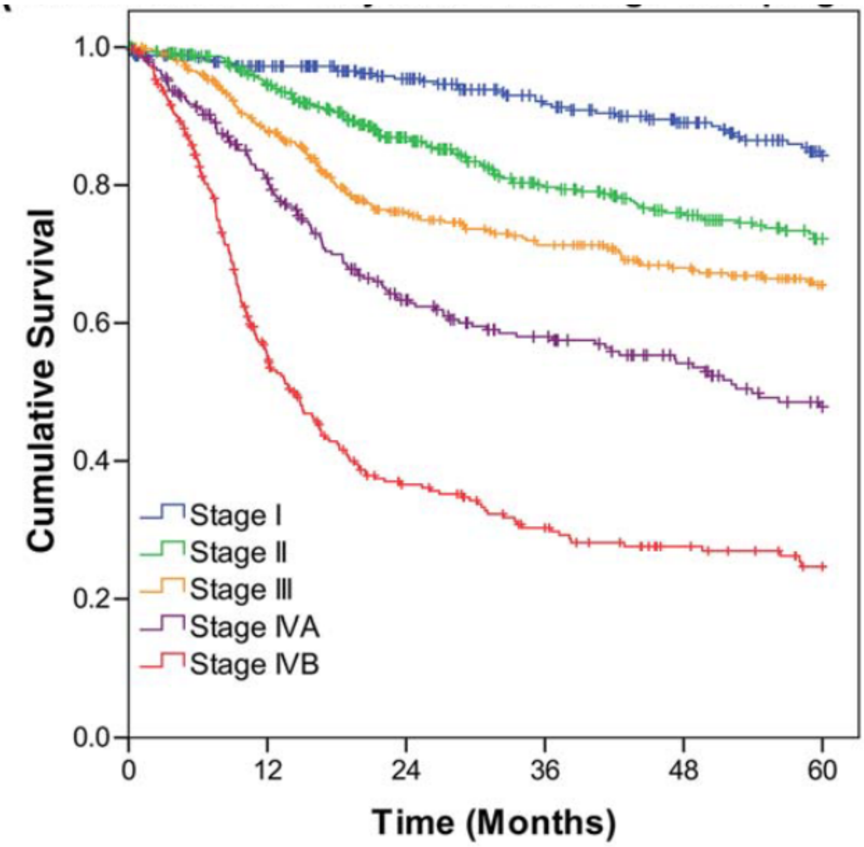


FIGURE 2.3: Kaplan-Meier survival curves for different stages of oral cancer [18].

The grade of a tumour is determined by a histopathologist tasked with microscopically examining sectioned tissue, and is a measure of how differentiated the cancer cells are compared to healthy cells. The first OSCC grading system was devised in 1920 by Broder, specifically for lip cancers [20] and is closely related to the the modern WHO system [21]. There are numerous studies which criticise grading systems, suggesting they have little to no prognostic value [22]. Despite this, they are still employed as an important factor in the risk stratification of OSCC.

Histopathological examination and diagnosis is a field with origins that can be traced back to the nineteenth century. Extracted tissue is routinely treated with a fixative such as formalin to halt autolysis (postmortem decay), then embedded in paraffin wax to displace water and preserve the structure of tissue. The melting point of histology grade paraffin (57 °C) is crucial as this temperature does not change the structure and morphology of the tissue. The tissue is stored in formalin-fixed, paraffin-embedded (FFPE) blocks. Thin sections (4 - 5 µm) can be easily extracted from FFPE tissue blocks with a precision knife known as a microtome. Transfer onto a microscopy slide is achieved by floating freshly cut sections onto the surface of a water bath and adhering these to a clean glass slide.

Conventional microscopic images rely on chromatic contrast in order to reveal morphological structures. Paraffin wax offers very little contrast, so is poorly suited to histopathology. For this reason, samples are often dewaxed and stained with a coloring reagent which reveals specific contrast. Haemotoxylin and eosin (H&E) is a dye routinely used in histopathology, and is the most widely used stain in medical diagnostics [23]. The haemotoxylin component of the dye stains cell nuclei blue, with the eosin component staining protein rich components such as the extracellular matrix and cytoplasm pink. This gives stained tissue the desired contrast between important organelles, a critical augmentation to a histopathologist in the diagnosis of disease. Other stains are commonly used to offer differing contrast, such as toluidine blue and Masson's trichrome.

The five year survival rate of patients with oral cancer over the previous three decades has only marginally increased, with current figures between 50% and 55% for all stages [24]. This relatively poor survival rate can be in part attributed to the fact that most affected persons report symptoms to the clinic at a late stage, due to the difficulties in early recognition at a naive individual level. OSCC is frequently preceded by a spectrum of abnormalities, collectively termed *potentially pre-malignant oral epithelial lesions* (PPOELs). These vary from small, flat white patches on the tongue, to large and irregular red patches on the floor of the mouth. Particular emphasis should be put on the early detection of PPOELs, as timely diagnosis and therapy is absolutely essential in minimising the risk of malignant transformation.

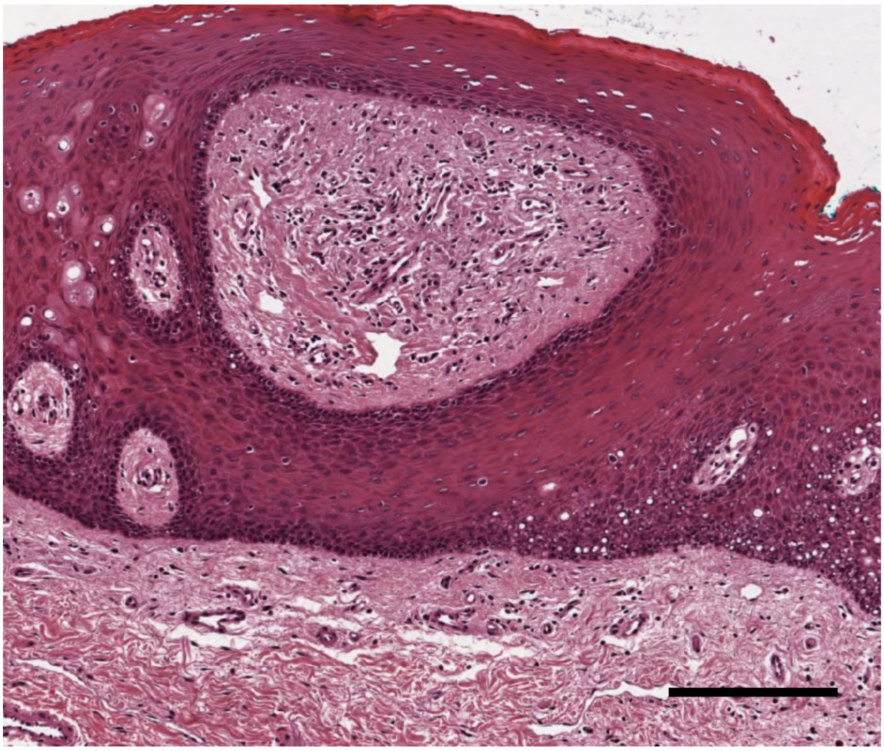


FIGURE 2.4: Microscopic image of H&E stained tissue section from dysplastic epithelium. Individual nuclei are discernible from the surrounding connective and epithelial tissue. Scale = 200 μ m.

2.2 Potentially pre-Malignant Oral Epithelial Lesions

The term ′potentially pre-malignant oral epithelial lesion′ is relatively new, and evolved from ‘oral potentially malignant disorder′ (OPMD). This evolution in terminology arises from the fact that not all of these conditions have any potential to transform into cancer, with malignant transformation rates (MTRs) ranging from as low as $\approx$ 0.1% in benign leukoplakia [25], to over 50% in some erythroplakias [26]. The current section will focus on various types of PPOELs, associated risk factors, clinical and histopathological features, modes of detection and their respective flaws.

2.2.1 Clinical Lesions

Oral Leukoplakia

Oral leukoplakia (OL) has been most recently defined as ′white plaques of questionable risk having excluded (other) known diseases or disorders that carry no increased risk for cancer′ [27]. This definition indicates that the OL diagnosis is a diagnosis of exclusion, and it should be emphasised that there are a plethora of benign conditions that share common characteristics but bear no malignant risk, making it difficult to distinguish from common inflammatory disease. OL is diagnosed in the clinic: there are no pathological criteria that can be examined by microscopy, although biopsy and histopathological assessment is often recommended to probe for associated disorders. OL's account for 95% of histopathologically defined dysplastic lesions [28].

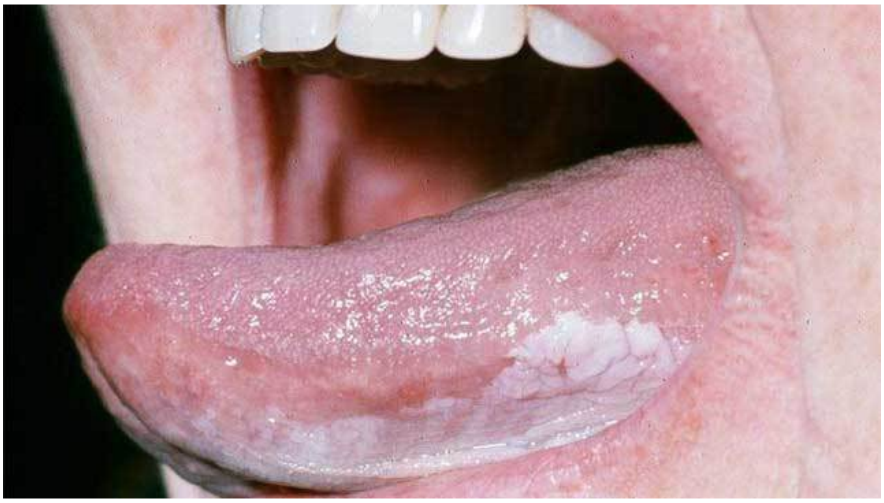


FIGURE 2.5: Oral leukoplakia of the lateral tongue. OL is characterised by a flat, homogeneous white appearance.

Common sites of OL vary by region and culture. In western cultures, lesions are frequently identified in the floor of mouth and on the lateral tongue due to carcinogen pooling from smoking and alcohol, whereas in Asian populations lesions in the buccal lining (inner cheek) are more abundant [29], attributed to the chewing of betel quid. A study in Spain in 2010 [30] aimed to correlate clinical and pathological diagnosis of OL from a patient cohort (n=54) spanning different age groups, genders, sites and appearances. The male to female ratio of 1.45:1 (32 vs 22) implies that OL is slightly more prevalent in males, potentially attributed to higher levels of smoking in males. They also found higher occurrence of OL with a homogeneous white appearance, and those situated on the lateral aspect of the tongue, with the latter being associated with a higher risk of malignant transformation.

Another study by Kuribashi *et al* [31] implemented a long term 'wait and see' policy to try and identify important factors in the development of OL. They segmented a total of 237 lesions from 218 patients into five distinct groups: unchanged, reduced, disappeared, expanded and malignantly transformed. From long-term follow-up between 2001-2010, over half of the set (57.0%) remained unchanged, 12.7% reduced in size, 18.6% had disappeared and 7.2% had clinically deteriorated or spread. The remaining 11 lesions (4.6%) had developed OSCC. They found that non-homogeneous appearance and tongue-localisation were linked to transformation.

A recent systematic review [32] aimed to ascertain the MTR of OL and related risk factors. Its findings showed a vast range of reported MTRs, between 0.13% and 34.0%, with a mean of 3.5%. It is important to note that the systematic review is of retrospective studies, which carry potential sources of bias, necessitating carefully designed prospective studies.

Oral Erythroplakia

Oral erythroplakia (OE) can be characterised as 'a fiery red patch that cannot be characterised clinically as any other definable disease'. They are often situated in close proximity to leukoplakia lesions, but have much higher potential to be malignant, with MTRs ranging from 14% to 50% [33]. OE prevalence ranges from 0.02% and 0.83%, mainly in older age groups and men. The floor of the mouth, soft palate and buccal mucosa are reported as the most common sites [34]. The relatively high MTR of OE necessitates surgical excision and close monitoring for patients who harbour the condition.

There is poor understanding of the causes and mechanisms of OE progression. Tobacco and alcohol consumption remain important carcinogenic risk factors [35].

2.2.2 Oral Epithelial Dysplasia

The aforementioned PPOELs (OL, OE) are clinically diagnosed conditions. This means they are detected at the point of primary care based on reported symptoms, visual appearance and associated patient health records. They are both diagnoses of exclusion, whereby a decision is informed by the absence of other possible conditions. The clinician may deem it necessary to refer the patient for histopathological assessment. The histopathologist will first determine whether there is any cancer amongst the tissue, making a diagnosis at this point if necessary. In the absence of full-blown cancer, the histopathologist will examine in and around the epithelial layer to inspect the physical appearance of the squamous cells.

Oral epithelial dysplasia (OED) is a histopathologically defined lesion, based on the fulfilment of a range of architectural and cytological criteria [36]. The WHO defines OED as *'a precancerous lesion of stratified squamous epithelium characterized by cellular atypia and loss of maturation and stratification short of carcinoma in situ'*. The presence of dysplastic cells within the epithelial layer is believed to be associated with a likely progression to cancer [37], with reported MTR's in the range of 1.4 - 36% for all grades [38], [39], however normal epithelial cells can bypass the dysplastic phase and progress to OSCC.

Clinical Determinants

Despite the primary role histopathological grading plays in OED diagnostics, it is important to understand how clinical factors influence the potential of malignant transformation. In this context, 'clinical factors' refers to information learned about the patient and condition at the point of primary care. It encompasses patient metadata such as age, gender, smoking-status, and physiological data such as the size, site and appearance of a lesion. Numerous studies have attributed various clinical factors to malignant transformation of PPOELs. Female gender [40], [41], tongue/floor of mouth sub-site [42]–[44], the existence of multiple accompanying lesions [45] and large size [46]–[48] are just a few risk factors that were found to be significant predictors.

The majority of these studies focus on a particular clinical lesion, such as leukoplakia, rather than histopathologically diagnosed OED. A relatively recent study by Ho *et al* [49] aimed to ascertain the clinical determinants which correlated with malignant transformation of OED. A total of 91 patients who fulfilled strict inclusion criteria were recruited for the study, and lesions were graded by two oral histopathologists. They reported an MTR of 22%, which is relatively high compared with other hospital-based studies [38], [50]. The most significant predictors in the study emerged as non-smoker status, non-homogeneous appearance and size of lesion. Of less significance was the consensus histopathological grade of the lesion. The prognostic weight attributed to histopathological grading is a controversial topic, with various conflicting reports in the literature [38], [50]–[53]. A somewhat surprising finding was that non-smokers were 7.1 times more likely to progress to cancer, which the authors speculate is linked with intrinsic factors. It is consistent with the finding that idiosyncratic leukoplakia is correlated with a higher risk of transformation [40], [53].

Diagnosis

The current convention for OED diagnosis is that lesions are assigned one of three grades: *mild*, *moderate* or *severe*, reflecting the extent to which the criteria in table 2.2 are fulfilled. In this model, OED is part of a longer progression pathway, starting at *hyperplasia* (increased number of cells) and terminating at *invasive carcinoma*, whereby the now cancerous cells depart the epithelial tissue and propagate elsewhere. Each step in the pathway is a more severe manifestation of the underlying biomolecular transitions that occur during OSCC carcinogenesis. There are numerous alternative systems for OED grading, such as the Ljubjana classification system [54], and a more recent binary adaptation of the WHO system [55].

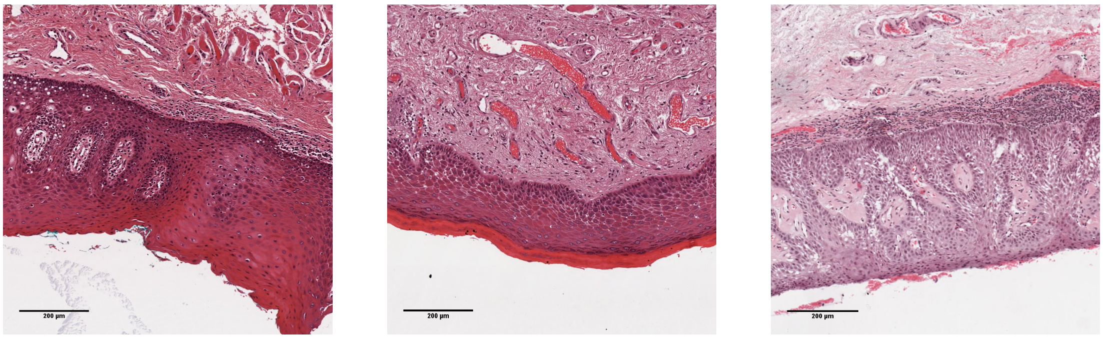
TABLE 2.2: WHO architectural and cytological criteria used for grading OED.

Architecture criteria	**Cytology criteria**
Irregular epithelial stratification	Abnormal variation in nuclear size
Loss of polarity of basal cells	Abnormal variation in nuclear shape
Drop-shaped rete ridges	Abnormal variation in cell size
Increased number of mitotic figures	Abnormal variation in cell shape
Abnormally superficial mitoses	Increased nuclear- cytoplasmic ratio
Keratin pearls within rete ridges	Increased nuclear size
Atypical mitotic figures	Hyperchromatism
Premature keratinization in single cells	Increased number and size of nucleoli

A brief description of each of the dysplastic grades in the context of table 2.2 follows.

(i) *Mild dysplasia*: Architectural disturbances limited to the lower third of the epithelium. Minimal cytological atypia.

(ii) *Moderate dysplasia*: Architectural disturbances extending into the middle third, marked atypia may indicate severe dysplasia, mildly atypical may merit mild dysplasia.



(A) Mild dysplasia (B) Moderate dysplasia (C) Severe dysplasia

FIGURE 2.6: Examples of dysplasia grading

(iii) *Severe dysplasia*: Greater than two thirds of epithelium.

Figure 2.6 shows an example of each of three different dysplasia gradings, from moderate through to severe. In Fig. 2.6a, the abnormal cells and architectural changes are confined to the lower third of the epithelium, resulting in a diagnosis of mild dysplasia. Conversely, Fig. 2.6c shows major abberations throughout the epithelium, with an abundance of cytological variation, resulting in a diagnosis of severe dysplasia. Moderate dysplasia (Fig. 2.6b lies somewhere between the two, with both cytological and architectural changes extending to the middle third of the epithelial layer.

One of the key issues with OED grading is the inter- and intra-observer variability amongst histopathogists. Since grading of OED remains the principal prognostic predictor, diagnostic disagreement is problematic. OED development is generally viewed as a continuous progression of abnormal alterations, and present grading systems like the one shown in table 2.2 attempt to quantise the condition based on subjective criteria.

While there is universal understanding of the criteria shown in table 2.2 amongst pathologists, the interpretation of the degree and significance of each of the criteria is essentially subjective, giving rise to great variability amongst observers [37], [55]. In the study by Kujan *et al* [55], moderate agreement was observed between four observers for the following criteria: increased number of mitotic figures; drop-shape rete ridges, increased nuclear size and abnormal variation in cell shape. The highest disagreement was observed for irregular epithelial stratification, loss of polarity of basal cells, abnormal variation in nuclear size, atypical mitotic figures and hyperchromatism. Interestingly, the same study showed that a cumulative scoring system resulting in a binary grade of *high* or *low* risk led to a substantial increase in agreement. This implies that the histopathologists in the study compensated for differences in their scoring of individual criteria. They hypothesised this could be due to each observer initially screening the sample and arriving at a prospective diagnosis, which would bias the eventual diagnosis. Another study by Krishnan *et al* [56] concluded that the inter-observer agreement ranged from poor to moderate for several grading systems. These findings suggest that histopathological grading is not an exact science, and would benefit from more objective tools to augment the process of diagnosing OED.

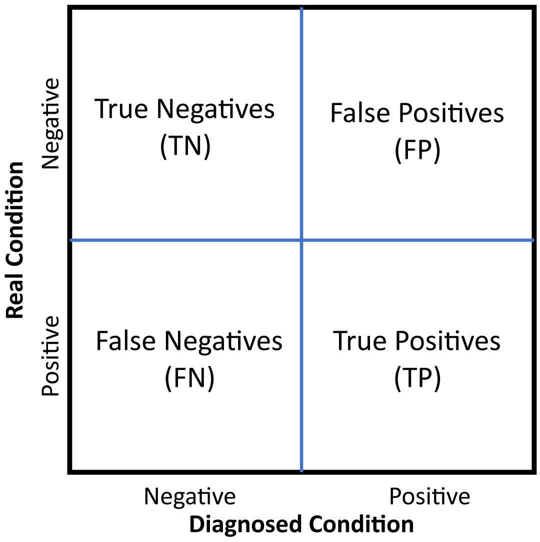
2.2.3 Testing

Performance metrics

The gold standard for diagnosis of OED and OSCC remains histopathological assessment, which carries disadvantages as previously discussed, particularly the invasive nature of extracting tissue, subjectivity of diagnosis and long waiting times. Several screening methods have been proposed to augment the clinical oral examination process in a bid to improve diagnostic accuracy and aid the workflow in an already overloaded system.

A testing or screening method should be able to accurately detect both positive (disease) and negative (healthy) cases. If many positive (disease) cases are missed, then there is potential for a high human cost - more people will have the condition undetected and will not receive necessary treatment. If many people who do not have the disease are being incorrectly diagnosed as having the disease, there is a high economic and wellbeing cost - unnecessary time and money is being spent on treating an absent disease, whilst the patient receiving treatment will experience heightened anxieties for a condition they don't have. Diagnostic performance can be quantified by several different performance metrics, each conveying a different quality of the test. Each quantity can be derived from a confusion matrix, depicted in Fig. 2.7.

The key performance metrics are summarised in Eqs. (2.1) to (2.4)



Real Condition \ Diagnosed Condition	Negative	Positive
Negative	True Negatives (TN)	False Positives (FP)
Positive	False Negatives (FN)	True Positives (TP)

FIGURE 2.7: Confusion matrix of a diagnostic test

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{2.1}$$

$$\text{specificity} = \frac{TN}{TN + FP} \tag{2.2}$$

$$\text{Positive Predictive Value} = \frac{TP}{TP + FP} \tag{2.3}$$

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN} \tag{2.4}$$

Sensitivity quantifies how well a test can diagnose a positive case. For example, a diagnostic test for oral cancer detection with high sensitivity will rarely fail to detect those who have cancer. On the other hand, specificity (or true negative rate) quantifies how well a test can correctly reject a positive diagnosis. Both are very important for disease diagnostics, and it is important to strike a balance between the two measures. A 100% sensitive test with poor specificity means that a negative case has a high probability of being diagnosed as having the condition. Conversely, a 100% specific test with poor sensitivity means that a positive case has a high probability of being incorrectly diagnosed as not having the condition.

Consider a quantitative diagnostic test which gives a Gaussian distribution of scores for both patients with a disease (positive class), and patients without the disease (negative class). The patients with disease tend to score higher, whereas healthy patients have a lower score, as shown in the distributions in Fig. 2.8. There is a region of overlap, where there are similar scores for a small subset of both positive and negative patients. By varying the decision threshold of the test, one can determine a new confusion matrix from which a unique set of performance metrics can be calculated. The receiver operating characteristic (ROC) curve is a parametric curve which displays the sensitivity and specificity at different thresholds. Figure 2.8 contains three thresholds shown on the distributions in (a), with the corresponding sensitivity and specificity shown on the ROC curve in (b).

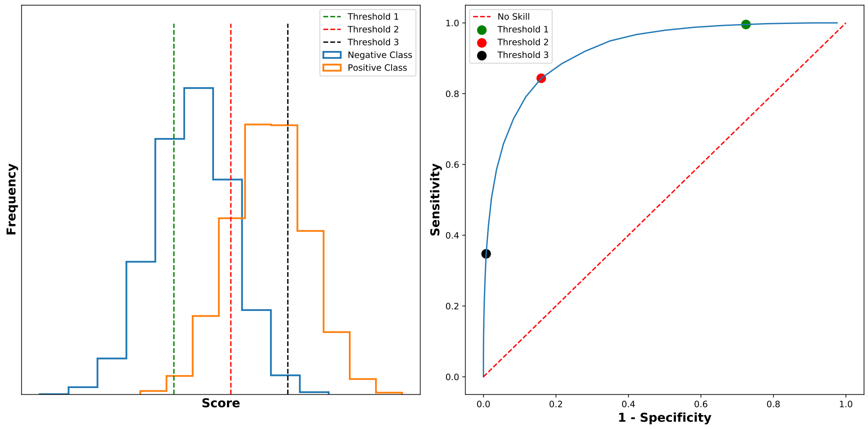


FIGURE 2.8: Simple depiction of a diagnostic test and the resulting ROC curve

Figure 2.8 demonstrates the trade off between sensitivity and specificity in a binary diagnostic test. One can enhance either metric at the expense of the other. A test with absolutely no diagnostic skill is characterised by a diagonal line bisecting the plot and intersecting the origin (red line in Fig. 2.8b). Let's say for instance, that Fig. 2.8a represents the distribution of scores for a test to identify patients at high risk of developing lung cancer. If the threshold was shifted towards the right, the number of true and false negatives will increase, with decreasing true and false positives. In such a test, the specificity will be high, compensated for by a drop in sensitivity. If the test diagnosed the patient as being high risk, there is a heightened degree of certainty of that patient actually going on to develop lung cancer. On the other hand, shifting the threshold to the left will increase the sensitivity at the expense of the specificity, producing a test where a negative diagnosis is very likely to truly be negative. Most would argue that the latter scenario in the case of risk stratification is preferred over the former, where many high risk patients will go undetected.

A hypothetical test where specificity is preferred over sensitivity is one which informs a decision on whether to perform dangerous surgery or treatment (e.g. radiotherapy, chemotherapy) on a patient. In this case it is crucial to be certain that it is necessary to carry out such a procedure. Most of the time, a good diagnostic test will favour thresholds which yield sound values for both measures.

Clinical Adjunctive Tests

There have been several promising clinical adjunctive tests proposed for the detection of OED and OSCC. The different techniques can be categorised into three groups: *optical imaging devices*, *high resolution microscopy* and *vital staining techniques*. These tools are developed with the understanding that they will aid the diagnostic process, rather than replace it.

An example of an optical imaging device used is the VELScope® (Visually Enhanced Lesion Scope) [57], which is based on the principles of autofluoresence. Autofluoresence is the emission of light from biological structures after the absorption of light, without the need for other fluorescent markers. Illumination of the mucosa with a light source emitting 400-460 nm results in green autofluorescence for normal tissue, whereas abnormal tissue absorbs the incident light. Fluorescent substances such as collagen, flavin adinine dinucleotide (FAD) and nicotinamide adenine dinucleotide (NADH) are responsible for this effect. Collagen fibre linkage and the reduction of FAD and NADH in malignant tissue relative to normal tissue reduces the emission of autofluorescence, which results in a dark patch in abnormal regions [58]. A meta-analysis of the limited number of studies associated with the technology report sensitivities ranging from 0.73-0.97 and specificities from 0.22-0.87 [59]. Other light based technologies include ViziLite® and Microlux®.

High resolution microendoscopy (HRME) is another example of a clinical adjunctive test. This technique allows for the *in vivo* acquisition of microscopic images of oral mucosa by means of an endoscopic probe and contrast agents such as acetic acid or proflavin solution. The morphological features used by histopathologists to diagnose cancer and dysplasia can then be used as a basis for diagnosis, whilst allowing for the quantification of properties such as the nucleic density and nuclear-cytoplasm ratio.

Vital staining techniques such as toluidine blue (TB) staining are amongst the most frequently reported adjunctive tests. Either by rinsing the mouth with the dye or direct staining with a swab, TB imparts a blue colour to acid rich regions of oral mucosa. Abnormal tissue retains the dye after washing with acetic acid due to the increased nucleic density (and hence relative abundance of nucleic acid), whilst normal tissue does not retain the dye. The sensitivity of this technique ranges from 0.74-0.90, with a specificity of 0.59-0.79 [59]. False positives arise predominantly as a result of pooling within natural crevices within the mouth, or uptake by inflammatory tissue or benign ulcerative conditions, whereas false negatives are primarily caused by a thick superficial keratinised layer preventing the penetration of the dye. The technique is most effective at diagnosing high grade dysplasia and carcinoma, lower grade dysplasia tends to suffer from a drop in sensitivity [60], [61].

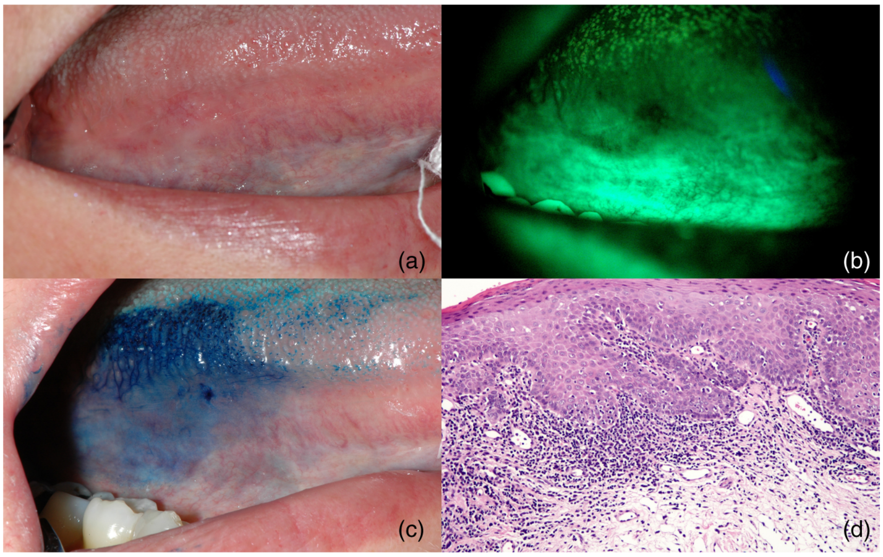


FIGURE 2.9: Velscope and Toluidine blue can locate abnormal dysplastic tissue *in vivo*. (a) Photograph of lateral tongue ROI with leukoplakia. (b) Photograph of ROI illuminated with Velscope. (c) Photograph of ROI stained with toloudine blue. (d) Microscopic image of HE stained tissue confirming the presence of dysplasia.

Despite the good sensitivities reported by the studies, the specificity of such tests is still sub-optimal. This confounds the feasibility of the described techniques to support clinical decision making, as false positives lead to unnecessary increased economic cost and anxiety in those patients. TB shows promising results in identifying high risk lesions with poor outcomes, but the clinical studies had small numbers of patients, therefore more extensive research is required before reaching meaningful conclusions. The main drawback with these techniques is that they still depend on subjective interpretation of results, they do not address the issue that demands objective and accurate automatic tests are required.

Biomarkers

The inability to accurately predict the progression of OED on the basis of histopathological grade significantly hampers the field. Despite the promising clinical adjuncts summarised in section 2.2.3, these are still fundamentally subjective, requiring interpretation from clinicians and pathologists. Some of the adjuncts, especially the light based ones, reveal the presence of PPOELs rather than actually indicate the prognosis of underlying conditions. This makes the suitable as tools to specifically locate a lesion with borders initially difficult to resolve, but not much more.

A systematic review [62] aimed to assess the use of biomarkers as prediction tool for OED progression into OSCC. It concluded that there is a lack of strong evidence for biomarker usage in OED prognosis due to a limited number of clinical studies. Nevertheless, they found that the loss of heterozygosity and allelic instability at specific loci in chromosomes significantly increases the risk of progression to cancer in the limited studies available. Cytogenic location of genes is specified using a standardized address, which first specifies the chromosome number (1-22 or X,Y), followed by a p or a q which encodes whether the gene appears on the short or long arm respectively. For example, a location of $9p$ indicates the presence of the gene on the long arm of chromosome 9. A loss of heterozygosity on $3p$, $8q$, $9p$ and $11p$ were all correlated with a relative risk of progression to cancer. Heightened DNA content was also found to be a significant predictor in OED transformation.

2.3 Summary and Motivations for Work

This chapter has summarised some of the many clinical and biological issues that surround the diagnosis and prognosis of oral cancer and PPOELs. Despite a wealth of knowledge that exists within the field, there is need for tests which can augment the flawed diagnostic workflow in a rapid, automatic and objective manner.

The motivations and objectives for this thesis are two-fold. Firstly, by exploiting objective chemical imaging techniques, the biochemistry of oral cancer and OED can be investigated and compared with the existing consensus. Secondly, a new tool for OED prognosis based on the rich information acquired by such chemical imaging modalities can be built and assessed using a framework of data processing, machine learning and optimisation. The prospective outcome of such studies will hopefully pave the way for wider studies, as well as contribute to the vast knowledge pool that already exists.

The next chapters will cover the relevant experimental and analytical techniques, with accompanying data and results to support the objectives outlined previously.

Chapter 3

Theoretical Concepts

3.1 Introduction

This chapter will focus on the theoretical principles that the techniques used in this work are built upon. Starting from the fundamentals of light-matter interaction, different modalities of infrared (IR) imaging will be described, with comparisons made to other prevalent techniques. Statistical analysis and machine learning will also be introduced, with particular focus on the pre-processing and classification of IR data. Relevant literature in the field will also be summarised.

3.2 Infrared Spectroscopy and Imaging

3.2.1 Fundamentals

Light-Matter Interactions

At the core of any light-based technique is the concept of electromagnetic (EM) waves, and the rules that govern the interaction of such waves with matter. This extensive field is termed *optics*, and is one of the most fundamental and established branches of physics, relying on both classical electromagnetism and quantum mechanical principles to explain observed phenomena.

An EM wave is composed of an oscillating electric field and magnetic field, which exist in planes perpendicular to each other. Propagating EM waves travel in a direction which is mutually orthogonal to both the electric field and magnetic field. The periodic change in polarity of the electric and magnetic vector leads to a sinusoidal wave as a function of both position and time. This periodic nature allows EM waves to defined in terms of their frequency and wavelength, with a fixed speed (in a vacuum) of approximately $c_0 = 3 \times 10^{10}\,\mathrm{m\,s^{-1}}$.

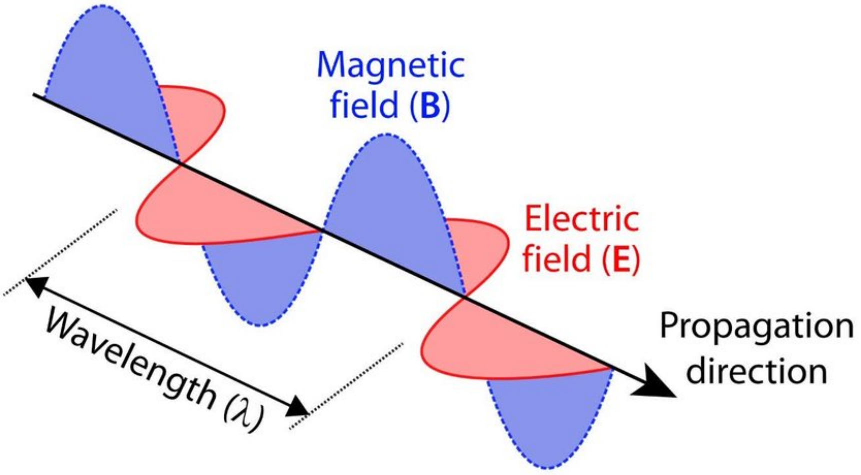


FIGURE 3.1: Basic depiction of an EM wave [63]

The behaviour of EM radiation at the boundary between media and within matter is governed by the materials complex refractive index: $\underline{n} = n + i\kappa$. The real component of this quantity (n) indicates the phase velocity of EM waves through a particular medium, whilst the imaginary component (κ) conveniently quantifies the absorption of radiation in the same medium. Equation (3.1) and Eq. (3.2) show the plane wave equation for the electric (E) and magnetic (B) fields both travelling in the z direction.

$$E(z,t) = E_0 \exp[i(\underline{k}z - \omega t)], \tag{3.1}$$

$$B(z,t) = B_0 \exp[i(\underline{k}z - \omega t)]. \tag{3.2}$$

Here,

$$\underline{k} = \frac{2\pi \underline{n}}{\lambda}. \tag{3.3}$$

Substituting $\underline{n} = n + i\kappa$, this yields:

$$E(z,t) = \exp(\frac{-2\pi\kappa z}{\lambda}) \cdot E_0 \exp[i(kz - \omega t)], \tag{3.4}$$

and similarly for the magnetic field. This proves that the imaginary component of the refractive index leads to absorption: materials with higher κ values result in a more rapid exponential decay, as implied by the first term of the product shown in Eq. (3.4). The real component of the refractive index is given by $n = c_0/c_m$, where c_m is the speed of light through the material.

The spatial frequency of an electromagnetic wave (number of complete cycles per unit distance) is expressed by a quantity called the *wavenumber*. Spectroscopists conventially define electromagnetic radiation in terms of wavenumber ($\overline{\nu}$) in units of inverse centimeters (cm^{-1}) rather than wavelength. The relationship between wavelength and wavenumber is given by Eq. (3.5).

$$\overline{\nu} = \frac{10000}{\lambda(\mu\mathrm{m})}. \tag{3.5}$$

The complex refractive index ($\overline{n}$) is a function of the wavenumber:

$$\overline{n}(\overline{\nu}) = n(\overline{\nu}) + i\kappa(\overline{\nu}). \tag{3.6}$$

This dependence leads to the observation that different wavelengths of EM radiation interact differently when they propagate through a given medium. Snell's law (Eq. (3.7)) dictates the relationship between the angle of incidence (θ_i) and the angle of refraction (θ_T) when a ray of light crosses the boundary from one medium to another. Both angles are measured anti-clockwise from the normal to the interface. Equation (3.7) is the reason why visible light disperses when it propagates through a prism.

$$\frac{n_1(\overline{\nu})}{n_2(\overline{\nu})} = \frac{\sin\theta_t}{\sin\theta_i}. \tag{3.7}$$

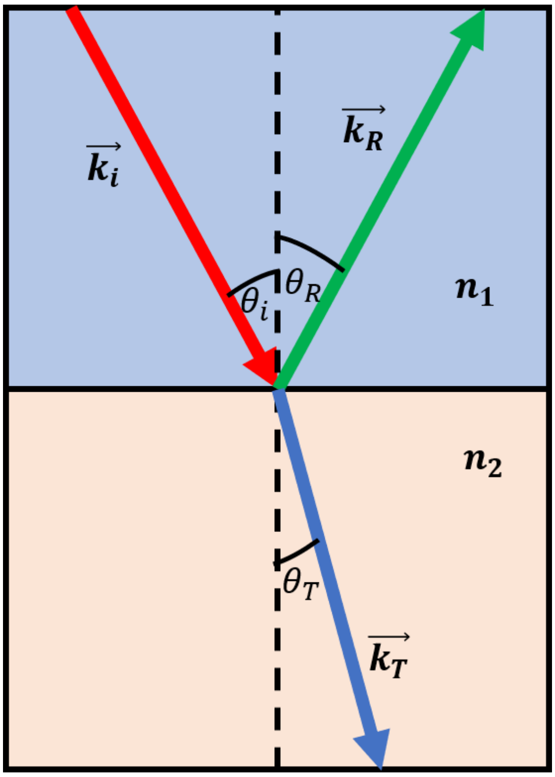


FIGURE 3.2: Behaviour of EM radiation at a boundary between two media with refractive indexes $n_1(\overline{\nu})$ and $n_2(\overline{\nu})$.

The extent to which an EM wave is absorbed in a medium is governed by the first exponential factor in Eq. (3.4). Since the intensity (I) is proportional to the square of the amplitude of the electric field, one can deduce the relationship between the fractional transmittance (T), the refractive index and the distance the wave has travelled in the medium:

$$T(\overline{\nu}, z) = \frac{I(z)}{I(0)} = \exp(-\alpha(\overline{\nu})z). \tag{3.8}$$

Here, $\alpha(\overline{\nu})$ is the *linear absorption coefficient*, and is equal to $\frac{4\pi\kappa(\overline{\nu})}{\lambda}$. The absorbance of a pure sample of thickness b is subsequently calculated by taking the logarithm to the base 10 of 1/T, giving:

$$A(\overline{\nu}) = \frac{1}{\ln 10}\alpha(\overline{\nu})b = a(\overline{\nu})b. \tag{3.9}$$

Here, $a(\overline{\nu})$ is the absorptivity of the pure sample. For a mixture consisting of N absorbing species, each with a defined absorptivity $a_i(\overline{\nu})$ and concentration c_i, Eq. (3.9) is modified to yield the *Beer-Lambert law*, given by Eq. (3.10) [64]:

$$A(\overline{\nu}) = \sum_{i=1}^{N} a_i(\overline{\nu}) b c_i. \tag{3.10}$$

This shows that the absorbance, rather than the transmittance, is directly proportional to the concentration of absorbing media. It also shows that there exists a linear relationship between absorbance and optical path length. The refractive index and associated quantities are, in essence, macroscopic representations of an ensemble of underlying microscopic mechanisms. The velocity of an EM wave changes in different media due to their differing electric susceptibilities, which cause the electrons within the material to oscillate, leading to the radiation of an EM wave with the same frequency but (usually) a different phase difference. The superposition of these secondary waves leads to a wave with the same frequency but different wavelength, hence the change in phase velocity conveyed by the real component of the refractive index.

Absorption of EM light occurs due to energy transitions within atoms and molecules within the material. Visible light is absorbed through a process by which a quantum of energy corresponding to an electrons energy level within an atom is absorbed, exciting the electron and causing it to vibrate, imparting thermal energy into its surroundings. For lower energy EM radiation, a single quantum of energy is not enough to excite an electron. However, lower energy shifts, such as those associated with molecular vibrations and rotations, correspond to the energy of quanta from the IR and microwave region of the EM spectrum. The interaction between IR radiation and molecular vibrations can be exploited to offer a rich source of chemical information.

Molecular Vibrations

The number of ways a molecule with N atoms is allowed to move is equal to $3N$. There are three translational and three rotational degrees of freedom in 3D space. The remaining $3N - 6$ degrees of freedom represent the number of ways the atoms within the molecule are able to vibrate. In the special case of a linear molecule, there is no rotational motion about the longitudinal axis, leaving $3N - 5$ vibrational degrees of freedom. For instance a diatomic molecule (which must be linear), can only vibrate in one way, whereas a triatomic non-linear molecule, has three vibrational states. The different vibrational states available to specific molecule are also known as its *vibrational modes*.

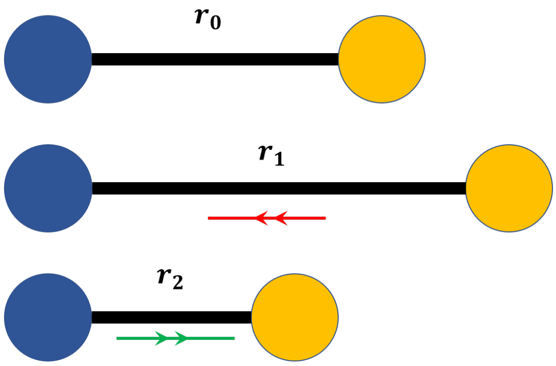


FIGURE 3.3: Stretching vibrations of a simple linear diatomic molecule. r_0 is the equilibrium inter-atomic separation, $r_{1,2}$ are the inter-atomic separations at points of maximum extension and compression respectively.

Figure 3.3 is a simple depiction of a diatomic molecule vibrating in the only mode accessible. The system of two atoms joined by a chemical bond can be modelled as a spring connecting two masses, which is assumed to follow Hooke's law (Eq. (3.11)):

$$F = -kx. \tag{3.11}$$

This law describes the relationship between a spring's displacement from equilibrium (x) and its exerted force (F). The linear relationship is scaled by a constant of proportionality called the *spring constant* (k):

In the case of Fig. 3.3, x can be calculated by subtracting the equilibrium inter-atomic separation r_0 from the inter-atomic separation. Considering molecular vibrations occur at an atomic scale, the potential energy of the system must be described quantum mechanically using Schroedingers equation, resulting in a set of discrete energy levels, dictated by Eq. (3.12):

$$V = (v + \frac{1}{2})h\nu. \tag{3.12}$$

Here, V is the potential energy, v is the vibrational quantum number (can take integer values $0, 1, ...$), h is Planck's constant and ν is the vibrational frequency. Given that a system which obeys Hooke's law exhibits simple harmonic motion when displaced from equilibrium, ν is related to the mass (m) and spring constant (k) as follows:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}}. \tag{3.13}$$

In addition to a quantised set of allowed energy levels, Eq. (3.12) implies that a molecule can never have zero energy, a fundamental characteristic of quantum mechanical systems which has its origins in Heisenberg's uncertainty principle. Figure 3.4 shows the potential energy V as a function of the inter-atomic separation r, with the allowed quantised energy levels.

The promotion of a molecule to a higher energy vibrational state can be achieved through the absorption of IR radiation. The possibility of such an event is entirely dependant on the following rules:

- A quantum of IR radiation must *exactly* match the energy difference between adjacent energy levels for the excitation to occur.
- The dipole moment (μ) associated with the molecule must have a non-zero rate of change with respect to position: $\frac{\partial \mu}{\partial r} \neq 0$.

The rule that there must be a changing dipole moment implies that homopolar diatomic molecules (identical atoms, same charge distribution) are not IR active, as the change in charge distribution from one atom always cancels the other out. When IR radiation of energy $h\nu$ interacts with the oscillating dipole of same frequency, the radiation is absorbed and the vibrational amplitude increases. Therefore the vibrational amplitude is quantised.

The simple harmonic oscillator described here is in fact only the first order approximation of a vibrational mode. The force required to compress a bond by a certain

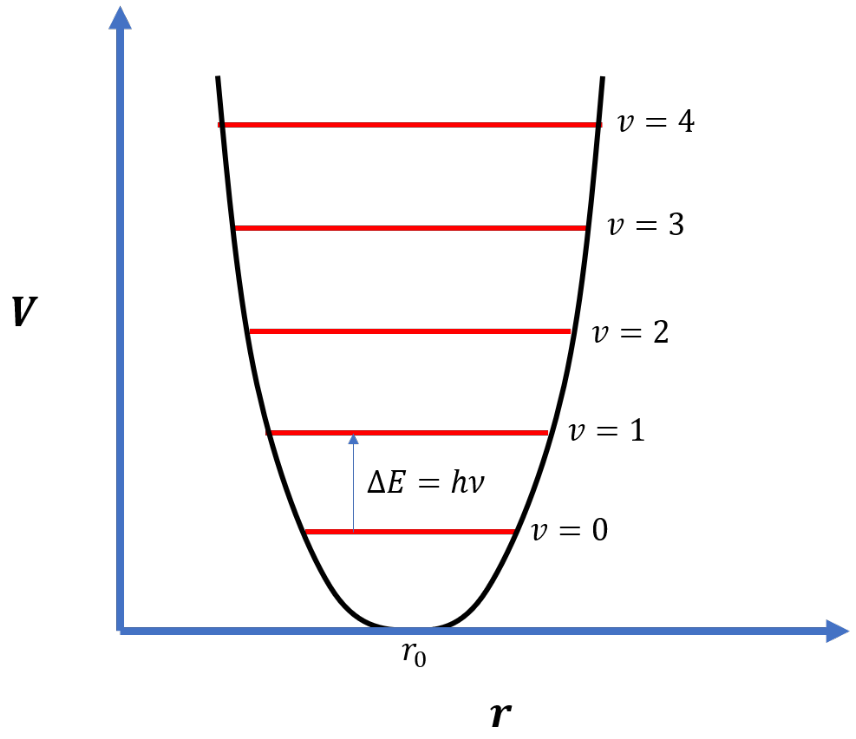


FIGURE 3.4: Harmonic oscillator potential energy (V) as a function of inter-atomic separation (r). Red lines indicate discrete energy levels allowed in a quantum harmonic oscillator

distance is actually greater than the force required to stretch it, resulting in an anharmonic potential, described by Fig. 3.5 and shown Eq. (3.14):

$$V = (v + \frac{1}{2})h\nu - (v + \frac{1}{2})^2 h\nu x. \tag{3.14}$$

In Eq. (3.14), x represents the anharmonicity constant. As r increases, the potential energy asymptotically approaches the spectroscopic dissociation energy (green dashed line), the energy at which the molecule dissociates. In contrast to a quantum harmonic oscillator, where only transitions of $\Delta\nu = \pm 1$ are allowed, molecules are able to transition to higher energy levels, with $\Delta\nu = \pm 2, 3$ amongst others allowed.

The vibrational motion of a polyatomic molecule is described by a set of vibrational modes each with a discrete frequency. In the harmonic approximation, the overall motion can be expressed as a superposition of the independent modes. For example, a non-linear triatomic molecule such as water has three vibrational modes: symmetric

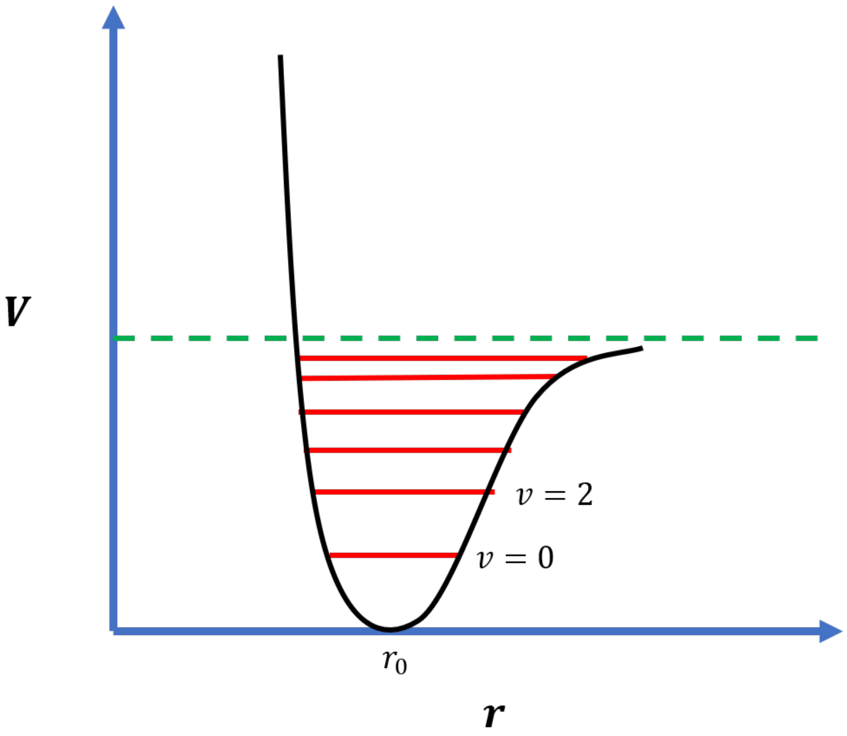


FIGURE 3.5: Anharmonic oscillator potential energy (V) as a function of inter-atomic separation (r).

stretching, asymmetric stretching and bending, each with their own characteristic fundamental frequency. Another example, which is the focus of much of the discussion in this thesis, is the amide group of vibrations. These are vibrations about an amide functional group, shown in Fig. 3.6.

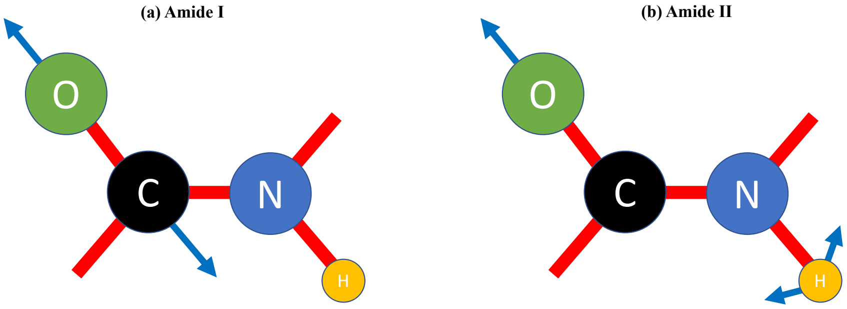


FIGURE 3.6: The amide I (a) and amide II (b) vibrational modes. The amide I mode ($\approx$ 1650 cm^{-1}) is a stretching vibration centred on the C=O bond, and the amide II mode ($\approx$ 1550 cm^{-1}) is a bending vibration of the hydrogen atom.

The fact that every molecule has a unique dipole structure and mass, resulting in a characteristic set of vibrational modes, allows IR active molecules to be identified based on the frequencies of IR it absorbs. This information can be acquired from a family of techniques called vibrational spectroscopy. Fourier Transform IR (FTIR) spectroscopy and Raman spectroscopy are the two most established techniques in the field, and serve to complement each other on many levels. The work described in this thesis utilises FTIR spectroscopy due to the ease of imaging, which will be discussed in section 3.2.3.

3.2.2 Fourier Transform Infrared Spectroscopy

Spectroscopy is a broad field concerned with obtaining the frequency dependence of how EM radiation interacts with or is emitted from matter. As discussed in section 3.2.1, molecular vibrations can be probed using a family of techniques termed vibrational spectroscopy. FTIR spectroscopy is a technique which is able to simultaneously acquire spectral data spanning a wide spectral range, allowing for the relatively rapid acquisition of a sample's IR spectrum. This section will introduce the theoretical principles surrounding FTIR spectroscopy, before discussing its adaptations and uses in the biochemical sciences.

Sources

A resistively heated silicon carbide rod, commercially termed a '*globar*' is used in most modern FTIR systems. A globar can be approximated as a Planck radiator or blackbody, which has an emission spectrum at temperature T governed by Planck's equation (Eq. (3.15)):

$$U_\nu(T) = \epsilon(\nu) \cdot \frac{C_1 \nu^3}{\exp(C_2 \nu / T) - 1}, \tag{3.15}$$

where

$$C_1 = 1.19 \times 10^{-12} \quad \mathrm{W\,cm^{-2}} \cdot \mathrm{sr} \cdot (\mathrm{cm}^{-1})^4$$

and

$$C_2 = 1.439 \quad \mathrm{K} \cdot \mathrm{cm}$$

.

The standard operating temperature for a globar is $\approx$ 1300K. Since no material can be described as a perfect blackbody, the emission spectrum is scaled by the frequency dependant emissivity ($\epsilon(\nu)$) of the object, which is between 0.83 and 0.85 for a Globar source [65].

Equation (3.15) implies that hotter sources (leading to higher spectral density, $U_\nu(T)$) would be ideal candidates for an FTIR system because the signal-to-noise-ratio (SNR) is a critical factor in FTIR spectroscopy. However, there are other factors such as source size, stability and emissivity that must be considered. For instance, a tungsten lamp filament can reach temperatures of up to 3000 K, but needs to be sealed within a glass vacuum chamber, which has low transmission in the mid-IR (400 - 4000 cm^{-1}). Nernst glowers are another commonly used source which have a superior spectral energy density to globars, however the emissivity above 2000 cm^{-1} is poor, leading to high SNRs at high wavenumbers. Another consideration is the size of the source, especially for microscopic applications. Sources would ideally match the size of the sample to maximise the flux through the sample.

Interferometer

Conventional spectroscopic systems vary the wavelength of light by means of a dispersive prism or tuneable laser. The obvious drawback with this approach is the time taken to acquire a large enough signal-to-noise-ratio at each wavelength, in addition to the time taken to change the wavelength. FTIR spectroscopy overcomes this obstacle by decomposing the broadband radiation into a time domain signal by means of an inter ferometer, before transmitting this through a sample and applying a Fourier transform (FT) to resolve the signal in the frequency domain.

The interferometer utilised in FTIR systems is based on a Michelson interferometer (MI). A simple MI divides a beam into two orthogonal rays, one ray is incident on a mirror of fixed position, the other on a mirror moving along the axis of the beam at a known velocity. The introduced path difference between the rays as they recombine creates a condition where constructive or destructive interference can occur, generating a signal as a function of path length, which is a time domain signal.

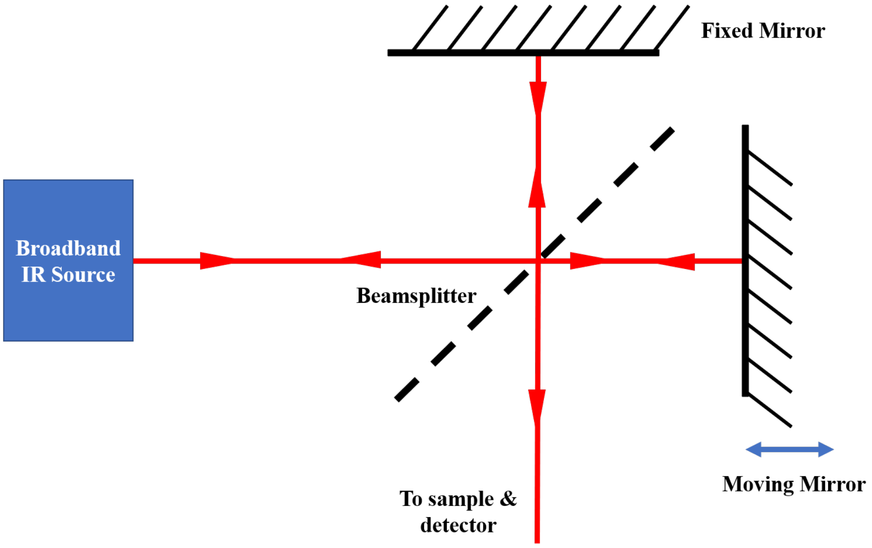


FIGURE 3.7: A schematic of a Michelson Interferometer.

The basic principles of an MI can be conveyed using a monochromatic narrow beam as an example. A source emitting monochromatic light at intensity $I(\overline{\nu_0})$ is split into two orthogonal rays, with the intensity shared equally across the two beams (for a perfect beamsplitter). If the two mirrors are equidistant from the beamsplitter, there is no introduced path difference (δ) when the beams recombine. Considering the fact that phase changes introduced by reflections from the beamsplitter and mirrors will the same for both rays, and therefore have no effect, the two rays will constructively interfere when δ = 0.

If the moving mirror is displaced by $\lambda_0/4$, where λ_0 is the wavelength, the path difference between the two beams when they recombine will be $\lambda_0/2$, corresponding to a phase difference of π radians. The rays are said to be in anti-phase with each other and destructively interfere, resulting in zero signal propagating towards the detector. If the mirror is moved again by $\lambda_0/4$, constructive interference would occur. This sinusoidal dependence of the detected intensity on the path difference can be expressed in Eq. (3.16):

$$I'(\delta) = 0.5I(\overline{\nu_0})(1 + \cos(2\pi\overline{\nu_0}\delta)), \tag{3.16}$$

where the AC term is known as an interferogram. If the mirror moves at a constant velocity v, the path difference can be expressed as $\delta = 2vt$. Substituting for δ in the interferogram gives Eq. (3.18):

$$I'(t) = 0.5I(\overline{\nu_0})\cos(2\pi\overline{\nu_0} \cdot 2vt). \tag{3.17}$$

In order to acquire the spectral intensity variation $I(\nu)$, an FT can be applied to the time domain function $I'(t)$:

$$I(\nu) = \int_{-\infty}^{+\infty} I'(t)\exp(-2\pi i\nu t) \quad dt, \tag{3.18}$$

where $\nu = 2v\overline{\nu}$ is the Fourier frequency.

For the case of broadband radiation, the observed interferogram is the sum of the interferograms that result from each individual wavenumber. The capability to simultaneously acquire data across the entire spectrum increases efficiency and signal to noise ratio compared to conventional dispersive instruments.

Detectors

A sensitive and low noise method to detect IR photons is an integral part to any FTIR spectrometer. IR detectors can be sub-categorised into two groups: *thermal detectors* and *quantum detectors*. A thermal detector operates on the principle that materials absorbing IR radiation experience an increase in temperature. This change in temperature can be indirectly detected through, for example, the change in resistance of a conductor or semiconductor. Thermal detectors are not well suited to FTIR spectroscopy due to the relatively large response time ($\approx 10^{-3}$ s) [64], far too long for the high frequencies within the interferogram.

Quantum detectors offer a highly tuneable and sensitive approach to detecting IR radiation. They are named as such due to the quantum interactions between IR photons and electrons within the detector, resulting in a measurable electric signal. For example, photomultiplier tubes (PMTs) work on the principle of photoemission, which occurs when a photon imparts enough energy to an electron to overcome the workfunction of the material. The liberated electrons then flow through a vacuum towards an anode to generate a current. The work-function of photoemissive materials is much higher than the energy of an IR quantum, therefore the use of PMTs and other photoemissive techniques is prohibited for IR detection.

The tuneable and sensitive nature of semiconductor detectors make them ideal for the detection of IR radiation. The defining feature of a semiconductor is the existence of a relatively small energy gap between the valence and conduction bands of the material. This is in contrast to an insulator which has a large energy gap (typically $>$ 5 eV), and to a conductor, where there is significant overlap between the valence and conduction bands, allowing for the free movement of charge carriers which give rise to an electric current. By choosing a semiconductor with a band gap similar to the energy of that of the radiation that is to be detected, an incident photon promotes an electron from the valence band to the conduction band.

The detectors generally used in the detection of IR radiation are $p - n$ junction semiconductors. These consist of two adjoining semiconductors, one with a relative abundance of electrons (n-type), the other with a relative deficiency of electrons, or abundance of holes (p-type). Electrons and holes migrate across the interface between the two semiconductors until the built up charge prevents any further diffusion. The resulting depleted region near the interface consists of recombined electrons and holes, whereby electrons exist predominantly in the valance band, preventing conduction. The depleted region is maintained by applying a reverse bias voltage across the semiconductor. When a photon is incident on the detector, valence band electrons are promoted to the conduction band and are attracted towards the positive terminal, and vice-versa for the generated electron holes. They are prevented from recombining by the reverse bias voltage.

Mercury cadmium telluride (MCT) is an alloy consisting of mercury telluride (HgTe) and cadmium telluride (CdTe). HgTe is a semi-metal with an overlap between the conduction and valence band, whilst CdTe is a semiconductor with a native band gap of approximately 1.5 eV. Careful selection of the relative compositions in the alloy allows for the continuous tuning of the band gap between -0.3 eV and 1.6 eV [66]. Due to the relatively low energy required to excite electrons in MCT detectors, cryogenic cooling is generally used to minimise thermal excitation of electrons.

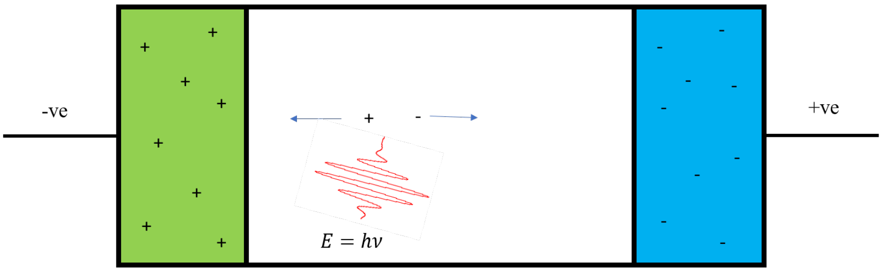


FIGURE 3.8: Semiconductor detector schematic. When a photon with an energy greater than bandgap of the semiconductor is incident on the detector, an electron is excited to the conduction band, leaving a hole in the valence band. Movement of these charge carriers leads to a measurable electric current.

3.2.3 Microscopy and Imaging

The core technique of FTIR spectroscopy as described is a powerful tool for analysing the chemical composition of bulk, homogeneous materials. On the other hand, it is not ideally suited to the analysis of heterogeneous samples, especially those which vary at a microscopic level, as the output data is an indication of the absorption characteristics integrated over a large volume. FTIR microspectroscopy (FTIR-MS) is the combination of FTIR spectroscopy and conventional microscopy. The utilisation of FTIR-MS enables the analyst to acquire spatially located spectral data by exploiting conventional microscope optics and array detector technology.

FTIR-MS can be achieved by making a few essential modifications and additions to the fundamental FTIR spectrometer configuration. The major differences are in the light propagation and detection mode. A broadband source coupled to an interferometer is initially used to generate an interferogram, and the IR light is then focused onto the sample. In a manner specific to the mode of operation, the transmitted signal is detected by an array of single element detectors called a focal plane array (FPA).

