Raman Spectroscopy and Colorectal Cancer: Towards early diagnosis and personalised medicine

Jenkins, Cerys A.

Raman Spectroscopy and Colorectal Cancer: Towards early diagnosis and personalised medicine

Cerys Anne Jenkins

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy.

Swansea University Medical School

September 2018

For Nan and Granddad.

Grand told me to "Give them hell." - I try my best to

Abstract

The development of healthcare technologies to streamline patient referral systems and diagnose the early onset of disease is of great importance for improving cancer survival and is the basis of this work. This thesis details the development of Raman spectroscopy as a triage tool for urgent suspected colorectal cancer referrals. In this work the development of high-throughput, cost effective standardised platforms for the analysis of biofluids with Raman spectroscopy has been shown. The platforms developed allow the analysis of both dry and liquid biofluid samples.

The optimal liquid biopsy for colorectal cancer applications was found to be serum due to its ability to be stored and transported without the formation of precipitates within the samples. Serum samples were then used to optimise dry and liquid HT-platforms for reproducible spectral collection. Principal component analysis (PCA) was used to investigate and optimise inter-user measurements to ensure a robust measurement platform. PCA analysis showed that patient fasting status and sex could have potential effects on spectral reproducibility and diagnostic capability.

The liquid HT platform developed had less sensitivity for colorectal cancer detection than a dry platform. However, it showed lower inter-user spectral variations and the overall analysis time for each sample was faster. It was also less susceptible to freeze-thaw sampling effects in terms of diagnostic capability. This made it the method of choice when considering a translatable technology. The limits of the liquid HT platform were investigated with random forest based machine learning to develop diagnostic models for serum spectra. It was established that the technique could be used for the detection of precursor cancer lesions when tested against healthy control patients with a positive predictive value (PPV) of 40.00% and a negative predictive value (NPV) of 88.89%. The technique could also detect CRC in a large cohort of test patients against healthy controls with a NPV of 94.44%. This approaches the NPV of approximately 98% for the gold standard diagnostic test (colonoscopy) for colorectal cancer. The thesis concludes by discussing the clinical translation of the technique as an effective diagnostic based upon the results presented.

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans after expiry of a bar on access approved by the Swansea University.

Signed ..... (candidate)

Date .....

Acknowledgements

None of this work would have been possible if it were not for the combined support of my fantastic supervisory team. Prof. Dean Harris, thank you for all of your energy, enthusiasm and providing me with more samples than I'd ever hoped for! Prof. Pete Dunstan, thank you for persuading me to embark on this project and then giving me the freedom to explore new ideas and drive this project. I'd like to thank Prof. Cathy Thornton for her support, guidance and extensive knowledge, as well as many uplifting conversations about food. I'd also like to thank Prof. Paul Lewis for the support given at the beginning of this project.

This PhD has granted me the opportunity to make many new friends that have supported me throughout. Specifically, thanks to Dr Georgina Menzies who taught me so much and was a constant sounding board and great friend throughout this project - SGFTW!

A special thank you to the the people on the ILS1 second floor and in CNH. In particular Amanda, Charlie and Gemma for being a constant support network as we came through our PhDs together. I'd like to thank the people who put up with all of the emotions that I brought away from the university. In particular, I must thank my carer Jordan and my surrogate golf mother Jess for all of the laughter that has kept me going over the past 4 years. More recently, the support of the Scottish wifeys who have helped me push through the final stages of this work - thank you.

This PhD would not have been possible without the backing of my family. Mam, Dad you have been there throughout my many years at university, and always supported me to follow the career that made me happy. You have spurred me on to continue working hard, and I will continue to do so in the knowledge I always have you all to call upon if I need help.

Finally, Jacob; You have supported me through this entire project, loved me, picked up my slack and put up with the PhD induced crazy. Without your unwavering support through me living in Swansea and in Scotland this PhD would not have been possible. Multan dankon. Mi amas vin.

"We know what happens to people who stay in the middle of the road...They get run over." - Nye Bevan

Contents

Abstract **v**

Declaration **vii**

Acknowledgements **ix**

1 Introduction **3**

1.1 Cancer . . . 3

1.2 The structure of the colorectum . . . 5

1.2.1 The colorectum . . . 5

1.2.2 Incidence and risk factors . . . 6

1.3 Molecular pathways to Colorectal cancer . . . 9

1.3.1 Genetic Instability . . . 10

1.3.2 The Chromosomal Instability Pathway . . . 11

1.3.3 CpG Island Methylator Phenotype Pathway . . . 12

1.3.4 Microsatellite Instability Pathway . . . 14

1.4 Molecular classification of colorectal cancer . . . 15

1.4.1 Consensus molecular subtypes of colorectal cancers . . . 16

1.4.2 Tumour staging . . . 18

1.5 The route to colorectal cancer diagnosis . . . 20

1.5.1 Symptoms and Screening . . . 20

1.5.2 Precursor lesions and the detection of precursor lesions . . . 25

1.5.3 Other screening tests available outside of the UK . . . 25

1.6 Investigating and diagnosing colorectal cancer . . . 27

1.6.1 Conclusion . . . 29

Bibliography . . . 30

2 Vibrational Spectroscopy **37**

2.1 Raman spectroscopy . . . 37

2.1.1 A Raman spectrum . . . 38

2.2 Classical Theory of Raman Scattering 40
2.2.1 Enhanced Raman Spectroscopy 42
2.3 Infrared Absorption and Fourier transform infrared spectroscopy . 44
2.3.1 FTIR spectroscopy 46
2.4 Interpreting vibrational spectra 48
2.4.1 Choosing the right spectroscopic method 50
2.5 Current role of Vibrational spectroscopy in detecting colorectal cancer 52
2.6 Clinical applications of Raman spectroscopy in colorectal cancer . 53
2.6.1 Tissue analysis 55
2.7 Detection of colorectal cancer in blood samples 57
2.7.1 Surface enhanced Raman scattering detection methods . . 58
2.7.2 Clinical applications of Surface enhanced Raman spectroscopy for colorectal cancer 59
2.7.3 The limitations of Raman and surface enhanced Raman spectroscopy in clinical applications 61
2.8 Conclusions from review of the literature 62
Bibliography 64

3 Experimental principles, materials and methods 73
3.1 General Considerations 73
3.2 Raman Microspectroscopy 74
3.2.1 Raman spectrometer overview 74
3.2.2 Taking a Raman Spectrum 77
3.3 Cleaning methods (substrates) 80
3.4 Attenuated total reflectance spectroscopy 80
3.4.1 Environmental contributions 81
3.4.2 Measurement conditions 82
Bibliography 83

4 Spectral processing and analysis 85
4.1 Introduction 85
4.1.1 Software considerations 86

4.2 Data preprocessing . . . 87
4.2.1 Quality Control . . . 87
4.2.2 Wavenumber standardisation . . . 89
4.2.3 Baseline subtraction . . . 92
4.2.4 Normalisation . . . 98
4.3 Chemometric spectral analysis and machine learning . . . 99
4.3.1 Unsupervised vs supervised techniques . . . 99
4.3.2 Supervised classification algorithms and machine learning techniques . . . 106
4.3.3 Cross validation and diagnostic performance . . . 113
4.4 Automated spectral analysis Matlab app . . . 117
4.5 Conclusion . . . 123
Bibliography . . . 125

5 Baseline studies 129
5.1 Introduction . . . 129
5.1.1 Variability . . . 129
5.1.2 Aims and objectives . . . 130
5.2 Materials and Methods . . . 131
5.2.1 Sample collection . . . 131
5.2.2 FTIR spectroscopy . . . 131
5.2.3 Raman Spectroscopy . . . 131
5.2.4 Spectral Analysis . . . 133
5.3 Results and discussion . . . 135
5.3.1 Characterising peripheral blood samples with vibrational spectroscopy . . . 135
5.3.2 Plasma vs serum . . . 135
5.3.3 Assessment of plasma and serum post freeze-thaw . . . 139
5.3.4 Vibrational band assignments for serum samples . . . 142
5.3.5 Optimisation of Dry measurement platform for Raman spectroscopy . . . 144
5.3.6 The Vroman effect . . . 146

5.3.7 Development of a measurement platform for liquid data acquisition . . . 149
5.3.8 Cooling the well plate . . . 151
5.3.9 Choosing spectroscopic method . . . 153
5.3.10 Pre-analytical considerations . . . 154
5.3.11 Sample handling . . . 154
5.3.12 Freeze-Thaw stability of serum . . . 155
5.3.13 Investigating inter-operator variability . . . 158
5.3.14 Investigating patient demographic effects on Raman spectra 162
5.3.15 Fasted vs non fasted samples . . . 162
5.3.16 Medication . . . 171
5.3.17 Sex . . . 173
5.3.18 Age and smoking status . . . 178
5.3.19 Discussion and future considerations . . . 179
Bibliography . . . 182

6 Optimisation of serum Raman spectroscopy for colorectal cancer detection 187
6.1 Introduction . . . 187
6.2 Aims and objectives . . . 187
6.3 Materials and Methods . . . 188
6.3.1 Raman spectroscopy . . . 189
6.3.2 Spectral Analysis . . . 190
6.4 Results and discussion . . . 194
6.4.1 Investigating feasibility of liquid and dry serum Raman spectroscopy for CRC detection. . . . 194
6.4.2 Pilot study for fresh dry serum Raman spectroscopy . . . 194
6.4.3 Pilot study of fresh liquid serum Raman spectroscopy for CRC detection. . . . 198
6.4.4 Optimisation of pre-processing methods for diagnostics . . 205
6.4.5 The effects of sample modality and patient demographics on diagnostic models . . . 207

6.5 Conclusions and future work 215
Bibliography 217

7 Establishing the limits of diagnostic capability for colorectal cancer with serum Raman spectroscopy 219
7.1 Introduction 219
7.1.1 Aims and objectives 220
7.2 Materials and methods 221
7.2.1 Cohort information 221
7.2.2 Raman microspectroscopy 225
7.2.3 Data pre-processing 226
7.2.4 Data analysis 227
7.3 Results and discussion 230
7.3.1 Investigating precursor lesions and the limit of sensitivity . 231
7.3.2 Application of liquid serum Raman to a binary cancer vs control cohort 235
7.3.3 Feature selection comparisons between polyp and colorectal cancer RF models 239
7.3.4 Comparison to MS characterisation of CRC from blood samples 240
7.3.5 Investigating a non-binary random forest diagnostic model including polyp, control and cancer patients 241
7.3.6 The effects of inflammatory diseases on diagnostic capability 244
7.3.7 Disease monitoring 247
7.3.8 Investigating a multi laser 785 nm and 532 nm diagnostic model 252
7.3.9 Application of methodologies to other cancer types 257
7.4 Conclusions and further work 263
Bibliography 266

8 Conclusions and future outlook 269
8.1 Development of method and apparatus for high throughput data collection and analysis 269

8.1.1 Data analysis routines 271
8.1.2 Further developments 272
8.2 Clinical validation study 272
8.2.1 Future work towards translation 274
8.2.2 University spin out company 278
8.3 Contributions and Publications 278
8.3.1 Poster presentations 278
8.3.2 Oral contributions 278
8.3.3 Publications 279
Bibliography 280

Appendix **282**

A Patient metadata database **283**

B Spectral normalisation **285**

C Matlab code **287**
C.1 Data importing and preprocessing 287
C.1.1 Import 287
C.1.2 Standardisation 289
C.1.3 Rolling circle filter 292
C.1.4 Derivatives 297
C.1.5 Rubber-band baseline subtraction (FTIR specific, not self produced) 299
C.1.6 Iterative polynomial baseline subtraction (Written by G.A.Lloyd) 302
C.1.7 Vector normalisation 305
C.1.8 Normalisation to the phenylalanine peak/min/max (Original code by Royston Goodacre) 306
C.2 Data analysis 307
C.2.1 RF machine learning code 307
C.3 PLS-DA 313

D Additional baseline studies **315**

D.1 Whole blood Raman spectrum 315

D.2 Dry substrate comparison 316

D.3 Peltier cooling system base plate 317

E Limits of the liquid platform **319**

E.1 Polyp patient metadata 320

E.2 Vector normalised combined model confusion matrices 321

E.3 Non-binary model training confusion matrix 322

E.4 Disease monitoring 323

E.5 PCA analysis of different sample collection methods 325

List of Figures

1.1 Anatomy of the colon and rectum. 6

1.2 Distribution of Colorectal cancer by site and sex in the UK. . . . 8

1.3 Specific base pairings that lead to the double helix shape. Taken from [19]. 10

1.4 The position of DNA, genes and chromosomes in relation to a cell. Adapted from [19]. 11

1.5 A schematic drawing of the adenoma to carcinoma sequence for colorectal cancer. 12

1.6 A schematic drawing of the adenoma to carcinoma sequence for colorectal cancer. 14

1.7 Summary of the molecular calssifications of CRC adapted from [27] 16

1.8 A schematic for the CMS of colorectal cancer. 18

1.9 TNM staging of colorectal cancer. Adapted from [31]. 19

1.10 Sensitivity, specificity and negative likelihood ratios of individual symptoms, taken from [36]. 21

1.11 Schematic of the current referral pathways to cancer through symptomatic presentation in the UK. 23

1.12 A comparison of other screening tests being developed for the colorectal cancer screening market. Taken from [52]. 26

1.13 Colorectal treatment pathway for patients referred with suspected CRC. Adapted from [12]. 27

2.1 Different types of light scattering and the corresponding lines on a theoretical spectrum. Adapted from [1] 38

2.2 Raman spectrum of carbon disulphide taken at (a) -12°C and (b) 45°C. Figure is adapted from [2] 39

2.3 Resonance Raman Scattering occurs when the incoming EM radiation is near to or equal to the electronic energy states of an atom. 43

2.4 A typical FTIR spectrum of dried human serum. 47
2.5 Summary of spectral band features for both Raman and FTIR. . . 49
2.6 Vibrational modes of CO2 molecule. 50
2.7 Representative Raman and FTIR spectra of polystyrene foam. . . 51

3.1 Renishaw InVia Raman Microscope beam path. 75
3.2 Beamline shape for the 785nm laser line with pinhole in (a) and pinhole out(b). 76
3.3 Wire software measurement setup. 78
3.4 HiPlan objective (10x) spectral contribution. 79
3.5 PerkinElmer Spectrum Two Beampath. 81
3.6 Energy Spectrum of the background components of the FTIR spectrum. 82

4.1 Flow chart showing the different components of spectral analysis needed in order to interpret spectra for diagnostic purposes. . . . 85
4.2 Selection parameters for cosmic ray removal within the Wire 4.1 software. 88
4.3 Steps of the Wire cosmic ray removal process 89
4.4 Summary of the different methods of shifting wavenumbers with straight shifting(top) interpolation(middle) and the hybrid shift and interpolation (bottom). 91
4.5 Wavenumber calibration. 92
4.6 Representative raw FTIR and Raman spectra of dried serum. . . 93
4.7 Iterative polynomial baseline subtraction. 94
4.8 Example first and second derivative FTIR and Raman spectra of dried serum. 96
4.9 Rolling circle filter background subtraction. 97
4.10 Spectral normalisation methods 99
4.11 PCA takes the original d-dimensional basis set for the data and projects it onto a new orthogonal bases with reduced dimensions such that the variance in the data is maximised. 101

4.12 The PC score projection is a result of the dot product between the loading vectors for each PC and the mean centred data matrix. . 103
4.13 Step by step HCA. 104
4.14 Example HC dendrogram. 105
4.15 PLS-DA variables, adapted from [30]. 107
4.16 Example tree learner, with some of the nodes labelled along a 'branch' for clarity. 111
4.17 RF out of bag classification error convergance. 112
4.18 Gini importance against average spectrum, the higher the importance peak the more useful the wavenumber in classification. . . . 113
4.19 Data splitting for spectral model building and optimisation. . . . 114
4.20 K-fold cross validation; partitioning of the training set to select the validation set. The error from each fold is then combined. . . 114
4.21 ROC curve for an example model, AUC indicates the overall learning performance of the classifier (0.91 classes as very good) and the orange spot denotes the position along the curve of the cross-validated results of the model. 116
4.22 GUI import data files steps. 118
4.23 GUI pre-processing option selection. 119
4.24 GUI plotting functions 120
4.25 GUI analysis suite examples. 121
4.26 GUI plotting functions 122

5.1 Chemical composition of Plasma 136
5.2 Comparison of spectra from fresh liquid serum and plasma samples excited with (a) the 785 nm laser line, (b) the 532 nm laser line. . 138
5.3 Comparison between dried fresh serum and plasma samples excited with the 785 nm laser line. 139
5.4 Example of plasma cryo-precipatate at 785 nm excitation. 140
5.5 Comparison between fresh dried plasma sample and post-freeze-thaw plasma sample with cryo-precipitate droplets. 141

5.6 Example of the multi-well aluminium plate used for all of the dried spectra in this work. 146
5.7 Representative Raman Intensity Map over a dry serum droplet. . 147
5.8 PCA imaging over dry droplet 148
5.9 Optimal sampling position for dry serum droplets (red). 149
5.10 Exemplary example of a stainless steel well plate. 150
5.11 Representative depth profile through the sample, step sizes given in μm. Note that raw data are shown therefore cosmic rays have not been removed. 151
5.12 USB powered Peltier cooling system for the stainless steel well plate. The base plate schematic can be found in Appendix D.3. . 152
5.13 Representative example of raw spectral data from the same patient under cooled and room temperature conditions. Spectral data collection was repeated for 5 different wells using control patient samples. 153
5.14 Example serum FTIR spectra for a liquid and dry samples. 153
5.15 PC score, loading and hierarchical cluster analysis of the fresh and freeze-thawed samples. 156
5.16 PC score, loading and hierarchical cluster analysis of the fresh and freeze-thawed samples. 157
5.17 Inter-user liquid dataset PC Scores,loading and HC clustering analysis. 159
5.18 Inter-user dry dataset PC Scores,loading and HC clustering analysis. 161
5.19 Average serum spectra comparing fasted vs non-fasted patients for 785 nm and 532 nm. 165
5.20 PC scores for CRC patients fasted vs non-fasted. 166
5.21 PC score plots for fasted vs non fasted control patients. 167
5.22 Loadings on PCs for fasted vs non fasted control patients. 167
5.23 PC scores and loadings plots for 532 nm spectra of fasted vs non-fasted cancer patients. 169
5.24 PC score plot for fasting status of control patients. 170

5.25 Comparison of raman spectra with 785 nm and 532 nm excitation from patients on different medications. 172
5.26 Male vs. female average and difference spectra 174
5.27 Male vs. female difference spectra compared to combined cohorts of patients. 175
5.28 PCA scores and loadings for sex of patient with 785 nm laser line. 177
5.29 PCA scores and loadings for sex of patient with 532 nm laser line. 178

6.1 Approach for optimising pre-processing parameters. 192
6.2 Mean and standard deviation 785 nm spectra of cancer vs control with difference spectra 195
6.3 PLS-DA cross validation and ROC curves for fresh dry 785 nm data. 196
6.4 Calculated response for dry fresh 785 nm PLS-DA and loadings on LVs 197
6.5 Mean, standard deviation and difference spectra for fresh liquid 785 nm and 532 nm excitation for cancer vs controls. 199
6.6 ROC curves for liquid pilot study model. 200
6.7 Calculated response for fresh liquid 785 nm PLS-DA and loadings on LVs. 201
6.8 ROC curves for liquid pilot study model. 203
6.9 Calculated response for fresh liquid 532 nm PLS-DA and loadings on LVs. 204
6.10 Representative example of polynomial baseline correction method and normalisation 206
6.11 Comparison between PLS-DA latent variable loadings for a 785 nm dataset. 209
6.12 PLS-DA loadings comparison between fresh and FT liquid samples. 211
6.13 Comparison between 532 nm loadings on LV1-3 for PLS-DA models constructed from fresh and FT datasets. 212

7.1 785 nm mean, standard deviation and difference spectra for polyp vs control spectra. 231

7.2 ROC and classifier position for polyp vs control samples and Gini importance for the overall predictor importance for the large RF binary model. 233

7.3 ROC for larger RF diagnostic model and Gini importance for binary RF classification model. 236

7.4 Literature study summary of metabolite changes between CRC and controls 240

7.5 Predictor importance plot for the polyp, control and cancer RF model. 242

7.6 Predictor importance plot for the polyp, control and cancer RF model. 245

7.7 Average 785 nm patient spectra for a patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. 248

7.8 Average 532 nm patient spectra for a patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. 250

7.9 Average 785 nm patient spectra for the second patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. 251

7.10 Example combined 785 nm and 532 nm spectrum 252

7.11 Cross validated ROC curve comparison between a 785 nm model, a 532 nm model and a combined model. 254

7.12 Gini importance for the combined multi-modal model 257

7.13 Mean, standard deviation and calculated difference spectra between pancreatic cancer and control samples and between pancreatic and colorectal cancer. 259

7.14 PCA score plot (PC1 vs PC2) and associated loadings for pancreatic cancer, colorectal cancer and control data. 260

7.15 ROC and CV calculated response for pancreatic cancer vs control patients. 261

7.16 Cross validated calculated ROC and calculated responses vs scores for pancreatic versus colorectal cancer 262

8.1 Comparison between the current USC referral pathway and a shortened USC pathway by using Raman spectroscopy as a triage tool. 276

A.1 Example of the patient meta data database created during this work. 283

D.1 Example of liquid whole blood Raman spectrum. The blood is said to be 'liquid' however during the measurement time the blood coagulated. 315

D.2 Comparison of different substrate Raman activity. 316

D.3 Schematic of the aluminium base plate design used in the cooling system. This base plate replaces the standard renishaw plate. . . 317

E.1 Full polyp patient dataset 320

E.2 Average 785 nm patient spectra for the second patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. 323

E.3 Average 532 nm patient spectra for the second patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. 324

E.4 PC1 vs PC2 score plot for sample processing methods. 325

List of Tables

2.1 Comparison of detection techniques 52
2.2 Table summarising the clinical applications of Raman spectroscopy to Colorectal Cancer. 54

4.1 Summary of minimum intensities for different sampling modalities 88
4.2 Example of spectral dimensions and coordinate positions for FTIR and Raman spectra for different spectra (S). 90
4.3 Confusion matrix example, with diagonal elements being the correctly predicted spectra and the off-diagonal elements being incorrectly identified spectra. 115

5.1 Data acquisition parameters for different sampling modes. 132
5.2 Raman serum spectral band assignments. 143
5.3 FTIR spectral band assignments for serum, adapted from [3]. Where ν is stretching, δ is bending, s is symmetric and as is asymmetric stretch. 144
5.4 A comparison of the different experimental substrates available for spectroscopic analysis of biological samples. 145
5.5 Cohort details for fasting vs non-fasting patients to investigate effects on Raman spectra 163
5.6 Age and study number of the control patients used for investigation to the effect of sex on spectra. 173

6.1 Patient details for pilot study. 189
6.2 Optimised data acquisition conditions for different sampling modes. 189
6.3 Optimised parameters for different pre-processing methods for dry serum Raman spectral data taken with 50x objective and the 785 nm laser. 191
6.4 Optimised parameters for different pre-processing methods for liquid serum Raman spectral data taken with 10x objective. 191

6.5 Comparison of sensitivities (Sens) and specificities (Spec) for different combinations of pre-processing methods for PLS-DA models. 205
6.6 Comparison of the sensitivity and specificity values for different sampling methods and different laser excitations. 208
6.7 A comparison between liquid models from data from male and females participants. 214

7.1 Patient details for the binary polyp vs control study. 222
7.2 Patient details for RF binary patient study with cancer and healthy control patients 223
7.3 Patient details for the non-binary model including healthy control, cancer, inflammatory, and polyp patients. 224
7.4 Patient details for the multi-laser excitation cancer vs control model 225
7.5 Optimised data acquisition conditions for 785 nm and 532 nm liquid excitation. 226
7.6 Example model result decision key. The 'positive' class for models, i.e. cancer or polyp is 1 and a negative result is 2. 228
7.7 Spectral band assignments for regions within the top 50 most important wavenumbers from the RF predictor importance, [7]. . . . 234
7.8 Patient-wise confusion matrix for polyp spectra vs control spectra. 235
7.9 Spectral band assignments for regions within the top 50 most important wavenumbers from the RF predictor importance, [7]. . . . 237
7.10 Large RF model cross-validation result on a spectrum-wise basis. . 238
7.11 Patient-wise confusion matrix for the detection of colorectal cancer vs control patients in a 49 patient blind model testing set. 239
7.12 Confusion matrix for spectrum wise-analysis from an independent test set of 225 spectra. 242
7.13 Table of sensitivities, specificities, NPV and PPV calculated from the spectrum-wise independent test set confusion matrix in 7.12 . 243
7.14 Per patient results from non-binary model testing polyp, control and cancer patients. 243

7.15 Confusion matrix and calculated sensitivity, specificity, NPV and PPV values for a binary classification including inflammatory control patients. 246
7.16 Spectrum-wise confusion matrix for a combined laser excitation model. 253
7.17 Comparison of the sensitivities, specificities, NPV and PPV values of individual excitation based models and a combined model. . . . 256
7.18 Cross-validated confusion matrix for PLS-DA model on a spectrum-wise basis for pancreatic cancer versus control. 262
7.19 Cross validation confusion matrix for spectrum-wise PLS-DA model discriminating between pancreatic cancer and colorectal cancer. . 263

8.1 Comparison between calculated sensitivities, specificities, susceptibility to user variability and also total sample measurement time for one sample (including pipetting, drying, etc) 270
8.2 Comparison between calculated sensitivities and specificities of the frozen samples, and the relative change in sensitivity and specificity due to freeze-thawing serum samples prior to analysis. 271
8.3 Comparison between the maximum performance of the liquid serum RS diagnostic model compared to current screening and diagnostic methods. Diagnostic vales taken from [2–6]. 273

E.1 Training spectra confusion matricies 321
E.2 Per spectrum testing set confusion matrix 321
E.3 Per patient confusion matrix for the testing set. 321
E.4 Training confusion matrix for non-binary model including polyp, control and cancer spectra. 322
E.5 Calculated sensitivity and specificity for each class from the non-binary model 322

Chapter 1

Introduction

Colorectal cancer (CRC) is a general term that describes a malignant tumour that has developed in either the wall of the large intestine or the rectum. CRC is the fourth most common cancer in the UK and is the second largest cause of cancer related death in the UK after lung cancer [1]. This is largely due to detection at a late stage of disease leading to a poorer prognosis relating at least in part to the invasive nature of many of the current diagnostic tests. The work during this PhD has explored the prospect of using vibrational spectroscopy to address this issue using peripheral blood samples taken from patients.

To better understand the current state of CRC diagnostics and before going onto a more rigorous description of vibrational spectroscopy this chapter will give a general description of cancer. It will then focus on CRC and give the risk factors, incidence and general biology of the disease including the anatomy of the colon and rectum and the molecular classifications of the disease.

The final part of this chapter will include a critical review of the pathways to diagnosis of CRC in the UK. The current diagnostic technology available will be described highlighting particular strengths/weaknesses of these.

1.1 Cancer

Cancer is defined a disease that is caused by the uncontrolled division of a malignant growth or tumour [2]. The term malignant is used to describe the fact that the cells that are dividing in an uncontrolled manner are abnormal. Cancers normally start with abnormal changes in one or a few cells from within the body. These changes can be due to epigenetic, genetic and other factors that cause a cell to not perform it's normal role within the body. For example, the changes will cause a change in cell shape or cause the cells to divide uncontrollably. The type

of cell that has changed and its location help to classify the type of cancer that is present within the body. For example, if cells within the brain are the initial abnormal cells the cancer will be brain cancer. Within this classification there are then subtypes of cancer depending on which cell the cancer has originated from. A few of the larger groups of cancer that are grouped by the type of cell that the cancer originated from are summarised as follows;

- Carcinoma - This is the most common type of cancer, it originates from epithelial cells (cells that create a barrier between the inside and outside of the body).

- Sarcoma - Originates from cells that make up soft tissues and bone such as muscle, fat and blood vessels.

- Leukaemia - This cancer originates in the blood-forming tissue of the bone marrow. Leukaemias do not form a solid tumour, instead large numbers of abnormal white blood cells accumulate in the blood and bone marrow.

- Lymphoma and myeloma - Both lymphoma and myeloma are from immune cells. Lymphoma originates in (T and B lymphocytes) and myeloma forms in plasma cells.

It should also be noted that not all abnormal growths within the body are malignant. A malignant growth has the ability to invade into surrounding tissue and spread into other areas of the body. Instead some tumours are benign, meaning that generally they are slow growing and do not invade into nearby tissue. Colorectal cancer is a heterogeneous disease caused by different molecular pathways that in turn lead to different phenotypes and therefore different classifications [3]. Colorectal carcinoma is the most common subtype of colorectal cancer [1]. More than 90% of colorectal cancers are adenocarcinomas that originate from normal epithelial cells in the colorectal mucosa [4]. It is well established that the majority of these develop from benign adenomas with a small portion developing from serrated or hyper plastic polyps. The remainder of colorectal malignant tumours that are not adenocarcinomas include carcinomas and lymphomas. The work car-

ried out in this thesis will focus mainly on the detection of colorectal carcinoma, however other carcinomas will also be studied such as pancreatic carcinoma.

1.2 The structure of the colorectum

To understand the classification of tumours of the colon and rectum it is first appropriate to consider the anatomy. The colorectum consists of the small intestine, the large intestine and the rectum. In the UK cancer of the small intestine is rare (just under 1300 cases) compared to large intestine and rectal cancers ($\approx$ 46000 each year) [1]. It is the higher incidence of cancers of the colon and rectum that motivates this research to specialise in the detection of colorectal cancer.

1.2.1 The colorectum

In the human body once food has been swallowed, it passes down the oesophagus into the stomach. Once in the stomach the digestive process begins, the food then passes into the small intestine where the digestive process continues. Once digested food has travelled through the small intestine it passes into the large intestine. The small intestine is the longest part of the bowel at around 20ft [1], it is termed 'small' due to it's narrow diameter compared to the large intestine. The large intestine (colorectum) is around 5ft long and can be divided into four main regions; the caecum, the colon (ascending, transverse, descending and sigmoid), the rectum and the anal canal (Figure 1.1). The caecum is a pouch approx 2/3 inches in length. The colon has a diameter that varies between 1-2 inches and the rectum has a diameter that is larger than the colon and its use is primarily as a storage reservoir. Colorectal cancer (CRC) occurs in all regions of the colorectum.

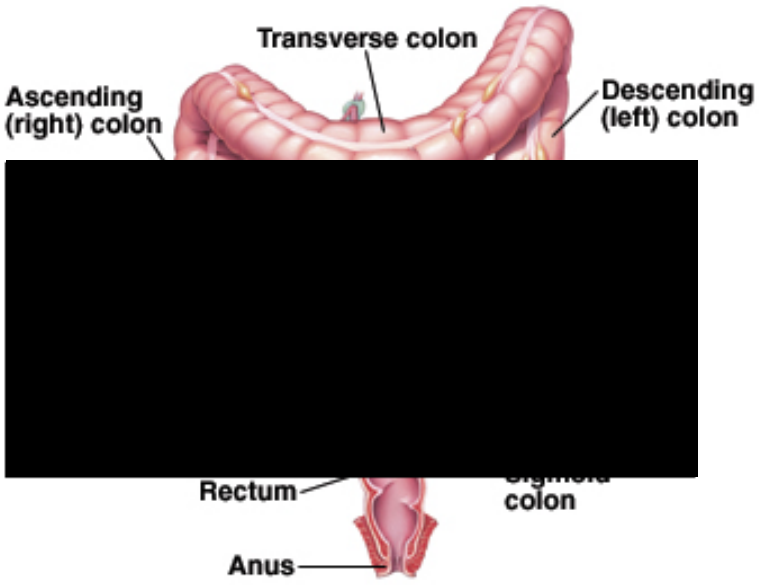


Figure 1.1: Anatomy of the colon and rectum. Taken from [5]

1.2.2 Incidence and risk factors

Worldwide, CRC is the third most common cancer in men (746,000 cases) and the second in women (614,000 cases) [6]. However, incidence is not uniformly distributed throughout the world [6, 7]. It seems that CRC is most common in developed countries, with the incidence rate in places such as Australia, New Zealand and Western Europe (more than 40 per 100,000 people) varying up to 10-fold compared to some parts of Asia and Africa (less than 5 per 100,000 people) [8, 9]. This difference may be partly due to the varying data quality worldwide [6], but it may also reflect the different risk factors, screening programs, and diagnostic methods that people in different countries have access to [1].

Environmental risk factors

Colorectal cancer incidence varies not just within populations but it can also vary within populations living in one community. There is some evidence in the UK that age-standardised incidence rates are 13% higher for males living in the most deprived areas compared with those in the least deprived areas [1]. Further to

this, studies on groups of migrants moving from low-risk to high-risk communities show that migrants tend to rapidly adopt the risk of developing CRC associated to their new communities [7], [9]. The results of these studies suggest that there are environmental factors that have a role in colorectal carcinogenesis. The definition of what exactly constitutes an environmental factor remains unclear. However there are some risk factors that are recognised by the World Health Organisation (WHO) to contribute to CRC development. These include 'lifestyle factors'. In the UK around 54% of colorectal cancers are linked to lifestyle [10] with diet being one of the largest factors contributing to the risk of contracting colorectal cancer. Typically, high incidence is observed in areas that are considered to have a 'Western style' diet i.e with high caloric foods rich in animal fat , coupled with a sedentary lifestyle. Smoking, alcohol consumption and red meat consumption have been identified as risk factors. Vegetable consumption along with exercise have been identified as risk lowering factors [11].

Non-modifiable risk factors

Further to the modifiable or environmental risk factors that play a role in determining an individuals overall risk of developing a colorectal cancer, there are a number of non-modifiable risk factors. These include but are not limited to sex, height, age, personal history of cancer/other diseases of the bowel and inherited genetic risk.

Sex Across western countries men generally have a higher risk of developing colorectal cancer. In the UK in 2014 there were 22,844 reported cases of colorectal cancer in males compared to 18,421 cases in females [1]. The age-standardised incidence rate in England was 54.4% higher in males than females [12]. Sex also seems to affect the distribution of colorectal cancer within the bowel. For example males have higher percentage of rectal cancer whereas the percentage of female patients who had right-sided tumour was higher than men (Figure 1.2).

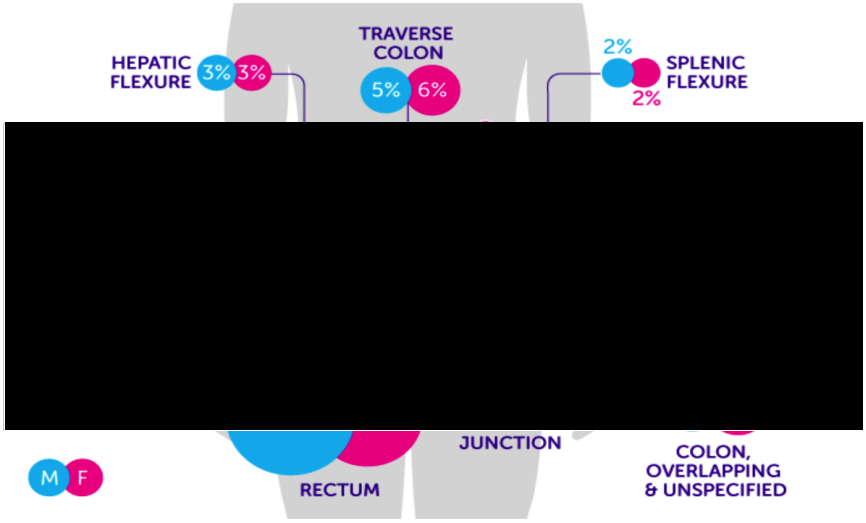


Figure 1.2: Distribution of Colorectal cancer by site and sex in the UK. Taken from [1]

Age/Height In the UK CRC is most frequently diagnosed in people aged 60 and over. CRC has a number of clinical symptoms and for some of these, the risk of achieving a firm diagnosis varies with age [13]. A few studies have shown a correlation between an individuals adult attained height and the risk of developing CRC with individuals that are taller being at increased risk.

Pre-existing conditions of the Bowel Chronic inflammatory bowel diseases represent a significant risk factor in developing colorectal cancer, with risk increasing after a patient suffering from the disease for more than 8-10 years [14]. Some pre-existing conditions such as Crohns disease and Ulcerative colitis are associated with increased risk for an individual to develop CRC. Patients with Crohn's disease have a three times higher risk than normal of developing a malignant tumour in both the small and the large intestine. Ulcerative colitis is considered a pre-malignant disorder and 1% of Colorectal cancers are thought to be due to Ulcerative colitis [1].

Genetic susceptibility Colorectal cancer has both sporadic and familial (inherited) forms, therefore some genetic conditions can lead to an increased risk in developing CRC. The most common inherited CRC associated syndrome is Lynch syndrome also known as Hereditary Nonpolyposis Colorectal Cancer (HNPCC).

Lynch syndrome accounts for around 3-5% of all CRC cases. The lifetime risk of developing CRC if you have Lynch syndrome has been calculated to be around 66% for men and 43% for women [15]. Familial adenomatous polyposis (FAP) is another inherited colorectal cancer syndrome but it is much less prominent than Lynch syndrome. FAP accounts for approximately 1% of CRC cases and affects both men and women equally. If a patient with FAP is left untreated the risk of developing CRC is approximately 100% [16]. Other genetic syndromes associated with CRC include Peutz-Jeghers syndrome and Juvenile polyposis syndrome which both have a cumulative lifetime risk of developing CRC of around 40% [4].

Previous cancer Patients who have a history of colorectal cancer are at an increased risk of developing a second primary colorectal cancer [17]. In some cohort studies the risk of developing a secondary primary colorectal cancer is higher if a patient is a survivor of head and neck cancers, cervical cancer, lung cancer, breast and oesophageal cancers [18].

1.3 Molecular pathways to Colorectal cancer

It is generally accepted that CRC results from the transformation of normal mucosal cells to adenoma (abnormal growth) and then to carcinoma (malignant growth). It is a heterogeneous disease that has both sporadic and familial forms (hereditary). Approximately 95% of all CRC cases can be considered at least partially sporadic [3]. Traditionally, three distinct molecular pathways have been recognised to lead to carcinoma in the colorectum in sporadic cases. These are known as the chromosomal instability pathway (CIN), the microsatellite instability pathway (MSI) and the CpG island methylator phenotype pathway (CIMP). This section will summarise the different molecular pathways to colorectal cancer and how the disease can be classified according the the molecular characteristics of the cancer. It will also summarise the current route a patient takes to diagnosis of CRC in the UK and therefore the motivation behind this work.

1.3.1 Genetic Instability

Instability in the genome plays a large part in the development of CRC. The accepted molecular pathways to CRC generally involve a series of genetic and epigenetic changes that lead to carcinoma. To better understand these changes it is appropriate to first consider the role of deoxyribonucleic acid (DNA).

DNA is a linear polymer which encodes the genetic information for living organisms. It has a very simple structure which consists of smaller units called nucleotides. Each nucleotide consists of a nitrogen base (monomer) with a backbone built of repeating sugar-phosphate units.

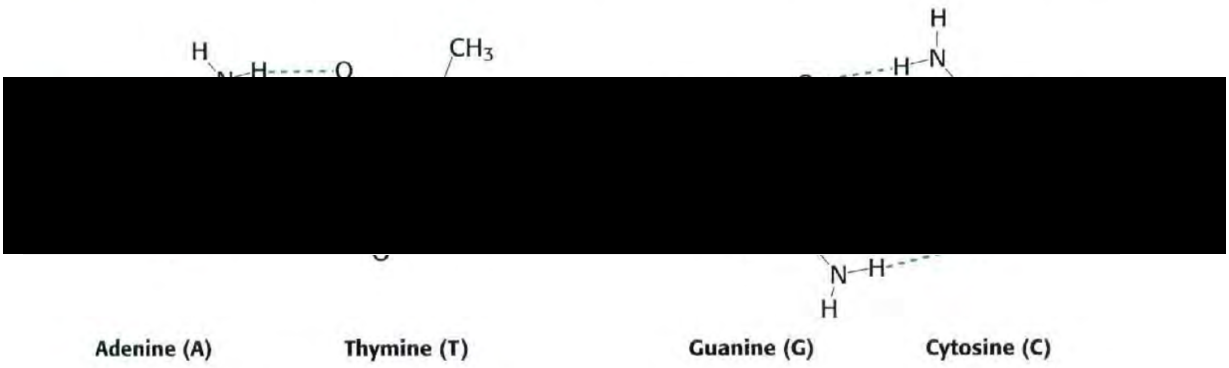


Figure 1.3: Specific base pairings that lead to the double helix shape. Taken from [19].

The molecules that make up this backbone are mono saccharides known as deoxyribose, from which DNA gets its name, and a phosphate group. There are only four nitrogen bases that join to the deoxyribose to form nucleotides in DNA, these are Adenine(A), Thymine(T), Guanine(G) and Cytosine(C) see Figure (1.3).

The bases are covalently attached to the deoxyribose components in the backbone to form chains of nucleotides which then form strands of DNA. The sequence in which the nucleotides are arranged along a strand of DNA codes for specific genes. The genes are then arranged along chromosomes (Figure 1.4) in the nucleus of cells within the body. Humans have 23 pairs of chromosomes making a set of 46. Each chromosome has a longer arm (q arm) and a shorter arm (p arm) which can be used to identify the location of genes.

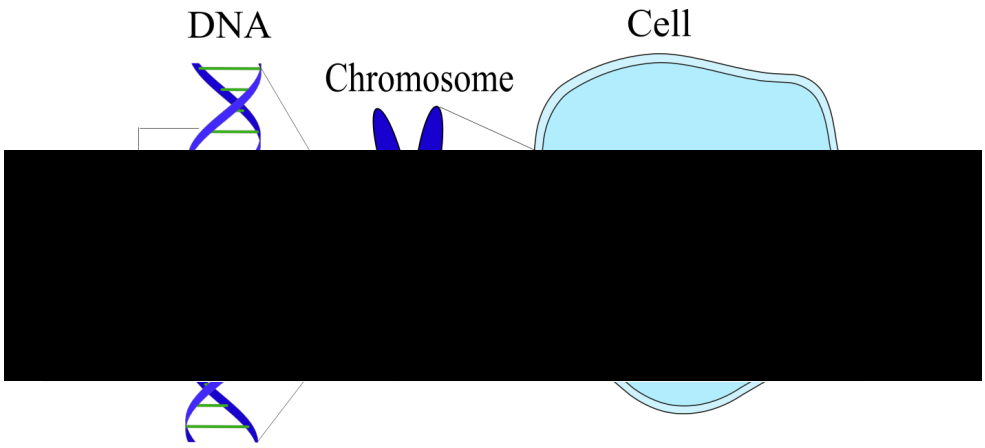


Figure 1.4: The position of DNA, genes and chromosomes in relation to a cell. Adapted from [19].

1.3.2 The Chromosomal Instability Pathway

The Chromosomal instability pathway (CIN) is the most common pathway to CRC. The CIN pathway accounts for 65-70% of all sporadic CRC cases [3]. CIN is characterized by a gain or loss of whole chromosomes or chromosomal regions. This causes an imbalance of chromosome number (aneuploidy), sub-chromosomal genomic amplifications, and a high frequency of loss of heterozygosity (LOH).

The CIN model is an updated model of the 'adenoma to carcinoma' sequence originally posed in 1990 by Vogelstein et. al. (Figure 1.5) [20]. The model consisted of a homogeneous set of genetic mutations that led to a progression from normal mucosa to the formation of an adenoma to carcinoma.

In the CIN pathway the earliest identifiable lesion is the dysplastic aberrant crypt focus (ACF) this is a microscopic lesion that can lead to the development of a polyp [21]. Dysplasia in the ACF is mostly associated with inactivation of the Adenomatous Polypsis Coli (APC) gene through mutation and/or loss of chromosome 5q that contains the APC gene. Progression to CRC occurs in 30-60% of cases by activation of the Kirsten-Rat Scarcoma (K-RAS) group of proto-oncogenes through mutation. K-RAS gene activation can affect many different cellular pathways that control mechanisms such as cellular growth, differentiation and survival, thus contributing to colorectal tumurogensis. The role of K-RAS in

tumorogenesis is not unique to CIN pathway CRCs. It also plays an important role in the CpG Island Methylator Phenotype pathway (CIMP) which is detailed below. In the CIN model, K-RAS activation is then followed by loss of chromosome region 18q and deletion of chromosome 17p which contains the tumour suppressor gene 53 (TP53). TP53 codes for the p53 protein, p53 normally acts to slow down the cell cycle to allow DNA sufficient time to repair itself. Loss or impairment of the gene usually occurs late in the cycle but enhances the transition from adenoma to carcinoma.

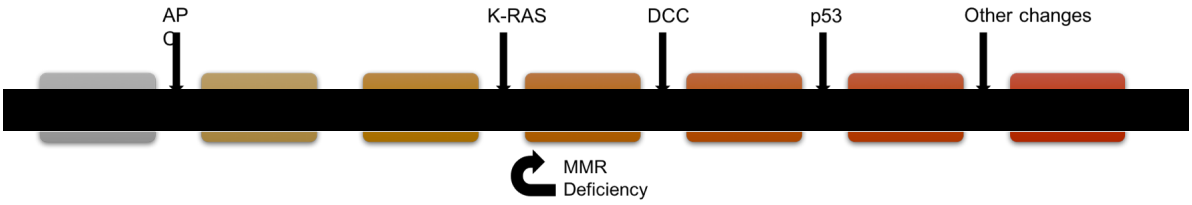


Figure 1.5: A schematic drawing of the adenoma to carcinoma sequence for colorectal cancer. Where MMR is the mis-match repair deficiency described in section 1.3.3. It should be noted that the time frame of this model spans over many years as this is a gradual process. Adapted from [22].

Not all CIN positive tumours have this exact sequence of aberrations. Studies into human genome sequences are currently on-going should provide a more-detailed analysis of the chromosomal instability pathway. Nevertheless, the CIN pathway and the accompanying adenoma to carcinoma sequence has provided a foundation for the molecular classification of CRC. In particular, it has provided a reference frame for other molecular CRC profiles such as the CIMP and microsatellite instability (MSI) pathways.

1.3.3 CpG Island Methylator Phenotype Pathway

Epigenetic alterations in the genome are alterations that can affect the function or expression of a gene without changing its DNA sequence. One of the most common epigenetic alterations is methylation of CpG (Cytosine preceeding guanine) dinucleotides. Around 70% of all CpG dinucleotides in the human genome are

heavily methylated, the remainder GpG sites are typically found in CpG 'islands' (regions of $\leq$200bp) that occur commonly in the promoter region of genes [23]. Methylation of CpG islands can therefore inhibit gene transcription. In the context of CRC methylation of CpG dinucleotides can inhibit the transcription of tumour suppressor genes and therefore provide an alternative mechanism for loss of function of these genes [23]. It is important to note that methylation can occur as a part of the normal function of a cell and does so in many somatic cells, however hypermethylation can cause problems and complete inactivation of genes. In CRC genes affected by DNA hypermethylation at CpG islands include APC, MLH1 and MGMT. The CpG island methylator phenotype pathway (CIMP) is characterised by the hypermthylation of marker genes. There is a panel of five marker genes, CACNA1G, IGF2, NEUOG1, RUNX3 and SOCS1. CIMP positive CRCs are defined by having methylation of at least three of the five marker genes [24]. CIMP positive tumours account for around 15-20% of all sporadic CRCs. The precursor to CIMP CRCs is usually a sessile serrated adenoma (SSP). SSP lesions appear to be smoother and flatter than ACF lesions because the cells do not display dysplasia. Despite the difference in pathology some CIMP-positive CRCs are closely linked to CIN pathway tumours because both pathways can affect the same tumour suppressor genes such as the APC and KRAS genes. Figure 1.6 shows a simplified model of the link between the CIN and the CIMP pathways to CRC.

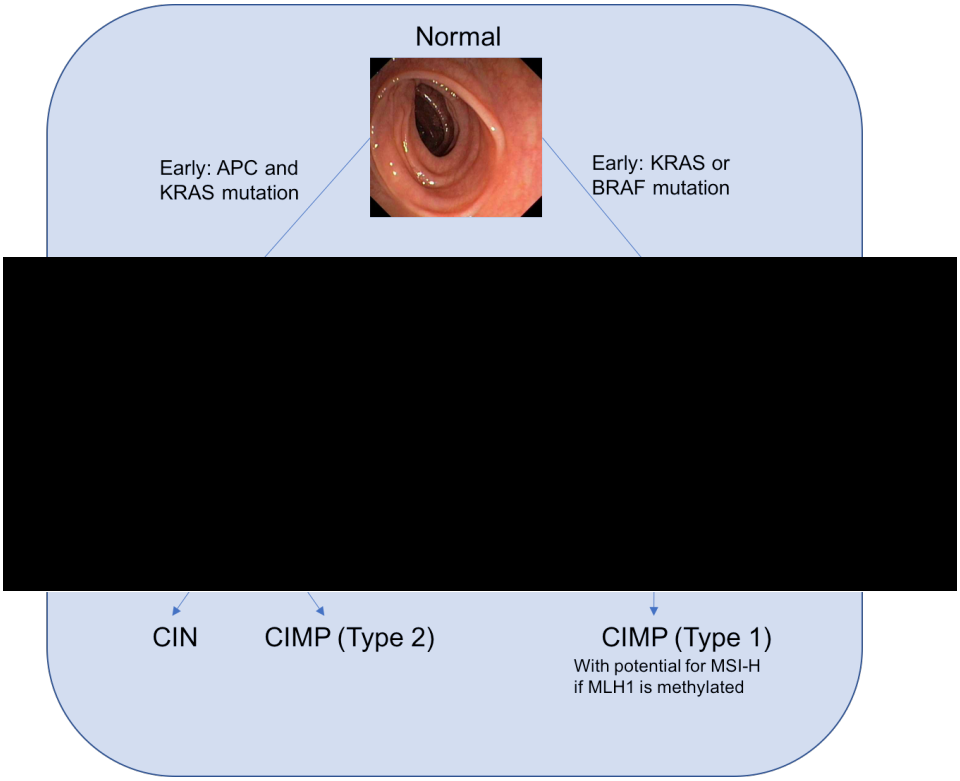


Figure 1.6: A schematic drawing of the adenoma to carcinoma sequence for colorectal cancer. This is a gradual process and the time frame of this model spans over many years as this is a gradual process. Adapted from [25] with pictures provided by Dean A Harris.

1.3.4 Microsatellite Instability Pathway

Microsatellite Instability (MSI) occurs in approximately 15-20% of all CRC cases [26]. Microsatellites are short repeating nucleotide sequences that occur across the entire genome. Microsatellites tend to have a higher rate of mutation/error in replication due to them consisting of many repeat sequences. The DNA mis-match repair (MMR) system is designed to recognise these mutations and repair them, in cases of tumours with MSI the MMR system fails to recognise and fix errors in the DNA replication process. This failure of the MMR system results in a discrepancy in the number of nucleotide repeats in the microsatellite region of the tumour DNA compared to normal DNA. The level of instability (MSI) of the MMR system is measured by the level of instability in a panel of 5 microsatellite sites. The sites include two mononucleotide (BAT25, BAT26) and

three dinucleotide microsatellites (D5S346, D2S123 and D17S250). MSI-high is characterised by instability in more than two of the 5 sites. MSI-low is characterised by instability in one of the microsatellite sites. MSI-stable is where no instability can be detected in the five sites. MSI-high CRC is often associated with silencing of the MLH1 gene which is part of the MMR system [3].

1.4 Molecular classification of colorectal cancer

Having a clear classification system of CRCs in terms of tumour molecular markers aids the process of managing the CRC as well including the likelihoods of different outcomes for the patient. Previously, CRCs have been classified according to the pathway the CRC took to becoming a carcinoma i.e. CIN,CIMP,MSI. However, not all CRCs follow a distinct pathway to tumourogenesis. The molecular classification has undergone much development in the past decade. In 2007, Jass proposed a clasification system based on the underling genetic instability of the cancer with five groups:

1. CIMP-High, MSI-High, methylation of MLH1 and BRAF mutation;
2. CIMP-high, MSI-Low or MS-Stable, BRAF mutation;
3. CIMP-Low/MSI-Stable ot MSI-Low/KRAS mutation;
4. CIMP-negative/MSI-Stable;
5. CIMP-negative/MSI-high.

Groups 1 and 2 normally originate in sessile serrated polyps (SSPs), group 3 originates in either SSPs or adenomas and groups 4 and 5 originate in adenomas [27]. However, there was still a problem with this classification as many of the classes of cancer had overlapping characteristics. The genomic analysis of CRC tumours highlighted a further two groups of CRC, hypermutated cancers and non-hypermutated cancers (Figure 1.7).

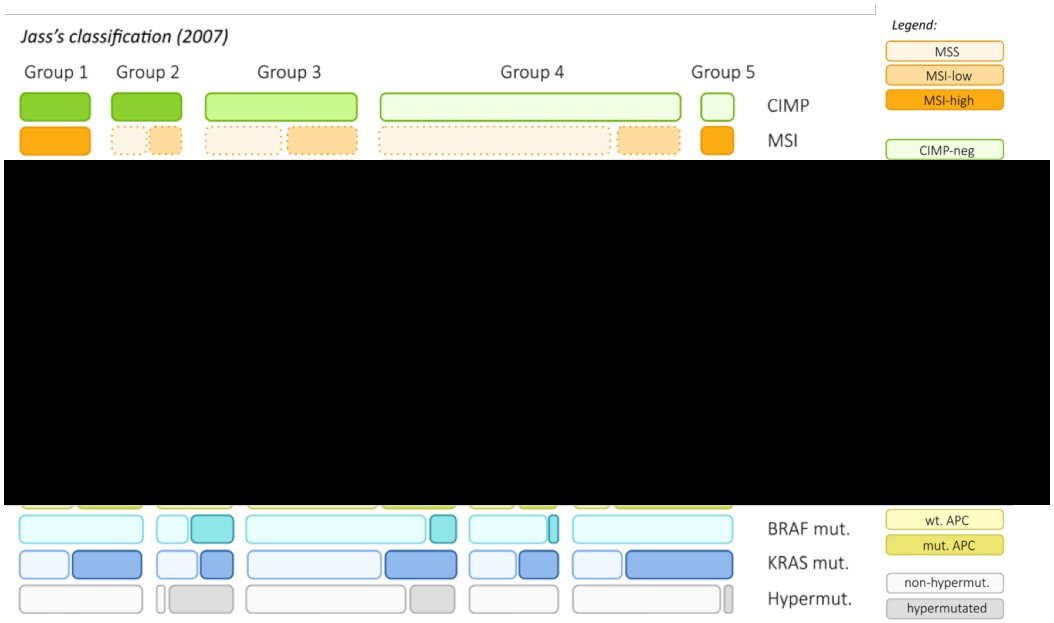


Figure 1.7: Summary of two of the classification methods for molecular classification of CRCs. The size of boxes indicates the number of relative cases of each subtype. The 2014 classification does not include the final two groups of unmutated and other cancers.

Following the Cancer Genome Atlas in 2013, an alternative classification model was proposed by Domingo et al [28]. It consisted of six groups:

1. MSI and BRAF mutations;
2. CIN and/or TP53 mutation, wild type KRAS and PIK3CA;
3. KRAS and/or PIK3CA mutations, CIN, wild-type TP53;
4. KRAS and/or PIK3CA mutations, CIN-negative, wild type P53;
5. no mutations;
6. others.

1.4.1 Consensus molecular subtypes of colorectal cancers

To tackle the problems with classifying CRCs, following the cancer genome atlas study in 2014 there has been an international effort towards resolving inconsisten-

cies in previously reported gene-based CRC classifications. The consensus molecular subtypes of colorectal cancers was the result of a large-scale data sharing and analysis project. Guinney et al showed that by combining 18 large datasets (n = 4,151) in a central repository and applying network based analysis using six different independent classification systems that four main classification groups can be identified. These groups, the consensus molecular subtypes (CMS) are split via their molecular distinguishing features as follows [29]:

CMS1 (microsatellite instability immune 14%), hypermutated, microsatellite unstable and strong immune activation;

CMS2 (canonical 37%), epithelial, market WNT and MYC signalling activation

CMS3 (metabolic 13%), epithelial and evident metabolic dysregulation;

CMS4 (mesenchymal, 23%), prominent transforming growth factor-β activation, stromal invasion and angiogenesis

Mixed (mixed feature samples 13%), not falling into any one classification and is considered either a transition in phenotype or intratumoural heterogeneity.

Figure 1.8 shows a schematic representation of the CMS for colorectal cancer. The schematic was created by Dienstmann et al. It shows that MSI is linked the hypermutation, hypermethylation, immune filtration, activation of RAS and BRAF mutations and is also associated to the proximal colon locations. The CIN tumours show a higher heterogeneity at the gene-expression level with different activation pathways from both CMS2 and CMS4. CIN tumours are mainly found in the left colon or rectum and the environment of the tumour tends to be poorly inflamed/immunogenic, with marked stromal inflation. Finally, a subset of CRC tumours enriched for RAS mutations have strong metabolic adaption (CMS3), intermediate levels of mutation as well as methylation and copy number events.

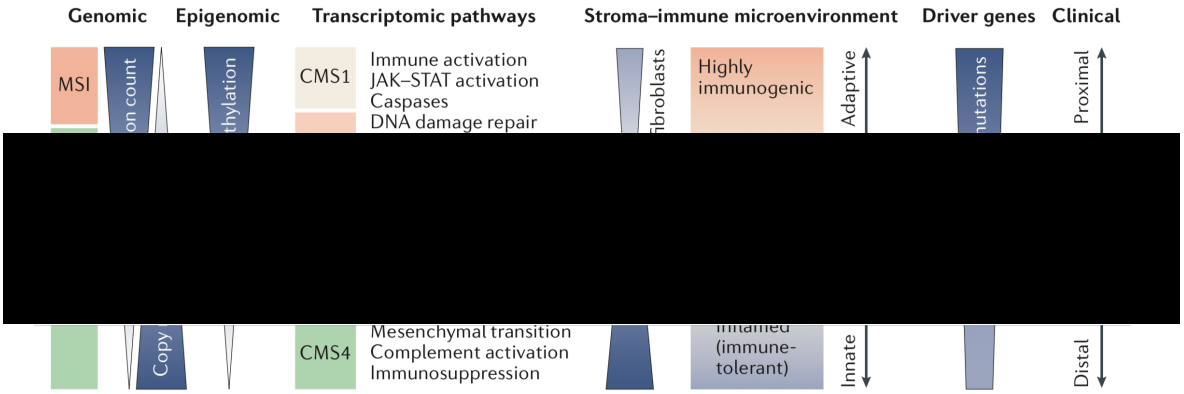


Figure 1.8: A schematic representation of the CMS for CRC. Where JAK, Janus kinase; STAT, signal transducer and activator transcription; TGF-β, transformation growth factor-β; VEGF, vascular epidemal growth factor. Taken from [30]

This system still does not include information such as response to treatment so the classification systems are still in development. The complex nature of the classification of disease and the fact that it is still in development highlights the heterogeneous nature of CRCs. It also highlights how difficult it is to classify a particular patient's disease. This in turn makes detecting and correctly classifying the disease difficult.

1.4.2 Tumour staging

The stage of progression of a cancerous tumour within the body plays an import role in the classification of a tumour and acts as a predictor for clinical outcome. There are two main systems of tumour classification that are used; Duke's and TMN, both based on the level of invasion of a tumour.

Duke's tumour classification

The Duke's tumour classification was originally used for colorectal cancer classification. It was split into four groups depending on how far a tumour has invaded into a patients bowel wall. The staging consisted of 4 main groups defines as follows:

- Dukes A - Tumour has invaded but is limited to mucosa.

- Dukes B1 - Invasion extends to the muscular tissue of the bowel wall but not through it and no lymph nodes are involved.
- Dukes B2 - Tumour has penetrated through the muscularis propria but no lymph nodes are involved.
- Dukes C1 - Tumour extends into the muscularis propria but does not go through it, however there are lymph nodes involved.
- Dukes C2 - Tumour has penetrated the muscularis propria and lymph nodes are involved.
- Dukes D - Tumour has metastasised.

TNM tumour classification

The TNM staging system is commonly used to reference the stage of colorectal cancer. It is also used with different criteria for different parts of the body. In the colorectum the layers of tissue that make up the wall of the bowel can be split into layers as in Figure 1.9.

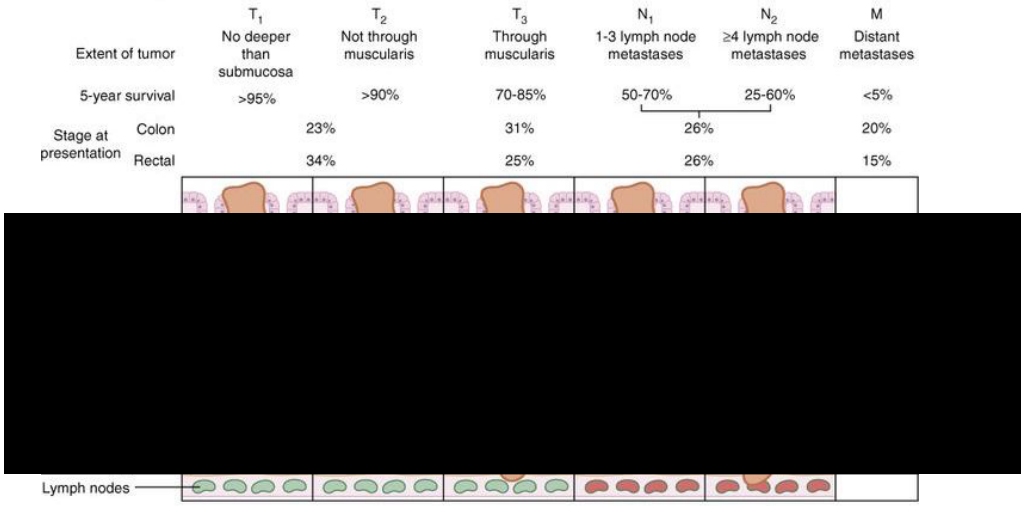


Figure 1.9: TNM staging of colorectal cancer. Adapted from [31].

The TNM system also classifies a tumour depending on the level of invasion through the layers of the colorectum as well as if the tumour has come into contact with regional lymph nodes. For example, a T3N1M0 classification would

imply that the tumor has invaded through the muscularis propria (smooth muscle), has 1-3 lymph node metastases and has no distant metastases. The correct classification of the tumour is vital for the prognosis and treatment of the cancer. To correctly classify the tumour a patient typically goes through a series of tests. The typical staging modalities for colorectal cancer include CT and MRI scanning. However these are performed once a patient is already in secondary care with suspected colorectal cancer or already confirmed cancer via other methods. The following section describes how CRC is detected in the UK following the current patient referral and treatment pathways for CRCs.

1.5 The route to colorectal cancer diagnosis

To understand the advantages/disadvantages of many of the technologies used to detect colorectal cancer it is important to consider the pathway by which a patient is diagnosed in the UK. This is particularly important when considering translation of Raman or other new technology into a treatment pathway as it has to be acceptable to both patients and NHS staff. Healthcare facilities in the UK follow guidelines given by the National Institute for Health and Care Excellence (NICE) for referral and treatment guidelines [32], [33]. The current treatment pathway is initially dependent on either an emergency presentation of colorectal cancer (i.e. acute large bowel obstruction) or presentation in primary care with suspected colorectal cancer. The majority of patients with bowel cancer in the UK are diagnosed following a GP referral (55%). Just under 10% of patients are diagnosed following a referral from the bowel screening programme [34].

1.5.1 Symptoms and Screening

The majority of CRCs present symptomatically in a primary care setting [35]. The relationship between initial symptoms and mortality as a diagnostic indicator have been discussed in depth [36], [37], [38], [35]. Many symptoms that present locally include abdominal pain, change in 'normal' bowel habit (looser, more frequent stools) and rectal bleeding [1]. There are also systemic symptoms

experienced by patients such as weight loss and mild anaemia [38]. A study by Hamilton et al (2006) found no relationship between the duration that symptoms had been experienced by patients and the staging of the disease or mortality. However, rectal bleeding alone as an initial symptom is associated with lower mortality rates whereas mild anaemia (haemoglobin of 10-12.9 g dl^{-1}) is associated with more advanced tumour staging and worse mortality rates [36]. Unfortunately initial symptoms of suspected CRC can also be symptoms of benign diseases [33], but currently there is no diagnostic test available in primary care that has sufficient differentiation to inform referral [36]. Furthermore, single symptoms that present alone tend lack diagnostic value as seen in Figure 1.10.

Table 1. Sensitivity, specificity, positive and negative likelihood ratios of unpaired symptoms

Symptom and study	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)
Rectal bleeding				
Hamilton *et al*, 2005[27]	42 (37.2 to 47.8)	96 (94.8 to 96.7)	10.13 (7.85 to 13.08)	0.60 (0.55 to 0.66)
Hamilton *et al*, 2009[28]	16 (14.6 to 16.6)	99 (98.7 to 98.9)	12.97 (11.62 to 14.48)	0.85 (0.84 to 0.86)
Panzuto *et al*, 2003[11]	44 (28.5 to 60.3)	60 (53.3 to 66.1)	1.09 (0.75 to 1.60)	0.94 (0.70 to 1.25)
Summary estimates	17 (16.4 to 18.4)	98 (98.3 to 98.6)	5.31 (1.65 to 17.07)	0.77 (0.57 to 1.03)
	I^2 = 98.6%, P<0.001	I^2 = 99.6%, P<0.001	I^2 = 98.7%, P<0.001	I^2 = 96.7%, P<0.001
Abdominal pain				
Hamilton *et al*, 2005[27]	42 (37.2 to 47.8)	91 (89.2 to 92.0)	4.54 (3.75 to 5.49)	0.64 (0.58 to 0.70)
[redacted]				
Panzuto *et al*, 2003[11]	68 (51.9 to 81.9)	83 (77.5 to 87.4)	3.98 (2.81 to 5.64)	0.38 (0.24 to 0.60)
Change in bowel habit				
Hamilton *et al*, 2008[28]	11 (10.4 to 12.1)	99 (98.9 to 99.1)	11.47 (10.12 to 13.00)	0.90 (0.89 to 0.91)
Panzuto *et al*, 2003[11]	20 (8.8 to 34.9)	80 (73.8 to 84.4)	0.95 (0.49 to 1.86)	1.01 (0.86 to 1.19)
Bloating				
Panzuto *et al*, 2003[11]	54 (38.7 to 67.9)	39 (33.4 to 45.6)	0.88 (0.63 to 1.15)	1.18 (0.79 to 1.64)

Figure 1.10: Sensitivity, specificity and negative likelihood ratios of individual symptoms, taken from [36].

This leaves GPs referring patients according to who they consider to be at the highest risk, which is calculated on a combination of symptoms (Figure 1.11) and age. As can be seen the current pathway is highly complicated and the routes to secondary care diagnostic pathways for CRCs rely on symptoms and basic testing in primary care such as Faecal occult blood testing and basic blood tests to test for anaemia. Rapid diagnosis of CRC is crucial to patient outcomes, the 5-year survival rate for colorectal cancers detected in early stages are $> 90\%$, however the 5-year survival rate for later-stage cancers is $< 10\%$ [1], [39]. This highlights the clinical need for earlier diagnosis of CRC. In 2006, a national screening programme was introduced using the faecal occult blood test (FOBT) to reduce the mortality rate of colorectal cancer in the UK by aiding the referral information available to GPs.

Currently there are two types of screening methods for CRC and late-stage adenoma namely measurement of markers in faecal samples and flexible sigmoidoscopy (FS) [40].

FOBT

The current screening test used in England and Wales is the Guaiac Faecal Occult Blood test (gFOBT). The gFOBT test is based on the fact that faecal matter passes through the colon to the rectum and it can disturb the surface of a tumour or polyp and cause a small amount of blood to be present in faeces.

Unfortunately, gFOBT is susceptible to false positive results as the diet of a patient can affect results. The guaiac based test is not human-specific and any heme present in a sample or anything containing a peroxidase could cause a false positive result. Since there is haeme in red meat and some fruits and vegetables (e.g. radishes and broccoli) contain peroxidase, dietary restrictions have to imposed on anyone taking the test, this can be inconvenient for the participant and difficult to be certain of compliance. Also, it has been shown that taking vitamin C tablets or eating foods with high levels of vitamin C can give false positive gFOBT results [41]. These problems with gFOBT are solved using immunochemical faecal occult blood tests (iFOBT). In general, iFOBT tests detect human specific haemoglobin in faecal samples using monoclonal antibodies.

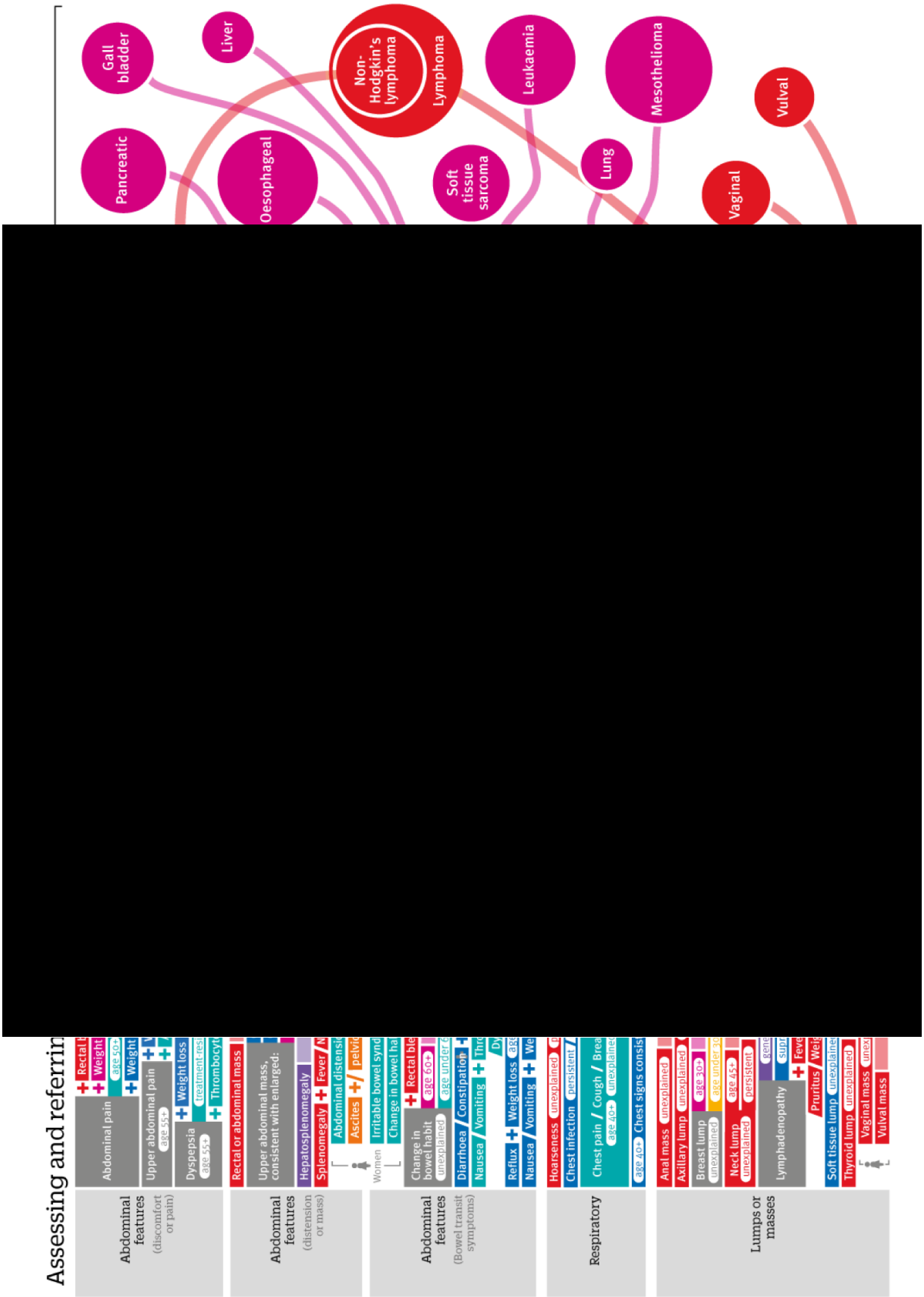


Figure 1.11: Schematic of the current referral pathways to cancer through symptomatic presentation in the UK.

Investigations and meta-analysis studies have shown iFOBT to have improved sensitivity in the detection of colorectal cancers and late-stage adenomas in both high-risk and average-risk populations compared to gFOBT without compromising on specificity [42], [43], [44]. However, this test still relies on screening faecal samples which may be in part responsible for the nationally low uptake of CRC screening in England and Wales. Furthermore, this test is not able to be performed on patients with rectal bleeding which leaves it unable to cover all patients.

Flexible sigmoidoscopy (FS)

An alternative screening method is flexible sigmoidoscopy. This is a procedure carried out using an endoscope passed through a patient's anus. The endoscope is used to directly visualise the mucosa for polyps and malignant lesions. To perform the procedure air or carbon dioxide is introduced into the intestine and a small camera mounted into the scope is used to relay images back to an external screen. FS examines only part of the large bowel (sigmoid and rectum). The test takes around 10-20 mins and requires a healthcare professional to perform the procedure. In some cases a sedative may need to be administered to the patient which carries a risk of respiratory depression and aspiration pneumonia. The procedure can cause cramping and some discomfort in patients in which case carbon dioxide is used to try and ease pain or discomfort. In approximately 1 in 2000 cases FS causes perforation in the colon and surgery is required to fix this [45]. FS also requires bowel preparation before the procedure; for optimum results the bowel must be empty. Generally, patients need to stop eating solid food one day previous to having the procedure and in some cases they may have to have an enema. A UK study into the effectiveness of a single FS procedure as a screening tool was conducted by Atkin et al (2010). With just one procedure for patients aged 55-64 years the study saw a 33% reduction in CRC cases and a 43% mortality reduction, however this is only applicable with respect to the distal colon [46]. The cost of FS is much higher than tests that look for markers in faecal samples mainly because of trained staff time [47]. Moreover, the results of tests reliant on the opinion of a practitioner can be subjective depending on

the experience of the practitioner.

1.5.2 Precursor lesions and the detection of precursor lesions

The heterogeneity of CRCs is reflected in the precursor lesions (polyps) that develop before a cancer. As an example, Figure 1.6 showed two examples of different polyps aetiology that result in different forms of CRC (adenoma and sessile serrated polyps). The current methods of polyp detection are the same as for CRC (faecal testing, endoscopic procedures and imaging methods). However, these methods are not optimised for detecting all types of polyps. As an example, the sessile serrated polyps are flattened lesions which can be harder to spot via endoscopic procedures and may require dyes and narrow band imaging techniques. Flattened polyps are also less likely to bleed than an adenomatous polyp and hence be detected via faecal testing methods, making them harder to detect than an adenomatous polyp. It is the hope of this work that the different expression of genes and proteins of polyps from different origins may cause different metabolites to be released into the blood when a polyp is developing. The use of Raman spectroscopy as a tool to detect metabolic changes in blood samples could potentially circumvent the current issues in detecting precursor lesions by detecting the differences in metabolic profile in blood samples compared to visual inspection.

1.5.3 Other screening tests available outside of the UK

Outside of the United Kingdom there are other diagnostic tests are also in use such as the tumour M2-pyruvate kinase (tumour M2-PK) test. This is a faecal test that is more sensitive than FOBT [48]. The cost effectiveness and the sensitivity and specificity of M2-PK in comparison to the iFOBT needs to be established in the United Kingdom as published results conflicting [49]. There is also a real time polymerase chain reaction based blood test that is available outside of the United Kingdom that detects methylated Septin 9 (mSept9). This blood test has sensitivity and specificity ranging from 50%-90% and 88%-91%, respectively [50].

A blood test is potentially more attractive option for patients compared to faecal and colonoscopy tests so studies are underway to determine if the higher cost of mSept9 would be recovered by higher screening uptake [51]. However, PCR based tests are currently still significantly more expensive than the current available options [52]. Figure 1.12 shows a comparison of different blood based tests for CRC that are currently being developed as an alternative to screening. The methods in development generally rely on PCR and ELISA based platforms. There techniques still all have a minimum cost per test of $91.

Test Name	Company	Cost per test ($)	Method	Marker	Availability/Regulatory
Epi proColon	Epigenomics	91-141	qRT-PCR	methylated Septin 9 DNA	CE (EU), FDA approved in April 2016
ColonSentry	Gene News	350	qRT-PCR	7 gene Panel	Available in 1 CLIA approved in US
ColoMarker	EDP Biotech	95	ELISA	CA11-19	CE (EU)
[redacted]					end 2016
SimpliPRO	Applied Proteomics	448	Proprietary Mass Spectrometry platform	11-protein biomarker panel	Own CLIA Approved Lab

Table 7: CRC Blood Based Tests (Data source: Evolution modelling of Company Announcements)

Figure 1.12: A comparison of other screening tests being developed for the colorectal cancer screening market. Taken from [52].