Figure 3.9 is a basic representation of the how modulated light from the interferometer is propagated through an FTIR microscope operating in transmission mode. Light from the interferometer reflects off a hyperbolic secondary mirror which distributes the light onto a primary parabolic mirror. The parabolic mirror then focuses the light

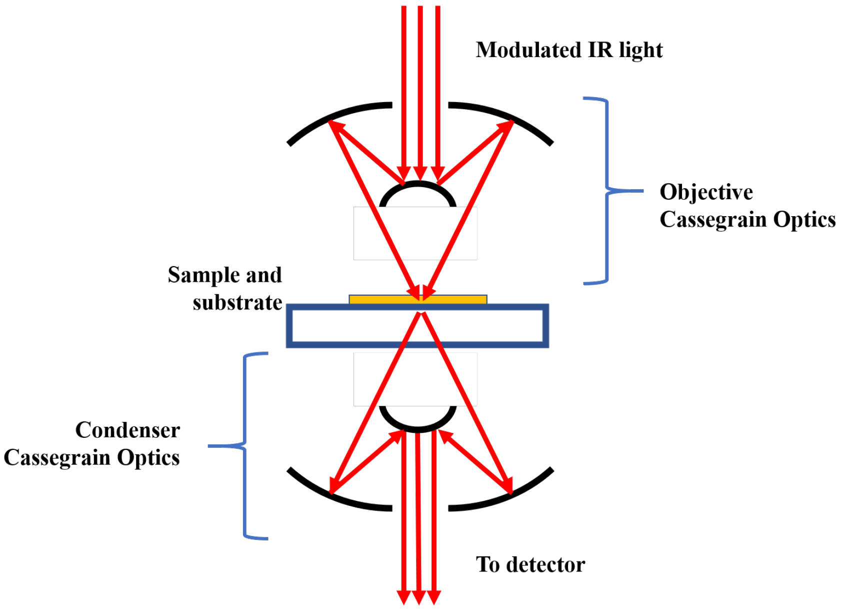


FIGURE 3.9: Basic schematic of an FTIR microscope in transmission mode.

onto the sample. In the case of transmission FTIR, the transmitted signal is propagated through another mirror configuration in order to collimate the light for detection. Reflective optics are generally preferred to refractive optics (lenses) as glass has poor transmission for IR wavelengths, and IR transmissive lenses forged from materials such as calcium fluoride are expensive. Another consideration with a refractive optical system is chromatic abberation, a phenomenon whereby the wavelength dependent refractive index leads to dispersion according to Snell's law (Eq. (3.7)) and therefore different wavelengths of radiation will have dissimilar foci.

Conventional far-field optics places a fundamental limit on the spatial resolution one can obtain from an optical system. A point source of light cannot produce an equivalent point in the resulting image due to the diffraction pattern observed in optical systems with a finite diameter, such as lenses, mirrors and apertures. This diffraction pattern is known as an Airy pattern [67]. If the Airy patterns from two sources are too close, they coincide with each other and the sources are no longer resolvable. The Rayleigh criterion [68] places a theoretical limit on the separation between two points before their Airy patterns are indiscernible, defined by:

$$\Delta r = \frac{1.22\lambda}{2 \cdot \mathrm{NA}}. \tag{3.19}$$

Here, Δr is the spatial resolution, λ is the wavelength, and NA is the numerical aperture, defined as $\mathrm{NA} = n \sin(\alpha)$, where n is the refractive index and α is the half angle of the cone of light emerging from or entering the optical system. In the case of FTIR imaging in the mid-IR spectral region (2.5 µm - 10 µm), Rayleigh's criterion would impose a spatial resolution limit of *absolutely* no less than $\approx$ 5 µm. The diffraction limit can be overcome by exploiting near-field characteristics of light, which will be further discussed in section 3.4.

Hyperspectral Imaging

FPA technology has enabled FTIR-MS to flourish. At its inception, apertures would be used to specifically restrict the beam to a small region of the sample, so that only a relatively small fraction of energy was reaching the sample, which results in a low SNR. In order to increase the SNR, multiple readings would usually be co-added, however this approach can only ever increase the SNR by a factor of $\sqrt{N}$, where N is the number of readouts at the detector. To add to this, formation of an image would require the raster scanning of the sample stage in order to focus the light on different regions of the sample. Both of these factors dramatically limit the speed of the instrument. Use of FPAs has allowed this liitation to be overcome. An FPA is essentially a two-dimensional array of single element detectors situated in the focal plane of the objective. Each detector corresponds to a pixel in the resultant image. This eliminates the requirement to guide the beam using an aperture, as the entire field of view (FoV) can be simultaneously imaged without the need to raster scan the stage. The capability to simultaneously acquire both spectral *and* spatial data provides a very high-throughput chemical imaging instrument, and is termed hyper-spectral imaging (HSI).

The data format emerging from an HSI instrument consists of two spatial dimensions and one spectral dimension. This data, commonly referred to as a hyperspectral datacube, or *hypercube*, can be thought of as a sequence of planes, each one encoding the absorption profile at a particular resolution element in the spectrum. The data can be alternatively but equivalently visualised as a 2D arrangement of one dimensional spectral vectors.

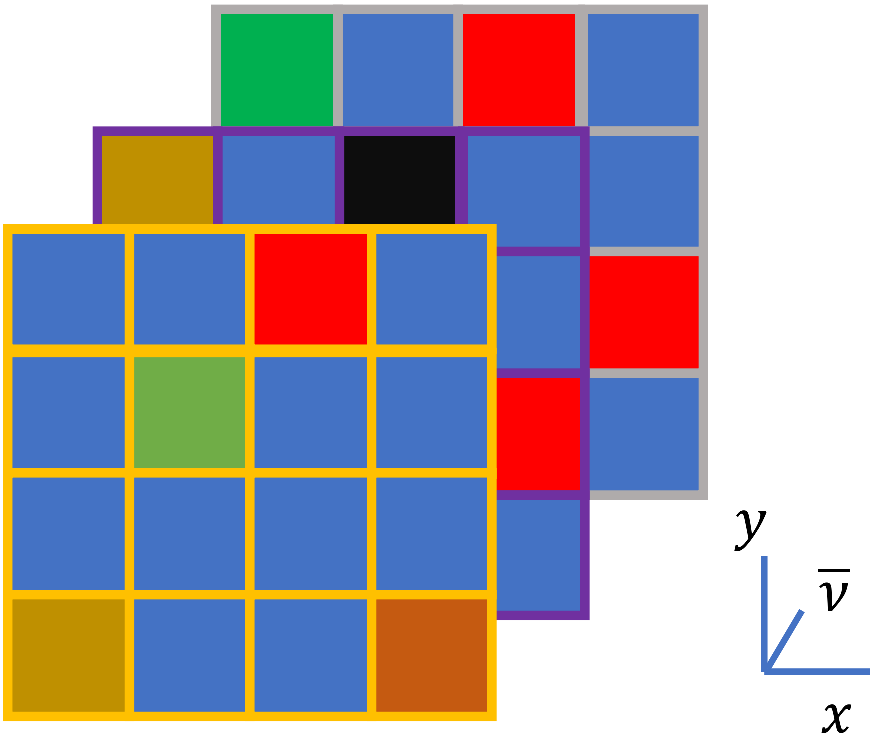


FIGURE 3.10: Depiction of hyperspectral datacube resulting from a 4×4 FPA. Each slice in the data is a unique array of values corresponding to the absorbance at a particular wavenumber.

Most FPAs have either 64×64 or 128×128 elements. If an FTIR microscope with an FoV of 1 mm^2 came equipped with a 128×128 element FPA, the pixel size in the image would be approximately 7.8 µm. There is a trade off between pixel size and SNR, the smaller the pixel element, the more signal is required to achieve reasonable SNR, hence a brighter source or longer acquisition times are required. Brighter sources such as quantum cascade lasers (QCLs) and synchrotrons can be used to address this compromise. The resolution limit imposed by Eq. (3.19) should also be considered, as there is little benefit in oversampling with pixels that have dimensions smaller than the spatial resolution of the instrument.

The size of the FPA, spectral range and spectral resolution are the factors which determine the size of the hypercube. For instance, an FTIR imaging instrument with a 128×128 FPA detector, and a spectrometer able to record data between 1000 cm^{-1}

and 4000 cm^{-1} at a a spectral resolution of 2 cm^{-1} will yield a hypercube with dimensions 128×128×1500, amounting to an array with in excess of 20 million elements. This abundance of high dimensional data often necessitates the utilisation of sophisticated visualisation tools and pattern recognition algorithms to reveal important features within the data. This will be covered in much more detail in section 3.3.2.

3.2.4 Application to Cancer Diagnostics and Characterisation

It has long been realised that there is high potential for the application of vibrational spectroscopic techniques in the biomedical paradigm. Of particular emphasis is the exploitation of the high-throughput and objective nature of FTIR-MS in cancer diagnostics. Since as early as 1952 [69], infrared spectroscopy has been explored as a tool to characterise tissue and cells that are in an altered, pathological state. There have been numerous studies related to the characterisation of specific cancers, such as colorectal [70], prostate [71], lung [72] and oral cancers [73].

Broadly speaking, biological specimens all have a very similar chemical structure at the scale of the spatial resolution attainable with FTIR-MS. The infrared spectra associated with different tissue and cells will therefore contain contributions from similar regions of the spectrum. Figure 3.11 shows a typical biological spectrum, with important peaks identified and categorised depending on the biomolecule with which they are associated.

The two dominant peaks centered at 1650 cm^{-1} and 1550 cm^{-1} are called the amide I and amide II bands respectively. These are both sum peaks of underlying contributions that arise from proteins exhibiting varying structural conformations. The amide I peak predominantly consists of C=O and C-N stretching vibrations. The amide II peak is built up of N-H bending, C-N stretching and C-C stretching vibrations. The peak position is sensitive to the backbone conformation of the protein itself, with protein structures such as alpha (α) helices, beta (β) sheets and random-coils giving rise to subtly different peak shapes. The presence of proteins in all biological tissue means that these two dominant peaks are an important feature of the architecture of an IR spectrum derived from biological material. Hydrocarbon groups such as the different vibrating modes of CH_2, are also commonplace in biological spectra due to the

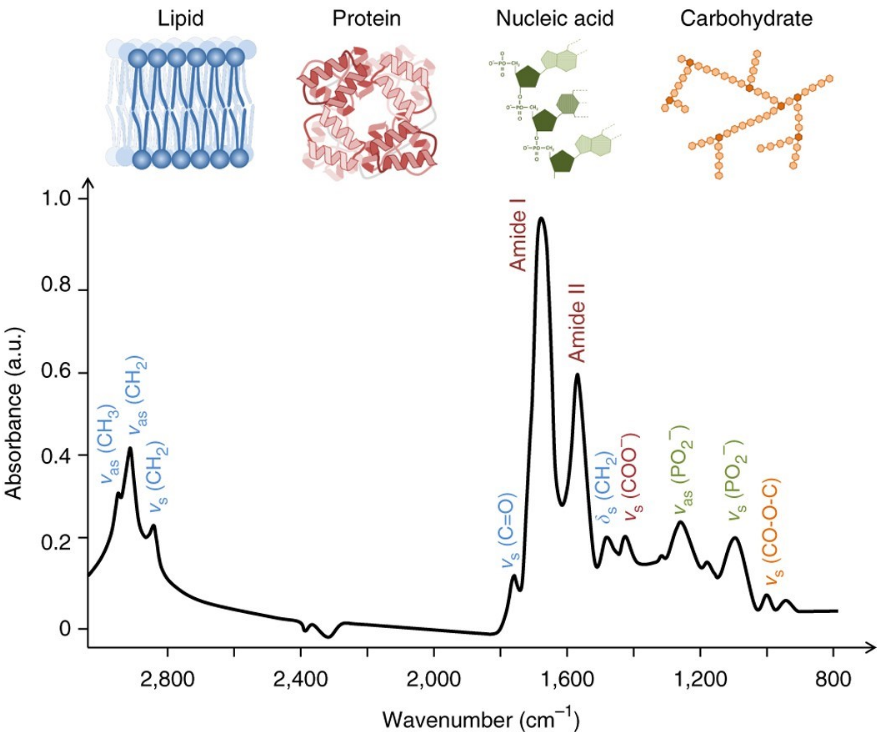


FIGURE 3.11: Typical biological spectrum and corresponding peak assignments. Type of vibrations signified by a v (stretching), δ (bending), s (symmetric) and as (asymmetric). Extracted with permission from [74]. Colour of peak label indicates group of biomolecules each spectral biomarker is related to (lipid: blue, protein: red, nucleic acid: green, carbohydrate: yellow.)

abundance of lipids in biological material. Peaks arising from the symmetric and asymmetric stretching modes of phosphate groups indicate the presence of DNA, and vibrations such as C-O-H bending are abundant in glycogen rich samples [75].

This abundance of chemical information is somewhat of a double edged sword. The complexity of biological tissue translates to a complicated convoluted set of spectral profiles that may differ only very subtly from one spectrum to the next. For this reason, it is usually necessary to employ sophisticated analytical methods in order to extract the useful biochemical information from a biological spectroscopic dataset (section 3.3.2).

Despite the undeniable promise of what is termed by many as 'clinical spectroscopy', there is still a long road to traverse before its introduction into a clinical setting, with multiple bottlenecks. The abundance of instrumentation and analytical methods available has led to a significant lack of consensus and standardisation of techniques. For a tool to be implemented in the clinic, large scale, multi-centre validation studies must be carried out in order to verify the performance and robustness of such an approach. The application of FTIR-MS coupled with rigorous analytical pipelines to samples derived from patients exhibiting OSCC and OED is the basis of this work. A more specific summary of the work surrounding clinical spectroscopy for oral cancer can be found in the experimental chapters of this thesis.

3.3 Analytical Techniques

Various analytical steps required to extract meaningful information from raw spectral data derived from biological samples. This workflow can be thought of as a pipeline, or sequence of steps that are required to infer patterns and differences between spectra originating from different types of cell or tissue. The workflow can be divided into two distinct stages: pre-processing and model construction.

Pre-processing methods can be broadly defined as the steps are required to mitigate for unwanted spectral artefacts and aberrations that may conceal or warp results further down the line. They include the smoothing of noisy signals, correction for scattering artefacts, and the careful selection of the features that contain the most relevant information.

Model construction refers to the building of models using either labelled or unlabelled data. These techniques belong to a wider subset of algorithms termed *machine learning* (ML), a field which is experiencing monumental growth in a huge array of industries. These data-driven approaches make efficient use of the vast processing power and storage capabilities of modern computers in order to build complex functions without being explicitly programmed to do so. Examples of ML applications in ordinary life include spam filters, facial recognition, autonomous cars and targeted advertisement. In a broad sense, ML can be defined as algorithms that enable computers to learn rules from data so they can make predictions when they encounter new data.

3.3.1 Pre-processing

Pre-processing methods for IR spectral data can be subdivided into a number of stages that aim to mitigate a particular undesirable trait within the data. The stages most commonly applied are *smoothing*, *baseline correction*, *normalisation*, *scaling* and *feature selection*.

Denoising

Random, high frequency noise is not reflective of the true sample signal, therefore it should be reduced. Various approaches are commonly employed in order to smooth out the unwanted noise from a spectrum. Savitzky-Golay (SG) smoothing [76] is a commonly utilised method whereby a polynomial is fitted to a small window of data points. The polynomial is then evaluated at the centre of the window and the window is shifted to the next data point. This process is repeated until the window has passed through all the data, yielding a much smoother signal. Figure 3.12 shows how SG smoothing can be applied to denoise a sine wave with high frequency normally distributed noise added to it.

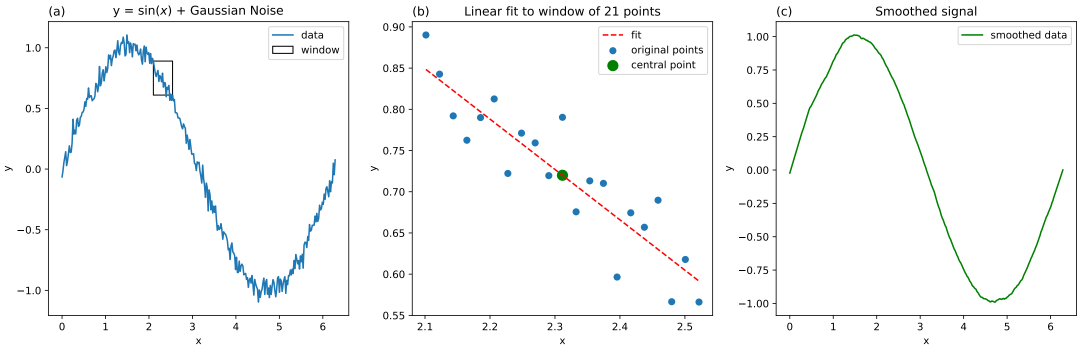


FIGURE 3.12: Basic depiction of SG smoothing. A linear polynomial is fitted to a window of 21 points. The window is propagated through the data to acquire a smoothed signal.

Another method frequently applied is principal components analysis (PCA) denoising [77]. PCA is an important technique that seeks to project the data scatter matrix onto a set of principal components (PCs) that maximise the variance within the dataset, through a process called singular value decomposition (SVD). This can be achieved with a small number of steps applying linear algebra to the data, summarised below.

The data X is contained within an $n \times p$ matrix, where n is the number of columns and p is the number of variables (wavenumbers). The first step is to mean centre this matrix by subtracting the column mean from each value, so that the mean of each column is zero. The covariance matrix C is given by Eq. (3.20):

$$\mathbf{C} = \frac{\mathbf{X}^T\mathbf{X}}{n-1} = \mathbf{VLV}^T. \tag{3.20}$$

Here, **L** is a diagonalised matrix of eigenvalues, and **V** contains the eigenvectors. SVD can be applied to obtain **L** and **VX** as follows:

$$\mathbf{X} = \mathbf{USV}^T, \tag{3.21}$$

where **U** is a unitary matrix so that $\mathbf{U}^T\mathbf{U}$ =I, and **S** is a diagonal matrix containing the singular values for the singular vectors **V**. Substitution of Eq. (3.21) into Eq. (3.20) gives:

$$C = \frac{\mathbf{VS}\overbrace{\mathbf{U}^T\mathbf{U}}^{\mathbf{I}}\mathbf{SV}^T}{n-1} = \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^T. \tag{3.22}$$

The diagonal matrix of singular values, **S**, therefore contains the eigenvalues of the covariance matrix of **X**, divided by $n-1$, where n is the number of variables. The eigenvectors are also known as the principal components, which contain the relative weight each untransformed variable contributes to the projection on the new axis. The eigenvalues (given by US in Eq. (3.21)) are also known as the scores, where each element (i, j) of the $n \times p$ matrix is the result of the linear transformation of row i of **X** onto loading j. The original data matrix can be reconstructed simply by taking the vector product $\mathbf{X} = \mathbf{TV}^T$, where **T** contains the scores.

By considering that the principal components are sorted in descending order of explained variance, a set of components that don't contribute significantly to the variance of the data can be removed in order to remove noise from the dataset. The factors that are retained will therefore describe the important variance within the data, and the data can be reconstructed on the basis of these principal components.

High frequency noise contributes negligibly to the total variance of a large, complex dataset. By applying PCA and discarding the PCs that have a negligible contribution to the total variance of the dataset, one can reconstruct the dataset with less noise.

Baseline Correction

Background absorption interferences such as scattering often manifest in a non-zero baseline for FTIR spectra. This baseline may lead to the misinterpretation of spectral differences, as the relative peak heights and peak positions will be offset by an arbitrary amount. There are several ways in which this can be mitigated. Rubber-band baseline correction fits and subtracts a convex polynomial (rubber-band) to the minima of peaks, aligning the spectrum to a mutual baseline. Spectral derivatives are also frequently used, as these intrinsically remove any non-wavenumber dependant baseline aberrations.

The baseline induced by scattering interferences is not always quantifiable using a simple approach such as rubber-band or spectral derivatives. The theory of Mie scattering was developed by Gustav Mie in 1908 [78], and it refers to the scattering of light by a homogeneous spherical particle. If the wavelength of the incident radiation is orders of magnitude different to the dimensions of the scattering particle, there exists accurate and simple approximations which describe the scattered light. However, for radiation scattered by objects with dimensions similar to the wavelength, the effects are more pronounced and require a more rigorous approach [79].

The dimensions of cells and other tissue components are of the same order of magnitude as the wavelengths used here ($\approx 10^{-6}$ m - 10^{-5} m), which unfortunately leads to an ideal scenario for Mie scattering. This is often manifest as a broad, sinusoidally undulating baseline, with dispersive artefacts towards the higher wavenumber (lower wavelength) end of the spectrum. Correction for the undulating baseline is achievable using conventional approaches such as rubber-band, however this does not account for the induced shift in peak position and shape artefacts.

The scattering efficiency of a non-absorbing dielectric sphere illuminated by radiation of wavelength λ can be approximated by:

$$Q = 2 - \left(\frac{4}{\rho}\right)\sin(\rho) + \left(\frac{4}{\rho^2}\right)[1 - \cos(\rho)], \tag{3.23}$$

where $\rho = 4\pi d(n-1)/\lambda$.

The first implementation of a Mie scatter correction was by Romeo and Diem [80], who fitted a scattering efficiency curve $Q(\nu)$ to the IR profile of a single biological cell. Kohler *et al* [81] integrated this approach into their extended multiplicative scattering correction (EMSC) in order to account for a range of refractive indexes and scattering diameters. Whilst the approach successfully mitigated for the undulating baseline, the distortion in peak position was still pronounced.

In order to understand the scattering mechanisms in biological samples, Bassan *et al* utilised polymethyl methacrylate (PMMA) micro-spheres of defined refractive index and size to model the scattering caused by cells [82]. They found that isolated spheres yield highly dispersed spectra which bear little resemblance to the spectra of the same spheres in a bulk arrangement. They termed this phenomenon 'resonant Mie scattering'. The development of correction algorithms based on the newly acquired understanding followed, with a resonant term added to the EMSC model proposed by Kohler [83], [84]. Figure 3.13 shows the effects of Mie scattering and the RMieSc scatter correction applied to prostate tissue spectra.

Normalisation

The Beer-Lambert law (Eq. (3.10)) implies a linear relationship between the absorption and path length of the material. Theoretically, this means that the FTIR spectra of thicker specimens will be scaled up compared to thinner samples. In chemical analysis, where the chemical concentration of a particular moiety is of interest, the variation in spectral intensity resulting from differing path lengths is of little interest. Normalisation is a group of approaches that intend to scale each spectrum to a common quantity or range, in order to be able to directly compare spectra originating from areas with different thicknesses.

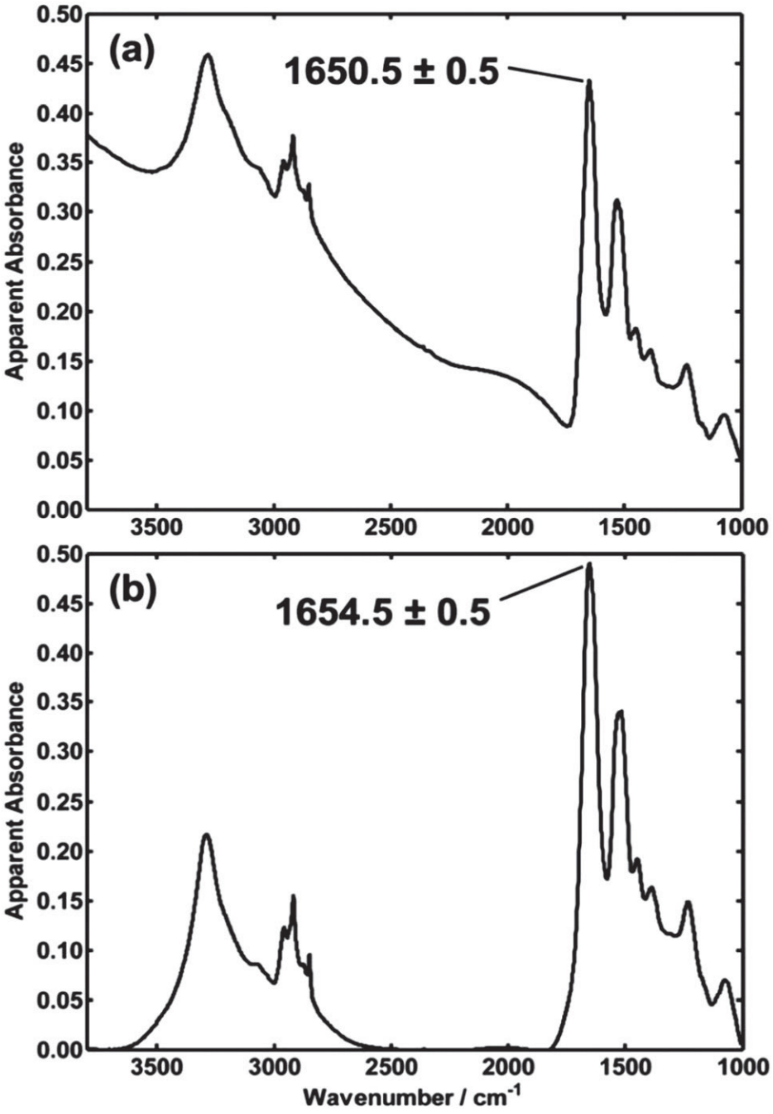


FIGURE 3.13: Example of raw (a) and Mie scattering corrected (b) spectra of prostate tissue. There is a prominent undulating baseline in (a), with peak shifts in the amide region. Reproduced with permission from [85].

There exist a few different established normalisation methods. Vector normalisation, for example, scales each spectrum by its Euclidean length. Min-max normalisation scales and offsets each spectrum so that each variable $x_i \in 0, 1$. Feature normalisation scales each variable by the intensity or area of a particular spectral region.

Feature Extraction

Identifying distinct spectral features extracted from FTIR spectra to indicate a characteristic disease is very difficult, because human tissue is composed of widely different molecular structures where overlapping of individual spectral peaks leads to formation of broader ones [6].

The selection of spectral features that characterise a particular tissue type is a difficult task because of the intrinsic complexity of biological tissue. The ensemble of overlapping peaks leads to broad peaks with a shape common to most tissue types, thus rendering the raw data difficult to interpret. Feature extraction methods serve as an important step to isolate and extract the important information from the raw dataset [86]. Feature extraction encapsulates both the selection of wavenumbers from the original dataset, and the construction of new variables in new domains for visualisation purposes. The reduction of the number of variables into a smaller subset decreases the load on model construction, reducing the risk of overfitting and increasing training speed [87].

Feature selection may entail the truncation of each spectrum to the range ($1000\,\mathrm{cm}^{-1}$, $1800\,\mathrm{cm}^{-1}$). This region contains the rich biochemical information that contains the unique spectral 'fingerprint' of the sample, leading to it being termed the *fingerprint region*. It also may include model based approaches that iteratively select the important features from a classifier, such as forward feature selection (FFS) [88].

Feature construction often refers to linear methods that generate a new set of features which are generated by a loadings matrix. Each feature is essentially a linear combination of the original input features, which in this case are the absorption at certain wavenumbers. The manner of the linear transformation depends on the technique used and whether the data is labelled or not.

PCA can be used as a powerful dimensionality reduction technique, as it can efficiently determine a much smaller set of dimensions that still describes much of the variance within the dataset. It does this without any knowledge of the origin of a spectrum (it is unlabelled), so has no inherent bias. Linear discriminant analysis (LDA) is a related technique that uses knowledge of class labels to determine a new set of dimensions, called linear discriminant vectors, onto which projected data has maximum inter-class separation and minimum intra-class variation.

3.3.2 Machine Learning

There has been a rapid emergence of ML applications to biomedical FTIR-MS datasets. The complex, multidimensional and similar nature of the data demand the deployment of sophisticated algorithms that can infer information about the data. This ideally suits the problem to ML, which has been shown to learn rules in complex datasets which would be very difficult to recognise without the mechanisms of ML.

The ML landscape is vast. Most ML algorithms can be categorised into one of two types: supervised and unsupervised learning. There are two other categories that blur the lines between the two, called semi-supervised and reinforcement learning, but only the two major types will be discussed further.

In a general sense, a set of data can be thought of as a 2D matrix ($\mathbf{X}$). Each column of the matrix represents a feature or variable within the data, each row represents a new observation, or feature vector. The size of the matrix is therefore given by $i \times j$ where i is the number of observations, or examples, and j is the number of features, or variables. In supervised learning, there is an additional variable ($\mathbf{y}$), a vector corresponding to the output of each example so that the model can attempt to formulate the function that predicts the outcome. Supervised learners can be further categorised according to the nature of $\mathbf{y}$. If $\mathbf{y}$ is a continuous variable, *regression* algorithms should be used. If $\mathbf{y}$ is a set of discrete labels indicating the identity of a data point, classification algorithms should be used. An example in the context of medical diagnostics of a problem where regression would be well suited may be the life-expectancy after a diagnosis. Similarly, a classifier in this field may binarise this variable into survived/not survived after 5 years.

In the context of FTIR-MS applied to cancer diagnostics, each row of $\mathbf{X}$ would correspond to a spectrum, with each column representing the absorption at a particular resolution element. The class membership vector **y** may then encode whether a spectrum originates from a healthy (0) or abnormal (1) region of the image. The class membership vector may not be binary; most classifiers can deal with multi-class datasets, which is especially advantageous for FTIR-MS as there may exist several different tissue or cell types within the same image.

Unsupervised Learning

Unsupervised learners identify patterns in the data without prior knowledge of the class distribution. Cluster analysis (CA) is a popular unsupervised learning algorithm, and it serves to group data into a predefined number of clusters based solely on similarities between the data. This can be especially useful for the initial visualisation of hyper-spectral images, as the process essentially flattens the third spectral dimension, giving a 2D map portraying the cluster identity at each pixel.

The k-means CA (KCA) is an algorithm which groups data together based on the Euclidean distance metric. In KCA, the user defines the number of clusters, k, for the data to be grouped in to. After this, k random feature vectors, termed *centroids* are generated, and the Euclidean distance between each observation and each centroid is computed. Observations are subsequently being grouped into the cluster with the nearest centroid. The mean of each cluster is then calculated and redefined as the new set of centroids for the next iteration. This process is repeated until cluster assignments remain unchanged from one iteration to the next [89].

The main disadvantage with unsupervised learning is that the performance of such an approach is difficult to assess. Since there is no ground truth, performance metrics such as accuracy, sensitivity and specificity are inaccessible.

Supervised Classifiers

Supervised classifiers require a target outcome variable **y** associated with each example in order to infer a function that can predict the outcome for unseen data. The majority of algorithms are model-based, which means that a classifier can be represented by a set of model parameters which are determined in the training phase of the model. The specific model parameters and the way in which they are determined are unique to the classifier that is being trained.

Take logistic regression as an example. A weighted sum of the input features (plus a bias term) is used as the argument in a sigmoid function (Eq. (3.24)):

$$S(z) = \frac{1}{e^{-z} + 1}, \tag{3.24}$$

where $z = \sum_{i=1}^{n} (w_i \cdot x_i) + b$. Equation (3.24) outputs a number between 0 and 1, which is defined as the probability of the sample belonging to the positive class. Here the model parameters are w_i and b_i, which are iteratively optimised during training.

Testing the resultant model parameters on the same data would return an overly-optimistic representation of the performance. For this reason, the data should be at the very least divided into a training set and testing set, so a more realistic evaluation of the performance can be determined. The training and testing sets are subsets of the initial dataset which should maintain a similar distribution of the data to mitigate against any biases within the data.

In addition to the model parameters, there often exists a set of parameters that influence the way in which a model is trained. These parameters, termed *hyperparameters*, differ from the model parameters in that they are not used in the final representation of the trained model. The hyperparameters must be defined by the user pre-training, but the optimal choice of hyperparameters is often ambiguous and ill-defined [9], and may require tuning, either by exhaustively trialling every permutation of hyperparameters or by using a more efficient implementation such as random search or Bayes search for higher dimensional problems [90].

The optimisation of hyperparameters should be performed on a dataset kept separate from the testing data, to ensure that the selection of hyperparameters isn't biasing the training. For this reason, the training set is often further subdivided into a training and validation set. This concept will be explained further in chapter 5.

A typical workflow for supervised ML experiments is shown in Fig. 3.14.

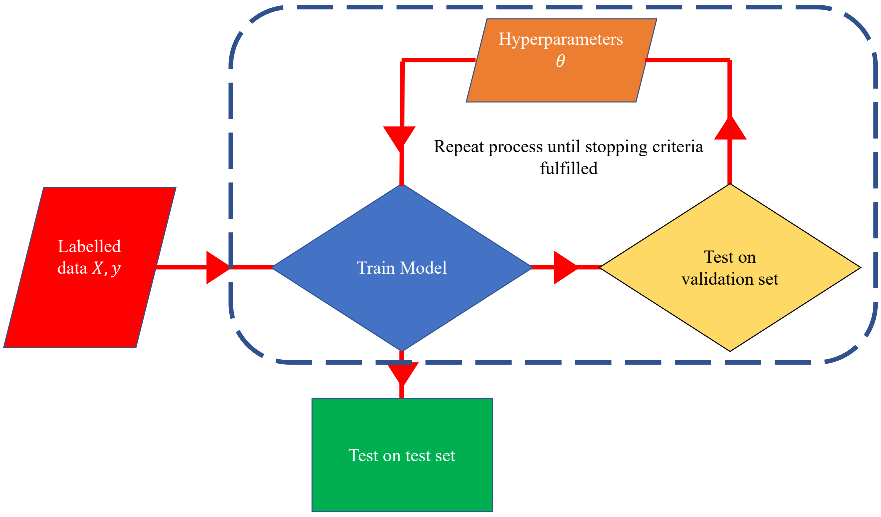


FIGURE 3.14: Typical workflow of supervised ML experiments. The optimisation of hyperparameters is often an iterative process which should be isolated from the test data to prevent bias leaking through to evaluation.

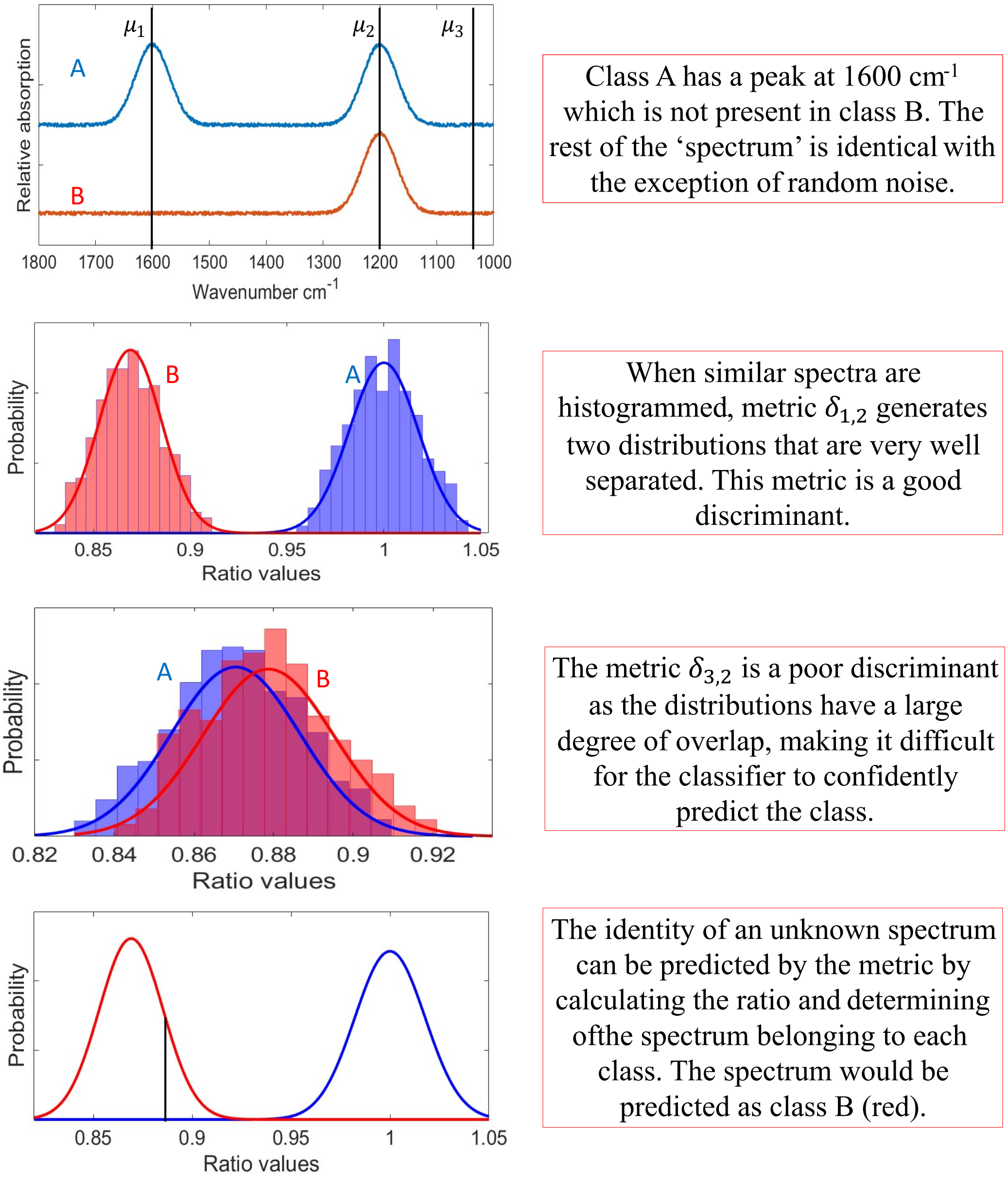
Metric Analysis

Metric Analysis (MA) is a novel ML algorithm originally developed by James Ingham within the SciaScan research group [91]. Further developments have been made by myself in order to improve the efficiency and robustness of the method. In its current phase, MA is written in MATLAB and is specifically designed for the analysis of labelled spectral data.

MA can be considered as a simple yet thorough, highly interpretable tool for spectral discrimination. At its core is the concept of a 'metric', which in this context refers to the absorbance ratio at two different wavenumbers (μ, ν), denoted herein as $\delta_{\mu,\nu}$. MA firstly calculates $\delta_{\mu,\nu}$ for every combination of wavenumber variables in the training data $\mathbf{X}$. The data can now be reframed as having n_λ features for each row of the original set. This new transformed set will be denoted $\mathbf{M}$.

As MA is a supervised classification algorithm, it requires access to a labels vector $\vec{y}$, which is the identity of each row (spectrum) in $\mathbf{X}$. MA fits a statistical distribution across each class for each metric, $\mathbf{M}$. By default, this is a Gaussian distribution, so each metric is parameterised using the mean (μ) and standard deviation (σ) for each unique class in $\mathbf{y}$.

The analysis then uses this information to determine which metrics give the best discrimination between the classes. Denoting the number of classes as n_{class}, there will be n_{class} distributions for each metric. Each of these distributions is a probability density function $\phi(\mathbf{M})$, that can be used to infer a probability that a given observation is a member of the class. Through this process, each of the training spectra can be labelled as the class with the highest probability. By doing this over all $\mathbf{M}$, the predicted labels ($\hat{\mathbf{y}}$) can be directly compared with the true labels $\mathbf{y}$ in order to evaluate a set of performance measures for each metric according to Eqs. (2.1) to (2.4).



Class A has a peak at 1600 cm^{-1} which is not present in class B. The rest of the ‘spectrum’ is identical with the exception of random noise.

When similar spectra are histogrammed, metric $\delta_{1,2}$ generates two distributions that are very well separated. This metric is a good discriminant.

The metric $\delta_{3,2}$ is a poor discriminant as the distributions have a large degree of overlap, making it difficult for the classifier to confidently predict the class.

The identity of an unknown spectrum can be predicted by the metric by calculating the ratio and determining ofthe spectrum belonging to each class. The spectrum would be predicted as class B (red).

FIGURE 3.15: Simple depiction of the formulation and assessment of a metric.

The metrics are subsequently ranked according to their 'score', which is calculated according to a combination of performance measures defined at the initialisation of the model. Each metric can now be considered as a standalone univariate classifier, where the variable is the absorbance ratio.

The final stage combines the highest scoring metrics into a soft-voting classifier ensemble [9]. With each addition of a metric, the probability each spectrum belongs to each class can be calculated by combining the individual probabilities derived from $\phi(\mathbf{M})$. The spectrum is assigned to the class with the highest probability across the ensemble. The score as a function of number of metrics can be calculated, leading to the determination of the optimal number of metrics. The optimally ordered and truncated sequence of metrics is the trained model. In order to assess the performance of the model, it is essential to to test it on *hold-out* data. Hold-out data is data which is kept completely separate from the training and optimisation process.

There has been considerable development made to MA over the course of this work. The algorithm was rewritten in a much more object-oriented (OO) manner, which improved the efficiency and eased the extraction of results. Parameters that were not originally accessible were reintroduced as hyperparameters which were easily modifiable in the function call. MA will be discussed further in chapter 4.

3.4 Scanning Near Field Optical Microscopy (SNOM)

The limit imposed by the Rayleigh criterion on the lateral spatial resolution of far field optics prevents the investigation of samples with a structure smaller than the wavelength of the light used to interrogate it. As discussed in section 3.2.3, this limit is a consequence of the diffraction observed when radiation propagates through an aperture. In the case of infrared imaging, this imposes quite a significant limit on the size of features that can be resolved. Imaging sub-micron detail with a conventional infrared microscope is impossible, prohibiting the investigation of sub-cellular organelles, nanoparticles, and many other interesting samples.

A mathematical description of the process which leads to image formation and the diffraction limit can be achieved through an approach based on Fourier optics [92].

An optical field from a monochromatic source can be described by Eq. (3.1), from which the temporal and spatial components can be separated to yield Eq. (3.25):

$$\mathbf{E}(x,y,z,t) = \exp[i(k_x x + k_y y + k_z z)] \cdot \exp(-i\omega t), \tag{3.25}$$

where $k_{x,y,z}$ are the spatial components of the wavevector $\mathbf{k}$. Considering a plane wave incident on an object situated on a transverse plane $(x,y,0)$, the spatial frequency spectrum of the electric field $\mathbf{E}(f_x,f_y)$ in the plane of the object can be obtained by applying a Fourier transform to the spatial component of Eq. (3.25). The electric field in the spatial domain can therefore be obtained by applying an inverse Fourier transform to the frequency spectrum (Eq. (3.26)):

$$\mathbf{E}(x,y,0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathbf{E}(f_x,f_y)\ \exp[2\pi i(f_x x + f_y y)]\ df_x df_y, \tag{3.26}$$

where $f_{x,y}$ is the spatial frequency and is equal to $k_{x,y}/2\pi$. The meaning of this relationship is that the electric field can be visualised as a superposition of plane waves with unique wavevectors $\mathbf{k} = (k_x,k_y,k_z)$. The component f_z can be determined by first considering $|f| = 1/\lambda$, which allows us to express f_z as follows:

$$f_z = \sqrt{\frac{1}{\lambda^2} - f_x^2 - f_y^2}. \tag{3.27}$$

In order to resolve the electric field at $z > 0$, a term accounting for the z direction propagation must be added to Eq. (3.26) in order to yield Eq. (3.28):

$$\mathbf{E}(f_x,f_y,z) = \mathbf{E}(f_x,f_y)\ \exp\left[2\pi iz\left(\sqrt{\frac{1}{\lambda^2} - f_x^2 - f_y^2}\right)\right] \tag{3.28}$$

In the case where $f_x^2 + f_y^2 < 1/\lambda^2$, the exponential in Eq. (3.26) remains imaginary, resulting in a propagating plane wave equation. This implies that low-frequency components of the field corresponding to structures in the object that are large compared to the wavelength are detectable in a distant image plane at z. On the other hand, consider $f_x^2 + f_y^2 > 1/\lambda^2$. Equation (3.27) would yield an imaginary number, which results in the exponential argument in Eq. (3.28) becoming real, corresponding to an exponential decay of the high-frequency components of the field. This implies that the fine structure cannot be observed at a finite distance z from the object.

These two conditions give rise to two distinct regions. The 'far field' is the term given to the region where $z > \lambda$, where the propagating components from the lower spatial frequency components of the field dominate. Resolution is limited by diffraction in this region. The region where $z << \lambda$ is termed the near field, and is characterised by the dominance of an exponentially decaying field containing high spatial frequency components. These are also known as *evanescent waves*, the confinement and detection of which can be achieved through a technique called *scanning near-field optical microscopy* (SNOM).

3.4.1 History

The inception of the technique used to exploit the rich information encoded within the near-field was a publication by Edward Synge in 1928 [93]. His proposal involved the illumination of a thin opaque screen with a 100 nm aperture. He postulated that the local illumination by the small aperture of a thin biological section situated at a distance similar to the diameter of the aperture would allow for the nano-scale imaging of the section, as the near-field components of the field dominate over the far-field. A basic schematic of such an instrument is shown in Fig. 3.16.

Unfortunately, Synge's concepts were decades ahead of their time. Fundamental technological difficulties prevented the application of his ideas, including the control of source-sample separation, generation of a strong signal, ability to raster scan in the nanometer regime, and the fabrication of a sub-wavelength aperture. It wasn't until the latter 20th century that these difficulties began to be addressed by the development of scanning probe microscopy techniques.

The validity of Synge's concept was experimentally demonstrated by Ash and Nichols in 1972 [94], who used radiation in the microwave regime to image a diffraction grating with a pitch of 0.5 mm. Achieving such a spatial resolution with microwaves of wavelength 3 cm proved that the concept was realisable, and this work is often cited as the birth of experimental near-field imaging.

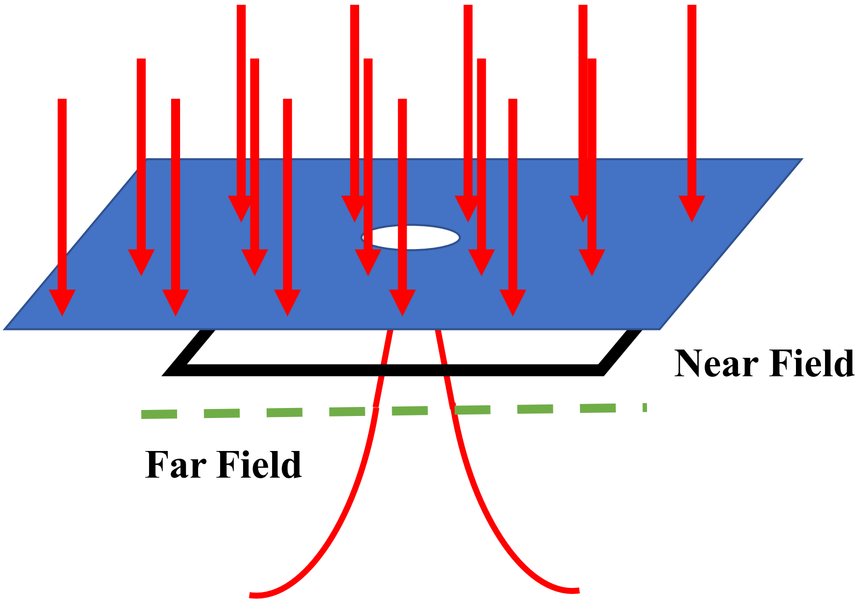


FIGURE 3.16: A sub-micron imaging system conceptualised by Edward Synge. A small aperture in an opaque screen is used as a nano-source, with a sample situated in the near field zone of the source.

Realisation of the technique with microwaves as compared to shorter wavelengths was relatively easy, as the desired sub-wavelength resolution is of the order 10^{-3} m. In order to create an image, either the aperture or sample must be incrementally scanned in the lateral plane, with each step corresponding to a pixel in the image. In the case of Ash and Nichols' experiment, positional transducers attached to a vibrating $x - y$ stage were used to control the scanning. The extension of this technique to shorter wavelengths was not realised until the invention of *scanning tunneling microscopy* in 1984 by Binning and Rohrer [95], who used a piezo-ceramic stage to scan the sample with sub-nm precision.

Technological advances motivated two independent groups to build upon Synge's proposal to build the first SNOM instruments. The two groups, one led by D.Pohl [96], the other by A.Lewis [97], independently developed apparatus capable of beating the diffraction limit with visible light.

3.4.2 Principles

The purpose of SNOM is to isolate and detect the non-propagating evanascent waves that exist within the near-field region. There exists two distinct modalities for this task: aperture SNOM and apertureless (scattering) SNOM. Aperture SNOM instruments are based on the original conception of local illumination or collection of light using a sub-wavelength aperture. Apertureless SNOM is a fundamentally different approach, in which a sharp metal tip acts as a scattering centre, converting the evanescent components of the EM field into propagating waves.

The general configuration of an aperture SNOM is to utilise an optical fiber tapered with a small aperture as either an emitter or collector. The former is analogous to Synge's postulation, whereby the local illumination of a sample is achieved by propagating far field light through a small aperture. In the latter method, the tip of the fiber is brought into the near-field zone of an illuminated sample. As discussed in section 3.4, the electric field in the plane of the sample is a superposition of waves with unique spatial frequencies, with the high spatial frequencies decaying exponentially with distance, rather than sinusoidally propagating into the far-field.

By bringing a limited object into the near field, the tip has access to those high spatial frequencies corresponding to the fine, sub-wavelength detail in the sample. A limited object is defined as an object with a sharp discontinuity leading to an infinite spatial Fourier spectrum [98]. The submersion of a limited object such as a sharp tip into the evanescent field of a high frequency object, will therefore lead to the production of both evanescent *and* propagating fields, which can be directed towards a detector. In theory, a limited object of any size will be able to convert evanescent waves into propagating ones, but the integrative effects of a large interaction cross-section will negate the super-resolution information one wishes to acquire.

The reciprocal process, illumination mode aperture SNOM, is theoretically equivalent [99]. The confined evanescent waves at the surface of a nano-emitter are disturbed by an extended object (the sample) with high frequency components. This disturbance leads to the production of evanescent and propagating components which can be detected in the far-field. There exists a linear relationship between the Poynting vector associated with the evanescent field and the intensity of the detected propagating wave, enabling the generation of an image which correlates with the fine spatial detail of a sample.

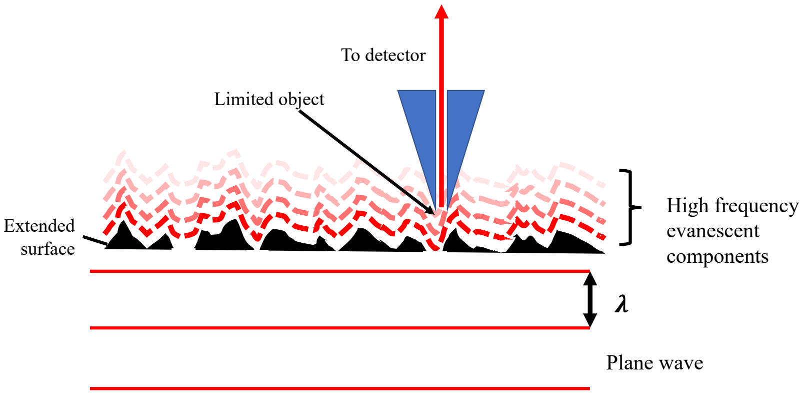


FIGURE 3.17: Depiction of aperture SNOM operating in collection mode. A plane wave is incident on a sample. The high frequency components in the diffracted light decay away rapidly with distance from the surface. A limited object can be used to convert the evanescent field into propagating waves which are directed towards a detector through an optical fiber.

A crucial aspect of SNOM is the maintenance of a constant tip sample distance. It is easy to understand why, the rapid decay of evanescent fields implies that the distance between the tip and the sample must be controlled as much as possible. If a constant distance is not maintained, intensity variations in the subsequent image will not be the result of solely the varying fine structure in the evanescent field, but it will be dominated by the varying tip-sample distance. To simply illustrate the notion of a decaying evanescent field, the tip-sample system can be modelled as an oscillating dipole. The field amplitudes at a point $P(\mathbf{r})$ contains dependencies on $|r|^{-1}, |r|^{-2}, |r|^{-3}$. The intensity of the field, which is just the square of the field amplitude, is therefore proportional to $|r|^{-2}, |r|^{-4}, |r|^{-6}$. The first term is a result of the familiar inverse square law of EM radiation, whereas the latter two illustrate the near field characteristics of dipole radiation. In order to extend this approach to more complicated objects, one can approximate the object as an ensemble of oscillating elementary dipoles. In this case, the Fourier optics approach outlined at the beginning of the section can be deployed [100].

Apertureless SNOM works on the principle that an EM field will be perturbed when it interacts with a dielectric material, leading to the excitation of the tip apex. This in turn leads to the elastic scattering of light, converting the evanescent wave into a propagating one. The main advantage of apertureless SNOM is the fact that it does not rely on an optical fibre to transmit light, therefore the tip diameter is no longer constrained. Modern technology allows for the manufacture of metal tips with atomically sharp tips, theoretically enabling atomic scale resolution of optical information, an enormous improvement on the diffraction limit imposed by far-field optics. The modality does come with some drawbacks, one is that the signal from the source must be isolated from the background. This can be achieved by modulating the light source and using an instrument such as a lock-in amplifier to isolate the signal at a particular temporal frequency. The superior popularity of the aperture compared with apertureless approach is probably due to the earlier realisation and adoption of the former technique: Pohl's and Lewis' original instruments were aperture based.

SNOM is part of a wider family of techniques called *scanning probe microscopy* (SPM), aptly named after the defining characteristic that images are formed by laterally displacing a probe relative to a sample in order to form a 2D representation of a particular quantity associated with the employed technique. For instance, *atomic force microscopy* (AFM) is an established technique that is able to image the mechanical properties of a sample (such as force, adhesion, viscosity) with an extremely high spatial resolution ($< 10^{-9}$ m) [101]. STM is a technique based on the quantum phenomenon of electron tunnelling when a bias voltage is applied between two conductors, enabling the imaging of the local electronic environment of a sample with a resolution similar to that of AFM [95].

SNOM is not a technique constrained to the visible wavelengths of the EM spectrum. In fact, Ash and Nicholls [94] used microwaves in their proof-of-concept paper in 1972. As discussed in section 3.2.1, chemical information based on characteristic molecular vibrations of different chemical species can be acquired using FTIR-MS. Unfortunately, the spatial resolution of FTIR-MS is fundamentally limited as it is a far-field technique. This is especially problematic as IR wavelengths are longer than visible wavelengths, rendering it impossible to acquire chemical images with a similar resolution to a conventional light-microscopy image.

The history and principles of SNOM imaging have been described in a broad, holistic manner. The instrumentation and application of IR-SNOM in the context of this work is discussed in more detail in section 4.4.2.

Chapter 4

Oral Cancer Analysis

The work contained within this chapter can be subdivided into three key phases. The first two phases apply FTIR-MS and MA to a cohort of oral cancer patients, and is primarily for assessing the utility of MA as a multi-class classifier, and feature extraction tool that can be used to determine spectral biomarkers. In the third and final phase, IR-SNOM is used to investigate the discriminatory features determined in the previous phase using an IR-SNOM. The MA method for analysing FTIR data has only previously been evaluated by applying it to cell lines derived from single patients [91].

4.1 Background

The motivation for developing more informative, objective and robust visualisation tools for oral tissue samples is largely due to the current lack of molecular markers and accurate clinical tests. Several studies applying FTIR spectroscopic techniques to oral cancer and associated tissue have been published over the previous two decades, presenting a variety of interesting findings which have helped to consolidate and progress understanding of the chemical processes which underpin the disease.

In a study by Schutz *et al* in 1998 [102], FTIR-MS was applied to cancerous specimens to investigate the abnormal biochemical changes in OSCC. The formation of keratin rich morphological structures known as keratin pearls is a common manifestation of OSCC. These form due to the loss of cohesion between abnormal squamous cells, leading to their concentric arrangement. These cells are functionally differentiated, symbolised by the production of keratin and loss of DNA content. Selected spectra indicated abnormal levels of protein in the keratin pearl, with the highest level concentrated in the centre. They also revealed the absence of DNA, which is in agreement with the current understanding of the structure. They also showed that surrounding collagen acts to contain and stabilise the keratin pearl structure. This study indicated the potential exploratory utility of FTIR-MS in oral histopathological studies .

Another study carried out the following year by Fukuyama *et al* [103] aimed to find the spectral markers that differentiate between OSCC and surrounding normal tissue (normal gingival, normal sub-gingival). They used FTIR spectroscopy rather than FTIR-MS to measure the bulk absorption of small sections extracted from the regions of interest. By measuring differences between spectra, they found several spectral markers that separate malignant from normal tissue. In particular, they found significant differences in intensity of features between 1482 - 1431 cm^{-1}, 1274 - 1183 cm^{-1} and 1368 cm^{-1}. This study demonstrated the potential of FTIR spectroscopy as a means to discriminate malignant from benign oral tissue.

Bruni *et al* [104] were able to attribute high DNA and high collagen content to proliferative and regressive states of OSCC tumours respectively, on the basis of chemical maps acquired by FTIR-MS. Spectral differences between normal and diseased tissue were observed at 970 cm^{-1}, 1026 $^{-1}$, 1550 cm^{-1} and 1735 $^{-1}$. These findings exemplify how spectroscopic changes can be correlated with subtle pathological alterations that may otherwise be overlooked.

Conti *et al* published two papers which extended the univariate analysis carried out by Bruni *et al* to a more thorough multivariate approach. In the first paper [105], samples were selected from several different regions of the oral cavity and lymph nodes presenting OSCC or OED of variable degree, and imaged using an FTIR microscope. The data were clustered using hierarchical cluster analysis (HCA), with those clusters labelled by a pathologist. They then compared spectral profiles of different clusters, using pre-defined characteristic peak positions to inform their selections. PCA analysis was also used on the data, which was able to separate the clusters determined by HCA over three principal components. The same authors followed this up in a second study, where HCA and PCA were again used to group and compare similar spectra, correlating the findings with pathological observation. They imaged normal and malignant cell lines, inoculated rat cancer tissue and human tongue tissue as samples. These robust analyses demonstrated how unsupervised techniques were able to categorise spectra into pathologically different groups based entirely on spectral similarity rather than any *a priori* knowledge of their origin. The spectral differences between the clusters are therefore implied to be 'fingerprints' of the pathologies represented by such clusters, which is supported by the comparable results from the PCA.

Sabbatini *et al* [106] built upon these results to attempt to ascertain the spectral characteristics that define different grades of OSCC, healthy and dysplastic tissue. HCA and PCA were again deployed as multivariate approaches to group, compare and contrast spectra. They correlated spectral differences in more malignant tissue with known molecular mechanisms, such as consumption of carbohydrates for proliferating cells that require a surplus of energy.

Meanwhile, the same group employed multiple multivariate image analysis (MIA) methods to analyse tissue specimens from multiple OSCC patients. The samples were arranged in an array of small circular samples which have been extracted from an FFPE tissue block (see section 2.1.2) called a tissue micro-array (TMA). The study, by Pallua *et al* in 2012 [107], used both HCA and KCA on the hyperspectral images to generate pseudo-colour images that correlated with the H&E stained counterparts of the same sample (Fig. 4.1). This further shows that, with the aid of MIA techniques, the high information content contained within hyperspectral images of oral cancer tissue can be leveraged to rapidly and objectively cluster into groups that are reflective of the pathology. In the same analysis, the group were able to obtain contrast in maps of peak area corresponding to glycoproteins and nucleic acids, although more contrast was acquired when using the MIA techniques. Demonstration of class separation on an entirely spectral basis was also achieved using PCA on a number of selected spectra.

A different study by Banerjee *et al* [108] aimed to determine the spectral biomarkers that distinguish between normal tissue, OL and OSCC. They realised that classification using fewer features selected by FFS attained superior scores than that of the entire feature space. For the OL vs OSCC classification, 81.3% sensitivity and 95.7% specificity was achieved using linear and quadratic support vector machines (SVMs).

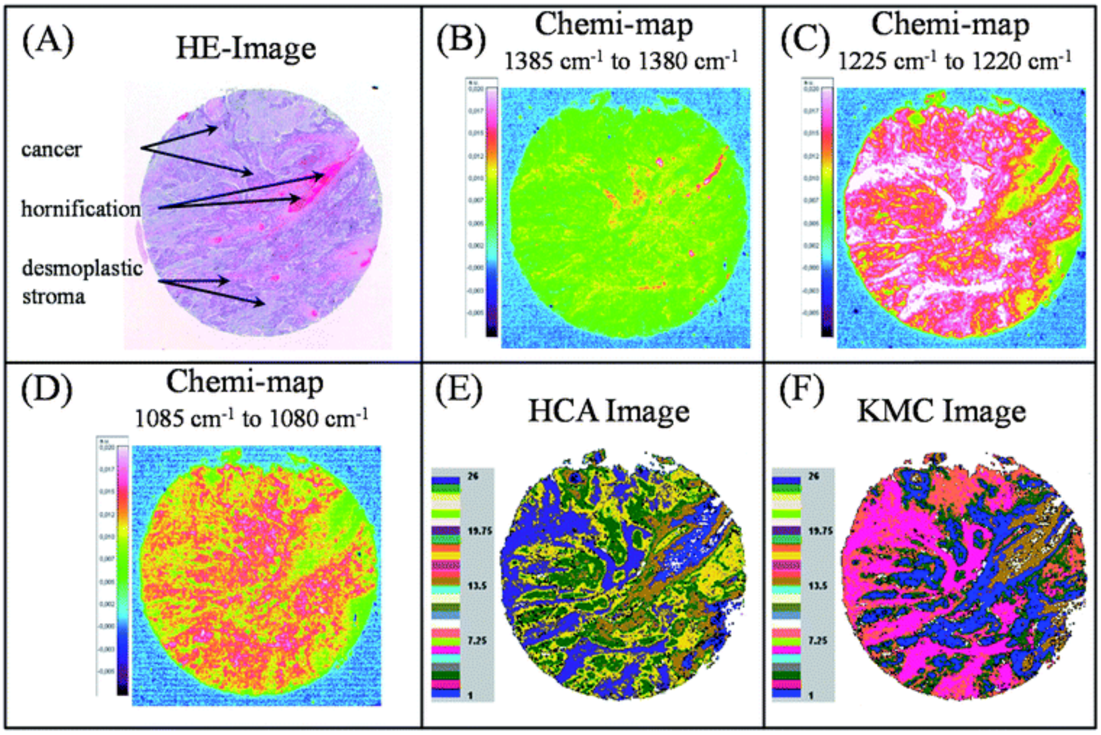


FIGURE 4.1: Pallua *et al* [107], fig. 7. (a) HE stained sample from TMA, regions of interest corresponding to cancer, hornification and desmoplastic stroma identified by histopathologist. (b) - (d) Peak area maps for 3 different wavenumber ranges: 1385 cm^{-1} - 1380 cm^{-1} (glycoproteins), 1225 cm^{-1} - 1220 cm^{-1} (nucleic acids), 1085 cm^{-1} - 1080 cm^{-1} (nucleic acids). (e) HCA pseudo-colour image. (d) KMC pseudo-colour image.

The features selected by FFS were attributed to glycogen and keratin content within the two different tissues. The successful discrimination between OSCC and OL suggests that FTIR spectroscopy could have utility in the risk management and stratification of PPMOELs.

Raman spectroscopy has also seen a variety of applications to oral tissue related studies [109]. Raman spectroscopy is less well suited to high volume histopathological studies than FTIR-MS because it takes longer to record an image (point spectroscopy raster scanning vs FPA hyperspectral imaging) and attain high signal to noise from a weak signal.

The survey of recent literature surrounding FTIR-MS applied to oral cancer reveals that there has been a lack of supervised learning classification techniques applied to entire images of oral tissue. The papers described previously either make use of unsupervised algorithms such as HCA, KMC and PCA to generate pseudo-colour images that can be correlated with histopathology. Whilst the results have shown high degrees of correlation, it is difficult to infer a quantative measure of how well the algorithm has performed, as there is no ground truth. Supervised techniques have been used to classify powdered samples formed from dried tissue sections [108] with good results, but the potential of the model as a labelling tool was not demonstrated. Interpretation of results derived from supervised models are also more reliable than that of unsupervised models, as the discrimination is guided by the identity of each spectrum.

The work in this chapter will be split into three sections. The first will be the application of multivariate techniques to images acquired by FTIR-MS, for the characterisation of various tissue types that are typically found in normal and malignant oral tissue specimens. The following two sections will be oriented around a recently published paper by Ellis *et al* [110] which uses metric analysis to identify spectral biomarkers that separate metastasis from lymphoid tissue, and subsequently using those biomarkers for higher resolution study with an IR-SNOM. The author of this thesis is the lead author of [110].

4.2 Methods

4.2.1 Sample Selection and Preparation

Specimens were selected from a cohort of patients displaying a mixture of primary OSCC (n=4), nodal metastasised OSCC (n=2) and associated normal tissue. Two adjacent $\approx 4\,\mu\text{m}$ sections were obtained using a Beecher MTA-1 microtome. One of the sections was floated onto a charged glass slide for conventional histopathological assessment, whilst the other was floated onto a 2 mm calcium fluoride (CaF_2) disk for FTIR-MS experiments.

Preparation for histopathological assessment entailed the routine deparaffinisation and HE staining of the thin section. The samples set on glass slides were initially submerged in histology grade xylene for 5 minutes to dissolve the paraffin. Following this, the samples were sequentially immersed in 100%, 95% and 70% ethanol for one minute each in order to gradually hydrate the sample. The slides were then rinsed in running water for 30 seconds to remove any remaining paraffin or other solid contaminants. The first component of the dye to be added was haemotoxylin, which the slide was fully submerged in for a total of 3 minutes, and subsequently rinsed with running water. The slide was then dipped into acid alcohol to prevent excess background staining, which was followed by treatment with Scott's tap water as a 'blueing' step. As a precursor to eosin staining, which is alcohol based, the sample was again immersed in 95% alcohol before submersion in eosin for 5 minutes. The samples were transferred back through 95% then 100% ethanol to dehydrate the section, followed by a final treatment with xylene for 1 minute. Stained samples were covered with a cover-slip.

The samples set on CaF_2 disks were not subject to the same deparaffinisation as the histopathological slides, the sections were instead retained in their original FFPE form. It was decided not to remove paraffin for a number of reasons, listed below [77]:

(i) The exhaustive process of paraffin removal may introduce unwanted sample variability influenced by the environment.

(ii) There is sufficient information outside of paraffin dominated spectral regions, so that these regions can be discarded. Important spectral biomarkers such as those resulting from proteins, DNA and carbohydrate are still present. Significant paraffin peaks exist within 3000 - 2800 cm^{-1} and 1490 - 1340 cm^{-1}.

(iii) The problem of Mie scattering (section 3.3) is largely alleviated when the sample is embedded in paraffin. This is because the refractive index of paraffin and tissue are very similar, which decreases the scattering efficiency in Eq. (3.23).

(iv) The stability and structural integrity may be preserved by paraffin.

Light micrscopy (LM) images of the stained sections were acquired using an Aperio CS2 scanner (Leica Biosystems). These images were used by an oral histopathologist to identify specific regions of interest that should be targeted in subsequent FTIR experiments.

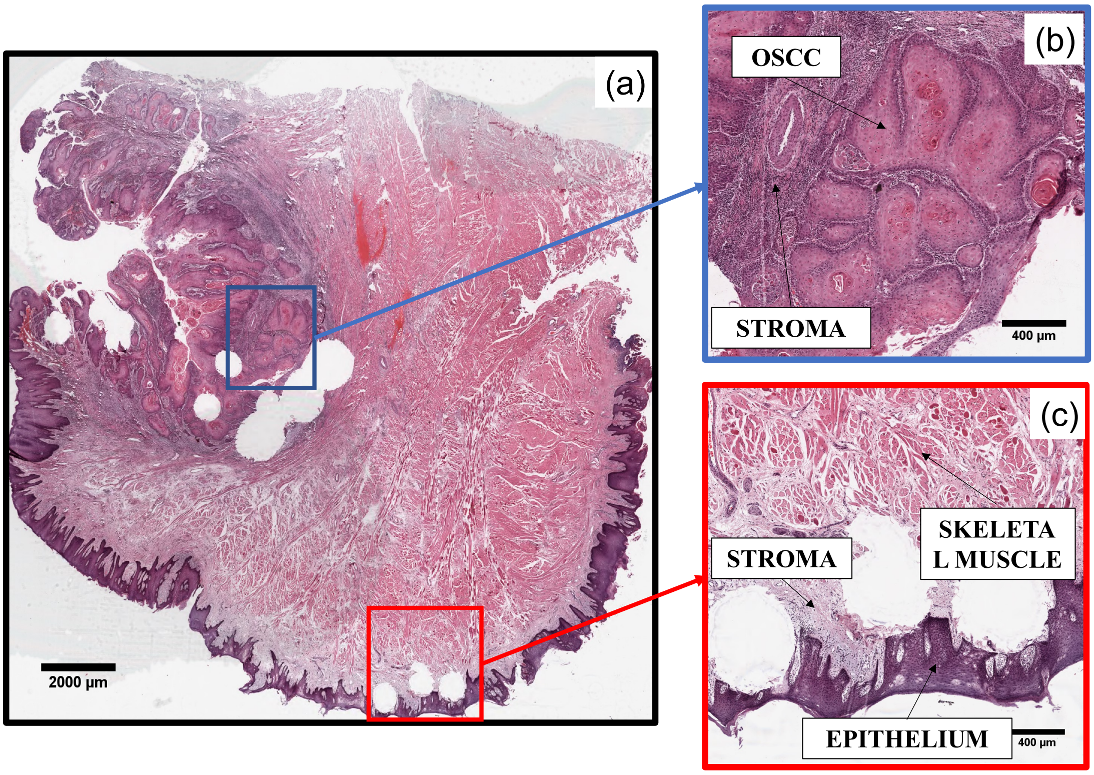


FIGURE 4.2: (a) HE of entire tissue section from OSCC patient, with regions of interest (ROI) defined by histopathologist. (b) ROI 1: Invasive OSCC with surrounding stroma tissue. (c) ROI 2: Normal epithelium with adjacent stroma and skeletal muscle tissue.

Figure 4.2 is a set of LM images of a section taken from the tongue of patient P2 with OSCC. The small holes located in the middle and bottom edge of the sample are where punches have been used to extract smaller cores for TMAs. The region enclosed by the blue square is shown at a higher magnification in Fig. 4.2b. The well-differentiated cancer cells have accumulated in a series of islands, situated amongst stroma tissue. Since well differentiated squamous cells produce keratin, there are areas that have been stained red by the H&E dye. Stroma is the general name given to the supportive framework of tissue which contains functional cells such as fibroblasts and immune cells. Stroma tissue in the vicinity of a tumour is abundant in cancer associated fibroblasts (CAFs), cells that stimulate the growth of tumours by the production of growth factors, the presence of which has been shown to correlate with stage and prognosis [111]–[113]. A list of patients and corresponding site is shown in table 4.1.

TABLE 4.1: Patients selected for imaging.

Patient	Site	N(ROIs)
D18	Tongue	4
A12	Lymph Node	3
B12	Retromolar	3
D2	Buccal	1
A13	Alveolus	3

4.2.2 FTIR Experiments

FTIR-MS experiments were carried out in Peter Gardner's laboratory, Manchester institute of biotechnology, Manchester, United Kingdom. The apparatus, shown in Fig. 4.3, consisted of a Varian 620 IR microscope coupled with an Agilent Cary (formerly Varian) 670 FTIR spectrometer. A schematic is also shown in Fig. 4.4 The sample stage was contained within a sealed chamber which is purged with dry-air in order to significantly reduce the presence of water vapour, which has strong presence in the mid-IR spectral region. The lateral position and height of the microscope's sample stage could be controlled by an external controller so that the chamber need not be opened in order to adjust sample position. The integrated system was controlled entirely by a PC operating the Agilent Resolutions Pro software.

The spectrometer previously described was also purged with the facility's dry-air supply to effectively remove any water vapour from the optical path. The mirrors within the interferometer were air bearing in order to reduce external vibrational noise and increase speed and throughput. The humidity and temperature of the system were monitored using a digital display humidity gauge and thermometer within the purge-chamber.

Considering the time taken to purge the apparatus to acceptable levels ($< 1\%$ relative humidity), opening the purge chamber to replace each sample after data acquisition would be inefficient. Furthermore, any background atmospheric correction would be inconsistent with current conditions, as the chamber will be purged to a lesser/higher extent. For these reasons, a sample holder capable of housing three disks was designed using CAD software and 3D printed. Two of the disks were to contain tissue, with a third, blank disk reserved for calibration and background scans.

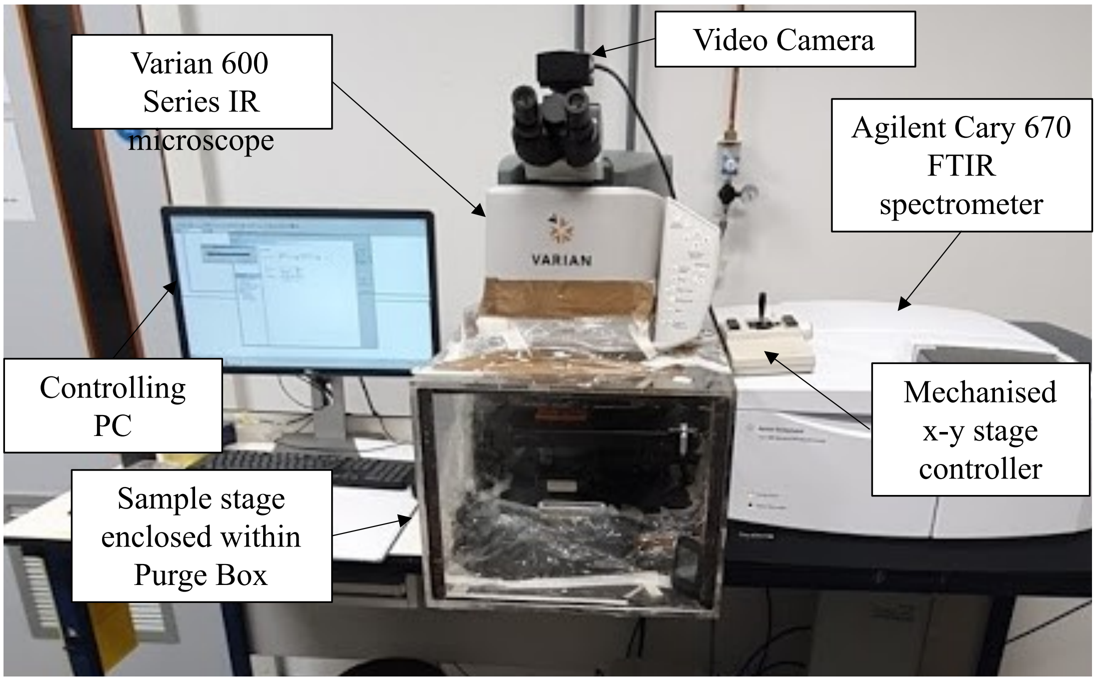

Varian 600 Series IR microscope
Video Camera
Agilent Cary 670 FTIR spectrometer
Controlling PC
Mechanised x-y stage controller
Sample stage enclosed within Purge Box

FIGURE 4.3: Labelled photograph of apparatus used for FTIR-MS experiments.

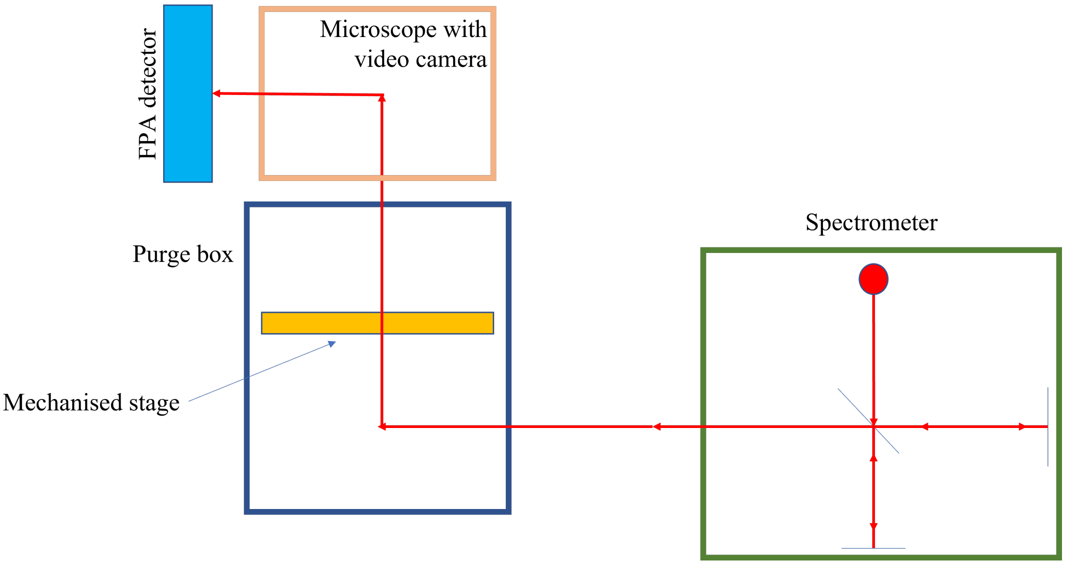


FIGURE 4.4: Schematic of the FTIR system. More detailed figures for the optical system are in Fig. 3.9 and Fig. 3.7.

Contained within the IR microscope was a 128x128 element MCT-FPA detector, which was able to simultaneously detect the IR light over a FoV governed by the microscope optics. For low magnification mode, a 15x objective was used, enabling a FoV of approximately $0.7 \times 0.7\,\mathrm{mm}^2$, corresponding to a 5.5 µm pixel size in the resultant hyperspectral image. The detector was situated next to a dewar which was filled with liquid nitrogen (LN_2) in order to reduce the thermal noise as explained in section 3.2.2. The effective FoV of the microscope could be extended by using a built in 'mosaic' function, which programmed the microscope to sequentially record a defined arrangement of hyperspectral images, which were concatenated together after data acquisition.

As depicted in Fig. 3.9, when the microscope is configured to transmission mode, the IR light propagates through two optical systems: the *condenser* and the *objective*. The condenser focusses the beam onto the sample, whereas the objective collects the transmitted light, recollimating it for detection at the FPA. It is crucial that these two optical systems are brought into mutual focus to ensure high image quality. To bring the objective into focus, the microscope was configured to image visible light using the video camera situated on the same light path as the IR beam. The microscope was set to transflection mode, whereby the optical path is configured to both illuminate the sample and collect the reflected beam through the objective. The condenser beneath the sample is therefore removed from the optical path, isolating the objective and enabling the focus to be found by adjusting the stage height. The operating mode was switched to transmission in order to focus the condenser. Coarse adjustments were made by adjusting the condenser height and monitoring the visible image quality.

Fine adjustments were made to the condenser height by monitoring the distribution of light incident on the FPA. Since the FPA detects IR photons, the optics were switched back from the visible path to the IR path. The condenser height was adjusted until a uniform intensity was observed on the FPA image, shown in Fig. 4.5, whilst changes to the intensity were made to ensure the detector was not saturated. The intensity at the detector was modified by changing the integration time or positioning a 25%, 50% or 75% attenuator in the light path. The FPA was calibrated by setting the intensity at each pixel to zero.

When there is zero path difference (ZPD) between the moving and fixed mirror, all constituent wavelengths of the recombined beam are in-phase with each other. This property is reflected in the interferogram, where there is a dominant peak at ZPD, as all the wavelengths are constructively interfering with each other. Before any spectral measurements are to be measured, the point of ZPD, termed the *centerburst*, should be reset to zero as a calibration step. A rapid background hypercube was acquired as a test to check the quality of the measured spectra. Spectra across the image should be similar in shape and intensity. The spectra were inspected for detector saturation, which manifests as sharp discontinuities at the top of the spectrum.

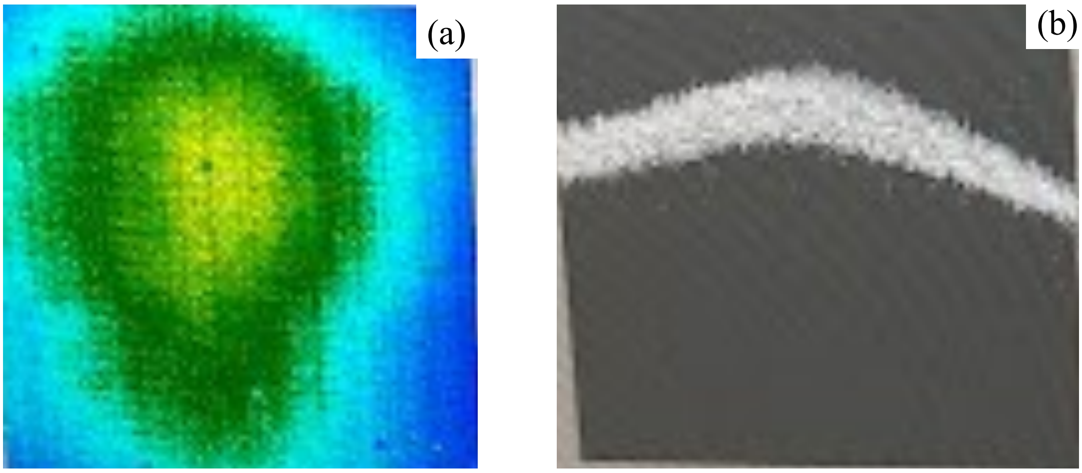


FIGURE 4.5: Representations of the illumination distribution on the FPA. The 2D intensity map is shown in (a), whilst the intensity as a function of flattened position is shown in (b). A tight, uniform intensity distribution is desired, represented by the compact, shallow curve in (b)

The software was used to configure various settings of the instrument prior to data acquisition. The Fourier transform applied to the interferogram (Eqs. (3.16) and (3.18)) has infinite limits, which implies that the mirrors move to an infinite distance. To account for this impossibility, an apodisation function is applied to the FT. This may take the form of a simple box car function, which is equal to 1 within the limits of the interforemeter and zero otherwise. The issue with this is that the FT of a boxcar function is a *sine cardinal* (sinc) function, which is characterised by a central peak with much smaller peaks (ripples) either side, which manifest as spectral noise in the data. Apodisation with a triangular function reduces this effect, but also broadens peak width, effectively worsening the spectral resolution. The Happ-Genzel apodisation function is frequently used in FTIR spectroscopy applied to solid samples as its FT displays a desirable trade-off between spectral resolution, peak height and spectral noise. For this reason, the Happ-Genzel function was selected for apodisation of the interferogram. Example apodization functions and their respective Fourier transforms are shown in Fig. 4.6.

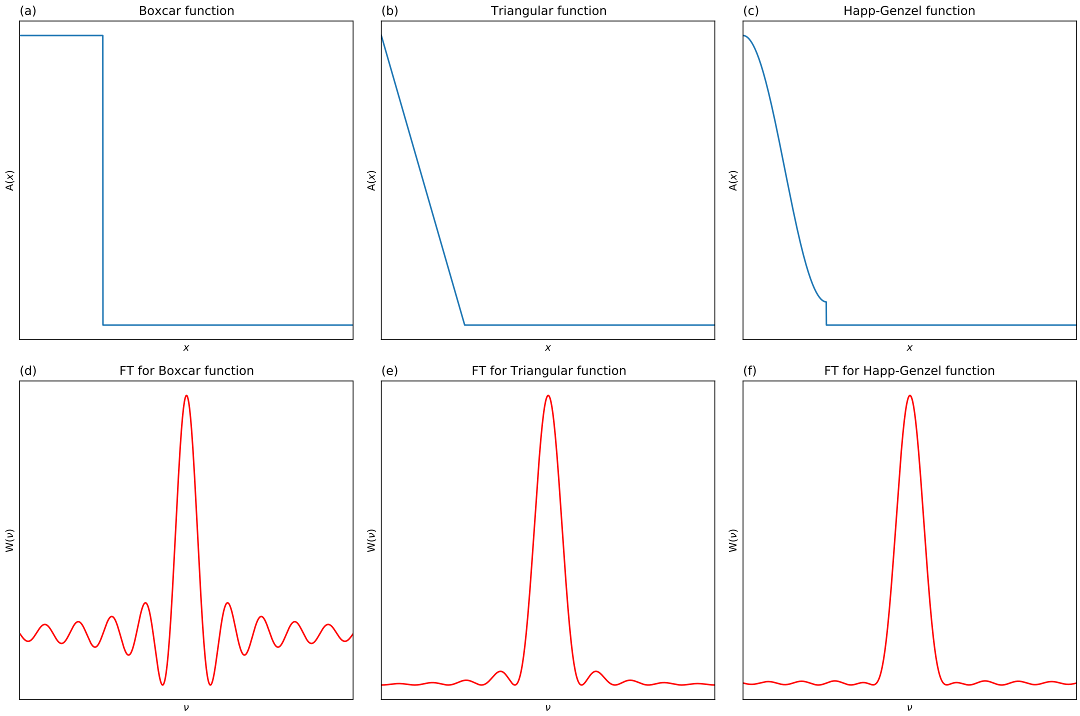


FIGURE 4.6: Plots of different apodisation functions (a-c) and respective Fourier transforms (d-f)

The spectral resolution is an important parameter which governs the minimum width of spectral features that can be resolved in the FTIR spectrum. This was set at 4 cm^{-1} in order to capture the narrower spectral bands and shoulders that often appear on the fringes of broader bands such as the amide I peak. The spectra were interpolated to produce a smoother line shape by using a technique called zero-filling, whereby N zeros are added to the end of the N point interferogram, which changes the data spacing from 4 cm^{-1} to 2 cm^{-1}. The spectral acquisition range was set as 900 - 3800 cm^{-1}.

Real spectral data could be acquired only when the sealed purge chamber had reached <1% humidity, and the detector had cooled to 78K. A background scan was first recorded by co-adding 256 hyperspectral images of the blank disk. A higher number of co-additions are required for the background scan than the sample scan to minimise the introduction of noise from the background, which is more suscpetible to noise artefacts due to the lack of absorbing media in comparison to the sample scans [64].

The transmission spectra of subsequent sample scans are divided by the transmission spectra of the background scan as, according to Eqs. (3.8) and (3.9), the absorption is calculated by taking the base 10 logarithm of the ratio between the absorbed and non-absorbed transmission spectrum. The background spectrum is a convolution of the source emission spectrum, the transmission spectrum of the CaF_2 disk and the transmission spectrum of atmospheric contributions such as carbon dioxide and water vapour. Figure 4.7 is the average spectrum taken from an example background hyperspectral image. The uniformity of the beam and spectral characteristics are reflected in the red dotted lines representing the standard deviation of the spectra.

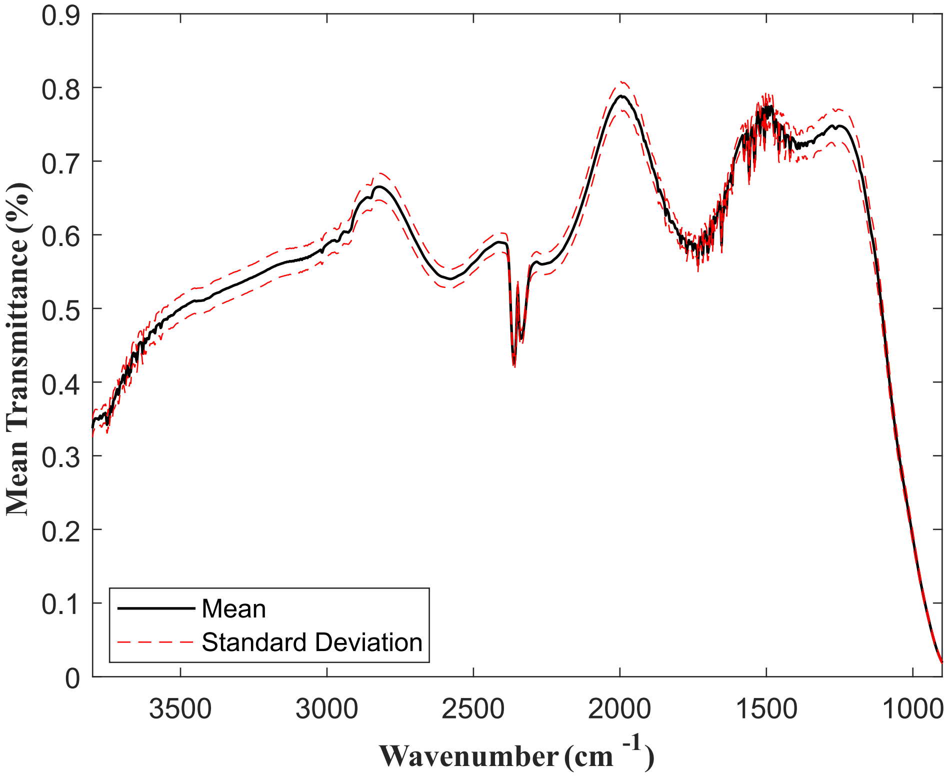


FIGURE 4.7: Mean and standard deviation of an example background transmission spectrum obtained from CaF_2 housed in a purged chamber.

Subsequent to the background scan, sample hyperspectral images were acquired. The built-in visible light video camera was first used to find the specific ROI(s) within the tissue section. Since the FoV of the video camera is very small compared to the H&E images used for co-registration, finding the exact location for imaging was difficult. To make this task easier, the mosaic function was used to concatenate multiple adjacent frames together to extend the FoV. The process is summarised in Fig. 4.8, whereby the H&E microscopic image is used to locate the ROI on the large visible mosaic image.

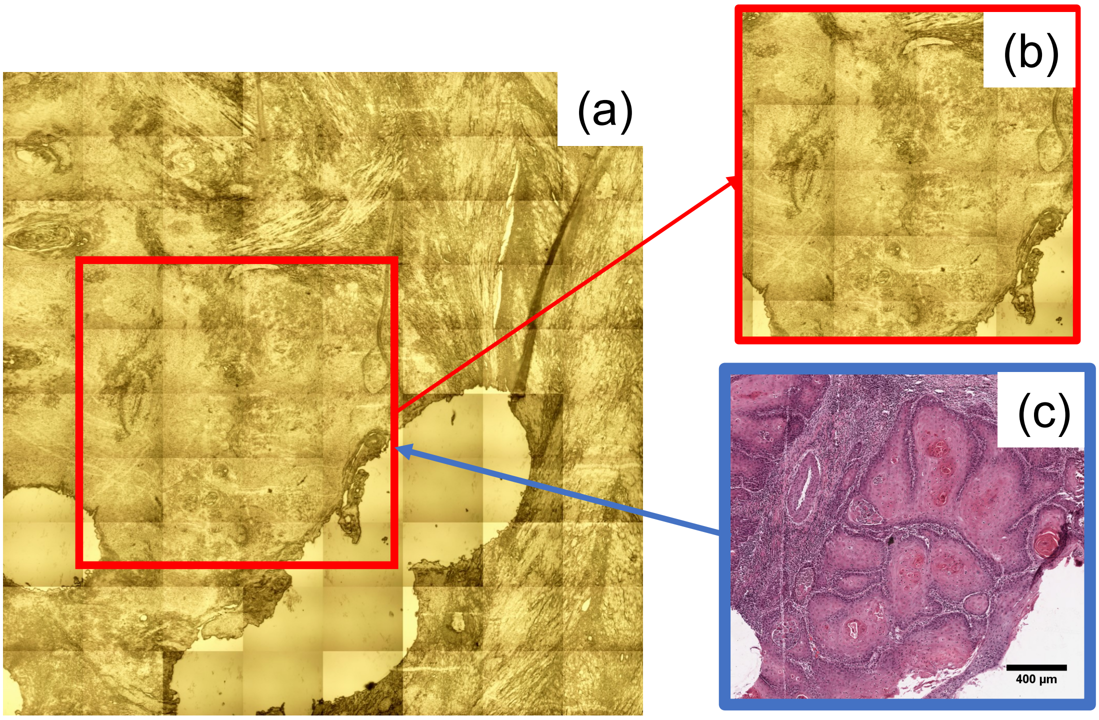


FIGURE 4.8: Procedure for finding and selecting ROIs using the visible mosaic function in the resolutions pro software. (a) is the large visible mosaic, (b) is the cutout selected for IR imaging, informed by the H&E counterpart in (c).

For the sample scans, all experimental parameters were kept the same other than the number of co-additions, which was reduced to 128 to increase the throughput of the experiments. The software automatically calculated the number of FPA images required to image the ROI defined according to Fig. 4.8. A background image was acquired every time two samples were replaced on the 3D printed slide holder, so that current atmospheric conditions were appropriately corrected for.

4.2.3 Data Labelling and Pre-processing

All data processing and analysis in this chapter was performed using `MATLAB` (MathWorks), a versatile programming language and platform which facilitates tasks such as image analysis, machine learning, matrix manipulation and graphical visualisation. The `ChiToolbox` [114] and `iRootLab` [115] frameworks were used to import the data, and apply various methods to process the data. A lot of the methods, however, were written by the author - including the revised MA algorithm, SNOM process methods amongst others.

The imported hyperspectral images were stored as MATLAB 3D matrices, each with $128n_x \times 128n_y \times 1506$ elements, where n_x and n_y are the number of tiles in the horizontal and vertical direction respectively. Regions in the hyperspectral image which originated from areas of no tissue or areas where the section is too thick were identified and excluded by applying a filter which required the absorbance at the amide I peak (1650 cm^{-1}) to lie between 0.1 and 2. After this, each image was subject to the same pre-processing procedure which included spectral truncation, PCA denoising (retaining 10 PCs), paraffin region removal, vector normalisation and rubber-band correction. It was decided not to apply the Mie-scattering correction for reasons summarised in sections 3.3.1 and 4.2.1.

Seventeen regions of interest (ROIs) were deemed suitable by a histopathologist to be used for analysis. They were selected based on pathological criteria, such as typical presentation of disease and biological heterogeneity. Samples with more than one pathology (e.g. OSCC and CS) were also preferred due. The archival samples are from the University of Liverpool's biobank. Amongst the seventeen ROIs, eight different tissue types were identified and labelled by a pathologist. These annotations were used to specifically extract and label spectra that correspond to the different tissue types. The identified tissue types are listed in table 4.2, accompanied by the corresponding images used for labelling and total number of spectra. An equal number of spectra were randomly sampled from each individual image to prevent image related bias from interfering with results. Representative LM images containing each identified tissue type are shown in Fig. 4.9.

The pre-processed mean spectra for each tissue are shown in Fig. 4.10. Clearly, each tissue type shares common spectral characteristics, such as the dominant amide bands between 1700 cm^{-1} and 1500 cm^{-1}. Despite the appearance of differences between mean spectra, it is difficult to derive meaningful interpretation based on mean spectra inspection. The inherent heterogeneity of biological tissue is portrayed in the standard deviation of the spectral profiles (green and red dotted lines).

TABLE 4.2: Identified tissue types and their respective ROIs and number of labelled spectra.

Tissue	Abbreviation	ROIs	$N_{spectra}$
Oral squamous cell carcinoma	OSCC	D18(2), D18(3), D2(4)	33534
Maturation layer of normal epithelium	NE	B12(1), B12(2), D18(1)	5691
Newly formed desmoplastic tumour stroma	CS	A13(1)	1897
Pre-existing supporting stroma	NS	A13(1), B12(1), B12(2), D18(1)	21752
Metastasised tumour	MT	A12(1), A12(2)	2698
Lymphoid nodal Tissue	LYM	A12(1), A12(2)	4322
Submucosal elements	SM	A13(2), D18(1)	14790
Basal layer of normal epithelium	BL	B12(1), B12(2)	2036

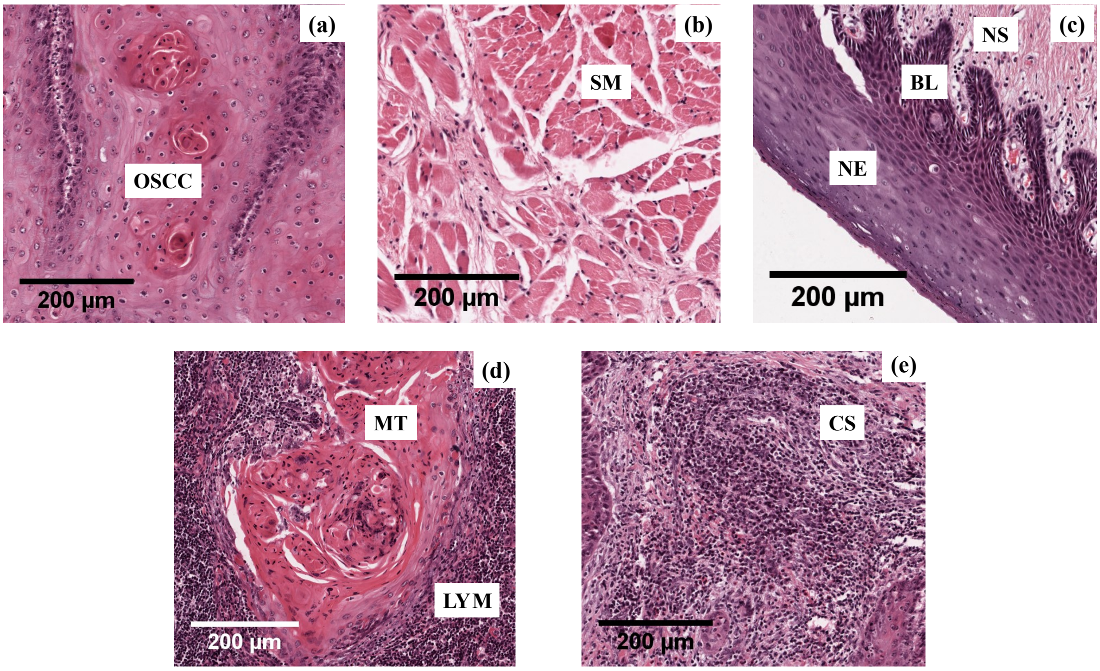


FIGURE 4.9: Representative LM images of H&E tissue containing the eight identified tissue types.

Despite the difficulties of interpreting mean spectra quantitatively, an initial qualitative impression can be obtained by comparing the spectral profiles in Fig. 4.10 and the representative images in Fig. 4.9. Of particular note is the difference in spectral shape in the nucleic acid region (1300 cm^{-1} - 1200 cm^{-1}) between epithelial tissue (OSCC,

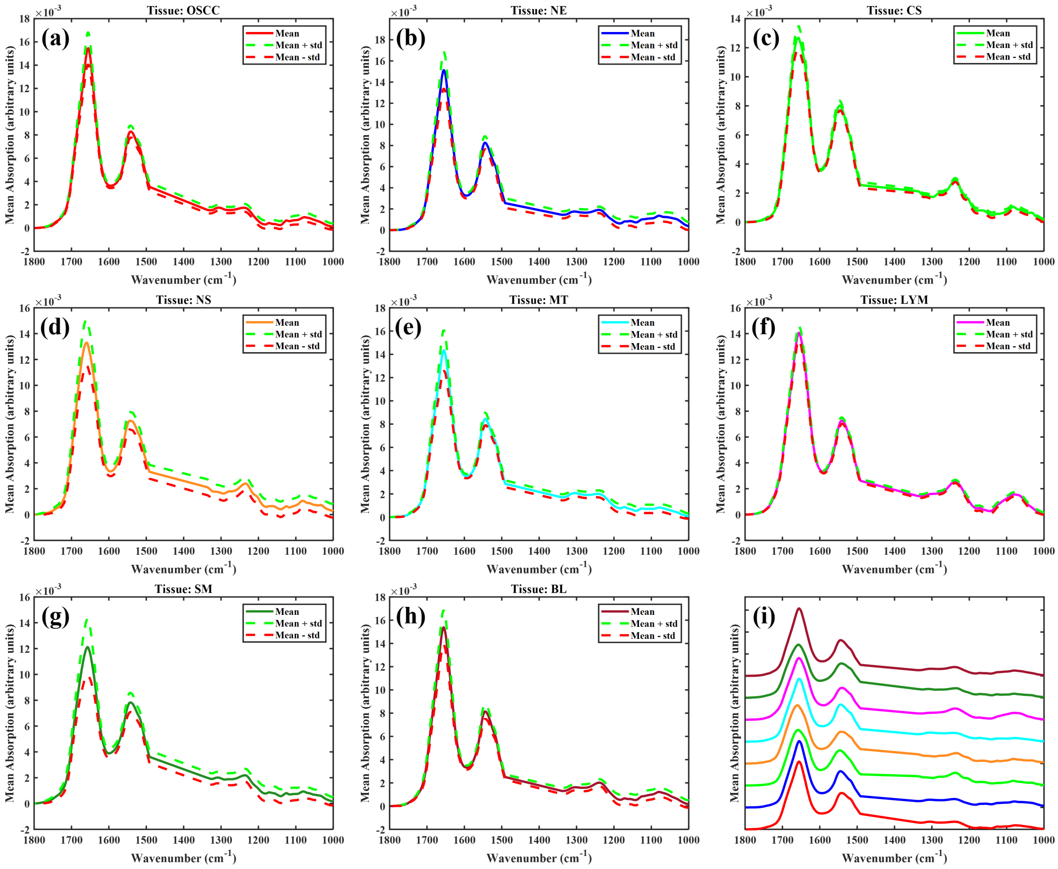


FIGURE 4.10: Mean spectra and standard deviation for (a) OSCC, (b) NE, (c) CS, (d) NS, (e) MT, (f) LYM, (g) SM, (h) BL. (i) stack of the average spectrum from each tissue for comparison.

NE, MT, BL) and supporting tissue (NS, CS, LYM). There appears to an enhanced signal in supporting tissue compared with epithelial tissue, perhaps arising as a result of the higher concentration of nuclei in Fig. 4.9a and b.

The MA method previously described in section 3.3.2 has been shown to be a powerful tool that both enables unbiased interpretation of spectral biomarkers that discriminate between tissue types, and as a diagnostic tool which is able to efficiently and accurately label data. For these reasons, MA was exploited in this work to determine spectral differences between the labelled tissue types, as well to label hyperspectral images to investigate whether the obtained information correlates with histopathology. MA was deployed on the labelled dataset to build two distinct models, detailed in subsequent sections.

4.3 Phase I: Multi-tissue Primary Site Model

4.3.1 Parameter selection and Model Training

This section will be based on the construction and interpretation of an MA model based on all primary site labelled tissue types. This includes all the tissue types in the data except for MT, which has, by definition, spread away from the primary site.

Metric analysis version 4 (MA.v4) is an object oriented adaptation of the previous iterations of MA. It also incorporates the hyperparameter tuning and hold out testing functionality detailed in section 3.3.2, so that the optimal set of parameters can be objectively determined. Table 4.3 lists and defines each tunable hyperparameter in MA.v4.

TABLE 4.3: MA.v4 hyperparameters

Hyperparameter (θ)	**Description**	**Possible Values**
Data spacing	The downsampling factor for the horizontal (wavenumber) axis. More downsampling increases the speed of the analysis.	$2, 4, 8$
Optimising score	The performance measure used to optimise the number of metrics in the model.	Youden's J, F1

The F1 and Youden's J scores both represent different combinations of the performance measures detailed in Eqs. (2.1) to (2.4). The F1 score (Eq. (4.1)) is the harmonic mean of the sensitivity and precision, whilst Youden's J (J) is a sum of the sensitivity and specificity Eq. (4.2). These scoring systems represent more balanced representations of the performance of a test.

$$F1 = 2 \cdot \left(\frac{\text{sensitivity} \times \text{ppv}}{\text{sensitivity} + \text{ppv}} \right) \tag{4.1}$$

$$J = \text{sensitivity} + \text{specificity} - 1 \tag{4.2}$$

A separate hold-out test set comprising 75% of the total data was kept aside from any model construction so that the optimised MA approach could be tested on unseen data. The partition was stratified so that the class proportions were constant in the train and test set. The training data was further divided into three equal sized class stratified partitions, so that each partition contained 1/3 of the training data. The model was then to be trained on two partitions, leaving the third out for testing of the selected hyperparameters. This was repeated three times so that the scores on all accessible data can be evaluated. This process, known as k-fold cross validation (where $k = 3$ in this instance), ensures the model has been evaluated on all data, addressing the risk of acquiring over/under-optimistic scores based on a single partition of data. Performance measures such as area under receiver operator curve (AUC), sensitivity and specificity can be expressed as a mean and standard deviation from the k iterations.

The hyperparameters were optimised by trialling every permutation of the hyperparameters detailed in table 4.3 in an approach known as grid-search. The permutation with the highest scores will be selected as the optimal model parameters. All the training data was then used to train the model which would be tested on the so-far unseen hold-out data, in order to evaluate the general performance of the model given new data. This approach reliably evaluates the diagnostic potential of the trained model.

Figure 4.11 shows the results of the hyperparameter optimisation. Three graphs depicting the distribution of scores for each tissue at the different hyperparameter selections are displayed in order to visualise the dependence of the scores on the different selections. The area under (ROC) curve (AUC) is a popular measure of the performance of a binary classifier, which is defined as the area under the ROC curve (described in detail in Fig. 2.8). A perfect classifier will have 100% sensitivity and 100% specificity, yielding an AUC of 1.00. Classifiers with an AUC between 0.8 and 1.0 are regarded as skillful, whilst anything less than 0.5 resembles a classifier with no skill. A trained MA model is a series of binary 'one vs all' classifiers, therefore the AUC and ROC curves are readily accessible.

From Fig. 4.11a, it is apparent that the submucosal elements (SM) classifier performs relatively poorly, yet in absolute terms the AUC is still good (lowest is $\approx$ 0.84 AUC.

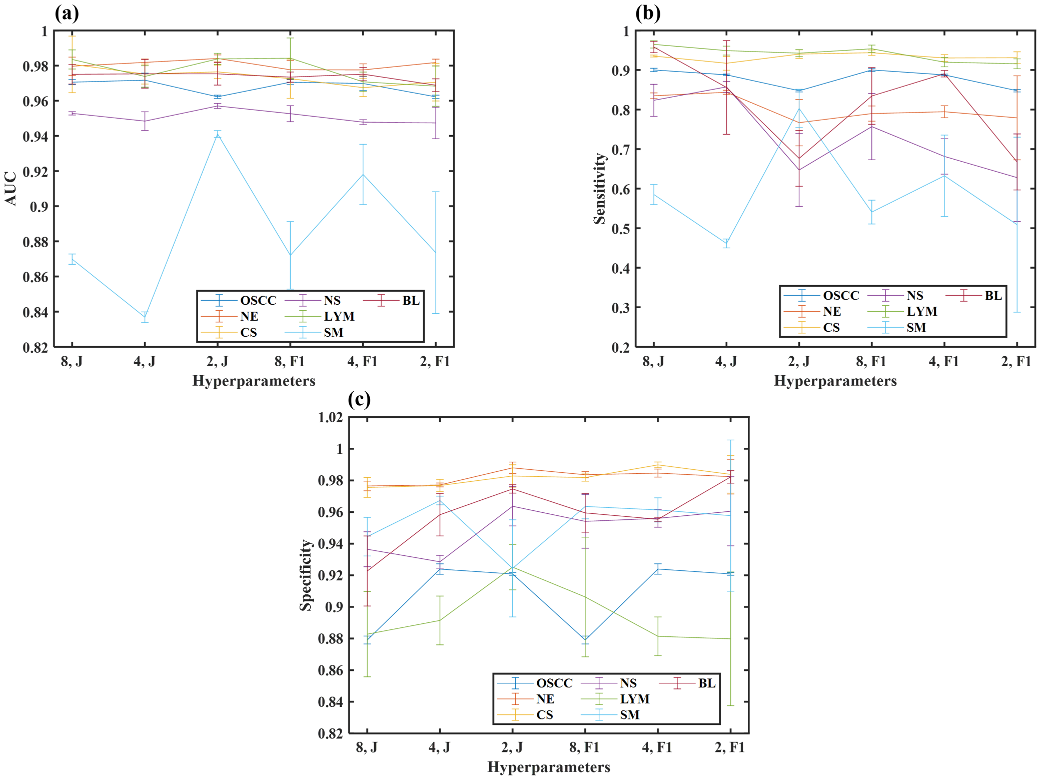


FIGURE 4.11: Plots of the mean (with error-bars) of the AUC (a), sensitivity (b), specificity (c) for each of the tissues within each different permutation of hyperparameters. The horizontal axis labels denote the hyperparameter permutation in the order [***down-sampling factor, optimising function***].

Also clear are the higher variances when the F1 score is used as an optimisation function. This may be attributed to the fact that the F1 score is very sensitive to class imbalances. F1 is dependent on the PPV, which is the ratio of correct positive predictions to incorrect positive predictions (Eq. (2.3)). This means that if the number of spectra belonging to the positive class is much less than the number of spectra remaining in the negative class(es) (often the case for multi-class models), one would obtain a poor PPV which is not reflective of the true performance of the model. On the other hand, Youden's J statistic is calculated from the sensitivity and specificity, which are both insensitive to class imbalances.

The remaining classifiers for each of the tissue types all have consistently high performances across the different permutations, with the AUC scores consistently higher than 0.92. The permutation with a down-sampling factor of 2 and Youden's J optimisation function was selected for training of the final model, which was performed on the entire training set.

4.3.2 Results

The trained optimised model was used to predict the identity of every spectrum held out for testing. The scores for each tissue are presented below in table 4.4.

TABLE 4.4: Results on hold-out test data

Tissue	Sensitivity (%)	Specificity (%)
OSCC	84.3	92.1
NE	71.8	99.2
CS	94.3	97.5
NS	57.2	97.9
LYM	91.0	93.2
SM	85.3	90.3
BL	71.1	96.7

The sensitivities reported in table 4.4 appear to be distributed across a greater range (57.2% - 94.3%) than the specificity scores (90.3% - 99.2%). Importantly, the OSCC model performs well, with a moderately high sensitivity of 84.3% and high specificity of 92.1%. This means that there is a high degree of confidence that spectra labelled as OSCC have been correctly labelled as such. The slightly lower sensitivity implies lower, yet still good confidence that spectra not labelled as OSCC are truly not of OSCC origin.

The metrics which form the optimum set for each tissue can interpreted to investigate the wavenumber features that separate the different tissue types in the data. The wavenumbers that appear in the optimum set of metrics can be histogrammed in an importance plot, which indicates the relative strength each wavenumber feature in the spectrum has in the discrimination. Figure 4.12 is the importance plot for this multi-class study.

At this stage, an initial impression of the characteristic features of each tissue type can be obtained. It appears that many of the important wavenumbers are situated in the amide I and II regions, implying that protein content is a key discriminator. There is also a cluster of important features emerging at $\approx$ 1080 cm^{-1}, indicating the importance of DNA content in characterisation of cancerous tissue. The importance of

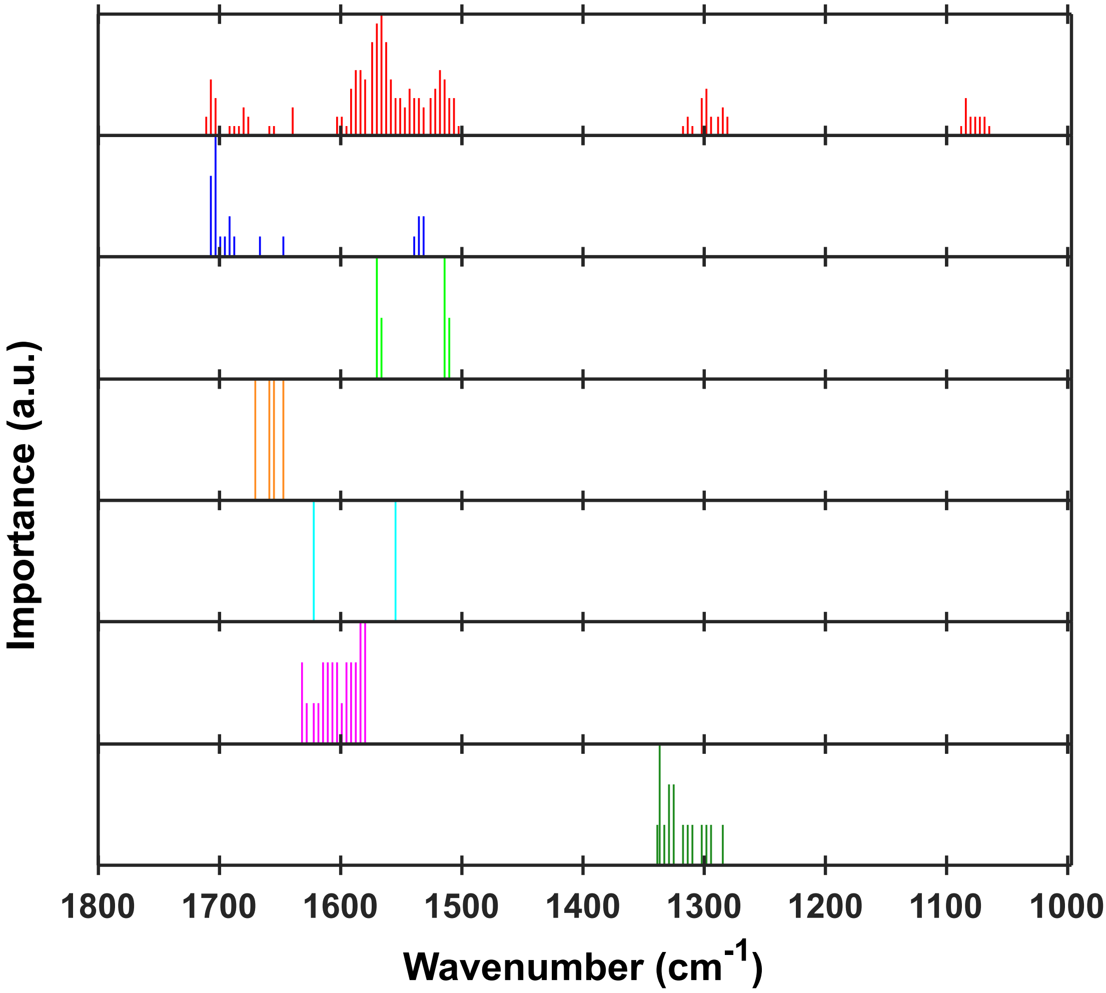


FIGURE 4.12: Importance plot for multi-class study. Tissue types from top to bottom: OSCC, NE, CS, NS, LYM, SM, BL.

wavenumbers ranging from 1285 cm^{-1} - 1340 cm^{-1} in OSCC and BL may indicate the importance of collagen in the discrimination of the two tissues from the rest of the set.

In order to gain further insight into the origin of importance for the top performing metrics, their distribution over the spectra for each tissue type can be plotted. table 4.5 reports the best metric $\delta_{\mu,\nu}$ for each of the seven tissues. Interestingly, the wavenumbers occuring in each metric are often from similar regions of the spectrum. This implies the analysis is identifying localised changes in spectral line shape as key discriminators. Figure 4.13 shows the distributions across the seven tissues for the best metric for OSCC and LYM.

The distributions in Fig. 4.13 show a good degree of separation for the positive class (OSCC in Fig. 4.13a and LYM in Fig. 4.13b) from the remaining six negative classes.

TABLE 4.5: Best metric for each tissue type in the multi-class study.

Tissue	**Top Metric**
OSCC	$\delta_{1518,1566}$
NE	$\delta_{1697,1703}$
CS	$\delta_{1514,1570}$
NS	$\delta_{1655,1658}$
LYM	$\delta_{1555,1622}$
SM	$\delta_{1591,1607}$
BL	$\delta_{1329,1333}$

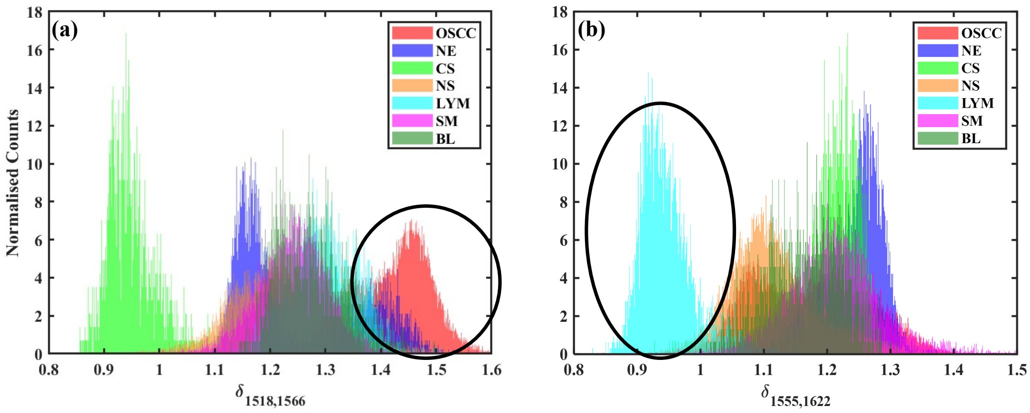


FIGURE 4.13: Distributions of the best metric for OSCC (a) and LYM (b). Positive class is circled in black.

Any spectrum that has a $\delta_{1518,1566}$ of greater than approximately 1.4 will be labelled as OSCC by the model, with a small fraction of false positives from the other classes (mainly NE). Similarly, LYM has a well separated distribution in Fig. 4.13b, whereby a spectrum that has a $\delta_{1555,1622}$ of less than approximately 1 will be labelled as LYM, again with a small number of false positives. It also appears that $\delta_{1518,1566}$ is a strong discriminator for CS, supported by the fact that a similar metric ($\delta_{1514,1570}$) is the strongest discriminator for CS. The distributions also help to explain why the specificity scores are consistently very high. Since the model for each tissue is a one vs rest classifier, the true negatives for metrics where the positive class is well separated will vastly outweigh the false positives, as the negative class contains all the spectra from $n-1$ classes.

The efficacy of the model can be demonstrated by using the trained model to label a hyperspectral image, which is fundamentally a two dimensional arrangement of

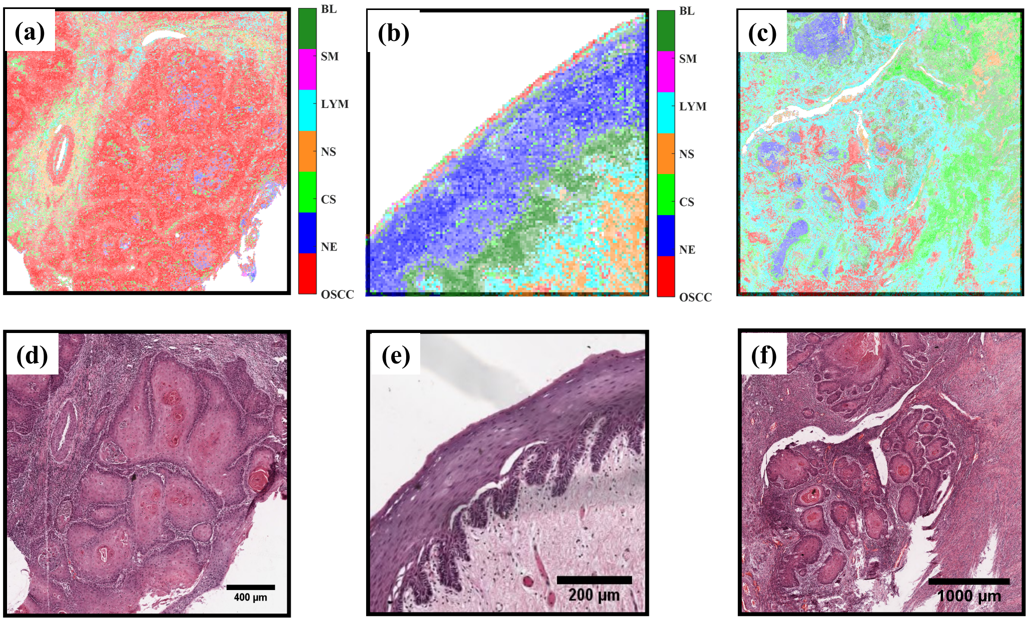


FIGURE 4.14: Backlabelled images. (a) and (b) were used to extract a portion of spectra used for model construction, (c) was not used for analysis. (d-f) are LM HE counterparts to (a-c) respectively for comparison.

spectra, generating a pseudo-colour map (where each colour denotes a tissue type). Shown below in Fig. 4.14 are a selection of pseudo colour images, two of which (a-b) are from images used to extract training spectra from and the other one (c) is from an image that was not used for training. H&E counterparts are shown in the second panel (d-f). The intensity of the colour indicates confidence of class assignment, found by dividing the probability the spectrum belongs to the assigned class by the sum of the probabilities of all other classes. The interpretation of spectral biomarkers and pseudo-colour images will be discussed in more detail in the subsequent section.

4.3.3 Discussion

The multi-tissue analysis has demonstrated that MA can attain good scores (table 4.4) when tasked with discriminating between different types of tissue. The pseudo-colour images also make evident the inherent skill of the trained MA model at identifying tissue types on the basis of the information used to build it. The highly interpretable nature of MA enables spectral biomarkers to be identified and further investigated, allowing for the postulation of chemical differences that give rise to the various features that drive the discrimination.

The scores associated with OSCC were high, with a sensitivity and specificity of 84.3% and 92.1% respectively. The optimum metrics, histogrammed in Fig. 4.12, show that the discrimination is primarily due to differences within the amide region, with additional contributions arising from regions further towards the lower wavenumber end of the spectrum at approximately 1080 cm^{-1}. The pseudo-colour image in Fig. 4.14a correlates with the morphology and pathology portrayed in its H&E counterpart shown in Fig. 4.14d, which consists of invasive OSCC arranged in large clusters, some of which have highly keratinised cores as a product of highly differentiated keratin producing OSCC. A minority of spectra within the invasive cluster are being labelled as other tissue types, mainly maturation layer (NE). This is not surprising considering both the small sample set and the similarities between 'normal' epithelial tissue and squamous cell carcinoma, given that they are both epithelial tissue. Due to the fact that samples reserved for FTIR imaging were from an adjacent, rather than identical, slice of the FFPE tissue block, the images produced by LM and FTIR-MS will never identically match each other, so it's impossible to make exact comparisons. It does appear, however, that the regions of NE labelling (blue) correlate with regions of higher keratinisation, where the H&E stains the tissue red. This may be because the normal epithelial cells within the maturation layer become increasingly keratinised as they migrate towards the superficial layer [116]. The false positive labelling of NE as OSCC is also evident in Fig. 4.13, where the distribution of NE spectra overlaps with the OSCC spectra in the region where the OSCC distribution dominates.

As previously stated, the collective term given to the supportive network of tissue is stroma, which contains functional cells such as lymphocytes and fibroblasts. Figure 4.14a indicates the presence of a mixture of predominantly CS (rich in fibroblast), LYM (rich in immune cells) and normal stroma (the underlying general supportive network). This is also true for the other two pairs of images. Figure 4.14b and e are FTIR and H&E images of non-malignant epithelium adjacent to normal stroma supportive tissue. The majority of the epithelial thickness has been labelled as NE, which is in agreement with the histopathology in Fig. 4.14b. Spectra from the deepest layer of the epithelium have been correctly labelled as BL, before transitioning to LYM followed by NS. Broadly, this agrees with the histopathology (Fig. 4.14e), where the basal layer of the normal epithelium protrudes into the normal stroma in structures known as rete ridges, with a higher density of immune cells (LYM) surrounding the epithelial border. On the other hand, the borders of the progenitor (BL) and keratinised maturation (NE) compartment of the normal epithelium differed between Fig. 4.14a and d. This can be attributed to the reduction in spatial resolution, potential mismatches in serial section content and small number of training images available for training.

Figure 4.14c and f are the pseudo-colour and H&E images of an ROI which was not used to extract spectra for model construction. The H&E image displays invasive carcinoma underneath the tear in the specimen near the centre of the image. This histology is not exactly reflected in the pseudo-colour image, where there is pronounced confusion between maturation layer, OSCC and basal layer. These tissue elements are all epithelial in origin, and given the relatively small sample set and heterogeneity of tissue it is not surprising that there is confusion between the tissue types. A more robust labelling may be achieved if a higher number of patients and images were used for training, with much more selective labelling. Selective labelling is very difficult when using a high resolution LM image of tissue morphology to co-register precise regions of the hyperspectral image, as there is a significant drop in spatial resolution due to the diffraction limit of longer wavelength IR radiation. Furthermore, the use of serial sections for image registration is challenging due to the lack of certainty that abnormalities in one image will be present in the other. One solution to this would be to image exactly the same section with LM and FTIR, but this would introduce new challenges associated with sample preparation. As detailed in section 4.2.1, H&E staining requires that the specimen is dewaxed, whereas preserving the tissue in wax alleviates issues related to the sample longevity and scattering efficiency in FTIR imaging. Standard glass slides used in histopathology have a positive charge to aid sample adhesion and prevent risk of ‘float’ of tissue, so the transfer of the specimen from the FTIR imaging disk to the slide is necessary, especially as the staining process involves several steps which may disturb the structural integrity of the sample. A study by Pilling *et al* [117] in 2016 showed that high scores can be achieved when discriminating between different tissue components that have been imaged using FTIR-MS on stained glass samples, a promising finding which elevates the feasibility of FTIR-MS as a clinical augmentation, due to the improvements in throughput, cost and image registration. The main drawback with utilising this approach is the fact that glass is opaque in the information rich fingerprint region, therefore scope for spectral biomarker interpretation is significantly diminished.

The origin of the best OSCC metric ($\delta_{1518,1566}$) may be the presence of shoulders on either side of the amide II peak, which would indicate the shifts of smaller underlying peaks . The absence of a shoulder either side of the amide I peak for CS (Fig. 4.15) is unique relative to the other tissues, which explains why the distribution of CS is so well separated from the remaining tissue distributions. There is a degree of ambiguity with respect to whether the shoulders present in the mean spectra arise as a result of tissue heterogeneity in the labelled regions. The metric histograms shown in Fig. 4.13 indicate that when the metric is the best distribution for a certain tissue type, then the distribution tends more towards a Gaussian shape, which indicates similarity between labelled spectra at high performing wavenumbers.

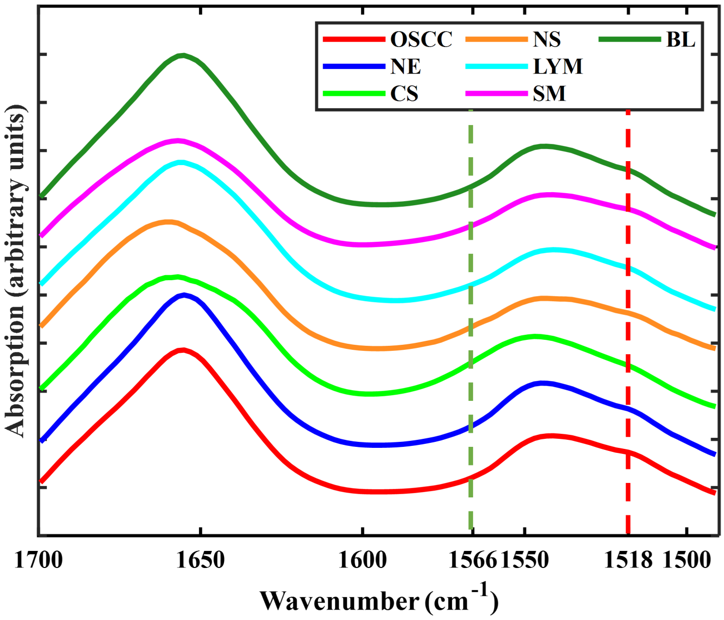


FIGURE 4.15: Tissue mean spectra in the region 1500 - 1700 cm^{-1}. Best metric for OSCC has been drawn with two dotted lines.

Interpretation of spectral biomarkers from a multi-class study is believed by some to be somewhat ambiguous [118]. The subtle differences between classes, convoluted nature of the IR spectrum and the ensemble of metrics required to achieve good scores renders the task of directly correlating results with spectral biomarkers challenging. With this in mind, it should be re-iterated that interpretation of multi-class models such as these is rather speculative, with more definitive information available when interpreting models with fewer classes. Nevertheless, the utility of MA as a so-called 'black-box' labelling tool has been demonstrated, and shows promise for the visualisation of tissue structure as a function of IR absorption.