1.6 Investigating and diagnosing colorectal cancer

Once a patient has either presented symptomatically or had a positive screening test they will typically then be referred for further diagnostic tests under the NICE treatment pathway (Figure 1.13). Currently, time between a patient with suspected cancer symptoms being referred by their GP and receiving definitive treatment is 62 days [33]. Once a patient is in secondary care under the referral pathway the patient will be assigned to a consultant who will commission further investigations and diagnostic tests.

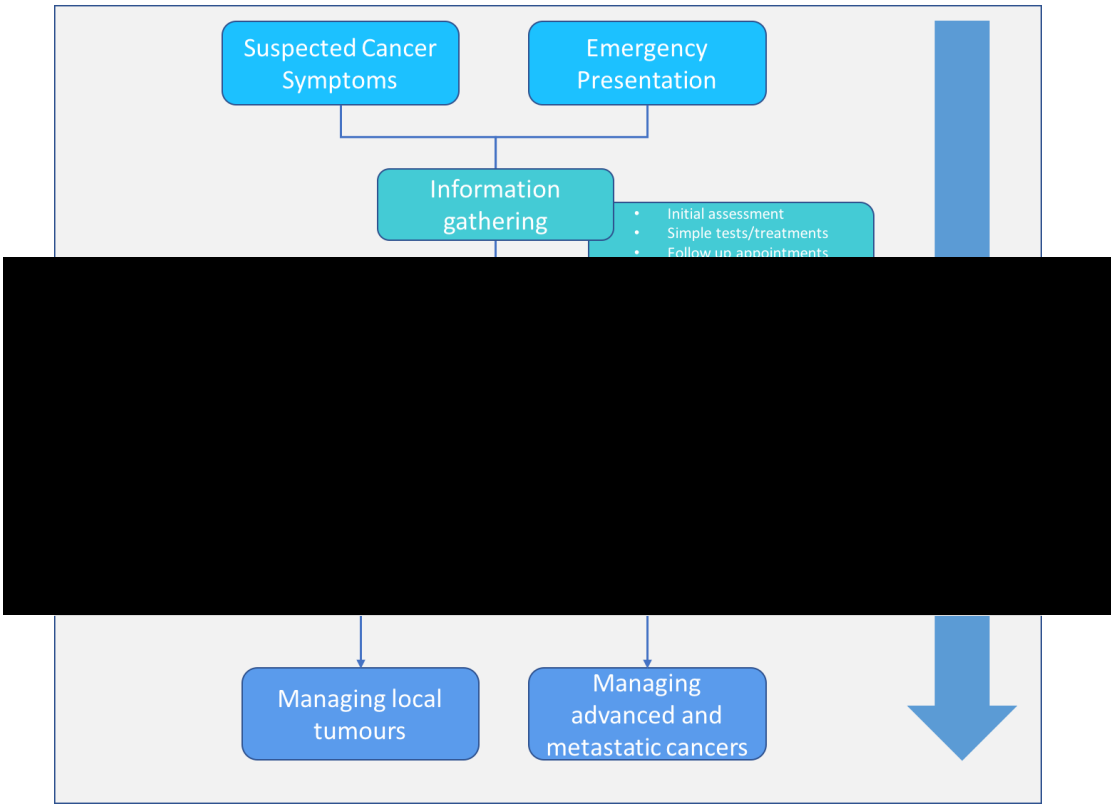


Figure 1.13: Colorectal treatment pathway for patients referred with suspected CRC. Adapted from [12].

The 'gold standard' secondary care diagnostic is colonoscopy [33]. This has sensitivity and specificity of around 90-95% [53]. Colonoscopy is a similar technique to FS, however, the entire colon is investigated. As with FS, colonoscopy

is an invasive procedure that requires highly skilled staff to perform the test. Colonoscopy also carries similar risks to FS of bowel perforations and infections as well as the potential of the result to be considered as practitioner dependent. Other diagnostic tests used in the UK include CT colonography and Barium enema. CT colonography has sensitivities and specificity of around 89% and 75% respectively [54].

During colonoscopy any suspicious tissue is biopsied for histopathology. Histopathological analysis of tissue biopsies is remains the gold standard for the diagnosis of malignancy. Biopsies of tissue are fixed, usually using formalin, and then this sections are cut and mounted onto glass slides. These sections are stained using various methods to determine TNM stage, tumour type, histologic grade and the level of vascular invasion. However, histopathology is a labour intensive process adding more time to the treatment planning process. Once a cancer has been diagnosed, and staged it is treated via the pathway in Figure 1.13.

Due to there being a reliance on symptoms and screening tests for referral many patents are referred to secondary care to 'rule out' cancer. The national bowel cancer audit (NBOCA) found that 55% of patients referred to secondary care were found to have cancer. Bowles et al found in a study of over 9000 colonoscopies that 40.4% reported all-clear results [55]. The large number of 'normal' patients who are undergoing these diagnostic tests leads to increased patient anxiety, longer waiting times for patients to get diagnosed and the associated poorer clinical outcomes.

It should be noted that there is currently a validated blood test available to clinicians associated to CRC (Carcinoembryonic antigen - CEA test). The blood test measures the levels of the CEA antigen in blood serum. Elevated serum levels of CEA can be detected in different malignant diseases. These include colorectal, pancreatic, gastric, lung and breast cancer. It is also observed in healthy heavy smokers, and in certain benign diseases such as diabetes, ulcerative colitis, pancreatitis and liver cirrhosis. CEA measurement is commonly used in clinical practice as part of follow up after a curative resection for CRC. It is used in conjunction with clinical evaluation, and radiological and colonoscopy examination. CEA is not specific for CRC and previous studies have shown that

CEA has poor sensitivity and specificity for early disease detection and therefore is not recommended for early detection or screening. Therefore, despite it being a straightforward blood test it does not fit into the current screening algorithm in the UK for CRC [56, 57].

1.6.1 Conclusion

In the UK there is a reliance on symptomatic presentation and screening to refer patients to secondary care for CRC diagnosis. Current screening tests available have poor patient uptake despite improved clinical outcomes. Despite patients referred via screening having a better outcome less than 10% of UK CRCs are diagnosed via that route [34]. This means a large number of patients get referred under the USC pathway for colorectal cancer who don't necessarily have cancer. On average globally the number of 'negative' finding colonoscopies is 63% [52]. This large proportion of negative colonoscopies causes a bottleneck of patients entering the waiting list for colonoscopy leading to increased costs, increases waiting times, and potential delays in diagnosis causing poorer outcomes. Despite the development of some alternative screening and diagnostic tests there is still a need to triage the colorectal referrals. The need for a triage tool for colorectal referrals has recently been highlighted as a global urgent clinical need [58]. The work in the rest of this thesis addresses the proposition that a Raman spectroscopy based blood test has the potential to provide more information to a GP to inform patient referral decisions.

Bibliography

[1] Cancer Research UK. Bowel cancer statistics, September 2016.

[2] Cancer Research UK. http://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts/types-of-cancer, 2016.

[3] Sam Al-Sohaily, Andrew Biankin, Rupert Leong, Maija Kohonen-Corish, and Janindra Warusavitarne. Molecular pathways in colorectal cancer. *J. Gastroenterol. Hepatol.*, 27:1423–31, 2012.

[4] Matthew Fleming, Sreelakshmi Ravula, Sergei F Tatishchev, and Hanlin L Wang. Colorectal carcinoma: Pathologic aspects. *J. Gastrointest. Oncol.*, 3(3):153–73, sep 2012.

[5] Seer training modules, anatomy of the colon and rectum, national cancer insitute.

[6] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5):E359–E386, 2015.

[7] P. Boyle and J. S. Langman. Abc of colorectal cancer: Epidemiology. *BMJ (Clinical research ed.)*, 321(7264):805–808, Sep 2000.

[8] A. B. Wilmink. Overview of the epidemiology of colorectal cancer. *Diseases of the colon and rectum*, 40(4):483–493, Apr 1997.

[9] Fatima A. Haggar and Robin P. Boushey. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(4):191–197, Nov 2009.

[10] D M Parkin, L Boyd, and L C Walker. 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *Br. J. Cancer*, 105 Suppl(S2):S77–81, dec 2011.

[11] Stanley R Hamilton and Lauri A Aaltonen. World Health Organization Classification of Tumours Pathology and Genetics of Edited by. 2000.

[12] Office for National Statistics. Cancer registration statistics england 2013, 10 July 2015.

[13] William Hamilton, Robert Lancashire, Debbie Sharp, Tim Peters, KK Cheng, and Tom Marshall. The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. *BMC Medicine*, 7(1):17, 2009.

[14] Stanley R Hamilton, Lauri A Aaltonen, International Agency for Research on Cancer, World Health Organization, et al. *Pathology and genetics of tumours of the digestive system*. IARC press Lyon, 2000.

[15] Elena Stoffel, Bhramar Mukherjee, Victoria M Raymond, Fay Kastrinos, Jennifer Sparr, Fei Wang, Sapna Syngal, and Stephen B Gruber. NIH Public Access. 137(5):1621–1627, 2010.

[16] Elizabeth Half, Dani Bercovich, and Paul Rozen. Familial adenomatous polyposis. *Orphanet J. Rare Dis.*, 4:22, 2009.

[17] Kavitha P Raj, Thomas H Taylor, Charlie Wray, Michael J Stamos, Jason A Zell, et al. Risk of second primary colorectal cancer among colorectal cancer cases: a population-based analysis. *Journal of carcinogenesis*, 10(1):6, 2011.

[18] Jérémie Jégu, Marc Colonna, Laetitia Daubisse-Marliac, Brigitte Trétarre, Olivier Ganry, Anne-Valérie Guizard, Simona Bara, Xavier Troussard, Véronique Bouvier, Anne-Sophie Woronoff, et al. The effect of patient characteristics on second primary cancer risk in france. *BMC cancer*, 14(1):1, 2014.

[19] Jeremy M Berg, John L Tymoczko, and Lubert Stryer. Biochemistry. 5th, 2002.

[20] Eric R. Fearon and Bert Vogelstein. A genetic model for colorectal tumorigenesis, 1990.

[21] Maria S Pino and Daniel C Chung. THE CHROMOSOMAL INSTABILITY PATHWAY IN COLON. 138(6):2059–2072, 2014.

[22] Kenneth W Kinzler and Bert Vogelstein. Lessons from hereditary colorectal cancer. *Cell*, 87(2):159 – 170, 1996.

[23] Justin Jong Leong Wong, Nicholas John Hawkins, and Robyn Lynne Ward. Colorectal cancer: a model for epigenetic tumorigenesis. *Gut*, 56(1):140–148, 2007.

[24] Daniel J Weisenberger, Kimberly D Siegmund, Mihaela Campan, Joanne Young, Tiffany I Long, Mark A Faasse, Gyeong Hoon Kang, Martin Widschwendter, Deborah Weener, Daniel Buchanan, et al. Cpg island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with braf mutation in colorectal cancer. *Nature genetics*, 38(7):787–793, 2006.

[25] Daniel L Worthley and Barbara A Leggett. Colorectal Cancer : Molecular Features and Clinical Opportunities. 31(May):31–38, 2010.

[26] Richard C Boland and Ajay Goel. Microsatellite Instability in Colorectal Cancer. *Gastroenterology*, 138(6):2073–2087, 2010.

[27] J R Jass. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*, 50(1):113–30, jan 2007.

[28] Enric Domingo, David N. Church, Oliver Sieber, Rajarajan Ramamoorthy, Yoko Yanagisawa, Elaine Johnstone, Brian Davidson, David J. Kerr, Ian P.M. Tomlinson, and Rachel Midgley. Evaluation of PIK3CA mutation as a predictor of benefit from nonsteroidal anti-inflammatory drug therapy in colorectal cancer. *J. Clin. Oncol.*, 31(34):4297–4305, 2013.

[29] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien de Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift

Nyamundanda, Paolo Angelino, Brian M Bot, Jeffrey S Morris, Iris M Simon, Sarah Gerster, Evelyn Fessler, Felipe De Sousa E Melo, Edoardo Missiaglia, Hena Ramay, David Barras, Krisztian Homicsko, Dipen Maru, Ganiraju C Manyam, Bradley Broom, Valerie Boige, Beatriz Perez-Villamil, Ted Laderas, Ramon Salazar, Joe W Gray, Douglas Hanahan, Josep Tabernero, Rene Bernards, Stephen H Friend, Pierre Laurent-Puig, Jan Paul Medema, Anguraj Sadanandam, Lodewyk Wessels, Mauro Delorenzi, Scott Kopetz, Louis Vermeulen, and Sabine Tejpar. The consensus molecular subtypes of colorectal cancer. *Nat. Med.*, 21(11):1350–1356, 2015.

[30] Rodrigo Dienstmann, Louis Vermeulen, Justin Guinney, Scott Kopetz, Sabine Tejpar, and Josep Tabernero. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer*, 2017.

[31] Anthony S Fauci et al. *Harrison's principles of internal medicine*, volume 2. McGraw-Hill, Medical Publishing Division, 2008.

[32] UK Government. Direct access to diagnostic tests for cancer. best practise referral pathways for gps. DH website, April 2012.

[33] National Institute for Health and Care Excellence from 1 April 2013.

[34] UK government. National bowel cancer audit report 2017, 2017.

[35] Jacqueline Barrett, Moyez Jiwa, Peter Rose, and William Hamilton. Pathways to the diagnosis of colorectal cancer: an observational study in three uk cities. *Family practice*, 23(1):15–19, 2006.

[36] Margaret Astin, Tom Griffin, Richard D Neal, Peter Rose, and William Hamilton. The diagnostic value of symptoms for colorectal cancer in primary care: a systematic review. *The British journal of general practice : the journal of the Royal College of General Practitioners*, 61(586):e231–e243, 2011.

[37] S Stapley, TJ Peters, D Sharp, and W Hamilton. The mortality of colorectal cancer in relation to the initial symptom at presentation to primary care and

to the duration of symptoms: a cohort study using medical records. *British journal of cancer*, 95(10):1321–1325, 2006.

[38] W Hamilton, A Round, D Sharp, and TJ Peters. Clinical features of colorectal cancer before diagnosis: a population-based case–control study. *British journal of cancer*, 93(4):399–405, 2005.

[39] Jessica B O'Connell, Melinda A Maggard, and Clifford Y Ko. Colon cancer survival rates with the new american joint committee on cancer sixth edition staging. *Journal of the National Cancer Institute*, 96(19):1420–1425, 2004.

[40] Michael J Duffy, Leo G M van Rossum, Sietze T van Turenhout, Outi Malminiemi, Catherine Sturgeon, Rolf Lamerz, Andrea Nicolini, Caj Haglund, Lubos Holubec, Callum G Fraser, and Stephen P Halloran. Use of faecal markers in screening for colorectal neoplasia: a European group on tumor markers position paper. *Int. J. Cancer*, 128(1):3–11, jan 2011.

[41] Immunochemical Fecal, Occult Blood Tests, Stool Screening, and Using Molecular Markers. Emerging Technologies in Screening for Colorectal Cancer : CT Colonography , Immunochemical Fecal. pages 44–55.

[42] Ming Ming Zhu, Xi Tao Xu, NIE Fang, Jin Lu Tong, Shu Dong Xiao, and Zhi Hua Ran. Comparison of immunochemical and guaiac-based fecal occult blood test in screening and surveillance for advanced colorectal neoplasms: A meta-analysis. *Journal of digestive diseases*, 11(3):148–160, 2010.

[43] Adolfo Parra-Blanco, Antonio Z Gimeno-García, Enrique Quintero, David Nicolás, Santiago G Moreno, Alejandro Jiménez, Manuel Hernández-Guerra, Marta Carrillo-Palau, Yoshinobu Eishi, and Julio López-Bastida. Diagnostic accuracy of immunochemical versus guaiac faecal occult blood tests for colorectal cancer screening. *Journal of gastroenterology*, 45(7):703–712, 2010.

[44] Dong Il Park, Seungho Ryu, Young-Ho Kim, Suck-Ho Lee, Chang Kyun Lee, Chang Soo Eun, and Dong Soo Han. Comparison of guaiac-based and quantitative immunochemical fecal occult blood testing in a population at

average risk undergoing colorectal cancer screening. *Am. J. Gastroenterol.*, 105(9):2017–25, sep 2010.

[45] Varut Lohsiriwat. Colonoscopic perforation: Incidence, risk factors, management and outcome. *World J. Gastroenterol.*, 16(4):425–430, 2010.

[46] Wendy S Atkin, Rob Edwards, Ines Kralj-Hans, Kate Wooldrage, Andrew R Hart, John M a Northover, D Max Parkin, Jane Wardle, Stephen W Duffy, and Jack Cuzick. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet*, 375(9726):1624–33, may 2010.

[47] J A Burch, K. Soares-Weiser, D J B St John, S. Duffy, S. Smith, J. Kleijnen, and M. Westwood. Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: a systematic review. *J. Med. Screen.*, 14(3):132–137, 2007.

[48] U Haug, D Rothenbacher, MN Wente, CM Seiler, C Stegmaier, and H Brenner. Tumour m2-pk as a stool marker for colorectal cancer: comparative analysis in a large sample of unselected older adults vs colorectal cancer patients. *British journal of cancer*, 96(9):1329–1334, 2007.

[49] Suresh Sithambaram, Ida Hilmi, and Khean Lee Goh. The diagnostic accuracy of the M2 pyruvate kinase Quick stool test-A rapid office based assay test for the detection of colorectal cancer. *PLoS One*, 10(7), 2015.

[50] Hye Seung Lee, Sang Mee Hwang, Taek Soo Kim, Duck-Woo Kim, Do Joong Park, Sung-Bum Kang, Hyung-Ho Kim, and Kyoung Un Park. Circulating methylated septin 9 nucleic Acid in the plasma of patients with gastrointestinal cancer in the stomach and colon. *Transl. Oncol.*, 6(3):290–6, 2013.

[51] Uri Ladabaum, John Allen, Michael Wandell, and Scott Ramsey. Colorectal cancer screening with blood-based biomarkers: Cost-effectiveness of methylated septin 9 DNA versus current strategies. *Cancer Epidemiol. Biomarkers Prev.*, 22(9):1567–1576, 2013.

[52] Frank Rinaldi. Global market assessment for Raman spectroscopy and colorectal cancer., 2017.

[53] Hooi C Ee, James B Semmens, Neville E Hoffman, et al. Complete colonoscopy rarely misses cancer. *Gastrointestinal endoscopy*, 55(2):167–171, 2002.

[54] PFC Lung, D Burling, L Kallarackel, J Muckian, R Ilangovan, A Gupta, M Marshall, P Shorvon, S Halligan, G Bhatnagar, et al. Implementation of a new ct colonography service: 5 year experience. *Clinical radiology*, 69(6):597–605, 2014.

[55] C J A Bowles, R Leicester, C Romaya, E Swarbrick, C B Williams, and O Epstein. A prospective study of colonoscopy practice in the UK today: are we adequately prepared for national colorectal cancer screening tomorrow? *Gut*, 53(2):277–83, 2004.

[56] Healthcare Improvement Scotland. Scottish Referral Guidelines for Suspected Cancer. *Scottish Exec.*, (August):1–54, 2014.

[57] D S Thomas, E-o Fourkala, S Apostolidou, R Gunu, A Ryan, I Jacobs, U Menon, and W Alderton. Evaluation of serum CEA , CYFRA21-1 and CA125 for the early detection of colorectal cancer using longitudinal preclinical samples. *Br. J. Cancer*, 113(2):268–274, 2015.

[58] Mark Lawler, Deborah Alsina, Richard A. Adams, Annie S. Anderson, Gina Brown, Nicola S. Fearnhead, Stephen W. Fenwick, Stephen P. Halloran, Daniel Hochhauser, Mark A. Hull, Viktor H. Koelzer, Angus G.K. McNair, Kevin J. Monahan, Inke Näthke, Christine Norton, Marco R. Novelli, Robert J.C. Steele, Anne L. Thomas, Lisa M. Wilde, Richard H. Wilson, and Ian Tomlinson. Critical research gaps and recommendations to inform research prioritisation for more effective prevention and improved outcomes in colorectal cancer. *Gut*, 67(1):179–193, 2018.

Chapter 2

Vibrational Spectroscopy

Spectroscopy is the study of the interactions between matter and electromagnetic (EM) radiation. Molecules that make up matter can be described in terms of their total internal energy. The total internal energy of a molecule can be resolved into the sum of the rotational, vibrational and electronic energy states. Vibrational spectroscopy is therefore the study of the interaction between EM radiation and the vibrational states of matter. This chapter will provide a summary of the theory behind two types of vibrational spectroscopy that have potential for use in biofluid analysis. It will discuss the theoretical basis of Raman spectroscopy and Fourier transform infrared spectroscopy (FTIR). The enhancement effects of surface enhanced Raman spectroscopy (SERS) and resonance Raman spectroscopy (RRS) will also be summarised briefly. This chapter will then provide a literature review on the current ways in which Raman and FTIR spectroscopy are being used to diagnose colorectal cancer focusing particularly on the evolution from SERS techniques back to basic Raman and FTIR based techniques.

2.1 Raman spectroscopy

Raman spectroscopy allows the user to gain molecular information about a sample through the scattering of incident light. In general, when light is passed through or onto a sample a small proportion of the photons are scattered. The majority of this is Rayleigh or elastic scattering; where the energy of the incoming photon is equal to the energy of the scattered photon (Figure 2.1)[23]. Around 1 in 10^7 of the incident photons are in-elastically scattered resulting in the incident photon and the scattered photon having a difference in energy. The inelastic scattering is a relatively weak effect which was first observed in 1928 by Sir CV Raman and is known as Raman scattering [24]. When scattered light is measured with a

spectrometer a series of lines are observed, the shift in the energy [measured by wavenumber (cm^{-1})] from the Rayleigh line (equal to incident energy) is known as the Raman shift. The shift recorded corresponds to specific vibrational or rotational modes of the molecule.

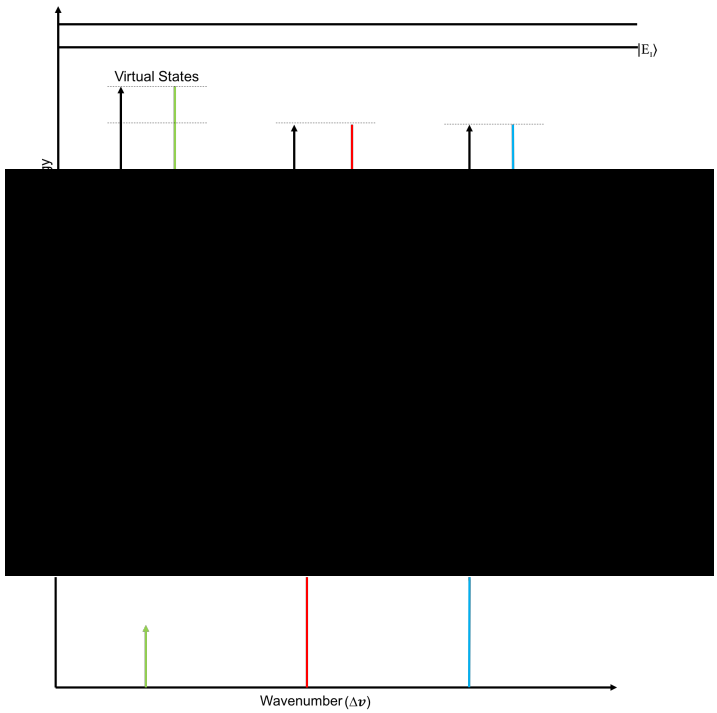


Figure 2.1: Different types of light scattering and the corresponding lines on a theoretical spectrum. Adapted from [1]

2.1.1 A Raman spectrum

Data collected from a Raman spectrometer usually is plotted as Raman shift ($\boldsymbol{\Delta\nu}$) against intensity of the scattered radiation in a spectrum. The incident EM radiation is normally from a laser, the part of the EM spectrum that the laser emits at is measured in terms of wavelength (λ). The relationship between wavenumber of the frequency shift and wavelength is

$$\boldsymbol{\Delta\nu} \propto \frac{1}{\lambda_{Raleigh}} - \frac{1}{\lambda_{Raman}}. \tag{2.1.1}$$

Figure 2.2 shows an example Raman spectrum of carbon disulphide around the excitation frequency of the laser light taken at different temperatures. It is

clear that the intensities of the peaks are different for all three frequencies. The

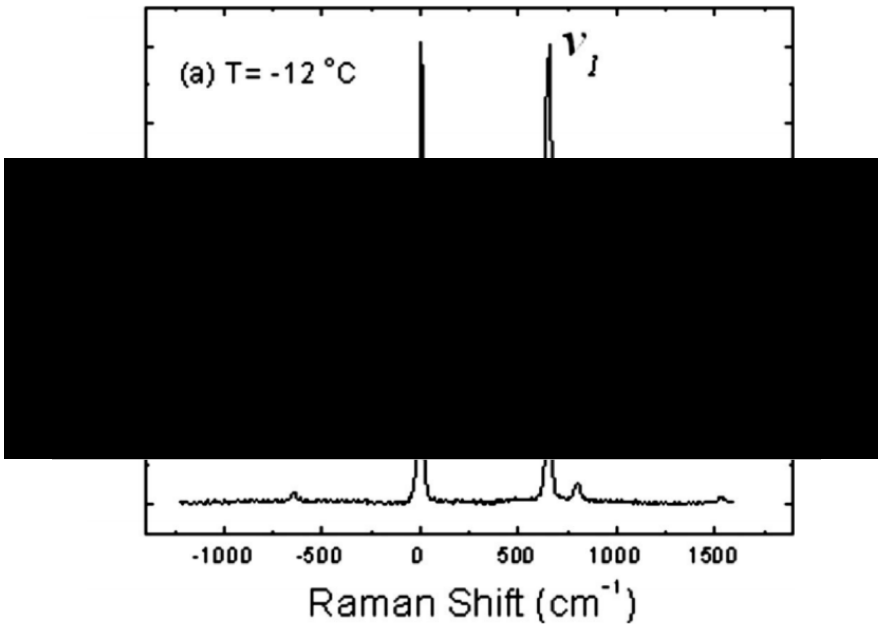


Figure 2.2: Raman spectrum of carbon disulphide taken at (a) -12°C and (b) 45°C. Figure is adapted from [2]

Rayleigh line is the most intense, followed by the the Stokes, with the Anti-Stokes Raman line the least intense. To fully explain these differences in intensity a quantum mechanical (QM) treatment of the interaction between the incident EM radiation and the atoms/molecules in a sample needs to be considered. Briefly, in QM the probability of a transition from one vibrational state to another (n $\rightarrow$ m) is given by the integral,

$$\vec{P}_{nm} = \langle \psi_m^* | \hat{\Omega} | \psi_m \rangle, \tag{2.1.2}$$

where $|\psi_m\rangle$ and $|\psi_n\rangle$ are wavefunctions of the vibrational states and n and m and $\hat{\Omega}$ is the operator that describes the perturbation or distortion of the electron cloud of the molecule by EM radiation. This operator has different properties depending on the type of perturbation and, therefore, can describe many different vibrational phenomena including Raman scattering and infrared absorption

intensities. For a full QM treatment of both Raman scattering and infrared absorption the reader is directed to [3], [4].

Some intuition about the cause of the difference between Stokes and anti-Stokes intensities without the full QM treatment can be gained. Consider Figure 2.1, for there to be an anti-Stokes transition the atoms in a sample must be in the first excited state rather than the ground state. The probability of an electron being in the first excited state in a molecule is given by the Maxwell-Boltzmann distribution

$$N_m(\epsilon, \nu, J) = \frac{\sum_m N_m}{Z} g_m \exp \frac{-E_m}{KT}, \tag{2.1.3}$$

where N_m is the number of electrons (population density) according to the electronic (ϵ), vibrational (ν)and rotational (J) states of the molecule. Z is the atomic number, K is the Boltzmann constant and T is the temperature. The sum over g_m is a statistical weight factor and E_m is the first excited state of an atom assuming the transition between vibrational modes would go (n$\rightarrow$m). At room temperature the probability of electrons being in the first excited state is lower than the probability of them being in the ground state which is the condition for Stokes scattering. Therefore the probability of there being a higher rate of Stokes scattering at room temperature is higher by a factor of exp $\frac{-E_m}{KT}$. Therefore it makes sense that a spectrum taken at room temperature would have higher Stokes-Raman scattering. This is also demonstrated in Figure 2.2, the ratio between Stokes and anti-Stokes lines changes according to temperature. Despite this, the Stokes Raman line always has a higher intensity at room temperature. Therefore, the Raman system used in this work only considers the Stokes-Raman scattering.

2.2 Classical Theory of Raman Scattering

The position of the shifted lines on a Raman spectrum can be explained using a classical description of Raman scattering. When radiation from an EM source is incident on a sample consisting of atoms, the EM radiation causes a distortion

of the electron cloud surrounding the atoms. This distortion creates oscillating dipoles which can themselves emit EM radiation. The incident EM radiation can be described via a harmonic oscillator

$$\vec{E} = \vec{E}_0 \cos(2\pi\nu_0 t), \tag{2.2.4}$$

where $\vec{E}$ is the electric field vector, and ν_0 is the initial frequency of the electric field[1]. The induced dipole oscillation in the atoms can then be described by

$$\tilde{\mu}_{induced} = \tilde{\alpha}(\nu) \cdot \vec{E}, \tag{2.2.5}$$

where $\tilde{\alpha}(\nu)$ is the polarizability tensor for the atoms. This is a measure of how easily an atom/molecule can be polarized along any direction. The polarizability tensor is time dependent as the distorted electron cloud varies with time in response to the nuclei of the atoms oscillating at a normal frequency ν_k[2]. By substituting 2.2.4 into 2.2.5 we obtain

$$\tilde{\mu}_{induced} = \tilde{\alpha}(\nu) \cdot \vec{E}_0 \cos(2\pi\nu_0 t). \tag{2.2.6}$$

The oscillating nuclei vibrate around a coordinate system Q, where Q_k is the normal coordinate for the vibration. As we are assuming that the whole system of the interaction is acting as a harmonic oscillator so it is possible to do a Taylor Series expansion on the polarizability tensor

$$\tilde{\alpha}(\nu) = \tilde{\alpha}_0 + \sum_k \left(\frac{\delta\tilde{\alpha}}{\delta Q_k}\right)_0 Q_k + \dots. \tag{2.2.7}$$

Focusing on the normal mode on one vibration (Q_k), equation 2.2.7 can be written in the form

$$\tilde{\alpha}_k = \tilde{\alpha}_0 + \tilde{\alpha}'_k Q_k, \tag{2.2.8}$$

[1]It is also common to express this in terms of angular frequency, (ω) where $\omega = 2\pi\nu$.

[2]This classical treatment assumes that the particle scattered or re-emitted is allowed to vibrate but not rotate freely in space so is therefore a 'fixed' vibration around some coordinate axis.

where $\tilde{\alpha}'_k$ is the derived polarizability tensor which has coordinates w.r.t. Q_k. Assuming that the vibration Q_k acts as a harmonic oscillator, the displacement of the atoms under the normal vibration can be described by

$$Q_k = Q_{k_0} \cos(2\pi\nu_k t). \tag{2.2.9}$$

A combination of equation 2.2.9 and 2.2.7 substituted into 2.2.8 yields that the induced oscillating dipole is given by

$$\tilde{\mu}_{induced} = \tilde{\alpha}_k \vec{E}_0 \cos(2\pi\nu_0 t) + \frac{\tilde{\alpha}'_k}{2}\Big[\vec{E}_0 Q_{k_0} \cos[(2\pi t(\nu_0 - \nu k)] + \vec{E}_0 Q_{k_0} \cos[2\pi t(\nu_0 + \nu k)]\Big]. \tag{2.2.10}$$

Evaluating equation 2.2.10 it is clear that the induced polarizability depends on three frequencies ν_0, $(\nu_0 - \nu_k)$ and $(\nu_0 + \nu_k)$ where ν_0 is the frequency of the incident EM radiation. Therefore, it corresponds to Raleigh scattering. The frequencies $(\nu_0 - \nu_k)$ and $(\nu_0 + \nu_k)$ correspond to Stokes Raman and Anti-Stokes Raman, respectively. In terms of a spectrum the lines on either side of the Raleigh line are then the Stokes and Anti-Stokes shifted light. The position of the lines is then governed by the size of the shift $(\Delta\nu)$ which is specific to the bond being interrogated. Therefore Raman spectroscopy can be used to study the structure of molecular bonds. It is also clear from equation 2.2.10 that Stokes Raman and Anti-Stokes Raman scattering are only possible when $\tilde{\alpha}' \neq 0$. This leads to the gross Raman selection rule that Raman scattering is only possible when there is a change in polarizability of the molecule.

2.2.1 Enhanced Raman Spectroscopy

Raman spectroscopy is an inherently weak technique with only 1 in 10^7 photons being scattered when incident on a single molecule in normal Raman scattering. However, methods have been developed which amplify the Raman intensity recorded by a spectrometer. One of the simplest ways to achieve enhancement is to use the electronic molecular structure of molecules.

Resonance Raman

When the incident EM radiation on a sample is close to or equal to the electronic transition energy of a molecule it is possible to achieve resonant Raman (RR) effects. There are three categories of RR; near, resonant and continuum. These three are achieved when the excitation radiation is near but slightly less than the electronic transition energy, equal to the transition energy or higher than the transition energy (Figure 2.3), respectively. The recorded spectrum can achieve

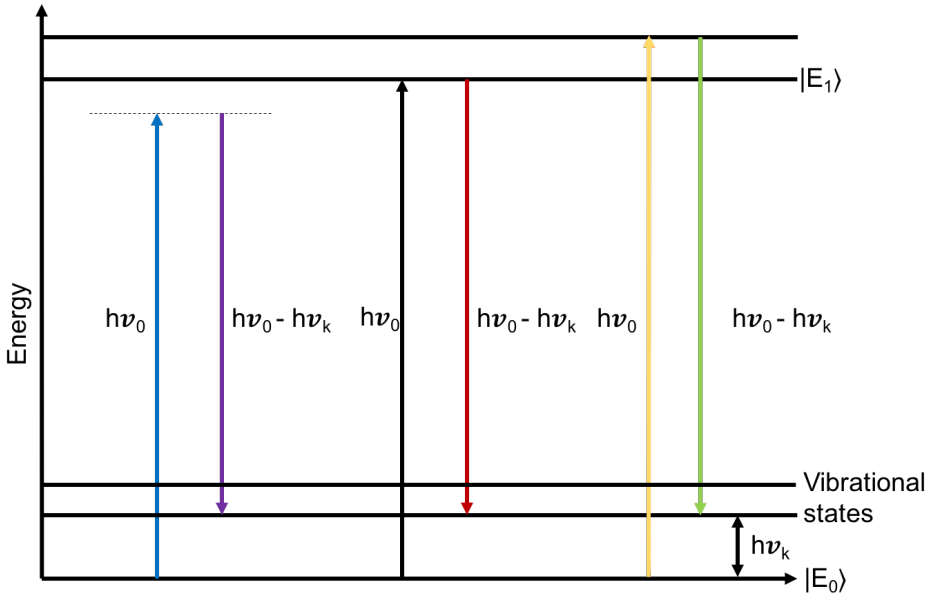


Figure 2.3: Resonance Raman Scattering occurs when the incoming EM radiation is near to or equal to the electronic energy states of an atom.

up to a $10^2 - 10^6$ rise in intensity compared to normal Raman [5]. To achieve this enhancement the laser wavelength being used to excite the molecules must be closely tuned to the energy of the electronic transition energy of the molecule. This has two consequences, firstly it can increase the likelihood of getting a fluorescence transition instead of a Raman transition, and secondly it means that the effect is highly wavelength and therefore laser dependant. This can be problematic: if the analyte is a mixture of molecules then you can only have an excitation wavelength that is resonant to one of the molecules. This means that spectral information molecules that are not resonant at the exciting wavelength can be masked.

SERS

A less wavelength dependent method of enhancing Raman signal is surface enhanced Raman spectroscopy (SERS). The basic principle of SERS is to amplify the Raman response of a given analyte. The SERS effect was first discovered in 1974, and understood to be an enhancement of Raman scattering in 1977 [6, 7]. This is generally achieved by having an analyte attached or close to the surface of a nanoscale metal substrate causing an enhancement factor of up to 10^{11} [8, 9]. The exact mechanism of SERS enhancement is still an area of active research, however it is generally accepted that two mechanisms contribute to the enhancement [10]. One is based on electromagnetic field enhancement due to excitation of EM resonances in the SERS active nanoscale metal substrate. The other is known as chemical enhancement which is a result of the metal electrons causing a charge transfer between the metal substrates and the adsorbates. The result of the combined enhancement is an extremely powerful technique that combines ultra sensitive detection limits with the molecular structure information from Raman spectroscopy giving the possibility of single molecule detection [11].

2.3 Infrared Absorption and Fourier transform infrared spectroscopy

If incident EM radiation on a sample is in the infrared (IR) region of the EM spectrum then the IR radiation normally interacts with the vibrational modes of molecules. This is because IR radiation has a relatively lower energy than UV or visible light which tend to interact more with the electronic states of molecules. If the molecules have a change in polarizability under the IR radiation then IR Raman scattering is possible, however absorption of IR radiation is also possible.

The gross selection rule for IR absorption to occur within a molecule being irradiated by IR light is that the molecule must undergo a change in dipole moment. The selection rule can be described by considering equation 2.1.2. In IR the transition from one vibrational state to another ($n \rightarrow m$) within a molecule is caused by the absorption of a photon in the IR region of the EM spectrum.