4.4 Phase II & III: Binary Metastasis in Lymph Node Analysis

For reasons outlined in the previous section, interpreting multi-class model results is somewhat challenging and ambiguous. This section will focus on a more directed approach, whereby a new MA model will be trained on a small set of spectra originating from two distinct classes in the dataset. These two classes were chosen to be OSCC that has metastasised to the lymph nodes (MT) and lymphoid nodal tissue (LYM), due to their close proximity and sharp delineation between the two tissues in H&E images.

Given that MT and LYM tissue can be easily distinguished using histopathology, any results will bear little clinical significance. The main motive behind this study is to obtain a set of defining wavenumbers for translation to discrete frequency IR-SNOM imaging, which will be described in much more detail later in this section. The material covered will therefore provide insight into two novel techniques in MA and aperture IR-SNOM.

4.4.1 Spectral Biomarker Extraction

The experimental procedure leading up to the training of the MA model was identical to that of phase I of this work. For the model training, a modified approach to the one used previously in section 4.3.1 was used, whereby no hyperparameters were optimised as every accessible resolution element in the data was to be used to search for the optimal metric.

Data from both images (detailed in table 4.2) were used to train a single MA model. As shown in (table 4.6), the resultant model could discriminate between MT and LYM with near perfect scores by using only one metric, specifically $\delta_{1252,1285}$. Similar to the multi-class study, the discriminatory power of the resultant metric can be visually demonstrated by plotting the distributions of $\delta_{1252,1285}$ for each of the two tissues (Fig. 4.16c). Immediately evident is the high degree of separation between MT and LYM, which is manifest in the scores contained within table 4.6. Also interesting is the additional discriminating power acquired when calculating a ratio between two wavenumbers, compared with solely the absorbance at each of the two wavenumbers (Fig. 4.16a,b).

TABLE 4.6: Results for MT vs LYM binary model

Highest ranked metric	$\delta_{1252,1285}$
Sensitivity	98.7%
Specificity	99.9%
AUC	0.99

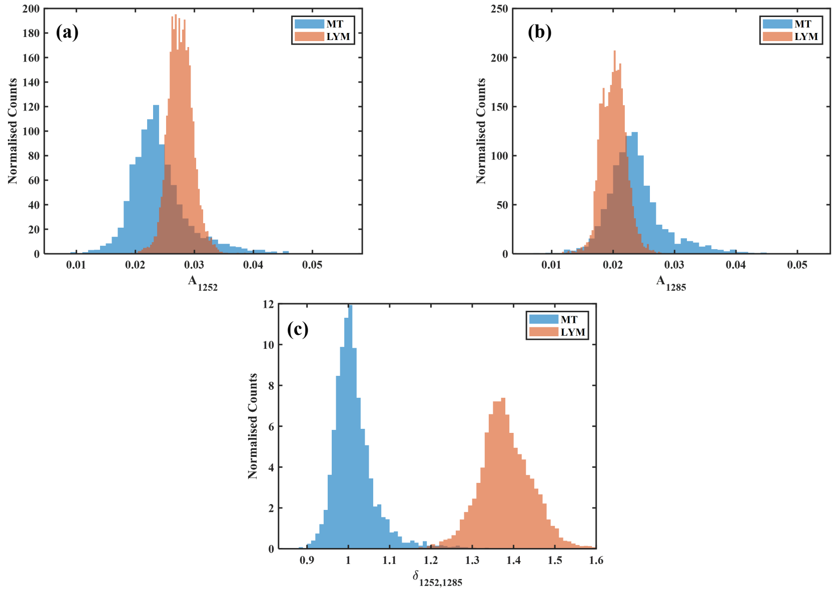


FIGURE 4.16: Distributions of absorbance at (a) 1252 cm^{-1} and (b) 1252 cm^{-1}, which both form the best metric $\delta_{1252,1285}$, the distribution of which is shown in (c).

Instead of using the metric $\delta_{1252,1285}$ as a component in an ensemble classifier, it is possible to measure the performance of univariate classifiers based solely on the absorbance of the component wavenumbers A_{1252} and A_{1285}. This can be achieved by treating their distributions (shown in Fig. 4.16a and b) as probability density functions.

The sensitivity and specificity for $A_1$252 would be 89.3% and 73.3% respectively, and for the A_{1285} the same measures would yield 90.4% and 54.4% respectively.

4.4.2 IR-SNOM Instrument

The aperture IR-SNOM (herein referred to as SNOM) instrument at the University of Liverpool, UK was used to exploit the characteristics of the near-field as described in section 3.4. The SNOM in its current configuration is a bespoke amalgamation of a benchtop quantum cascade laser (QCL) IR light source, an SPM system loosely based on constant distance AFM to control the lateral and vertical positioning of the probe, and an integrated IR detection system. The current section will detail each aspect of the instrument, which can be subdivided into different systems:

(i) SNOM head system: Components which control the height of the tip as it scans the sample.

(ii) QCL IR source: Production of IR light and propagation towards sample and detector.

(iii) Optical fibre and detector: Detection of light and conversion into a signal that can be recorded in an image.

(iv) Stage and microscope: The piezoelectric stage that controls and monitors the precise positioning of the sample beneath the probe. The light microscope is also used to spatially locate regions of interest on the sample.

(i) SNOM Head System

Contained within the SNOM head system are the components that detect and control the probe-sample separation. The probe is the primary detector in the SNOM, and if a tapered optical fibre is used, both the topography and the near-field can be detected. The probe is adhered with glue to a driving bimorph and sensing bimorph. A bimorph contains piezoceramic materials, which either oscillates when an AC voltage is applied, or produces an AC voltage during oscillation. Each bimorph contains two piezoceramics separated by a nonactive layer of titanium. One side of the driving bimorph has an AC voltage applied to it, whilst the other side has the inverse

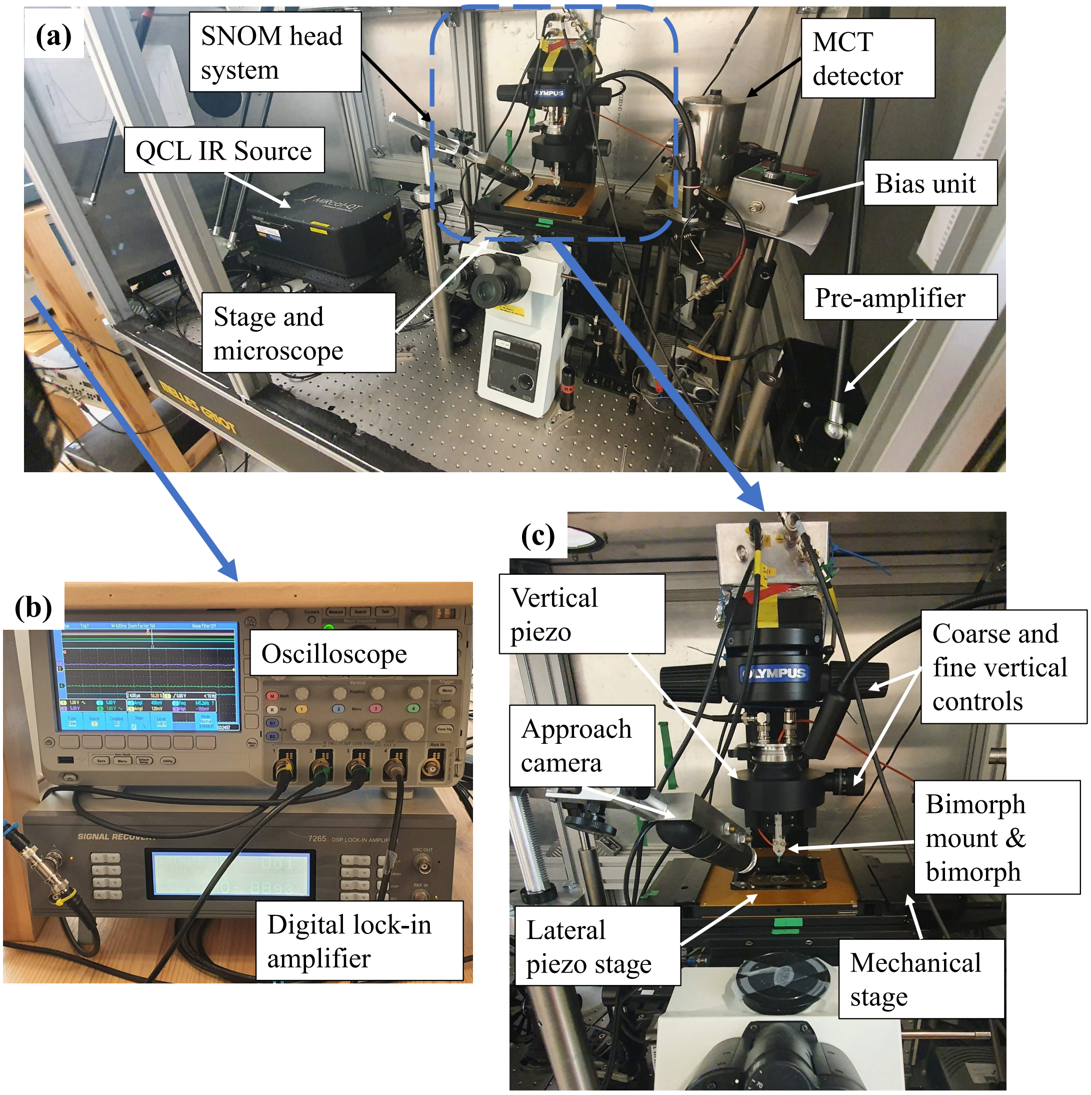


FIGURE 4.17: Labelled photographs of the SNOM instrument used in this work. (b) is a photograph of the oscilloscope unit that is used to monitor multiple signals from the QCL, MCT and lock-in; and the lock-in amplifier which is used to isolate a periodic signal from noise with a continuous frequency spectrum. (c) is a zoomed in view of the SNOM head/stage and microscope system.

(anti-phase) of the AC signal applied, so that when one side contracts the other expands, forcing the bimorph to oscillate. The amplitude and frequency of the system is detected by the sensing bimorph, and can be monitored by the SNOM software.

The tip-bimorph system has a resonance frequency dependant on the distribution of mass throughout the system. The resonance frequency can be found by adjusting the driving frequency until the amplitude of the signal produced by the sensing bimorph is a maximum. When the tip is brought close to a surface, the amplitude is dampened

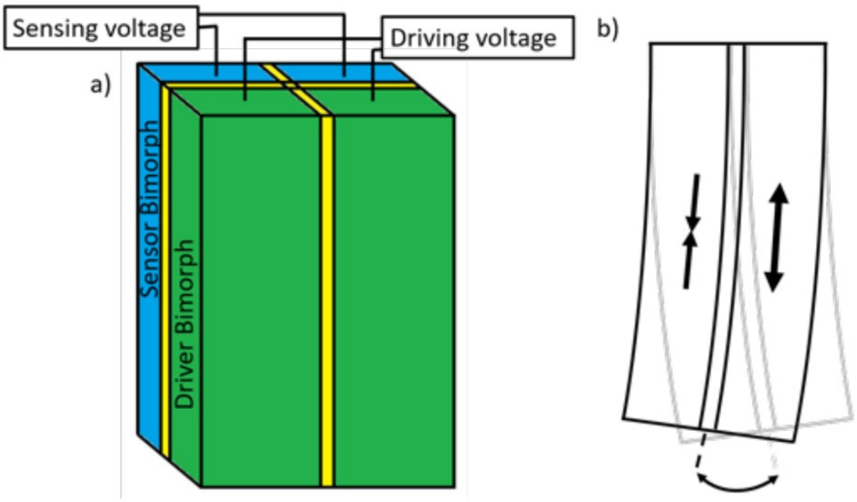


FIGURE 4.18: a) Schematic diagram of the driving and sensing bimorph which are driven into oscillation by an external voltage. Extracted with permission from James Ingham's PhD thesis [119]. (b) Shows how the contraction and expansion of either side of the driving bimorph results in oscillation.

by shear forces that exist very close to the sample [120], this drop in amplitude can be registered on the SNOM software. In order to implement a constant distance feedback system, a reference amplitude can be set so that the vertical position of the SNOM head system (controlled by the z-piezo) can be adjusted to maintain the reference amplitude of the sensing bimorph signal. The response speed of the z-piezo to the changes in amplitude of the bimorph can be controlled by adjusting the gain. For samples with large variations in height at high spatial frequencies, the gain should be set high so that the tip responds quickly to sudden changes in topography, reducing the risk of collision with the surface. For flatter samples it is advisable to keep the gain low so that small variations in height aren't over corrected for, which would introduce more noise into the image.

(ii) QCL IR Source

QCLs are a relatively recent technology which are able to lase mid-IR photons over a broad, tuneable range. First demonstrated by Faist *et al* at the AT&T Bell laboratories in 1994 [121], based on a principle first proposed by Kazarinov and Suris in 1971 [122]. They are fundamentally different to semiconductor diode lasers, which emit photons when high energy conduction band electrons and low energy valance band holes recombine in a p-n doped semiconductor. The recombined electron occupies a lower energy state and this excess energy is emitted in the form of a photon. This spontaneous emission of a photon will drive the recombination of electron hole pairs in close proximity, causing the stimulated emission of another photon with the same phase, frequency and polarisation, hence 'lasing' coherent light. The energy (and therefore wavelength) of the photon is solely dependant on the energy gap of the semiconductor, therefore are not instantly tunable.

The aforementioned diode lasers operate through interband transitions which are responsible for the emission of a photon. QCLs, on the other hand, are instead driven by intersubband transitions. Rather than using a single bulk semiconductor as the lasing medium, QCLs consist of a periodic arrangement of thin layers of materials which have different composition, known as a *superlattice*. The superlattice imposes a non-uniform electric potential, which results in the probability of an electron occupying a certain position to vary along the length of the device. This multiple quantum well confinement has the effect of splitting of the band into discrete subbands. When an electron traverses across the medium, it transitions into a lower subband, resulting in the emission of a photon. This electron can then tunnel into the next layer, where the above process is repeated. This process is where the terminology 'quantum cascade' originates. The wavelength of the emitted photon depends on the layer thickness, rather than the material bandgap as for diode lasers. This property enables the broad tuning of QCLs.

The QCL source used for SNOM experiments is a MIRCat-QT$^{\mathrm{TM}}$ equipped with four modules that enables selective tunability across a wide spectral range. The laser comes equipped with three modules which span a unique spectral range, with an effective tunable range of approximately 1150 cm^{-1} - 2000 cm^{-1}. The power output spectrum of the QCL is shown in Fig. 4.19. The QCL was operated in pulsed mode, whereby the device would lase at a certain pulse rate (frequency, not to be confused with frequency of emitted light). The laser would be triggered by either an external or internal square wave trigger which instructs the laser to begin lasing for a defined time, known as the pulse-width. The ratio of the pulse width of the laser to the period of the square wave trigger is called the duty-cycle, which should be kept as low as 5% to prevent overheating and preserve the QCL.

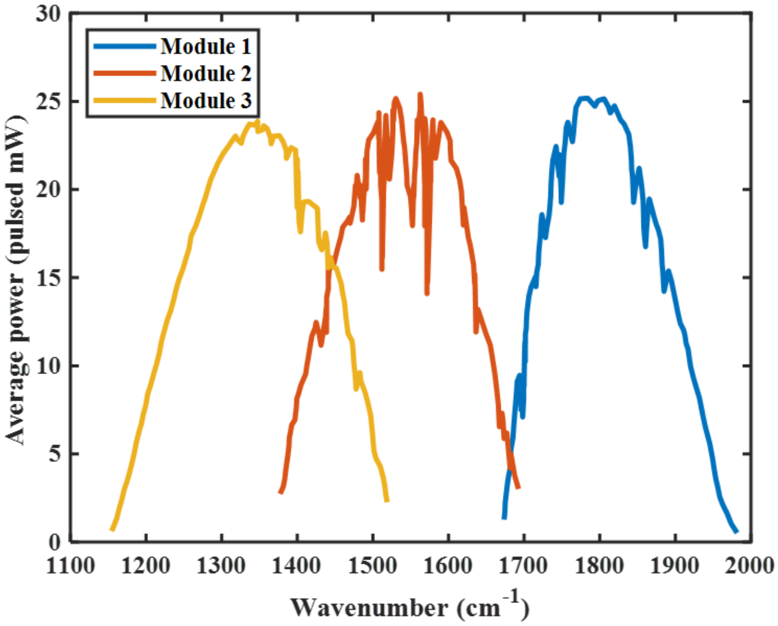


FIGURE 4.19: QCL emission spectrum.

The beam emerging from the QCL was directed towards the sample compartment by using a single Au mirror with an SiO_2 protective layer positioned directly beneath the sample. The mirror has an IR reflectance of $> 96\%$, making it an ideal canditate to direct the IR beam towards the sample in an efficient manner. The angle of the mirror was configured to be 45° so as to reflect a ray that is perpendicular to the incident ray. The lateral position of the mirror can be manipulated with coarse and fine control in order to adjust the position of the beam. Figure 4.20 is a schematic diagram of the SNOM.

(iii) Optical Fibre and Detection

The use of optical fibres for the transmission of visible wavelengths in sectors such as telecommunications has been realised for some time, however only until recently have optical fibres that are transmissible in IR wavelengths become commercially available. The fibre used in this work (Coractive, USA) has a 100 µm inner core fabricated from selenide glass, which is surrounded by a 170 µm layer of cladding which protects the

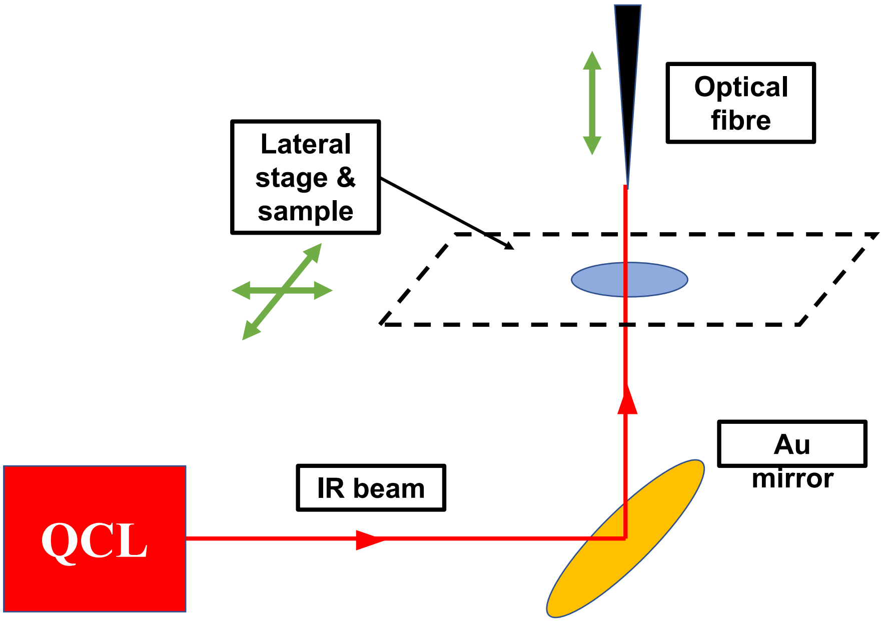


FIGURE 4.20: Optical path taken by the IR beam from the QCL to the optical fibre tip.

inner core and increases transmission efficiency. The inner core and cladding are encircled by an acrylate layer and a plastic sheath which act to stengthen and protect the fragile fibre from outside interferences.

The optical fibres direct the light towards the detector using the principle of total internal reflection. Snell's law (Eq. (3.7)) can be rearranged to account for the condition that the transmitted ray is refracted by an angle of at least 90°, so that it is effectively reflected back into the medium with refractive index n_1.

$$n_1 \sin(\theta_i) = n_2 \sin(90^\circ) \tag{4.3}$$

By considering that $\sin(90^\circ)$ is 1, and terming the incident angle for total internal reflection as the critical angle θ_c, Eq. (4.3) can be rearranged for θ_c in Eq. (4.4), setting up the condition that the refractive index of the second medium must be greater than that of the first medium in order for total internal reflection to become a possibility.

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{4.4}$$

In order for the near field to be detected, the fibre was etched to create a nano-antenna as described in section 3.4. The outer plastic layer and fibrous layer were first stripped to expose approximately 5 cm of the acrylate clad fibre. The fibre was subsequently mounted onto a fixture prior to submersion in dichloromethane for 10 minutes until the acrylate swells and softens. The acrylate was then removed with wire strippers to expose the inner two layers (inner core and cladding). 4200 µl of sulphuric acid was then mixed with 1800 µl of hydrogen peroxide to form so called 'piranha solution', a strong corrosive substance used to dissolve a variety of materials. 600 µl of tetramethylpentadecane (TPMD) was then added to form a thin suspension on top of the piranha. 2 mm of the exposed fibre was immersed in the solution to commence the chemical etching of the fibre. The boundary between the piranha and the TPMD layer create convection currents which force the acid to erode the fibre near the boundary, so that a regular, sharp apex can be formed.

The output of the optical fibre was attached to an MCT detector, which is described in more detail in section 3.2.2. The MCT detector is housed within a vacuum sealed dewar filled with liquid nitrogen, to noise induced by thermal excitations in the semiconductor detector. A bias current of 7.5 mA was applied to the MCT so that a measurable electric current could be detected after the generation of electron hole pairs after photon interaction. Weak signals travelling through coaxial wires are prone to low signal to noise ratios, so the signal was amplified immediately after the output of the MCT. The pulsed signal from the MCT was fed into a lock-in amplifier, which isolated components of the signal with the same frequency as the reference (the same square wave used to trigger the QCL). The sensitivity of the amplifier could be adjusted to reflect the magnitude of the input signal from the MCT, which was heavily dependant on the wavelength (see Fig. 4.19) and pulse-width of the QCL.

The alignment of the QCL beam with the optical fibre is an essential step to ensure maximum efficiency in coupling the fibre with the optical field. The QCL is equipped with an auxiliary helium-neon (HeNe) red laser which is mutually aligned with the IR optical path. This is used to guide the coarse adjustments required to bring the beam into coincidence with the apex of the tip. Once this has been achieved, the optics were switched back to IR in order to maximise the throughput of the light up the fibre. Through close monitoring of both the output of the MCT and the output of the lock in amplifier, fine adjustments to the lateral position of the mirror were made until the signals were maximised, indicating the maximum throughput up the optical fibre.

(iv) Stage and Microscope

The sample housing is mounted upon an inverted optical microscope equipped with four objectives (4×, 8×, 16× and 32×) and a manual mechanical stage. The microscope was important so that the end of the tip could be visually aligned with the region of the sample that is to be scanned, as well as to locate the specific region of interest within the sample.

The mechanised sample stage was manual and with poor spatial precision, making it unsuitable for SPM measurements, where reliable and precise knowledge of the sample location relative to the tip is crucial. Instead, the fine lateral position of the sample is monitored and controlled by a piezoelectric actuator, a device which is has been significantly responsible for the emergence of SPM due to its high displacement resolution, rapid response and large actuating force [123]. The device, herein called the 'lateral piezo stage' has a range of motion of 500 µm, and is bi-directional, meaning it is able to scan both forwards and backwards at the same rate, enabling the acquisition of two sets of images. The relative motion of the tip relative to the sample is depicted in Fig. 4.21. It is important to note that the lateral position of the probe is fixed relative to the observer, it is the lateral piezo stage that moves.

Despite the high spatial resolution, the positional accuracy of the device is severely hampered by the existence of hysteresis, where there is a non-linear relationship between the driving voltage (and hence perceived position) and the true position of the stage. In the context of SPM, where the size, shape and position of structures within an image is of interest, it is important that these distortions induced by hysteresis are corrected for. Since the piezo stage is moving in opposite directions for each pair of scans, an independent correction for both forwards and backward scans is required.

In order to calibrate the response of the lateral piezo stage in both directions, a calibration sample containing features with known dimensions (Fig. 4.22) was used. Regularly spaced gold rectangles of well defined size were placed atop a silicon base so that a calibration curve of real position to measured position could be measured. Since

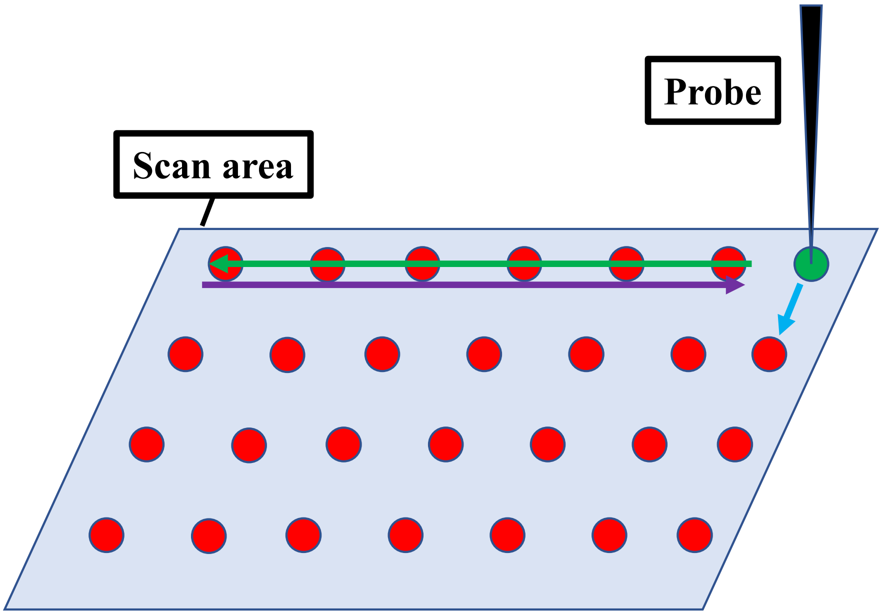


FIGURE 4.21: Path taken by probe relative to the sample surface. For the forward scan (green), the position of the sample is stepped to the right until the end of the line, at which point it is stepped all the way back to the left for the backwards scan (purple). At each point (red dot), the signal is recorded so that an image can be formed. At the end of each line, the stage steps up so the line beneath can be obtained.

only the topography was to be measured, a small offcut of sharp metal wire rather than an etched optical fibre was used as the probe. Figure 4.23 shows the forward (a) and backward (b) topography scans of the calibration grid.

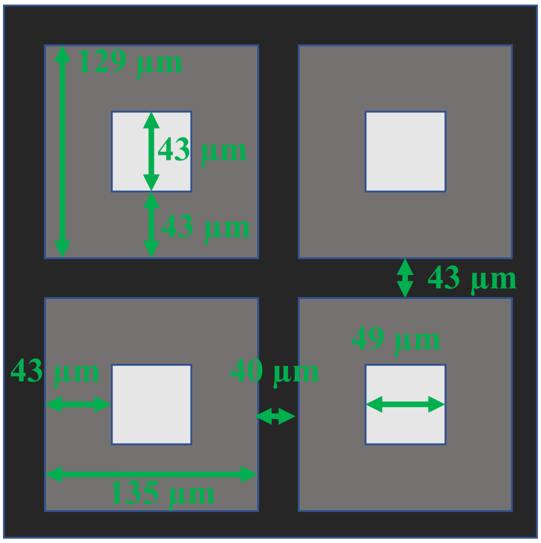


FIGURE 4.22: Size of the features of the hysteresis calibration grid.

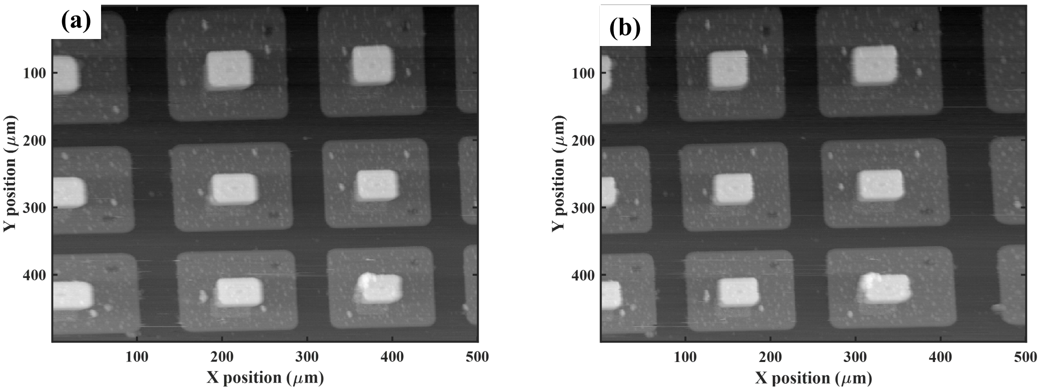


FIGURE 4.23: Forwards (a) and backward (b) topography scans of the calibration grid.

Inspection of Fig. 4.23 reveals glaring vertical and horizontal distortions in both directions. The sample also appears to have been scanned at a slight rotational offset, which should be taken into account when calculating relative positions on the image. Three lines intersecting each row and column of the calibration grid were used to plot the observed positions as a function of the true positions, which are shown for the horizontal and vertical directions in Fig. 4.24. From Figs. 4.23 and 4.24, it is noticeable that the extent of hysteresis depends on the direction of the scan. In the forward scan, where the probe scans from left to right (relative to the images), the distortions are much more pronounced towards the left of the image, whereas the opposite is true for the backwards directed scan. The vertical distortion is virtually identical for both forward and backwards as the probe steps vertically in the same direction (top to bottom relative to the image) at the end of each line.

A third order polynomial was fitted to build a regression model for each of the curves. The model was used to transform the original images into a new set of images that had been corrected for hysteresis, shown in Fig. 4.25. The result is that the feature sizes are consistent across the image, indicating that the calibration has effectively corrected hysteresis in the images.

4.4.3 Experimental

The key spectral biomarkers determined from the MA model in phase II were used to guide the experiments on the IR-SNOM instrument in phase III. The instrument in

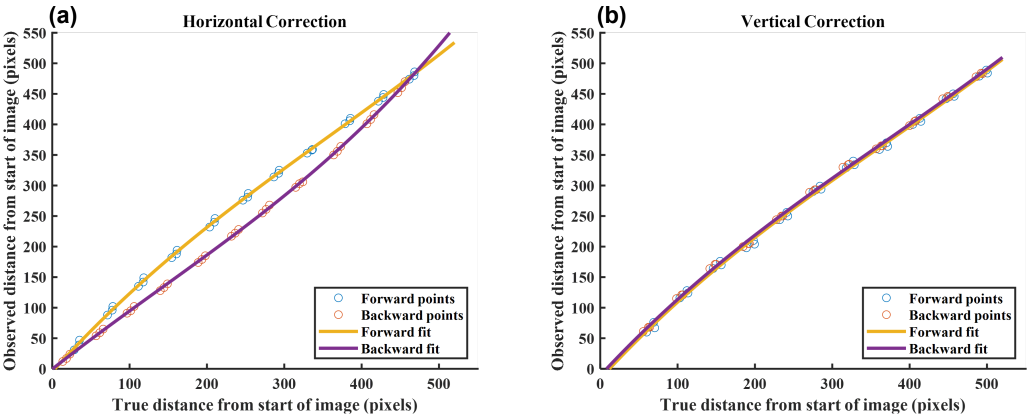


FIGURE 4.24: Curves of observed position as a function of true position for horizontal (a) and vertical (b) orientations.

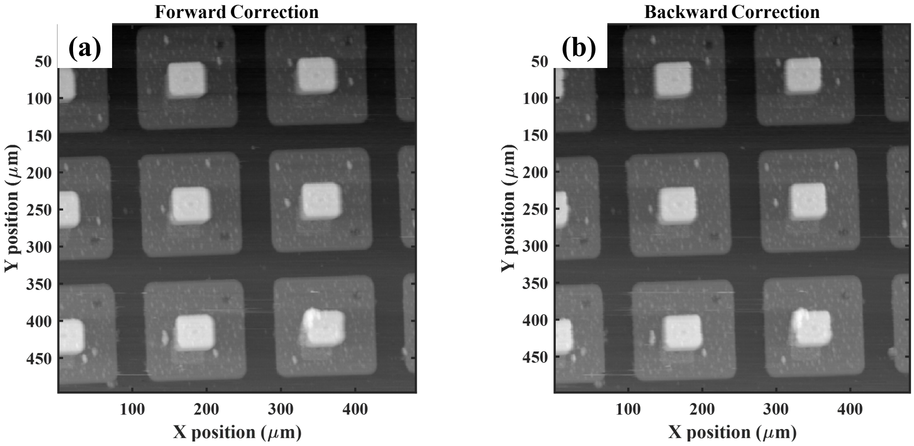


FIGURE 4.25: Corrected forward (a) and backward (b) scans.

its current configuration operates in discrete-frequency (DF) mode, but future studies with the instrument may utilise the integrated spectroscopic capabilities of the QCL. In addition to the pair of spectral biomarkers (1252 cm^{-1} and 1285 cm^{-1}), wavenumbers associated with strong contributions from key biomolecules were also selected for study, shown in table 4.7.

TABLE 4.7: Wavenumbers selected for SNOM study

Wavenumber (cm^{-1})	**Associated Biomolecule**
1252	MA
1285	MA
1650	Protein [124]
1369	Nucleic Acids [124]
1751	Lipids [124]

Sample Preparation and Model Validation

The large tissue specimens imaged in the FTIR experiments were sub-optimal targets for imaging with SNOM. This is due to the much smaller FOV of the SNOM, which is directly related to the range of motion of the lateral piezo, which is fixed at 500 μm. The FTIR microscope utilised in these experiments came equipped with tools that aided the co-registration of IR image with the H&E image, such as the mosaic function and integrated visible microscope. As such, small 1 mm diameter cores were extracted from the same tissue sample that was imaged with the FTIR microscope using a Beecher MTA-1 tissue microarrayer, so that features can be compared across images with a heightened degree of confidence. Serial sections were extracted from the constructed tissue microarray (TMA) block, consisting of two sections on CaF_2 disks for FTIR IR-SNOM experiments, sandwiched between two pairs of sections on glass slides stained with H&E and IHC for pan-cytokeratins using the AE1AE3 antibody (Agilent DAKO, Stockport, UK) and a Bond RX^{TM} autostainer (Leica Biosystems, Milton Keynes, UK). The histopathology and IHC slides were scanned using the same apparatus as for the H&E images in phase I.

A bespoke dewaxing methodology [125] was implemented in order to remove the paraffin from the section prepared for SNOM experiments, where it was important that the wax is thoroughly removed from the sample for a number of reasons. One of the target wavenumbers, 1369 cm^{-1}, is obfuscated by contributions from paraffin spectra, therefore for valid interpretation of images at this wavenumber there should be no paraffin present. The surface of paraffin embedded sections are smooth and flat as a result of the wax filling the spaces between structures within the tissue. This is problematic for an SPM instrument that operates in constant distance mode such as the SNOM described here, as the true topography of the tissue will not be followed so the strength of signal from the real sample will vary across the image.

Dewaxing an FFPE sample with a hydrocarbon solvent such as xylene will inevitably remove some of the native hydrocarbons from the tissue itself. Free, unbound lipids will be removed by the process, whilst solvent resistant lipids remain in the sample as they are effectively locked into protein-lipid complex matrices. The residual lipids are detectable by spectroscopy and may therefore still be of diagnostic use [126].

The LM images (H&E and IHC) of the sample chosen for SNOM imaging are shown in Fig. 4.26. The images reveal specific loci of tumour in both cores, stained dark brown for IHC and red for H&E. A positive (brown) stain in (b) indicates the presence of epithelial tissue due to the well differentiated and heightened expression of cytokeratins in the epithelium [127]. The areas contained within the green squares were chosen for SNOM imaging as they consisted of well defined regions of lymphoid tissue, epithelial metastasised tumour and a highly keratinised core, and it is hypothesised that contrast at spectral biomarkers associated with proteins and DNA should be observed in subsequenct SNOM images. The metric $\delta_{1252,1285}$ has also been trained to distinguish between metastasised tumour and lymphoid tissue, so it is instructive to investigate the contrast using a different imaging modality using this representative region.

In order to visualise the discriminatory power of the derived spectral biomarkers on regions not used in training, an FTIR image of the core shown in Fig. 4.26a and b was obtained using the same methodology as described in section 4.2.2. The corresponding absorbance and ratio maps at the discriminatory wavenumbers is shown in Fig. 4.27, which reveals a high degree of correspondence between the ratio image and IHC image in Fig. 4.26.

SNOM Experiments

The light microscope was used to align the upper right corner of the region of interest with the apex of the probe, so that it is covered by the range of motion detailed in Fig. 4.21. The bimorph-probe system was driven into resonance frequency, which was found to be 4.688 kHz with a non-dampened amplitude of 0.020 V. The reference amplitude was defined as 95% of the original amplitude (0.018 V), so that the vertical piezo in the SNOM head will act to restore this amplitude by adjusting the vertical position of the probe relative to the sample. In order for the feedback to activate, the SNOM head was lowered with a fine adjustment screw until the dampened amplitude was equal to the reference amplitude, at which point the tip was described as being ‘in contact’ with the surface of the sample. During the tip lowering, it was important that the gain was high to facilitate the rapid response between the bimorph and lateral

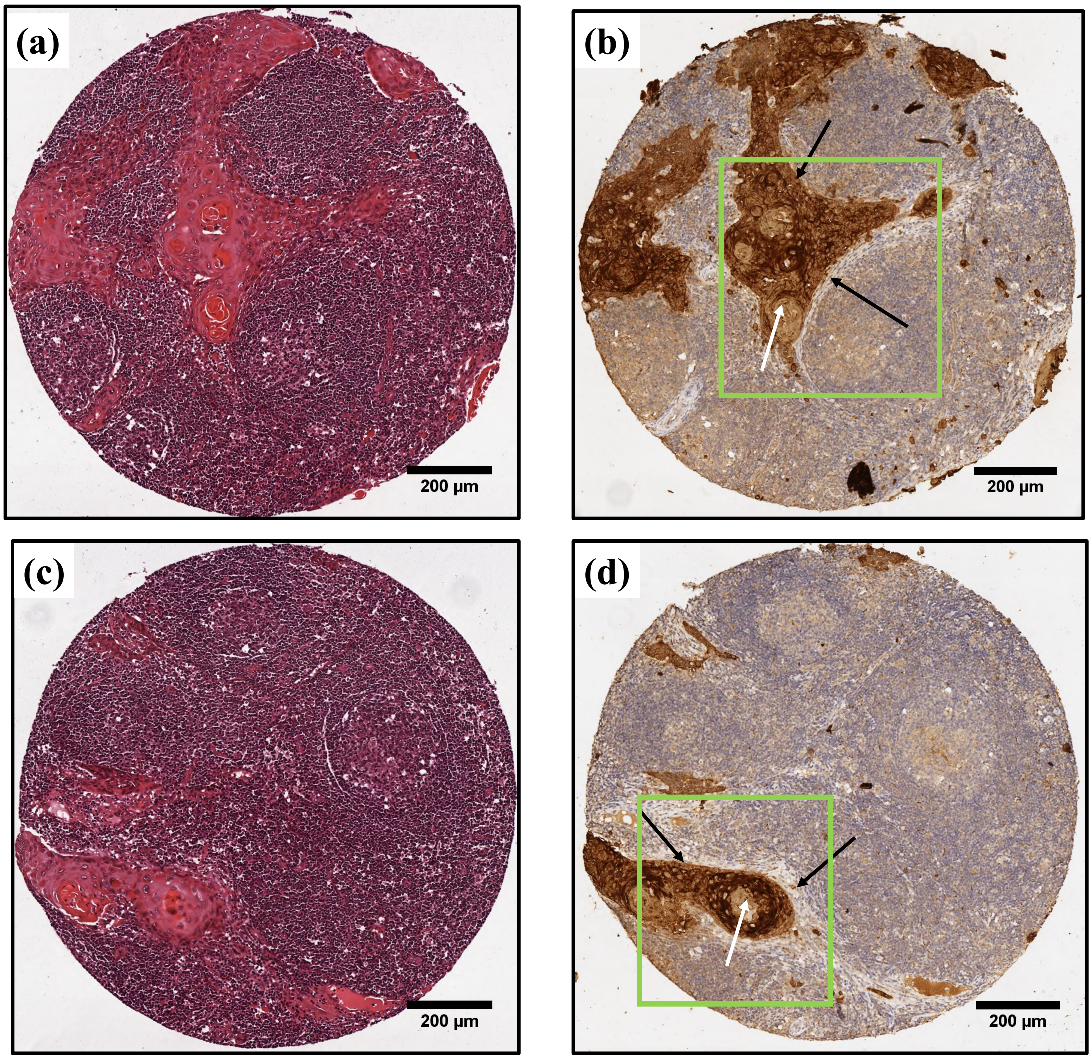


FIGURE 4.26: H&E (a,c) and IHC (b,d) images of the cores extracted for study. Black arrows indicate periphery of tumour, white arrow indicates keratin pearl within tumour. The green boxes indicate the region of interest (ROI) that will be targeted.

piezo, however the gain was lowered to a level where the voltage applied to the vertical piezo did not diverge far from equilibrium. The nature of the sample surface (flat, homogeneous tissue) justified the choice of low gain, as rapid corrections in height are not necessary.

The sample-probe alignment was again checked by ensuring MCT output when the QCL was active. After this, the QCL was tuned to the wavenumber of interest (see table 4.7), with the QCL and detector parameters for the specific wavenumber adjusted to optimise signal. The pulse width of the QCL was adjusted in conjuction with the sensitivity of the lock-in amplifier to acquire a stable output from the lock-in amplifier

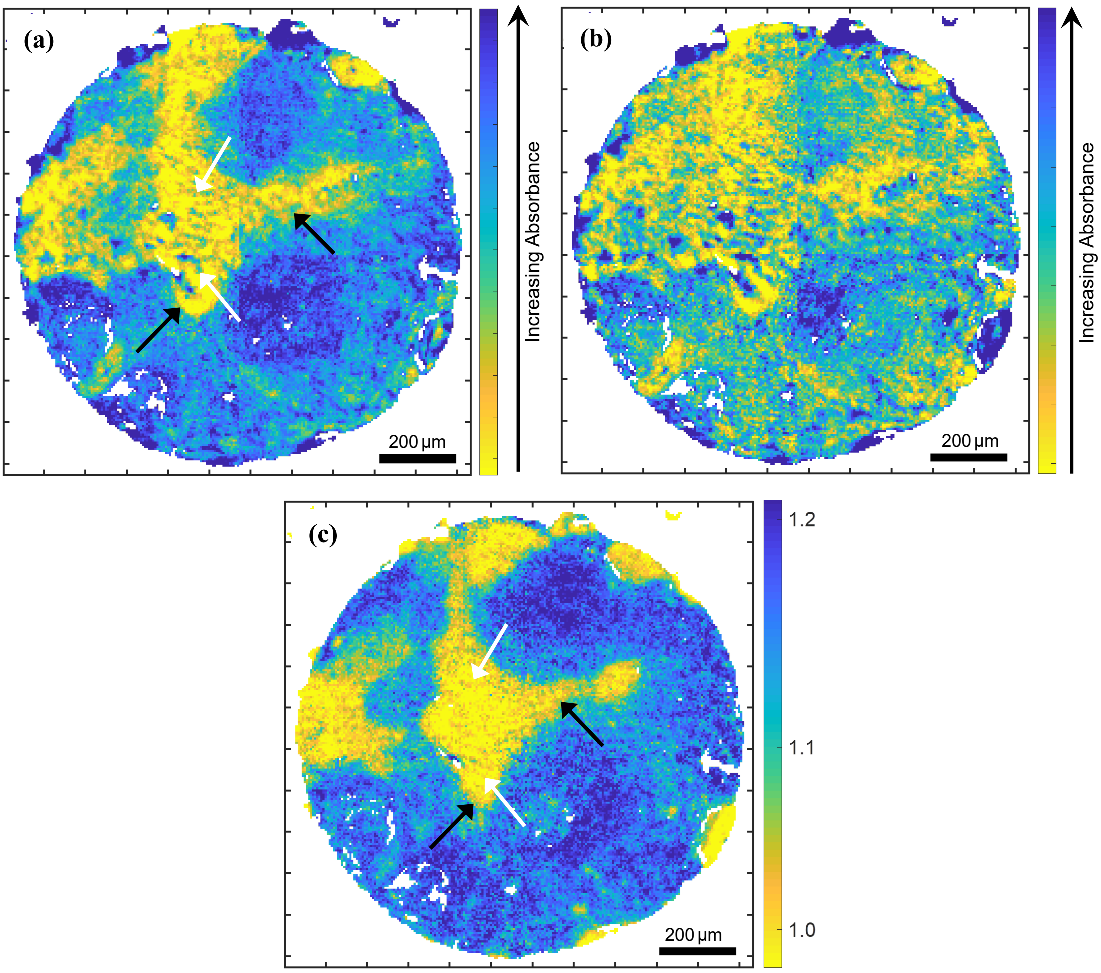


FIGURE 4.27: Core 1 absorbance images at (a) 1252 cm^{-1} and (b) 1285 cm^{-1}. The ratio between the two images is shown in (c), corresponding the optimum metric emerging from section 4.4.1. Black and white arrows used to indicate the same features as in Fig. 4.26b and d. Each FTIR image is plotted with a colour table covering the 5th to 95th percentiles of the image intensity range.

that is well below the limit of the analogue to digital converter of ≈10 V. Table 4.8 shows the parameters selected for each wavenumber. For all wavenumbers, the QCL was configured to pulse at a frequency of 80 kHz, corresponding to a period of 1.25 µs. The time constant of the lock in amplifier was also an important consideration, as it represents the integration time over which the signal is determined. If the time constant is too small, the signal varies too sharply and rapidly so an accurate reading cannot be acquired. On the other hand, large time constants result in a poor response time between changes in the MCT signal and the output of the lock-in amplifier. The time constant was chosen to be 10 ms as this offered stable readings with sufficient response speed. The lateral piezo was configured to scan a 300 µm × 300 µm area at a scan rate of 20 Hz and step size of 2 µm.

TABLE 4.8: Wavenumber specific experimental parameters.

Wavenumber	**QCL pulse width**	**Lock-in sensitivity**
1252 cm^{-1}	200 ns	20 mV
1285 cm^{-1}	200 ns	50 mV
1650 cm^{-1}	500 ns	20 mV
1369 cm^{-1}	200 ns	20 mV
1751 cm^{-1}	200 ns	10 mV

Image Processing

The raw images acquired from the instrument were not immediately comparable to one another due to differences in power, tip artefacts and differences in registry between different scans. Simlar to FTIR spectroscopic analysis, normalisation can be used to scale the image to common range for each wavenumber so that the relative absorption patterns can be readily compared. This accounts for the marked differences in emmissivity from the QCL and fibre transmission at different wavenumbers in the spectrum. The images were min-max normalised so that the minimum value was subtracted from each pixel before being divided by the maximum value in the transformed array. This has the effect of scaling the image between 0 and 1. Streak removal and a Gaussian filter (FWHM = 2 pixels) were applied to the images to reduce spatial noise and remove image artefacts. The Gaussian filter replaces each pixel value with a convolution of the central pixel and a two-dimensional Gaussian distribution of defined width to the neighbourhood of pixels surrounding it.

Small lateral drifts in the piezo stage were observed between scans that should be accounted for before direct image comparison. A set of reference images that should be the same for each scan should be used to crop the set to a common area. The actual SNOM images from each wavenumber will have different features, so these mustn't be used as a co-registration reference. However, the topographic images recorded for each scan should have a very similar features, suiting them ideally as a guide for co-registration. The two dimensional cross-correlation is a mathematical operation which effectively translates one image (matrix) relative to a reference, then calculates the total product of both images (I_1, I_2) at various shifts (u, v) relative to each other. The cross-correlation (γ) between two matrices is defined in Eq. (4.5):

$$\gamma(u,v) = \sum_{x=0}^{n_x-1} \sum_{y=0}^{n_y-1} I_1(x,y) \cdot I_2(x-u, y-v). \tag{4.5}$$

The point of registry is indicated by a maximum in the resultant cross correlation matrix (γ). The relative shift of every image in the set was determined by calculating the cross correlation of its topographic scan with the topography of the first image in the set as a template. The shifted images were subsequently cropped to a common area for direct comparison.

4.4.4 Results

The LM images, topography and processed SNOM images are shown in Fig. 4.28. The ratio of images of 1252 cm^{-1}and 1285 cm^{-1} is also shown in (i) to investigate the contrast at the optimum metric $\delta_{1252,1285}$ determined in phase II. Unfortunately, the SNOM imaging of the core presented in Fig. 4.26a and b was destroyed during SNOM imaging, as the feedback was sub-optimally configured which led to the probe touching the sample, scraping away the delicate thin layer of tissue. Nevertheless, the ROI in the second core (Fig. 4.26c and d) contains well defined metastasis and lymphoid features, so the SNOM images obtained on this area will be the focus of discussion.

The images from the SNOM depict a distinct region in the bottom right corner of the ROI, indicating correspondence with the tumour mass found in the same region of the LM images (a,b). Another striking feature is the heterogeneity of the SNOM images, which supports the notion that spatial resolution is enhanced with this imaging modality, potentially offering a tool to invesitigate additional information not previously accessible with conventional FTIR-MS techniques.

To reveal the information contained within the images in more focussed detail, line profiles taken through the tumour in the SNOM images were extracted, and are plotted in Fig. 4.29. The profile from each image was 1 pixel wide, and traverses the width of the metastasis in the bottom right corner of the images shown in Fig. 4.28. Despite

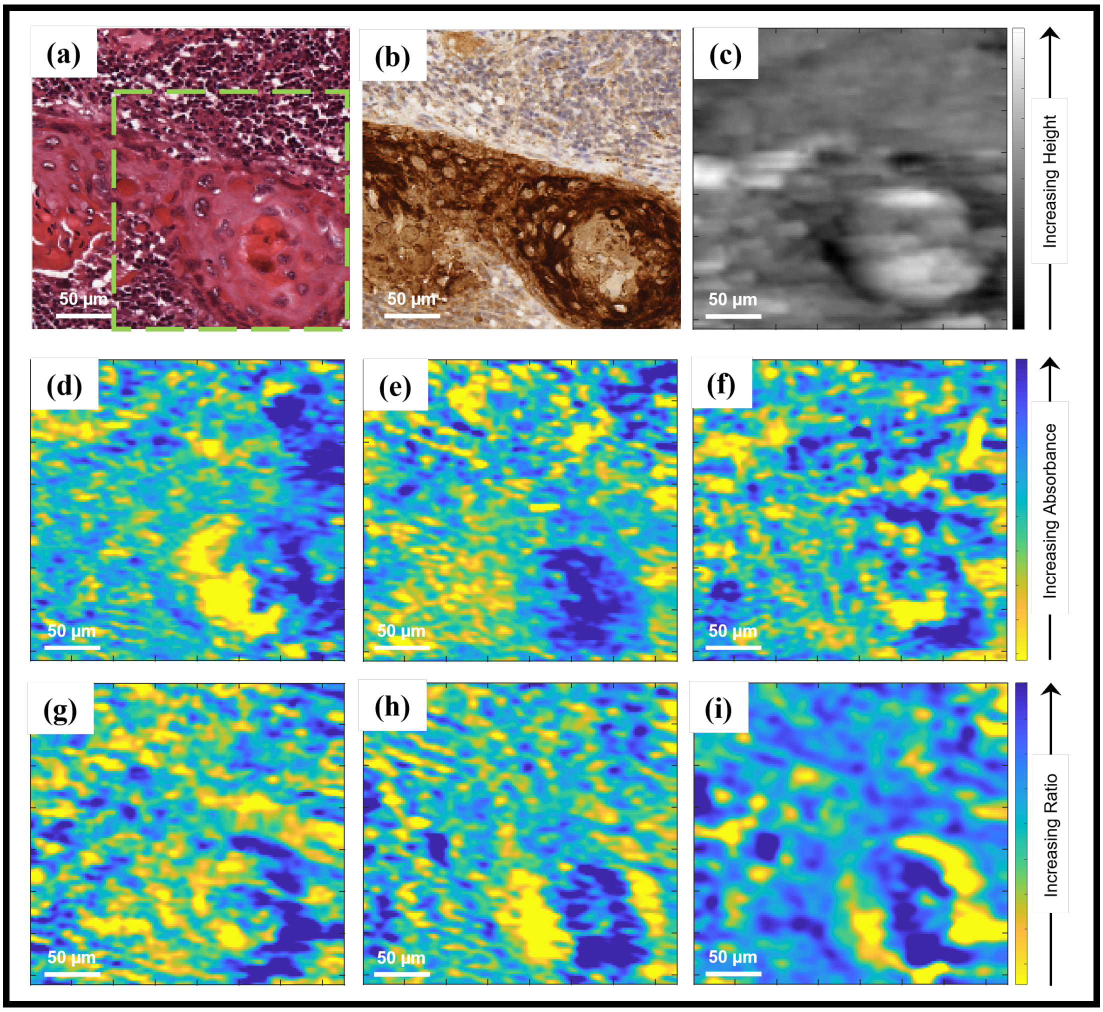


FIGURE 4.28: (a) H&E stained image, (b) IHC image stained for pan-cytokeratins (dark brown), (c) topography, IR SNOM images at (d) 1751 cm^1, (e) 1650 cm^1, (f) 1369 cm^1, (g) 1285 cm^1, (h) 1252 cm^1 and (i) ratio of 1252 cm^1/1285 cm^1 [i.e. (h)/(g)]. All images are 300 μm × 300 μm. Each SNOM IR image is plotted with a colour table covering the 5th to 95th percentiles of the image intensity range. Image (a) was obtained from a section adjacent to that used to obtain image (b), which was in turn adjacent to that used to obtain images (c) to (i).

the success of the hysteresis correction demonstrated in Fig. 4.25, it was difficult to obtain an exact co-registration for forwards and backwards directed scans for the same wavenumber with smaller sized scans. For this reason it was not suitable to calculate noise levels on the basis of differences between the two directions. Rather, noise was calculated by finding the difference between the smoothed images and raw images, then calculating the root-mean-squared (RMS) of the deviations. Comparing this value with the mean intensity of the raw image approximates the noise as $< 5\%$ for all wavenumbers.

Since each pixel in a raw SNOM image is directly related to the detected signal, they are essentially transmission maps. In order to present data that can be easily compared with FTIR images, which are a series of absorption maps for each wavenumber, the line profiles have been inverted so that peaks correspond to more absorption. The profiles are presented on vertical scales that have been corrected for image acquisition parameters such as detector sensitivity. Since the fibre and QCL characteristics vary between images, direct comparison between profiles does not indicate relative molecualar concentration differences. Instead, the spatial arrangement of each wavenumber should be discussed on an intra-image basis.

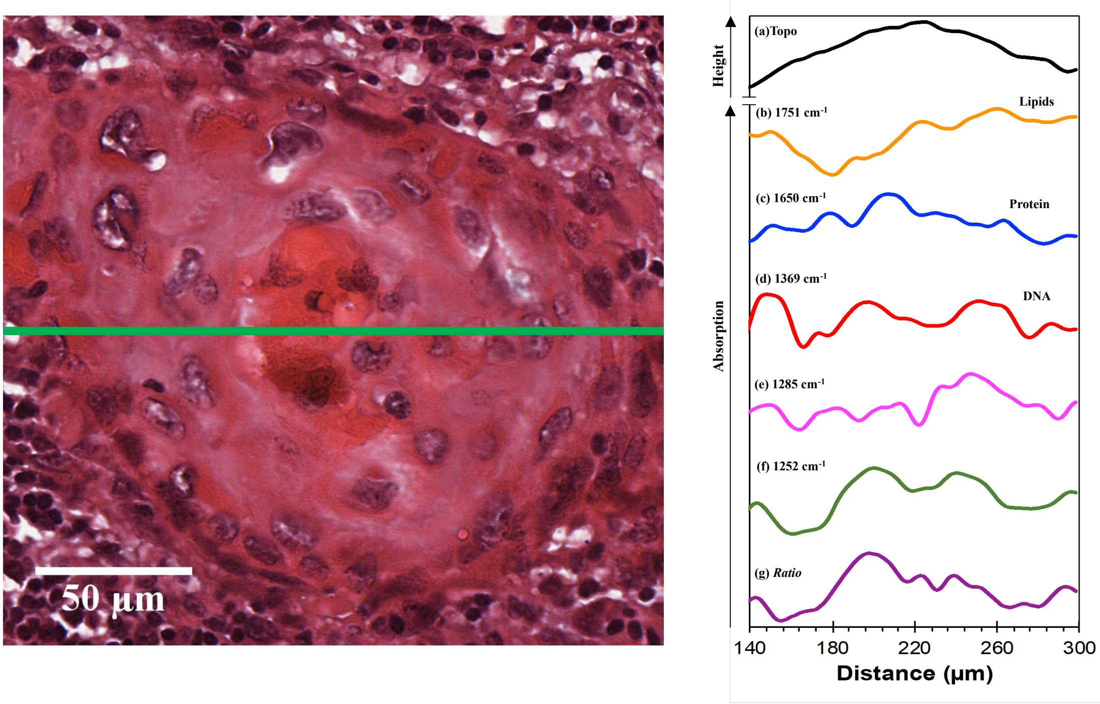


FIGURE 4.29: H&E stained image of region enclosed by green square in Fig. 4.28a (left) and line profiles (right) taken through the core at the white line showing (a) topography, (b) 1751 cm^{-1}, (c) 1650 cm^{-1}, (d) 1369 cm^{-1}, (e) 1285 cm^{-1}, (f) 1252 cm^{-1} and (g) ratio of 1252 cm^{-1}/1285 cm^{-1} [i.e. (f)/(e)]. H&E image (left) was obtained from a section adjacent to that used to obtain SNOM line profiles. Each line profile has been normalised to its min/max values.

The topographic image depicted in Fig. 4.28c indicates an increase in height in the centre of the tumour, where it is noticeable higher than the surrounding tissue. The increase in height appears to correspond with an increase in absorption in the 1650 cm^{-1} image (Fig. 4.28e). Comparison of SNOM line profiles to the topography profile in Fig. 4.29 reveals more marked differences in signal across shorter distances, implying the presence of subtle chemical differences in the metastasis that are not correlated with the topography.

4.4.5 Discussion

The presented results have shown the capability of MA to discriminate between FTIR spectra of metastatic cancer origin and that of lymphoid tissue origin using the metric $\delta_{1252,1285}$. The relatively weak correspondence between the LM images in Fig. 4.26a,b and the absorbance maps in Fig. 4.27a,b compared with Fig. 4.27c supports the notion that the metric provides stronger discrimination, which is clearly demonstrated in the distributions of absorbances and ratio in Fig. 4.16. These results imply that, with sufficient spatial resolution and control of the SNR, delineation of the two tissues can be achieved simply by directly inspecting the spectra in the region between 1250 cm^{-1} and 1289 cm^{-1}. Despite this interesting finding, which does carry academic merit, the robustness must be further tested in larger scale studies incorporating a multi-patient cohort. Unfortunately, at the time of study there was no access to more than a single patient harbouring nodal metastases, and the author of this thesis acknowledges the imposed limitations on the study.

It is very unlikely at present that the described methodology would replace standard histopathological protocol, as clinical adoption would require large scale clinical trials to test the efficacy of such a methodology. Further investigation as to whether the methodology assists in resolving histopathological diagnosic dilemmas such as the identification of isolated tumour cells or atypical micrometastases with conventional techniques like routine histology and immunohistochemistry. It is also important to iterate that the discrimination presented here may not necessarily generalise well to other cancers.

The lower absorption levels of 1252 cm^{-1}, which is highly absorbed by nucleic acids, in the tumour metastasis relative to surrounding lymphoid tissue was somewhat surprising, since increases in DNA ploidy is a common phenotype of OSCC and other solid tumours [128]. The increase in absorption in the surrounding tissue may be explained by the higher concentration of nuclei, as the quantity is directly proportional to the absorption according to Eq. (3.10). Despite the strong correspondence between the large structures in the LM images (Fig. 4.26a,b) and the ratio image (Fig. 4.27c), there is little to no contrast within the tumour itself, which is in itself non-homogeneous, with the metric inable to distinguish between components such as the highly keratinised core and periphery highly differentiated squamous cells.

The superior intrinsic spatial resolution of the SNOM images at important wavenumbers may provide some additional insight into the chemistry of individual tissues. The wavenumbers selected for study all have common attributions to important biomolecules, shown in table 4.7. 1252 cm^{-1} is strongly associated with the PO_2^- nucleic acid signal [129], whilst 1285 cm^{-1} is a characteristic spectral biomarker of collagen [124]. They have also been the subject of previous SNOM studies [125], [130], [131]. Variations in absorption depicted in Fig. 4.28 are on a much finer length scale than the FTIR images of a similar region, with a feature size of $\approx 4\mu m$. The ratio image shown in Fig. 4.28i indicates higher contrast between different areas of the tissue compared with that of the individual wavenumbers (Fig. 4.28g,h). Of particular note is that the centre of the tumour is bound between two broad arcs of tissue, highlighting the capability of SNOM to differentiate between the core and periphery of the metastasis.

The line profiles (Fig. 4.29) provide further insight into the chemical variations in the image. The topographic signal indicates that the keratinised core is higher than the periphery by approximately 1 μm, determined by considering the range of motion of the vertical piezo. Precise quantification of the height is difficult due to the non-linearity of the vertical piezo and dependance on tip geometry. The increase in height coincides with an increase in protein signal in the keratinised core (Fig. 4.28c), which is also reflected in the amide I line profile, which is attributed to the secondary structure (specifically α-helical) of cytokeratins [132], [133]. Alterations to the spatial distribution of cytokeratins and molecules associated with their production are expected between the often highly keratinised core and the less keratinised tissue surrounding it, which corresponds to the advancing front of the tumour, supported by the decrease in 1650 cm^{-1} absorbance on the edges of the inner core.

The smooth topographic and protein signals are in stark contrast to the more marked variations observed in the line profiles at other wavenumbers. This indicates that there may be subtler differences in the abundance of chemical species absorbing these wavenumbers. There are similarities between the lines for 1252 cm^{-1} and 1369 cm^{-1}, which is not surprising considering that they are both strongly associated with nucleic acids. The 1285 cm^{-1} is attributable to collagen, which shows less pronounced variation in the line through the core. This relative lack of contrast is supported by the 1285 cm^{-1} cross-section of the hyperspectral image taken of the first core (Fig. 4.27b), which shows much less contrast than that of the 1252 cm^{-1}. This finding, as well as the scores derived from the distributions in Fig. 4.16 implies that the discrimination between metastatic and lymphoid tissue in this case is dominated by nucleic acids.

Designating the peak of the topography as the reference point for the centre of the tumour, the 1252 cm^{-1} line shows a reduction in intensity in the centre, surrounded by two peaks $\approx$ 25 µm either side, which is consistent with heightened levels of keratin in the core of the tumour. Further reductions are observed $\approx$ 50 µm either side of the centre, approximately 25 µm in width, which corresponds to roughly 2-3 layers of cancer cells. This particular finding seems counter-intuitive as one would expect keratinised, highly differentiated tissue to contain fewer nuclei relative to the periphery [134]. However, attribution of 1252 cm^{-1} to phosphate groups means all nucleic acids [135], [136] and phospholipids [137] will show enhanced absorption at that wavenumber. This leads to the hypothesis that the increase in absorption may reflect a change in RNA signature or increase in endoplasmic reticulum corresponding with increased proteinosynthetic events in the advancing front.

The line profile from the 1285 cm^{-1} image represents quite a complex signal across the whole section, particularly immediately to the right of the keratinised core. Collagen attributes such as density, alignment and straightness vary between cancer types, and specific tumour sub-sites [138], and effect key processes such as invasion, metastasis and apoptosis. The concentration of collagen is also influenced by the tumour microenvironment, and affects the immune response [139]. The differences observed in the 1285 cm^{-1} line profile may therefore arise from subtle changes in collagen fibre structure, which could be further investigated in future studies.

4.5 Conclusion

This chapter has demonstrated the utility of IR imaging modalities and an adapted MA technique to the study of oral cancer histopathological specimens. Phase I indicated there is potential for using trained MA models to label hyperspectral images, however the author acknowledges that expansion of the patient cohort is essential before any conclusive remarks can be validly claimed. Future studies should focus on reinforcing the model with data from more patients that have been specifically selected for factors such as gender, age and lifestyle risk factors. Due to the paucity of histopathological specimens that are available for this style of study, multi-centre studies would probably be necessary to obtain the data in a reasonable time-frame. This would require more rigorous pre-processing and calibration to mitigate for the experimental variability introduced by different individuals performing experiments using different instruments [140].

Phase II III detailed the employment of MA in discriminating between OSCC nodal metastasis and surrounding lymphoid tissue. This was achieved using a single metric ($\delta_{1252,1285}$), which was determined without any prior assumptions about the important features within the data, emphasising the methods capabilities as a feature extraction technique in addition to supervised classification. Near perfect classification was achieved, with a sensitivity and specificity of 98.7% and 99.9% respectively. The author acknowledges that the very small sample size and homogeneous data are the main contributing factors to such high scores, but the task of discriminating between nodal metastases and lymphoid tissue is not a clinical problem. In fact, there is clear delineation between the LM images as shown in Fig. 4.26. Instead, phase II should be regarded as a proof of concept study for utilising MA as an interpretable tool that can guide further studies.

Phase III implements the results from phase II in more focussed discrete frequency IR-SNOM experiments. The results of these gave additional insight into the chemistry of OSCC metastasis, whereby the keratinised core, advancing front and surrounding lymph tissue were all discernible in the ratio image corresponding to $\delta_{1252,1285}$. The findings were in agreement with current biological understanding of the biochemical mechanisms at play, but additional questions such as the role of collagen should be explored in further studies. These questions will be addressed in future work but are not the subject of the remainder of this thesis. In the subsequent chapters, focus will shift from using FTIR-MS as an oral tissue labelling tool, to its pairing with a novel framework as a prognostic predictor for malignant transformation.

Chapter 5

Pipeline Optimisation Framework

5.1 Introduction

This chapter is based around a framework which objectively optimises the selection of pre-processing and classification methods and hyperparameters (HPs) for a given task. The program, which is a collaborative effort between myself and fellow PhD student Conor Whitley, was written in Python with the vision of making it open source in the future. A paper based on the framework is currently in review.

5.1.1 Background and Motives

Despite the undeniable promise of applying vibrational spectroscopic techniques and machine learning to the analysis of biomedical datasets, it is hampered by the lack of consensus regarding the choice of pre-processing and machine learning techniques. Pre-processing is a vital step in the analysis workflow, as it has been shown to generally increase performance of classification models [8], as well as to increase the validity and interpretability of results. As detailed in section 3.3, the process of applying multiple steps in order to transform a raw labelled dataset (X, y) into a set of predictions $\hat{y}$ can be considered a supervised 'pipeline'.

Interestingly, despite the potentially significant influence the choice of pre-processing methods imposes on results [141], robust selection of an optimised pipeline is not routinely implemented [142]. This is surprising since the optimal method has been shown to depend on the characteristics of the specific dataset, as well as the purpose of the analysis (binary classification, regression, multi-class classification, calibration).

A study by Engel *et al* [142] revealed that there was a 20% difference between the worst and best combination of pre-processing methods (out of ≈ 5000 theoretically reasonable combinations) applied to an FTIR dataset. This highlights the necessity in realising the optimal protocol if high scores are to be achieved, but it also exposes how an arbitrary approach based on intuition are flawed.

Jarvis and Goodacre [143] detailed an approach whereby a genetic algorithm (GA) was used to optimise a pre-processing pipeline. In this method, generations of pre-processing sequences were allowed to 'evolve' in a manner analogous to Darwinian evolution. The 'fittest' pipelines were allowed to cross-over with each other to produce 'off-spring' which can mutate and cross-over to produce more generations. The optimised pipeline resulted in 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 in run time by means of parallelisation is not possible.

In a more recent study by Butler *et al* [144], a trial and error approach was used on an attenuated total reflectance (ATR) FTIR biofluid dataset from brain cancer patients. Each permutation was tested using either a random forest (RF) classifier or a support vector machine with features selected by RF or GA. They concluded that there is a small fraction of permutations which are highly favourable compared to the rest of the dataset, which further highlights the need for optimisation.

Each stage within the pipeline may have several available methods to choose from. Normalisation, for instance, may refer to vector, min-max or feature normalisation, all of which are well suited to some tasks more than others. The choice of classifier is another aspect that should be optimised, as some classifiers perform better depending on the nature of the data. There may also exist a set of HPs associated with the method, which must also be carefully selected based on the chosen data and model. This multi-step, multi-method problem makes the task of optimising the pipeline extremely difficult to achieve in a trial-and-error approach. The total number of permutations (N) can be determined using Eq. (5.1).

$$N = \prod_{s=0}^{n_s-1} \left(\sum_{m=0}^{n_{m(s)}-1} n_{\theta(m)} \right) \tag{5.1}$$

where n_s is the number of steps in the pipeline, $n_{m(s)}$ is the number of methods associated with step s, and $n_{\theta(m)}$ is the size of the HP search space for method m. Consider a simple case where there are 5 steps, each with 3 methods, with each method having 2 HPs taking 3 possible values. Firstly, the size of the HP search space will be $3^2 = 9$ for each method, summing to 45 for each step. Finding the product of this grid across the five steps gives $45^5 \sim 2 \times 10^8$ total permutations. Furthermore, the evaluation of one single pipeline can be quite computationally expensive in itself, depending on the size of the data and complexity of each applied method. Sequentially processing this many pipelines would take an enormous amount of time, which is a major limitation if vibrational spectroscopy and ML are to be considered a feasible clinical adjunct.

As discussed in section 3.3, validating the optimal HPs should be implemented using a routine such as k-fold cross validation, so that HPs are not overfitted to a subset of data. Because an independent model is being trained for each cross validation iteration, this effectively scales up the runtime of the already computationally heavy optimisation by a factor of k. A framework that searches for the best pipeline with the best set of HPs using cross validation would execute many independent processes before combining and comparing the results to obtain the optimal pipeline. This makes the framework an ideal candidate for parallel computing, which processes independent tasks simultaneously, so the runtime can be dramatically reduced.

The work presented in this chapter proposes a novel method for objectively optimising an analytical pipeline for vibrational spectroscopy. It combines a trial-and-error approach with parallel computing and Bayesian optimisation (BO), a concept which will be discussed in the following section. The desired outcome is a robust framework that can efficiently search across a large optimisation space in order to maximise the performance of the pipeline. Despite being designed and developed with spectroscopic data in mind, it can be adoptable by other fields which have a large parameter space associated with multi-step pipelines.

5.1.2 Bayesian Hyperparameter Search

HP searches seek to determine the optimal HP vector $\boldsymbol{\theta}^*$ so that an objective function $f(\boldsymbol{\theta})$ is either minimised or maximised. For the optimisation of classification pipelines, a performance measure should be maximised, this is summarised in Eq. (5.2):

$$\boldsymbol{\theta}^* = \underset{\theta}{\operatorname{argmax}} \left[f(\boldsymbol{\theta}) \right] \tag{5.2}$$

Section 3.3.2 outlines some of the ways in which the optimal set of HPs can be selected in a machine learning algorithm. A set of HPs in a pipeline $\boldsymbol{\theta}_p$ can be though of in a similar way, whereby each method contributes its own set of HPs $\boldsymbol{\theta}_m$ to $\boldsymbol{\theta}_p$. For instance, consider a pipeline consisting of PCA denoising and vector normalisation as pre-processing steps, followed by a logistic regression classifier. The extent to which the data is denoised is determined by the number of retained PCs, so this is passed into θ_p, along with the regularisation strength in logistic regression which reduces overfitting to training data in model construction. The optimal set of HPs $\boldsymbol{\theta}_p^*$ exists at a point within the 2-dimensional space (each dimension represents a HP), and can be determined or estimated with the aid of search algorithms. The familiar grid search method searches every combination of discrete values associated with each HP. Whilst this is thorough, it is not ideal as it may transform what could be a continuous domain (regularisation strength) into a discrete, user defined set of possible values. Random searches differ from this in that they randomly sample a defined number of points from the original space in which the possible HP combinations exist.

Figure 5.1 is a basic comparison between HP optimisation using grid search or random search. The different colour stars represent the optimal solution in two scenarios as an example. The grid search is superior at converging on the first solution (yellow) than the random search, but inferior to converging on the second solution (green). This demonstrates the weaknesses associated with either technique. Grid searches across continuous domains carry the risk that too few points are specified to find the true optimum solution, whereas random searches require that enough points are sampled in order to be effective and to minimise the risk of missing the optimal solution. This implies (unsurprisingly) that more points are required to increase the likelihood that

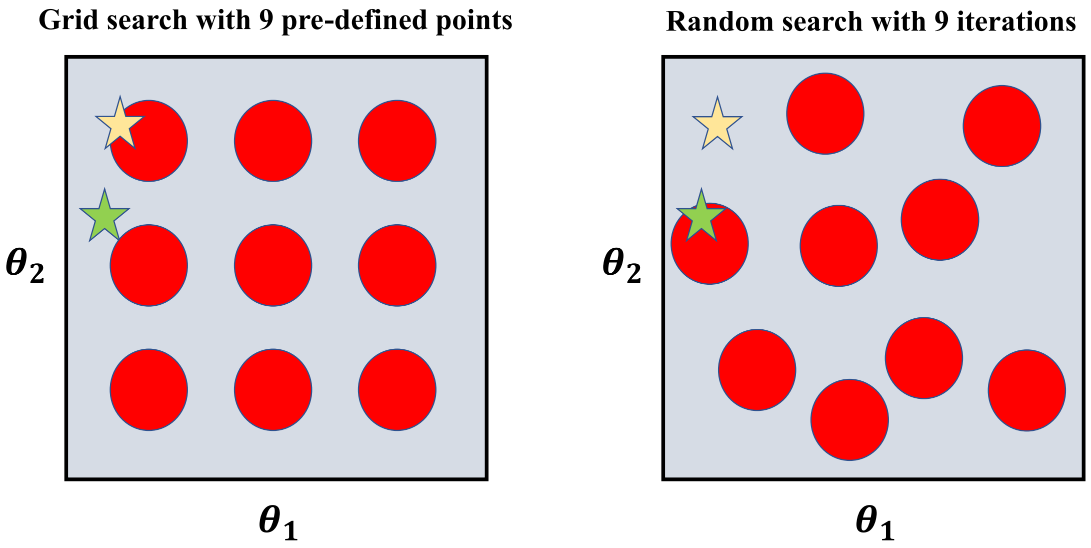

Grid search with 9 pre-defined points
Random search with 9 iterations
θ_2
θ_1
θ_2
θ_1

FIGURE 5.1: Comparison between a grid search and random search for HP optimisation. The two methods are used on the same problem, for which two different optimal combinations are depicted by the green and yellow stars. The grid search is configured to search over 9 user-defined positions in the space, whereas the random search is configured to randomly sample combinations from the space.

the best combination of HPs is converged upon. This dependence scales poorly as the number of dimensions in $\boldsymbol{\theta}_p$ increases, which suits the methods poorly to pipelines which typically have a large number of HPs.

Another limitation of the aforementioned approaches is that they are completely unguided by information from previous evaluations. This can lead to repeatedly trialling bad HPs, which decreases the efficiency of the process. Consider for example, the 2D function $f(\theta_1, \theta_2)$ shown in Fig. 5.2. Since the objective of a HP search is to find a global maximum of a function such as classification accuracy, the positions where $f(\theta_1, \theta_2)$ are low are of no interest. It would be very inefficient if a grid search was implemented to find this function's optimum value since there is only a small region within the 2D matrix that corresponds to high values. If random search was utilised, the iteration number would need to be high in order to increase the likelihood of locating the maximum.

A much more efficient way to locate the maximum of this function would be to use the information acquired over the course of the search to make informed decisions as to where to select the next set of HPs. This is the essence of Bayesian optimisation (BO), a

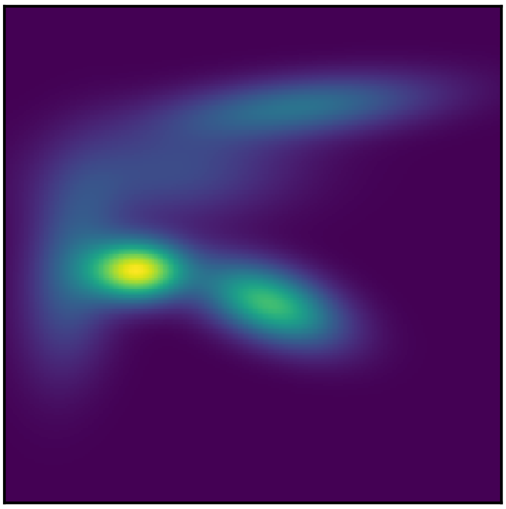


FIGURE 5.2: 2D function $f(\theta_1, \theta_2)$. Implementing random or grid search to find the maximum to this function would be inefficient.

term generally attributed to work by Jonas Mockus in the 1970s and 1980s [145], [146]. Bayesian optimisation carries utility in the following conditions:

(i) Only the global optimum is of interest. Knowledge of local optima is not required. True for HP optimisation in ML.

(ii) The function f is expensive to calculate. True for cross validation evaluations of pipelines.

(iii) No knowledge about the behaviour of f, i.e. it is a 'black-box'.

(iv) The number of dimensions is typically ≤ 20 [147]. True for analytical pipelines.

The mechanism by which BO builds up a knowledge base of the unknown function f whilst using that knowledge to guide future decisions is by fitting a probability surrogate model to f. The surrogate model is usually a Gaussian process (GP), which is a probabilistic model mapping inputs (in this case the HP vector $\boldsymbol{\theta}$) to the objective function f. A GP can be regarded as a distribution of functions, which has a mean function $m(\boldsymbol{\theta})$ and a covariance function $k(\boldsymbol{\theta})$, where each point in HP space has an associated mean μ and standard deviation σ. The mean function is the approximation of the relationship between the choice of HPs and the objective function, whereas the covariance function encapsulates the relationship between the points in the space. The Matern kernel is frequently used as a covariance function due to its ability to cope with a noisy objective function [147], which is likely to be the case when evaluating pipelines using average cross validation statistics. The equation for the Matern kernel is:

$$k\left(\boldsymbol{\theta}, \boldsymbol{\theta'}\right) = \sigma^2 \left(1 + \frac{\sqrt{3}r}{l}\right) \cdot \exp\left(-\frac{\sqrt{3}r}{l}\right) \tag{5.3}$$

where $r = |\boldsymbol{\theta'} - \boldsymbol{\theta}|$ is the euclidean distance between two points, σ^2 is the variance, and l is the length scale parameter. All combinations of points constructs the covariance matrix which governs the relationship between points in the model.

The Gaussian process is used in the determination of the next best point to evaluate in the optimisation. Given an initial set of n evaluations at $f(\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_n})$, the GP posterior can be updated to incorporate the new evaluations. The updated GP, which has a μ and σ for every $\boldsymbol{\theta}$, can be used to inform some acquisition function which determines which point next to evaluate. This acquisition function may take various forms, but the most widely implemented is perhaps the expected improvement algorithm (EI), which determines the next best point by considering both the mean and standard deviation functions of the posterior through Eq. (5.4).

$$EI(\boldsymbol{\theta}) = \begin{cases} \left(\mu\left(\boldsymbol{\theta}\right) - f\left(\boldsymbol{\theta}^*\right)\right)\Phi(Z) + \sigma\left(\boldsymbol{\theta}\right)\phi(Z) & \text{if } \sigma\left(\boldsymbol{\theta}\right) > 0 \\ 0 & \text{if } \sigma\left(\boldsymbol{\theta}\right) = 0 \end{cases} \tag{5.4}$$

Where $\mu\left(\boldsymbol{\theta}\right)$ and $\sigma\left(\boldsymbol{\theta}\right)$ are the mean and standard deviation of the Gaussian at $\boldsymbol{\theta}$, $f\left(\boldsymbol{\theta}^*\right)$ is the current maximum evaluation, and $\Phi(Z)$ and $\phi(Z)$ are the standard cumulative density function (CDF) and probability density function (PDF) at Z respectively, where:

$$Z = \frac{\mu\left(\boldsymbol{\theta}\right) - f\left(\boldsymbol{\theta}^*\right)}{\sigma\left(\boldsymbol{\theta}\right)} \tag{5.5}$$

The set of HPs with the maximum EI is selected and the function is evaluated at the new value. The GP model is subsequently updated with the new value. It is worth mentioning that the standard deviation at points where the function has been evaluated goes to zero, as there is no longer any uncertainty in the value of the objective at that point. The expression for EI in Eq. (5.4) gives two conditions where the EI will be high. First consider that the mean function at a point is high, i.e. the current GP model predicts the value of the objective function will be high. Sampling more points in this space will be beneficial, since the global optimum may exist in this region. This gives weight to the first term in Eq. (5.4), which is high if the mean function is larger than the best evaluation so far. If this were the sole factor in the acquisition function, there is a high probability of converging on a local optimum, since there is no incentive to evaluate any points that are distant from a small region of HP space. This 'exploitation' factor should instead be balanced with an 'exploration factor', which is what the second term in Eq. (5.4) addresses. Regions of space where relatively few evaluations have been made will have a relatively large standard deviation function, since the standard deviation goes to zero at points which have been evaluated, and will be similar at neighbouring points due to the covariance of the GP.

To demonstrate its effectiveness and reinforce intuition, BO can be used to attempt to find the optimum of the 2D function shown in Fig. 5.2. Each horizontal panel of Fig. 5.3 contains the surface plot representations of the true function (a), the mean function as predicted by the GP (b), the standard deviation function as predicted by the GP (c), and the EI acquisition function (d).

To intialise the optimiser, a GP is fit to 10 random points which have been evaluated using the function. The mean and standard deviation are updated accordingly. The first panel indicates that the model is already beginning to represent the overall trend in the data. The GP posterior is then used to calculate the expected improvement using Eq. (5.4), which then determines the next sampled point, shown using a red point in figure (d) of each panel. The true optimum of the data is shown with a red point in figure (a) and (b) of each panel. The regions of low standard deviation (dark blue) in figure (c) of each panel indicate points that have been evaluated, as the standard deviation reduces to zero in these cases. It can be seen in (c) and (d) of each panel that

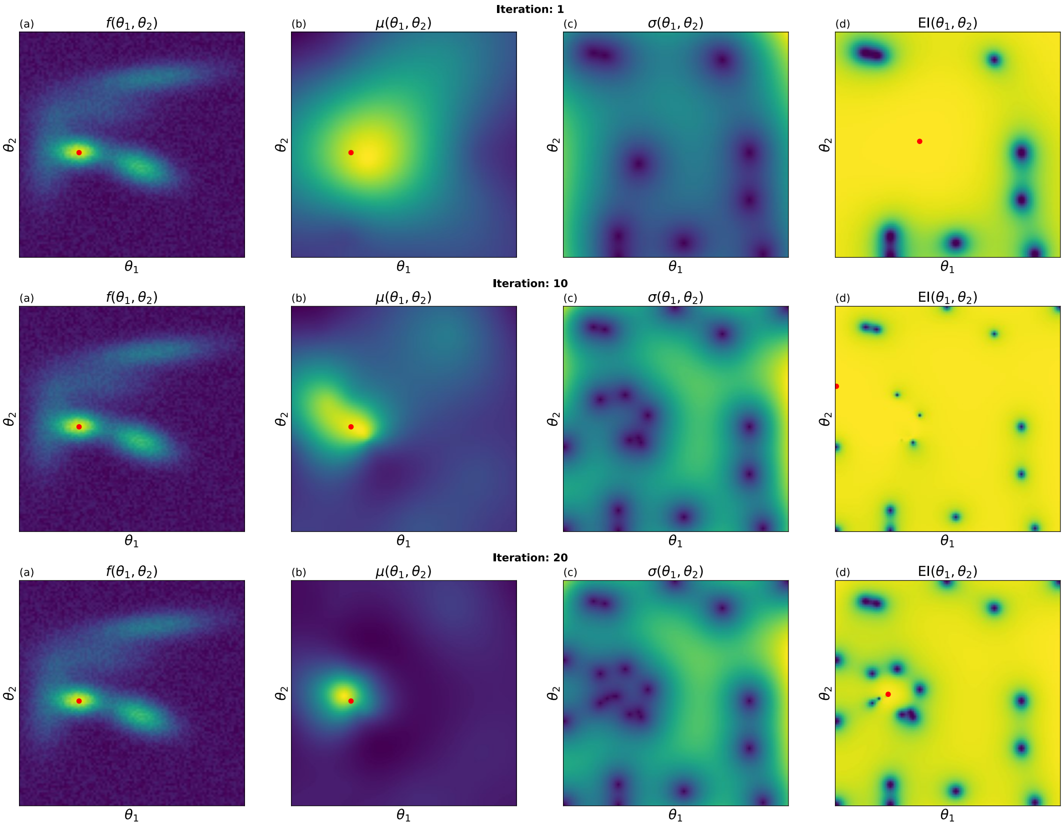


FIGURE 5.3: Finding the optimum of a 2D function using BO over 20 iterations. 1st, 10th and 20th iteration shown for brevity. Each panel shows the true function (a), mean function (b), standard deviation (c) and expected improvement (d).

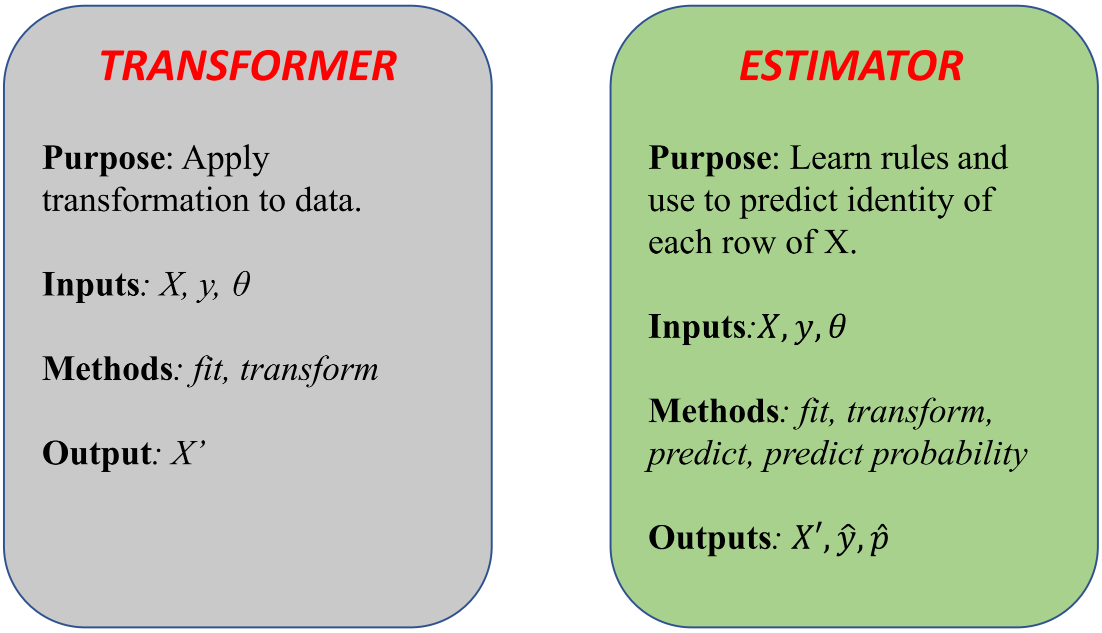
the model is sequentially guided to sample points which are near the maximum of the function. After 20 iterations, the global maximum has been accurately determined, demonstrating the efficiency and efficacy of utilising such an approach. If a grid search was used over the same domain ($0 \leq x \leq 100$, step 1), to get the same precision would require $100^2 = 10^4$ evaluations, indicating an efficiency markup of 500. Of course, it's not quite so simple to make such a direct comparison, as there are advantages and drawbacks to both approaches. A lot of HPs for pipelines are categorical/integer, for instance the window size in SG smoothing is constrained to be an odd number, and the number of trees in a random forest is constrained to be an integer. Gaussian processes can work with discrete domains such as these [148], however they are most definitely designed and optimised for continuous domains [147].

5.2 Framework Components

There were three main goals in mind when designing and developing the framework, which will be referred to herein as pipeline optimiser or simply '`PipeOpt`'. The first is to incorporate a hierarchical search across different methods and respective HPs in order to thoroughly optimise the protocol. The second is to utilise high throughput tools such as parallel processing in order to increase the efficiency of optimising over a large space. The other is to make `PipeOpt` as modular as possible, in order to maximise the versatility of said framework, so that methods and HPs can be easily added by either myself or other users. `PipeOpt` uses the object oriented (OO) programming paradigm to increase the modularity of the framework and to enable the monitoring of intrinsic processes and variables within.

5.2.1 Steps and Methods

To begin with, modules for each step within the framework were created. Each module contains a number of classes which apply a particular method to the data. The classes 'inherit' from the `scikit-learn` application programming interface (API) [149], a very popular python library containing tools for machine learning oriented tasks. Pre-processing steps are defined as `Transformer` classes, with the ultimate classification step defined as an `Estimator` class. Transformers take the data matrix $\mathbf{X}$ and apply some transformation (such as PCA denoising) to output a transformed data matrix $\mathbf{X}'$. Sometimes the label vector $\mathbf{y}$ is used in the transformation, but for most protocols (such as pre-processing of FTIR data) the transformers are unsupervised and $\mathbf{y}$ is passed through as a redundant variable. On the other hand, estimator classes take $\mathbf{X}$ and the labels $\mathbf{y}$ as an input, and learn rules that either map $\mathbf{X}$ to $\mathbf{y}$ (for supervised learning) or recognise patterns that clusters the data into distinct groups without knowledge of $\mathbf{y}$ (unsupervised learning). The estimator is then able to predict the labels ($\hat{\mathbf{y}}$) and associated probability ($\hat{\mathbf{p}}$) of each row of X. An estimator class may also involve a transformation, such as LDA and logistic regression, which both rely on a linear transformation of variables in order to obtain a prediction. A summary of both base classes is shown in Fig. 5.4. In addition to the data inputs ($\mathbf{X}$, $\mathbf{y}$), there may be a set of HPs $\boldsymbol{\theta}$ associated with the method, which can either be declared as default variables or require that they are explicitly passed with the function call.



TRANSFORMER

Purpose: Apply transformation to data.

Inputs: X, y, θ

Methods: *fit*, *transform*

Output: X'

ESTIMATOR

Purpose: Learn rules and use to predict identity of each row of X.

Inputs:X, y, θ

Methods: *fit*, *transform*, *predict*, *predict probability*

Outputs: $X', \hat{y}, \hat{p}$

FIGURE 5.4: Summaries of the transformer and estimator `scikit-learn` base classes.

Although the design of `PipeOpt` enables steps and methods to be defined with ease depending on the application, the version demonstrated here contains standard pre-processing and classification methods commonly used in FTIR data analysis. Below is a list of modules and respective classes which were integrated into the framework.

(i) Denoising: PCA, SG-Smoothing

(ii) Baseline: rubber-band

(iii) Normalisation: vector, min-max, amide I

(iv) Scaling: standardisation, robust, min-max

(v) Decomposition: PCA

(vi) Classification: logistic regression, random forest, xgboost

Most of the pre-processing methods listed above have been previously described in section 3.3.1. One set of routines not discussed is scaling, a step which scales each variable (wavenumber) to a common domain. This is a very important step for many classifiers, especially those based on a linear transformation such as logistic regression, LDA and neural networks [9]. The three methods defined above are standardisation, robust and min-max. For standardisation, the mean absorbance of each wavenumber is subtracted from each spectrum, prior to division by its standard deviation, so that the distribution over the data for each wavenumber has zero mean and unit variance. Robust scaling is similar to standardisation, but the median is subtracted rather than the mean, and it is scaled by the interquartile range rather than the standard deviation. This approach is less sensitive to outliers, which skew the mean and standard deviation in a negative way. Min-max scaling simply scales each wavenumber variable to range from 0 to -1, however this approach is very sensitive to outliers which may compress or stretch the distribution based on anomalously high or low absorption.

The methods contained within the final classification step have not yet been introduced. Logistic regression, random forest and XGBoost, the former two of which are available as packages in `scikit-learn`, whilst XGBoost is written independently, but inherits from the estimator base class from the `scikit-learn` API. The three classifiers were chosen due to their differences in complexity, with the number of searchable HPs in excess of 20 for XGBoost. The training of the classifier involves the optimisation of the weights w and bias term b that lead to the highest performance. RF and XGBoost are both based on decision trees, a type of supervised classifier which seek to determine a set of conditional statements which lead to a decision of what class the data belongs to. Figure 5.5 shows a rudimentary example of what a decision tree in the context of FTIR spectroscopy may look like.

The parent node at depth = 0 (top of Fig. 5.5) splits the data based on the condition that it has an aborbance at 1650 cm^{-1} greater than 0.8. The subset that does not fulfil this condition then steps to the left (false direction), where it is classified as cancer. If the condition is met, it steps down to the right, where it meets another conditional statement which queries if the absorbance at 1242 cm^{-1} is less than 0.8. If this is true, the spectrum is classified as dysplasia, if false it's labelled as healthy. Each node in the tree has an associated quantity known as the Gini impurity, which is calculated using Eq. (5.6).

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \tag{5.6}$$

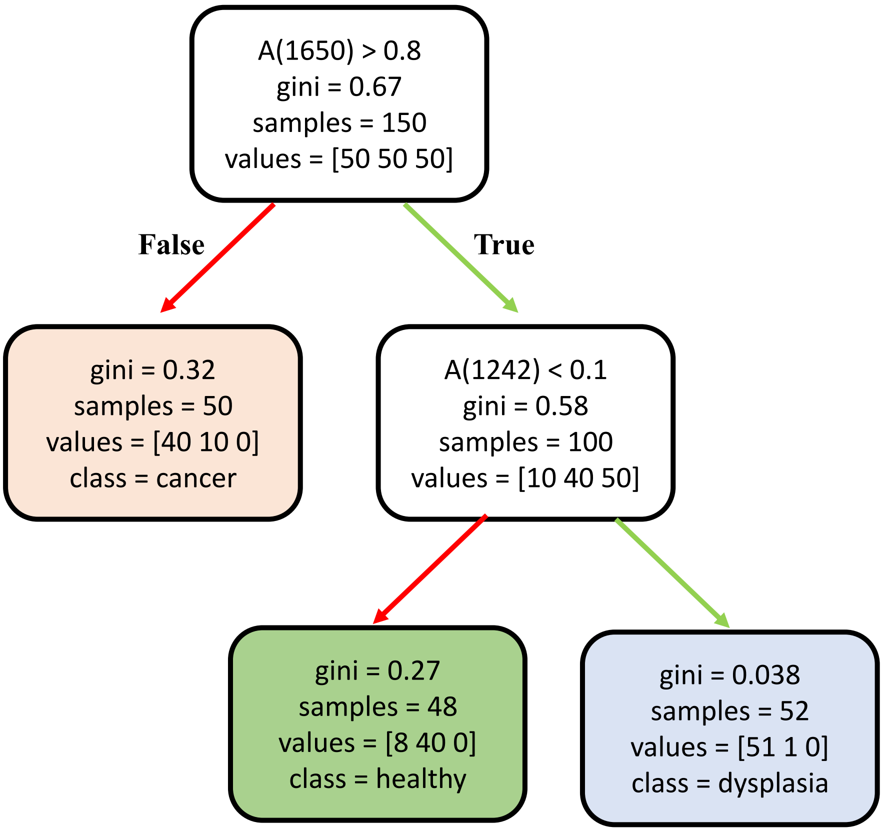


FIGURE 5.5: Example of a simple decision tree with two root nodes and three leaf nodes for a three class classifier.

where $p_{i,k}$ represents the ratio of the number of time the class k occurs in the i^{th} node. For instance, in the parent node at depth = 0, the number of instances for each of the three classes is 50, therefore the Gini impurity would be calculated by $1 - (1/3)^2 - (1/3)^2 - (1/3)^2 = 2/3$. This quantity is used in the training of the decision tree, which seeks to determine the features and thresholds which produce the purest nodes (those that produce nodes with a low Gini impurity). RF is essentially an ensemble of decision trees trained on a random subset taken from the training data, using a random selection of available features. Each of these independent classifiers votes for an overall classification. RF has been used on FTIR-MS biomedical datasets on numerous occasions [117], [150]–[152]. XGBoost [153], which stands for extreme gradient boosting, is another variation of decision tree ensemble, where classifiers are added sequentially to correct the mistakes made by the previous iteration.

5.2.2 Pipeline Construction

The next component of `PipeOpt` to be integrated was the mechanism by which the pipelines are constructed. Each pipeline was to be a unique combination of a defined set of methods, which utilised a Bayesian HP search to determine the optimal set of HPs. The steps and methods were defined in a python data structure known as a `dictionary`, which are associative collections that map values to identifiers known as keys. They are a very effective way of storing data in such a way that values can be accessed with something more intuitive or convenient than a numerical indexer.

In the case of `PipeOpt`, the keys are the steps that will make up each pipeline, and the corresponding values are a list of methods that are to be trialled for each step. Each element of the list (method) also contains another dictionary which specifies each HP and its respective search domain as key-value pairs. This structure can be visualised as a hierarchy, shown in Fig. 5.6. Every permutation of methods (taking one from each step) is then generated and can be passed as an argument to a pipeline object. A pipeline object is a `scikit-learn` estimator (Fig. 5.4), which executes all the transformations associated with the previous steps before training the estimator in the final step. In the example shown in Fig. 5.6, there are two steps each containing two methods, this would yield $2 \times 2 = 4$ pipeline objects. The HP search domains associated with each method in the pipeline are combined, and will be the subject of BO in subsequent stages.

5.2.3 Job Dispatching and Execution

As previously stated, the throughput of trialling many different pipelines can be dramatically increased by utilising parallel processing. The initial release of `PipeOpt` makes use of the popular open source high throughput computing software framework *HTCondor* [154], herein referred to as 'condor'. Condor utilises a pool of idle computers in a local network to process parallelisable, computationally expensive tasks, or 'jobs', and can be controlled with ease from a server. The University of Liverpool condor framework consists of a pool of 1900 computers situated in teaching centres, laboratories and libraries. They are only accessible when they are idle, and as soon as interaction is made with the computers input hardware (keyboard or mouse),

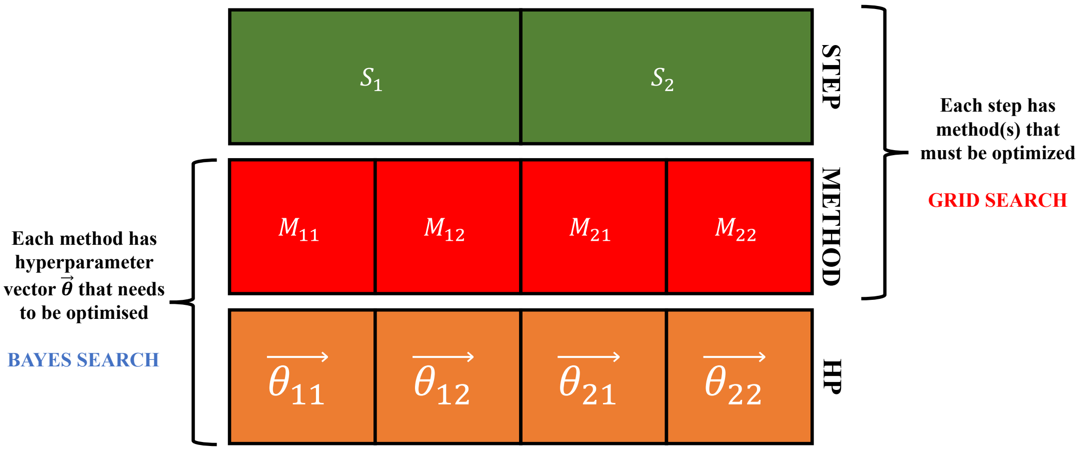

S_1
S_2
STEP
M_{11}
M_{12}
M_{21}
M_{22}
METHOD
$\overrightarrow{\theta_{11}}$
$\overrightarrow{\theta_{12}}$
$\overrightarrow{\theta_{21}}$
$\overrightarrow{\theta_{22}}$
HP
Each step has method(s) that must be optimized
GRID SEARCH
Each method has hyperparameter vector $\vec{\theta}$ that needs to be optimised
BAYES SEARCH

FIGURE 5.6: Structure of the pipeline search dictionary. All combinations of methods (taking one from each step) are generated, combining all HPs from the selected methods into one domain. The combined HP vector is subject to BO.

the computer is removed from the available pool. Each computer is equipped with an Intel core i3 (quad-core) processor runnning at 3.3 GHz, 8GB RAM and 120 GB storage.

The framework is configured to dispatch individual pipeline objects to a job. Accompanying the pipeline object are several mutual inputs that are required for job execution; these consist of the execution script, labelled train and test data, the parameters for BO, a python distribution containing all the required packages, and any other dependencies. The process is illustrated in Fig. 5.7.

Each job now contains all the required inputs to run the optimisation. The execution script is a python file which controls the processes within the job, such as the loading of the pipeline, running of the BO, and saving of results. For the running of BO, the function `BayesSearchCV` from the package `scikit-optimize` was utilised, which enabled the specification of parameters such as the objective function, number of iterations, and number of initial points to sample to intialise the optimiser. Furthermore, it allows for validation routines such as *k*-fold cross validation to be implemented, in order to robustly determine the optimum based on all the accessible training data. The choice of objective function to be optimised should be selected to suit the specific problem. For instance, a classification problem would be suited to performance measures such as the AUC, sensitivity or specificity.

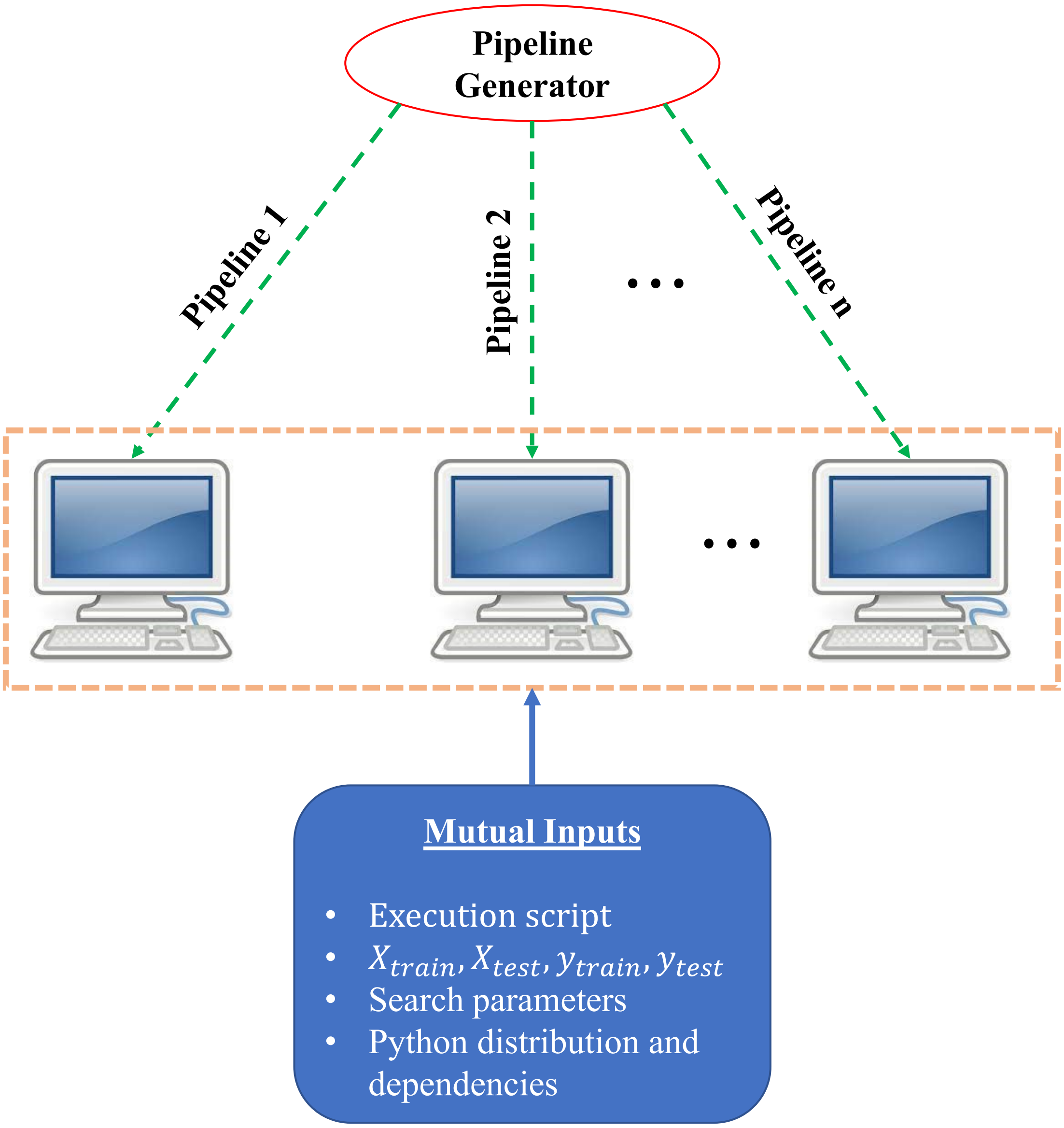


FIGURE 5.7: Process of dispatching n pipelines and a set of mutual inputs to each job in the condor framework.

Following HP optimisation using BO, the HPs with the best mean performance from the search are selected and the pipeline is retrained on all the training data, before testing on the independent hold-out data. The results for each specific job are saved to a file and returned to the directory, wherein the optimised pipelines can be recombined and compared.

5.3 Testing

5.3.1 Methods

For testing the functionality of `PipeOpt`, the framework was deployed in the analysis of a multi-patient real FTIR dataset. The dataset comprises of spectra taken from primary tumour sites of 28 patients. All samples were FFPE tissue biopsies collated into 1mm diameter core TMAs. Tissue annotation was performed by a maxillofacial pathologist using adjacent H&E stained sections to ensure accurate labelling. The dataset had previously been used to investigate the prognostic ability of other biomarkers [155]. The objective was to obtain the optimised pipeline with the best mean performance across a number of train-test splits. The task was to predict the prognosis of a patient as surviving beyond, or less than one year of the most recent review date. It is important to state that the purpose of this exercise is solely to test that `PipeOpt` functions as expected, rather than to derive any biological conclusions, since the data was not collected by myself. Chapter 6 will detail the analysis of a multi-patient dataset of dysplasia samples using `PipeOpt`, and will include discussion of biological and clinical implications.

The total number of pipelines generated can be easily calculated as the cumulative product of the number of methods in each step. An option to bypass the step completely was also included to investigate the effect this has on model performance. In the case of this demonstration, referring to the list in section 5.2.1 gives (including bypass option): 3 smoothing methods; 2 baseline methods; 4 normalisation methods; 4 scaling methods; 2 decomposition methods and 3 classifier methods, which yields 576 unique pipelines. Spectra from approximately 2/3 of the patients were sampled for model optimisation and training, with the remaining 1/3 of patients left out to test the model. In order to boost speed and reduce bias towards patients, the number of spectra sampled from each patient was set equal at 200. This whole process was repeated 50 times for different train-test splits, in order to acquire a general picture of the performance of the optimised pipelines. Figure 5.8 details the end-to-end optimisation process.

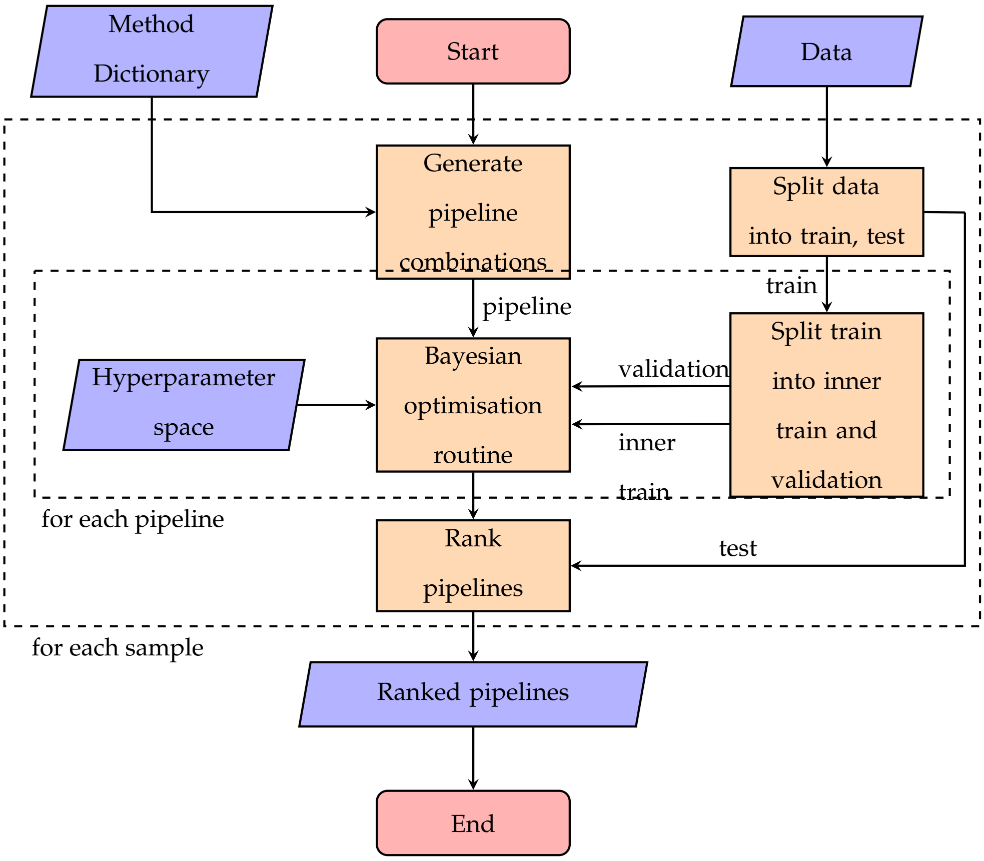


FIGURE 5.8: Flowchart of overall optimisation process

5.3.2 Results

The framework generated a total of $n_{pipe} \cdot n_{sample} = 574 \cdot 50 = 28700$ independent tasks to be executed using the HTCondor service. All processes took no longer than 48 hours to complete, and the results of each were subsequently recombined in order to compare performances. The pipelines were ranked by mean AUC to determine the optimum. Figure 5.9 shows various performance measures attributed to each of the 576 ranked pipelines.

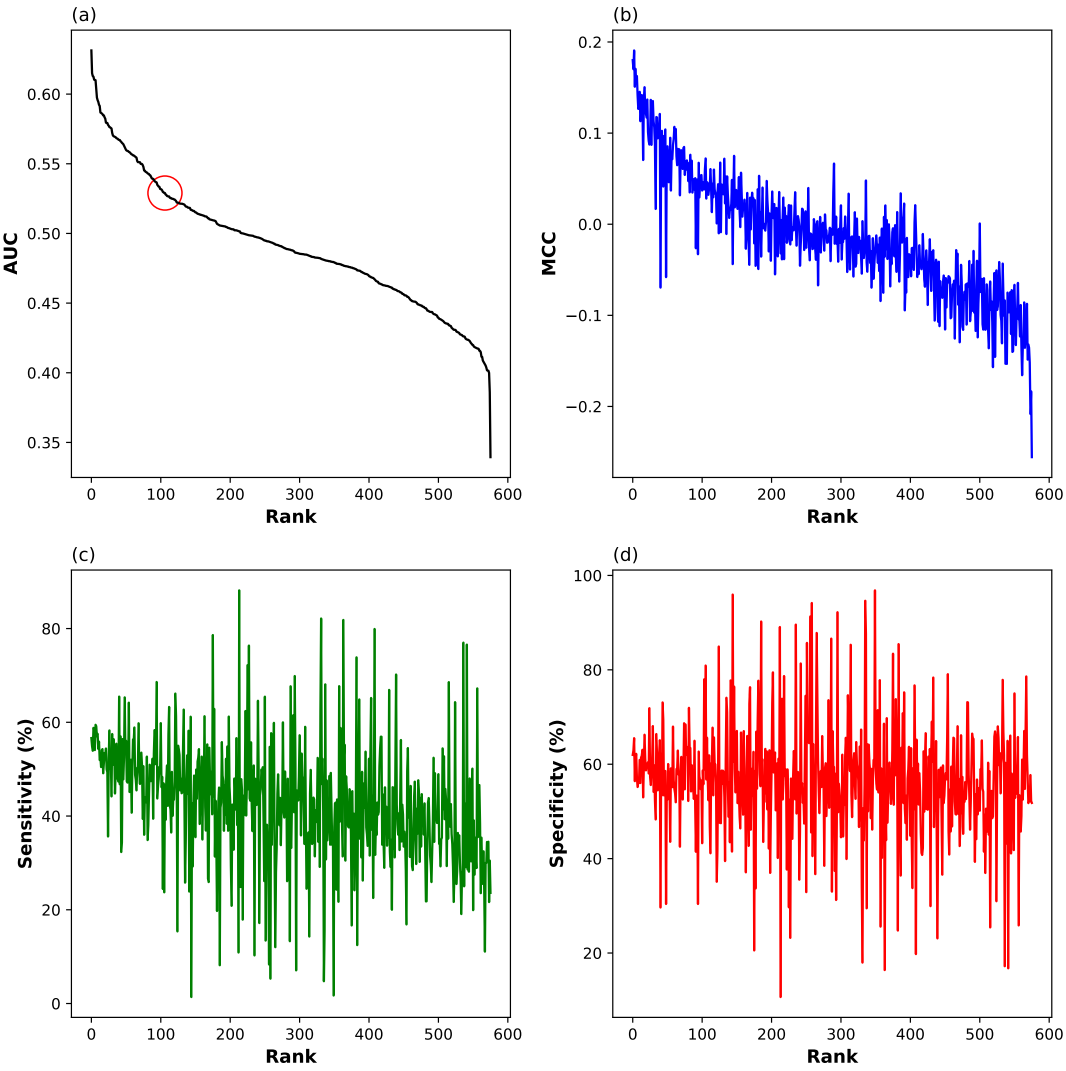


FIGURE 5.9: Mean scores of pipelines for various metrics [(a): AUC, (b): MCC, (c): sensitivity, (b): specificity] , ranked by AUC. Red circle in (a) marks pipeline which applies no pre-processing and logistic regression.

Table 5.1 summarises the various methods associated with each step in the top five pipelines. The number of HPs (n_θ) corresponding to each permutation is also shown. The mean performance of the best two pipelines over the 50 samples is portrayed in a confusion matrix (Fig. 5.10a,c) and ROC curve (Fig. 5.10b,d).

TABLE 5.1: Top ranking pipelines.

Smooth	Baseline	Norm	Scaling	FE	Classifier	n_θ	AUC
-	-	Amide I	Robust	-	LogReg	1	0.63 ± 0.02
SG	-	Min-Max	Standard	PCA	LogReg	3	0.62 ± 0.02
-	-	Vector	Robust	-	LogReg	1	0.61 ± 0.02
-	-	Amide I	Robust	-	LogReg	1	0.61 ± 0.02
-	Rubber	Vector	Standard	PCA	LogReg	2	0.61 ± 0.02

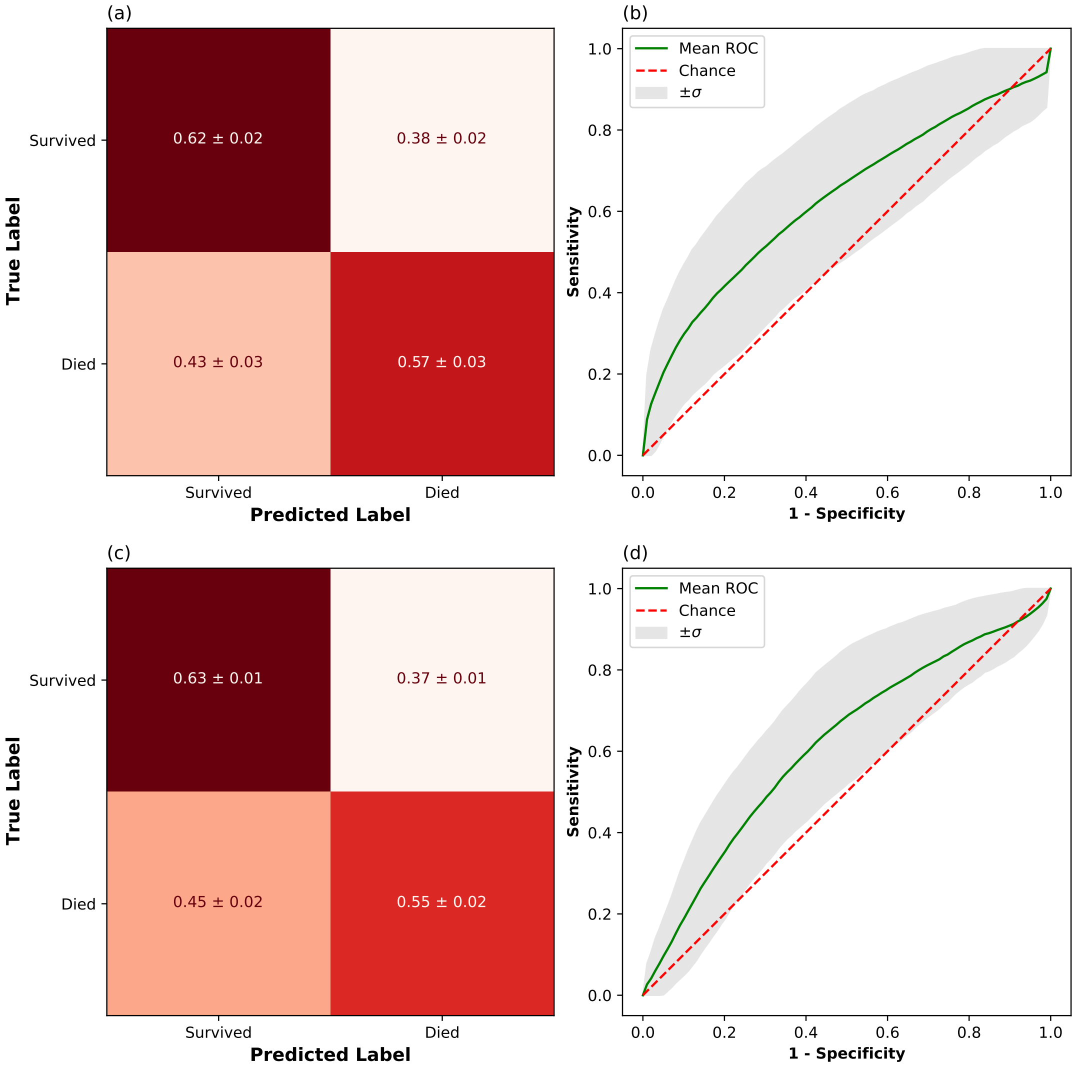


FIGURE 5.10: Mean confusion matrix and ROC curve shown with standard errors for best (a,b) and second best (c,d) pipelines as shown in table 5.1.

Figure 5.11 shows the value of the objective function at different points in HP space according to the GP model. As the HP space is 2-dimensional for this pipeline, the function can be displayed as a 'loss surface', which in this case is equal to $-\overline{AUC}$ (mean AUC score). The loss surface is apparently dependent on the number of components used in the feature extraction (PCA decomposition) step of the pipeline. On the other hand, the C parameter associated with the logistic regression classifier appears to have little influence over model performance. 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 indicated a high regularisation effect, this would likely be detrimental to the classification accuracy and so the score for that pipeline would be low.

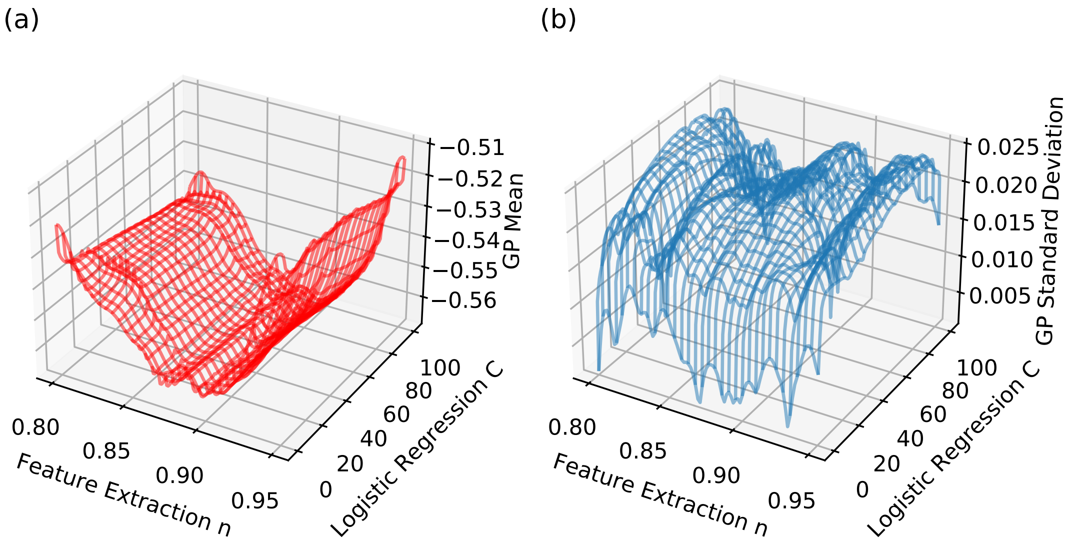


FIGURE 5.11: GP hyperparameter surfaces showing mean function in red and standard deviations in blue averaged across 50 sample iterations.

Each of the 50 samples have a unique set of optimised HPs, confirming that the choice of training and validation data has a significant impact on optimal selection. Histograms of the chosen HPs over the 50 samples of the best 2 pipelines is shown in Fig. 5.12.

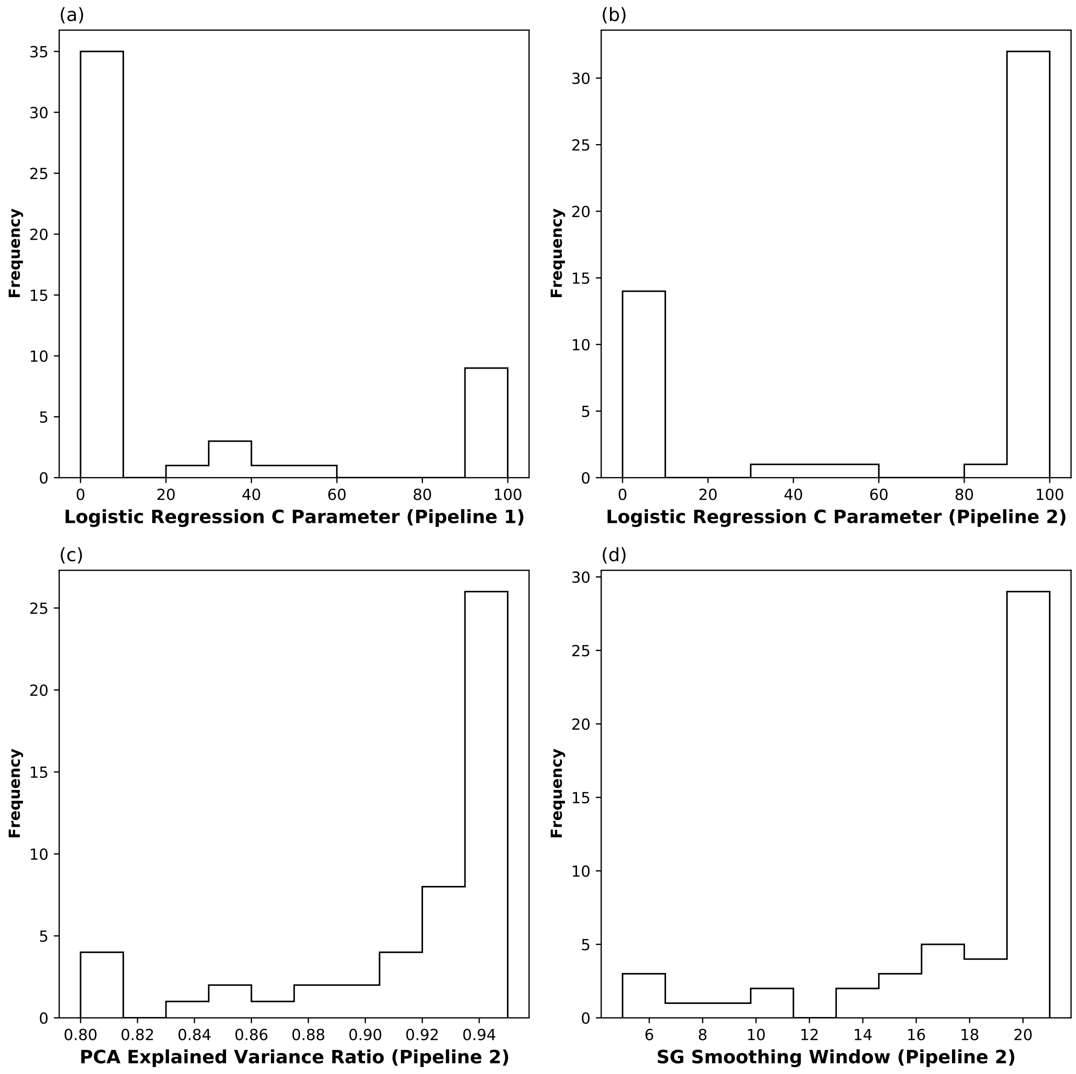


FIGURE 5.12: Histograms of optimum hyperparameters over the 50 train-test splits.

The strategy of randomly sampling 200 spectra from each patient was chosen so that patient related biases did not influence the choice of methods and HPs in the optimised model. To acquire a more complete measure of performance, the optimised pipelines were trained again with all spectra from each patient. The modal value from each of the parameters in Fig. 5.12 were used as the model parameters in this stage. The new results for the full dataset are shown in Fig. 5.13, where each of the sub-figures describes the same as in Fig. 5.10.

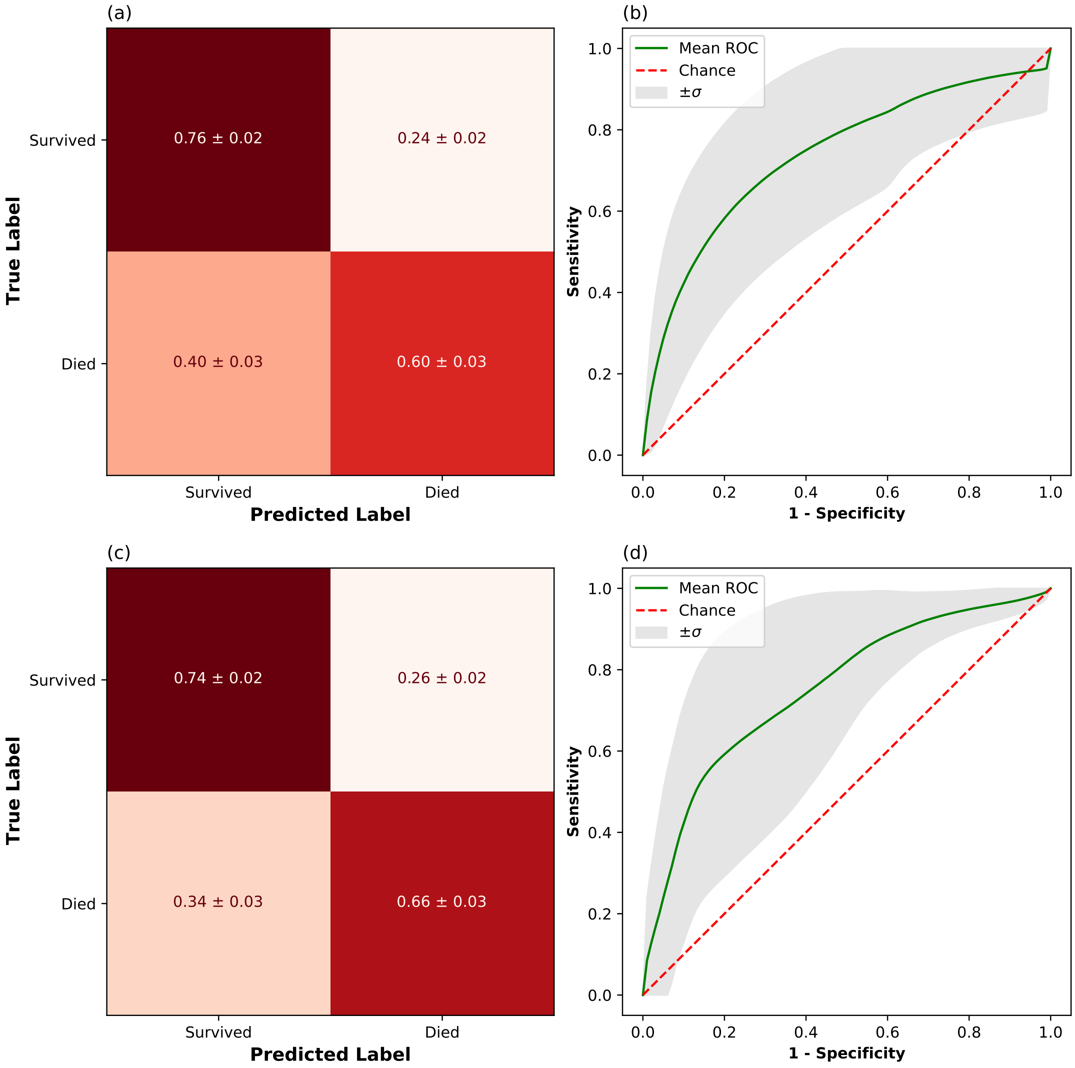


FIGURE 5.13: Mean confusion matrix and ROC curve shown with standard errors for best (a,b) and second best (c,d) pipelines trained and tested on full dataset.