Assuming that the EM radiation distorts the electron cloud around the molecule, as in Raman, the radiation will cause a perturbation of the electron cloud and therefore oscillating dipoles. In the QM description, the absorption process is controlled by the dipole moment operator $\hat{\mu}_q$ where

$$\hat{\mu}_q = \sum_i e_i \cdot \hat{q}_i, \tag{2.3.11}$$

with e_i being the effective charge at some atom i and q_k being the distance to the center of the molecule in terms of cartesian coodiates (x,y,z). In QM the interaction between the incident IR radiation and the molecule can be written as a dot product between the two systems, therefore similar to the induced dipole oscillation in Raman we can write

$$\vec{P} = \hat{\mu}_q \cdot \vec{E}, \tag{2.3.12}$$

where $\vec{P}$ is the perturbation caused by the IR radiation. By averaging over all of the directions in the Cartesian coordinate system the overall intensity of a transition can be written,

$$I_{n\to m} \propto ([\mu_x]^2_{nm} + [\mu_y]^2_{nm} + [\mu_z]^2_{nm}) \tag{2.3.13}$$

where

$$[\mu_q]_{nm} = \langle \psi^*_m | \hat{\mu}_q | \psi_n \rangle. \tag{2.3.14}$$

Therefore, a transition will only occur from $n \to m$ if $\hat{\mu}_q$ is non-zero. As with the polarizability tensor discussed previously, if we assume that the system is acting as a harmonic oscillator we can expand $\hat{\mu}_q$ with respect to the normal coordinates of the molecular vibration Q_k. This gives

$$\hat{\mu}_q = \mu^0_q + \sum_{k=1}^{3n-6} \hat{\mu}^k_q \, Q_k + \dots \tag{2.3.15}$$

with

$$\hat{\mu}_q^k = \left(\frac{\delta\mu_k}{\delta Q_k}\right)_0. \tag{2.3.16}$$

The sum in the expansion differs to that of the induced dipole oscillations in Raman scattering as the sum term in equation 2.3.15. This is because the sum is dependent on the number of degrees of freedom in the molecule being interrogated[3]. The probability of an IR transition to occur and therefore IR radiation to be absorbed can be written using equation 2.1.2,

$$[\hat{\mu}_q] = \langle\psi_m^*|\hat{\mu}_q|\psi_n\rangle \tag{2.3.17}$$

$$= \mu_q^0\langle\psi_m^*|\psi_n\rangle + \sum_{k=1}^{3n-6}\hat{\mu}_q^k\langle\psi_m^*|Q_k|\psi_n\rangle. \tag{2.3.18}$$

Therefore, from 2.3.17 for a transition to occur then $\hat{\mu}_q^k \neq 0$ and the induced dipole moment must change in time with respect to the normal coordinates of the vibration. This leads to the gross selection rule for IR absorption to occur being that there must be a change in dipole moment for a transition between vibrational states to occur.

2.3.1 FTIR spectroscopy

IR light that has been absorbed by a molecule also can be recorded in terms of a spectrum, usually in the form of absorbance vs wavenumber. When a molecule absorbs in the IR region the frequency of the absorbed radiation matches the frequency of the molecular vibration. Therefore different molecules and different bond types absorb at different parts of the IR spectrum. In order to produce a spectrum across the IR region an Michelson interferometer is commonly used. The output from this is known as an interferogram. A Fourier transform of the signal from the interferometer is required to obtain a spectrum with clear peaks. Figure 2.4 shows a typical FTIR spectrum of dried human serum.

[3]The DOF in this example sum (3N-6) is the number of degrees of freedom for a non-linear molecule. In the molecule was linear then the degrees of freedom would be (3N-5) [12].

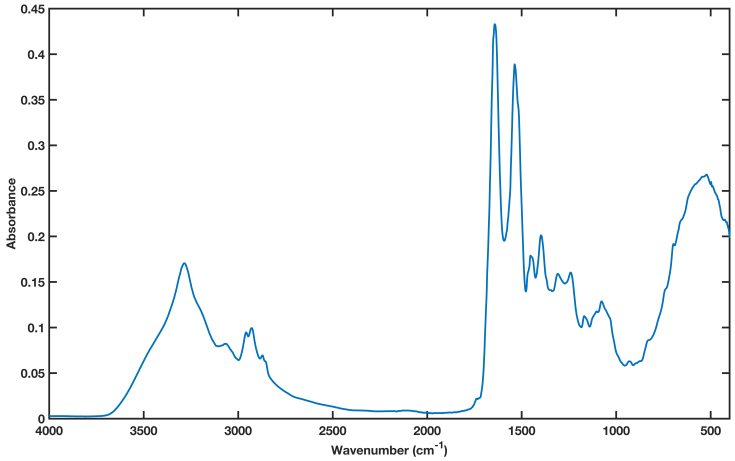


Figure 2.4: A typical FTIR spectrum of dried human serum.

The frequency of the absorption bands is proportional to the energy of the molecular transition so higher energy transitions are on the left hand side of the spectrum. To explain the proportion of light absorbed by a molecule then one must again consider an EM wave propagating through a medium,

$$A(\psi) = A_0\,(\psi)\,e^{i(\omega t-\delta)} \tag{2.3.19}$$

where A is the amplitude of the propagating wave, ψ is the polarization angle, ω is the angular frequency, δ is the phase angle and t is time [13]. This equation only holds when the medium that the light is passing through is non-absorbing; if the IR radiation passes through a molecule that absorbs some of the energy of the radiation then this equation is modified to include a complex term. When the molecule is absorbing the intensity of the EM radiation it can be written

$$I = I_0 e^{-al}, \tag{2.3.20}$$

where I is the transmitted light intensity, I_0 is the incident light intensity, a is the absorption coefficient (where $a = \frac{2\pi k}{\lambda}$)and l is the path length that the radiation passes through [14]. The Beer-Lambert law then follows directly from equation 2.3.20. According to the Beer-Lambert law the transmitted intensity from the molecule varies with the concentration of the molecules and the sample length as

$$I = I_0\,e^{-\sigma Nl}, \tag{2.3.21}$$

where σ is the absorption cross section or emissivity of the molecules and N is the number of molecules per unit volume (concentration of a sample). The amount of light absorbed by a molecule then has a linear relationship with the concentration of the molecules in the sample

$$Absorbance = \log\left(\frac{I}{I_0}\right)\sigma\, N\, l. \tag{2.3.22}$$

This means that IR absorbance can be used to determine the concentration of a molecule within a sample as well as its molecular bond information. It should be noted that the intensity of a Raman(Stokes) line on a spectrum can also be calculated in a similar manner using the Raman scattering cross-section of a molecule, the intensity is given by

$$I = I_0\, \sigma_{Raman}^{free}\, N. \tag{2.3.23}$$

In the case of Raman intensities, the Raman cross section is inversely proportional to the wavelength of the incident light

$$\sigma_{Raman} \propto \frac{1}{\lambda^4}, \tag{2.3.24}$$

therefore the intensity of the Raman scattered light,

$$I \propto \frac{I_0}{\lambda^4}. \tag{2.3.25}$$

Both Raman and FTIR spectroscopy techniques can provide the user with a wide range of information about the molecule being interrogated. Both techniques therefore lend themselves for use in characterising materials. However, correct spectral interpretation is very important, especially when considering complex biological samples.

2.4 Interpreting vibrational spectra

In general, vibrational spectra can be thought of as a 'molecular fingerprint' of the sample that is being interrogated. The spectral bands have three main properties: band shape, band intensity and the band position (frequency). Figure 2.5 summarises the spectral band features for both Raman and IR spectra.

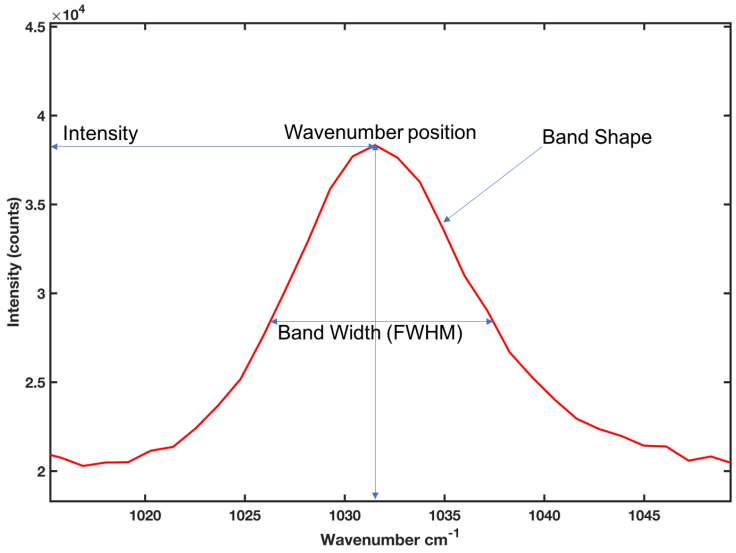


Figure 2.5: Summary of spectral band features for both Raman and FTIR.

The intensity of spectral bands for both IR and Raman can be used in order to determine the molar concentration of an analyte using the Beer-Lambert law and equation (2.3.23). The band shape is essentially a measure of the amount that the molecular vibrations within the sample are in phase with each other [15]. Vibrationally excited molecules in IR and Raman interactions relax to their ground state within a picosecond scale. This is known as the 'lifetime' of molecules. Initially the excited molecules will vibrate coherently, however differences in vibrational frequencies randomise the oscillations (dephasing), also within a picosecond scale. This is known as the 'coherence lifetime'. A spectrometer measures molecules while they are excited and coherently vibrating. The combination of the relaxing and dephasing effects is known as the effective lifetime τ. The effective lifetime is proportional to the band width and is related to the band shape parameter, the full width half maximum (FWHM), by

$$\Delta x = \frac{1}{\tau}. \tag{2.4.26}$$

The effective lifetime (τ) of the molecular vibration is highly environmentally dependant so changes in temperature, stress and the state (liquid,solid or gas) all affect the spectral band shape.

The position along the x axis that spectral bands appear are molecule dependent, the selection rules described previously govern if a molecule is IR or Raman active. As an example, consider a molecule of carbon dioxide. The molecule is linear and therefore has 4 vibrational modes. Two stretching modes and two bending modes (Figure 2.6).

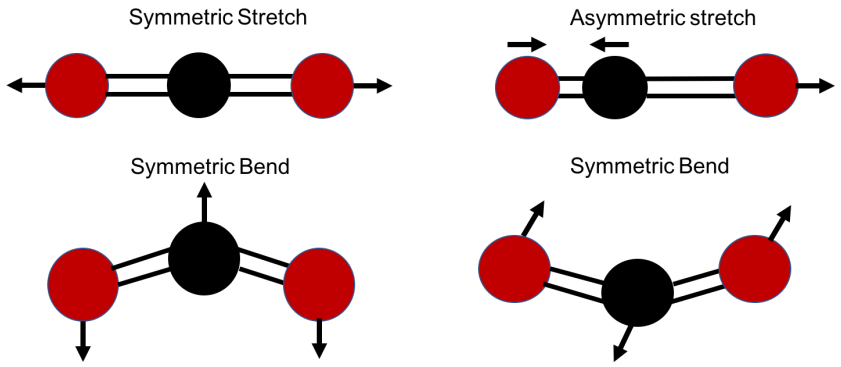


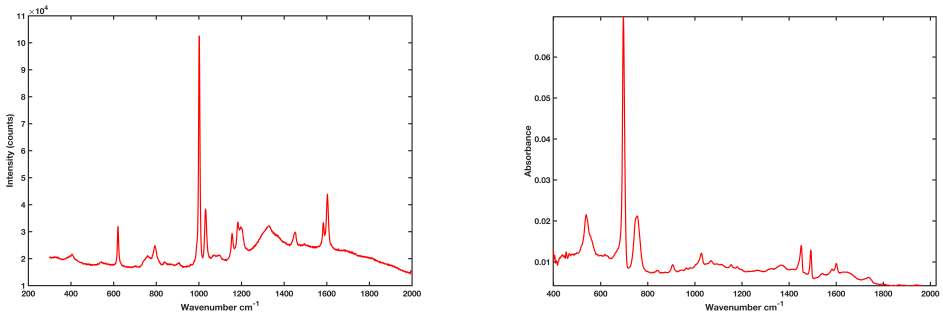
Figure 2.6: Vibrational modes of CO2 molecule.

The symmetric stretching mode causes a change in polarizability in the molecule and is therefore Raman active, the asymmetric stretch would not cause a change in overall polarizability, but would cause a change in dipole moment. Therefore the asymmetric stretch is active in the IR. The bending modes are also active in the IR, however, despite there being three vibrational modes that are active in the IR these modes only cause two bands on the IR absorption spectrum (peaks around 666cm^{-1} and 2350cm^{-1}). This is due to degeneracy effects, where the two bending modes are exactly opposite and therefore have the 'same' change in dipole moment.

2.4.1 Choosing the right spectroscopic method

When considering which spectroscopic technique is best for a particular sample type it is vital to consider the vibrational selection rules, environment of the sample and concentration of the analyte. Many molecules are active in both Raman and IR. Figure 2.7 shows representative Raman spectrum and a representative FTIR spectrum of polystyrene.

It is clear that although many of the vibrational modes are strong in both the



(a) Raman spectrum of polystyrene (b) FTIR spectrum of polystyrene

Figure 2.7: Representative Raman spectrum of polystyrene (a) and FTIR spectrum of the same piece of polystyrene (b). Wavenumbers that are very strong in the Raman Spectrum at 1600cm^{-1} are also present in the FTIR spectrum, however the strongest band at 100cm^{-1} is barely visible in the FTIR spectrum and the strong band at 700cm^{-1} is the FTIR spectrum is not visible in the Raman spectrum.

IR and the Raman spectra the peaks don't completely coincide with each other. Some peaks that are strong in Raman are weak in IR and vice versa depending on molecular structure. Generally, Raman is more sensitive to heteronuclear bonds whereas FTIR is more sensitive to homonuclear bonds and is very sensitive to OH bonds in water. Furthermore, Raman is more sensitive to how amorphous the crystal structure of the analyte is and less sensitive to IR overtones which leads to Raman peaks being generally sharper and more easily resolved than FTIR bands [13]. For a full treatment of the theoretical and experimental differences between Raman and FTIR spectra the reader is directed to [14]. When considering biological samples, both Raman and FTIR spectroscopy have been used previously to characterise protein structure [16–18]. Both techniques have also been used for disease classification in biological samples [19–22]. However, Raman holds many desirable properties for the application to a biofluid based method of disease detection. One of the biggest advantages of using Raman spectroscopy is that samples can be in aqueous solutions as water has a small Raman cross-section at near-infrared wavelengths. For this reason the following section will cover the current role of vibrational spectroscopy in detecting CRC with a focus on Raman spectroscopy.

2.5 Current role of Vibrational spectroscopy in detecting colorectal cancer

Current methods for detecting colorectal cancer are through screening and secondary care testing. In general these are invasive, expensive and cause high patient anxiety. There are currently no blood based tests in use in the UK for detecting CRC; histopathology of resected tissue specimens remains the 'gold standard' technique for diagnosis and staging. Raman and FTIR have been explored for diagnosis of CRC independently and also coupled with immunohistochemical staining [23, 24]. Table 2.1 provides an overview of the strengths and weaknesses of FTIR, Raman and traditional hematoxylin and eosin staining (H&E) [25–27] when used for clinical applications .

Table 2.1: Comparison of detection techniques

	Raman	**FTIR**	**H&E staining**
Method of Detection	Inelastic scattering of monochromatic (laser) light	Absorbance (polychromatic light source)	Combination of basic and acidic dyes
Real Time	Yes	Yes	No
Wavenumber range (cm^{-1})	50-4000	400-4000	N/A
Spatial Resolution	< 1 μm	5 μm	cellular
Enhancement Techniques	SERS, TERS, CARS, SORS, SRS	ATR	"Special" staining
Effect of water	Minimal	Large absorbance in NIR region	No
Destructive to Sample	No	No	Yes

SERS: Surface enhanced Raman spectroscopy; TERS: Tip enhanced Raman spectroscopy; CARS: Coherent anti-Stokes Raman spectroscopy; SORS: Spatially offset Raman spectroscopy; SRS: Stimulated Raman spectroscopy; ATR: Attenuated total reflection; FTIR: Fourier transform infrared spectroscopy; NIR: Near-infrared; N/A: Not available.

It suggests that for application to biological samples such as tissues and biofluids Raman could be the most favourable option. When used appropriately, Raman, like FTIR is non destructive and can be performed in real time compared to

H and E staining. The advantage of Raman over FTIR is the larger wavenumber range than FTIR and with better spatial resolution than both FTIR and H&E. Furthermore, Raman is a technique that relies on scattering so measurements can be taken with a wide range of instrumentation including microscopes and endoscopic probes. This is advantageous for *in-vivo* applications because it makes it possible to study samples that are optically too thick for transmission techniques which is traditionally the method for FTIR data collection in biological samples [28]. This, along with the need for a less invasive diagnostic tool that can analyse liquid biofluids leads to the remainder of this review focusing on the application of Raman technology to detecting CRC.

2.6 Clinical applications of Raman spectroscopy in colorectal cancer

Histopathological analysis of tissue biopsies is still considered the gold standard for the diagnosis of malignant tissues that have been surgically resected. Typically, after tissure fixation, sections of tissue are cut and mounted onto glass slides then stained using various methods to determine TNM stage, tumour type, histologic grade and the level of vascular invasion. However, histopathology is a slow process that requires a trained pathologist, it is also inherently subjective [29]. Raman spectroscopy offers the possibility of determining the presence of malignancy by detecting differences in Raman spectral features between normal and malignant tissue. Previously, Raman spectroscopy has been applied to *in vivo* probes that have the ability to discriminate multiple tissue types [30,31], biofluid analysis [32] and also analysis of cancerous cell lines for both discrimination and characterization [33]. The motivation behind using Raman used *in vivo* is to aid rapid diagnosis and help to identify possible areas of tissue for biopsy that might otherwise be missed. A summary of the literature and different applications of Raman towards clinical applications for CRC can be found in Table 2.2.

Method	**Sampling Type**	**Patient Number**	**Author/s**	**Year**	**Spectral Region (cm^{-1})**	**Laser excitation (nm)**	**Data analysis**
Probe	*In vivo* (tissue)	20	Shim et al.	2000	450-1800	785	PCA, PLS, ANN
Probe	*In vivo* & *ex vivo* (tissue)	9	Molckovsky et al.	2003	900-1800	785	PCA, LDA, LOOCV
Micro-spectrometer	*In vitro* (primary culture)	10	Chen et al	2006	500-1900	782.5	PCA
Probe	*Ex vivo* (tissue)	59	Widjaja et al.	2008	800-1800	785	PCA, SVM, LOOCV
Micro-spectrometer	*Ex vivo* (tissue)	54	Beljebbar et al.	2009	600-1800	785	SVM, PCA
Micro-spectrometer	*In vitro* (serum)	120	Li et al.	2012	800-1800	785	PCR, PLSR, LDA
Micro-spectrometer	*In vitro* (Cell lines)	N/A	Ranc et al.	2013	400-1800	532	PCA
Probe	*In vitro* (tissue)	177	Wood et al.	2014	800-1800	830	PCA, LDA, LOOCV
Micro-spectrometer	*In vivo* (tissue)	50	Bergholt et al.	2014	(800-1800) & (2800-3600)	785	PLS, LDA

Table 2.2: Table summarising the clinical applications of Raman spectroscopy to Colorectal Cancer. Where principal component analysis (PCA), linear differential analysis (LDA), principal least squares regression analysis (PLS/PLSR), leave one out cross validation (LOOCV), artificial neural network (ANN) and support vector machine (SVM).

Table 2.2 shows that most clinical applications of Raman have been in the near-infrared (NIR) region using 785 nm laser excitation for analysing tissue samples. This is likely due to reduced fluorescence of biological samples in the NIR region, furthermore photons from the 785 nm laser have lower energy than those in the visible region so are often less likely to damage a biological sample. The work is generally done in the "fingerprint region" of the Raman spectrum, i.e., 400-1800 cm^{-1} due to molecular bonds present in biological samples being Raman active in this region. All of the studies use chemometric data analysis so it seems for NR this is essential to differentiate the small differences in the Raman spectra when using biological samples. It is also clear that work in the field has been dominated by work towards the development of *in vivo* Raman probes for use during endoscopy but more recent work has used both Raman probes and micro-spectrometers.

2.6.1 Tissue analysis

Reviews dedicated to Raman spectroscopy for clinically useful *in vivo* probes for many types of tissue including tissues of the gastrointestinal (GI) tract are already available [25, 34, 35]. These *in vivo* probes were first introduced by Shim et al (2000) [36]. In general, the *in vivo* probes use a laser excitation source in the NIR wavelength range (light is non-mutagenic in this region) coupled to an optical fibre probe. The probe acts as both a source of light and a detector relaying a signal back to a charged coupled device detector (CCD) and computer for analysis. The development of probes is not a simple process, some materials that the probes are manufactured from have a large Raman cross section leading to design challenges regarding signal to noise ratio (gaining unwanted noise from the material). Other issues can be caused by tissue fluorescence signals being larger than the Raman signal making data acquisition difficult and spectral acquisition times impractical for clinical applications (more than 10 min). Nevertheless, some research groups have been successful in designing probes with short acquisition times, for specific use with gastrointestinal tissue for use in routine endoscopy that have short acquisition times [30, 37].

Shim et al (2000) successfully applied an *in vivo* probe to gain Raman spectra of colonic and oesophageal tissues from 20 patients [30]. The pressure spectral acquisitions were made with 5 s exposure time and repeated at normal and malignant sites. In the colonic tissue subtle differences between spectral area 1100-1800cm^{-1} were identified in normal versus malignant sites. PCA and LDA analysis was then used to determine accuracy for application to GI diagnostics however no specific results relating to diagnostics were published.

Molckovsky et al (2003) were the first to assess the diagnostic potential of near infrared Raman spectroscopy on colonic tissue by using adenomatous polyps as a model for dysplasia. The group used a custom-made fibre-optic Raman probe and used PCA-LDA analysis and LOOCV to analyse a total of 33 polyps from 8 patients [38]. After an initial *ex vivo* study of polypectomy specimens that involved a total of 54 spectra the analysis algorithm identified a sensitivity of 91% and a specificity of 95%. An *in vivo* study was then conducted with a total of 19 spectra form 9 polyps, after spectral analysis the algorithm identified adenomas with a sensitivity of 100% and specificity of 89%. A similar study involving a more *ex vivo* specimens was conducted by Widjaja et al (2008), the group were able to differentiate cancerous tissue with 100% sensitivity and 98.1%-99.7% specificity using a diagnostic algorithm using PCA and LDA [39].

Raman spectroscopy also has been investigated as a complementary technique to histopathology. The first application of Raman spectroscopy to discriminate between cancerous vs normal colonic tissue (among others) was by Feld et al (1995) [40]. The group looked at the difference spectra between normal and cancerous tissue. Results showed potential with spectral differences in the tissue that could be due to higher nucleic acid levels in the cancerous samples.

Beljebbar et al (2009) used a Raman micro spectrometer on 27 normal and 27 cancerous *ex vivo* frozen tissue samples. Unsupervised hierarchal cluster analysis to differentiate between normal and adeno-carcinomatous human colonic tissues was discussed [41]. The technique was based on the spatial distribution of molecular changes in colon constituents such as proteins, lipids and nucleic acids. The spectroscopic data were then used to create pseudo-colour Raman images of tissues for comparison with histopathological slides. Databases were created for the

purpose of comparing unknown specimens. Six extra frozen unknown samples were fed into the database and were correctly identified as either cancerous or normal [41]. This study showed the potential for Raman spectroscopy to aid histopathology by adding structural molecular information to the visible information gained from H and E staining. This study used frozen tissue samples but there are different fixation methods available. There have been studies on the effect that different fixation methods of both cell lines and tissue samples taken from resection has on the Raman spectral signature but these will not be discussed further in this review [42–44].

It should be noted that Raman spectroscopy used as an adjunct to histopathology has not just been applied to CRC. Some groups have investigated the use of stimulated Raman spectroscopy (SRS) to create spectral histopathological images for breast, brain and skin tissue among others [45–47]. For example, Satoh et al (2014) used SRS and PCA multivariate analysis to produce Raman map images of damaged liver tissue in mice [48]. The images could then be compared to different staining methods used in histopathology.

2.7 Detection of colorectal cancer in blood samples

The small Raman cross section of water is advantageous for the study of biofluid samples such as urine and blood as they can be analysed whilst still liquid. Berger et al (1999) introduced the idea that Raman had potential for the analysis of biofluids, in particular human blood [49]. Premasiri et al (2012) found that the Raman spectra of whole human blood are dominated by the normal modes of either oxygenated or deoxygenated haemoglobin porphyrin macrocycle and haemoglobin [50]. To combat this some studies were performed on blood derivatives. Harris et al (2009) discussed the potential of peripheral blood samples analysed by NR to provide cancer screening in head and neck cancer [51]. Using Raman spectroscopy and LDA analysis alone the study found this technique to have approximately 65% specificity and sensitivity for the discrimination of can-

cer in peripheral blood samples. The use of non-enhanced Raman spectroscopy for the discrimination of blood serum between normal and CRC was first reported by Li et al (2012) [52]. Li et al used clinical samples from 44 colon, 46 rectum and 30 healthy controls. Raman peak parameters and fluorescence background were used along with multivariate analysis techniques such as PCR and PLSR for the dimension deduction of spectral data. LDA on PCs was used to assess diagnostic performance. Three distinct Raman peaks were found to have significance: 1029cm^{-1}, 1538cm^{-1} and 1170cm^{-1}. The 1538cm^{-1} peak was assigned to beta-carotene and 1170cm^{-1} to tryptophan and phenylalanine. On comparing the average spectra from the three sample groups the latter two peaks were shown to decrease in the cancer compared to the control group. This was explained as a decrease in anti-cancer related molecules. This study showed that it is possible to discriminate between serum samples of patients with and without CRC. The result of the PCR-LDA analysis was promising as it identified normal samples with 87.5% accuracy and 96.7% specificity and colon cancer samples with a sensitivity of 84.8%.

2.7.1 Surface enhanced Raman scattering detection methods

Some Raman techniques are subject to interference from samples that exhibit fluorescence; a fluorescence signal is far greater than the Raman response so can mask Raman signals. Fluorescence is fairly common when using wavelengths less than 785 nm that are used commonly in biological studies. SERS offers a resolution to some of these issues as it reduces the effects of inherent fluorescence while increasing the intensity of the Raman response of the sample. Premasiri et al (2001) demonstrated the need for an enhancement mechanism to detect some low concentration analytes in urine using Raman spectroscopy [53]. Urea had a sufficiently high concentration to be analysed by NR in liquid form however, lower-level nitrogen compounds needed enhanced Raman spectroscopy for detection. Therefore, SERS was used as an enhancement mechanism in to enable detection of the compounds that were in concentrations too low to detect with Raman.

2.7.2 Clinical applications of Surface enhanced Raman spectroscopy for colorectal cancer

The main use of SERS in clinical applications for CRC has been as a detection method. Lin et al (2011) were the first to report SERS serum analysis for the detection of CRC [54]. In their study 43 nm spherical colloidal gold nanoparticles were simply mixed with serum samples from 38 patients and 45 control samples and dropped onto an aluminium substrate. This technique is known as label-free SERS. It generally relies on blood constituents being adsorbed onto the surface of metallic nanoparticles causing an enhanced Raman response. In the Lin et al study a Raman micro-spectrometer (Renishaw, Great Britain) fitted with a 785 nm diode laser was used to gain spectra in the 300-1800 cm^{-1} range. Spectra from the two groups were normalised to the integrated area under the curve in the 350-1750 cm^{-1} wavenumber range. The mean spectrum for the normal serum and the cancer serum were then compared to isolate wavenumbers that showed the most variation between the two groups. An empirical diagnostic algorithm based on peak intensities at 725 cm^{-1} and 638 cm^{-1} was then used to classify the normal and the cancer samples. These were chosen based on previous studies by Han et al (2008) that showed the ratio to be an important disease marker. The empirical algorithm technique was compared PCA-LDA multivariate analysis [55]. The PCA analysis then used whole spectra to discern the spectral components that had the largest variation. After comparison the group found PCA-LDA to be more effective at detecting CRC. Specificities for detecting cancer to be 68.4% and 97.4% for the empirical approach and the multivariate approach, respectively. This study showed that through simply mixing gold nanoparticles with serum that it is possible to discriminate between normal and cancer samples using SERS along with both empirical and multivariate analysis techniques. The potential for using whole spectra coupled with PCA-LDA as a screening technique for CRC was then discussed using the PCA-LDA methods in a second publication from the same group [56]. The two publications also included tentative peak assignments and major vibrational bands that have been observed previously in serum samples. The peak assignments are vital for an accurate description of

what is happening at the molecular level when a patient has a disease. However, due to the additive nature and complex compounds found in serum samples label-free SERS techniques are subject to huge amounts of variation. The field therefore developed into a specific molecule detection method.

The need for specific detection has motivated the development of labelled SERS probes. These probes have previously been used for detecting disease specific proteins in both tissue and serum samples [56–60]. They have also been used for the detection of circulating tumour cells [61]. However little is reported on the specific application of targeted SERS probes for use in detecting CRC [62]. In general, targeted SERS techniques rely on either aggregation of antibody-functionalised nanoparticles after exposure to a protein or they are used to form of sandwich immunoassay similar to that of an enzyme-linked immunosorbent assay (ELISA) setup but using SERS active probes rather than fluorescent-tagged antibodies. Chen et al (2013) developed a SERS based immunoassay for the detection of carcinoembryonic antigen (CEA) in serum of patients with CRCs. CEA antibody functionalised glass slides were used in conjunction with SERS active probes that were also functionalised with anti-CEA [63]. A range of concentrations from 5 10-3-5 105 ng/mL CEA were prepared and the SERS response monitored. The SERS intensity was correlated linearly with the concentration of CEA in the characteristic peak of the Raman reporter molecule at 1077 cm^{-1} and hence a calibration curve was established. CEA concentrations in serum samples form 26 patients with CRC were then analysed with both SERS immunoassay and electrochemical luminescence. The results were then compared and the two techniques had similar agreement. Using the calibration curve for the patient samples a detection limit of 5 pg/ mL was achieved. This is the only study specifically using CEA for the detection of CRC, however there is other work in the literature using CEA conjugated antibodies but in other disease settings [58], [59]. Another slightly different approach to use SERS as a characterisation/ validation tool when developing other nanoscale devices for clinical use such as the work done by da Paz et al (2012) [64]. In this study SERS was used as a characterisation tool in the development of maghemite nanoparticles as a theragnostic device for CRC. The SERS active nano- particles used in this study were functionalised with Anti-

CEA antigen in the hope that they can be used to detect primary and metastatic CRC, the authors hoped that these nanoparticles could then be developed for a variety of applications including magnetic resonance imaging (MRI) enhancement and targeted drug delivery although nothing further has been published.

2.7.3 The limitations of Raman and surface enhanced Raman spectroscopy in clinical applications

Raman and SERS based tools have shown potential that they will have a place as either an alternative or an adjunct to current diagnostic methods. The development of SERS biomarker detection could also lead to its use in personalized medicine. However, there are still some limitations of Raman and SERS techniques that will need to be overcome before they are used routinely in a clinical setting. These include:

- Many Raman studies involve costly equipment and expensive substrates, there will need to be investigations into cost reduction for large scale applications;
- Raman and SERS studies that are carried out in a laboratory will require sample handling and storage, the effect of handling samples and storage techniques on the performance of Raman based tools will need to be quantified;
- Thermal damage thresholds of *in vivo* tissue and *ex vivo* tissue samples from the colon and rectum will need to be established;
- Many clinical studies use different analytical techniques, and require the skill of the user to interpret the results. User-friendly software for diagnostic analysis of the spectra will need to be developed and tested for multi-user reliability;
- Inter-equipment and inter-user variability studies will need to be carried out, Raman equipment can often be susceptible to variability from external factors such as room temp, laser stability, etc.;

- SERS based techniques have been subject to reproducibility issues, CRC is a heterogeneous disease so if immunoassay style tools are to be used then large scale studies with clinical samples will need to be carried out.

- SERS based methods still rely on named biomarkers so will always cost more than normal Raman and other label-free techniques.

2.8 Conclusions from review of the literature

In the field of cancer detection in general Raman spectroscopy and SERS have gone through a period of rapid progress in the last decade. The use of Raman in clinical applications for CRC has been dominated previously by the discrimination of cancerous vs non-cancerous tissue with only a few studies on the use of Raman with biofluids for CRC detection. Currently, there are successful *in vivo* Raman tools for real-time use during endoscopy. These tools can be used to gain molecular information through Raman imaging and traditional spectroscopy. Therefore, they aid current endoscopic techniques by providing molecular information that could be missed using traditional methods. However, endoscopic Raman tools remain expensive to produce and require specialist knowledge to operate the machinery. This does not improve the referral process as a patient still undergoes an invasive procedure. Furthermore, thermal thresholds for the damage of GI tissue need to be properly established before these tools can used in a routine clinical setting. Future research into the large-scale manufacture (and miniaturisation) of endoscopic Raman tools needs to be carried out to investigate variability between sites and the cost effectiveness of Raman tools compared to current technology.

SERS has emerged as an alternative method to Raman for detection of low concentration analytes as it enhances signals and reduces fluorescence compared to Raman. Currently, research uses different techniques are used to gain a SERS response from samples. One of the limitations of SERS based techniques has been variation in the plasmon resonance of nano-structures which are subject to large variability. Therefore, in SERS methods, research into reducing the variability in

SERS response even across a single sample will need to be investigated. Another method of gaining a SERS response is through a SERS based immunoassay; this has been used successfully to detect the current accepted biomarker for CRC CEA. However, SERS based immunoassay to detect CRC are still reliant on having a named biomarker to detect. Reviews questioned the effectiveness of CEA as a screening tool [65]. This along with the variability of SERS and the reliance of protein or metabolite based biomarkers means it would potentially be difficult to translate especially into a primary care setting.

The least studied application for Raman in detecting CRC is the potential of using normal Raman for biofluid analysis. Label free analysis holds the greatest potential for clinical translation as a primary care referral tool. This is due to the fact that blood tests have the potential to be less invasive, quicker and cheaper than the current diagnostic tests [66]. There is also a clinical need in the UK for tests that aid the referral process by providing GPs more information, that have a quick turnaround, and are simple to interpret result. There are currently other CRC detection tests in development that are more advanced than Raman based blood tests such as mSept9 blood based testing. However, the cost of these tests their availability in the UK are yet to be confirmed [66].

There are no current studies that use non-enhanced Raman spectroscopy for CRC detection on liquid serum samples. Resonance Raman has been used previously with fluorescence to detect CRC but for the technique to be translated a larger cohort of samples is needed. Studies using other vibrational spectroscopy techniques such as FTIR have suggested that biofluid spectroscopy holds huge promise in being translated to a clinical setting [1]. Therefore the work in this thesis will concentrate on the development of serum Raman spectroscopy for the detection of CRC. Specifically the potential of serum Raman spectroscopy to be used as a diagnostic aid to primary care clinicians as a referral triage tool.

Bibliography

[1] Matthew J. Baker, Shawn R. Hussain, Lila Lovergne, Valérie Untereiner, Caryn Hughes, Roman A. Lukaszewski, Gérard Thiéfin, and Ganesh D. Sockalingum. Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chem. Soc. Rev.*, 45(7):1803–1818, 2016.

[2] Dake Wang, Kathryn Mittauer, Nicholas Reynolds, Dake Wang, Kathryn Mittauer, and Nicholas Reynolds. Raman scattering of carbon disulfide : The temperature effect Raman scattering of carbon disulfide : The temperature effect. 1130, 2009.

[3] Derek A Long. *The Raman Effect*. Wiley,, 2002.

[4] Johannes Arnoldus Koningstein. *Introduction to the Theory of the Raman Effect*. Springer Science & Business Media, 2012.

[5] Fathima S Ameer, Charles U Pittman Jr, and Dongmao Zhang. Quantification of resonance raman enhancement factors for rhodamine 6g (r6g) in water and on gold and silver nanoparticles: Implications for single-molecule r6g sers. *The Journal of Physical Chemistry C*, 117(51):27096–27104, 2013.

[6] Martin Fleischmann, Patrick J Hendra, and A James McQuillan. Raman spectra of pyridine adsorbed at a silver electrode. *Chemical Physics Letters*, 26(2):163–166, 1974.

[7] M Grant Albrecht and J Alan Creighton. Anomalously intense raman spectra of pyridine at a silver electrode. *Journal of the american chemical society*, 99(15):5215–5217, 1977.

[8] Katrin Kneipp, Harald Kneipp, Ramasamy Manoharan, Irving Itzkan, Ramachandra R Dasari, and Michael S Feld. Surface-enhanced raman scattering (sers)a new tool for single molecule detection and identification. *Bioimaging*, 6(2):104–110, 1998.

[9] Eric Le Ru and Pablo Etchegoin. *Principles of Surface-Enhanced Raman Spectroscopy: and related plasmonic effects*. Elsevier, 2008.

[10] W Suetaka. *Surface infrared and Raman spectroscopy: methods and applications*, volume 3. Springer Science & Business Media, 2013.

[11] Katrin Kneipp, Yang Wang, Harald Kneipp, Lev T Perelman, Irving Itzkan, Ramachandra R Dasari, and Michael S Feld. Single molecule detection using surface-enhanced raman scattering (sers). *Physical review letters*, 78(9):1667, 1997.

[12] Ewen Smith and Geoffrey Dent. *Modern Raman spectroscopy: a practical approach*. John Wiley & Sons, 2013.

[13] K. Janssens. *Section V Methods 4: Elemental Analysis*. 2003.

[14] Peter Larkin. *Infrared and Raman spectroscopy: principles and spectral interpretation*. Elsevier, 2017.

[15] Brian C Smith. *Infrared spectral interpretation: a systematic approach*. CRC press, 1998.

[16] A. Rygula, K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk, and M. Baranska. Raman spectroscopy of proteins: A review. *J. Raman Spectrosc.*, 44(8):1061–1076, 2013.

[17] Huayan Yang, Shouning Yang, Jilie Kong, Aichun Dong, and Shaoning Yu. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat. Protoc.*, 10(3):382–396, 2015.

[18] Parvez I. Haris and Feride Severcan. FTIR spectroscopic characterization of protein structure in aqueous and non-aqueous media. *J. Mol. Catal. - B Enzym.*, 7(1-4):207–221, 1999.

[19] Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, Michael J Walsh, Martin R McAinsh, Nicholas Stone, and Francis L Martin. Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.*, 11(4):664–687, 2016.

[20] Matthew J Baker, Ehsan Gazi, Michael D Brown, Jonathan H Shanks, Peter Gardner, and Noel W Clarke. Ftir-based spectroscopic analysis in the identification of clinically aggressive prostate cancer. *British journal of cancer*, 99(11):1859, 2008.

[21] Francisco Santos, Sandra Magalhaes, Magda C Henriques, Margarida Fardilha, and Alexandra Nunes. Spectroscopic features of cancer cells: Ftir spectroscopy as a tool for early diagnosis. *Current Metabolomics*, 6(2):103–111, 2018.

[22] Inês P Santos, Elisa M Barroso, Tom C Bakker Schut, Peter J Caspers, Cornelia GF van Lanschot, Da-Hye Choi, Martine F van der Kamp, Roeland WH Smits, Remco van Doorn, Rob M Verdijk, et al. Raman spectroscopy for cancer detection and cancer surgery guidance: translation to the clinics. *Analyst*, 142(17):3025–3047, 2017.

[23] Liu Dong, Xuejun Sun, Zhang Chao, Shiyun Zhang, Jianbao Zheng, Rajendra Gurung, Junkai Du, Jingsen Shi, Yizhuang Xu, Yuanfu Zhang, and Jinguang Wu. Evaluation of FTIR spectroscopy as diagnostic tool for colorectal cancer using spectral analysis. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.*, 122:288–294, 2014.

[24] Qing-Bo Li, Zhi Xu, Neng-Wei Zhang, Li Zhang, Fan Wang, Li-Min Yang, Jian-Sheng Wang, Su Zhou, Yuan-Fu Zhang, Xiao-Si Zhou, Jing-Sen Shi, and Jin-Guang Wu. In vivo and in situ detection of colorectal cancer using Fourier transform infrared spectroscopy. *World J. Gastroenterol.*, 11(3):327–330, 2005.

[25] Catherine Kendall, Martin Isabelle, Florian Bazant-Hegemark, Joanne Hutchings, Linda Orr, Jaspreet Babrah, Rebecca Baker, and Nicholas Stone. Vibrational spectroscopy: a clinical tool for cancer diagnostics. *Analyst*, 134:1029–1045, 2009.

[26] Rachel E Kast, Stephanie C Tucker, Kevin Killian, Micaela Trexler, Kenneth V Honn, and Gregory W Auner. Emerging technology: applications of Raman spectroscopy for prostate cancer. *Cancer Metastasis Rev.*, 2014.

[27] Andrew H. Fischer, Kenneth A. Jacobson, Jack Rose, and Rolf Zeller. Hematoxylin and eosin staining of tissueand cell sections. *Cold Spring Harb. Protoc.*, 3(5):4986–4988, 2008.

[28] a J Berger, Y Wang, and M S Feld. Rapid, noninvasive concentration measurements of aqueous biological analytes by near-infrared Raman spectroscopy. *Appl. Opt.*, 35(1):209–212, 1996.

[29] Matthew Fleming, Sreelakshmi Ravula, Sergei F. Tatishchev, and Hanlin L. Wang. Colorectal carcinoma: Pathologic aspects, 2012.

[30] Martin G. Shim and Brian C. Wilson. Development of an In Vivo Raman Spectroscopic System for Diagnostic Applications. *J. Raman Spectrosc.*, 28(23):131–142, 1997.

[31] A Mahadevan-Jansen, M Mitchell, N Ramanujam, U Utzinger, and R Richards-Kortum. Development of a fiber optic probe to measure NIR Raman spectra of cervical tissue in vivo. *Photochem Photobiol*, 68(3):427–431, 1998.

[32] Xiaozhou Li, Tianyue Yang, and Siqi Li. Discrimination of serum Raman spectroscopy between normal and colorectal cancer using selected parameters and regression-discriminant analysis. *Appl. Opt.*, 51(21):5038, 2012.

[33] Václav Ranc, Josef Srovnal, Libor Kvítek, and Marian Hajduch. Discrimination of circulating tumor cells of breast cancer and colorectal cancer from normal human mononuclear cells using Raman spectroscopy. *Analyst*, 138(20):5983–8, 2013.

[34] E Hanlon, R Manoharan, T Koo, K Shafer, J Motz, M Fitzmaurice, J Kramer, I Itzkan, R Dasari, and M Feld. Prospects for in vivo Raman spectroscopy. *Phys. Med. Biol.*, 45(2):R1–59, 2000.

[35] Urs Utzinger and Rebecca R Richards-Kortum. Fiber optic probes for biomedical optical spectroscopy. *J. Biomed. Opt.*, 8(1):121–47, 2003.

[36] M G Shim, L M Song, N E Marcon, and B C Wilson. In vivo near-infrared Raman spectroscopy: demonstration of feasibility during clinical gastrointestinal endoscopy. *Photochem. Photobiol.*, 72(1):146–150, 2000.

[37] J. J. Wood, C. Kendall, J. Hutchings, G. R. Lloyd, N. Stone, N. Shepherd, J. Day, and T. A. Cook. Evaluation of a confocal Raman probe for pathological diagnosis during colonoscopy. *Color. Dis.*, 16(9):732–738, 2014.

[38] Andrea Molckovsky, Louis Michel Wong Kee Song, Martin G. Shim, Norman E. Marcon, and Brian C. Wilson. Diagnostic potential of near-infrared Raman spectroscopy in the colon: Differentiating adenomatous from hyperplastic polyps. *Gastrointest. Endosc.*, 57(3):396–402, 2003.

[39] Effendi Widjaja, W E I Zheng, and Zhiwei Huang. Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines. pages 653–662, 2008.

[40] M S Feld, R Manoharan, J Salenius, J Orenstein-Camdona, T J Romer, J F Brennan III, R R Dasari, and Y Wang. Detection and Characterization of Human Tissue Lesions with Near Infrared Raman Spectroscopy. In *Adv. Fluoresc. Sens. Technol. II, SPIE*, volume 2388, pages 99–104, 1995.

[41] a Beljebbar, O Bouché, M D Diébold, P J Guillou, J P Palot, D Eudes, and M Manfait. Identification of Raman spectroscopic markers for the characterization of normal and adenocarcinomatous colonic tissues. *Crit. Rev. Oncol. Hematol.*, 72(3):255–64, dec 2009.

[42] Philip R T Jess, Daniel D W Smith, Michael Mazilu, Kishan Dholakia, Andrew C. Riches, and C. Simon Herrington. Early detection of cervical neoplasia by Raman spectroscopy. *Int. J. Cancer*, 121(12):2723–2728, 2007.

[43] James W. Chan, Douglas S. Taylor, and Deanna L. Thompson. The effect of cell fixation on the discrimination of normal and leukemia cells with laser tweezers Raman spectroscopy. *Biopolymers*, 91(2):132–139, 2009.

[44] Aidan D. Meade, Colin Clarke, Florence Draux, Ganesh D. Sockalingum, Michel Manfait, Fiona M. Lyng, and Hugh J. Byrne. Studies of chemical fixation effects in human cell lines using Raman microspectroscopy. *Anal. Bioanal. Chem.*, 396(5):1781–1791, 2010.

[45] Richa Mittal, Mihaela Balu, Tatiana Krasieva, Eric O. Potma, Laila Elkeeb, Christopher B. Zachary, and Petra Wilder-Smith. Evaluation of stimulated raman scattering microscopy for identifying squamous cell carcinoma in human skin. *Lasers Surg. Med.*, 45(8):496–502, 2013.

[46] M D Schaeberle, V F Kalasinsky, J L Luke, E N Lewis, I W Levin, and P J Treado. Raman chemical imaging: histopathology of inclusions in human breast tissue. *Anal. Chem.*, 68(11):1829–1833, 1996.

[47] Minbiao Ji, Daniel a Orringer, Christian W Freudiger, Shakti Ramkissoon, Xiaohui Liu, Darryl Lau, Alexandra J Golby, Isaiah Norton, Marika Hayashi, Nathalie Y R Agar, Geoffrey S Young, Cathie Spino, Sandro Santagata, Sandra Camelo-Piragua, Keith L Ligon, Oren Sagher, and X Sunney Xie. Rapid, label-free detection of brain tumors with stimulated Raman scattering microscopy. *Sci. Transl. Med.*, 5(201):201ra119, 2013.

[48] Shuya Satoh, Yoichi Otsuka, Yasuyuki Ozeki, Kazuyoshi Itoh, Akinori Hashiguchi, Ken Yamazaki, Hiroyuki Hashimoto, and Michiie Sakamoto. Label-free visualization of acetaminophen-induced liver injury by high-speed stimulated Raman scattering spectral microscopy and multivariate image analysis. *Pathol. Int.*, 64(10):518–526, 2014.

[49] a J Berger, T W Koo, I Itzkan, G Horowitz, and M S Feld. Multicomponent blood analysis by near-infrared Raman spectroscopy. *Appl. Opt.*, 38:2916–2926, 1999.

[50] Author Manuscript, Blood Plasma, and Bioanalytical Sensing. NIH Public Access. 116(31):9376–9386, 2013.

[51] Andrew T Harris, Anxhela Lungari, Christopher J Needham, Stephen L Smith, Michael a Lones, Sheila E Fisher, Xuebin B Yang, Nicola Cooper,

Jennifer Kirkham, D Alastair Smith, Dominic P Martin-Hirsch, and Alec S High. Potential for Raman spectroscopy to provide cancer screening using a peripheral blood sample. *Head Neck Oncol.*, 1:34, jan 2009.

[52] Xiaozhou Li, Tianyue Yang, and Siqi Li. Discrimination of serum Raman spectroscopy between normal and colorectal cancer using selected parameters and regression-discriminant analysis. *Appl. Opt.*, 51(21):5038–43, jul 2012.

[53] W R Premasiri, R H Clarke, and M E Womble. Urine analysis by laser Raman spectroscopy. *Lasers Surg. Med.*, 28(September 2000):330–334, 2001.

[54] Duo Lin, Shangyuan Feng, Jianji Pan, Yanping Chen, Juqiang Lin, Guannan Chen, Shusen Xie, Haishan Zeng, and Rong Chen. Colorectal cancer detection by gold nanoparticle based surface-enhanced Raman spectroscopy of blood serum and statistical analysis. *Opt. Express*, 19(14):13565–77, jul 2011.

[55] H. W. Han, X. L. Yan, R. X. Dong, G. Ban, and K. Li. Analysis of serum from type II diabetes mellitus and diabetic complication using surface-enhanced Raman spectra (SERS). *Appl. Phys. B*, 94(4):667–672, 2008.

[56] Juqiang Lin, Rong Chen, Shangyuan Feng, Jianji Pan, Buhong Li, Guannan Chen, Shaojun Lin, Chao Li, Li-qing Sun, Zufang Huang, and Haishan Zeng. Surface-enhanced Raman scattering spectroscopy for potential noninvasive nasopharyngeal cancer detection. *J. Raman Spectrosc.*, 43(4):497–502, apr 2012.

[57] Xiaohui Ji, Shuping Xu, Lianying Wang, Min Liu, Kai Pan, Hang Yuan, Lan Ma, Weiqing Xu, Jinghong Li, Yubai Bai, and Tiejin Li. Immunoassay using the probe-labeled Au/Ag core-shell nanoparticles based on surface-enhanced Raman scattering. *Colloids Surfaces A Physicochem. Eng. Asp.*, 257-258:171–175, may 2005.

[58] Ji-Wei Chen, Yong Lei, Xiang-Jiang Liu, Jian-Hui Jiang, Guo-Li Shen, and Ru-Qin Yu. Immunoassay using surface-enhanced Raman scattering based

on aggregation of reporter-labeled immunogold nanoparticles. *Anal. Bioanal. Chem.*, 392(1-2):187–93, sep 2008.

[59] Hyangah Chon, Sangyeop Lee, Sang Wook Son, Chil Hwan Oh, and Jaebum Choo. Highly sensitive immunoassay of lung cancer marker carcinoembryonic antigen using surface-enhanced Raman scattering of hollow gold nanospheres. *Anal. Chem.*, 81(8):3029–34, apr 2009.

[60] Chunyuan Song, Linghua Min, Ni Zhou, Yanjun Yang, Boyue Yang, Lei Zhang, Shao Su, and Lianhui Wang. Ultrasensitive detection of carcinoembryonic antigen by using novel flower-like gold nanoparticle SERS tags and SERS-active magnetic nanoparticles. *RSC Adv.*, 4(78):41666–41669, aug 2014.

[61] Yanping Chen, Xiongwei Zheng, Gang Chen, Chen He, Weifeng Zhu, Shangyuan Feng, Gangqin Xi, Rong Chen, Fenghua Lan, and Haishan Zeng. Immunoassay for LMP1 in nasopharyngeal tissue based on surface-enhanced Raman scattering. *Int. J. Nanomedicine*, 7:73–82, jan 2012.

[62] Michael Y Sha, Hongxia Xu, Michael J Natan, and Remy Cromer. Surface-enhanced Raman scattering tags for rapid and homogeneous detection of circulating tumor cells in the presence of human whole blood. *J. Am. Chem. Soc.*, 130(51):17214–5, dec 2008.

[63] Gang Chen, Yanping Chen, Xiongwei Zheng, Cheng He, Jianping Lu, Shangyuan Feng, Rong Chen, and Haisan Zeng. Surface-enhanced Raman scattering study of carcinoembryonic antigen in serum from patients with colorectal cancers. *Appl. Phys. B*, 113(4):597–602, may 2013.

[64] Mariana Campos da Paz, Maria De Fátima M Almeida Santos, Camila M B Santos, Sebastião W. da Silva, Lincoln Bernardo de Souza, Emília C D Lima, Renata C. Silva, Carolina M. Lucci, Paulo César Morais, Ricardo B. Azevedo, and Zulmira Guerrero Marques Lacava. Anti-CEA loaded maghemite nanoparticles as a theragnostic device for colorectal cancer. *Int. J. Nanomedicine*, 7:5271–5282, 2012.

[65] Thomas F Imperiale and Charles J Kahi. Cost effectiveness of new biomarkers for colorectal cancer screening–futility or call for innovation?, 2017.

[66] Frank Rinaldi. Global market assessment for Raman spectroscopy and colorectal cancer., 2017.

Chapter 3

Experimental principles, materials and methods

This chapter will include a description of the general materials and methods used throughout this work. Sample handling and storage will be considered the general experimental principles behind Raman and FTIR spectroscopy will be covered. Any parameters that were kept consistent during the work will be presented, with any parameters that required optimisation are detailed in the relevant chapters.

3.1 General Considerations

Personal protective equipment including safety glasses, laser safety goggles, laboratory coats and disposable gloves were utilised as appropriate. Laser safety training and hepatitis B immunisation were carried out prior to work commencing. All reagents and other waste products were disposed of in accordance with Swansea University's health and safety protocols.

Ethical approval

Ethical approval for sample collection during this study was granted by Health and Care Research Wales Research Ethics Service (REC reference 14/WA/0028). Patients were prospectively recruited from a university teaching hospital after written informed consent. Blood samples were collected from each patient by a trained phlebotomy. On sample collection patient data were anonymised and stored with a reference number.

Patient inclusion criteria

Patients were recruited prospectively into seven research groups as detailed in the Appendix A. All patients in the colorectal cancer cohort had their tumour status pathologically verified by a consultant histopathologist. All patients in the control group had a normal colonoscopy with exclusion of adenomatous polyps and malignancy. Smoking status, co-morbidities and current medications were also recorded.

Sample handling/storage

Five millilitres of blood was collected by venous blood collection using serum separator gel tubes (BD Vacutainer, USA) unless otherwise stated. All blood samples were left to clot for 30mins before being spun at 1300g for 10 minutes in a swing head centrifuge. Fresh samples (250 μl aliquots) were then kept on ice until data collection within 24 hours of blood collection as to prevent sample degradation. The remainder of the samples were aliquotted again at 250 μl and stored at -80°C until thawed for data collection. Frozen samples were thawed on ice and data were collected within 24h of thawing.

3.2 Raman Microspectroscopy

3.2.1 Raman spectrometer overview

The Renishaw InVia Raman spectrometer (Reflex) was used to obtain all Raman data. The system consists of two laser sources that are coupled to an upright microscope in a backscatter collection configuration. The microscope is within a sealed pod that includes a motorised stage that allows the user to focus laser light onto a sample for investigation with out being exposed to laser light. The system is controlled by the Renishaw Wire software which automatically controls laser beam path alignment into the microscope and allows the user to control the alignment, focusing and data acquisition. Figure 3.1 shows an overview of the beam path set up in the Renishaw system. Briefly, the two laser light sources

are a visible 532 nm (Nd:YAG) and a near infra-red 785 nm (diode). Both light sources have separate beam paths which are aligned automatically when a source is selected using the Wire software. Once a light source has been selected and the corresponding optical mirrors aligned the laser beam enters the system. On entry, the beam encounters a set of neutral density (ND) filters. These allow for a selection of the percentage attenuation of the laser beam into the system which can be selected using the Wire interface. The percentage attenuation has 16 possible options ranging from 0.00000005% to 100%. This is an especially important feature when using the system to acquire data on soft matter materials such as blood products which can be photosensitive or damaged by too much light exposure. Once the beam has passed through the ND filters it is redirected using a motorised mirror into the system.

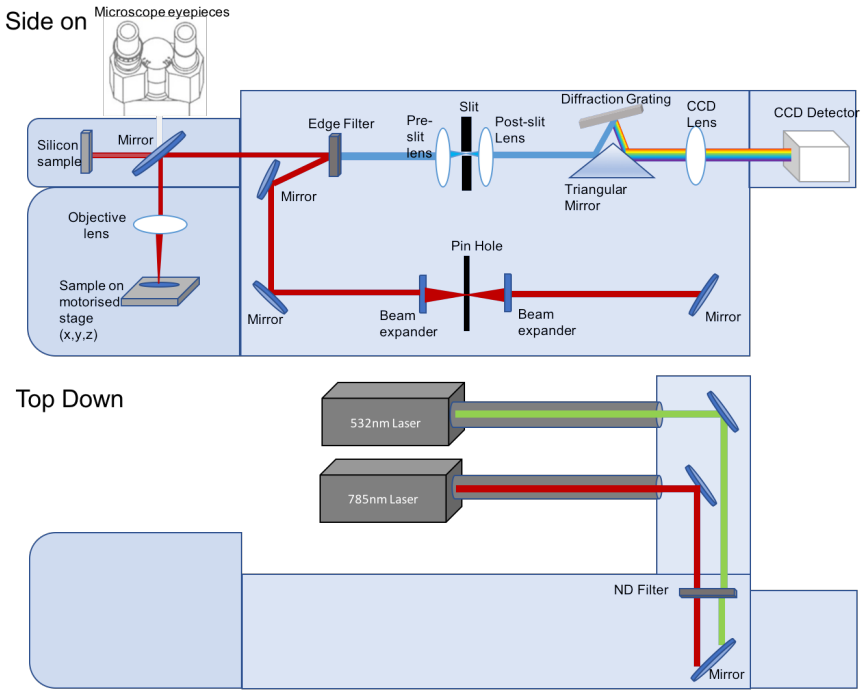


Figure 3.1: Renishaw InVia Raman Microscope beam path.

In the system, the beam passes through a beam expander and an optional pin hole. The pin hole does not affect the beam path of the 532 nm laser. However, it can be used with the 785 nm laser in order to create an attenuated point laser spot rather than the line beam created by the laser. The maximum output from the 785 nm laser at the source is 300 mW, the maximum output at the source

for the 532 nm laser is 150 mW.

(a) Pin hole out (b) Pin hole in

Figure 3.2: Beamline shape for the 785nm laser line with pinhole in (a) and pinhole out(b).

After the beam has passed the pin hole the beam is re-directed by another motorised mirror onto another series of mirrors to focus the beam onto a sample through the microscope objective. At the sample the laser power for the 785 nm laser and the 532 nm lasers were in the range between 165-185 mW and 45-55 mW, respectively without the pinhole and at 100% laser power.

The choice of objective is important as the objective is responsible for focusing the laser beam onto a sample as well as collecting the backscattered Rayleigh and Raman light. The beam of light then gets directed through a Raman edge filter. The edge filter is a type of long pass filter that absorbs all light up to a certain wavelength e.g. for a 785 nm laser will absorb up to 787 nm. It then transmits all light above that wavelength allowing only the Raman shifted light to pass. Therefore the beam of reflected light from the sample has all Rayleigh scattered light filtered off allowing detection of only the Raman scattered light. The InVia system is equipped with an edge filter for both 785 nm and 532 nm light.

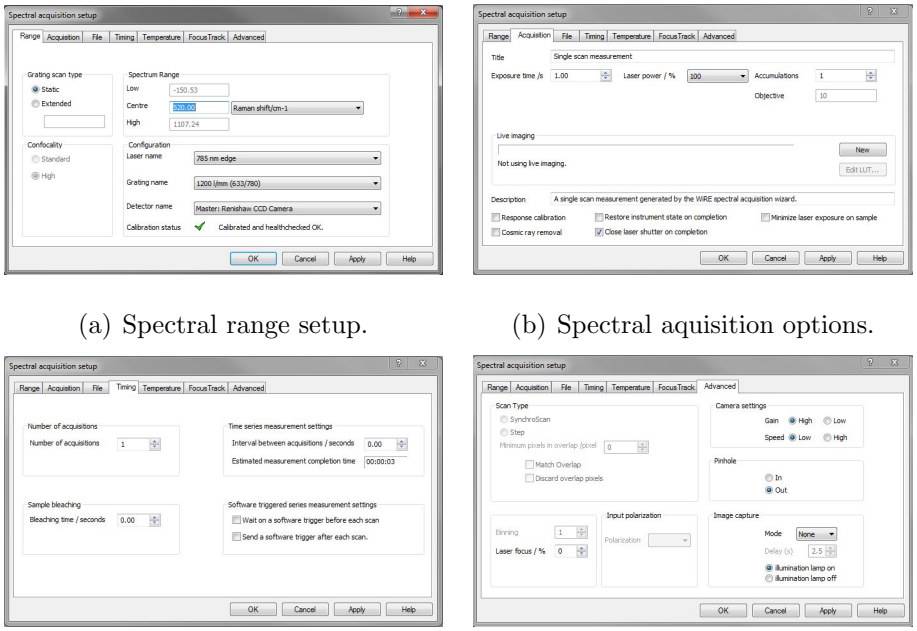
Following the edge filter, the beam then passes through a pre-slit lens which focuses the light into a slit. The light is then re-dispersed via a post-slit lens and is then collimated before passing into a triangular mirror where it is then directed onto a diffraction grating. The diffraction grating then disperses the collimated light according to wavelength. The Renishaw InVia system is equipped with two diffraction gratings. The choice of grating is important as it affects the spectral resolution that can be achieved with the instrument. As the Renishaw is an ‘off

the shelf' system the entrance slit aperture, focal length, laser wavelength and CCD pixels per nm are fixed meaning that the only 'variable' element that affects the spectral resolution is the diffraction grating. The system has 2 grating options of 1200 g/mm and 2400 g/mm which the system is optimised for use with the 785 nm and 532 nm lasers respectively. The Wire software automatically selects the most appropriate grating for the chosen wavelength. Once the light has been dispersed in the spectrometer by the diffraction grating it is then directed onto the CCD array via the second face of the triangular mirror. The CCD detector is cooled by a Peltier cooling system with a small fan to assist with airflow. The individual measurement parameters including the grating choice, laser line, acquisition times and laser power will be detailed in the relevant results chapters.

3.2.2 Taking a Raman Spectrum

The Raman system is controlled entirely by the Renishaw Wire software. Therefore all measurement parameters are selected from within the software. The software has a simple measurement setup window in which all of the spectral parameters for the measurement are selected. Figure 3.3 is a screenshot of the spectra measurement set-up window for a basic spectral acquisition. The spectral setup is split into seven tabs regions which have settings in each. The Raman system used in this work did not have a temperature cell therefore the temperature setting were set as default as were the Focus-Track settings.

When setting up a spectral acquisition, there are different scan types that can be performed from the Wire software, extended scans span across the entire wavelength range of a spectrum (100-3200 cm^{-1}) but require a longer laser light exposure time (≥ 10 s). They are performed by moving the diffraction grating and the charge across the CCD detector simultaneously. The other type of scan that is possible in the Renishaw system is a 'static' scan where the grating remains in a fixed position and the scan interrogates a smaller spectral range. The work in this thesis used static scans throughout. This is partly due to the majority of the spectral information from biological samples being in the fingerprint region of the spectrum (610-1720 cm^{-1}). It was also in part due to wanting to minimise laser



(a) Spectral range setup.

(b) Spectral aquisition options.

(c) Timing settings.

(d) Advanced measurement settings.

Figure 3.3: Screenshots of each stage of a spectral measurement setup.

exposure time on the samples and the fact that static scans can be performed a lot more quickly than extended scans. The other parameters in each tab were optimised for each type of spectral measurement and substrate used in this work and the details of the measurement conditions are detailed in each chapter that Raman spectral measurements were used. Setting that are not mentioned in the optimisation of each chapter were kept as the default settings in figure 3.3.

Microscope Objectives

The system used in this work includes 5x, 10x, 20x, 50x and 100x dry objectives and one 20x water immersion lens. The choice of lens in this work was dependant on the maximisation of light collection rather than spatial resolution in most cases as this work did not generally involve imaging. Dry spectra in this thesis were taken using a 50x objective unless otherwise stated. Liquid spectra were taken with 10x objective unless otherwise stated.

All of the objectives within the system were Leica objectives, however, they differed in their lens coating. Some objectives were N Plan objectives and one was a Hi Plan objectives (10x objective). In general the N plan objectives have a

lower spectral contribution than Hi Plan objectives due to the differences in their coatings causing a lower internal contribution to spectra from the lens. Despite this, the 10x objective was selected for liquid measurements due to its longer working distance. This allowed higher throughput of measurements which are essential for translation. An example of the objective contribution from the 10x Hi Plan objective can be seen in Figure 3.4.

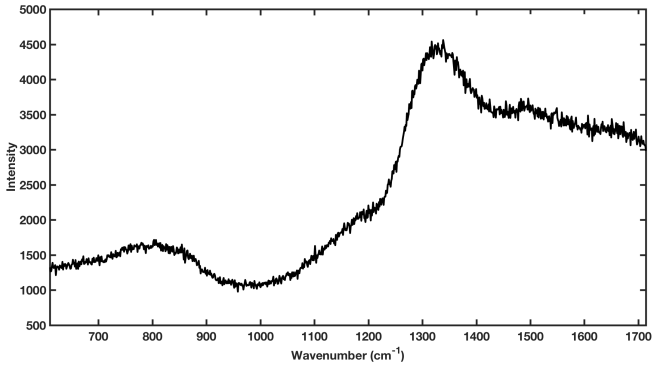


Figure 3.4: HiPlan objective (10x) spectral contribution.

Internal Calibration

The Renishaw system includes an internal silicon reference sample which can be used to calibrate the beam path and CCD area along with a daily check of relative laser power. The sample is located in a sealed part of the pod above the microscope. The position of the sample is therefore only adjustable via automated calibration. An internal silicon reference scan was taken daily before measurements for each laser line. The system can also be calibrated with a Neon calibration source (NeAr) which was conducted at the commencement of the work to calibrate the system (x axis) to Neon spectral lines.

Internal reference parameters

An internal reference spectrum was taken each day for each laser before measurement. After each reference spectrum a curve was fitted to obtain the intensity and position of the reference peak. The spectrum was calibrated daily to produce spectra with an intensity of higher than 18,000 and 15,000 counts for the 785 nm

and the 532 nm laser lines respectively. The position of the peak was calibrated to 520 $cm^{-1} \pm 0.5\ cm^{-1}$ daily.

Data handling

All Raman spectral data were collected using the Renishaw Wire software. Spectra were automatically saved into .wdf format. Spectra were then converted to .txt files before exporting into MATLAB for spectral pre-processing and analysis (detailed in Chapter 4).

3.3 Cleaning methods (substrates)

A stainless steel multi-well plate has been developed in this work (detailed in Chapter 5). Post spectral measurements the plate was washed with washing up liquid (Fairy, UK) and each well scrubbed. It was then rinsed with deionised water, ultrasonicated for 5 mins in ethanol, then ultrasonicated for 5 mins in deionised water. The well plate was then dried with pressured air.

3.4 Attenuated total reflectance spectroscopy

The PerkinElmer spectrum two attenuated total reflectance (ATR-FTIR) spectrometer was used for all ATR-FTIR data acquisition during this work. The system is closed and completely controlled via the PerkinElmer Spectrum software. The spectrometer system comprises an infra-red light source (GLOWBAR) that emits an effervescent wave through the system as in Figure 3.4. The IR light is then directed via a mirror to a Michelson interferometer, after this the beam is directed further around the system via a series of mirrors. The beam then passes through a Jacquinot stop (aperture) to limit the size of the spot size on the detector. The beam then goes past another mirror into the removable element of the spectrometer. For this work a diamond ATR element configuration (PerkinElmer,USA) was used. The ATR acts as a substrate for the sample as well as an element of the beampath. The beam passes through the ATR element (single bounce through the diamond crystal) with path length of 3 μm above the

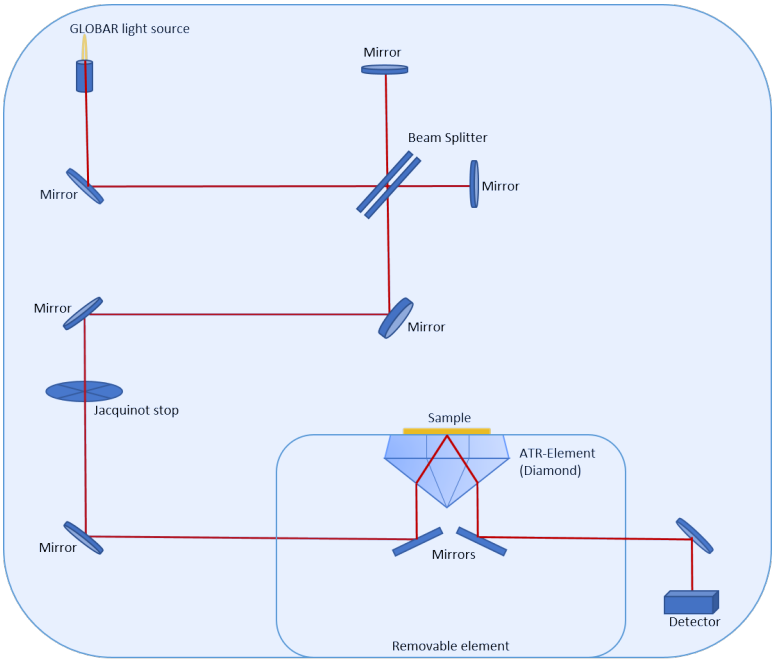


Figure 3.5: PerkinElmer Spectrum Two Beampath.

crystal. If the sample is IR active it will absorb IR radiation, the remaining transmitted light is then passed back onto another mirror and onto the detector. The detected interferogram then gets transformed into a spectrum via a fast Fourier transform (FFT) algorithm that is in-built into the Spectrum^{TM}software. For a full treatment of Fourier Transforms of an interferogram the reader is directed to [1].

3.4.1 Environmental contributions

The beam that enters the detector (the interferogram) contains contributions from the instrument, ATR element and also any air coming into contact with the beam (CO_2 molecules are highly IR active see Chapter 2). As a result, the spectrum produced by the spectrometer will contain all of this information which has the potential to mask the contribution from the sample of interest. As a result a background reference spectrum must be taken before placing a sample onto the ATR element. This background is then subtracted from subsequent measurements to negate the environmental contribution effects to the spectrum.

Background scan parameters

Background spectra were taken between each patient sample via the automated background collection within the Spectrum ^{TM}software. Background spectra were in the full spectral range 400-4000 cm^{-1} and each scan consisted of 24 accumulated scans. The background was automatically subtracted from the data for each sample. Figure 3.5 shows an example of the background spectrum from the Spectrum Two that was used for the work during this project. The spectrum is a transmission energy spectrum. The spectrum is dominated by the spectral bands (c-c) from the diamond ATR-emelment with peaks in the region from 1900-2400 cm^{-1} as well as a broad peak in the 2400-2800 cm^{-1} region of the spectrum. However, this spectrum also includes a spectrum of the environment surrounding the surface of the ATR crystal i.e air. There are therefore also some bands in this spectrum that can be attributed to CO2 at 2350 cm^{-1} and 667 cm^{-1} [2].

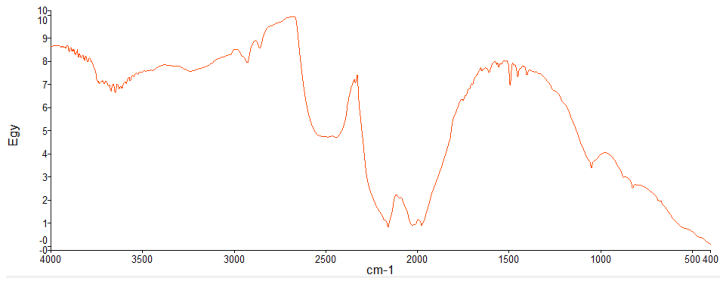


Figure 3.6: Energy Spectrum of the background components of the FTIR spectrum.

3.4.2 Measurement conditions

All spectra were acquired in the Spectrum software package using the full wavenumber range of 400-400 cm^{-1}. Each spectrum consisted of 24 accumulations and was exported as a .csv file after each measurement acquisition for analysis in the MATLAB environment. For each serum sample 2 μl was pipetted onto the measurement substrate before data acquisition. For liquid FTIR measurements serum samples were pipetted directly onto the ATR crystal and spectra were taken to avoid immediately to avoid evaporation effects.

Bibliography

[1] Brian C Smith. *Fundamentals of Fourier transform infrared spectroscopy.* CRC press, 2011.

[2] Peter Larkin. Infrared and Raman Spectroscopy. *Infrared Raman Spectrosc.*, Chapter 2:7–25, 2011.

Chapter 4

Spectral processing and analysis

4.1 Introduction

Raman and FTIR spectral signal processing is often divided into two main aspects: pre-processing and analysis. Pre-processing is the process of removing unwanted features from the spectral dataset and preparing it for analysis. The need for pre-processing arises from the dependence of vibrational spectroscopy on light scattering or absorbance effects that can have small variations between measurements causing small spectral differences even if the test sample is identical. Pre-processing spectra involves a range of steps including basic quality control, calibration, normalisation and background subtraction of the data.

Once a dataset is pre-processed it can then be analysed by means of univariate or multivariate chemometric analysis to investigate spectral differences between different cohorts of interest. The pre-processed data can also then be used to develop machine learning based models such as random forest, support vector machines or artificial neural networks. The goal of the model building is to utilise a large library of previously acquired spectra from samples within known target groups (e.g. cancer, control) to form a model 'training dataset'. This then allows for spectra from an unknown group as a part of a 'testing dataset' to be fed into the model and compared to the training dataset in order to produce a comparative diagnostic result (Figure 4.1).

Figure 4.1: Flow chart showing the different components of spectral analysis needed in order to interpret spectra for diagnostic purposes.

This chapter will firstly describe principles and methods of spectral pre-processing techniques used during this work. Secondly, it will describe the main principles of the chemometric analysis techniques used for unsupervised classification of spectral data. Thirdly, the chapter will briefly describe the two main supervised classification techniques for model building and analysis and the methods for measuring diagnostic accuracy and results and how this data is usually visualised. Finally, this chapter will address the issue of automated processing. Automated processing is crucial for the translation of spectroscopic techniques; the automated spectral pre-processing and analysis application that has been developed for use in MATLAB will be described. This application allows a general user to pre-process raw data, perform univariate and chemometric analysis and finally compare processed spectra to a diagnostic model that has been built using a training dataset from this work.

4.1.1 Software considerations

All Raman spectra were collected using an InVia Raman spectrometer with the Wire 4.1 software as detailed in Chapter 3. Spectra were individually saved in the default format (.wdx) in order to maintain full spectral information. The spectra were then exported as .txt files for all analysis outside of the Wire environment. All FTIR spectra were collected using the Spectrum software (PerkinElmer, USA) and saved in the default format (.spc). Spectra were also exported as .csv files for all pre-processing and analysis. The spectral processing and analysis from both the Raman spectrometer and the attenuated total reflectance (ATR)-FTIR spectrometer were performed in the MATLAB environment. To perform operations on the data, individual codes (a mixture of scripts and functions) were developed to perform each of the steps required in the data analysis. The codes were then packaged into a MATLAB App for automated spectral processing using the application developer in the MATLAB software. All codes that have been developed can be seen in Appendix C. Code that has been used from open source software and edited or code developed by outside agencies will have the authors name in the title of the attached code.

4.2 Data preprocessing

Diagnostic models must be able to be expanded to multiple instruments and potentially multiple sites in order to become valid. Data pre-processing is an essential step in this process as it helps to minimise the variation in spectra caused by external sources such as laser power, cosmic rays, etc. and allows for maximising variation caused by changes in the sample.

4.2.1 Quality Control

The first step in preprocessing is ensuring that spectral data is above some minimum quality threshold. Therefore, quality control procedures must be implemented to ensure model performance is high. This is true for any type of medical test. As an example, a blood pressure monitor that has taken data from an adult using a cuff meant for a child will give invalid results, therefore the blood pressure monitor has in-built systems to detect if this is the case and instruct a user to re-take data using the correct methods. To ensure the quality of any spectral analysis a basic spectral quality control procedure was implemented. The quality control procedure required spectra to meet two basic requirements- minimum intensity and the absence of cosmic rays.

Minimum Intensity

The measurement parameters for each measurement vary depending on the sample wavelength and if the sample has been analysed in dry or liquid form. Details of the optimised measurement parameters for serum samples are detailed in chapter 5. Table 4.1 gives the summary of the minimum intensities required to pass the quality control for each sampling modality used in this work. The intensity of the 532 nm excited sample is considerably higher than the other sample modes as the 532 nm spectrum is a resonance spectra naturally causing much higher intensities. Any spectra that did not meet these minimum requirements were repeated wherever possible and excluded from analysis if repeat measurement was not possible.

Table 4.1: Summary of minimum intensities for different sampling modalities

	785nm Laser		**532nm Laser**	
	Liquid	Dry	Liquid	Dry
Minimum intensities (counts)	5,000	10,000	500,000	N/A

Cosmic ray removal

Cosmic radiation can hit a CCD detector causing spectral interference. Cosmic rays cause large peaks in spectral datasets not due to the molecular properties of the sample. A crucial part of the quality control for spectra during this work was ensuring that all spectral data were free of cosmic rays before undergoing spectral analysis. This was achieved via manual detection and also software based cosmic ray removal. Manual cosmic ray removal was used when a cosmic ray was detected during the spectral acquisition of data, in this case acquisition was repeated immediately until spectra without cosmic rays were acquired. In the case where datasets were acquired over longer periods of time such as during mapping measurements the Wire 4.1 automated cosmic ray removal software was used. The removal of cosmic rays within the software depends on the software detecting the cosmic rays based on user defined width and intensity of the cosmic ray feature (Figure 4.2).