Figure 5.13 shows a significant increase in both sensitivity and specificity when the optimised pipelines and HPs are deployed on the full dataset.

5.3.3 Discussion

The performance measures associated with each of the trialed pipelines show marked variation in when they are used to analyse this particular dataset, which is clearly demonstrated in Fig. 5.9. AUC and MCC are performance measures which take both sensitivity and specificity into account, which explains the similar trend shown in

Fig. 5.9a and b. 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 AUC or MCC favours pipelines with balanced sensitivity and specificity.

The trace in Fig. 5.9a begins with a small number of relatively high scoring pipelines, before levelling off towards a mid region of scores which are distributed about an AUC of 0.5 and an MCC of 0.0. This implies that the pipelines in this central region and beyond have little to no classification skill whatsoever. This draws attention to high ranking side of Fig. 5.9, with the highest scoring pipelines summarised in table 5.1.

Table 5.1 makes it clear that the optimal classifier for this dataset is logistic regression, with various choices of pre-processing options preceding this step. Normalisation and scaling are never bypassed, suggesting this is an essential step if any subsequent classification is going to achieve decent scores. Two instances in the top five classifiers utilise PCA to transform the high dimensional data in to a smaller space, suggesting that this step is not 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 5.

In order to gain more insight into 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 plotted. Here, the reference score is the median score of all pipelines in the analysis.

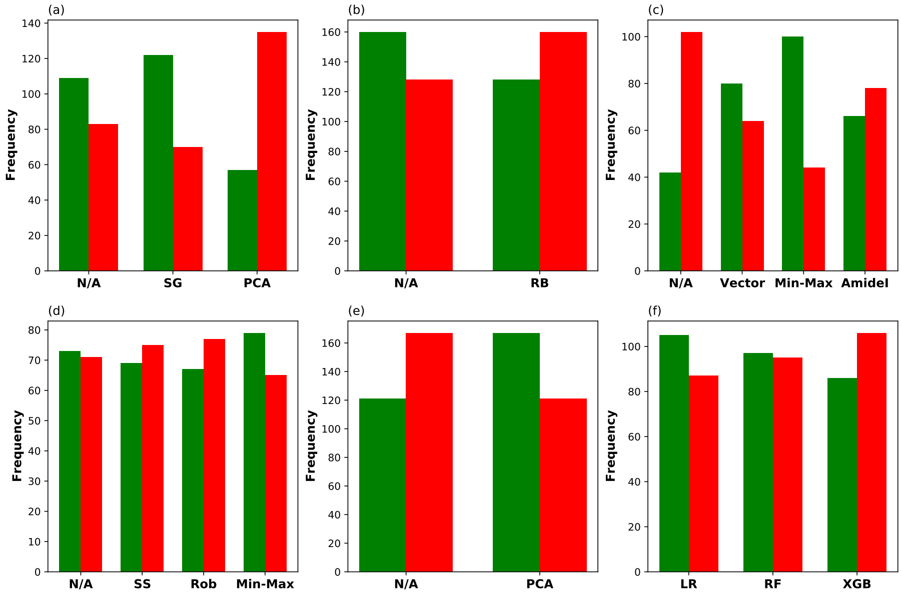


FIGURE 5.14: Frequency each method either enhances (green) or diminishes (red) relative to the median score (AUC = 0.48). Steps are (a) smoothing, (b) baseline, (c) normalisation, (d) scaling, (e) feature-extraction, (f) classifier.

Figure 5.14 shows some interesting insights into the effects of various methods to the performance of an analysis pipeline. The choice of smoothing method evidently has a significant effect, the majority of pipelines which utilise PCA denoising 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 subject to scatter correction prior to the analysis, negating the requirement to perform a baseline correction. Normalisation is evidently a step that can not be bypassed, a reasonable result as spectra originate from different samples, each with dissimilarities in sample thickness. 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 important. This is an interesting finding, and it indicates that the normalisation of each feature might be beneficial for logistic regression for this task. This conclusion is in fact in agreement with much of the consensus around algorithms which utilise a sequential optimiser, such as *gradient descent*, to converge on a solution. Gradient descent is an algorithm intimately related to the training of deep learning models, and governs the degree to which feature weightings are adjusted on each training iteration. For this reason it is often advantageous for the features to be scaled to within the same range.

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 AUC = 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 Fig. 5.12 show the frequency of values of the pipeline specific hyperparameters for each of the two top performing pipelines. Interestingly, it reveals that the logistic regression C value in pipeline 1 (Fig. 5.12a) converges to a much lower value ($\approx$ 0.01) than for pipeline 2, where it appears to converge towards 100. The lower the C value, the more regularisation, or less overfitting, there is to the training data. Pipeline 2 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 2, as it is already being deployed on a dataset which is difficult to overfit to.

Taking the modal hyperparameter selections from Fig. 5.12 and training and testing on all available data (using the same patients for each of the 50 train-test splits) enhances the scores significantly, exemplified in the mean confusion matrices and ROC curves in Fig. 5.13. For pipeline 1, there is a 14% increase in mean specificity, and a 3% increase in mean sensitivity. Pipeline 2 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 the all the available data from different patients should be used.

5.4 Conclusion

The work presented here demonstrates a versatile framework capable of determining a near optimal data pre-processing and classification pipeline in a holistic way. 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.

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

Chapter 6

Dysplasia Transformation Analysis

The work presented in this chapter is based on a project funded by cancer research UK (CRUK), which set out to investigate how IR techniques can be exploited to predict the transformation of oral epithelial dysplastic lesions into oral squamous cell carcinoma. I personally have been responsible for the sample preparation, data acquisition and data analysis in the project. At the time of submitting this thesis, a paper based on this work is in press for publication [156].

6.1 Background

As discussed in chapter 2, the risk stratification and management of patients with OED is primarily guided by histopathological grading of extracted tissue. The high subjectivity and inter/intra-observer variability of histopathological grading influences the accuracy of malignant transformation prediction. Furthermore, there is ambiguity and conflicting opinion as to whether much significance can be attributed to grading with respect to malignant transformation of OED [38], [50]–[53]. Given the clinical significance of OED, improvements to current diagnostic and prognostic pathways are desirable, and studies into potential biomarkers are widespread [62], [157]–[159].

Chapter 4 demonstrates the utility of combining vibrational spectroscopic techniques with machine learning algorithms as a tool to predict the histopathological identity of spectra derived from oral cancer tissue. There have been a limited number of studies applying this alternative methodology - predominantly Raman - to datasets involving OED. A study by Ibrahim *et al* demonstrated that high sensitivities and specificities could be achieved using Raman microscopy to discriminate between mild, moderate and severe dysplasia using discriminant analysis. Despite their exciting findings, the size of the patient cohort (n=4) was small, which implicates a lack of sufficient biological variability needed for any conclusive results. The follow-up study expanded the size of this cohort (n=57), but there was a marked reduction in sensitivity and specificity [160], which may be limited by the heterogeneity and size of the cohort. The same study used patient metadata such as smoking status, site of lesion, gender and alcohol status to subdivide the spectra. They found that spectra originating from patients with different alcohol consumption or different genders could be achieved. Discrimination between epithelial tissue from patients with different smoking status led to sound discrimination (AUC = 0.76), which indicates that smoking has an influence on the spectra in this cohort. Excellent discrimination was achieved when discriminating connective tissue with a high degree of inflammation, with increased nucleic acid and decreased collagen signals emerging as prominent predictors.

A study by Behl *et al* [161] set out to investigate whether Raman spectroscopy of exfoliated cells from dysplastic lesions could be differentiated from normal donor cells using partial least squares discriminant analysis (PLS-DA). They found that the spectra taken from the cell nuclei could be discriminated with a sensitivity of 86% and specificity of 85%, whilst models built from spectra derived from the cytoplasm of the cells increased the sensitivity to 96%. The discrimination was predominantly driven by lipidic contributions in the cytoplasm, which they attributed to the upregulation of the lipid metabolism within the cell to provide energy to the abnormal cells. Extraction of cytology specimens with brush cytology offers a minimally invasive method compared with repeated tissue biopsy extraction for monitoring PPOELs. Characteristics such as the nuclear area, cytoplasmic area and nuclear to cytoplasmic ratio form part of a routine cytopathological examination, however it is understood that various patient factors, such as age, gender and lifestyle habits influence morphological features and are not exclusively tied to the presence of dysplasia [162]–[164]. This study by Behl *et al* found that no discrimination could be achieved based on these patient factors, indicating there is some value in using vibrational spectroscopic techniques that objectively probe the chemical composition rather than subjective evaluation of morphological characteristics to diagnose dysplasia. This demonstrates a promising step in the right direction for early detection of OED.

Ghosh *et al* combined Raman and FTIR spectra of normal, dysplastic and malignant oral exfoliated cells in a novel integrated method for the complimentary techniques. The study recruited non-smokers and smokers to form two control groups, in addition to two patholigical groups consisting of patients with histopathologically defined OED and OSCC. They found that increase in DNA, protein and lipid content was correlated with malignancy, with corresponding features in the two sets of spectra emerging as important variables in a PCA-LDA model. Use of FTIR and Raman spectra independently led to classification accuracies of 85% and 82% respectively, whereas the integrated method yielded significantly enhanced scores of 98%, which elucidates the complimentary nature of the two techniques.

The inter- and intra-observer variability in grading OED discussed in chapter 2 imposes a significant limitation on studies which aim to discriminate between grades with alternate methodologies to the gold standard. The labels used to guide model training (dysplasia grade) exist within a subjective domain, so there is an inherent degree of uncertainty that propagates through model training and testing. Consensus labelling by independent specialists is desirable in order to reduce the mislabelling risk. A different approach would be to use soft classification where the uncertainty of each class label is incorporated into the model [165]. Furthermore, in addition to the imperfection in the grading system, there is ambiguity in whether histopathological grading is a strong predictor for malignant transformation risk [38], [50]–[53].

Instead of using ML to attempt to diagnose and grade dysplasia, there have been numerous studies into using clinical end-point as labels, effectively removing the several layers of ambiguity relating dysplasia grading to OSCC transformation. For pre-malignant conditions such as OED, the clinical end-point is a binary variable which simply states whether the patient has developed cancer in a defined timeframe. Models which can successfully delineate transforming and non-transforming lesions would not only offer a promising contribution to early diagnostics, they may also stimulate discussion around the reasons why the discrimination arises, especially if the choice of algorithm has a degree of interpretability. Saintigny *et al* [166] performed a study on 86 patients with PPOELs, attempting to use a combination of univariate and multivariate algorithms to predict the transformation on the basis of genetic sequencing data. The model which incorporated the genetic data outperformed one based solely on clinical and histopathological factors, indicating that additional predictive value can be acquired from techniques that are not currently part of the gold standard. The authors also derived important individual predictors from the analysis, attributing proteasome machinery and ribosomal components as the gene sets which contribute significantly to the prediction.

Liu *et al* [167] used exfoliative cytology from patients with oral leukoplakia to obtain a DNA index, which reflected the number of cells with an abnormal number of chromosomes, a phenotype which has been correlated with malignant transformation [168]. The authors combined this data with clinical factors and statistical modelling to produce a quantitative measure of transformation risk that could risk stratify the lesions that had potential to transform with very good specificity.

Given that vibrational spectroscopy objectively probes the chemistry of a sample, it is possible that the genetic expression and molecular pathways that are related to malignant transformation of OED can be determined and used in a similar statistical model as the studies described previously. Despite this, there has only been a small number of studies which leverage this rich information source as a tool to predict disease. In fact, to the author's knowledge, there has been no published work on the prediction of OED to OSCC transformation using either FTIR or Raman spectroscopy.

The work within this chapter presents a retrospective pilot study carried out on a small cohort of patients presenting with PPOELs with histopathologically defined OED. The primary objective of the study is to assess the feasibility of utilising FTIR-MS as an early diagnostic adjunct.

6.2 Experimental

6.2.1 Sample Selection & Preparation

Thirty patients with OED that had been diagnosed by a pathologist were included for the study. The patients were seeded from a larger cohort used in a study to determine the clinical risk factors associated with malignant transformation [49]. The patients in the study had all given written informed consent to an NHS Research Ethics Committee. The selection of patients was limited by an inclusion criteria which required absence of previous OSCC, follow-up data without lesion excision, and the availability of FFPE tissue in the biobank at the University of Liverpool.

For each of the selected patients which transformed (T), a single archival FFPE tissue block was obtained at the closest timepoint to transformation, which ranged from 2 to 43 months prior to transformation. Patients who had not transformed from 43-108 months were defined in this study as non-transformers (NT), as they had not developed cancer from lesion that had not been excised.

Four 5 µm serial sections from the FFPE block were obtained using a manual microtome, and in a similar protocol to the one detailed in the previous study (chapter 4), the first and second were reserved for routine H&E staining and histopathological examination. The remaining two sections were mounted onto two CaF_2 disks for FTIR-MS experiments. LM images of each H&E stained section were acquired using an Aperio CS2 scanner (Leica Biosystems), and were examined by a histopathologist to define regions and grade of dysplasia. Regions which were representative examples of the different dysplastic grades were identified and marked as target ROIs for subsequent FTIR-MS experiments.

6.2.2 FTIR Experiments

FTIR-MS experiments were conducted in a very similar way as described previously in section 4.2.2. The only difference in protocol was halving the number of background and sample scans from 256 to 128 and 128 to 64 respectively. This decision was made after considering the much larger cohort of patients and limited time using the instrument.

6.2.3 Labelling & Quality Checks

The raw hyperspectral image (HSIs) contained multiple different tissue types aside from OED. In order to build valid models that exclusively use spectra derived from OED, a much more specific labelling than the one used in chapter 4 was required. A hybridised approach which utilised both histopathological opinion and unsupervised analysis was designed and employed, so that specific labelling of similar spectra could be achieved.

The unsupervised analysis branch of the method involved applying a two-tiered *k*-means cluster analysis (KCA) to each image. Each image was first pre-processed using a standard protocol of a 2nd order SG differentiation (window = 5, order = 2), spectral truncation removing the paraffin region before vector normalisation. The number of clusters was initially set to 5, then under the guidance of a histopathologist the number of clusters was adjusted until the borders of the surface epithelium could be delineated from the surrounding tissue.

The spectra contained within the cluster(s) identified as epithelium were subject to another round of KCA in order to identify the dysplastic layers within. The origin of squamous cells is in the basal layer of the epithelium, and all grades of dysplasia are defined as having abnormalities in at least this layer [169]. For this reason, only spectra from that layer were to be extracted and used in subsequent modelling. The process is outlined below in Fig. 6.1.

The H&E image was examined to determine the histopathological grade of the OED lesion. Given the continuum of dysplastic changes and the preferences of the histopathologist, 5 grades of OED were used to group the spectra: *mild* (G1), *mild-moderate* (G2), *moderate* (G3), *moderate-severe* (G4) and *severe* (G5). A binary labelling system was also used, which defined G1 and G2 lesions as *low-grade* (L) and G3-G5 as *high-grade* (H).

The labelled spectra from each individual HSI were subject to a statistical quality check. This entailed the mean centering and decomposition of the spectra into five principal components using PCA, followed by using Hotelling's T^2 test to determine whether each spectrum agrees with the general covariance within the data. It is the multivariate generalisation of the univariate Student's *t*-test, which is used to determine statistical differences between sample means. The scores t and corresponding standard deviation σ of the PCA model are used in the determination of the Hotelling's T^2 statistic, described by Eq. (6.1).

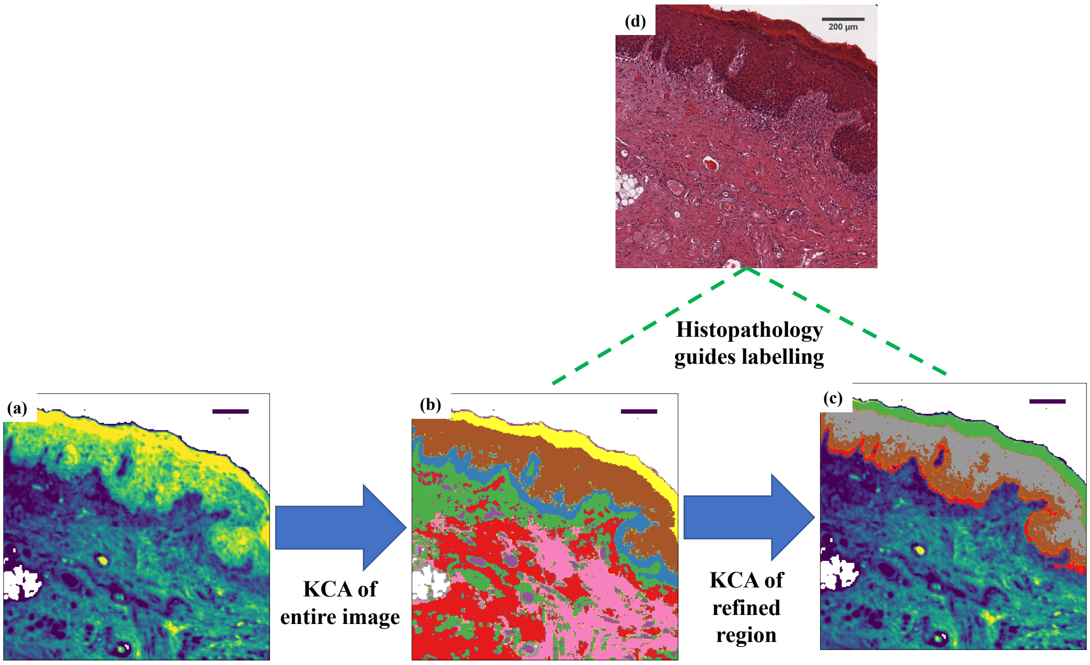


FIGURE 6.1: Workflow for the labelling protocol. Pre-processed image (a) is subject to two tiers of KCA (b-c), with the epithelial and dysplastic regions guided and confirmed by histopathologist informed by H&E image (d). The dysplastic layer in this image were defined as the brown cluster.

$$T^2 = \sum_{i=1}^{n_{components}} \frac{t_i^2}{\sigma_i^2} \tag{6.1}$$

Spectra which yielded a statistic that lay outside of the 95% confidence bounds of the distribution were discarded. This is a popular choice of pre-processing in the vibrational spectroscopic field due to it's automation and suitability to multivariate data [140], [170].

The projection of each spectrum onto the first three PCs in Fig. 6.2a shows a central red cluster that resembles normally distributed variables. This is to be expected, as the labelled spectra are from the same distinct pathology from within the same image, so systematic influences arising from biological and instrumental variations are not present. The outlier spectra are marked as black crosses, and are found on the periphery of the 3D scatter. The origin of the outliers are marked as red dots in Fig. 6.2b, which show that outlier spectra originate from localised clusters rather than a random

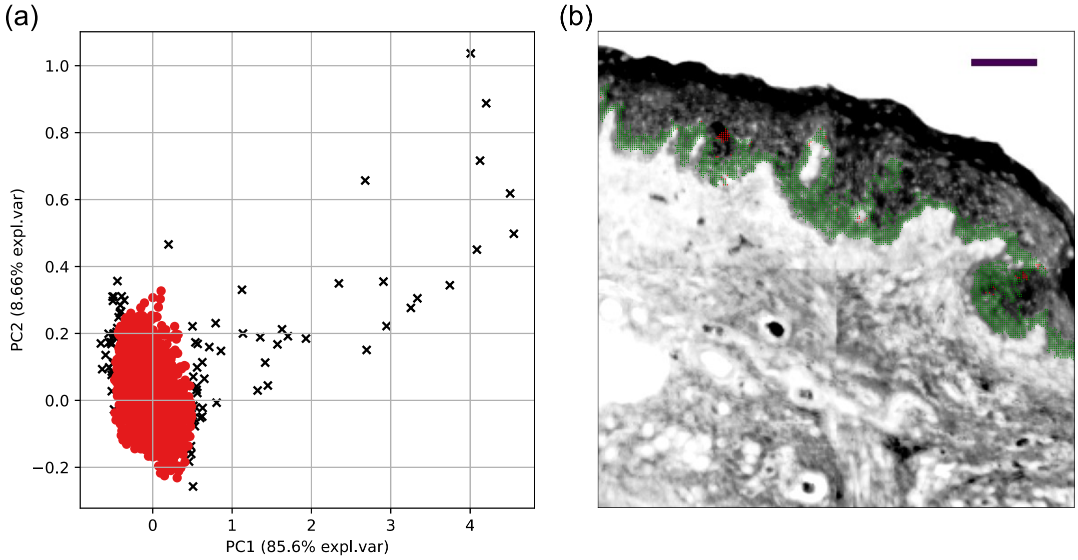


FIGURE 6.2: Example of a quality control test for data from a single image. Black crosses in (a) and red pixels in (b) indicate data points which are deemed outliers by the test. (b) IR image at 1650 cm^{-1}. Green dots indicate non-outliers. Scale bar = 200 µm.

scattering. These localised clusters may be regions of sample that lead to high scattering, mis-labellings, or even debris on the bottom of the CaF_2 disk. Regardless of the origin, these points were discarded from the data, as extreme points can significantly skew pre-processing and classification [140].

6.2.4 Pre-processing

The data was split in a number of ways to build different models to discriminate between transforming and non-transforming spectra for different groups of patients, which will be described in more detail in the relevant sections. The `PipeOpt` framework described in chapter 5 was used to rigorously optimise the analytical pipeline for each independent model. Prior to model optimisation, the data was copied and subsequently run through an extended multiplicative signal correction (EMSC) Mie correction algorithm described by Kohler *et al* [81]. This particular algorithm was preferred due to relatively high throughput compared to Bassan's RMieSC-EMSC algorithm. The resonant component which leads to dispersive artefacts is also far less pronounced for non-isolated cells and tissue [82], especially if the latter is embedded in paraffin wax [85].

The EMSC approach utilised in the Mie correction algorithm first calculates the scattering efficiency curves (Eq. (3.23) for a range of particle sizes across the defined wavenumber range. The simulated curves are decomposed using PCA into a predefined number of PCs. The algorithm then optimises the parameters for an EMSC model which minimises the differences between each apparent spectrum and the reference spectrum offset and scaled by the scattering efficiency model. These optimised parameters are then used to correct the apparent spectrum. This approach was chosen to account for the different sizes of scattering objects that may be present in heterogeneous tissue.

The computational cost of correcting each spectrum leads to prohibitively long runtimes for large numbers of spectra. To circumvent this issue, the HTCondor service described in chapter 5 was used to dispatch smaller numbers of spectra to run in parallel, as the correction of each spectrum is independent of the dataset as a whole. The reference spectrum used in the correction for each HSI was defined as its mean spectrum.

The method and HP search spaces used in the `PipeOpt` optimisation phase are the same for each model described in this chapter. Figure 6.3 depicts the different trialled steps and methods.

The Mie correction step precedes `PipeOpt` since the algorithm relies on HTCondor to efficiently process the spectra, so it is not possible to integrate into the framework in its current form. Each pre-processing step (including Mie correction) has a bypass option, yielding a total number of $2 \cdot 3 \cdot 3 \cdot 4 \cdot 3 \cdot 2 \cdot 3 = 1296$ possible pipelines. The HPs within each pipeline are optimised over a maximum number of 30 iterations, with the mean AUC acting as the objective function to be optimised.

6.3 Results

6.3.1 All Patients Model

The primary objective of this project is to build a model which is able to predict the outcome of patients with histopathologically defined OED based solely on spectral information, disregarding clinical and histopathological factors. In order to approach

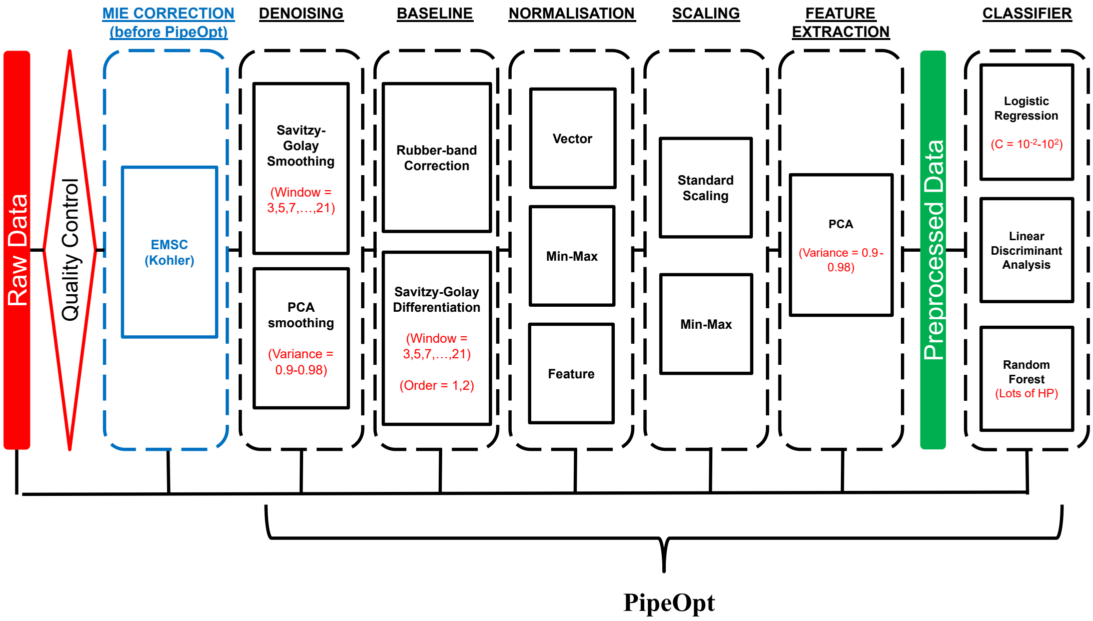


FIGURE 6.3: The method and HP search spaces used to optimise each model.

this problem, the model described in this section will include data acquired from all patients in the study. A table of patients and all associated metadata can be found in the appendix. A total of 30 patients (14 non-transformers, 16 transformers) were selected for model training, based on the inclusion criteria detailed previously.

Figure 6.4 shows the mean spectra for different groupings of data. It reveals very similar spectral line shapes, which is to be expected considering all are from the same pathological origin. Furthermore, even if there were clear differences in the mean spectra, the variance within the dataset (indicated by the shaded regions) would suggest it implausible to identify any statistically significant differences between the spectra from each group. The spectral variance may be attributed to the fact that no pre-processing has yet been applied to the data, therefore physical effects such as differing sample thicknesses and scattering interferences have not been mitigated for and significantly influence the appearance of spectra. It also reflects the biological heterogeneity in the dataset.

In order to acquire a robust measure of model performance, the optimisation process was repeated for every combination of one non-transforming and one transforming patient forming the hold-out test set. The data from the remaining patients in each

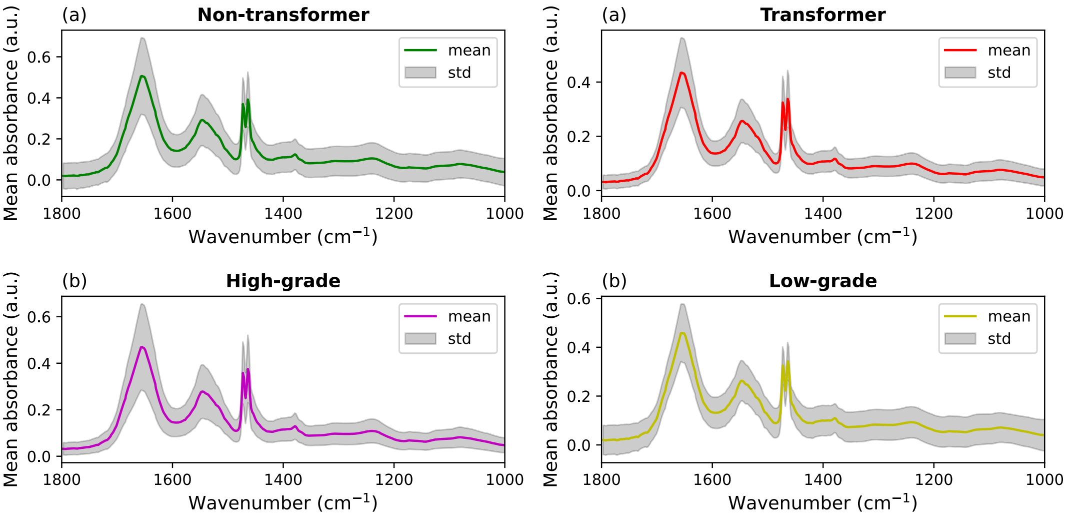


FIGURE 6.4: Mean spectra when the data is grouped by (a) outcome and (b) grade.

split was used in the Bayesian HP optimisation and model training. This approach essentially mimics the clinical scenario where the test would be on one single patient, but instead a patient from each group is used in order to attain both sensitivity and specificity scores. The total number of patient pairings can be easily determined by considering that each combination p_T, p_{NT} is an element in an $N(p_{NT}), N(p_T)$ matrix, where $N(p_{NT})$ amd $N(p_T)$ are 14 and 16 respectively, leading to a total of 224 patient pairings. This leads to a total number of pipelines to be optimised of $224 \cdot 1269 = 284296$, each of which was dispatched as an independent job to `PipeOpt`. An equal number of spectra from each biopsy were sampled to prevent patient or image bias from influencing optimisation results.

$$\begin{bmatrix} \boldsymbol{T_1, NT_1} & \cdots & \boldsymbol{T_{16}, NT_1} \\ \vdots & \ddots & \vdots \\ \boldsymbol{T_1, NT_{14}} & \cdots & \boldsymbol{T_{16}, NT_{14}} \end{bmatrix}$$

FIGURE 6.5: A matrix depicting the patient pairings from the two outcome groups.

Figure 6.6 shows the mean MCC (a), AUC (b), sensitivity (c) and specificity (d) for each pipeline, ranked in order of decreasing MCC score. The red circle in (a) marks the pipeline which applies no pre-processing and a logistic regression classifier, which in this case has a low score compared to the majority of pipelines, demonstrating the detrimental impact not pre-processing data has on results. The trace for AUC in (b) shows a much similar trend to MCC, as these are both similar performance measures which aim to quantify the overall efficacy of a test, considering both sensitivity and specificity. On the other hand, the sensitivity and specificity plots show a much more noisy trend, reflecting the dependence on the pipeline used and the trade-off between the two measures.

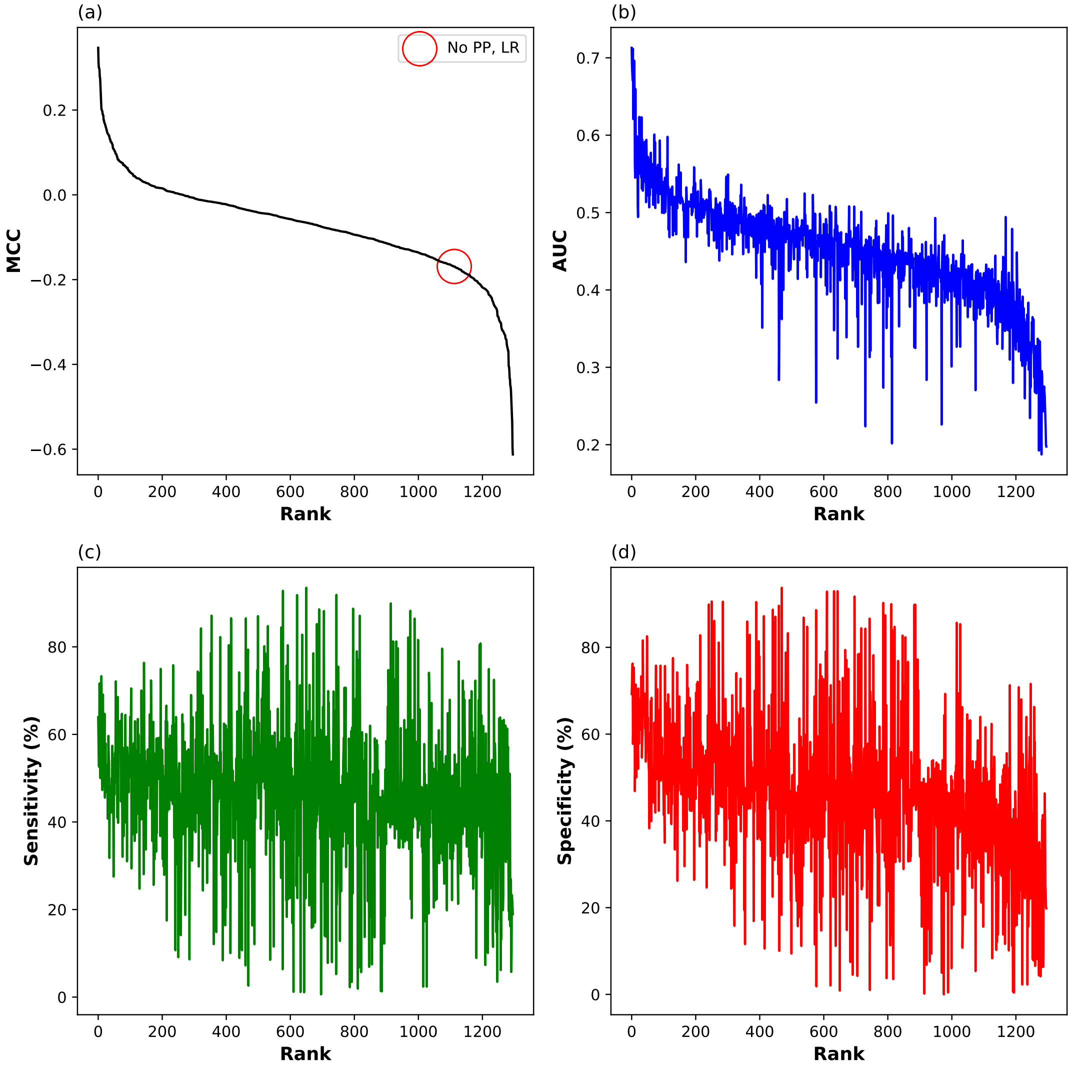


FIGURE 6.6: Mean MCC (a), AUC (b), sensitivity (c) and specificity (d) for pipelines (sorted in order of decreasing mean MCC). All patients incorporated into study.

Table 6.1 lists the top 5 pipelines (ranked by MCC) returned by PipeOpt. The classifier option in each of the five cases is random forest (RF), which implies that this classifier is the best choice for this particular dataset. The normalisation option is not bypassed in any of the best pipelines, which is in agreement with the conclusion made from Fig. 5.14, confirming that normalisation is a vital step to mitigate for differences in sample thicknesses in a heterogeneous multi-sample dataset.

TABLE 6.1: Top ranking pipelines for all dysplasia patients model.

Mie	Smooth	Baseline	Norm	Scaling	FE	Classifier
Yes	SG	N/A	Vector	N/A	N/A	RF
Yes	N/A	N/A	Vector	N/A	N/A	RF
Yes	N/A	SG diff	Feature	N/A	N/A	RF
No	N/A	SG diff	Feature	N/A	N/A	RF
Yes	PCA	N/A	Vector	N/A	PCA	RF

Figure 6.6 suggests that modest scores may be attainable when using an optimised pipeline to predict transformation of any grade of dysplasia, but it does not convey the variance of results over the different patient splits. Figure 6.7 shows the mean (green) and median (blue) ROC curves, as well as the individual ROC curves for all 224 splits of the data shown in paler colours. There is a marked difference between the mean curve and median curve, which is reflective of the fact that the curves from each split are not normally distributed, there are clearly more models that are concentrated in the upper left half of the plot, indicating that the model has moderate to excellent skill in predicting the outcome of the patients in those particular splits. There is clearly a significant variation in results depending on the choice of patients used for training and testing.

If this methodology were to be adopted in the clinic, patients and practitioners would require an outcome prediction on a patient basis, rather than each spectrum. Two approaches are proposed to use spectral predictions to determine a patient score: *hard voting* and *soft voting*. For every spectral prediction, there is an associated probability of it belonging to either class (NT or T). The class with the maximum probability is assigned as the predicted class. In the hard voting method, all spectral predictions for each patient are counted, with the class with the most votes N_{NT}, N_T assigned as the patient outcome prediction. Soft voting, which is also described in section 3.3.2, calculates the sum of the probabilities $\hat{y}_T, \hat{y}_{NT}$ over all the predictions for each patient to determine the outcome. Equations (6.2) and (6.3) summarises the two methods.

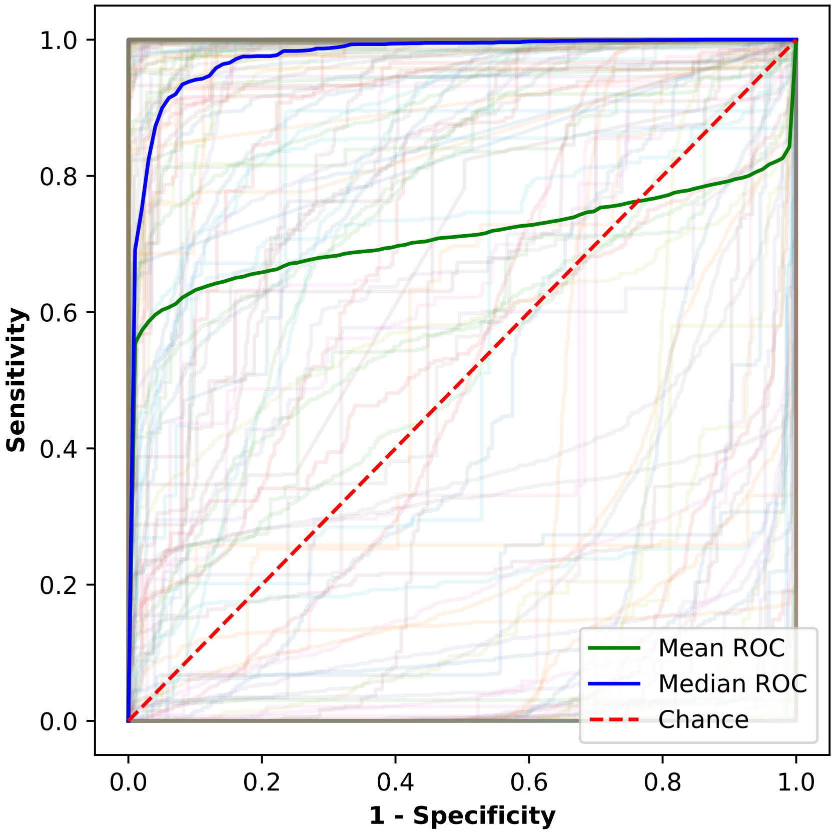


FIGURE 6.7: Mean and median ROC curves over the 224 individual splits, which are shown in paler colours.

$$y_{hard} = \underset{NT,T}{\operatorname{argmax}} \{N_{NT}, N_T\} \tag{6.2}$$

$$y_{soft} = \underset{NT,T}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n_{spectra}} \hat{y}_i(NT), \sum_{i=1}^{n_{spectra}} \hat{y}_i(T) \right\} \tag{6.3}$$

The advantage of using the soft voting method is that predictions with a high probability carry much higher weight than the more ambiguous ones. Consider a patient with 21 spectra, 10 of which are labelled as transforming each with an 80% probability, the remaining 11 are labelled as non-transforming each with a 55% probability, The hard voting scheme would predict the patient outcome as non-transforming as this is the class with the most votes, whereas the soft voting scheme would predict a transforming outcome, as the sum of probabilities is $0.8 \cdot 10 = 8$, versus $0.2 \cdot 10 = 2$ for non-transforming. Figure 6.8 shows the number of times patient outcomes were correctly predicted over the 224 splits for both voting strategies.

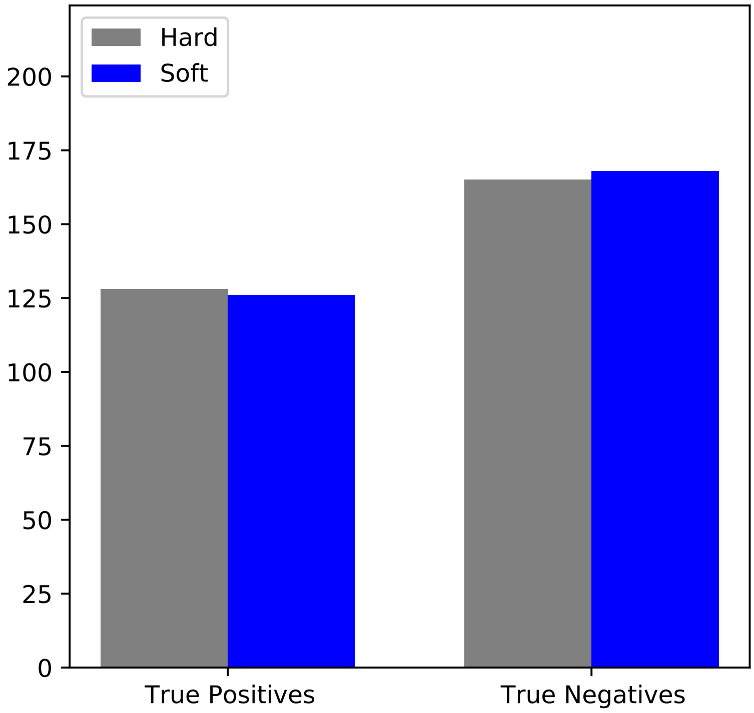


FIGURE 6.8: Number of times patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies.

The results shown in Fig. 6.8 indicate that there is very little difference between the results acquired using either hard or soft voting, and can be used to estimate the expected sensitivity and specificity of the model when used to predict the outcome of a new patient. The likelihood of correctly identifying a patient who will transform based on these results is 56-57%, whilst the likelihood of correctly identifying a patient who won't transform is 74-75%.

These findings would suggest, according to this limited dataset, that there is little feasibility in using this method as a means to predict patient outcome, regardless of their histopathological grade. This is perhaps unsurprising, considering the OED grading system measures architectural and cytological alterations that are manifestations of underlying biochemical changes. Combining patients with different grades of OED into one model may lead to overfitting to features characteristic of the grade, rather than achieve the goal of discriminating based on features that reflect propensity to malignant transformation.

For this reason, it was decided to investigate whether stratifying the dataset according to the binary grading system applied (high grade, low grade), would improve results. The rationale for this study is that one of the sources of biochemical variation would be controlled for, forming a more homogeneous set, allowing for the model to discriminate on the basis of features that drive transformation. The binary system also reduces the ambiguity and granularity of the 5-tiered grading system intially employed to grade each lesion, and such an approach has been shown in other studies to improve inter- and intra-observer variability [171].

6.3.2 Low Grade Patients Model

The process of optimisation and model training was repeated for patients with a histopathological grading of G1 and G2, who were defined as low-grade (L) in the binary grade system used in this study. Consequently, the size of the patient dataset was dramatically reduced from 30 (14 NT, 16 T) to 13 (7 NT, 6 T). The same splitting method was used to generate all pairings of patients from both groups, resulting in $7 \cdot 6 = 42$ iterations to evaluate. The same PipeOpt search spaces and settings were used to determine the optimum analytical pipeline. Figure 6.9 are the score traces for this particular task, showing the same metrics as Fig. 6.6.

Figure 6.9 clearly shows a significant improvement in model performance when training on patients with similar histopathological characteristics. The five pipelines with the highest MCC are further detailed in table 6.2.

TABLE 6.2: Top ranking pipelines for low grade dysplasia patients model.

Mie	Smooth	Baseline	Norm	Scaling	FE	Classifier
Yes	SG	SG diff	N/A	Standard	PCA	LR
Yes	N/A	SG diff	N/A	Standard	N/A	LR
Yes	N/A	SG diff	N/A	N/A	PCA	LR
Yes	SG	SG diff	N/A	Standard	N/A	LR
Yes	N/A	SG diff	N/A	Standard	PCA	LDA

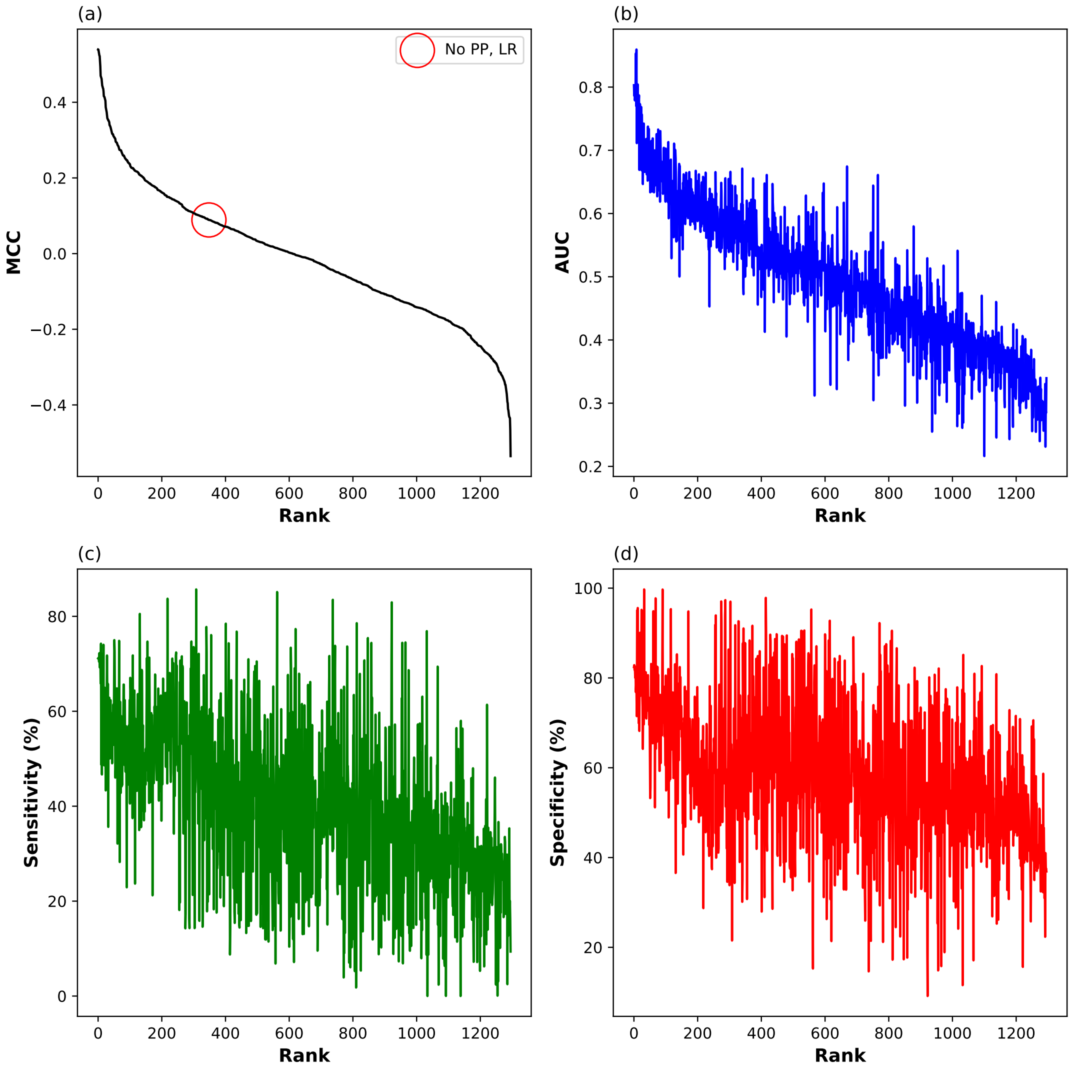


FIGURE 6.9: Mean MCC (a), AUC (b), sensitivity (c) and specificity (d) for pipelines (sorted in order of decreasing mean MCC). Low grade patients study.

Interestingly, spectral normalisation is a step that has been bypassed in each of the top performing pipelines. This may be explained by the fact that the data is being standardised in four out of five pipelines, which scales each wavenumber variable in the data to have a mean of zero and standard deviation of 1. This effectively transforms the data to exist within a common domain, which could be accounting for the differences in scale observed in uncorrected spectra from samples of different thickness. The importance of standardisation in these results is supported by the choice of classifier, which is either logistic regression or LDA. Both of these models apply a linear operation to transform a spectrum into a new domain, generally by determining a set of co-efficients which minimises a loss function in model training. Standardisation equalises the potential of the weight each wavenumber has in the model [142]. RF models, on the other hand, evidently do not benefit from standardised variables as they do not use linear transformations to make decisions, rather they make decisions based on an ensemble of rules.

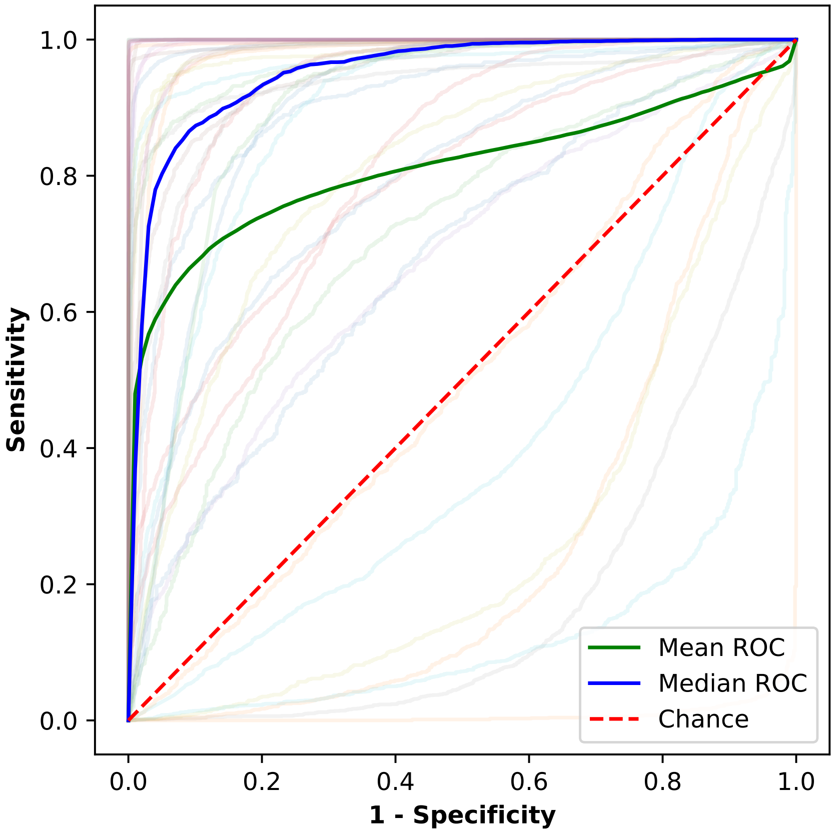


FIGURE 6.10: Mean and median ROC curves over the 42 individual splits (shown in paler colours) for low grade patients.

The mean, median and individual ROC curves are all shown in Fig. 6.14. The plots show a marked improvement in both the variation and performance of the model over the 42 patient pairings. The individual ROC curves also show that seven models (16.7%) have no classification skill, with some splits resulting in an AUC of much less than 0.5, which indicates that test spectra are being labelled incorrectly more often than not. In order to investigate the potential sources of the low scoring models, the mislabelling rate for both groups of patients is shown in Fig. 6.11. Five out of six who did go on to develop OSCC were incorrectly predicted in less than 30% of the models, with one patient (12092) achieving a 100% success rate in transformation prediction.

There is one patient who did not transform (12182) with very poor outcome prediction accuracy, whilst three out of four of the non-transforming patients were correctly predicted as such 100% of the time.

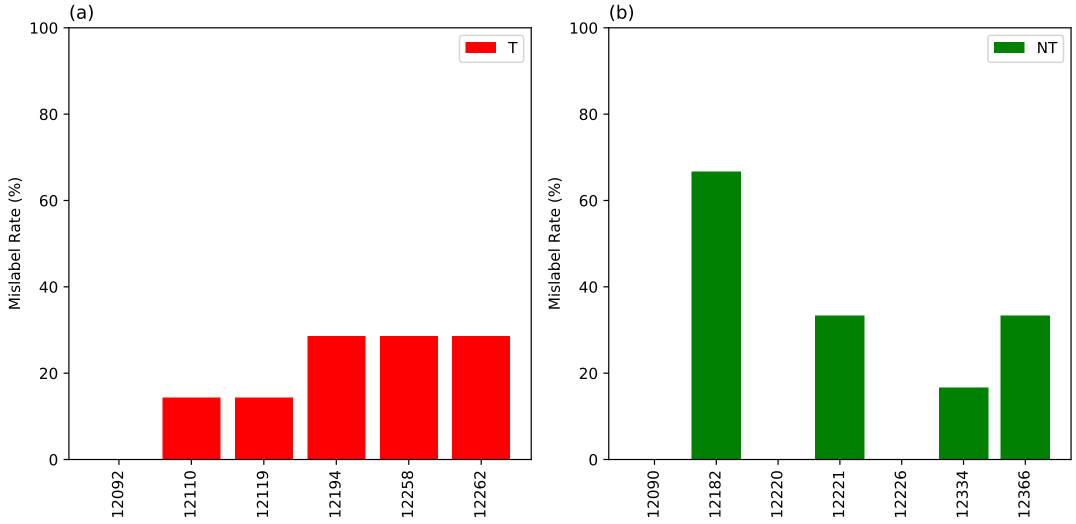


FIGURE 6.11: (a) Mislabelling rates for each transforming (a) and non-transforming (b) patient.

The overall sensitivity and specificity associated with the hard-voting and soft-voting strategies are shown in Fig. 6.12. The results from either approach appear to be identical, which reiterates that the predictions made by each model have a high degree of confidence. Based on these results, the likelihood of correctly predicting the outcome of a patient with low-grade dysplasia as transforming is 80-81%, whilst the likelihood of correctly predicting the outcome as non-transforming is 78-79%.

6.3.3 High Grade Patients Model

The final group of patients to be investigated were ones presenting with histopathologically defined OED of moderate (G3) to severe (G5) severity. Similar to the low-grade cohort, this sub-grouping was constructed to better control for the biochemical variance attributed to different grades of dysplasia. The same PipeOpt configuration was used to optimise the pipeline methods and HPs for predicting transformation of spectra extracted from the lesion. There are 70 patients in the high grade cohort, 10 of whom transformed to OSCC and 7 of whom had not. Using the same data splitting

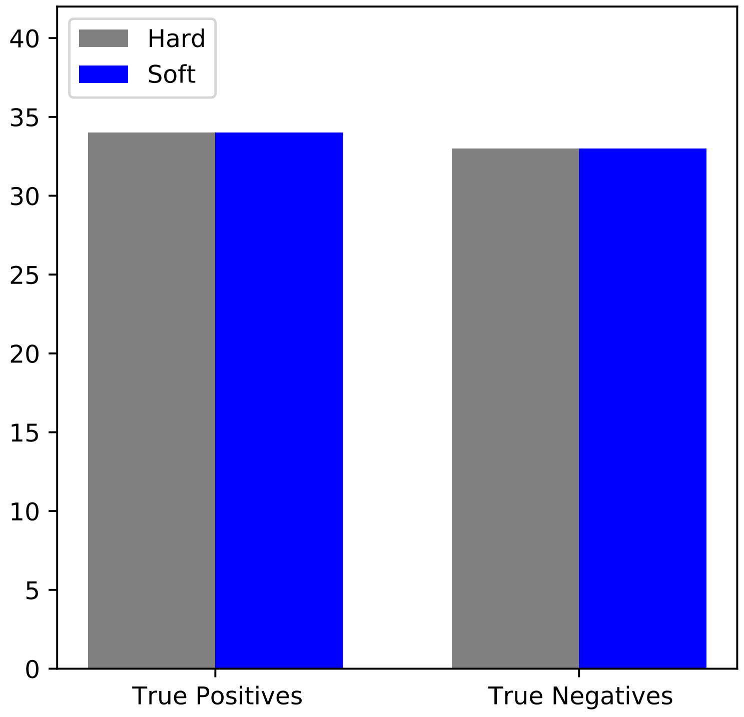


FIGURE 6.12: Number of times low-grade patients patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies.

strategy as in the other models would generate $10 \cdot 7 = 70$ combinations of T and NT patients.

Figure 6.13 shows the various performance measures of each pipeline ranked by MCC score. The general trend of the MCC trace is similar to Figs. 6.6 and 6.9, where there is a small region of relatively high performing pipelines followed by a steep drop-off into much more mediocre scores, before another steep drop off into pipelines with very poor performance. This variance again emphasises the importance of optimising the analytical pipeline, whilst also drawing attention to the fact that some combinations are not suitable for analysis of this dataset, which will be discussed in section 6.4

Figure 6.13 again shows a marked improvement in performance compared to the model which incorporated all patients regardless of OED grade. There is a slight drop

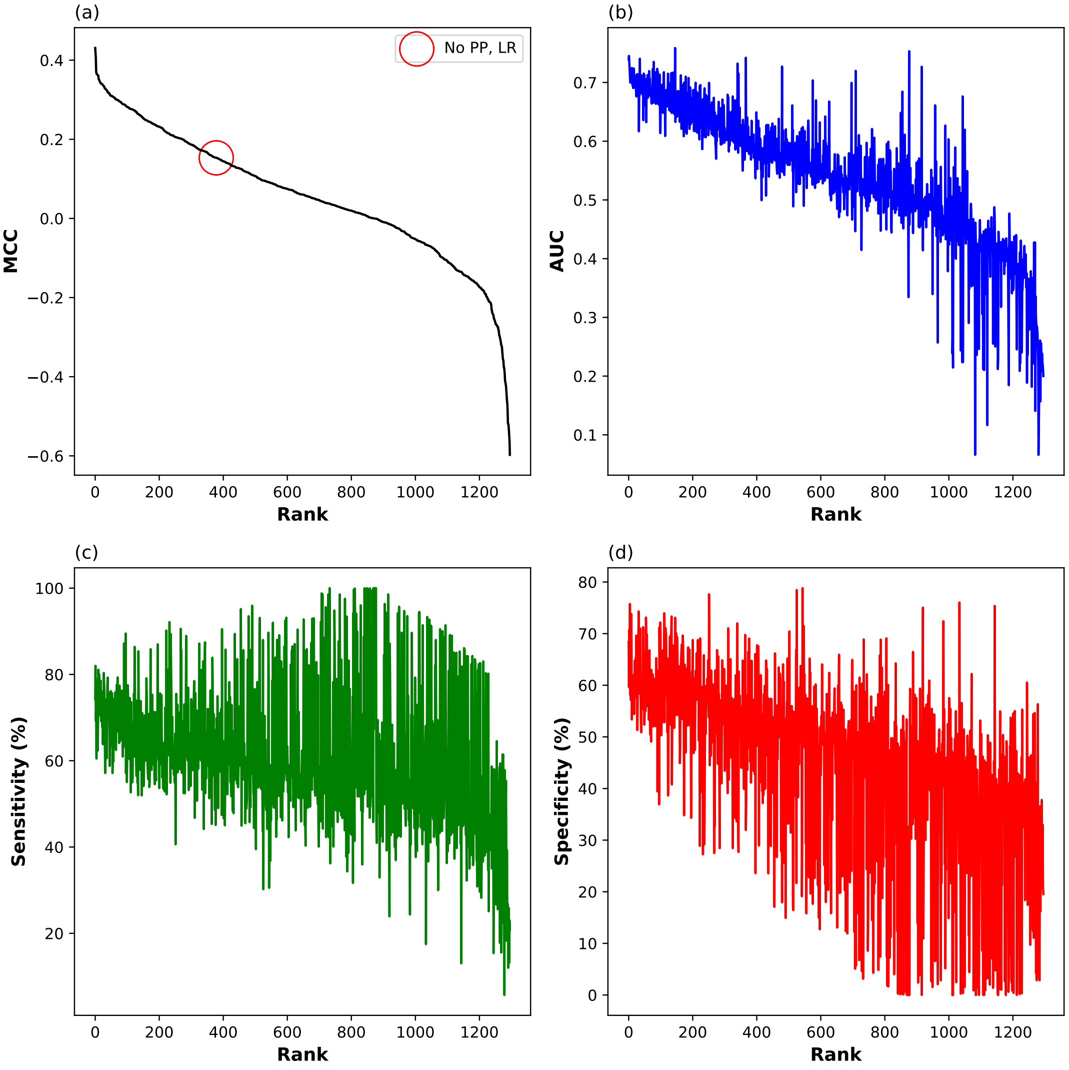


FIGURE 6.13: Mean MCC (a), AUC (b), sensitivity (c) and specificity (d) for pipelines (sorted in order of decreasing mean MCC). High grade patients.

in scores relative to the low grade model, implying that with this cohort transformation prediction is generally easier in patients presenting less severe OED. Table 6.3 summarises the methods associated with each of the top 5 best pipelines.

The methods selected by PipeOpt show noticable differences to those selected by the low grade model. Firstly, the best pipeline in the set does not apply a Mie correction to the data as an initial step. On the other hand, the four remaining pipelines do apply a Mie correction, which implies that a scatter correction may be important for models which apply a logistic regression classifier to the data, but not LDA. The ROC curves from the best pipeline for each pairing of patients is shown in Fig. 6.14. Again,

TABLE 6.3: Top ranking pipelines for high grade dysplasia patients model.

Mie	Smooth	Baseline	Norm	Scaling	FE	Classifier
No	SG	SG diff	Vector	N/A	PCA	LDA
Yes	PCA	SG diff	Vector	N/A	PCA	LR
Yes	N/A	RB	Vector	N/A	PCA	LR
Yes	N/A	RB	N/A	N/A	PCA	LR
Yes	N/A	RB	Amide I	N/A	N/A	LR

the individual ROC curves indicate that there is a lot of variation in the results when holding out different sets of patients. The shape and position of the median curve does indicate that the majority of trained classifiers can accurately predict the outcome of spectra. 12/70 (17%) of the curves have an AUC of less than 0.5, indicating models which have no skill in predicting the identity of data within the test set.

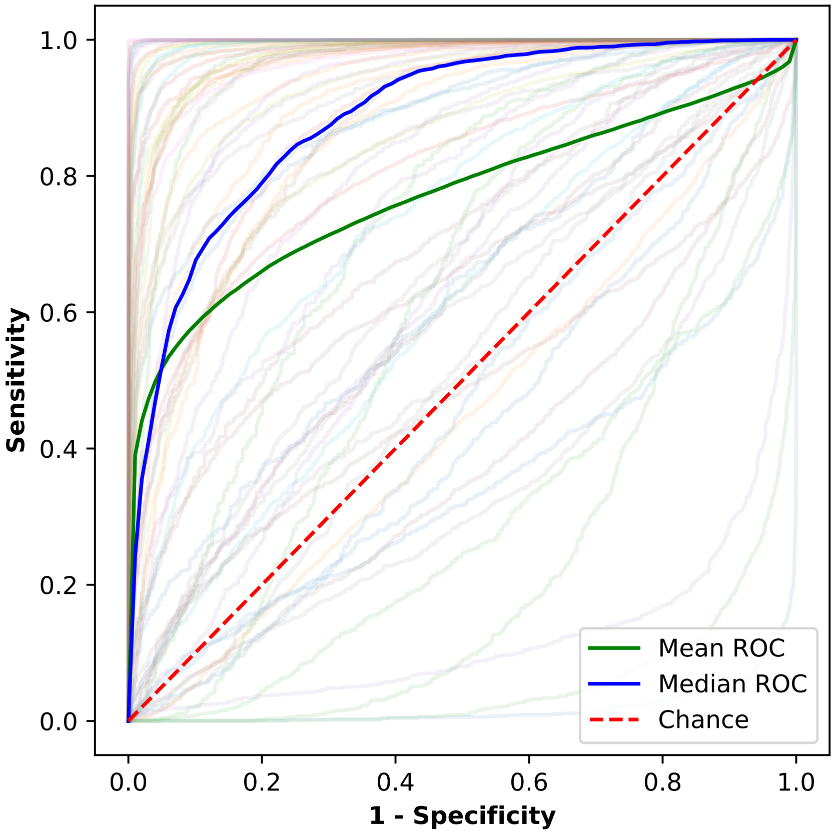


FIGURE 6.14: Mean and median ROC curves over the 70 individual splits (shown in paler colours) for high grade patients.

Figure 6.15 shows the fraction of instances a patient from either group was incorrectly labelled as the opposite outcome. Half of the transforming patients who were correctly labelled 100% of the time, whilst only two of the patients were incorrectly labelled over half of the time. This implies that the technique has some potential utility in the transformation prediction of high risk lesions. Furthermore, all patients who did not transform were correctly labelled as such over 50% of the time, emphasising that this technique not only identifies patients with a propensity to transform, but can also predict a negative outcome the majority of the time.

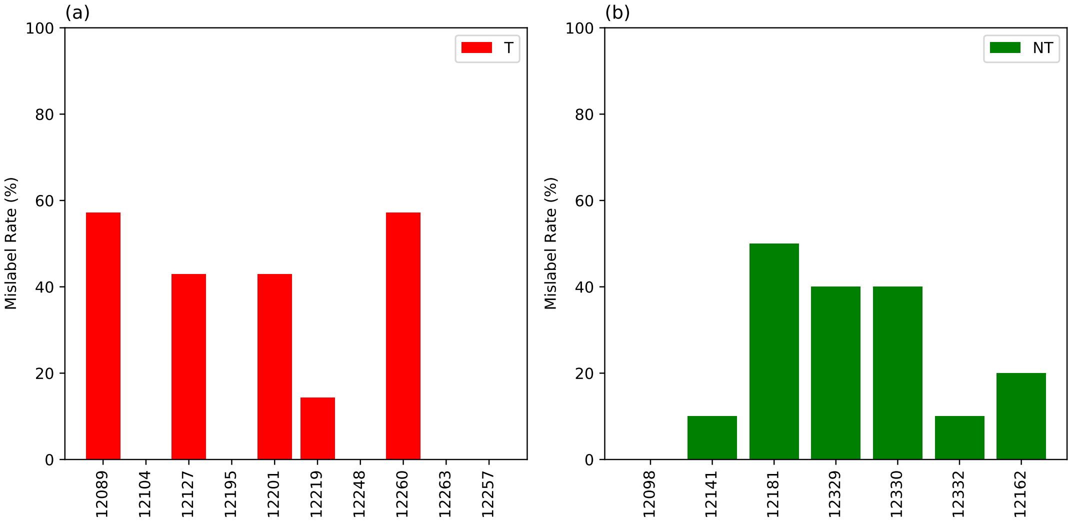


FIGURE 6.15: Mislabelling rates for each transforming (a) and non-transforming (b) patient.

The overall results for both hard and soft voting strategies for high grade patients is shown in Fig. 6.16. Out of the 70 models, transformers were correctly labelled as such 55 times (79%), whilst non-transformers were correctly labelled 53 times (77%), demonstrating that this model can predict OED outcome with moderately good accuracy.

6.4 Discussion

The study described in this chapter aims to provide a contribution to the clinical dilemma of accurately predicting the risk of OED lesions, which currently relies on histopathological grading as the primary indicator. The results presented have shown that there is potential predictive value attributed to vibrational spectroscopic data of lesions that are processed and analysed in an appropriate manner. Caution has been

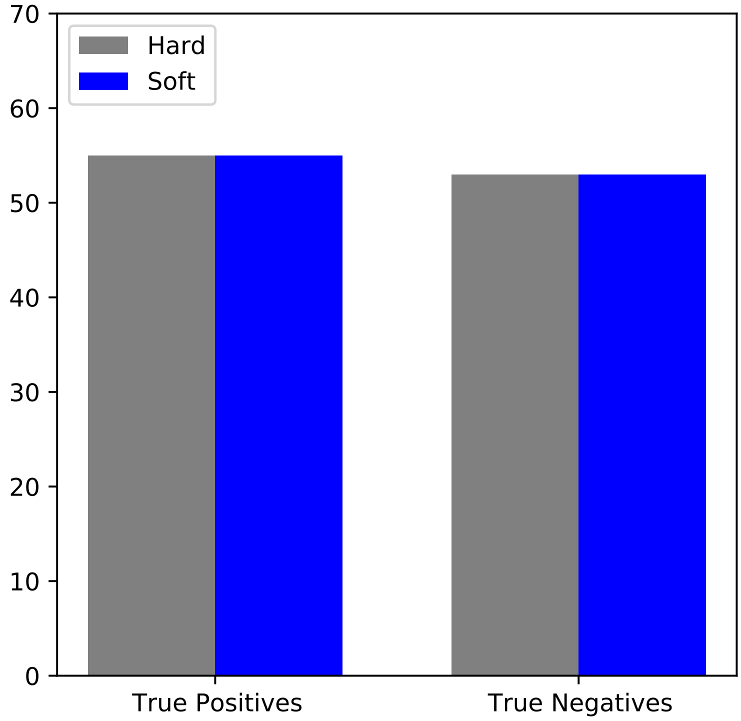


FIGURE 6.16: Number of times high-grade patients patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies.

taken to emphasise that the significance of these findings is significantly hampered by the fact that the patient cohort is far below the levels required to be recognised as a feasible alternative methodology. However, it's the author's view that the results presented here are noteworthy, as they provide scope and focus for future related studies.

6.4.1 All Patients Model

The first model, described in section 6.3.1, draws stark attention to the difficulty in using the proposed methodology to predict the outcome of any grade of dysplasia. From Fig. 6.6 it is clear that the vast majority of pipelines trialled in `PipeOpt` have no skill, indicated by a classifier with an MCC of ≤ 0, or an AUC of ≤ 0.5. It also appears that applying no pre-processing to the data has a negative impact on results. The optimum pipelines, summarised in table 6.1, indicate that random forest (RF) perform better than LDA and LR for this particular dataset. Many of the ROC curves shown in Fig. 6.7 show sub-optimal performance, with the mean curve re-iterating this interpretation. The unfamiliar shape of each ROC curve, characterised by sharp edges rather than a smooth curve, is a result of both the non-linear decision boundary formed by RF classifiers and the heterogeneous nature of the train and test data.

The sensitivity and specificity across every patient iteration, shown in Fig. 6.8, shows that only poor-moderate skill can be expected when using this approach to predict malignant transformation. The sensitivity, which is not much larger than 50%, suggests that this model will not be able to confidently rule out transformation if the patient has been labelled as a 'non-transformer'. This immediately raises concern with respect to the credibility of using this test as an adjunctive test. If a patient's outcome is predicted as no malignant transformation, but ultimately does develop OSCC $\approx$ 50% of the time, then the test is unsuitable for clinical adoption and further considerations should be taken to either expand the dataset or use a different approach. On the other hand, the overall specificity ($\approx$ 75%) does show a marked improvement relative to sensitivity. This means that patients who are predicted to develop OSCC by the model are incorrectly predicted 1/4 of the time. This confidence in disease prediction has economic benefit, as costs will be correctly allocated to follow up and treat the correct cases. Furthermore, higher specificities will reduce the detrimental impact false negatives have on patient wellbeing and mental health, due to the unnecessary anxiety induced by positive result.

Despite the ethical and ecomonic importance of a highly specific test, there is an abundance of tests with poor specificity that are already commonplace within the clinic, especially in early diagnostics. For instance, 50-61% of women who undergo annual mammography to screen for breast cancer can expect to have a false positive result [172], [173]. Here, a false positive result is one which flagged a positive result but had no positive gold-standard (histopathological) diagnosis after one year. False positives in this context may lead to biopsy, which is an expensive and invasive procedure, causing unnecessary discomfort and anxiety. Prostate-specific antigen (PSA) tests, which are screening tools for prostate cancer, also tend have a high false positive rate [174].

The threshold for a patient to be predicted as transforming can be altered to vary the specificity or sensitivity of the test, based upon its desired outcome. Screening tests, such as mammography and PSA, are weighted towards high sensitivity so that a negative decision can be made with the degree of certainty required. Not following up a positive patient may result in devastating human cost, such as severe injury or death. Figure 6.17 shows the effect of varying the threshold in this cohort of patients. For hard voting, the threshold is the fraction of T predictions to the total number of predictions, whilst for soft voting it's the sum of $p(T)$ across all folds. By default, and as shown in Fig. 6.8, these are both set to 0.5. Decreasing the threshold should increase the sensitivity, as the condition for being labelled as the positive class (transforming) is artificially relaxed to include more patients. The opposite is true when the threshold is increased, as the condition becomes a lot more stringent. This is reflected in Fig. 6.17, where a threshold of 25% (a) leads to an increase in sensitivity and decrease in specificity for both voting regimes, whilst increasing the threshold to 75% in (c) leads to an increase in specificity and decrease in sensitivity.

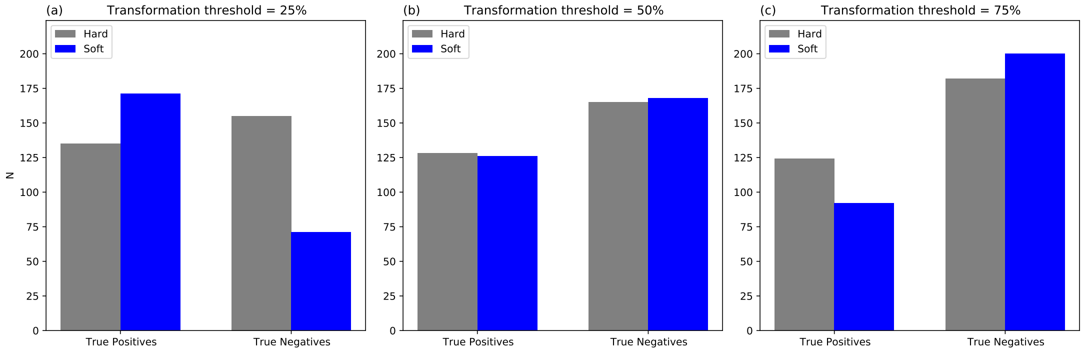


FIGURE 6.17: Number of times patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies, shown for different thresholds.

Despite the enhanced sensitivity shown in (a), there is a marked drop in specificity, which for soft voting decreases the score from 168/224 to 71/224 correctly predicted healthy patients. Operating at this threshold, the test would capture a lot more positive cases, but with a specificity less than 50%, the majority of patients with good outcomes will be unnecessarily followed up.

An interesting feature of Fig. 6.17 is that there is little change in the scores at different thresholds when using hard voting as the ensemble mode. Soft voting, on the other hand, yields the expected pattern (reducing the threshold reduces the number of true negatives, whilst increasing the number of true positives, and vice-versa for increasing the threshold). The nuances of how each protocol works can explain this. Hard voting will treat every classification with the same weight, regardless of how 'confident' the classification is. This can lead to cases where barely confident wrong spectrum-wise classifications can outweigh highly confident spectrum-wise classifications, which will in turn lead to an overall incorrect lesion classification. By implementing a soft-voting regime, the relative certainty of a classification is accounted for, which will provide higher weight to the spectra with high confidence.

6.4.2 Low Grade Patients Model

Stratifying the patients according to the severity of OED has a positive impact for both low and high grade patients. For low grade patients, the mean sensitivity over the 42 independent patient splits is $\approx$ 81%, a significant increase from the 56% observed using the entire cohort. The specificity also increases from 75% to 79%, showing a more incremental increase in the models' ability to correctly identify a patient who will not transform. Figure 6.9 shows that pipelines can achieve an AUC > 0.8 or MCC > 0.5, depending on the sequence of pre-processing methods applied to the data and classifier used. The top 5 pipelines, shown in table 6.2, apply similar methods to the data, suggesting that optimal performance can only be achieved with a small range of pipelines. All pipelines fist apply a Mie correction to the raw data, differentiate the spectra, and use a linear technique (LR or LDA) to discriminate between the two outcomes. Since the pipeline is optimised for each of the patient splits, each trained model has a unique optimised hyperparameter vector. Figure 6.18 contains histograms for the optimal value for each of the hyperparameters in the best pipeline from table 6.2.

There appears to be an almost unanimous decision on each of the HPs, except for the smoothing window, which influences to what extent each spectrum is smoothed. A high regularisation strength of 100 would suggest that the logistic regression model performs better on hold-out data when the model is better regularised, which reduces overfitting to the training data. This is a logical finding, as the small training set and patient variability means that that overfitting may lead to features unique to certain

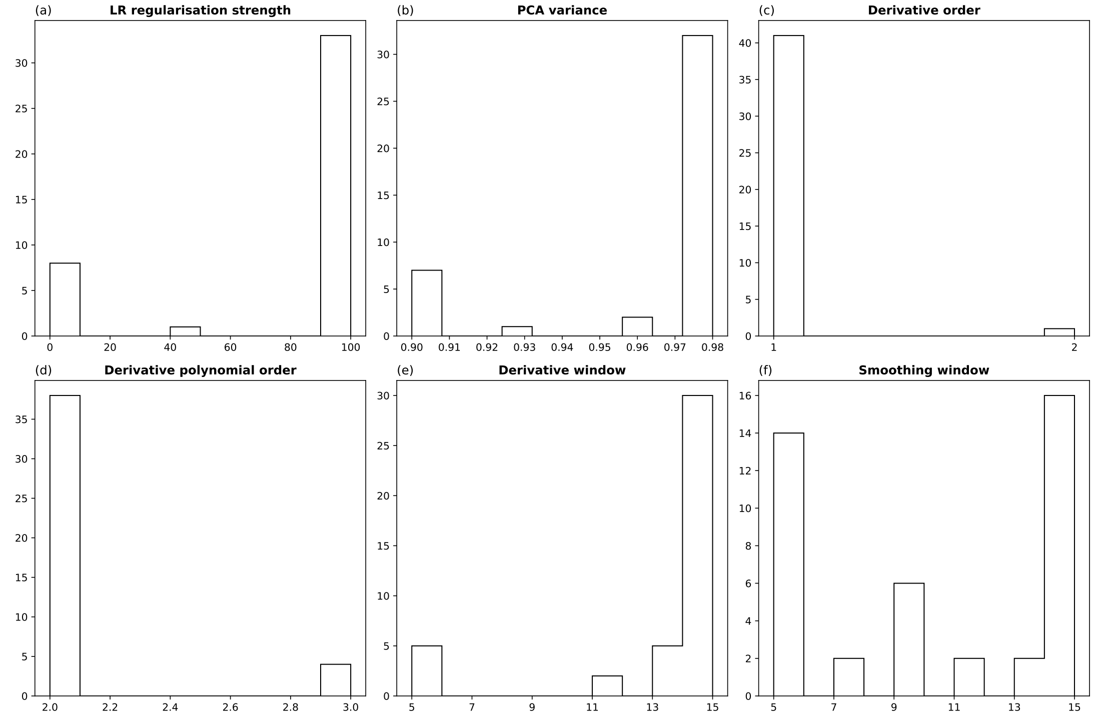


FIGURE 6.18: Number of times high-grade patients patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies, shown for different thresholds.

patients, therefore do not generalise well when tested on hold-out data from new patients. The choice of smoothing window length - an odd number where higher numbers indicate a higher level of smoothing - seems to be more evenly distributed. Intriguingly, the window lengths with the most occurrences are 5 and 15, implying that some models favour less smoothing, whilst others favour more, which again may reflect the inhomogeneity of the sample set, with patient images containing spectra with different signal-to-noise ratio. It appears that this step accounts for much of the sample inhomogeneity, manifest in the much sharper distributions seen for the other HPs. The spectra are differentiated to first order in 41/42 of the models, strongly indicating that this is the best approach for this particular dataset.

There is a notable difference between the median and mean ROC curves for each of the patient splits using the top pipeline (Fig. 6.14), which suggests that the distribution of ROC curves is skewed towards the top left corner of the plot, which represent skilful models. The position of the median curve also indicates that $>$ 50% of the independent models have very good skill, whereas the performance is much more varied for the lower half of models. Figure 6.11 suggests that the poor performance associated with some models can be attributed to the presence of a particular patient in the test set. For instance, patient 12092 from the transformation group was never incorrectly predicted as non-transforming, whilst three other transforming patients were incorrectly predicted in ≈ 30% of the models in which they formed part of the test set. On the other hand, one non-transforming patient (12182) was incorrectly predicted as transforming in the majority of models, with three patients enjoying 100% success rates.

To further investigate potential sources of patient mislabelling, the spatial arrangement of spectral predictions for a sample can be shown. An image from each outcome group were selected for poorest performance (T: 12194; NT: 12182) as a demonstration. Each non-transforming patient in the cohort appears in 10 models (paired with 6 transformers), whilst each transformer is paired with 7 non-transformers.

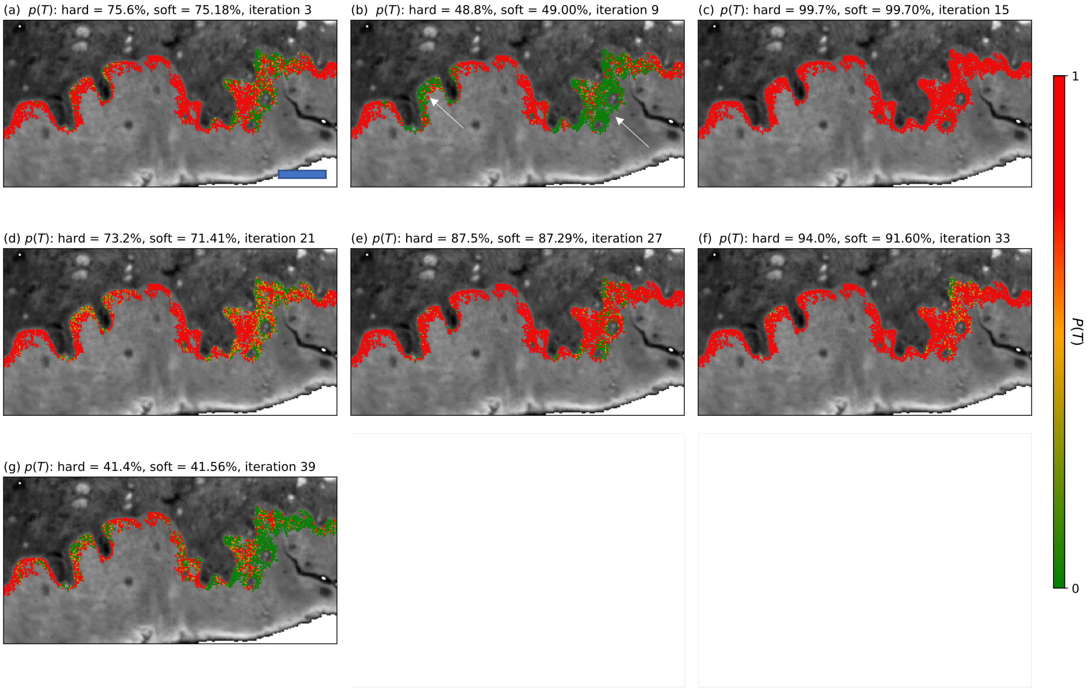

(a) p(T): hard = 75.6%, soft = 75.18%, iteration 3
(b) p(T): hard = 48.8%, soft = 49.00%, iteration 9
(c) p(T): hard = 99.7%, soft = 99.70%, iteration 15
(d) p(T): hard = 73.2%, soft = 71.41%, iteration 21
(e) p(T): hard = 87.5%, soft = 87.29%, iteration 27
(f) p(T): hard = 94.0%, soft = 91.60%, iteration 33
(g) p(T): hard = 41.4%, soft = 41.56%, iteration 39
1
P(T)
0

FIGURE 6.19: The spatial arrangement of test spectra super-imposed on the absorbance map at 1650 cm^{-1}, for each of the models where 12194 (T) appears in test set, of which there are 7 (a-g). The color encodes the transformation probability for each spectrum. Hard and soft voting probabilities are also shown to inform the patient outcome decision. Scale bar in (a) = 200 µm.

The sample shown in Fig. 6.19 was incorrectly predicted as non-transforming twice (b, g) out of the seven iterations it appeared as part of test set. For the two iterations in which the lesion was predicted incorrectly, the associated probabilities of transformation $p(T)$ were borderline relative to the 50% threshold, whereas iterations where the lesion was predicted as transforming had a much higher confidence. This is supported by the presence of a well defined region in each image where there may exist spectra being labelled as non-transforming. This is marked clearly in the corresponding histology image shown in Fig. 6.19, where regions with a low $p(T)$ are marked with a white arrow.

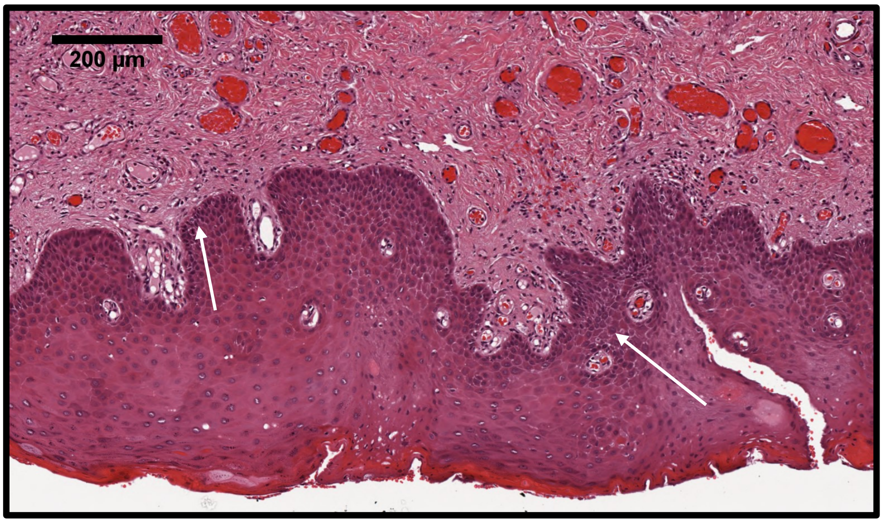


FIGURE 6.20: H&E stained sample of ROI shown in Fig. 6.19. White arrows are pointing the basal layer of the epithelium, which contains regions found confusing by the model.

The arrow on the right of Fig. 6.20 points to a region which appears to present subtle changes in morphology with respect to the rest of the basal layer, especially in terms of the shape and density of nuclei. Since FTIR spectra reflect the relative concentrations of various moieties (such as DNA, glycogen, proteins), it would not be surprising if these changes are manifesting as subtle shifts in absorption at certain wavenumbers, which would in turn adjust the probability of transformation depending on the model parameters for that particular iteration.

Various factors may be contributing to the apparent mislabelling in some of the models. Given that the number of patients within the training set is small, each patient has a significant influence on the respective model, so it is not surprising that the scores are so dependent on the patient train-test split.

A possible explanation for the specific clusters of high transformation potential in some of the images is that dysplastic cells in these regions of the epithelium may genuinely be undergoing different biochemical alterations, resulting in spectral changes, than cells located in the low transformation probability areas. Genetic and epigenetic mutations may not be uniform throughout the macroscopic vicinity of the lesion. In future studies, imposing stricter inclusion criteria on the size and features of the labelled area may lead to a lower variance in results.

Figure 6.21 shows the spatial arrangement of predictions for the lowest scorer in the NT outcome group (12182) which, as shown in Fig. 6.11, has a mislabelling rate greater than 60%. Apart from the model predictions shown in Fig. 6.21a, the mislabellings consistently occur in the same region of the image. This can be compared with the corresponding H&E image (Fig. 6.22).

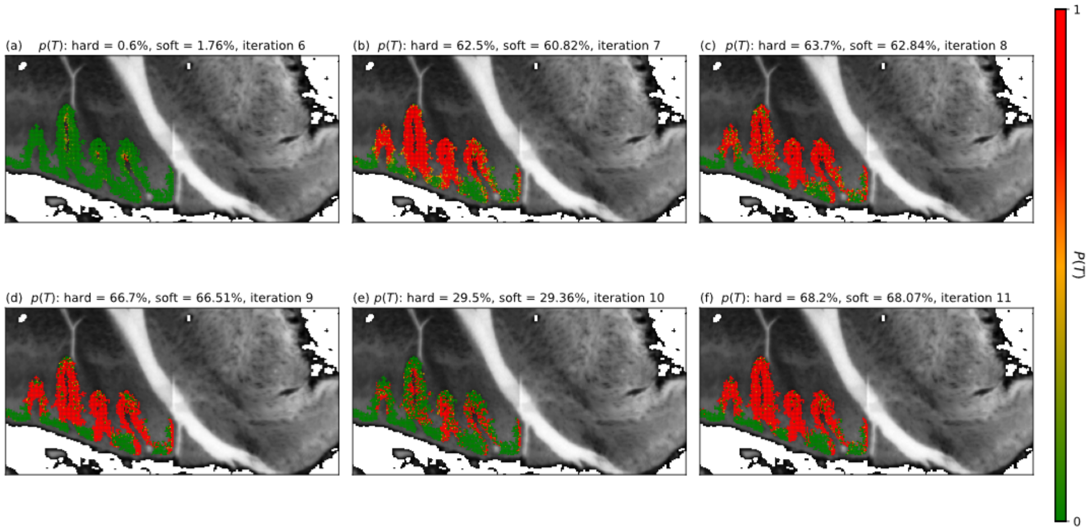

(a) $p(T)$: hard = 0.6%, soft = 1.76%, iteration 6
(b) $p(T)$: hard = 62.5%, soft = 60.82%, iteration 7
(c) $p(T)$: hard = 63.7%, soft = 62.84%, iteration 8
(d) $p(T)$: hard = 66.7%, soft = 66.51%, iteration 9
(e) $p(T)$: hard = 29.5%, soft = 29.36%, iteration 10
(f) $p(T)$: hard = 68.2%, soft = 68.07%, iteration 11
1
$P(T)$
0

FIGURE 6.21: The spatial arrangement of test spectra super-imposed on the absorbance map at 1650 cm^{-1}, for each of the models where 12182 (T) appears in test set, of which there are 6 (a-f). The color encodes the transformation probability for each spectrum. Hard and soft voting probabilities are also shown to inform the patient outcome decision.

Inspection of the H&E image shown in Fig. 6.22 does reveal that sub-optimal image registration may be the source of the relatively poor scores for this particular lesion.

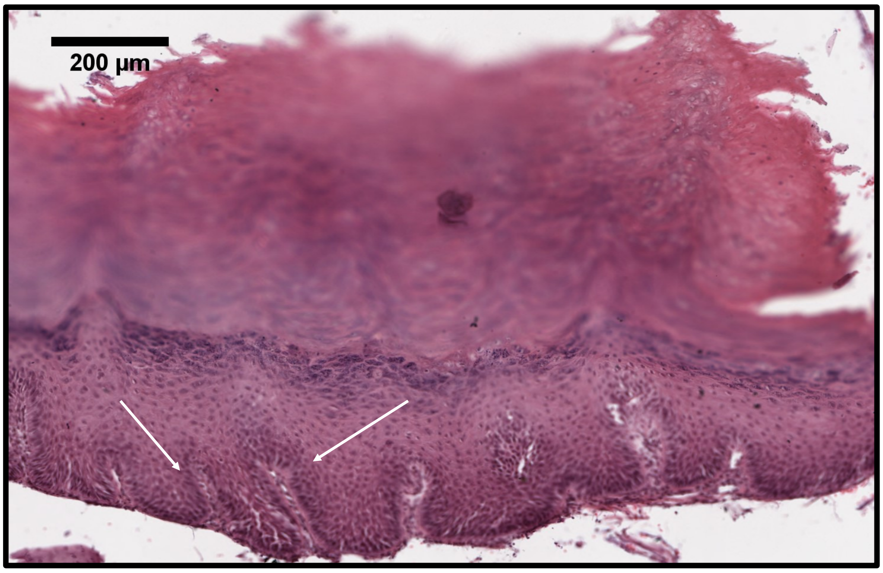


FIGURE 6.22: H&E stained sample of ROI shown in Fig. 6.21. The white arrows are pointing toward the rete ridges of the basal layer, which contains regions which the model finds confusing.

There are artefacts in the FTIR images shown in Fig. 6.21 which are not present in Fig. 6.22, making it difficult to precisely compare the two images. However, the mislabelling region appears to roughly coincide with the edges of the rete ridge shaped features (marked by white arrows). Morphologically, the cells that line the edge of the rete ridges appear to be similar with those that form the main epithelial-stroma interface, however the edges may be somewhat ill-defined as a result of the spatial resolution. Specific labelling is more difficult when there is a lack of a sharp interface between dissimilar substrates (such as epithelia to stroma, or tissue to blank slide).

The threshold used to predict malignancy in a lesion can be varied as in Fig. 6.17 to artificially decrease or increase the sensitivity and specificity. Given that lower grade lesions will not be as closely followed up as those with higher grades, there should be more emphasis placed on ruling out low grade lesions with a low propensity to transform, so that these patients can either be excluded from follow-up, or assigned a lower priority and managed as such. This will not only improve patient experience but streamline the clinical workflow, which has the potential to impact positively both economically and indirectly to other aspects of healthcare. As such, the transformation potential should be set low (25% in Fig. 6.23).

Alternatively, a ‘rule in’ test may be instead desired, where low-grade lesions with a high propensity to transform are risk stratified into the high risk groups, which often leads to excision/ablation of the lesion [175]. Rule in tests require a high specificity, consequently the threshold for transformation prediction should be set high. In this case, a threshold of 75% would lead to only two false positives.

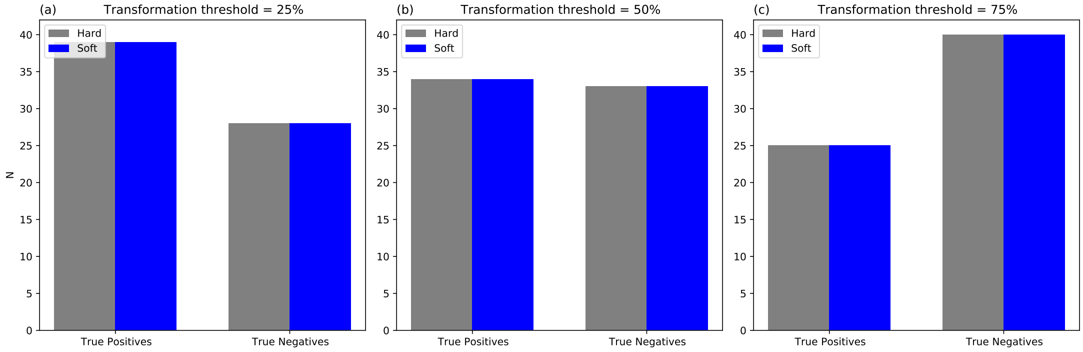


FIGURE 6.23: Number of times low grade patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies, shown for different thresholds.

6.4.3 High Grade Patients Model

Similar to the low grade patients, prediction of outcome for high grade patients has marked increase of sensitivity (57% to 79%) and marginal increase in specificity (75 to 77%) when compared with the baseline model incorporating all the samples. Figure 6.13 shows that similarly high scores to the low grade model can be achieved by carefully choosing an optimal sequence of pre-processing and classification steps to apply to the dataset. Table 6.3 goes further to highlight that an objective optimisation tailored to a particular dataset is important, given that the pipelines with the highest MCC score are different to the models described in sections 6.3.1 and 6.3.2.

Of particular note is the fact that, in contrast to the low grade model, each of the top five pipelines summarised in table 6.3 applies normalisation rather than scaling. This represents a return to the widespread consensus that normalisation of FTIR spectra is necessary to mitigate for those unwanted physical distortions to the absorbance of each spectrum. The position of the 'baseline' pipeline (marked with a red circle in Fig. 6.13a) indicates that the marked drop-off in both MCC and AUC results from pipelines with steps that are not compatible with each other, inferred from the fact that no pre-processing has less of a detrimental impact on performance. Table 6.4 shows the 5 pipelines with the poorest MCC in the high grade model.

TABLE 6.4: Worst pipelines for high grade dysplasia patients model.

Mie	Smooth	Baseline	Norm	Scaling	FE	Classifier
Yes	PCA	SG diff	-	Standard	PCA	LR
Yes	PCA	RB	-	Standard	PCA	LR
Yes	PCA	RB	-	Min-Max	PCA	LR
No	PCA	RB	-	Standard	-	LR
Yes	PCA	SG diff	-	Standard	-	LR

Table 6.4 reinforces that normalisation is crucial to this subset of the data, given that the five worst performers all bypass the normalisation step. A PCA denoising step is also present in each of the worst pipelines, implying that usage of this method is not suitable, especially when normalisation is bypassed. Interestingly, there are striking simililarities between the worst performing high grade pipelines shown here and the best low grade pipelines shown in table 6.2, which further reiterates the importance of optimisation for a given dataset. The significance of normalisation in the high grade subset may be a result of the greater number of patients (n=17), as well as the wider histopathological variance across the dataset, with three dysplasia severity groups (G3, G4, G5) as opposed to the two in the low grade model (G1, G2).

Visualisation of predicted transformation potential on different samples can be attained by plotting the pixel-wise scores and comparing with the corresponding H&E image. Figure 6.24 shows the spatial arrangement of predictions for a low scorer in the NT outcome group (12329), which has a mislabelling rate of $\approx$ 40%. Figure 6.25 is the corresponding H&E image of the depicted region.

Figure 6.24 does not appear to indicate any 'borderline' regions that lead to the outcome prediction changing for the lesion, unlike what was seen in Fig. 6.19. With the exception of perhaps Fig. 6.24d, the probability of transformation is rather uniform across each lesion, which implies that, for this lesion, prediction of transformation is difficult. This may be due to the fact that severity of dysplasia may be one of the factors that each model intrinsically accounts for, and given that the lesion shown in

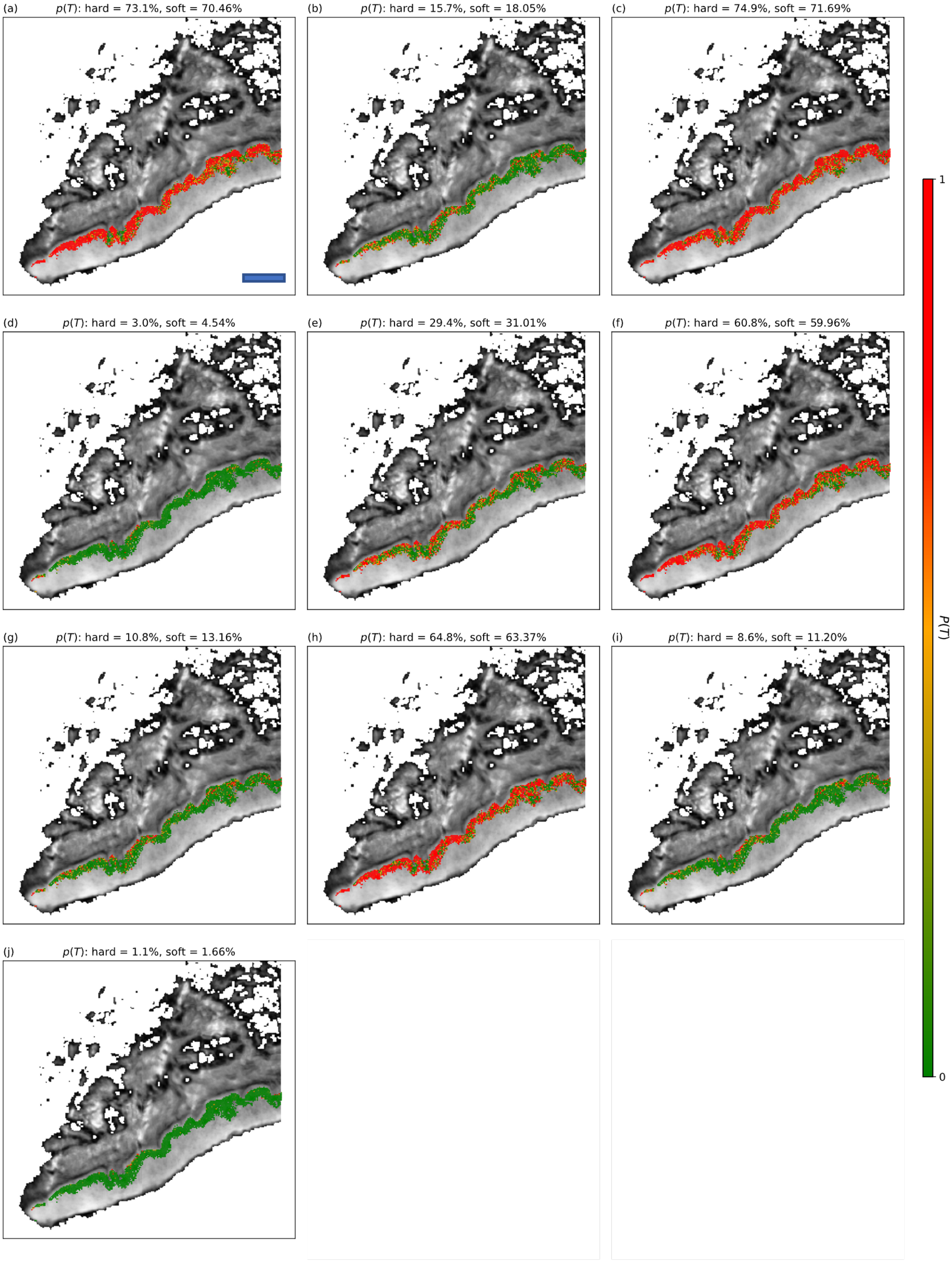

(a) p(T): hard = 73.1%, soft = 70.46%
(b) p(T): hard = 15.7%, soft = 18.05%
(c) p(T): hard = 74.9%, soft = 71.69%
(d) p(T): hard = 3.0%, soft = 4.54%
(e) p(T): hard = 29.4%, soft = 31.01%
(f) p(T): hard = 60.8%, soft = 59.96%
(g) p(T): hard = 10.8%, soft = 13.16%
(h) p(T): hard = 64.8%, soft = 63.37%
(i) p(T): hard = 8.6%, soft = 11.20%
(j) p(T): hard = 1.1%, soft = 1.66%
1
P(T)
0

FIGURE 6.24: The spatial arrangement of test spectra super-imposed on the absorbance map at 1650 cm^{-1}, for each of the models where 12329 (NT) appears in test set, of which there are 10 (a-j). The color encodes the transformation probability for each spectrum. Hard and soft voting probabilities are also shown to inform the patient outcome decision. Scale bar in (a) = 200 µm.

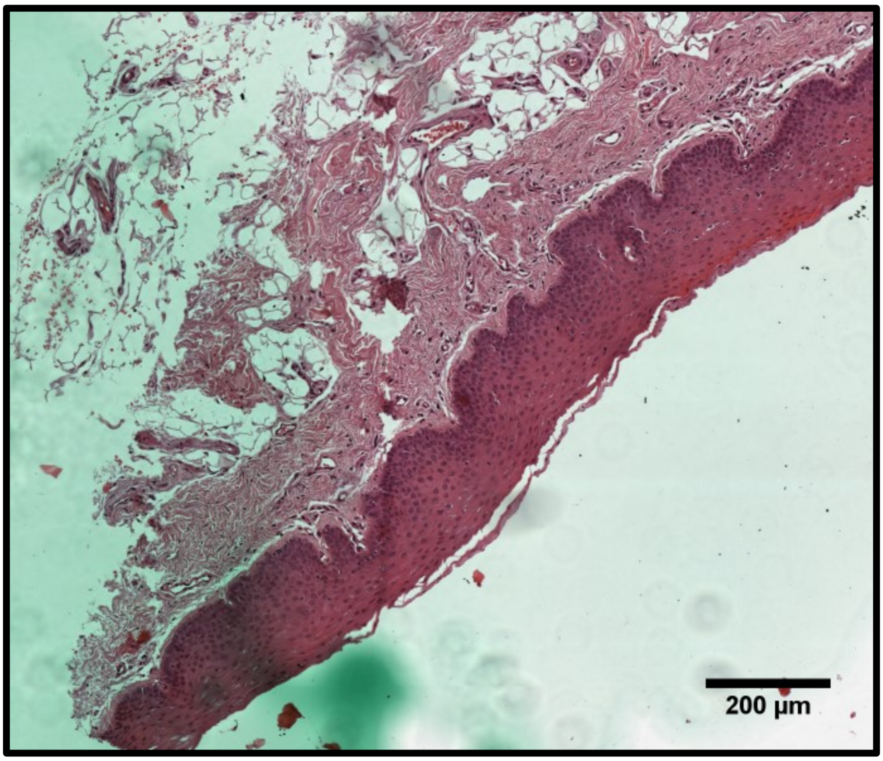


FIGURE 6.25: H&E stained sample of ROI shown in Fig. 6.24

Fig. 6.25 contains moderately dysplastic cells, this may lead to the apparent ambiguity in transformation potential.

A similarly poor scoring lesion in the T outcome group is 12127, which also has a mislabelling rate of $\approx$ 40%. Pixel-wise probabilities are shown below (Fig. 6.26) for visualisation. A corresponding photograph of the H&E stained sample is shown in Fig. 6.27.

For cases where the lesion has been incorrectly predicted as non-transforming, there appears to be distinct regions close to the basal layer which are dominated by spectra with low transformation probability. It may be hypothesised that these particular areas do not contain dysplastic cells which harbour a high transformation potential, especially given the heterogeneous nature of histology specimens.

The malignancy threshold can be varied (as in Figs. 6.17 and 6.23) in order to alter the sensitivity and specificity of the test. Considering the current management strategy of moderate and severe OED is excision/ablation [175], a more optimised protocol which was able to precisely identify negative cases is desirable. This means that the number

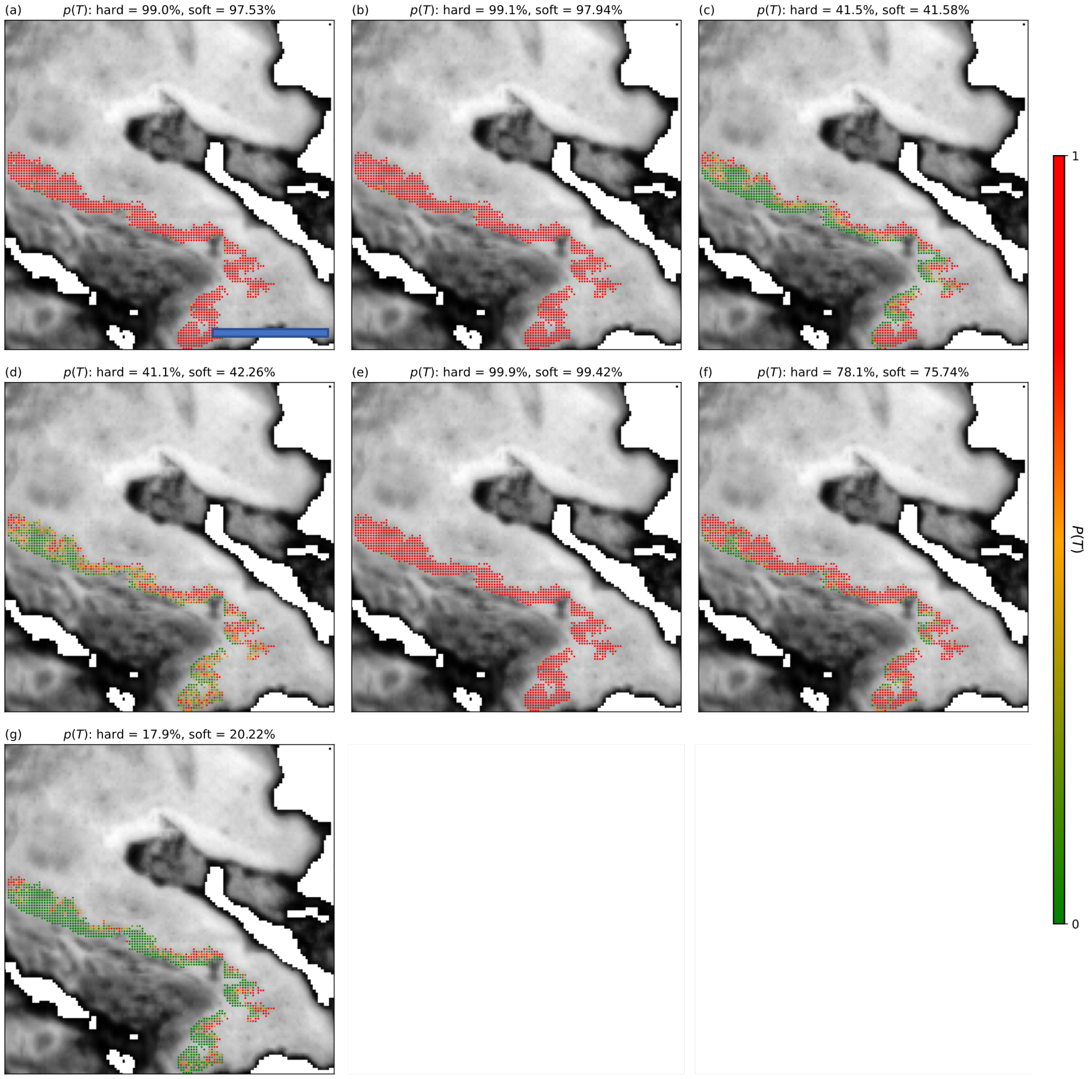

(a) p(T): hard = 99.0%, soft = 97.53%
(b) p(T): hard = 99.1%, soft = 97.94%
(c) p(T): hard = 41.5%, soft = 41.58%
(d) p(T): hard = 41.1%, soft = 42.26%
(e) p(T): hard = 99.9%, soft = 99.42%
(f) p(T): hard = 78.1%, soft = 75.74%
(g) p(T): hard = 17.9%, soft = 20.22%
1
P(T)
0

FIGURE 6.26: The spatial arrangement of test spectra super-imposed on the absorbance map at 1650 cm^1, for each of the models where 12127 (T) appears in test set, of which there are 7 (a-g). The colour encodes the transformation probability for each spectrum. Hard and soft voting probabilities are also shown to inform the patient outcome decision. Scale bar in (a) = 200 μm

of false negatives must be very low, requiring a test with high sensitivity. With such a test, patients who definitely do not require excision can be optimally stratified as such, rather than undergoing unnecessary and potentially harmful procedures such as surgical excision.

Setting the threshold low (25%) for this leads to a scenario where the number of times transforming lesions are predicted as non-transforming (false negatives) is low (2/70). This form of test would be ideal for this particular case, where it's envisaged that this

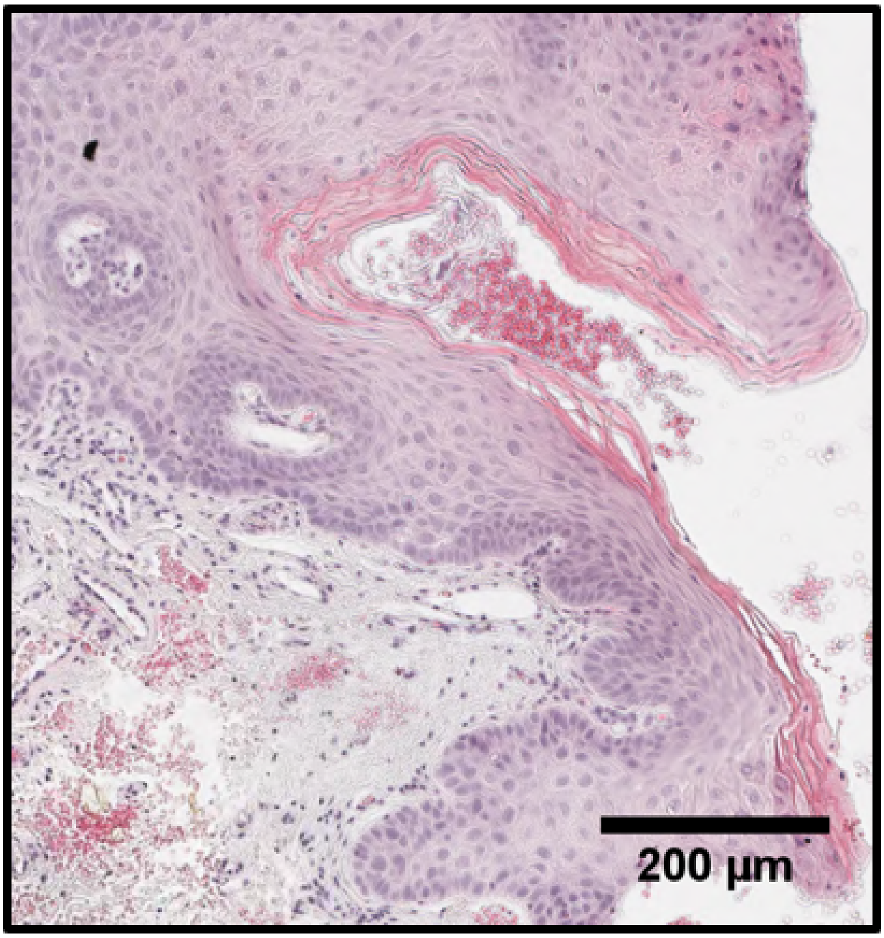


FIGURE 6.27: H&E stained sample of ROI shown in Fig. 6.26

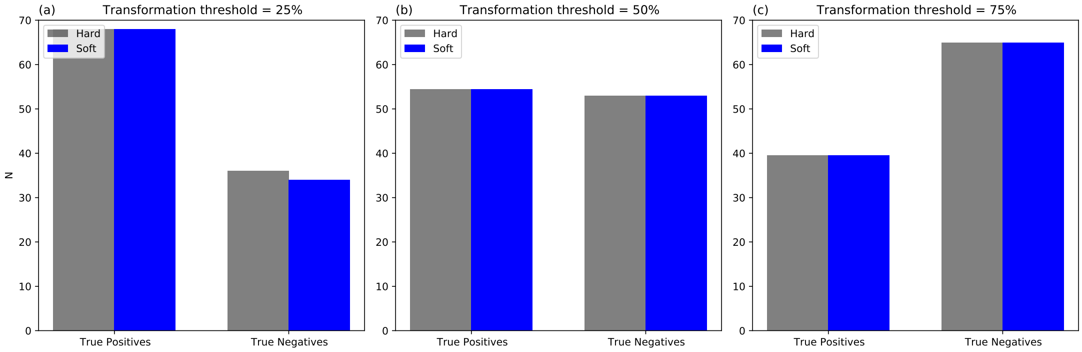


FIGURE 6.28: Number of times patients were correctly labelled as either transforming (true positive) or non-transforming (true negative) for both voting strategies, shown for different thresholds.

tool would be deployed as a 'rule out' test. Lesions which score less than the chosen threshold would be stratified into a lower risk group, which requires less rigorous and invasive follow-up. In addition to the obvious benefit this would bring on a patient level, the economic benefit of confidently and automatically ruling out patients presenting with a high grade lesion would be significant, given that the current diagnostic pipeline is time consuming, stress inducing and expensive. More considerations for the future landscape of machine learning in healthcare are discussed in chapter 7.

6.4.4 Spectral Biomarker Analysis

The well performing classifiers (LDA, logistic regression) deployed in this analysis share common characteristics. The main one is that they are statistical methods which seek to express a categorical variable (in this case, malignant transformation) in terms of a set of continuous variables (IR absorbance at each wavenumber). This encourages reasonable interpretation of the parameters emerging from fitted models, as they encode the importance each wavenumber contributes to a discriminating model, which in turn permits careful discussion of the spectral biomarkers associated with malignant transformation.

Trained logistic regression models have a coefficient vector and bias term which transforms each spectrum into a single number, which is subsequently used as the argument in a sigmoid function (Eq. (3.24)) to estimate class membership probabilities. As an example, Fig. 6.30a outlines the models decision boundaries for one of the patient splits in the low-grade analysis, for which the ROC curve is shown in Fig. 6.30b. The model is used to predict the identity of each test spectrum, which have been marked in Fig. 6.30a as either green or red dots, indicating the correct labelling of NT or T spectra respectively. Spectra which have been incorrectly labelled by the model are marked with a black cross.

The weight attributed to each wavenumber variable in the model can be probed as a measure of importance. However, there is a PCA step which precedes logistic regression (PCA-LR) in the top performing pipeline. The weights in the classifier model (W_{LR}) are therefore in the principal component domain, rather than the wavenumber domain, thus they should be transformed using the weight vector of the PCA model (W_{PCA}) using Eq. (6.4):

$$W_c = W_{PCA} \cdot W_{LR}. \tag{6.4}$$

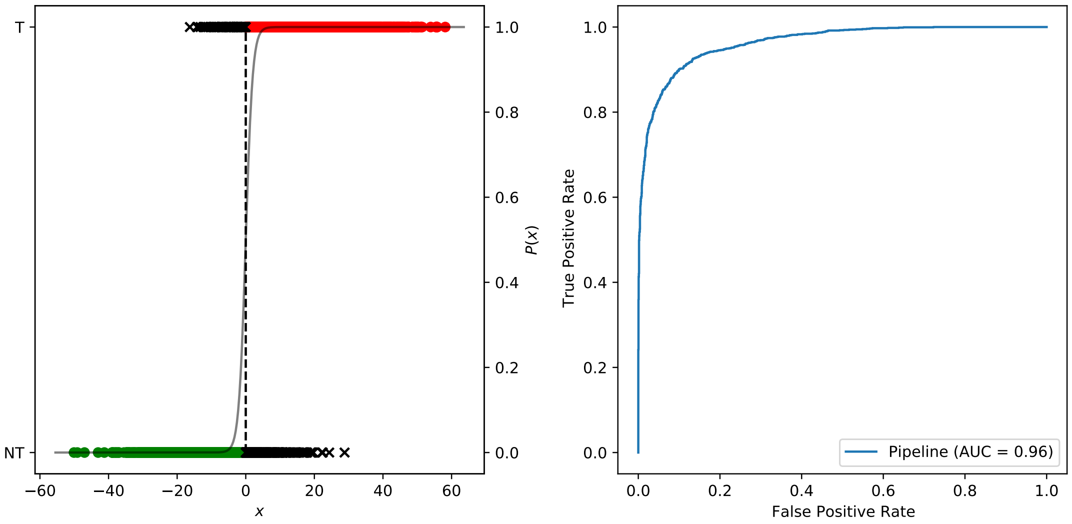


FIGURE 6.29: (a) Data points (spectra) transformed into predictions by the trained sigmoid function (see Eq. (3.24)) for one of the train-test splits. Those with a sigmoid output of > 0 are predicted as the positive class. (b) The ROC curve for the transformed points shown in (a).

FIGURE 6.30: (a) Logistic regression model for a single train-test split, for which the ROC curve is shown in (b)

The combined weight vector for the same model as shown previously (Fig. 6.30) is shown in Fig. 6.31. The magnitude of the weight reflects the importance of the variable in the model, where positive weights indicate that relative increases in the pre-processed signal lead to an increase in the sigmoid argument, which in turn raises the probability of the spectrum being labelled as transforming. The opposite is true for negative weights. However, since the input to the PCA-LR is a pre-processed dataset which has been smoothed, differentiated and standardised, the weights can not be directly attributed to relative changes in absorbance. Instead, the weight should be interpreted as the importance attributed to changes in gradient at points across the spectrum. In the instance shown in Fig. 6.31, the wavenumber with the highest weight is at about 1650 cm^{-1}, which is the centre of the amide I peak, with a weight of ≈ -2. This implies that relative decreases in the gradient at 1650 cm^{-1} leads to a higher probability of transformation.

Next, the important features for the best high-grade pipeline (table 6.3 can be investigated. Trained PCA-LDA classifiers have two sets of model parameters, in the form

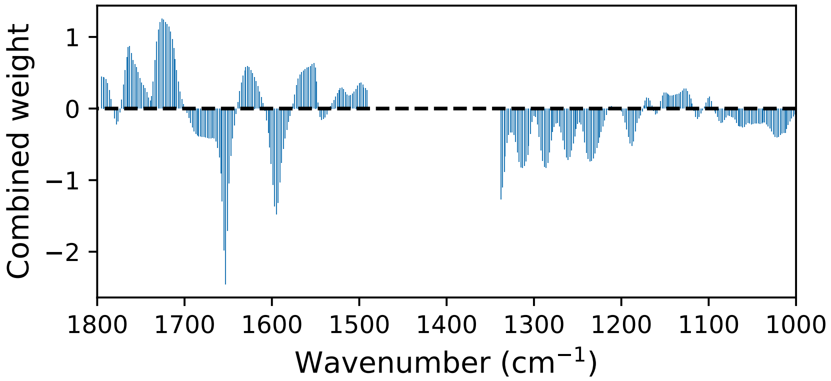


FIGURE 6.31: Combined weight plot for PCA-LR model for low-grade patients.

of linear transformation matrices, also known as weights (W), through which the preprocessed spectra are propagated. The linear discriminant vectors therefore exist in the domain of the principal components, therefore to extract feature importance the two matrices should be multiplied together (Eq. (6.5)) like for the PCA-LR classifier described in Eq. (6.4).

$$W_c = W_{PCA} \cdot W_{LDA} \tag{6.5}$$

Figure 6.32 shows both the decision boundary and the feature importance for an LDA model fit to a partition of the data.

Both Fig. 6.15 and Fig. 6.11 emphasise that prediction of malignant transformation in FTIR spectra from OED lesions is intrinsically complex and requires a pragmatic, multivariate approach in order to acquire reasonable scores. The weight plots for both grades are similar to some extent, with significant importance attributed to wavenumbers within both the amide I (1600 - 1700 cm^{-1}) and amide II (1500 - 1600 cm^{-1}) for both models. Less importance is attributed to lower wavenumbers in the low-grade model than the high-grade model, where there are prominent peaks at both 1030 cm^{-1} and 1252 cm^{-1}, which are characteristically absorbed by glycogen and nucleic acids respectively. Furthermore, more importance is attributed to the spectral region beyond amide I, ranging from 1700 - 1800 cm^{-1}, which can be attributed to the stretching vibrations of C=O bonds in lipids.

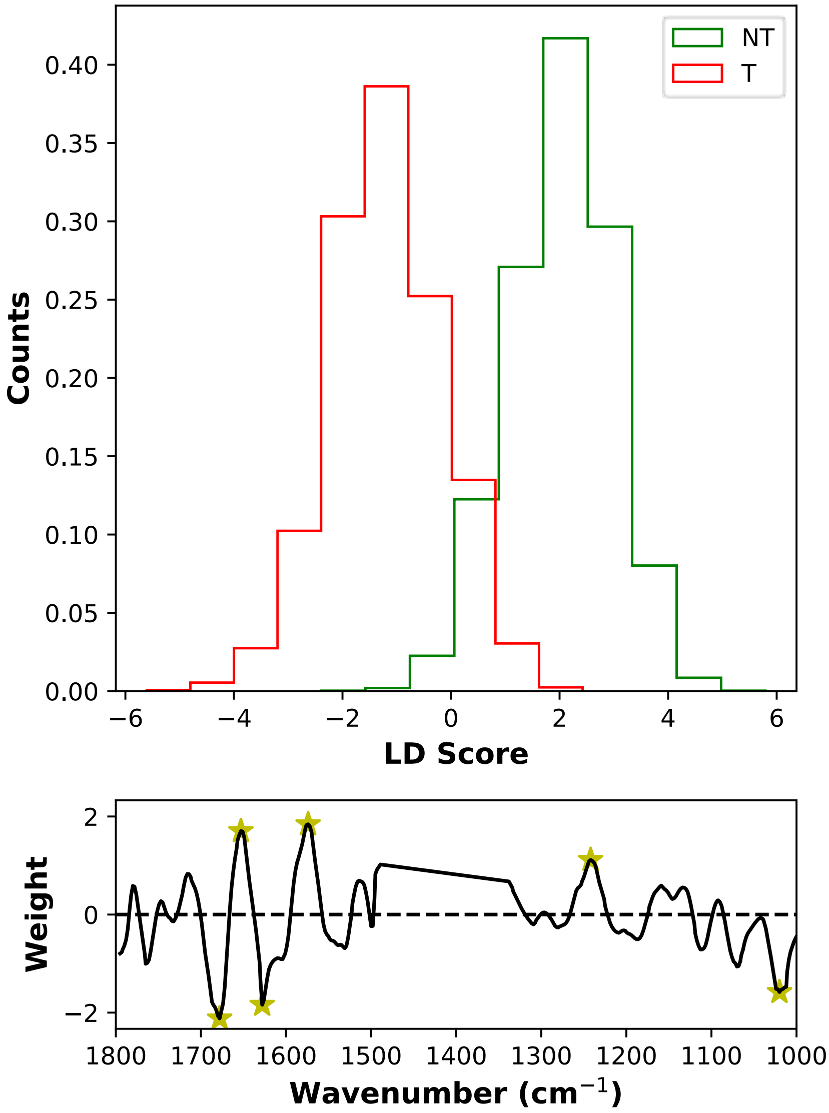


FIGURE 6.32: Decision boundary (top) and combined weight plot for PCA-LCA model (bottom) for high-grade patients. The stars delineate the wavenumbers with the highest weight amplitude, listed in table 6.5

The extent of pre-processing necessary to acquire reasonable results leads to a compromise in model interpretability, given that the classifiers are constructed on the basis of spectra which have been transformed to different domains that may no longer directly reflect the underlying biochemistry. For example, changes in intensities of a first derivative spectrum indicate that the gradient of an absorbance peak is different, rather than the absorbance itself. Furthermore, the convoluted nature of IR absorbance peaks leads to ambiguities in their origin. For instance, the amide I peak represents protein content, but there are multiple underlying absorbances resulting from from differently structured proteins which contribute towards the convolution [176]. It is therefore important that careful biochemical interpretation is made on a high level, rather than forensically examining the importance attributed to each potential spectral biomarker. The six peaks with the largest magnitude on the high grade feature plot (Fig. 6.32) have been marked with a star and tabulated in table 6.5.

TABLE 6.5: Important wavenumbers for outcome prediction

Rank	Wavenumber (cm^{-1})	Biomarker
1	1678	Amide I
2	1574	Amide II
3	1628	Amide I
4	1653	Amide I
5	1020	Glycogen
6	1242	Nucleic Acids

It is clear that a lot of discriminatory power is derived from wavenumbers in the amide regions of the spectrum, which indicates that protein content and expression plays an important role in malignant transformation. Protein phosphorylation is a process by which a molecule of ATP donates a phosphate group to an amino acid residue, catalysed by a protein kinase. Phosphorylation and dephosphorylation play an important role in the regulation of cell signalling pathways, acting as molecular switches, facilitating protein-protein interaction or translocating target proteins to different parts of the cell. Inappropriate activity of protein kinases may lead to the dysregulation of signalling pathways and malignant transformation may occur [177]. Overexpression of tyrosine phosphorylated epidermal-growth-factor-receptor (EGFR) is associated with an enhanced risk of malignant transformation in PPOELs [178]. Since protein phosphorylation manifests in structural changes, the subtle relative shifts in the amide regions may be the result of transforming lesions exhibiting increased expression of EGFR.