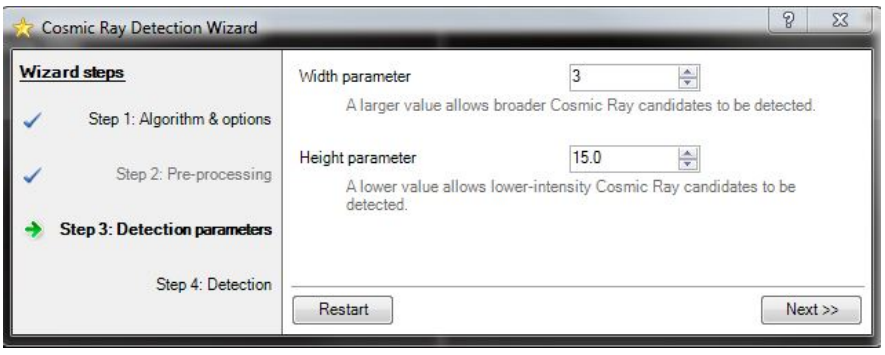
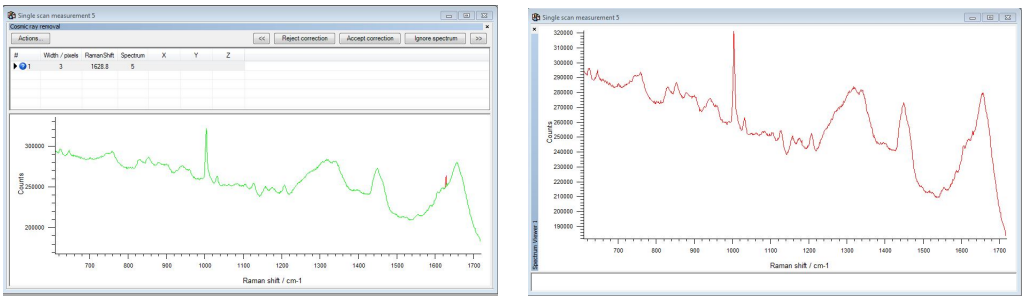


Figure 4.2: Selection parameters for cosmic ray removal within the Wire 4.1 software.

The software scans the spectrum for spectral features that fit into selected

cosmic ray parameters and highlights the features it suspects to be cosmic rays (figure 4.3a). The peaks that the software then suspects to be cosmic rays are highlighted and if the peak is indeed a cosmic ray the suspect peaks are accepted as cosmic rays and removed automatically within the software as in figure 4.3. Cosmic ray corrected spectra were then saved as .txt files for further analysis provided that they also met the other quality control criteria.



(a) (b)

Figure 4.3: Example of the cosmic ray detection and removal; spectrum with cosmic ray detected (red) with remaining spectral features (green)(a) and the resulting cosmic ray corrected spectrum (b) from the Wire 4.1 software.

4.2.2 Wavenumber standardisation

A single Raman or FTIR spectrum is a two dimensional object consisting of n pairs of (x,y) coordinates where x = wavenumber cm^{-1} and y = Intensity/Absorbance. A dataset of multiple spectra will then be a m by n array of data with m being the number of spectra. When comparing large datasets of spectra it is important that the positions of the x coordinates along a spectrum are consistent so that peak positions and intensities can be interpreted correctly. FTIR spectra are collected using a lithium tantalate mid infrared detector and therefore the x coordinates across a dataset will always be constant. However, there may be differences between instruments. When using dispersive Raman spectrometers that rely on diffraction gratings and CCD cameras for photon detection, the overall dimensions of the spectral dataset can change (table 4.2). The overall number of data points is constant, e.g. 1015 coordinate points between 610 cm^{-1} and 1720 cm^{-1} spectrum, but the spacing of the data points can be irregular depend-

ing on the diffraction grating, laser wavelength and 'binning' of the CCD array [1]. Unlike the FTIR spectrum the wavenumber coordinates on the spectrum are not always constant for one instrument. Therefore FTIR data taken from different instruments and all Raman spectra must be wavenumber standardised to allow spectral comparisons of peak intensities.

Table 4.2: Example of spectral dimensions and coordinate positions for FTIR and Raman spectra for different spectra (S).

	FTIR	**Raman**
S	$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$	$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
$S_{(1)}$	$\{(400, y_1), (401, y_2), \ldots, (4000, y_n)\}$	$\{(610.00, y_1), (611.22, y_2), \ldots, (1719.21, y_n)\}$
$S_{(2)}$	$\{(400, y_1)_2, (401, y_2)_2, \ldots, (4000, y_n)_2\}$	$\{(610.07, y_1)_2, (611.29, y_2)_2, \ldots, (1719.26, y_n)_2\}$
$\vdots$	$\vdots$	$\vdots$
$S_{(m)}$	$\{(400, y_1)_m, (401, y_2)_m, \ldots, (4000, y_n)_m\}$	$\{(610.01, y_1)_m, (611.24, y_2)_m, \ldots, (1719.11, y_n)\}_m$

Typically there are two methods for shifting spectra; the wavelength axis can be shifted by a set distance to align to a reference peak or the the wavelength axis can be interpolated to a brand new wavenumber axis. The first method is straightforward and does not affect the intensities of the spectra, however depending on the size of the shift this can lead to needing a cutoff at either end of the spectrum which has the potential to loose some spectral data at either end of the spectral window (Figure 4.4). Furthermore, this method doesn't correct unequal spacing along the wavenumber axis of the spectra so is unsuitable for a group of spectra that has different amounts that need to be shifted [2]. Interpolating the wavenumbers directly onto a wavenumber axis gives the advantage of having evenly spaced wavenumbers making spectral comparison easier. However depending on the position in the new wavenumber axis the maximum peak of the spectrum can shift position [2]. A hybrid method of first finding the maximum in a peak and then aligning and interpolating the spectra was developed for the processing of data presented in this thesis. The hybrid method of shifting spectra first finds the position of the maximum peak in the region of 990-1050 cm^{-1}, then the spectra are shifted and interpolated to standardise the wavenumber. It was

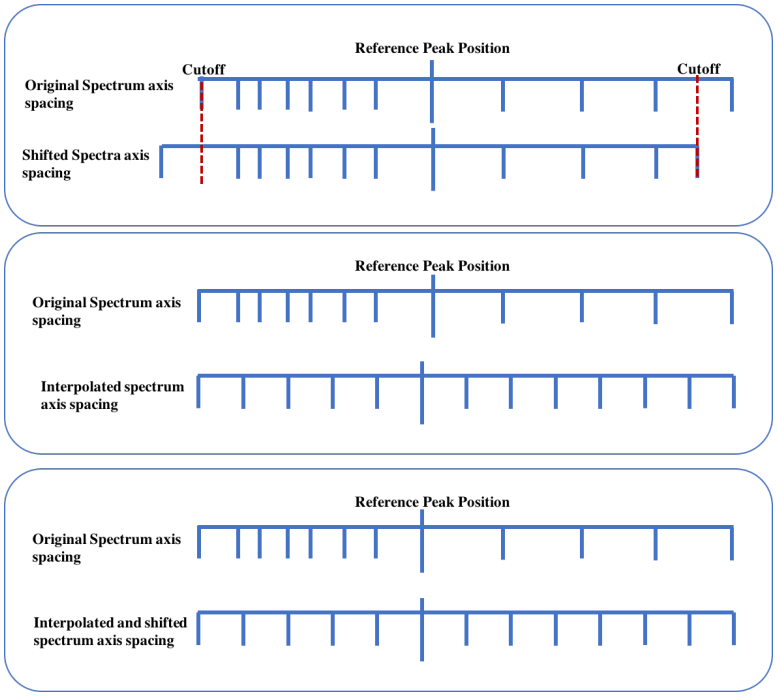
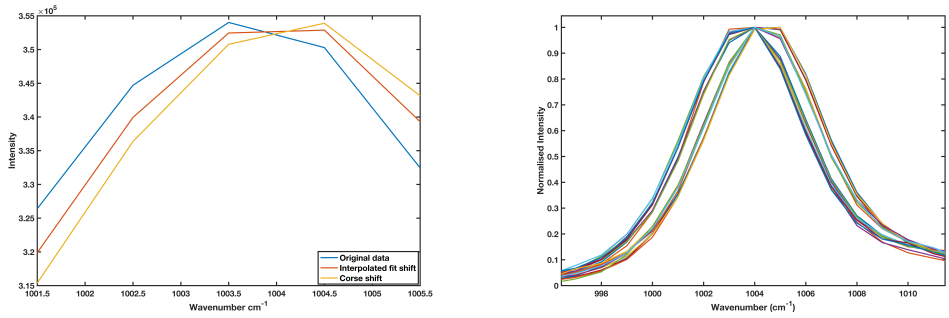


Figure 4.4: Summary of the different methods of shifting wavenumbers with straight shifting(top) interpolation(middle) and the hybrid shift and interpolation (bottom).

found that the actual maximum of the peak was too coarse a measurement so a 4th order polynomial was fitted to the absolute maximum of the data points and its two nearest neighbours above and below the maximum point. Figure 4.5 shows the effect of the hybrid correction method on the peak of the phenylalanine peak compared with the standard interpolation method. This method achieves spectra that have standardised wavenumber axes and spectra with a peak maximum for phenylalanine in the same position. All spectra that were imported into MATLAB underwent this wavenumber standardisation method.



(a) (b)

Figure 4.5: Comparison between original, coarsely interpolated and interpolated-shifted spectra (a) and an example of normalised, shifted data using the hybrid method developed for this thesis (b). Example data were taken from dry serum spectra with the 785 nm laser.

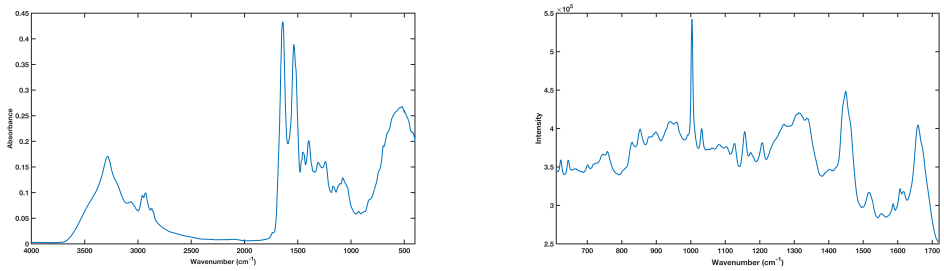
4.2.3 Baseline subtraction

Baseline subtraction is the process of removing slow varying background signals that are present in both Raman and IR spectra. The ideal removal of a background contribution should leave as much spectral information in the spectrum as possible whilst removing unwanted features. Therefore, the ideal method does not remove any small Raman features that could contribute to classification. There are many different options available for baseline subtraction within the literature such as polynomial fitting [3–5], Savitzky-Golay derivatives (S-G derivatives) [6], rolling circle filters [7], wavelet transformations [8], asymmetric least squares fitting [9], Fourier transforms and multiplicative scatter correction methods [10]. One of the main issues regarding the success of background subtraction methods is that generally most techniques rely on input parameters that are fed into an algorithm and visually or computationally inspected afterwards. This means that there is an element of subjectivity within most background subtraction methods. The performance of these techniques has been reviewed in terms of creating model datasets [11, 12]. The results show that different methods work well for spectra of different types and to choose the best performing method it is essential to understand the origin of backgrounds within spectra.

The IR background is generally a slowly oscillating background caused by Mie

scattering, however this effect is amplified when using sample specimens such as cells and cellular components that are of similar magnitude to the wavelengths of IR light being investigated [13, 14]. Serum samples do not contain any cellular components and therefore appear to an IR substrate as a much flatter more uniform sample than cells so the background effects are reduced. Nevertheless, the spectra in this work were background subtracted using derivative and polynomial baseline subtraction methods to account for any potential contributions from scattering effects.

The Raman background can originate from a variety of sources such as the internal fluorescence of a sample, the experimental substrate and/or other sources such as thermal fluctuations in the CCD detector; these factors have an effect on the shape of a spectral baseline [3]. Figure 4.6 shows typical raw spectrum from dry serum using ATR-FTIR and Raman spectroscopy that exhibit typical backgrounds seen during this work. The IR spectrum has a fairly flat baseline shape with the baseline close to zero whereas the Raman spectrum exhibits a baseline shape (due to sample autofluorescence) that slopes from left to right making comparisons of intensity across the spectrum difficult. To process the spectral data within this thesis, it was decided iterative polynomial background subtraction, derivative spectra and a rolling circle filter would be tested and optimised for the dataset background subtraction.

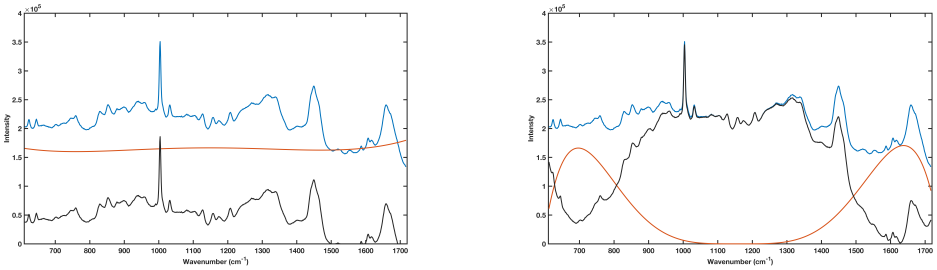


(a) Typical raw FTIR serum spectrum (b) Typical raw Raman spectrum

Figure 4.6: Typical raw FTIR spectrum of dried serum with the ATR-FTIR with a small background contribution (a). Example of a Raman spectrum of dried serum exhibiting background fluorescence (b).

Iterative polynomial baseline subtraction

One of the most common methods of background subtraction of both Raman and FTIR data is a polynomial background subtraction. The method fits an n^{th} order polynomial beneath the spectra to estimate the shape of the background [5]. This method is often used to batch process spectra so that the baseline shape can be fitted using an iterative method to gain the best estimation of the baseline shape. Figure 4.7 gives examples of raw spectra and baseline estimations of different polynomial orders to the spectra. It is clear that the order of the polynomial chosen for the spectra is crucial to gaining a good baseline estimation as a change between a 4th order and 7th order polynomial has an enormous difference and can lead to badly fitted baselines.



(a) 4th order polynomial fit.

(b) 7th order polynomial fit.

Figure 4.7: Example of a baseline fitted from a fourth order polynomial (a) and a seventh order polynomial (b). Polynomial fitting was based on five spectra; one is shown for clarity.

Furthermore, as the iterative polynomial subtraction method fit estimates a baseline for the particular batch of spectra that are being analysed, the method is subject to different polynomials being fitted for each batch of spectra that could lead to model over-fitting. Derivative spectra and rolling circle filters are not susceptible to these issues so these were chosen for the work within this thesis.

Derivative spectra

First and second order spectral derivatives are commonly used as a background subtraction method for both IR and Raman datasets. The first derivative of a spectrum calculates the rate of change of spectral intensity with respect to (w.r.t.) the wavenumber axis, therefore it is given by

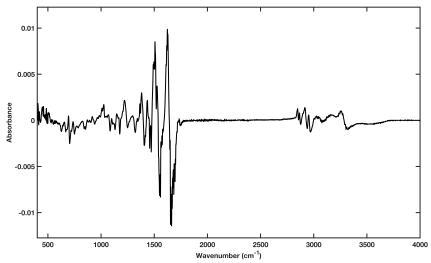
$$I' = \frac{\delta I}{\delta x}, \tag{4.2.1}$$

where I' is the first derivative of the spectral intensity with respect to x, therefore is the gradient of the tangential line at each point in the spectrum, I is the original intensity vector and x is the wavenumber vector. The resultant derivative spectrum gives the global maxima and minima of spectral peaks so can be used to successfully resolve sharp narrow spectral features[1]. Similarly, the second derivative of the intensities is then given by

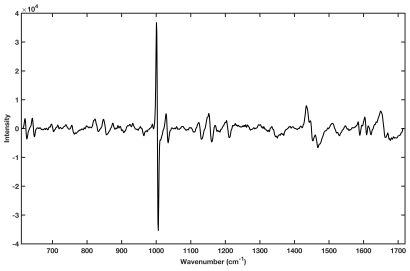
$$I'' = \frac{\delta^2 I}{\delta x^2}. \tag{4.2.2}$$

The second derivative gives the local maxima and minima features along the spectrum so can therefore be used for deconvolution of spectral peaks. This method is mainly used in FTIR spectra as the bands tend to be broader and more convoluted than sharper Raman features. Figure 4.8 shows the first and second derivatives on the Raman and IR spectra in Figure 4.6. It is clear that the second derivative in the case of the Raman spectra introduces lots of spectral noise only making deconvolution of the largest and broadest spectral bands.

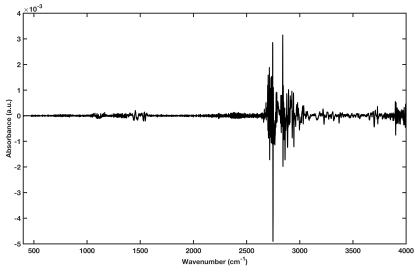
[1]This is due to the amplitude of the derivative of a peak being inversely proportional to the width of a peak [15].



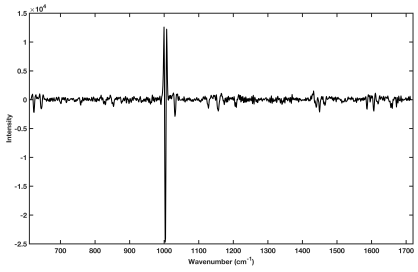
(a) First derivative FTIR serum spectrum.



(b) First derivative Raman spectrum.



(c) Second Derivative FTIR Spectrum.



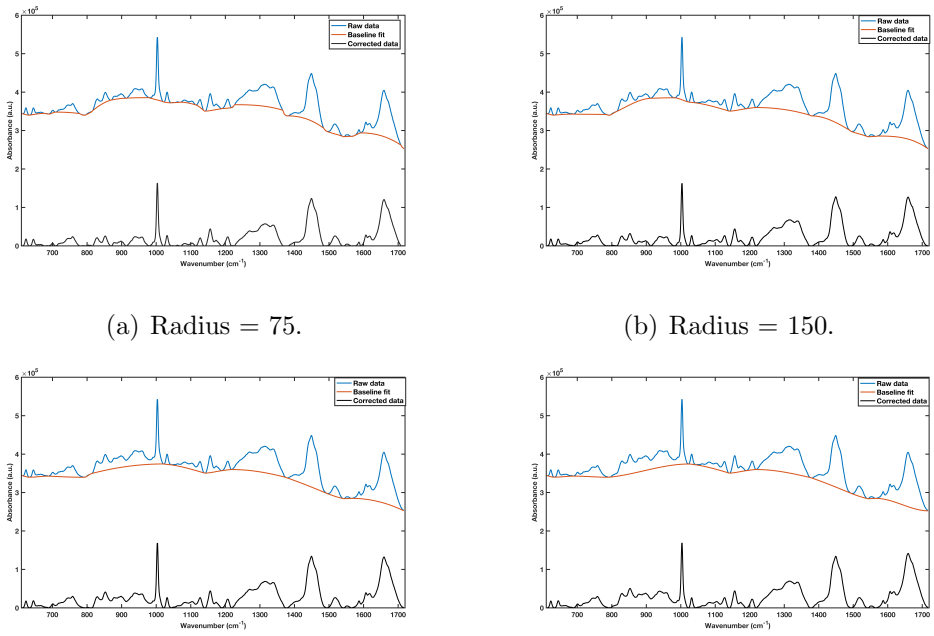
(d) Second Derivative Raman spectrum.

Figure 4.8: Example first and second derivative FTIR and Raman spectra of dried serum.

The most commonly used method of spectral differentiation is the Savitzky-Golay(S-G) algorithm [6]. The technique estimates a polynomial through the intensity points on a spectrum based on a window of wavenumber points then estimates the derivative from the polynomial fit. This technique has the advantage that as the derivative is calculated from a window of data points so the spectral noise introduced by performing the derivative is lowered [16]. However, the noise introduced is not eradicated. The method also carries the disadvantage of having more input parameters than polynomial fitting as the user must optimise the window length, the derivative order and also the polynomial order to produce effective results. Furthermore, it brings the disadvantage of loosing spectral information at either end of the spectrum. As an example, if an SG derivative is calculated using a window length of 7 points, then the spectra will 'loose' 7 data points at either end of the spectrum. A code was developed for calculating first and second order spectral derivatives. To allow for the loss of data at either side of the spectrum, zeros are added along the data points lost to keep the spectral dimensions consistent.

Rolling circle filter subtraction

A rolling circle filter based background subtraction method had previously been developed within the University for application to precision Raman measurements for the Katrin experiment in Germany and also in application to SERS pH sensor development [17, 18]. The technique has proven successful for batch processing of biological Raman spectra especially those with many peaks and generally non-uniform curved background shapes. The technique relies on a rolling 'circle' that rolls along the base of the spectrum. The algorithm then selects the minimum distance between the circle and the data along the radius of the circle. Figure 4.9 shows examples of the RCF background subtraction method using different radius sizes, the larger the radius the less the circle penetrates into the spectral features. The radius is the only parameter that is selected within this algorithm; for all spectra processed during this work the default radius was set to 150.



(a) Radius = 75.

(b) Radius = 150.

(c) Radius = 225.

(d) Radius = 300.

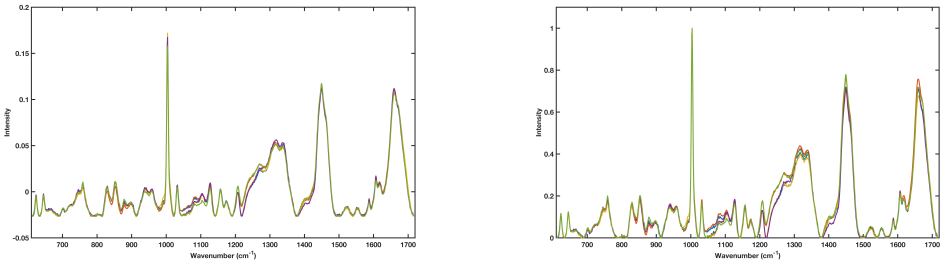
Figure 4.9: Examples of the rolling circle filter background with a radius of 75 (a), 150 (b), 225 (c) and 300(d).

This then effectively determines the contact points from which to calculate the baseline, which was then calculated using a piecewise cubic hermite interpolating

polynomial (pchip) from the contact points. The main advantage of the rolling circle based method is that the radius of the circle remains constant for all data put into the algorithm regardless of whether it has been batch processed or not. This allows consistent contact point selection and baseline removal across different batches of datasets.

4.2.4 Normalisation

Once spectra have been quality tested, wavenumber standardised and background corrected, the final step in pre-processing is usually to normalise the dataset. This is the process of removing differences in scattering intensity and absorbance due to sample thickness, laser power, etc. There are many methods of spectral normalisation for use when building machine learning algorithms such as area under the curve normalisation, vector normalisation and min/max normalisation [16]. For this work vector normalisation and min-max normalisation methods were chosen and codes that include the equations describing the both methods can be found in Appendix B. Briefly, the vector normalisation transforms the spectrum such that the length of the spectrum is equal to a unit vector of length 1 and the min/max normalisation transforms the data by subtracting the minimum spectral intensity of each spectrum from all of the spectral intensities then dividing by the difference between the maximum and minimum value in each spectrum thus giving the new data a range between 0 and 1 where the maximum point in the spectrum is at 1. The min/max normalisation method used in this work was modified for Raman datasets so that the spectral intensities were always normalised to the intensity of the phenylalanine (Phe) peak (1003-1004 cm^{-1}). Figure 4.10 shows an example of average datasets that have been vector normalised and a dataset which has been normalised to the maximum intensity of the Phe peak. The spectra that were vector normalised are subject to some parts of the spectrum becoming negative. Therefore for Raman spectra in this work the Phe normalisation was used for Raman spectra, min/max normalisation was used for FTIR spectra and vector normalisation was only used when derivative background subtraction had been used as the baseline correction method.



(a) Vector normalised spectra

(b) Phe peak normalisation.

Figure 4.10: The effects of vector normalisation on representative Raman spectra (a) and normalisation to the Phe peak (b).

4.3 Chemometric spectral analysis and machine learning

The datasets collected for the work within this thesis are large multidimensional arrays of data. In the case of Raman data the typical spectral dimensions were [n x 1015 x 2] and the FTIR datasets had dimensions [n x 3000 x 2]. To efficiently analyse the relationships between all of the variables it was appropriate to use multivariate chemometric data analysis techniques to reduce the dimensionality of the data and allow for better visualisation and analysis. This work used a combination of univariate, multivariate and classification algorithms to analyse the spectral data and build diagnostic models.

4.3.1 Unsupervised vs supervised techniques

Classification and clustering are two important concepts in analysing spectral data. Classification generally relies on giving a learning algorithm the classes that samples belong to to optimise modelling of already known classes and then predict the result of unknown samples. Clustering methods rely on 'unsupervised' learning and therefore do not take into account which class a sample belongs to (i.e. cancer or control).This is useful when results need to be unbiased for optimisation or quality control measurements. It is also a useful method for the discovery of new groups or clusters within a dataset. Both types have been used

within the data analysis for this thesis. Firstly, unsupervised methods such as PCA will be described followed by supervised learning techniques such as partial least squares discriminant analysis and random forest learners.

Univariate analysis

Basic spectral analysis can be performed via univariate methods. These take into account one aspect of the spectra such as looking at mean spectra with spectral standard deviation or peak positions/shifts. For this type of analysis during this work the inbuilt MATLAB mean, standard deviation and variance functions were used. However, it is useful to define the properties of these functions as the principles are used in many of the multivariate analysis techniques. For a one-dimensional set of data the mean ($\bar{X}$) is given by;

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \tag{4.3.3}$$

where n is the number of samples. The standard deviation from the mean of the dataset (spread of the dataset) is given by;

$$SD = \sqrt{\frac{\sum_{1}^{n}(X_i - \bar{X})}{(n-1)}}. \tag{4.3.4}$$

Finally, the variance of a dataset is given by the standard deviation squared;

$$Var(X) = \frac{\sum_{1}^{n}(X_i - \bar{X})}{(n-1)}. \tag{4.3.5}$$

Principal component analysis

Principal component analysis (PCA) is a robust method of reducing dimensionality from within a dataset. The technique does this by performing a matrix transformation on a multidimensional data set ($\mathbf{X}$) with dimensions (d) onto a new basis ($\mathbf{Y}$) with reduced dimensions (m). The transformation vector ($\mathbf{P}$) of the data matrix is such that the variance within the dataset is maximised and the correlation between attributes (or wavenumbers) is minimised [19]. In terms of a

spectrum, maximises the wavenumbers in the spectrum which cause the largest amount of variance within the dataset. Once the new reduced basis has been found, the original data matrix can be projected onto the new basis with reduced dimensions (Figure 4.11). The projection of the data onto the new basis set can

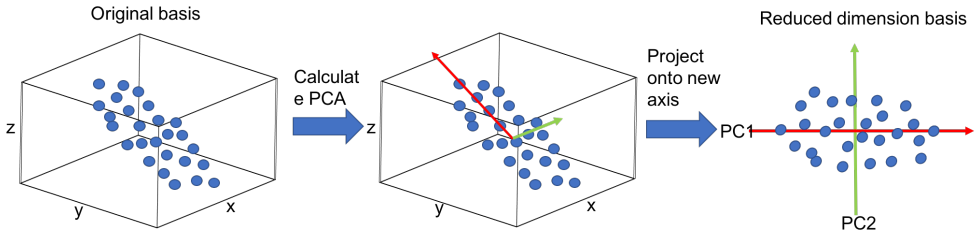


Figure 4.11: PCA takes the original d-dimensional basis set for the data and projects it onto a new orthogonal bases with reduced dimensions such that the variance in the data is maximised.

be written as a dot product between the unknown transformation vector $\mathbf{P}$ and the data matrix $\mathbf{X}$ [2].

$$\mathbf{Y} = \mathbf{P} \cdot \mathbf{X}, \tag{4.3.6}$$

where the coordinates of the transformed data onto the new basis set $\mathbf{Y}$ are known as the PC scores, and the rows of the transformation matrix $\mathbf{P}$ are known as the loading vectors for each principle component. To compute the matrix $\mathbf{P}$ such that the variance is maximised in the dataset, the covariance matrix of the new basis set must be considered. The covariance is the correlation between the different attributes of the data matrix, in terms of a spectrum this is a measure of the correlation between the intensities at each wavenumber within each spectrum for all spectra in the dataset. The covariance matrix of the data matrix may be defined as;

[2]The data matrix $\mathbf{X}$ is mean centered for all data within the PCA algorithm i.e $\mathbf{X} - \mu$, where μ is the mean of each attribute in the data matrix. If the data matrix is not mean centered before entering into PCA then the first computed principal component will be approximately equal to the mean of the dataset.

$$Cov(\mathbf{X}) = \mathbf{\Sigma_X} = \frac{1}{n-1} X \cdot X^T = \begin{bmatrix} Var(x_{1,1}) & Cov(x_{1,2}) & \dots & Cov(x_{1,d}) \\ Cov(x_{2,1}) & Var(x_{2,2}) & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_{n,1}) & Cov(x_{n,2}) & \dots & Var(x_{n,d}) \\ . & & & \end{bmatrix} \tag{4.3.7}$$

Therefore, the diagonal elements (n=d) of the matrix are equal to the variances of X and the off-diagonal elements are equal to the covariance of X. For the transformation of the data to be in the form that maximises variance and minimises correlation $\mathbf{\Sigma_X}$ must be a diagonal matrix with $Cov(x_{n,d}) = 0$. $\mathbf{\Sigma_X}$ can be diagonal if acted upon by the eigenvectors of $\mathbf{\Sigma_X}$ [19][3]. The eigenvectors satisfy the eigen-equation,

$$\mathbf{\Sigma_X}\, \hat{\mathbf{e}} = \lambda\, \hat{\mathbf{e}}, \tag{4.3.8}$$

where $\hat{\mathbf{e}}$ are the eigenvectors of $\mathbf{\Sigma_X}$ and λ are the corresponding eigenvalues. The eigenvalues with the largest value will correspond to the eigenvectors that explain the largest variance within the dataset. Furthermore, $\mathbf{\Sigma_X}$ is a symmetric matrix so the eigenvectors can be chosen to be unit length and in an orthonormal basis (i.e. $\hat{\mathbf{e}}^{-\mathbf{1}} = \hat{\mathbf{e}}^{\mathbf{T}}$).

The covariance matrix of the new dataset must also satisfy the same properties as the covariance matrix of X. Therefore, it also needs to be a diagonal matrix. This suggests that the unknown transformation vector P must be a rotation of Y such that $\mathbf{\Sigma_Y}$ is diagonal. The covariance of the new matrix basis (Y) can therefore be written in terms of the original data matrix covariance X;

$$\mathbf{\Sigma_Y} = \mathbf{P}\, \mathbf{\Sigma_X}\, \mathbf{P^T}. \tag{4.3.9}$$

[3]The equation for the diagonal matrix D, is given by; $D = \hat{\mathbf{e}}^{\mathbf{T}} \mathbf{\Sigma_X} \hat{\mathbf{e}}$, the eigenvectors are also often found using a singular value decomposition (SVD) this treatment can be found here [20].

Equation (1.3.3) and (1.3.6) suggest that

$$\mathbf{P} = \hat{\mathbf{e}}^{\mathbf{T}}. \tag{4.3.10}$$

So the unknown transformation vector is equal to the eigenvectors of the covariance matrix of the original data $\mathbf{\Sigma_X}$ transposed. Substituting this into equation (1.3.3) we find that the projection onto the new principal component basis;

$$Y = \hat{\mathbf{e}}^{\mathbf{T}} \cdot X. \tag{4.3.11}$$

so the PC scores = Loadings · X, where X is the mean centred original data matrix. The order of the principal component scores is then determined by the order of the eigenvectors which is sorted by the size of the corresponding eigenvalues. The eigenvector with the largest eigenvalue corresponds to PC1 (largest explained variance), PC2 the eigenvector with the second largest explained variance and so on.

The PCA function within MATLAB was used for all PCA in this thesis. The function returns the PC loadings, scores and the explained variances of each principal component. This is useful for analysing spectral data as the scores can be plotted to show the relationships between samples based on spectral variation on a reduced dimensional plot (Figure 4.12).

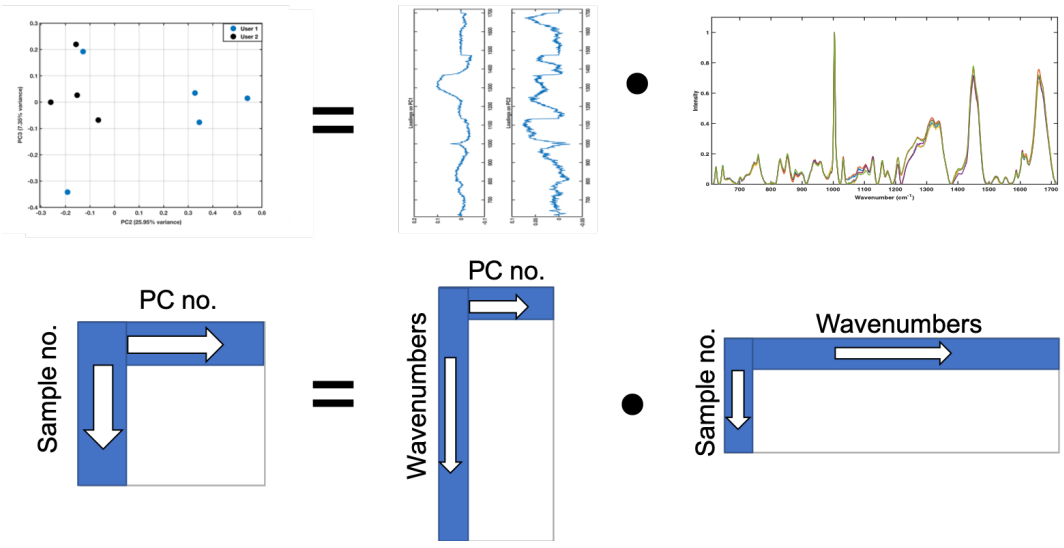


Figure 4.12: The PC score projection is a result of the dot product between the loading vectors for each PC and the mean centred data matrix.

Close PC scores then denote samples that are more similar to one another and differences in PC scores represent variation between samples. The loading vectors are useful in determining the source of the spectral variation (Figure 4.12) as when they are plotted against wavenumber they can show the spectral source of variation that contributes to that PC. The variances output (eigenvalues) can be used to compute the amount of overall variance explained by each principal component. This can be used to create a principal component number cut off one the cumulative explained variance is e.g. $\geq 98\%$. This process helps to cut out unnecessary PCs that describe the minutia of variance within the spectral dataset and often only contains spectral noise.

Agglomerative hierarchical clustering

Another method of reducing the dimensionality of a dataset is by using clustering methods. Clustering methods group similar data in clusters with different data in another cluster. There are many methods available for analysing spectral data such as k-mean clustering, density based clustering and hierarchical clustering methods [21,22]. During this work hierarchical clustering (HC analysis) was used to cluster data.

HC analysis can be split into two methods, divisive and agglomerative hierarchical clustering. This work used agglomerative HC in which each sample initially represents its own cluster, then clusters of similar samples are successively merged until the overall structure is obtained. The HC analysis is calculated using a three step algorithm which has in-built functions in the MATLAB environment.

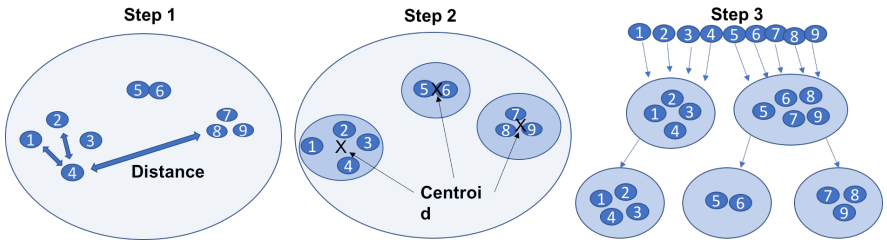


Figure 4.13: Step by step HCA.

Figure 4.13 shows an overview of the HCA algorithm, the first step in the algorithm calculates the Euclidean distances between each sample, the Euclidean

distance between two samples A and B is given by;

$$||a - b||_2 = \sqrt{\sum_i (a_i - b_i)^2}, \tag{4.3.12}$$

where a_i and b_i are the i^{th} attribute (wavenumber) of the spectrum for each sample. The second step in the algorithm determines the linkage or clustering between the samples. It starts with pairs of objects that are close together and pairs them into small clusters, then takes the small clusters and joins them to each other and larger clusters that are linked more closely together. This is calculated using the centroid of each smaller cluster and the centroid distances between the smaller clusters are joined to form larger clusters in the form of a hierarchical tree.

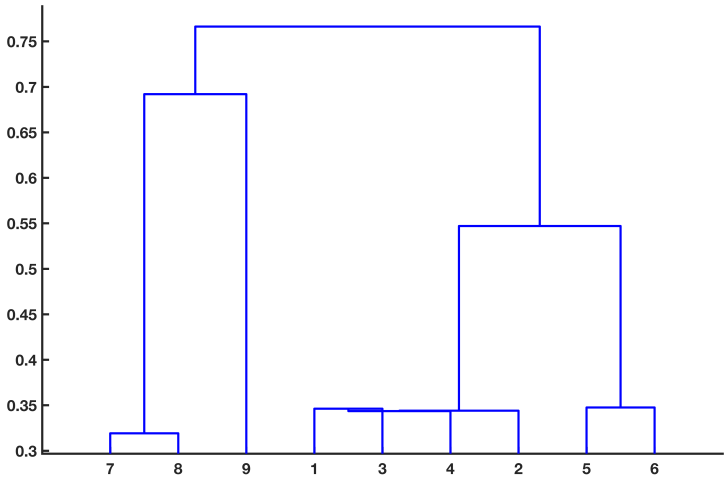


Figure 4.14: Example HC dendrogram.

The final step involves 'pruning' this tree to determine a cutoff and then plotting the results in the form of a dendrogram (Figure 4.14). The 'pruning' step involves specifying the maximum number of clusters allowed within the dataset. For example, in Figure 4.13 the amount of clusters could be cut to two, then the two most similar clusters out of the three would then be grouped into one large cluster. HC analysis was used during this work for external verification of PCA, to give an alternative clustering method to show the underlying relationships between the samples.

4.3.2 Supervised classification algorithms and machine learning techniques

Although unsupervised methods are very useful for determining large causes of variance between groups, they are not as useful for predicting the outcome of samples of unknown origin. To classify samples, supervised classification algorithms or machine learning algorithms are needed. However, these are not stand alone techniques; the data inserted into the model building techniques must go through the same treatment as the data for discriminant analysis. Therefore the overall methodology for creating diagnostic models should be considered as the overall process in Figure 4.1. The basic premise of these techniques is similar for all; a model is built on parameters from a library or 'training set' of data, the model aims to find a discriminator, separator or decision function to discriminate between the classes, the model performance is then cross-validated using a 'testing dataset' which is a set of samples with known results that are fed into the model and the model predicts which classification group a sample belongs to. Finally a cross-validated dataset can be used to predict the outcome of samples belonging to unknown groups. There are many different methods of building classification models. Within the literature for vibrational spectroscopy alone there are many such as PCA-linear discriminant analysis (PCA-LDA) [23], Euclidean centroid distances (EDC) [24], support vector machines (SVM) [25], artificial neural networks (ANN) [26] and random forest (RF) classifiers [27]. During this work two main classification algorithms were used, partial least squares discriminant analysis (PLS-DA) and random forest techniques (RF). The following section will briefly describe the algorithms used to build the models and it will also cover the rationale for the use of the two different methods for different applications within this work.

PLS-DA

PLS-DA is a multivariate analysis technique that can be used to investigate causes of variances within datasets[4]. It is based on partial least squares regression (PLSR) and can be used on datasets that have binary groups (e.g. cancer vs control) [29]. PLS regression can be used to form a linear multivariate model between two matrices (X and C), where in our case X is the spectral dataset and C is a set of classification labels. The discriminant analysis or PLS-DA is used where C is known and a PLS regression model is built between a dataset matrix (X) and a label matrix (C) where the label matrix contains numbers that correspond to groups within the dataset e.g. (-1 = Cancer, +1 = Control) and its length is equal to the number of samples in X. The goal of the algorithm is to

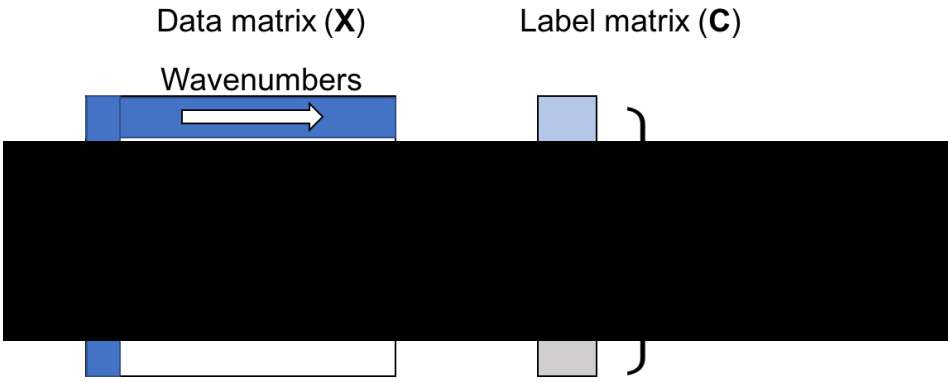


Figure 4.15: PLS-DA variables, adapted from [30].

draw a straight line (in the case of two variables) or a flat hyperplane (in the case of more than two variable) in space to separate two regions. In terms of spectral processing the algorithm aims to draw a hyperplane to distinguish a line between spectra of samples from control patients and spectra from cancer patients. The

[4]It should be noted that there are two types of PLS-DA algorithm (PLS1 and PLS2), all references to PLS-DA within this thesis refers to the PLS-1 method, for more information on the differences please see [28].

equations that describe the data and label matrix can be defined as;

$$\mathbf{X} = \mathbf{TP} + \mathbf{E}, \tag{4.3.13}$$

$$\mathbf{C} = \mathbf{Tq} + \mathbf{f}, \tag{4.3.14}$$

where $\mathbf{X}$ is the mean centered data matrix as in the PCA algorithm described previously, $\mathbf{C}$ is the label matrix, $\mathbf{T}$ is a common score matrix (PLS components), $\mathbf{E}$ and $\mathbf{f}$ are residual matrices and similar to PCA $\mathbf{P}$ and $\mathbf{q}$ are loading vectors. PLS-DA, in the algorithm described here differs from PCA as the components of $\mathbf{T}$ are orthogonal but the rows of the loadings matrix $\mathbf{P}$ are not and the eigenvectors do not necessarily reduce in succession in PLS-DA. The PLS-DA components are computed using the following algorithm;

1. Calculate the PLS weight factor;

$$\mathbf{w} = \mathbf{X}'\mathbf{C}. \tag{4.3.15}$$

2. Calculate the PLS score matrix components;

$$\mathbf{T_i} = \mathbf{t} = \frac{\mathbf{Xw}}{\sqrt{\sum w^2}}. \tag{4.3.16}$$

3. Calculate the x loadings via;

$$\mathbf{P_i} = \mathbf{p} = \frac{\mathbf{t}'\mathbf{X}}{\sum t^2}. \tag{4.3.17}$$

4. Calculate the c loadings;

$$q = \frac{\mathbf{C}'\mathbf{t}}{\sum t^2}. \tag{4.3.18}$$

5. Rearranging eq(4.3.13) and (4.3.14) the residual data matrices are then;

$$\mathbf{X_{resid}} = \mathbf{X} - \mathbf{tp}, \tag{4.3.19}$$

$$\mathbf{c_{resid}} = \mathbf{C} - \mathbf{tq}. \tag{4.3.20}$$

6. If more components are required, replace $\mathbf{X}$ and $\mathbf{c}$ by their residuals and return to step 1.

In the case where there is more than one PLS component the weight factor vector can be built into a weight matrix (**W**). The number of components required in step 6 can be computed by calculating the number of components that minimises the cross validation error for that particular model. For all data within this thesis this was the method employed and the number of components (latent variables) used will be stated for each model. The terms PLS-1 and LV1 will be used interchangeably to mean the latent variable on PLS 1.

Once the model has been built it can be used to predict a value for the label matrix **c** for unknown samples. To do this the predicted value of c is calculated using

$$\mathbf{c} = \mathbf{Xb} + \mathbf{f} = \mathbf{Tq} + \mathbf{f}, \tag{4.3.21}$$

where **b** is the regression coefficient given by,

$$\mathbf{b} = \mathbf{W}(\mathbf{PW})^{-\mathbf{1}}\mathbf{q}. \tag{4.3.22}$$

The unknown sample's value of $\hat{c}$ can then be estimated by

$$\hat{c} = \mathbf{xb}. \tag{4.3.23}$$

Once a value for $\hat{c}$ has been predicted then a decision rule must be put in place to determine the threshold between the two classes. The class labels are usually in the form of (+1,-1) or (0,+1) therefore for the first example, the easiest decision rule would be to assign all values < 0 to the cancer class (group B) and all samples with a $\hat{c}$ value above 0 to be assigned to group A (control). For more information on the positioning of the decision rule for datasets in the case where sample sizes are not equal or there are more than one class please see [31].

It should be noted that in many cases the results of PLS-DA can be equivalent to that of other arguably more simple techniques such as PCA-LDA and EDC [31]. Furthermore, the other techniques can indeed be more effective and better suited to building classification models for spectral datasets such as when the number of samples $<<$ number of variables (wavenumbers), where there is more than a binary classification system or where one classification group is larger than the other such as when a disease is only 10% likely within a population. However, the

output of PLS-DA provides similar information to that of PCA wherein latent variables are computed within the model and can be used in an analogous way to PCA loading to determine the regions of the spectrum that are causing separation between subject groups. So PLS-DA is a good method of relating the variables within a plot to the actual classifier. Other methods such as PCA-LDA and EC do not allow this information to be easily accessed despite producing equivalent or sometimes better suited classification models [30]. Therefore, PLS-DA was used half-way linking method between clustering methods such as PCA and classification algorithms because it is good for the discovery of 'spectral biomarkers' using the loading-type plots and the spectral biomarkers (variables) are related to the underlying classification groups. Therefore, PLS-DA has been used in this thesis to link underlying variables to classification performance and to compare the underlying contributing variables after different sample or data processing techniques e.g.(before and after freezing (Chapter 6)) for equal sized datasets with binary classification systems. For building a classification algorithm for disease characterisation with more than two classes with unequal sample sizes Random forest classification was used.

Random forest classification

A random forest is a classifier based on a collection of decision tree classifiers[5] to make an ensemble of decision trees. Therefore trees which are weak learners can be combined to create a strong learning algorithm. There are many different types of ensemble learning methods for tree classifiers such as boosting and bagging methods, generally RF is considered an enhanced version of the bagging method (sometimes known as bootstrap aggregation). Briefly, the bagging method involves averaging the resultant decision over a number of trees so the decision function for the learner becomes better than for a single tree. The collection of trees that make up the bagged ensemble are constructed with a random subset of the variables in a data matrix X. The random subset of variables is

[5]For brevity the details of tree learners will not be discussed in this thesis, for more information the reader is directed to [32].

constructed for each tree using bootstrap samples (random sample from the n variables with replacement) is taken and the trees are grown. Therefore at each tree branch (node) a bagged tree ensemble randomly searches over all of the available variables to create a new node that best splits the data, it then repeats this process until the final node is reached for the decision.

The enhancement of RF over bagging is that RF does not randomly search all variables at each branch of the tree, for a classification RF a random subset of the variables is chosen at each node and only those features are considered for best splitting the data. The number of random variables to choose from at each branch is dependent on the number of variables in the data matrix X, the number is by default $\sqrt{N}$ where N is the number of variables (or wavenumbers) in the dataset. Figure 4.16 shows an example of a single tree constructed as a part of a larger forest.

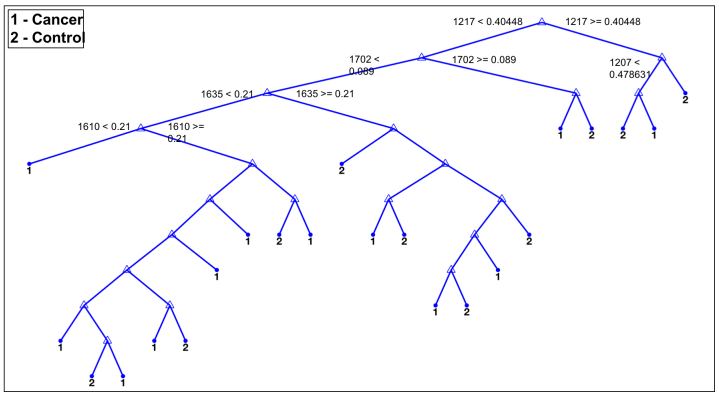


Figure 4.16: Example tree learner, with some of the nodes labelled along a 'branch' for clarity.

As with bagging, in RF each tree that makes up the forest then casts a vote on the classification of a certain sample $\mathbf{x_i}$, so if a forest is made up of 30 tree learners and 10 of the learners vote for a sample to be classed as group 1 (cancer) and the others vote for a sample to be classed as group 2 (control) the classification outcome would be a control. The advantage of this method over bagging methods is that for a multiclass system with a large number of variables the random ensemble of tree learners protects the model from becoming over-fitted as more trees are added [33]. It does this by de-correlating the trees that make up the forest by introducing the random subset of the total number of variables to search

over at each node of the decision tree. The out of bag (OOB) error rate or the error rate of the samples not included in the random subset at each node is monitored during the algorithm and the OOB error of the model will converge as the number of trees grows in the forest [6]. The convergence of the OOB error for the model can be plotted easily for each RF model. Figure 4.17 shows an example of the OOB error rate converging between 300 to 500 trees. Therefore the number of trees used in the models in this work was optimised to the point where the OOB error converges for each model (between 300-500 trees). The exact number of trees for each RF model in this work is stated in the methods for each chapter/result that the model has been used in.

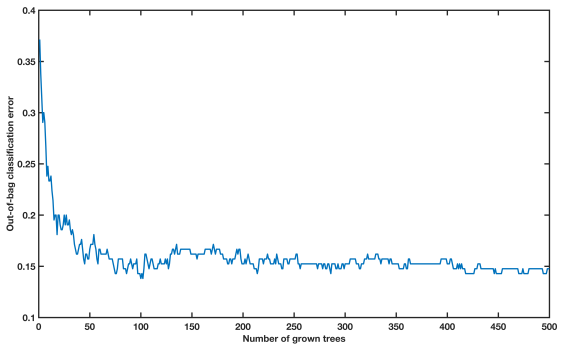


Figure 4.17: RF out of bag (OOB) classification error convergance as number of tree learners grows. The number converges at around 450 learners for the methods used in this thesis.

The RF algorithm can also be set to monitor the variables (wavenumbers) that most often contribute to the correct classification when training a model. The 'popularity' of these wavenumbers in the RF ensemble for classifying samples correctly can be plotted in the form of a gini importance plot (Figure 4.18). This is useful for finding 'spectral biomarkers' and investigating the underlying biological differences between the samples as with the PLS-DA methods. RF methods therefore are well suited to multiclass problems, for datasets with large numbers of variables and it can be used to monitor spectral biomarkers. Therefore RF

[6]This is due to the strong law of large numbers and the tree structure which are beyond the scope of this thesis but can again be found here [32].

has been used as the main form of machine learning model used for classification during this thesis.

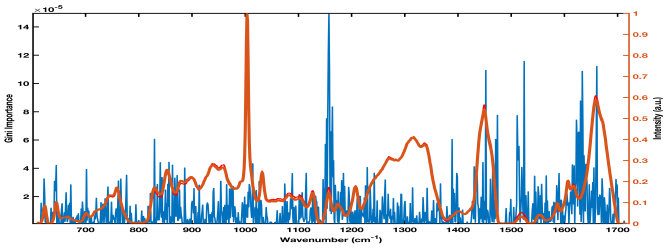


Figure 4.18: Gini importance against average spectrum, the higher the importance peak the more useful the wavenumber in classification.

4.3.3 Cross validation and diagnostic performance

Once a classification model has been constructed it is important to calculate the performance of the model using cross-validation methods. To validate a model the dataset for the model needs to be split to create three overall categories of data for the machine learning process:

- Training set: A subset of the total dataset used to train the model and fit parameters to the model e.g. in Random forest the number of trees.
- Validation set: A set of data that is partitioned out of the training set and used to cross-validate the model and fine-tune parameters e.g. in PLS-DA the number of loadings/latent variables used.
- Testing set: An independent set of data that is not used in the model building that is used to assess the diagnostic performance of the dataset.

Cross-validation is the process building a model, then leaving out a proportion of the training data matrix (validation set) and then testing the performance of the model against the validation set. The results of the cross-validated dataset can then be compared against different model parameters to fine tune the model building process or to fine tune the preprocessing methods used for the data (Figure 4.19). Once the model has been optimised the testing dataset can be used to test the overall performance of the classifier by re-calculating the confusion matrix and other performance indicators.

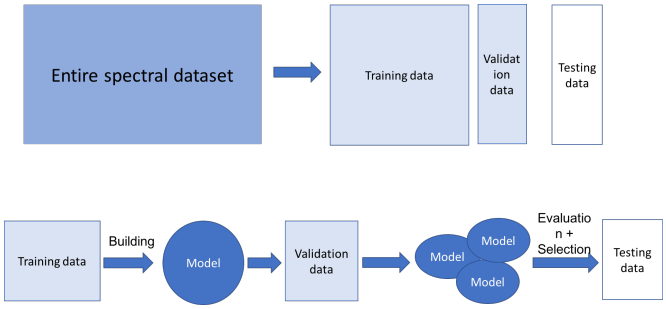


Figure 4.19: Data splitting for spectral model building and optimisation.

Cross validation methods

As with the other methods in analysis of vibrational spectroscopic data there are many different cross validation methods available for use including leave-one-out (LOOCV) [34] and k-fold cross validation [35,36]. K-fold cross validation involves an iterative process of splitting the training data into training sets and validation sets. The training data matrix is partitioned into k folds and the iteration done k times (Figure 4.20). The performance of each model iteration with different validation sets of data from the 'fold' is then combined to produce an overall performance for the model including the error in each iteration.

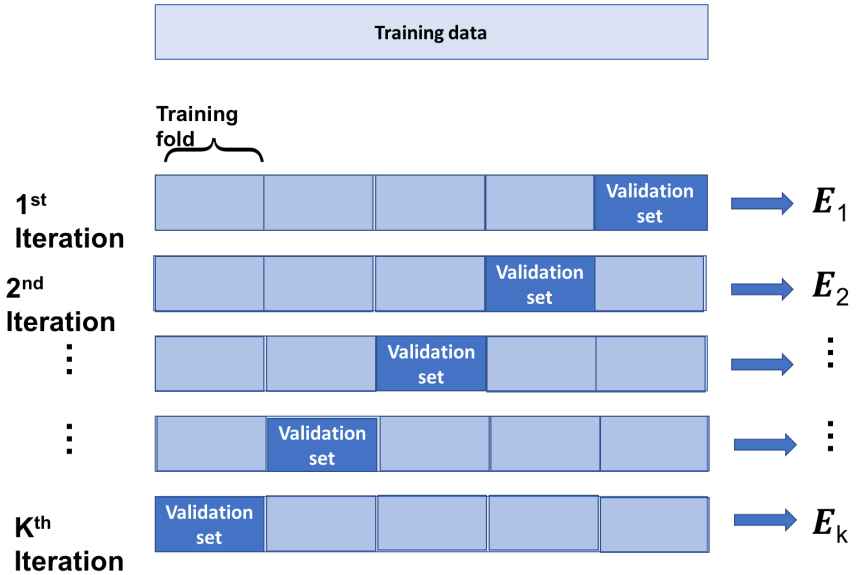


Figure 4.20: K-fold cross validation; partitioning of the training set to select the validation set. The error from each fold is then combined.

The sum of the errors (**E**) is then calculated and the overall classification error

and the cross validation error of the model can be computed. The error in the cross validation will depend on the number of fold in the model and the model parameters. For all of the models in this thesis k-fold cross validation was used with 5 folds.

Diagnostic performance: Sensitivities, Specificities and ROC curves

The result of the cross validation is normally given in terms of a confusion matrix which displays the correctly and incorrectly predicted spectra from the validation (Table 4.3). The confusion matrix can then be used to calculate certain diagnostic performance measures such as sensitivity, specificity and receiver operating characteristic analysis (ROC analysis).

Table 4.3: Confusion matrix example, with diagonal elements being the correctly predicted spectra and the off-diagonal elements being incorrectly identified spectra.

		Diagnosis			
	n = 49	**Cancer**	**Control**		
Predicted	**Cancer**	**15**	**16**	**NPV**	**94.44%**
	Control	**1**	**17**	**PPV**	**48.39%**
		Sensitivity	**Specificity**		
		93.75%	**51.52%**		

The sensitivity of the model is calculated from the confusion matrix and is defined as the true positive rate (number of cancer spectra identified by the model as cancer). Mathematically it is defined by:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{4.3.24}$$

where TP is the number of true positives and FN is the number of false negatives for that particular class. The specificity is defined as the measure of the true negative rate, mathematically defined as:

$$Specificity = \frac{TN}{TN + FP}, \tag{4.3.25}$$

where TN are the true negative values and the FP are the false positives (negative spectra incorrectly identified as positives). The ROC analysis is a measure of the distributions of the TN and TP populations of the results for a given model. The ROC analysis produces a curve that is a plot between the sensitivity as a function of the false positive rate (1-specificity). Therefore, each point along the curve corresponds to a specific sensitivity and specificity pair for each decision threshold for a particular model. Figure 4.21 shows an example of a ROC curve generated during this work, the area under the ROC curve (AUC) denotes the performance of the discrimination ability of the model. A model with perfect discrimination between two populations will have an AUC of 1 and will pass through the upper left corner of the ROC curve.

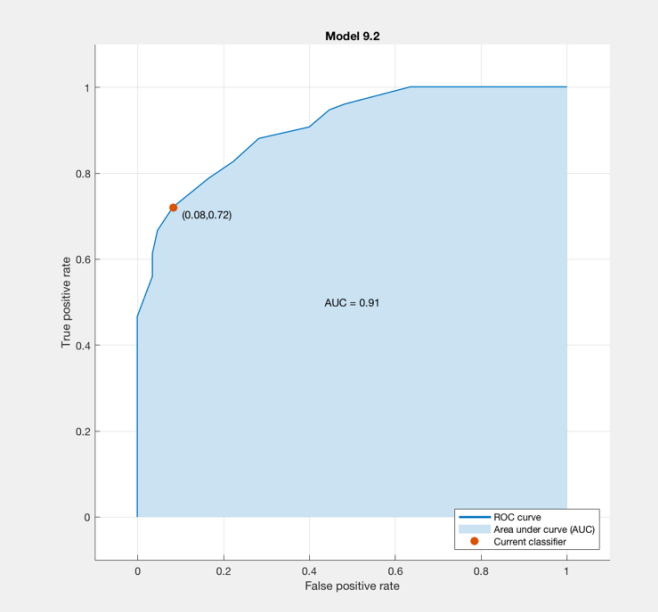


Figure 4.21: ROC curve for an example model, AUC indicates the overall learning performance of the classifier (0.91 classes as very good) and the orange spot denotes the position along the curve of the cross-validated results of the model.

An area of 0.5 or less denotes a diagnostic test that has no better ability to produce a result than flipping a coin or just a chance result. The position along the ROC curve that a diagnostic model lies is given by the cross validated performance of the model (i.e. the position of the sensitivity and specificity of the CV model).

Sample sizes for classification

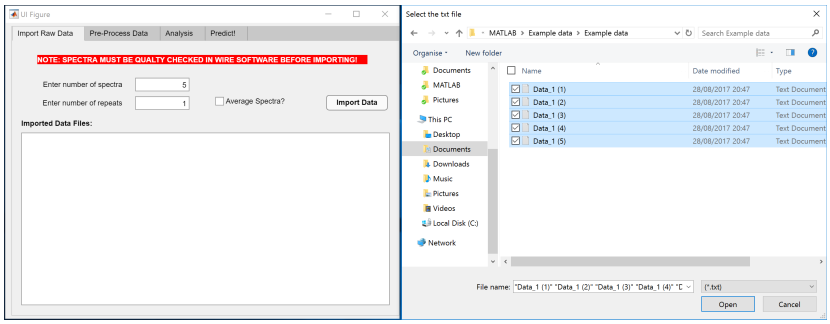
Spectra throughout this thesis were analysed via a spectrum-wise basis. It has been previously shown that an independent sample size of 75-100 cases is needed as an independent blind testing set [36]. Due to the nature of the patient data in this thesis there was not enough patients for this size of independent testing set. Therefore, rather than spectra being sampled on a patient-wise basis each individual spectrum was analysed in both the PLS-DA and RF methods and the individual spectra were used as 'independent' test sets to reduce the classification error and improve the learners due to lack of independent classifiers. To get results on a patient-wise basis the results of the outputs were averaged across the individual spectra for each patient to gain an overall decision. The decision functions for each model will be stated within each models methods.

4.4 Automated spectral analysis Matlab app

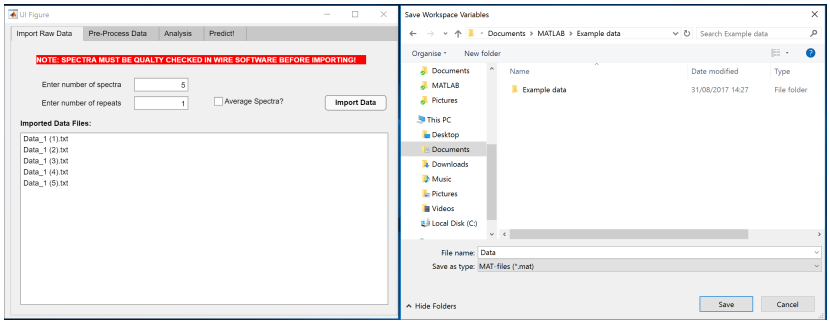
Automated processing and analysis of large spectral datasets is crucial when considering translation into a clinical environment. The overall spectral analysis routine including the quality checking, pre-processing and comparison to an already established model would ideally be completely automated into a 'black box' configuration that simply imports data and produces a diagnostic result ready for clinical interpretation. As a first step in this process a MATLAB spectral processing and analysis app has been developed. The app allows the user to import, pre-process and do some basic analysis on spectral data. It also allows the user to predict the result of optimised processing measures against RF and PLS-DA diagnostic models that are built-in to the app.

Figure 4.22 shows the opening tab within the graphical user interface (GUI) for importing raw Raman data that has been saved in .txt format. The GUI accepts data that has already been quality checked within the Wire software. The user simply inputs the number of spectra or number of patients and the number of repeat spectra per patient, then can select multiple files. If the option to average the data is selected the app will automatically average the spectra per

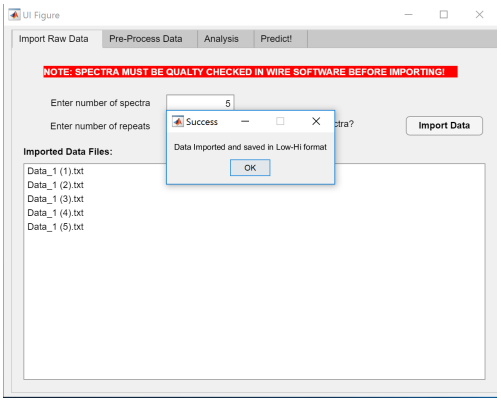
patient. The imported data from the selected files is then automatically saved as a .mat file for leaving the app and returning to data analysis later. The app also has in-built success and error messages to warn the user of the progress of the operation they are trying to perform.



(a) GUI interface with import file selection



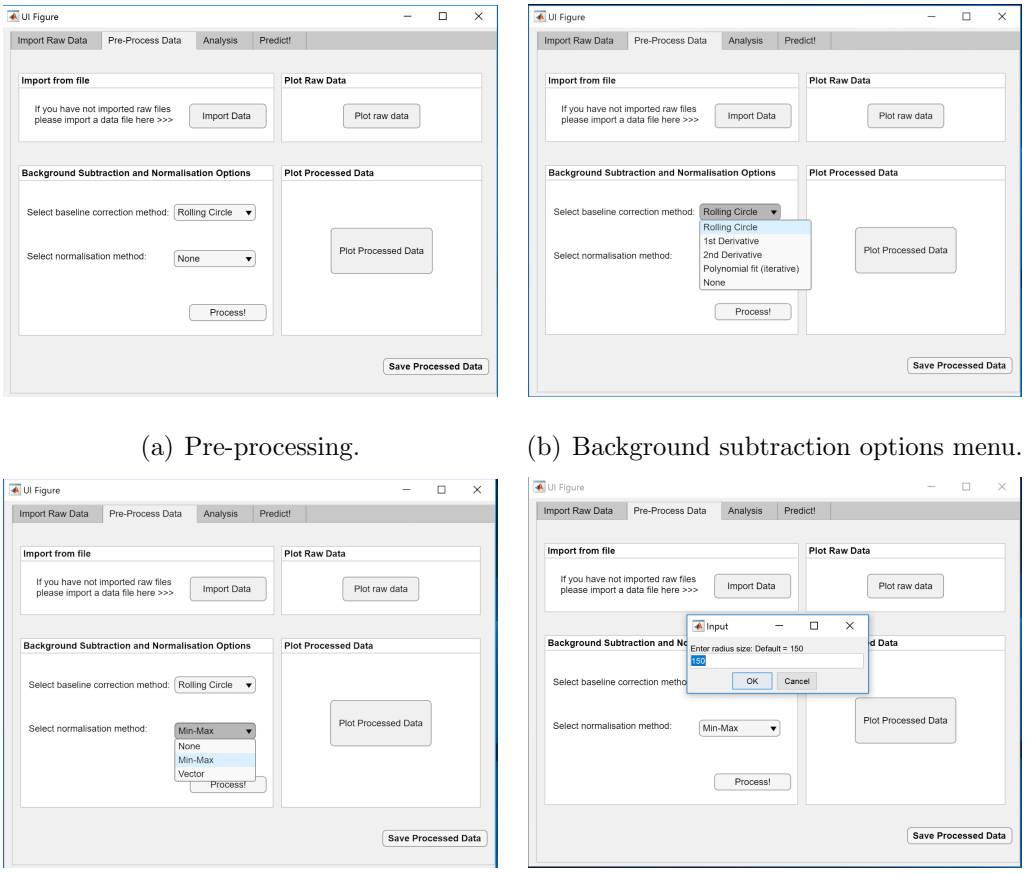
(b) Screenshot of data saving function within the import tab.



(c) GUI successful import message to alert user that process is completed.

Figure 4.22: GUI interface tab number 1, with file select (a), raw data saving (b) and import confirmation (c). Not shown in this image is an error message that displays if the number of spectra entered does not match the number of files selected.

Once spectra have been exported and saved they can then be pre-processed via the second tab in the app (Figure 4.23). The tab contains options to import previously saved data, plot raw data and perform background subtraction and normalisation. The application includes options to select different combinations of pre-processing and normalisation options, depending on the options e.g. RCF. The user is then presented with user input options such as radius size, polynomial order, and window length (derivatives). The pre-processed data are then easily visually examined via the plot processed data button (Figure 4.24). Pre-processed data can then be saved for later use as a .mat file including a wavenumber vector and processed spectra (in rows).



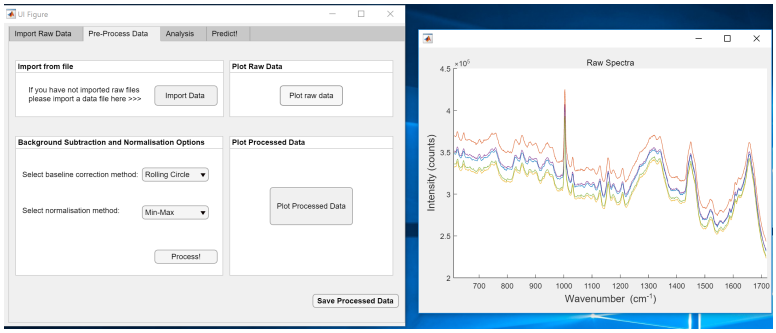
(a) Pre-processing.

(b) Background subtraction options menu.

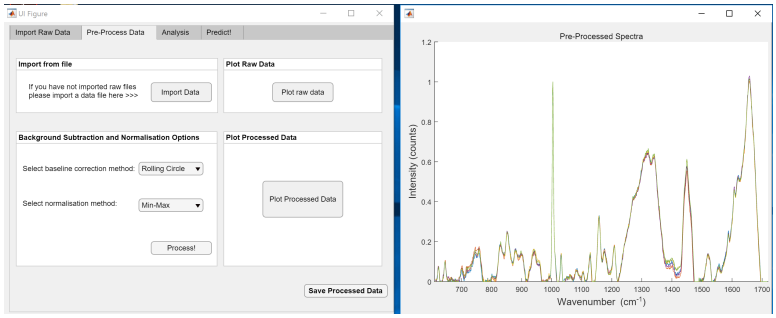
(c) Normalisation dropdown menu.

(d) Extra input option prompts.

Figure 4.23: GUI interface pre-processing tab overview (a), with dropdown options for inbuilt background subtraction (b) and normalisation (c) and the user input option for the selected methods (d). The interface also offers a user to import previously saved data (.mat format) imported and saved from the import data tab.



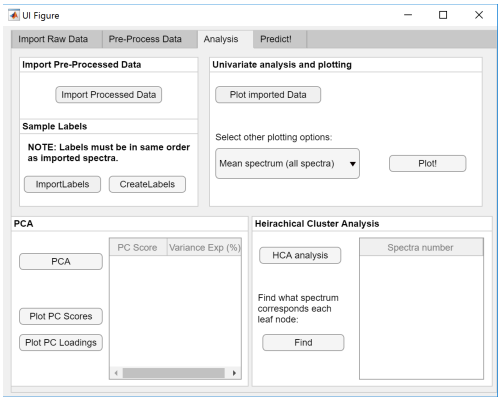
(a) Example of raw data plotting.



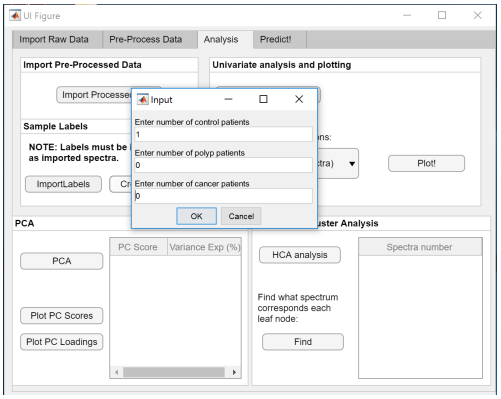
(b) Example of pre-processed data plotting.

Figure 4.24: Example outputs from the raw data plotting (a) and the pre-processed data plotting (b). The plot functions are dynamic so that users can change pre-processed data then get immediate feedback on the results. Figures are also automatically saved as .fig files to allow later editing and change of proportions, colours etc.

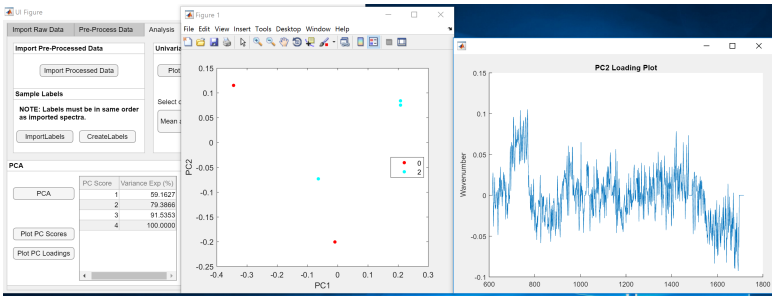
The application can also perform both basic univariate and multivariate data analysis including plotting the mean and SD of the spectra. The application allows for the importation or creation of label vectors to perform HCA and PCA analysis. Figure 4.25 shows an example output of PCA and HCA analysis of a dummy dataset. In the PCA analysis the application gives the options for viewing the explained variances, PC loadings and PC scores plots. For the HCA analysis the app gives the user the ability to perform HCA and then view the samples belonging to each end node in the dendrogram output. As with the pre-processing options the analysis options within the GUI are dynamic to allow the user to change group labels, figures are again automatically saved for later use.



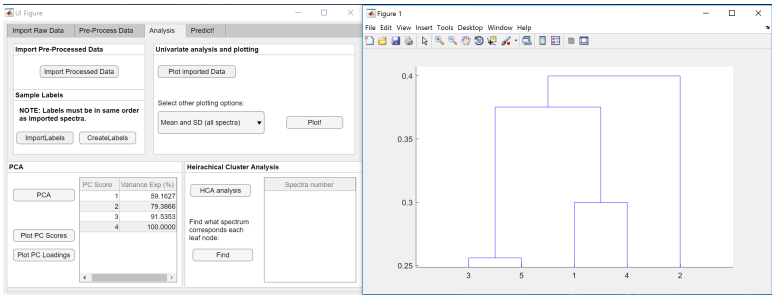
(a) Analysis options.



(b) Import/create group labels for analysis.



(c) PCA analysis output and plotting.

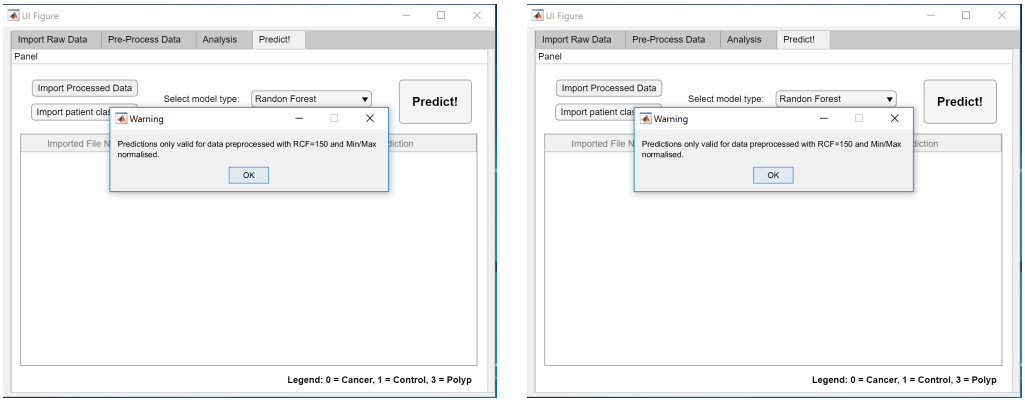


(d) HCA analysis and plotting.

Figure 4.25: Overview of the analysis options (a). Creation or import labels for samples ready for analysis (b). Example outputs of the PCA analysis showing explained variance, scores plots including groupings and loading plots (c). Example output from the HCA analysis (d), the HCA analysis also gives the user the option to see which labelled spectrum corresponds to each leaf node on the dendrogram plot (not shown here).

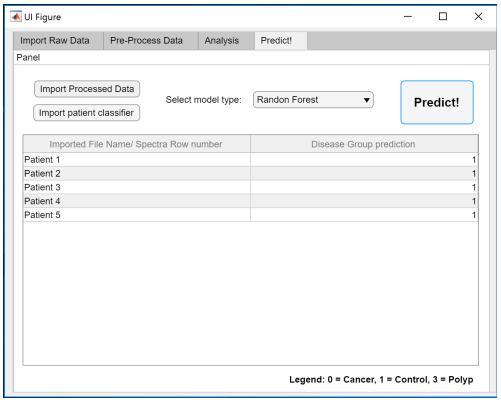
If spectra have been imported and pre-processed using the default parameters RCF background subtraction (radius = 150) and normalised to the phenylalanine peak at approx 1004 cm^{-1} the spectra can be classified using the predict tab of the application. Figure 4.26 shows the details of the prediction tab, it allows either

use of spectra that have been processed within the application or previously saved pre-processed data. The application includes an option to either compare against a model created using PLS-DA or using a random forest model algorithm (details in Chapter 6). The output can then be exported with the patient file name or imported identifier and a disease group classification per spectra as a csv file.



(a) Diagnosis prediction options.

(b) Prediction warning message.



(c) Example of the prediction output including patient identifier/original file name and prediction result with legend.

Figure 4.26: Example outputs from the raw data plotting (a) and the pre-processed data plotting (b). The plot functions are dynamic so that users can change pre-processed data then get immediate feedback on the results. Figures are also automatically saved as .fig files to allow later editing and change of proportions, colours etc.

4.5 Conclusion

The processing and analysis of vibrational spectroscopic data are crucial to gaining well built diagnostic models. There are many options for both pre-processing and analysis methods and before building a diagnostic model different iterations must be tried on the data to maximise the diagnostic capability of a model (Chapter 6). Both unsupervised and supervised classification methods were used during the work within this thesis, in particular PCA, HCA and PLS-DA were utilised during optimisation of spectral pre-processing, analysis and sample handling. To build a diagnostic model RF classification was used as it offers protection from over-fitting by using an ensemble of classification trees and it also allows the identification of spectral biomarkers using gini importance plots.