In advanced stages of OSCC transformation, epithelial cells migrate from the stratified epithelium and infiltrate surrounding tissue. Adhesion between normal cells is predominantly maintained by E-cadherin, a transmembrane glycoprotein. The abundance of E-cadherin can be regarded as a tumour-suppressor gene due to it's role in the negative regulation of cell proliferation [179]. A recent study concluded that the loss of E-cadherin can be utilised as a marker for increased susceptibility for oral cancer progression from oral leukoplakia [180]. Changes in the relative concentration of E-cadherin, which has a complex secondary structure, will lead to subtle shifts in the constituent bands that contribute to the convoluted amide bands.

The importance at 1020 cm^{-1} indicates that glycogen plays an important role in predicting malignant transformation in high-grade lesions. The abundance of glycogen has been shown to deplete in pre-malignant tissue as a result of the increased proliferation of abnormal cells, which require more energy than tissue in a less proliferative state [181]. It is plausible that the lesions in this dataset which did eventually transform were in a proliferative state, especially those of a higher histopathological grade.

The importance of the spectral region attributed to nucleic acids (1242 cm^{-1}) is also interesting. DNA aneuploidy refers to the case where there are an abnormal number of chromosomes in a givne cell, and is a known event in oral carcinogenesis [168]. A plausible hypothesis is that lesions with higher propensity to transform may be part characterised by this mechanism, leading to the changes in this region of the spectrum.

The decision not to use the metric analysis (MA) method described in section 3.3.2 is due to a number of reasons. Principally, the method is not suitable for the `PipeOpt` framework, given that it is written in a different language and application programming interface (API). The brute force nature of MA, which trials every pair of wavenumbers in a defined spectral domain, also retracts from it's utility in this study, given that the LOPOCV method depicted in Fig. 6.5 generates from 42 to 224 independent data partitions depending on the cohort. The parallel processing framework (HTCondor) has access to a finite number of processors, so utilisation of an expensive algorithm would impose a computational bottleneck on the analysis.

6.5 Conclusion

In this chapter, FTIR-MS of tissue harbouring OED has been used in conjunction with an objective and and pragmatic analytical pipeline, described in chapter 5, to build classification models which can predict malignant transformation. Despite promising results, it is important to note that the size of the dataset is limited, which limits the impact. Nevertheless, it is the author's opinion that this work is significant as it indicates feasibility of the approach, which will hopefully lead to more a more robust analysis on wider cohorts of patients.

In addition to demonstrating a promising tool in the scope of oral cancer prediction, it was also shown that the method can provide intelligible reasoning in terms of what spectral features (and hence, biochemical moieties) drive discrimination between high and low risk lesions. The characteristic wavenumbers from this study find origin in biomolecules such as proteins, nucleic acids and glygogen, which are generally understood as major components in malignant transformation. The next chapter will focus on critically discussing the impact and limitations of this thesis.

Chapter 7

Future and Outlook

It is the author's belief that there is novelty in the work undertaken in this thesis. Chapter 4, the first experimental chapter, described a series of experiments where two techniques, Fourier transform infrared (FTIR) imaging and infrared scanning near-field optical microscopy (IR-SNOM) were exploited to gain further insight into primary and metastatic oral cancer tissue. The first study utilised a novel classification algorithm, 'metric analysis' (MA), to discriminate amongst several histopathologically defined classes. The results showed that, for the first time, the method is robust to semantically label hyperspectral data that can be directly cross-referenced with histopathological images.

The remaining two experiments in the chapter involved investigating the use of MA and FTIR as a means to inform the configuration of IR-SNOM experiments. IR-SNOM is a discrete-frequency infrared (DF-IR) technique, where the planning and data acquisition time scale is significant in comparison to FTIR. The results show that the techniques can give deep and complementary insight into the chemical processes involved with oral cancer metastasis, such as the concentration and spatial arrangement of collage, and increased protein production in the advancing front of a metastatic tumour.

Chapter 5 describes the concept, design and construction of a novel framework, *PipeOpt*, which is able to efficiently and thoroughly search for the best protocol to pre-process and classify labelled spectral data. Despite the generally accepted notion that this process should be optimised objectively, in reality this approach is unfortunately not widespread. Furthermore, the data and labels will have a heavy influence on the optimal solution. For these reasons, it is recommended that researchers in the field adopt this approach, or something similar. The unification of methodological approaches and a departure from subjective decisions made on intuitive bases will, in the author's opinion, only serve to advance research as scientists can collaborate and communicate far more effectively.

Chapter 6 describes the investigation into the feasibility of using IR techniques to predict malignancy in patients with histopathologically diagnosed oral epithelial dysplasia, which is the most common precursor to frank oral squamous cell carcinoma (OSCC). The work aimed to address some of the issues surrounding oral pre-malignant disorders (OPMDs), such as inter and intra-observer disagreement, subjective and ambiguous gold-standard diagnosis and high costs. The experiments leveraged the framework described in chapter 5 in order to maximise the performance of the predictive model. To the author's knowledge, this was the first case of using FTIR-MS to directly predict malignant transformation of OED. One of the key strengths of this study is the use of hard, certain labels (malignant progression within a well defined time-frame), rather than those based on a subjective diagnosis. Through a robust statistical analysis, the study demonstrated that the approach has potential utility, but aspects such as clinical implementation (rule-in/rule-out) must be considered.

One confounding factor within the work contained in this thesis is the cost of obtaining high quality data. Despite the advances of FTIR technology over the last few decades, acquiring hyperspectral images of excised human tissue is very expensive, in terms of both labour and capital. Aside from the specialised, multi-staged process of tissue extraction, storage and sample preparation, state of the art FTIR microscopes are costly and in high demand, often meaning that extended periods of experiments are infeasible. This is one of the main reasons as to why FTIR imaging data is scarce, especially in the public domain. Data is the most important resource for the construction of reliable, robust and valid models, and without an abundance of good quality data, it is impossible to definitively assess the efficacy of a technique. The author is entirely aware of these issues, and therefore the described work should act as intermediate, proof-of-concept studies which may aid to pave the way for future large scale studies.

In the following sections, various ways in which these limitations can be addressed are considered.

7.1 Multi-Centre Studies

A multi-centre study is carried out in multiple different locations, by different researchers on different subjects. Their importance in the context of medical trials cannot be understated, and the benefits are abundant [182]. One major advantage is the increased number of participants, as the public reach will scale with the number of centres participating in the study. A consequence of both the increase in quantity and reach of patients is the introduction of a more biologically diverse and representative population. Furthermore, it is essential to perform external validation of a model using data from a separate population, in order to assess generalisability [183]. However, design and conduction of multi-centre trials requires a plurality of logistical and scientific expertise, and therefore, resources. It is thus imperative that planning of wider scale studies should be backed by smaller studies, such as the ones contained in this work, rather than wasting time and money.

As discussed in chapter 5, and in [77], one of the major obstacles in realising the potential of infrared imaging is the striking diversity between methodologies. Without unification, multi-centre studies would be incomparable due to the intrinsic influence sample preparation, instrumental configuration and data analysis imposes on results. It is of the author's view that the work contained within chapter 5 has the potential to contribute towards overcoming this barrier. Comparing objectively optimised results across studies would lead to more robust discussion and consideration into the feasibility of the research.

7.2 Different Sample Domain

Another consideration for the future would be the type of sample used to acquire data. In this work, thin tissue (histopathological) sections obtained from excised tissue was used. There were numerous reasons to utilising this sample domain, one of which was the availability of oral cancer tissue specimens from the the University of Liverpool biobanking facility. Another was that the apparent morphological structures within an infrared image of tissue enabled easy spatial co-registration with a histopathological image, offering clarity in terms of the semantics within the data. Experiments in Chapters 4 and 6 extract labelled spectra under the guidance of histopathological labelling, which wouldn't be possible using a different sample domain.

On the other hand, there are drawbacks to this approach. Firstly, acquiring a large image of tissue (> 1 mm) takes at least 15 minutes, a factor which does not scale well for a large number of samples. Furthermore, a small minority of the imaged region is actually extracted for analysis. Tissue microarrays (TMAs) are closely arranged small (diameter $\approx$ 1 mm) tissue extracts which offer an efficient alternative to imaging multiple sections. Future experiments should incorporate the use of TMAs to increase the throughput of data acquisition.

Another option would be to change to sample medium entirely. Liquid biopsies, which are samples of biological fluid (usually blood serum), offer an alternative approach to conventional surgical biopsies. Compared with their surgical counterpart, liquid biopsies are non-invasive, homogeneous, and easy to obtain, making them ideal for early diagnostic and screening. In the context of FTIR spectroscopy, liquid biopsies remove the requirement for imaging, due to the presence of a homogeneous medium with a known chemical composition and concentration. A recent systematic review by Anderson *et al* [184] summarises that vibrational spectroscopy for clinical cancer diagnosis shows high potential, but argues that standardised and uniform reporting of results is paramount, in addition to increasing patient numbers and performing external validation, if its clinical value is to be realised.

7.3 Deep Learning

The supervised classification algorithms applied in this work, such as metric analysis (MA), logistic regression (LR) and linear discriminant analysis (LDA), are fundamentally simple compared with other methods in machine learning. Metric analysis is an ensemble of bivariate classifiers, with a total of $2n$ parameters (the mean and standard deviation of each metric ratio in the ensemble). Metric analysis is limited by the prior assumption that the variables are normally distributed, which is one of the key reason why it was not applied for the analysis of the more heterogeneous data. LR and LDA are based on determining a set of parameters to incorporated into a linear model which can discriminate between given classes. This constrains the model to learn simple, linear relationships between a small number of features, which is sub-optimal for problems where there may be more nuanced and complex relationships in the data.

Deep learning is a subset of machine learning algorithms which aim to progressively learn features which characterise data [185]. It does this by passing the data through non-linear operations in successive layers, where each layer would extract higher level features than the previous. The design of these algorithms are generally based around artificial neural networks (ANNs), a revolutionary concept inspired by the processes underpinning biological neurological processes.

One major advantage of using neural networks is that they are robust enough to discriminate data with very complex decision boundaries. This is because there are typically millions of trainable parameters within the network, each representing a connection between two features in consecutive layers. The necessity for abundant and diverse data is even more potent for deep learning than for shallower methods, given that the high number of parameters increases susceptibility to overfitting, which leads to poor generalisation accuracy.

Computer vision (CV) is a very active branch of research which merges principles from deep learning and image analysis to emulate processes in the human visual cortex, in order to infer from visual inputs [186]. At the heart of CV are convolutional neural networks (CNNs), which learn the coarse and fine spatial features which characterise imagery. In addition to classification, the versatility of CNNs allow adaptation to other tasks, such as object detection [187] and segmentation [188]. Object detection aims to identify a localised region of interest from a much larger field of view, whilst the purpose of segmentation architectures is to semantically label pixels in an image.

Assuming an abundance of multi-centre, multi-modal data, future studies should make use of these CV architectures to maximise performance and efficiency. An end-to-end, multi-task approach should be adopted, utilising different CV architectures to localise dysplasia and predict outcome. This may take the form of an object detection framework which extracts a crop of dysplastic tissue from an hyperspectral image. This can be followed by a classifier which predicts outcome based on information in the *entire* image. This holistic approach to classification using deep learning is one of the key advantages compared with using the more supervised methods in this thesis, rather than specifically labelling the data to pipe into the ML model, the CV model is able to take into consideration the spatial information, and peripheral features. Furthermore, specifically labelled dysplastic regions will no longer be required, rather just regions of interest to train the object detector, and transformation labels for dysplastic images to train the classifier. This will remove the significant overhead of histopathological segmentation and pixel level labelling.

The future of cancer care and management is reliant on the emergence of state of the art technology, a passionate workforce, and cutting edge research. Research efforts must focus on early diagnostics as population growth and resource scarcity stretch healthcare services worldwide. The future is bright, but the adoption of automated techniques into diagnostics requires collaboration and compromise.

Bibliography

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 0, no. 0, pp. 1–41, 2021, ISSN: 0007-9235. DOI: 10.3322/caac.21660.

[2] B. W. Stewart, C. Wild, International Agency for Research on Cancer, and World Health Organization, *World cancer report 2014*, p. 630, ISBN: 9789283204299.

[3] W. Geneva., "Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. Geneva, World Health Organisation.," *World Health Organization*, 2020.

[4] *American Lung Association. State of Lung Cancer*, 2019. [Online]. Available: https://www.lung.org/research/state-of-lungcancer.

[5] *American Cancer Society. Cancer Facts and Figures*, 2018.

[6] R. Valand, S. Tanna, G. Lawson, and L. Bengtström, "A review of Fourier Transform Infrared (FTIR) spectroscopy used in food adulteration and authenticity investigations," 2019, ISSN: 1944-0057. DOI: 10.1080/19440049.2019.1675909. [Online]. Available: https://www.tandfonline.com/action/journalInformation?journalCode=tfac20.

[7] M. A. Mallah, S. T. H. Sherazi, M. I. Bhanger, S. A. Mahesar, and M. A. Bajeer, "A rapid Fourier-transform infrared (FTIR) spectroscopic method for direct quantification of paracetamol content in solid pharmaceutical formulations," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 141, pp. 64–70, Apr. 2015, ISSN: 13861425. DOI: 10.1016/j.saa.2015.01.036.

[8] P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," *Chemometrics and Intelligent Laboratory Systems*,

vol. 117, no. May, pp. 100–114, 2012, ISSN: 01697439. DOI: 10.1016/j.chemolab.2012.03.011.

[9] A. Géron, *Hands-on Machine Learning*. 2017, ISBN: 978-1-491-96229-9.

[10] Cancer Research UK, *Cancer Research UK Cancer incidence statistics*, 2018. DOI: 10.1002/ijc.20103.

[11] N. W. Johnson, P. Jayasekara, A. A. Amarasinghe, and K. Hemantha, "Squamous cell carcinoma and precursor lesions of the oral cavity: Epidemiology and aetiology," *Periodontology 2000*, 2011, ISSN: 09066713. DOI: 10.1111/j.1600-0757.2011.00401.x.

[12] R. Dwivedi, R. Pandey, S. Chandra, and D. Mehrotra, "Apoptosis and genes involved in oral cancer - A comprehensive review," *Oncology Reviews*, vol. 14, no. 2, pp. 108–120, 2020, ISSN: 19705565. DOI: 10.4081/ONCOL.2020.472.

[13] C. Rivera, *Essentials of oral cancer*, 2015.

[14] X. Jiang, J. Wu, J. Wang, and R. Huang, *Tobacco and oral squamous cell carcinoma: A review of carcinogenic pathways*, 2019. DOI: 10.18332/tid/105844.

[15] D. P. Lane, *p53, guardian of the genome*, 1992. DOI: 10.1038/358015a0.

[16] N. Azad, M. Kumari Maurya, M. Kar, M. M. Goel, A. K. Singh, M. Sagar, D. Mehrotra, and V. Kumar, "Expression of GLUT-1 in oral squamous cell carcinoma in tobacco and non-tobacco users," *Journal of Oral Biology and Craniofacial Research*, 2016, ISSN: 22124268. DOI: 10.1016/j.jobcr.2015.12.006.

[17] V. C. Angadi and P. V. Angadi, "GLUT-1 immunoexpression in oral epithelial dysplasia, oral squamous cell carcinoma, and verrucous carcinoma," *Journal of Oral Science*, 2015, ISSN: 18804926. DOI: 10.2334/josnusd.57.115.

[18] W. M. Lydiatt, S. G. Patel, B. O'Sullivan, M. S. Brandwein, J. A. Ridge, J. C. Migliacci, A. M. Loomis, and J. P. Shah, "Head and neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual," *CA: A Cancer Journal for Clinicians*, 2017, ISSN: 1542-4863. DOI: 10.3322/caac.21389.

[19] V. B. Wreesmann, N. Katabi, F. L. Palmer, P. H. Montero, J. C. Migliacci, M. Gönen, D. Carlson, I. Ganly, J. P. Shah, R. Ghossein, and S. G. Patel, "Influence of extracapsular nodal spread extent on prognosis of oral squamous cell carcinoma," *Head and Neck*, 2016, ISSN: 10970347. DOI: 10.1002/hed.24190.

[20] A. C. Broders, "Squamous-cell epithelioma of the lip: A study of five hundred and thirty-seven cases," *Journal of the American Medical Association*, 1920, ISSN: 23768118. DOI: 10.1001/jama.1920.02620100016007.

[21] A. K. El-Naggar, J. K. C. Chan, J. R. Grandis, T. Takata, and P. J. Slootweg, *WHO Classification of Head and Neck Tumours-WHO/IARC Classification of Tumours*, 2017.

[22] A. Almangush, I. O. Bello, H. Keski-Säntti, L. K. Mäkinen, J. H. Kauppila, M. Pukkila, J. Hagström, J. Laranne, S. Tommola, O. Nieminen, Y. Soini, V. M. Kosma, P. Koivunen, R. Grénman, I. Leivo, and T. Salo, "Depth of invasion, tumor budding, and worst pattern of invasion: Prognostic indicators in early-stage oral tongue cancer," *Head and Neck*, 2014, ISSN: 10970347. DOI: 10.1002/hed.23380.

[23] M. Titford, "The long history of hematoxylin," *Biotechnic and Histochemistry*, 2005, ISSN: 10520295. DOI: 10.1080/10520290500138372.

[24] L. Feller and J. Lemmer, "Oral Squamous Cell Carcinoma : Epidemiology , Clinical Presentation and Treatment," vol. 2012, no. August, pp. 263–268, 2012.

[25] A. Panwar, R. Lindau, and A. Wieland, *Management for premalignant lesions of the oral cavity*, 2014. DOI: 10.1586/14737140.2013.842898.

[26] S. P. Reddi and A. T. Shafer, *Oral Premalignant Lesions: Management Considerations*, 2006. DOI: 10.1016/j.coms.2006.08.002.

[27] S. Warnakulasuriya, N. W. Johnson, and I. Van Der Waal, *Nomenclature and classification of potentially malignant disorders of the oral mucosa*, 2007. DOI: 10.1111/j.1600-0714.2007.00582.x.

[28] M. A. Jaber, S. R. Porter, P. Speight, J. W. Eveson, and C. Scully, "Oral epithelial dysplasia: Clinical characteristics of western European residents," *Oral Oncology*, 2003, ISSN: 13688375. DOI: 10.1016/S1368-8375(03)00045-9.

[29] S. Warnakulasuriya and A. Ariyawardana, "Malignant transformation of oral leukoplakia: A systematic review of observational studies," *Journal of Oral Pathology and Medicine*, 2016, ISSN: 16000714. DOI: 10.1111/jop.12339.

[30] R. Vázquez-Álvarez, F. Fernández-González, P. Gándara-Vila, D. Reboiras-López, A. García-García, and J. M. Gándara-Rey, "Correlation between clinical and

pathologic diagnosis in oral leukoplakia in 54 patients," *Medicina Oral, Patologia Oral y Cirugia Bucal*, 2010, ISSN: 16984447. DOI: 10.4317/medoral.15.e832.

[31] Y. Kuribayashi, F. Tsushima, K. I. Morita, K. Matsumoto, J. Sakurai, A. Uesugi, K. Sato, S. Oda, K. Sakamoto, and H. Harada, "Long-term outcome of non-surgical treatment in patients with oral leukoplakia," *Oral Oncology*, 2015, ISSN: 18790593. DOI: 10.1016/j.oraloncology.2015.09.004.

[32] S. Warnakulasuriya and A. Ariyawardana, "Malignant transformation of oral leukoplakia: A systematic review of observational studies," *Journal of Oral Pathology and Medicine*, 2016, ISSN: 16000714. DOI: 10.1111/jop.12339.

[33] D. Ganesh, P. Sreenivasan, J. Ohman, M. Wallström, P. H. Braz-Silva, D. Giglio, G. Kjeller, and B. Hasséus, "Potentially malignant oral disorders and cancer transformation," *Anticancer Research*, vol. 38, no. 6, pp. 3223–3229, 2018, ISSN: 17917530. DOI: 10.21873/anticanres.12587.

[34] G. Yardimci, "Precancerous lesions of oral mucosa," *World Journal of Clinical Cases*, 2014, ISSN: 2307-8960. DOI: 10.12998/wjcc.v2.i12.866.

[35] C. Scully, *Oral and maxillofacial medicine: The basis of diagnosis and treatment*. 2013, pp. 1–435, ISBN: 9780702049484. DOI: 10.1016/C2011-0-04227-8.

[36] S. Müller, "Update from the 4th Edition of the World Health Organization of Head and Neck Tumours: Tumours of the Oral Cavity and Mobile Tongue," *Head and Neck Pathology*, vol. 11, pp. 33–40, 2017. DOI: 10.1007/s12105-017-0792-3.

[37] S. Warnakulasuriya, J. Reibel, J. Bouquot, and E. Dabelsteen, *Oral epithelial dysplasia classification systems: Predictive value, utility, weaknesses and scope for improvement*, 2008. DOI: 10.1111/j.1600-0714.2007.00584.x.

[38] H. Lumerman, P. Freedman, and S. Kerpel, "Oral epithelial dysplasia and the development of invasive squamous cell carcinoma," *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology and*, 1995, ISSN: 10792104. DOI: 10.1016/S1079-2104(05)80226-4.

[39] P. S. Ho, P. L. Chen, S. Warnakulasuriya, T. Y. Shieh, Y. K. Chen, and I. Y. Huang, "Malignant transformation of oral potentially malignant disorders in males: A retrospective cohort study," *BMC Cancer*, 2009, ISSN: 14712407. DOI: 10.1186/1471-2407-9-260.

[40] S. Silverman, M. Gorsky, and F. Lozada, "Oral leukoplakia and malignant transformation. A follow-up study of 257 patients," *Cancer*, 1984, ISSN: 10970142. DOI: 10.1002/1097-0142(19840201)53:3<563::AID-CNCR2820530332>3.0.CO;2-F.

[41] K. P. Schepman and I. van der Waal, "A proposal for a classification and staging system for oral leukoplakia: a preliminary study," *European Journal of Cancer. Part B: Oral Oncology*, 1995, ISSN: 09641955. DOI: 10.1016/0964-1955(95)00032-1.

[42] I. R. Kramer, L. El, and K. W. Lee, "The clinical features and risk of malignant transformation in sublingual keratosis," *British Dental Journal*, 1978, ISSN: 17417503. DOI: 10.1038/sj.bdj.4804055.

[43] M. A. Pogrel, "Sublingual keratosis and malignant transformation," *Journal of Oral Pathology & Medicine*, 1979, ISSN: 16000714. DOI: 10.1111/j.1600-0714.1979.tb01824.x.

[44] L. Zhang, C. F. Poh, W. L. Lam, J. B. Epstein, X. Cheng, X. Zhang, R. Priddy, J. Lovas, N. D. Le, and M. P. Rosin, "Impact of localized treatment in reducing risk of progression of low-grade oral dysplasia: Molecular evidence of incomplete resection," *Oral Oncology*, 2001, ISSN: 13688375. DOI: 10.1016/S1368-8375(00)00140-8.

[45] O. Hamadah, M. L. Goodson, and P. J. Thomson, "Clinicopathological behaviour of multiple oral dysplastic lesions compared with that of single lesions," *British Journal of Oral and Maxillofacial Surgery*, 2010, ISSN: 02664356. DOI: 10.1016/j.bjoms.2009.08.027.

[46] I. van der Waal, *Potentially malignant disorders of the oral and oropharyngeal mucosa; terminology, classification and present concepts of management*, 2009. DOI: 10.1016/j.oraloncology.2008.05.016.

[47] S. S. Napier, C. G. Cowan, T. A. Gregg, M. Stevenson, P. J. Lamey, and P. G. Toner, "Potentially malignant oral lesions in Northern Ireland: Size (extent) matters," *Oral Diseases*, 2003, ISSN: 1354523X. DOI: 10.1034/j.1601-0825.2003.02888.x.

[48] P. Holmstrup, P. Vedtofte, J. Reibel, and K. Stoltze, “Long-term treatment outcome of oral premalignant lesions,” *Oral Oncology*, 2006, ISSN: 13688375. DOI: 10.1016/j.oraloncology.2005.08.011.

[49] M. W. Ho, J. M. Risk, J. A. Woolgar, E. A. Field, J. K. Field, J. C. Steele, B. P. Rajlawat, A. Triantafyllou, S. N. Rogers, D. Lowe, and R. J. Shaw, “The clinical determinants of malignant transformation in oral epithelial dysplasia,” *Oral Oncology*, vol. 48, no. 10, pp. 969–976, 2012, ISSN: 13688375. DOI: 10.1016/j.oraloncology.2012.04.002.

[50] P. G. Arduino, A. Surace, M. Carbone, A. Elia, G. Massolini, S. Gandolfo, and R. Broccoletti, “Outcome of oral dysplasia: A retrospective hospital-based study of 207 patients with a long follow-up,” *Journal of Oral Pathology and Medicine*, 2009, ISSN: 09042512. DOI: 10.1111/j.1600-0714.2009.00782.x.

[51] S. Warnakulasuriya, T. Kovacevic, P. Madden, V. H. Coupland, M. Sperandio, E. Odell, and H. Møller, “Factors predicting malignant transformation in oral potentially malignant disorders among patients accrued over a 10-year period in South East England,” *Journal of Oral Pathology and Medicine*, 2011, ISSN: 09042512. DOI: 10.1111/j.1600-0714.2011.01054.x.

[52] G. Pitiyage, W. M. Tilakaratne, M. Tavassoli, and S. Warnakulasuriya, “Molecular markers in oral epithelial dysplasia: Review,” *Journal of Oral Pathology and Medicine*, vol. 38, no. 10, pp. 737–752, 2009, ISSN: 09042512. DOI: 10.1111/j.1600-0714.2009.00804.x.

[53] K. P. Schepman, E. H. Van Der Meij, L. E. Smeele, and I. Van Der Waal, “Malignant transformation of oral leukoplakia: A follow-up study of a hospital-based population of 166 patients with oral leukoplakia from The Netherlands,” *Oral Oncology*, 1998, ISSN: 13688375. DOI: 10.1016/S1368-8375(97)00097-3.

[54] N. Gale, V. Kambic, L. Michaels, A. Cardesa, H. Hellquist, N. Zidar, and M. Poljak, *The Ljubljana classification: a practical strategy for the diagnosis of laryngeal precancerous lesions.* 2000. DOI: 10.1097/00125480-200007040-00006.

[55] O. Kujan, R. J. Oliver, A. Khattab, S. A. Roberts, N. Thakker, and P. Sloan, “Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation,” *Oral Oncology*, vol. 42, no. 10, pp. 987–993, Nov. 2006, ISSN: 13688375. DOI: 10.1016/j.oraloncology.2005.12.014.

[56] L. Krishnan, K. Karpagaselvi, J. Kumarswamy, U. S. Sudheendra, K. V. Santosh, and A. Patil, "Inter- and intra-observer variability in three grading systems for oral epithelial dysplasia," *Journal of Oral and Maxillofacial Pathology*, 2016, ISSN: 1998393X. DOI: 10.4103/0973-029X.185928.

[57] D. Sawan and A. Mashlah, "Evaluation of premalignant and malignant lesions by fluorescent light (VELscope)," *Journal of International Society of Preventive and Community Dentistry*, 2015, ISSN: 2231-0762. DOI: 10.4103/2231-0762.159967.

[58] S. Warnakulasuriya, "Diagnostic adjuncts on oral cancer and precancer: An update for practitioners," *British Dental Journal*, 2017, ISSN: 17417503. DOI: 10.1038/sj.bdj.2017.883.

[59] R. Macey, T. Walsh, P. Brocklehurst, A. R. Kerr, J. L. Liu, M. W. Lingen, G. R. Ogden, S. Warnakulasuriya, and C. Scully, *Diagnostic tests for oral cancer and potentially malignant disorders in patients presenting with clinically evident lesions*, 2015. DOI: 10.1002/14651858.CD010276.pub2.

[60] K. H. Awan, Y. H. Yang, P. R. Morgan, and S. Warnakulasuriya, "Utility of toluidine blue as a diagnostic adjunct in the detection of potentially malignant disorders of the oral cavity - a clinical and histological assessment," *Oral Diseases*, 2012, ISSN: 1354523X. DOI: 10.1111/j.1601-0825.2012.01935.x.

[61] P. Cancela-Rodríguez, R. Cerero-Lapiedra, G. Esparza-Gómez, S. Llamas-Martínez, and S. Warnakulasuriya, "The use of toluidine blue in the detection of premalignant and malignant oral lesions," *Journal of Oral Pathology and Medicine*, 2011, ISSN: 09042512. DOI: 10.1111/j.1600-0714.2010.00985.x.

[62] J. Smith, T. Rattay, C. McConkey, T. Helliwell, and H. Mehanna, *Biomarkers in dysplasia of the oral cavity: A systematic review*, 2009. DOI: 10.1016/j.oraloncology.2009.02.006.

[63] G. J. J. Verhoeven and V. Archaeology, *The reflection of two fields – Electromagnetic radiation and its role in (aerial) imaging*, 2017.

[64] P. R. Griffiths and J. A. D. Haseth, *Fourier Transform Infrared Spectrometry*, 4621. 2007, vol. 222, p. 704, ISBN: 0470106298. DOI: 10.1002/047010631X.

[65] V. M. Zolotarev, "High-temperature spectral emissivity of SiC in the IR range," *Optics and Spectroscopy (English translation of Optika i Spektroskopiya)*, 2007, ISSN: 0030400X. DOI: 10.1134/S0030400X07100116.

[66] B. D. Guenther and D. G. Steel, *Encyclopedia of modern optics*. 2018, ISBN: 9780128149829. DOI: 10.5860/choice.43-0036.

[67] G. B. Airy, "On the Diffraction of an Object-glass with Circular Aperture," *Transactions of the Cambridge Philosophical Society*, vol. 5, p. 283, 1835.

[68] Rayleigh, " XXXI. Investigations in optics, with special reference to the spectroscope," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1879, ISSN: 1941-5982. DOI: 10.1080/14786447908639684.

[69] D. L. Woernley, "Infrared Absorption Curves for Normal and Neoplastic Tissues and Related Biological Substances," *Cancer Research*, 1952, ISSN: 15387445.

[70] E. Kaznowska, J. Depciuch, K. Szmuc, and J. Cebulski, "Use of FTIR spectroscopy and PCA-LDC analysis to identify cancerous lesions within the human colon," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 134, 2017, ISSN: 1873264X. DOI: 10.1016/j.jpba.2016.11.047.

[71] M. J. Baker, E. Gazi, M. D. Brown, J. H. Shanks, N. W. Clarke, and P. Gardner, "Investigating FTIR based histopathology for the diagnosis of prostate cancer," *Journal of Biophotonics*, 2009, ISSN: 1864063X. DOI: 10.1002/jbio.200810062.

[72] P. D. Lewis, K. E. Lewis, R. Ghosal, S. Bayliss, A. J. Lloyd, J. Wills, R. Godfrey, P. Kloer, and L. A. Mur, "Evaluation of FTIR Spectroscopy as a diagnostic tool for lung cancer using sputum," *BMC Cancer*, 2010, ISSN: 14712407. DOI: 10.1186/1471-2407-10-640.

[73] R. Wang and Y. Wang, "Fourier transform infrared spectroscopy in oral cancer diagnosis," *International Journal of Molecular Sciences*, vol. 22, no. 3, pp. 1–21, 2021, ISSN: 14220067. DOI: 10.3390/ijms22031206.

[74] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin, "Using Fourier transform IR spectroscopy to analyze biological materials," *Nature Protocols*, vol. 9, no. 8, pp. 1771–1791, 2014, ISSN: 17502799. DOI: 10.1038/nprot.2014.110.

[75] T. D. Wang, G. Triadafilopoulos, J. M. Crawford, L. R. Dixon, T. Bhandari, P. Sahbaie, S. Friedland, R. Soetikno, and C. H. Contag, "Detection of endogenous biomolecules in Barrett's esophagus by Fourier transform infrared spectroscopy.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 40, pp. 15 864–15 869, 2007, ISSN: 0027-8424. DOI: 10.1073/pnas.0707567104.

[76] A. Savitzky and M. J. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, 1964, ISSN: 15206882. DOI: 10.1021/ac60214a047.

[77] E. Goormaghtigh, "Infrared imaging in histopathology: Is a unified approach possible?" *Biomedical Spectroscopy and Imaging*, vol. 5, no. 4, pp. 325–346, 2017, ISSN: 22128808. DOI: 10.3233/BSI-160151. [Online]. Available: http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/BSI-160151.

[78] F Ehrenhaft and L Lorenz, "Der physik. 1.," 1908.

[79] C. F. Bohren, *Absorption and scattering of light by small particles*. 1983, ISBN: 9783527406647. DOI: 10.1088/0031-9112/35/3/025.

[80] M. Romeo and M. Diem, "Correction of dispersive line shape artifact observed in diffuse reflection infrared spectroscopy and absorption/reflection (transflection) infrared micro-spectroscopy," in *Vibrational Spectroscopy*, 2005. DOI: 10.1016/j.vibspec.2005.04.003.

[81] 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*, 2008, ISSN: 00037028. DOI: 10.1366/000370208783759669.

[82] P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, and P. Gardner, "Resonant Mie scattering in infrared spectroscopy of biological materials - Understanding the 'dispersion artefact'," *Analyst*, vol. 134, no. 8, pp. 1586–1593, 2009, ISSN: 13645528. DOI: 10.1039/b904808a.

[83] P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, and P. Gardner, "Resonant Mie Scattering (RMieS) correction

of infrared spectra from highly scattering biological samples," *Analyst*, 2010, ISSN: 00032654. DOI: 10.1039/b921056c.

[84] P. Bassan, A. Kohler, H. Martens, J. Lee, E. Jackson, N. Lockyer, P. Dumas, M. Brown, N. Clarke, and P. Gardner, "RMieS-EMSC correction for infrared spectra of biological cells: Extension using full Mie theory and GPU computing," *Journal of Biophotonics*, 2010, ISSN: 1864063X. DOI: 10.1002/jbio.201000036.

[85] M. J. Pilling, A. Henderson, B. Bird, M. D. Brown, N. W. Clarke, and P. Gardner, "High-throughput quantum cascade laser (QCL) spectral histopathology: a practical approach towards clinical translation," *Faraday Discuss.*, vol. 187, pp. 135–154, 2016, ISSN: 1359-6640. DOI: 10.1039/C5FD00176E. [Online]. Available: http://xlink.rsc.org/?DOI=C5FD00176E.

[86] S. Lee, K. Kim, H. Lee, C. H. Jun, H. Chung, and J. J. Park, "Improving the classification accuracy for IR spectroscopic diagnosis of stomach and colon malignancy using non-linear spectral feature extraction methods," *Analyst*, 2013, ISSN: 13645528. DOI: 10.1039/c3an00256j.

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

[88] R. Mankar, M. J. Walsh, R. Bhargava, S. Prasad, and D. Mayerich, "Selecting optimal features from Fourier transform infrared spectroscopy for discrete-frequency imaging," *Analyst*, vol. 143, no. 5, 2018, ISSN: 13645528. DOI: 10.1039/c7an01888f.

[89] A. Burkov, "The Hundred-Page Machine Learning Book-Andriy Burkov," *Expert Systems*, 2019, ISSN: 0266-4720.

[90] M. Claesen and B. De Moor, "Hyperparameter Search in Machine Learning," pp. 10–14, 2015. [Online]. Available: http://arxiv.org/abs/1502.02127.

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

Physics and Technology, 102. https://doi.org/, vol. 102, 2019, ISSN: 13504495. DOI: 10.1016/j.infrared.2019.103007.

[92] J. Goodman, *Introduction to Fourier Optics 3ed*, 2005.

[93] E. Synge, " XXXVIII. A suggested method for extending microscopic resolution into the ultra-microscopic region," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1928, ISSN: 1941-5982. DOI: 10.1080/14786440808564615.

[94] E. A. Ash and G. Nicholls, "Super-resolution aperture scanning microscope," *Nature*, 1972, ISSN: 00280836. DOI: 10.1038/237510a0.

[95] G. Binnig and H. Rohrer, "Scanning tunneling microscopy," *Surface Science*, 1983, ISSN: 00396028. DOI: 10.1016/0039-6028(83)90716-1.

[96] D. W. Pohl, W. Denk, and M. Lanz, "Optical stethoscopy: Image recording with resolution $\lambda/20$," *Applied Physics Letters*, 1984, ISSN: 00036951. DOI: 10.1063/1.94865.

[97] A. Lewis, M. Isaacson, A. Harootunian, and A. Muray, "Development of a 500 Å spatial resolution light microscope. I. light is efficiently transmitted through $\lambda/16$ diameter apertures," *Ultramicroscopy*, 1984, ISSN: 03043991. DOI: 10.1016/0304-3991(84)90201-8.

[98] E. Wolf and M. Nieto-Vesperinas, "Analyticity of the angular spectrum amplitude of scattered fields and some of its consequences," *Journal of the Optical Society of America A*, vol. 2, no. 6, p. 886, 1985, ISSN: 1084-7529. DOI: 10.1364/josaa.2.000886.

[99] K. Imura and H. Okamoto, "Reciprocity in scanning near-field optical microscopy: illumination and collection modes of transmission measurements," *Optics Letters*, 2006, ISSN: 0146-9592. DOI: 10.1364/ol.31.001474.

[100] D. Courjon and C. Bainier, "Near field microscopy and near field optics," *Reports on Progress in Physics*, vol. 57, no. 10, pp. 989–1028, 1994, ISSN: 00344885. DOI: 10.1088/0034-4885/57/10/002.

[101] G. Binnig, C. F. Quate, and C. Gerber, "Atomic force microscope," *Physical Review Letters*, 1986, ISSN: 00319007. DOI: 10.1103/PhysRevLett.56.930.

[102] C. P. Schultz and H. H. Mantsch, "Biochemical imaging and 2D classification of keratin pearl structures in oral squamous cell carcinoma.," *Cellular and molecular biology (Noisy-le-Grand, France)*, 1998, ISSN: 01455680.

[103] Y Fukuyama, S Yoshida, S Yanagisawa, and M Shimizu, "A study on the differences between oral squamous cell carcinomas and normal oral mucosas measured by Fourier transform infrared spectroscopy.," *Biospectroscopy*, vol. 5, pp. 117–126, 1999, ISSN: 1075-4261. DOI: 10.1002/(SICI)1520-6343(1999)5:2<117::AID-BSPY5>3.0.CO;2-K.

[104] P. Bruni, C. Conti, E. Giorgini, M. Pisani, C. Rubini, and G. Tosi, "Histological and microscopy FT-IR imaging study on the proliferative activity and angiogenesis in head and neck tumours," *Faraday Discussions*, vol. 126, no. 1, pp. 19–26, 2004, ISSN: 13645498. DOI: 10.1039/b306787b.

[105] C. Conti, E. Giorgini, T. Pieramici, C. Rubini, and G. Tosi, "FT-IR microscopy imaging on oral cavity tumours, II," *Journal of Molecular Structure*, vol. 744-747, no. SPEC. ISS. Pp. 187–193, 2005, ISSN: 00222860. DOI: 10.1016/j.molstruc.2004.10.042.

[106] S. Sabbatini, C. Conti, C. Rubini, V. Librando, G. Tosi, and E. Giorgini, "Infrared microspectroscopy of Oral Squamous Cell Carcinoma: Spectral signatures of cancer grading," *Vibrational Spectroscopy*, 2013, ISSN: 09242031. DOI: 10.1016/j.vibspec.2013.07.002.

[107] J. D. Pallua, C. Pezzei, B. Zelger, G. Schaefer, L. K. Bittner, V. A. Huck-Pezzei, S. A. Schoenbichler, H. Hahn, A. Kloss-Brandstaetter, F. Kloss, G. K. Bonn, and C. W. Huck, "Fourier transform infrared imaging analysis in discrimination studies of squamous cell carcinoma," *Analyst*, vol. 137, no. 17, pp. 3965–3974, 2012, ISSN: 00032654. DOI: 10.1039/c2an35483g.

[108] S. Banerjee, M. Pal, J. Chakrabarty, C. Petibois, R. R. Paul, A. Giri, and J. Chatterjee, "Fourier-transform-infrared-spectroscopy based spectral-biomarker selection towards optimum diagnostic differentiation of oral leukoplakia and cancer," *Analytical and bioanalytical chemistry*, 2015, ISSN: 16182650. DOI: 10.1007/s00216-015-8960-3.

[109] H. J. Byrne, I. Behl, G. Calado, O. Ibrahim, M. Toner, S. Galvin, C. M. Healy, S. Flint, and F. M. Lyng, "Biomedical applications of vibrational spectroscopy:

Oral cancer diagnostics," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 2021, ISSN: 13861425. DOI: 10.1016/j.saa.2021.119470.

[110] B. G. Ellis, C. A. Whitley, A. Jedani, C. I. Smith, P. J. Gunning, P. Harrison, P. Unsworth, P. Gardner, R. J. Shaw, S. D. Barrett, A. Triantafyllou, J. M. Risk, and P. Weightman, "Insight into metastatic oral cancer tissue from novel analyses using FTIR spectroscopy and aperture IR-SNOM," *Analyst*, pp. 6–9, 2021. DOI: 10.1039/d1an00922b.

[111] G. S. Karagiannis, T. Poutahidis, S. E. Erdman, R. Kirsch, R. H. Riddell, and E. P. Diamandis, *Cancer-associated fibroblasts drive the progression of metastasis through both paracrine and mechanical pressure on cancer tissue*, 2012. DOI: 10.1158/1541-7786.MCR-12-0307.

[112] H. Fu, H. Yang, X. Zhang, and W. Xu, *The emerging roles of exosomes in tumor–stroma interaction*, 2016. DOI: 10.1007/s00432-016-2145-0.

[113] V. V. Boras, A. Fucic, M. Virag, D. Gabric, I. Blivajs, C. Tomasovic-Loncaric, Z. Rakusic, V. Bisof, N. Le Novere, and D. V. Vrdoljak, *Significance of stroma in biology of oral squamous cell carcinoma*, 2018. DOI: 10.5301/tj.5000673.

[114] A. Henderson, *ChiToolbox: MATLAB toolbox for handling hyperspectral data generated by SIMS, FTIR and Raman instruments.* 2014.

[115] J. Trevisan, P. P. Angelov, A. D. Scott, P. L. Carmichael, and F. L. Martin, "IRootLab: A free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis," *Bioinformatics*, vol. 29, no. 8, pp. 1095–1097, 2013, ISSN: 13674803. DOI: 10.1093/bioinformatics/btt084.

[116] B. Perez-Ordonez and N. Marchese, *Diagnostic Pathology: Head and Neck*, 9. 2013, vol. 66, pp. 830–830, ISBN: 9780323392556. DOI: 10.1136/jclinpath-2011-200515.

[117] M. J. Pilling, A. Henderson, J. H. Shanks, M. D. Brown, N. W. Clarke, and P. Gardner, "Infrared spectral histopathology using haematoxylin and eosin (H&amp;E) stained glass slides: a major step forward towards clinical translation," *The Analyst*, vol. 142, no. 8, 2017, ISSN: 0003-2654. DOI: 10.1039/C6AN02224C.

[118] J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott, and F. L. Martin, "Extracting biological information with computational analysis of Fourier-transform

infrared (FTIR) biospectroscopy datasets: current practices to future perspectives," *The Analyst*, vol. 137, no. 14, p. 3202, 2012, ISSN: 0003-2654. DOI: 10.1039/c2an16300d. [Online]. Available: http://xlink.rsc.org/?DOI=c2an16300d.

[119] J. Ingham, "Towards cancer diagnosis via tissue discrimination using various infrared spectroscopy techniques.," Ph.D. dissertation, University of Liverpool, 2018.

[120] G. Shang, C. Wang, H. Lei, and C. Bai, "Detection of shear force with a piezoelectric bimorph cantilever for scanning near-field optical microscopy," *Surface and Interface Analysis*, 2001, ISSN: 01422421. DOI: 10.1002/sia.1057.

[121] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser," *Science*, 1994, ISSN: 00368075. DOI: 10.1126/science.264.5158.553.

[122] KAZARINOV RF and SURIS RA, "Possibility of the amplification of electromagnetic waves in a semiconductor with a superlattice," *Sov Phys Semicond*, 1971.

[123] J. Gan and X. Zhang, *A review of nonlinear hysteresis modeling and control of piezoelectric actuators*, 2019. DOI: 10.1063/1.5093000.

[124] A. C. S. Talari, M. A. Martinez, Z. Movasaghi, S. Rehman, and I. U. Rehman, *Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues*, 2017. DOI: 10.1080/05704928.2016.1230863.

[125] S. Al Jedani, C. I. Smith, P. Gunning, B. G. Ellis, P. Gardner, S. D. Barrett, A. Triantafyllou, J. M. Risk, and P. Weightman, "A de-waxing methodology for scanning probe microscopy," *Analytical Methods*, 2020, ISSN: 17599679. DOI: 10.1039/d0ay00965b.

[126] C. Hughes, L. Gaunt, M. Brown, N. W. Clarke, and P. Gardner, "Assessment of paraffin removal from prostate FFPE sections using transmission mode FTIR-FPA imaging," *Analytical Methods*, 2014, ISSN: 17599660. DOI: 10.1039/c3ay41308j.

[127] C. Belaldavar, D. R. Mane, S. Hallikerimath, and A. D. Kale, *Cytokeratins: Its role and expression profile in oral health and disease*, 2016. DOI: 10.1016/j.ajoms.2015.08.001.

[128] P. R. Bueno, L. N. Gias, R. G. Delgado, J. D. Cebollada, and F. J. Díaz González, "Tumor DNA content as a prognostic indicator in squamous cell carcinoma

of the oral cavity and tongue base," *Head and Neck*, 1998, ISSN: 10433074. DOI: 10.1002/(SICI)1097-0347(199805)20:3<232::AID-HED8>3.0.CO;2-1.

[129] P. Zucchiatti, E. Mitri, S. Kenig, F. Bille, G. Kourousias, D. E. Bedolla, and L. Vaccari, "Contribution of Ribonucleic Acid (RNA) to the fourier transform infrared (FTIR) Spectrum of eukaryotic cells," *Analytical Chemistry*, 2016, ISSN: 15206882. DOI: 10.1021/acs.analchem.6b02744.

[130] C. I. Smith, M. R. Siggel-King, J. Ingham, P. Harrison, D. S. Martin, A. Varro, D. M. Pritchard, M. Surman, S. Barrett, and P. Weightman, "Application of a quantum cascade laser aperture scanning near-field optical microscope to the study of a cancer cell," *Analyst*, 2018, ISSN: 13645528. DOI: 10.1039/c8an01183d.

[131] J. Ingham, T. Craig, C. I. Smith, A. Varro, D. M. Pritchard, S. D. Barrett, D. S. Martin, P. Harrison, P. Unsworth, J. D. Kumar, A. Wolski, A. Cricenti, M. Luce, M. Surman, Y. M. Saveliev, P. Weightman, and M. R. Siggel-King, "Submicron infrared imaging of an oesophageal cancer cell with chemical specificity using an IR-FEL," *Biomedical Physics and Engineering Express*, 2019, ISSN: 20571976. DOI: 10.1088/2057-1976/aaea53.

[132] P. M. Steinert, W. W. Idler, and M. L. Wantz, "Characterization of the keratin filament subunits unique to bovine snout epidermis.," *The Biochemical journal*, 1980, ISSN: 02646021. DOI: 10.1042/bj1870913.

[133] H. H. Bragulla and D. G. Homberger, "Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia," in *Journal of Anatomy*, 2009. DOI: 10.1111/j.1469-7580.2009.01066.x.

[134] M. Bryne, "Is the invasive front of an oral carcinoma the most important area for prognostication?" *Oral Diseases*, 1998, ISSN: 1354523X. DOI: 10.1111/j.1601-0825.1998.tb00260.x.

[135] H. P. Wang, H. C. Wang, and Y. J. Huang, "Microscopic FTIR studies of lung cancer cells in pleural fluid," *Science of the Total Environment*, 1997, ISSN: 00489697. DOI: 10.1016/S0048-9697(97)00180-0.

[136] B. Rigas, S. Morgello, I. S. Goldman, and P. T. Wong, "Human colorectal cancers display abnormal Fourier-transform infrared spectra," *Proceedings of the National Academy of Sciences of the United States of America*, 1990, ISSN: 00278424. DOI: 10.1073/pnas.87.20.8140.

[137] H. Fabian, M. Jackson, L. Murphy, P. H. Watson, I. Fichtner, and H. H. Mantsch, "A comparative infrared spectroscopic study of human breast tumors and breast tumor cell xenografts," *Biospectroscopy*, 1995, ISSN: 15206343. DOI: 10.1002/bspy.350010106.

[138] S. Xu, H. Xu, W. Wang, S. Li, H. Li, T. Li, W. Zhang, X. Yu, and L. Liu, *The role of collagen in cancer: From bench to bedside*, 2019. DOI: 10.1186/s12967-019-2058-1.

[139] S. M. Kakkad, M. Solaiyappan, B. O'Rourke, I. Stasinopoulos, E. Ackerstaff, V. Raman, Z. M. Bhujwalla, and K. Glunde, "Hypoxic tumor microenvironments reduce collagen I fiber density," *Neoplasia*, 2010, ISSN: 14765586. DOI: 10.1593/neo.10344.

[140] C. L. Morais, M. Paraskevaidi, L. Cui, N. J. Fullwood, M. Isabelle, K. M. Lima, P. L. Martin-Hirsch, H. Sreedhar, J. Trevisan, M. J. Walsh, D. Zhang, Y. G. Zhu, and F. L. Martin, "Standardization of complex biologically derived spectrochemical datasets," *Nature Protocols*, vol. 14, no. 5, pp. 1546–1577, May 2019, ISSN: 17502799. DOI: 10.1038/s41596-019-0150-x.

[141] A FAMILI, W SHEN, R WEBER, and E SIMOUDIS, "Data preprocessing and intelligent data analysis," *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 3–23, Jan. 1997, ISSN: 1088-467X. DOI: 10.1016/S1088-467X(98)00007-9.

[142] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, and L. M. Buydens, "Breaking with trends in pre-processing?" *TrAC Trends in Analytical Chemistry*, vol. 50, pp. 96–106, Oct. 2013, ISSN: 0165-9936. DOI: 10.1016/J.TRAC.2013.04.015.

[143] R. M. Jarvis and R. Goodacre, "Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data," *Bioinformatics*, vol. 21, no. 7, pp. 860–868, 2005, ISSN: 13674803. DOI: 10.1093/bioinformatics/bti102.

[144] H. J. Butler, B. R. Smith, R. Fritzsch, P. Radhakrishnan, D. Palmer, and M. J. Baker, "Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy," *The Analyst*, 2018, ISSN: 0003-2654. DOI: 10.1039/C8AN01384E. [Online]. Available: http://pubs.rsc.org/en/Content/ArticleLanding/2018/AN/C8AN01384E.

[145] J. Mockus, V. Tiesis, and A. Zilinskas, "The application of Bayesian methods for seeking the extremum," *Towards Global Optimisation*, 1978.

[146] J. B. Mockus and L. J. Mockus, "Bayesian approach to global optimization and application to multiobjective and constrained problems," *Journal of Optimization Theory and Applications*, 1991, ISSN: 00223239. DOI: 10.1007/BF00940509.

[147] P. Frazier, "Tutorial: Optimization via simulation with Bayesian statistics and dynamic programming," in *Proceedings - Winter Simulation Conference*, 2012, ISBN: 9781467347792. DOI: 10.1109/WSC.2012.6465237.

[148] E. C. Garrido-Merchán and D. Hernández-Lobato, "Dealing with categorical and integer-valued variables in Bayesian Optimization with Gaussian processes," *Neurocomputing*, vol. 380, pp. 20–35, 2020, ISSN: 18728286. DOI: 10.1016/j.neucom.2019.11.004.

[149] 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*, vol. 12, pp. 2825–2830, 2011.

[150] L. S. Leslie, T. P. Wrobel, D. Mayerich, S. Bindra, R. Emmadi, and R. Bhargava, "High definition infrared spectroscopic imaging for lymph node histopathology," *PLoS ONE*, 2015, ISSN: 19326203. DOI: 10.1371/journal.pone.0127238.

[151] S. Mittal, T. P. Wrobel, L. S. Leslie, A. Kadjacsy-Balla, and R. Bhargava, "A four class model for digital breast histopathology using high-definition Fourier transform infrared (FT-IR) spectroscopic imaging," in *Medical Imaging 2016: Digital Pathology*, 2016, ISBN: 9781510600263. DOI: 10.1117/12.2217358.

[152] S. Mittal, K. Yeh, L. Suzanne Leslie, S. Kenkel, A. Kajdacsy-Balla, and R. Bhargava, "Simultaneous cancer and tumor microenvironment subtyping using confocal infrared microscopy for all-digital molecular histopathology," *Proceedings of the National Academy of Sciences of the United States of America*, 2018, ISSN: 10916490. DOI: 10.1073/pnas.1719551115.

[153] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[154] D. Thain, T. Tannenbaum, and M. Livny, *Distributed computing in practice: The Condor experience*, 2005. DOI: 10.1002/cpe.938.

[155] J. Dhanda, A. Triantafyllou, T. Liloglou, H. Kalirai, B. Lloyd, R. Hanlon, R. J. Shaw, D. R. Sibson, and J. M. Risk, "SERPINE1 and SMA expression at the invasive front predict extracapsular spread and survival in oral squamous cell carcinoma," *British Journal of Cancer*, 2014, ISSN: 15321827. DOI: 10.1038/bjc.2014.500.

[156] B. Ellis, C. Whitley, A. Triantyfflou, P. Gunning, C. Smith, S. Barrett, P. Gardner, R. Shaw, P. Weightman, and J. Risk, "Prediction of malignant transformation in oral epithelial dysplasia using infrared absorbance spectra (in press)," *PLOS*, 2022.

[157] P. Nankivell and H. Mehanna, *Oral dysplasia: Biomarkers, treatment, and follow-up*, 2011. DOI: 10.1007/s11912-010-0150-z.

[158] X. Zhang, S. Han, H. Y. Han, M. H. Ryu, K. Y. Kim, E. J. Choi, I. H. Cha, and J. Kim, "Risk prediction for malignant conversion of oral epithelial dysplasia by hypoxia related protein expression," *Pathology*, 2013, ISSN: 14653931. DOI: 10.1097/PAT.0b013e3283632624.

[159] M. W. Ho, M. P. Ryan, J. Gupta, A. Triantafyllou, J. M. Risk, R. J. Shaw, and J. B. Wilson, "Loss of FANCD2 and related proteins may predict malignant transformation in oral epithelial dysplasia," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, Jul. 2021, ISSN: 2212-4403. DOI: 10.1016/J.OOOO.2021.07.001.

[160] O. Ibrahim, M. Toner, S. Flint, H. J. Byrne, and F. M. Lyng, "The potential of raman spectroscopy in the diagnosis of dysplastic and malignant oral lesions," *Cancers*, 2021, ISSN: 20726694. DOI: 10.3390/cancers13040619.

[161] I. Behl, G. Calado, A. Malkin, S. Flint, S. Galvin, C. M. Healy, M. L. Pimentel, H. J. Byrne, and F. M. Lyng, "A pilot study for early detection of oral premalignant diseases using oral cytology and Raman micro-spectroscopy: Assessment of confounding factors," *Journal of Biophotonics*, 2020, ISSN: 18640648. DOI: 10.1002/jbio.202000079.

[162] J. Bánóczy, "Oral cancer and precancerous lesions.," *Fogorvosi szemle*, 1997, ISSN: 00155314. DOI: 10.3322/canjclin.52.4.195.

[163] K. Papamarkakis, B. Bird, J. M. Schubert, M. Miljković, R. Wein, K. Bedrossian, N. Laver, and M. Diem, "Cytopathology by optical methods: Spectral cytopathology of the oral mucosa," *Laboratory Investigation*, 2010, ISSN: 00236837. DOI: 10.1038/labinvest.2010.1.

[164] C. F. Poh, C. E. Macaulay, D. M. Laronde, P. Michele Williams, L. Zhang, and M. P. Rosin, "Squamous cell carcinoma and precursor lesions: Diagnosis and screening in a technical era," *Periodontology 2000*, 2011, ISSN: 09066713. DOI: 10.1111/j.1600-0757.2011.00386.x.

[165] Q. Nguyen, H. Valizadegan, and M. Hauskrecht, "Learning classification models with soft-label information," *Journal of the American Medical Informatics Association*, 2014, ISSN: 1527974X. DOI: 10.1136/amiajnl-2013-001964.

[166] P. Saintigny, L. Zhang, Y. H. Fan, A. K. El-Naggar, V. A. Papadimitrakopoulou, L. Feng, J. J. Lee, E. S. Kim, W. K. Hong, and L. Mao, "Gene expression profiling predicts the development of oral cancer," *Cancer Prevention Research*, 2011, ISSN: 19406207. DOI: 10.1158/1940-6207.CAPR-10-0155.

[167] Y. Liu, Y. Li, Y. Fu, T. Liu, X. Liu, X. Zhang, J. Fu, X. Guan, T. Chen, X. Chen, and Z. Sun, "Quantitative prediction of oral cancer risk in patients with oral leukoplakia," *Oncotarget*, 2017, ISSN: 19492553. DOI: 10.18632/oncotarget.17550.

[168] A. Donadini, M. Maffei, A. Cavallero, M. Pentenero, D. Malacarne, E. Di Nallo, M. Truini, R. Navone, P. Mereu, M. Scala, A. Santelli, S. Gandolfo, and W. Giaretti, "Oral cancer genesis and progression: DNA near-diploid aneuploidization and endoreduplication by high resolution flow cytometry," *Cellular Oncology*, vol. 32, no. 5-6, pp. 373–383, 2010, ISSN: 15705870. DOI: 10.3233/CLO-2010-0525.

[169] WHO, *Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019*, 2020. [Online]. Available: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death.

[170] R. Goodacre, V. Sergo, H. Barr, C. Sammon, Z. D. Schultz, M. J. Baker, D. Graham, M. P. Marques, J. Sulé-Suso, J. Livermore, K. Faulds, F. Sinjab, P. Matousek, C. J. Campbell, R. Dluhy, P. Gardner, C. Phillips, M. Diem, B. Wood,

A. Apolonskiy, S. Kazarian, L. Fullwood, K. Gough, W. Petrich, G. Lloyd, O. Ibrahim, G. Cinque, G. D. Sockalingum, N. Stone, C. Kendall, S. McAughtrie, D. Perez-Guaita, L. Clark, K. Gerwert, A. Bonifacio, I. Notingher, P. Lasch, R. Bhargava, G. Lepert, K. Mader, and C. Paterson, "Clinical Spectroscopy: general discussion," *Faraday Discuss.*, vol. 187, pp. 429–460, 2016, ISSN: 1359-6640. DOI: 10.1039/C6FD90013E. [Online]. Available: http://xlink.rsc.org/?DOI=C6FD90013E.

[171] K. Ranganathan and L. Kavitha, *Oral epithelial dysplasia: Classifications and clinical relevance in risk assessment of oral potentially malignant disorders*, 2019. DOI: 10.4103/jomfp.JOMFP{_}13{_}19.

[172] K. Rosenberg, *Ten-year risk of false positive screening mammograms and clinical breast examinations.* 1998. DOI: 10.1056/nejm199804163381601.

[173] R. A. Hubbard, K. Kerlikowske, C. I. Flowers, B. C. Yankaskas, W. Zhu, and D. L. Miglioretti, "Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography; a cohort study," *Annals of Internal Medicine*, 2011, ISSN: 15393704. DOI: 10.7326/0003-4819-155-8-201110180-00004.

[174] M. J. Barry, "Prostate-Specific–Antigen Testing for Early Diagnosis of Prostate Cancer," *New England Journal of Medicine*, 2001, ISSN: 0028-4793. DOI: 10.1056/nejm200105033441806.

[175] M. Awadallah, M. Idle, K. Patel, and D. Kademani, "Management update of potentially premalignant oral epithelial lesions," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 125, no. 6, pp. 628–636, 2018, ISSN: 2212-4403. DOI: 10.1016/j.oooo.2018.03.010. [Online]. Available: https://doi.org/10.1016/j.oooo.2018.03.010.

[176] A. P. Fellows, M. T. Casford, and P. B. Davies, "Spectral Analysis and Deconvolution of the Amide I Band of Proteins Presenting with High-Frequency Noise and Baseline Shifts," *Applied Spectroscopy*, 2020, ISSN: 19433530. DOI: 10.1177/0003702819898536.

[177] F. Ardito, M. Giuliani, D. Perrone, G. Troiano, and L. L. Muzio, "The crucial role of protein phosphorylation in cell signalingand its use as targeted therapy

(Review)," *International Journal of Molecular Medicine*, vol. 40, no. 2, pp. 271–280, 2017, ISSN: 1791244X. DOI: 10.3892/ijmm.2017.3036.

[178] M. T. Benchekroun, P. Saintigny, S. M. Thomas, A. K. El, V. Papadimitrakopoulou, H. Ren, W. Lang, Y.-h. Fan, J. Huang, L. Feng, J. J. Lee, E. S. Kim, W. K. Hong, M Faye, J. R. Grandis, and L. Mao, "NIH Public Access," vol. 3, no. 7, pp. 800–809, 2011. DOI: 10.1158/1940-6207.CAPR-09-0163.Epidermal.

[179] M. T. Lau, C. Klausen, and P. C. Leung, "E-cadherin inhibits tumor cell growth by suppressing PI3K/Akt signaling via B-catenin-Egr1-mediated PTEN expression," *Oncogene*, vol. 30, no. 24, pp. 2753–2766, 2011, ISSN: 09509232. DOI: 10.1038/onc.2011.6.

[180] I. I. Sathish, K. Asokan, C. L. Krithika, and A. Ramanathan, "Expression of E-Cadherin and Levels of Dysplasia in Oral Leukoplakia -A Prospective Cohort Study," *Asian Pacific Journal of Cancer Prevention*, vol. 21, no. 2, pp. 405–410, 2020, ISSN: 2476762X. DOI: 10.31557/APJCP.2020.21.2.405.

[181] H. Aizawa, S. i. Yamada, T. Xiao, T. Shimane, K. Hayashi, F. Qi, H. Tanaka, and H. Kurita, "Difference in glycogen metabolism (glycogen synthesis and glycolysis) between normal and dysplastic/malignant oral epithelium," *Archives of Oral Biology*, vol. 83, pp. 340–347, Nov. 2017, ISSN: 18791506. DOI: 10.1016/j.archoralbio.2017.08.014.

[182] A. Cheng, D. Kessler, R. Mackinnon, T. P. Chang, V. M. Nadkarni, E. A. Hunt, J. Duval-Arnould, Y. Lin, M. Pusic, and M. Auerbach, "Conducting multicenter research in healthcare simulation: Lessons learned from the INSPIRE network," *Advances in Simulation*, vol. 2, no. 1, Dec. 2017. DOI: 10.1186/s41077-017-0039-0.

[183] S. E. Bleeker, H. A. Moll, E. W. Steyerberg, A. R. Donders, G. Derksen-Lubsen, D. E. Grobbee, and K. G. Moons, "External validation is necessary in prediction research: A clinical example," *Journal of Clinical Epidemiology*, vol. 56, no. 9, pp. 826–832, Sep. 2003, ISSN: 08954356. DOI: 10.1016/S0895-4356(03)00207-5.

[184] D. J. Anderson, R. G. Anderson, S. J. Moug, and M. J. Baker, "Liquid biopsy for cancer diagnosis using vibrational spectroscopy: systematic review," *BJS Open*, vol. 4, no. 4, pp. 554–562, Aug. 2020, ISSN: 2474-9842. DOI: 10.1002/bjs5.50289.

[185] Y. Lecun, Y. Bengio, and G. Hinton, *Deep learning*, May 2015. DOI: 10.1038/nature14539.

[186] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep Learning for Computer Vision: A Brief Review,” 2018. DOI: 10.1155/2018/7068349. [Online]. Available: https://doi.org/10.1155/2018/7068349.

[187] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, “Object Detection with Deep Learning: A Review,” Jul. 2018.

[188] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image Segmentation Using Deep Learning: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2021.3059968.