The development of an automated spectral processing and analysis application for Raman processing of data for diagnostic purposes speeds up the process and allows for rapid optimisation of different pre-processing and analysis techniques. It is also the first stage towards a 'black box' system of spectral processing that will lead towards translation of the technology [37]. The app also allowed for a robust standardised workflow for spectral analysis as follows;

1. Spectra are quality tested for minimum intensity, SNR and cosmic rays visually.

2. Spectra are then wavenumber standardised.

3. Spectra are then imported into the processing application for wavenumber standardisation.

4. Within the application spectra are background subtracted with the rolling circle filter (R=150) and normalised to Phe;

5. If needed, spectra are analysed using unsupervised classification methods.

6. If needed, the group that spectra belong to is predicted using either the PLS-DA or RF models.

7. Processed spectra and prediction results were saved into a MATLAB workspace and for future use.

Having a standard method of processing data allows for easy training of new users and reduces the risks associated with inter-user variability. However, the data put into the spectral analysis routine must have robust methods associated to it to ensure model quality. The next chapter will discuss the development of robust protocols for the data collection for liquid and dry blood product samples to allow for diagnostic model building from spectroscopic data.

Bibliography

[1] G Maes. Handbook of Raman Spectroscopy. From the Research Laboratory to the Process Line, 2003.

[2] C Beleites and V Sergo. Hyperspec: A package to handle hyperspectral data in r. *R Package version 0.98-201-50304*, 2015.

[3] Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, Michael J Walsh, Martin R McAinsh, Nicholas Stone, and Francis L Martin. Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.*, 11(4):664–687, 2016.

[4] Jianhua Zhao, Harvey Lui, David I. Mclean, and Haishan Zeng. Automated autofluorescence background subtraction algorithm for biomedical raman spectroscopy. *Appl. Spectrosc.*, 61(11):1225–1232, 2007.

[5] Chad A. Lieber and Anita Mahadevan-Jansen. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra. *Appl. Spectrosc.*, 57(11):1363–1367, 2003.

[6] Abraham Savitzky and Marcel J.E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.*, 36(8):1627–1639, 1964.

[7] N N Brandt, O O Brovko, a Y Chikishev, and O D Paraschuk. Optimization of the rolling-circle filter for Raman background subtraction. *Appl. Spectrosc.*, 60(3):288–93, mar 2006.

[8] Zhi Min Zhang, Shan Chen, Yi Zeng Liang, Zhao Xia Liu, Qi Ming Zhang, Li Xia Ding, Fei Ye, and Hua Zhou. An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *J. Raman Spectrosc.*, 41(6):659–669, 2010.

[9] Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Tech. Instrum.*, 2(1):8, 2015.

[10] Kristian Hovde Liland, Achim Kohler, and Nils Kristian Afseth. Model-based pre-processing in Raman spectroscopy of biological samples. *J. Raman Spectrosc.*, 47(6):643–650, 2016.

[11] Thomas Bocklitz, Angela Walter, Katharina Hartmann, Petra Rösch, and Jürgen Popp. How to pre-process Raman spectra for reliable and stable models? *Anal. Chim. Acta*, 704(1-2):47–56, 2011.

[12] Georg Schulze, Andrew Jirasek, Marcia M.L. Yu, Arnel Lim, Robin F.B. Turner, and Michael W. Blades. Investigation of selected baseline removal techniques as candidates for automated implementation. *Appl. Spectrosc.*, 59(5):545–574, 2005.

[13] 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. *Appl. Spectrosc.*, 62(3):259–266, 2008.

[14] Paul Bassan, Achim Kohler, Harald Martens, Joe Lee, Hugh J Byrne, Paul Dumas, Ehsan Gazi, Michael Brown, Noel Clarke, and Peter Gardner. Resonant Mie scattering (RMieS) correction of infrared spectra from highly scattering biological samples. *Analyst*, 135(2):268–277, 2010.

[15] J. O'Neal. *Introduction to Signal Transmission*, volume 20. 1972.

[16] Åsmund Rinnan. Pre-processing in vibrational spectroscopy when, why and how. *Anal. Methods*, 6(18):7124–7129, 2014.

[17] Timothy M. James, Magnus Schlösser, Richard J. Lewis, Sebastian Fischer, Beate Bornschein, and Helmut H. Telle. Automated quantitative spectroscopic analysis combining background subtraction, cosmic ray removal, and peak fitting. *Appl. Spectrosc.*, 67(8):949–959, 2013.

[18] Adam Williams, Kevin John Flynn, Zhidao Xia, and Peter Roger Dunstan. Multivariate spectral analysis of pH SERS probes for improved sensing capabilities. *J. Raman Spectrosc.*, 47(7):819–827, 2016.

[19] Jon Shlens. A tutorial on principal component analysis: derivation, discussion and singular value decomposition. *Online Note httpwww snl salk edushlenspubnotespca pdf*, 2:1–16, 2003.

[20] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420, 1970.

[21] Peter Lasch, Wolfgang Haensch, Dieter Naumann, and Max Diem. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochim. Biophys. Acta - Mol. Basis Dis.*, 1688(2):176–186, 2004.

[22] Lior Rokach and Oded Maimon. Clustering methods. *Data Min. Knowl. Discov. Handb.*, pages 321–352, 2005.

[23] Wenjing Liu, Zhaotian Sun, Jinyu Chen, and Chuanbo Jing. Raman Spectroscopy in Colorectal Cancer Diagnostics: Comparison of PCA-LDA and PLS-DA Models. *J. Spectrosc.*, 2016, 2016.

[24] Náira Da Silva Campos, Kamila De Sá Oliveira, Mariana Ramos Almeida, Rodrigo Stephani, and Luiz Fernando Cappa De Oliveira. Classification of Frankfurters by FT-Raman spectroscopy and chemometric methods. *Molecules*, 19(11):18980–18992, 2014.

[25] Martina Sattlecker, Conrad Bessant, Jennifer Smith, and Nick Stone. Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics. *Analyst*, 135(5):895, 2010.

[26] L. Mariey, J. P. Signolle, C. Amiel, and J. Travert. Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics. *Vib. Spectrosc.*, 26(2):151–159, 2001.

[27] Saranjam Khan, Rahat Ullah, Asifullah Khan, Anabia Sohail, Noorul Wahab, Muhammad Bilal, and Mushtaq Ahmed. Random Forest-Based Evalu-

ation of Raman Spectroscopy for Dengue Fever Analysis. *Appl. Spectrosc.*, 71(9):2111–2117, 2017.

[28] Harald Martens and Paul Geladi. *Multivariate Calibration*. Number January. 2004.

[29] Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, 58(2):109–130, 2001.

[30] Richard G. Brereton and Gavin R. Lloyd. Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.*, 28(4):213–225, 2014.

[31] Richard G. Brereton. *Chemometrics for Pattern Recognition*. 2009.

[32] Leo Breiman. Randomforest2001. pages 1–33, 2001.

[33] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. *Elements*, 1:337–387, 2009.

[34] Naomi McReynolds, Fiona G. M. Cooke, Mingzhou Chen, Simon J. Powis, and Kishan Dholakia. Multimodal discrimination of immune cells using a combination of Raman spectroscopy and digital holographic microscopy. *Sci. Rep.*, 7(January):43631, 2017.

[35] Andrew T. Harris, Anxhela Lungari, Christopher J. Needham, Stephen L. Smith, Michael A. Lones, Sheila E. Fisher, Xuebin B. Yang, Nicola Cooper, Jennifer Kirkham, D. Alastair Smith, Dominic P. Martin-Hirsch, and Alec S. High. Potential for Raman spectroscopy to provide cancer screening using a peripheral blood sample. *Head Neck Oncol.*, 1:34, 2009.

[36] Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. Sample size planning for classification models. *Anal. Chim. Acta*, 760:25–33, 2013.

[37] Frank Rinaldi. Global market assessment for Raman spectroscopy and colorectal cancer., 2017.

Chapter 5

Baseline studies

5.1 Introduction

If a Raman spectroscopy based technique for detecting colorectal cancer in biofluids is to be translated into a clinical setting, the measurement parameters for spectral data collection must be optimised. This includes but is not limited to the choice of biofluid, the experimental substrates, sample handling protocols, measurement conditions and the sampling modality. The methods for spectral collection must be robust and wherever possible the processing of the samples should be as closely matched to what would be used in a clinical setting rather than in a controlled laboratory environment. For adoption into a clinical setting the protocols must also allow for data to be taken quickly, accurately and cheaply. This chapter will show the development of the robust experimental protocols for spectral data acquisition from both liquid and dry serum samples. The protocols were developed considering the potential sources of variation that could be introduced into a protocol when translating to a clinical setting.

5.1.1 Variability

Spectral variability can originate from a variety of sources within an experiment. The three main types of variation that have had an impact on the work carried out in this thesis are instrumental, environmental and biological. An example of instrumental variation was discussed previously in Chapter 4.5. Raman spectral collection in the Renishaw system relies on a CCD array to collect photons. The 'binning' of pixels on the CCD can vary depending on the system. One way this source of variability can be minimised is through daily calibration of the CCD area. As discussed previously methods of software analysis can also be used to

counteract some of the introduced variability. However, spectral processing methods are not able to correct for all sources of variation. When considering a system that relies on a machine learning algorithm it is important to minimise the contributions to a spectrum arising from the experimental set up and sample handling. This allows the algorithm to maximise the discrimination between biological differences between samples to give the best chance of detecting cancer from control samples. Another source of variation can come from environmental factors such as the temperature of a sample and if the sample is in a liquid or dry state. It can be difficult to isolate the factors of variation for biological samples because these have an intrinsic variation from person to person and in some cases from the same person at different times of day. These variations can mask the variation between different classes of patient to be compared i.e disease classification. Therefore, when developing a spectral measurement protocol it is important that the experimental conditions minimise experimental and environmental variation and capitalise on variation between the samples.

5.1.2 Aims and objectives

This chapter aims to show the development of measurement protocols for blood samples to minimise experimental variability and maximise biological variation between samples from different patients. The chapter will begin with a comparison of different blood sample types as candidates for use in spectral analysis. Raman and FTIR characterisation of serum including spectral band assignments for Raman and FTIR serum spectra will also be discussed. This chapter will then show the development of high throughput protocols for acquiring data from dry and liquid serum samples. The effect that sample preparation can have on spectral variability will be discussed and this chapter will demonstrate the robust nature of the protocols by investigating the effect of different operators acquiring data for the same samples. Finally this chapter will explore the effects of patient demographics such as age, sex, comorbidities, fasting status and medication on serum Raman spectra.

5.2 Materials and Methods

5.2.1 Sample collection

Serum samples used in this chapter were processed as outlined in Chapter 3.1. Plasma samples were obtained using the same protocol but instead of using serum separator collection tubes (SST, BD Vacutainer, USA), plasma separator tubes (PST, BD Vacutainer, USA) were used that contain an anticoagulant (EDTA). Whole blood samples were also taken using the PST tubes.

5.2.2 FTIR spectroscopy

Spectral data were acquired using the protocol described previously in Chapter 3.4.

Spectral analysis The wavenumber range of the spectral peaks was found using the peak-pick function in the Spectrum software before exportation to MATLAB. Peaks and peak ranges were then attributed to the corresponding vibrational bond and where possible the biological compound that bond is attributed using a literature search.

5.2.3 Raman Spectroscopy

The Renishaw InVia Raman spectrometer was used for all Raman data collection in this chapter. Three different types of scan were used: single point scanning, depth profile scans and mapping measurements. The methods and parameters for each type of measurement are outlined within the methods for this chapter.

Point spectra

Point spectra were taken by aligning the microscope within the system to a single point and taking a spectral measurement from that single position. The optimised spectral parameters for liquid and dry measurements for each laser are outlined in Table 5.1. Parameters were selected in order to gain the highest signal to noise

ratio within a reasonable acquisition time. The positioning and substrate of the measurements is optimised later in this chapter. All spectra from dry samples in this chapter were obtained with the 50x objective. All liquid point spectra were obtained with the 10x objective. The measurement parameters used meant samples were interrogated with 165-170 mW of power with the 785 nm laser and 45-55 mW with the 532 nm laser. This ensured that the sample was not damaged during data acquisition.

Table 5.1: Data acquisition parameters for different sampling modes.

	785 nm Laser		**532 nm Laser**	
	Liquid	Dry	Liquid	Dry
Wavenumber range (cm^{-1})	610-1720	610-1720	610-1720	610-1720
Grating (l/mm)	1200	1200	2400	2400
Exposure time (s)	5	1	0.6	1
Accumulations	30	30	120	10
Laser Power (%)	100	100	100	100
Pinhole in (Y/N)	N	N	N	N

Depth profile

Depth profile measurements were taken using the 10x objective. The microscope was focused onto the base of a substrate then, using the automated microscope stage the distance of the sample from the microscope was increased incrementally after each spectral acquisition. Depth measurements were taken using 1 μm increments from the base up through the sample. Depth profile measurements were performed using the parameters for liquid measurements as in Table 5.1. The optimisation of the substrate for the liquid measurements will be outlined within this chapter.

Mapping measurements

Serum samples (3 μl) were pipetted onto an aluminium foil substrate in triplicate as seen in Figure 5.6. The Wire software includes a feature called surface that

allows the capture of a montage of images across a droplet including adjusting the z axis using the motorised stage for focus. A surface over each droplet was created, then map scan operation was conducted in the software over the montaged image. Point spectra were acquired across each droplet at 50 μm intervals across a grid to cover the droplet area. The measurement parameters were the same for dry measurements apart from mapping measurements which were acquired with the 10x objective .

5.2.4 Spectral Analysis

The spectral analysis for each measurement type varies slightly as the structure of the data acquired is different. Point spectra are saved in terms of a list of wavenumbers and corresponding intensities for 1015 data points across the spectrum. Mapping measurements and depth profiles have spectral information that is saved in a similar way with xyz coordinate positions of the sample added to the file name in order to construct a map. Therefore, some of the spectral data in this chapter was analysed in the Wire software and some was exported to the MATLAB workspace.

Point spectra

All point spectra were processed in the MATLAB environment. In order to investigate spectral reproducibility and variance spectral pre-processing was kept consistent for all spectra within this chapter. Raw spectra were quality checked for minimum intensity and cosmic rays before exportation into the MATLAB Raman data processing app. Data were then wavenumber standardised, a rolling circle filter background subtraction with a radius of 150 for spectra taken with the 785 nm laser line and a radius of 250 for 532 nm spectra. Spectral data were then normalised to the intensity of the phenylalanine peak at (1003 cm^{-1}) and saved in a large matrix (spectra in rows) before spectral analysis.

Vibrational band assignments Before exportation to the MATLAB software, the wavenumber range of the spectral peaks was found using the peak-pick

function in Wire. With reference to the published literature, peaks and peak-ranges were then attributed to a corresponding vibrational bond and, where possible, biological compounds.

Mapping measurements

Mapping measurements were processed in the Wire software environment. Wire has in-built Raman mapping software that allows the superposition of false-colour Raman maps on top of the white light images of a sample. The software contains a function that allows PCA-Raman mapping wherein the regions of most spectral variance are highlighted on the map. PCA-Raman maps were generated across dry droplets. The PC score variances were superimposed onto white light images of dry droplets to investigate variability across the mapping measurements to investigate sample locations where the variances are minimised to mitigate against sample drying effects.

Depth profiles

Data from depth profile measurements taken to investigate the effects of laser focus through the sample. Data were exported to Matlab for further analysis. The intensity and background contributions of each spectra were investigated using the inbuilt plotting functions in MATLAB.

Spectral variation Spectral variations such as the differences between serum and plasma were explored using the standard deviation from the mean spectrum for three different patients. Difference spectra between sample types were also plotted to highlight spectral differences. When investigating experimental variation and patient demographic effects, in the case where there were very subtle changes, PCA score plots were used to investigate the variances and PCA loading plots were used to highlight the regions of spectra causing the variance. The loading plots from the PCA analysis were used to assign spectral features to the causes of spectral variance. To verify the PCA analysis for the experimental variations hierarchical cluster (HC) analysis was conducted using standard MATLAB parameters to assess which spectra were most similar.

5.3 Results and discussion

5.3.1 Characterising peripheral blood samples with vibrational spectroscopy

The choice of biofluid for analysis is important when considering the purpose and end application for a diagnostic tool. Therefore, a preliminary study into which blood based biofluid was most suited for use with vibrational spectroscopy was undertaken. Whole blood was immediately ruled out of this process as it is difficult to store for longer periods of time so repeat measurements would not be possible from the sample. Also, whole blood contains haemoglobin which is a large Raman scatterer, therefore when analysing whole blood the spectrum is dominated by the haem signature (Appendix D.1). Good quality spectra are achievable with the ATR-FTIR however the cellular content of the blood caused some scattering effects and large variability is introduced into the spectra which are not present in spectra of biofluids without cells [1]. As a result serum and plasma were tested for suitability for use with a Raman and FTIR based spectroscopic test.

5.3.2 Plasma vs serum

Biofluids such as plasma and serum previously have shown potential for use in vibrational spectroscopic methods for disease detection [2–4]. Both serum and plasma are used regularly in current testing protocols within hospital laboratories, therefore are both easily accessible. Serum and plasma are both blood derived products which are obtained from whole blood samples. Figure 5.1 gives a summary of the breakdown of the chemical composition of plasma/serum. It shows that the main constituent of plasma is water, with added proteins, inorganic salts and organic substances including lipids, carbohydrates and amino acids.

The main difference between plasma and serum samples is that serum samples do not contain any fibrinogen or metabolites from compounds involved in coagulation [6]. Serum is produced by allowing whole blood samples to clot after they have been taken, plasma samples are produced by not allowing the blood

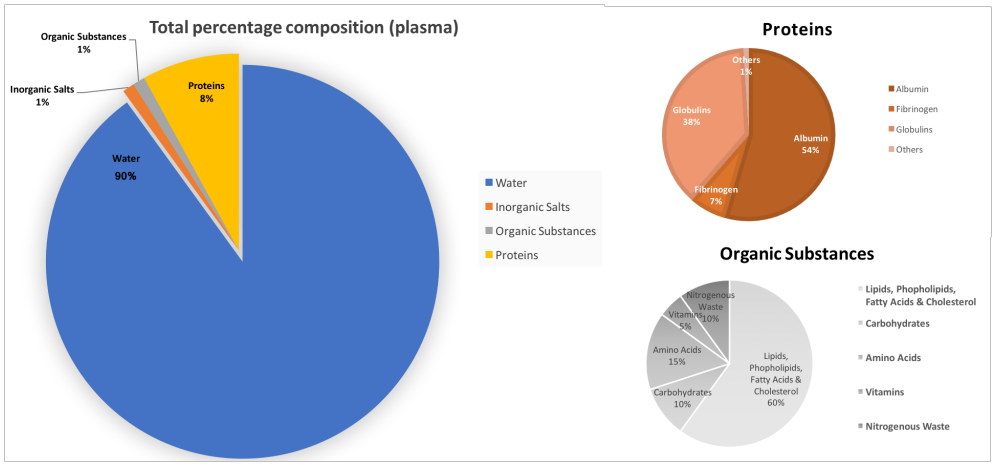


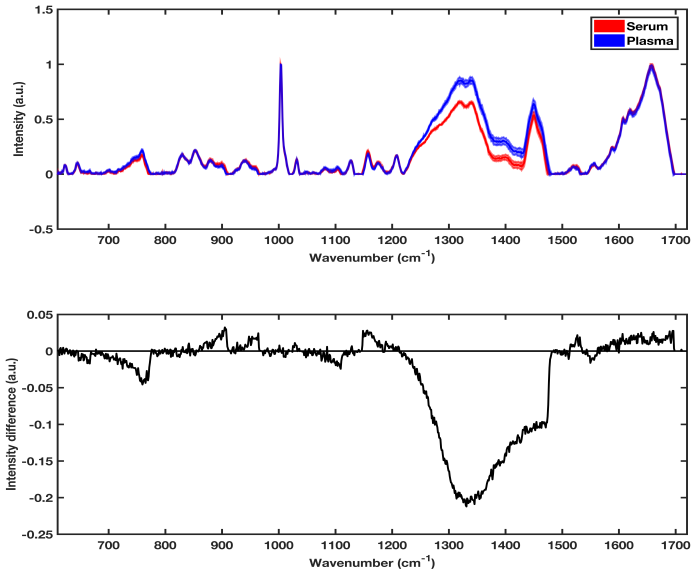
Figure 5.1: Example chart of the typical chemical composition of plasma, values taken from [5]. Serum exhibits similar values minus the clotting factors in protein e.g. fibrinogen.

to clot by adding an anti-coagulant agent to the blood sample in the collection tube. To find the optimum biofluid for use with vibrational spectroscopy it is important to consider the context that a spectroscopy based test would be used. The aim of this work is to develop a test which has flexibility to be used both at point of care in a primary setting but also within a central laboratory. Therefore the processing and storage of samples and the reproducibility of samples before and after storage is important. The rest of this section will consider the Raman spectral response of both serum and plasma and also consider the spectral reproducibility of both fluids. It will also assign biological substances to peaks found in the spectra where possible.

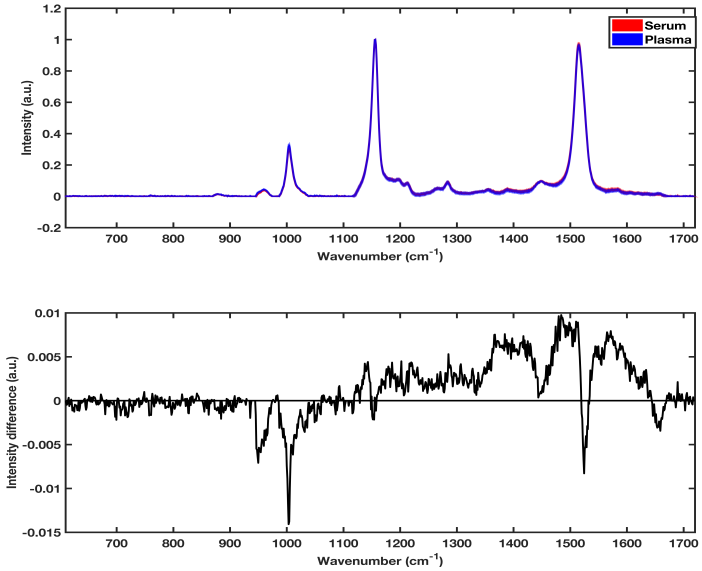
Assessment of fresh plasma vs serum

Fresh liquid serum and plasma samples taken from three different patients on the same day were excited with two different laser lines. Figure 5.2 (a) shows a representative example of the the mean and standard deviation of a fresh liquid serum sample and a plasma sample excited with a 785 nm laser source. The spectra are similar, with no clear sample type showing a higher variation than

the other. Plasma contains some proteins that are not in serum samples such as clotting factors e.g. fibrinogen (Figure 5.1) and the Raman intensity of blood plasma is clearly higher in the region from 1220-1400 cm^{-1}, this spectral region is attributed to protein amide bonds, cholesterol and lipids [7]. It is also higher at the peak at 1658 cm^{-1} attributed to Amide I bonds within protein molecules. This is confirmed when investigating the mean difference spectra between serum and plasma samples. Figure 5.2 (b) shows the mean and standard deviation of fresh serum and plasma samples excited with the 532 nm laser line. The 532 nm laser excitation gives a resonance Raman spectrum of the serum and plasma samples. The variation between the samples is very small and the peaks attributed to carotenoids (1156 and 1620 cm^{-1}) are indistinguishable between the two different sample types. To highlight this an average difference spectrum between the two samples types was plotted, this shows differences of order 0.01. This is a good indicator that fresh and liquid samples of both serum and plasma samples would be suitable for use in a Raman spectroscopy based triage tool for point of care use.



(a) 785 nm



(b) 532 nm

Figure 5.2: Comparison of mean and standard deviation spectra (upper) and difference spectra (lower) for fresh liquid serum and plasma samples excited with (a) the 785 nm laser line, (b) the 532 nm laser line. This is a representative spectrum from one patient of three that were investigated.

This investigation was repeated for dry serum and plasma samples pipetted in duplicate onto aluminium foil. Figure 5.3 shows that the variation between

dried serum and plasma samples is very hard to distinguish. In terms of spectral responsiveness and demonstrably better signal to noise ratio the 785 nm had a much better response than the 532 nm spectra. Furthermore, some samples became damaged when interrogated with the 532 nm laser line. In the interest of translatability and the main aim being for this to be non-destructive to samples only the 785 nm dry methodology was taken forward for dry spectral acquisition.

After drying, differences in the Amide III region disappear. The standard deviation in the plasma samples is higher than that of the serum but only marginally.

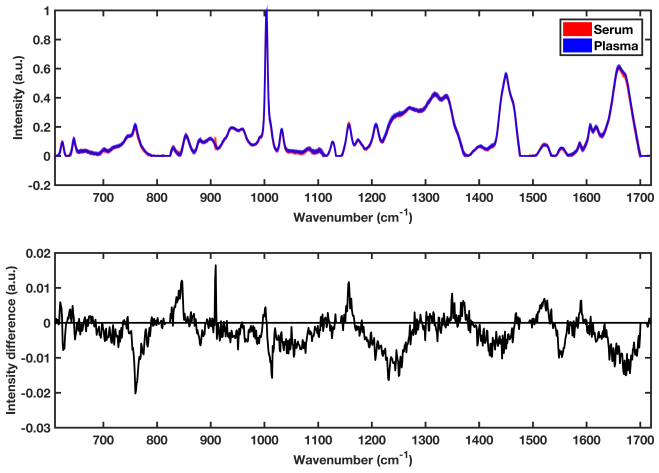


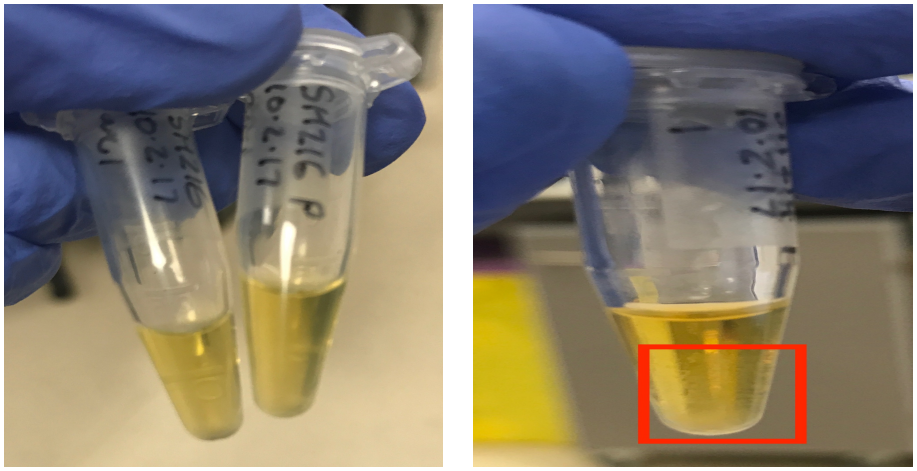
Figure 5.3: Comparison between dried fresh serum and plasma samples excited with the 785 nm laser line. Unfortunately, due to signal saturation issues from fluorescence, dry sample spectra could not be obtained from excitation with the 532 nm laser.

5.3.3 Assessment of plasma and serum post freeze-thaw

Samples used for medical diagnostics must have the ability to be analysed effectively so must be able to be transported for analysis and stored if repeat analysis is needed. Considering that this work is aimed at developing a tool that could be suitable for both fresh (e.g. point of care) and freeze-thawed samples it is important to investigate the effect that freezing has on the spectral response of serum vs plasma and also how this affects the spectral reproducibility.

Aliquots of the samples from the patients used in the investigation for fresh plasma vs serum were frozen at -80°C on the day of collection and stored for 3 months. These samples were then thawed and the measurements in the above

investigation repeated. Figure 5.4 is an example of one each of thawed serum and plasma samples. When compared to the serum sample, the plasma sample appears to be more cloudy. On closer inspection there are cream coloured droplets within the plasma reaction tube. The droplets are caused by the freezing process and are known as the cryo-precipitate [8].



(a) (b)

Figure 5.4: Comparison of thawed 785 nm serum and plasma sample from the same patient (a) and highlighted light cream coloured spots attributed to cryo-precipitate (b).

The cryo-precipitate contains fibrinogen, Von-Willebrand factor, factor VIII, factor XIII and fibronectin [9]. Due to it being a concentrated droplet of clotting factors the cryo-precipitate is often used to treat coagulation-related disorders [10]. In terms of use for a clinical spectroscopy the cryo-precipitate could cause unwanted/ spurious diagnostic results when using a a diagnostic system that relies on testing against a machine learning reference set. This is demonstrated in Figure 5.5 which shows that the raw spectra of the plasma samples that have droplets of cry-precipitate are highly variable and the spectrum has features which are not present in fresh plasma samples. The largest differences were in the amino acid region of the spectrum at 748 cm^{-1} and 758 cm^{-1}, the lipoprotein and triglyceride area between 1080-1096 cm^{-1}. There is also a peak at 1520 cm^{-1} that is attributed to clotting factors and a region between 1535-1600 cm^{-1} that can be attributed to fibrinogen [9].

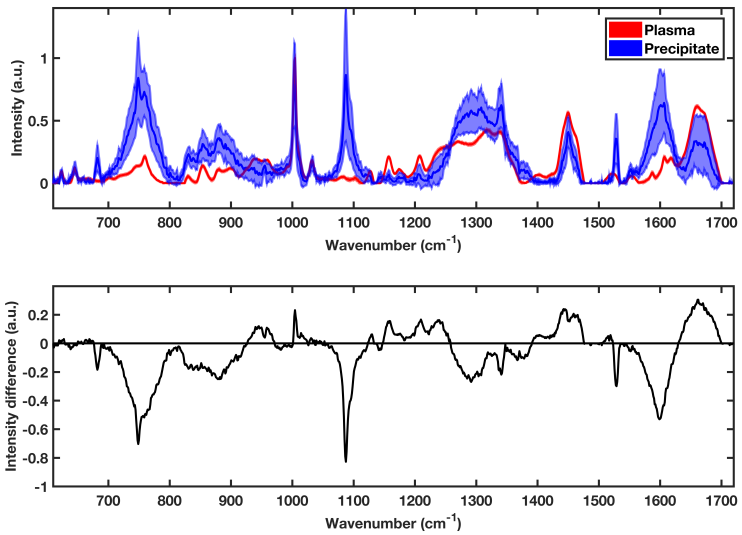


Figure 5.5: Average dry plasma spectrum and average dry precipitate spectrum excited with the 785 nm laser with the spectral standard deviations (upper) and difference spectra between the two components (lower).

Previous work within the research group has also investigated plasma and serum samples for use with Raman spectroscopy and has found that as well as there being issues with the development of cryo-precipitates in plasma samples due to activation of the fibrin cascade, different anti-coagulants have an effect on the spectral responses of plasma samples, particularly with samples excited by the 532 nm laser line [9]. It was found that plasma samples taken with citrate dextrose (ACD) and lithium heparin (LiHep) exhibited a very similar spectral response to serum (as in Figure 5.2). However, samples that used ethylene-diamine acid (EDTA) and sodium citrate (SC) caused a reduction in the resonance Raman response resulting in spectra that are more similar to that of samples excited with the 785 nm laser line. Furthermore, other studies have found that the ratio of anticoagulant to blood can also change spectral responses depending on anticoagulant and cause a detrimental affect when trying to diagnose disease in both fresh and freeze-thawed samples [11]. Anticoagulants appear to be a source of spectral variation that could therefore affect diagnostic capability of a spectroscopy based tool. Serum samples do not require an anticoagulant, have a similar reproducibility to fresh plasma samples and do not suffer the same

reproducibility as plasma samples post freeze-thaw cycle. Therefore, all further investigations during this project were carried out using fresh and freeze-thawed serum samples.

5.3.4 Vibrational band assignments for serum samples

Interpreting spectra correctly is crucial to understanding the underlying biological processes that are causing changes within the spectra. The wavelength of light used to excite the sample can lead to very different spectral responses. Table 5.2 gives a summary of the main spectral band assignments for human serum samples that have been excited with a visible (532 nm) and near infra-red (NIR;785 nm) laser source constructed from the literature [3, 7, 9, 12–14].

Serum that has been excited with NIR light gives an array of peaks that can be attributed to many biological components such as proteins, amino acids, lipids, nucleic acids and glycoproteins. The NIR spectrum of serum is generally characterised by a large central peak at 1004 cm^{-1} attributed to the aromatic ring breathing mode of phenylalanine. This is coupled with strong spectral features at 1447 cm^{-1} attributed to CH_2/CH_3 stretching modes and a strong Amide I peak at 1658 cm^{-1}.

In contrast, the Raman spectrum seen when a serum sample has been excited with visible light has far fewer spectral features. The main spectral features are due to molecules resonant at visible wavelengths e.g.carotenoids. It is characterised by the phenylalanine peak at 1004 cm^{-1} and the $(C\text{-}H)_n$ and $(C{=}C)_n$ stretches attributed to the carotenoid family of molecules [15]. It must be noted that the carotenoid peaks are also visible in the NIR spectra however, one of the peaks at 1516 cm^{-1} is shifted to 1520 cm^{-1} in the spectra from the visible light excited serum. The α-helix structure of Amide III in proteins can be distinguished in the resonance spectrum at 1283 cm^{-1} and 1298 cm^{-1}, this assignment cannot be found in the NIR spectrum of serum samples.

Peak position (cm^{-1})		Molecular assignment	Biological Component
785 nm excitation	532 nm excitation		
622	N/A		Phe
644-645	N/A		Phe, Tyr
700	N/A	$_{\nu}$(C-S)	Lipids
743	N/A		PPL; chol
758	N/A	(C-S), (CCN)	Amino acids
828-829	N/A	CH rock in CH_2	Phenol ring stretch (exposed Tyr)
853	N/A	(C-C) stretch	Tyr
898	N/A	$\delta(CH_3)$	Lipids; glycoproteins; fatty acid chains
938	N/A	(N-C)&(C-C) stretches	Proteins (α helical stretch)
958	960	CH_3 Def	
1003-1004	1004	Aromatic breathing	Phe
1032	N/A	(C-H) bending, (C-O) stretch	Phe; glycogen
1082	N/A	(C-N), PO_2	DNA
1127	N/A	(C-N) & (C-C) stretch	PPL; proteins; LDLs; HDLs
1157	1156	$(C\text{-}H)_n$	Carotenoids
1175	N/A		Trp, Phe (AAs)
1208-1209	1211	β-sheet	Amide III
1266-1271	N/A	(C-C),(C-N) stretching	PPL; Amide III
N/A	1283,1298	α-helix	Amide III
1319	N/A	(C-H) Def	Amide III; Adenine
1342	N/A		Amide III
1447	1448	CH_2, CH_3 stretching	Lipids
1520	1516	$(C=C)_n$	Carotenoids
1556	1535-1595	(N-H), (N-C) & NH_2	Tryp; DNA
1608	N/A	(C=C) stretch	Phe, Tryp
1658	1659	(C=O), (C-N), NH_2	Amide I

Table 5.2: Raman spectral band assignments (from centre of the feature). Where Tyr - tyrosine, Phe - phenylalanine, chol - cholesterol, PPL - phospholipids, AA - amino acids, LDLs and HDLs are high and low density lipoproteins respectively and Def is deformation. Constructed from [3, 7, 9, 12–14].

FTIR serum spectral band assignments

As previously shown in Chapter 2.3 (Figure 2.4), a typical FTIR spectrum of serum also shows peaks from a range of different types of molecules including; lipids, carbohydrates, fatty acids, protein conformational bonds and amino acids. The serum FTIR spectrum is characterised by two large peaks at 1550 cm^{-1} and 1660 cm^{-1} attributed to Amide II and Amide I protein bonds.

Band position (cm^{-1})	**Molecular assignment**	**Biological component**
1170-1120	ν(C-O) and ν(C-O-C)	Carbohydrates
1240	ν_{as}(P=O)	Nucleic acids
1400	$\nu(COO^-)$	Amino acids
1550	δ(N-H)	Protein (Amide II)
1660	ν(C=O)	Protein (Amide I)
1730-1760	ν(C=O)	Fatty acids
2840-2860	$\nu_s(CH_2)$	Lipids
2865-2880	$\nu_s(CH_3)$	Lipids
2920-2930	$\nu_{as}(CH_2)$	Lipids
2950-2960	$\nu_s(CH_3)$	Lipids
3050-3090	ν(C=H)	Lipids
3300	ν(N-H)	Protein (Amide A)

Table 5.3: FTIR spectral band assignments for serum, adapted from [3]. Where ν is stretching, δ is bending, s is symmetric and as is asymmetric stretch.

5.3.5 Optimisation of Dry measurement platform for Raman spectroscopy

Experimental Substrates

One of the main considerations of an experimental substrate is to reduce the autofluorescence of the sample and to minimise the background contribution from the substrate. Previously, serum samples have been investigated using a variety of substrates including glass, CaF slides, aluminium foil, silicon slides and plastic slides [9]. Table 5.4 outlines the main advantages and disadvantages of some of the most widely used substrates for serum analysis of dried samples. Spectra for each

of the considered substrates were taken and compared, the spectral comparison can be found in Appendix D.2.

	Cost (£)	**Advantages**	**Disadvantages**
Aluminium foil	0.01	Cheap; low background contribution	Subject to degradation over time
Glass slides	0.14	Cheap; already used regularly in pathology	Large florescence fingerprint region with NIR
CaF slides	70	Lowest background contribution	Highly expensive; not feasible for translation.
Silicon Chip	2.00	Low background in fingerprint region; low cost.	Strong peak at 520 cm^{-1}.
Plastic microscope slide	2.70	Cheap; Re-usable.	Large spectral contribution across the whole spectrum.

Table 5.4: A comparison of the different experimental substrates available for spectroscopic analysis of biological samples.

CaF slides have the lowest Raman background contribution but these slides are expensive at £70 per slide and the slides can degrade over time which does not lend them to clinical translation. Glass and plastic microscope slides are less expensive but they both have large background contributions in NIR Raman spectra, therefore heavy processing is required. Silicon chips are also low cost, but silicon has a large peak at 520 cm^{-1} in both visible and NIR excitation that is very strong and has the potential to mask other spectral features. Aluminium

foil is the most economical substrate with the lowest background contribution. Previously, studies have found aluminium foil to be a suitable substrate for Raman studies using biological samples [16]. Therefore, all of the dry data were taken using aluminium foil in a dimpled well as the substrate (Figure 5.6). Impressing the aluminium foil into a dimpled well pattern allowed multiple samples to be dried in a uniform spacing for higher throughput and even drying.

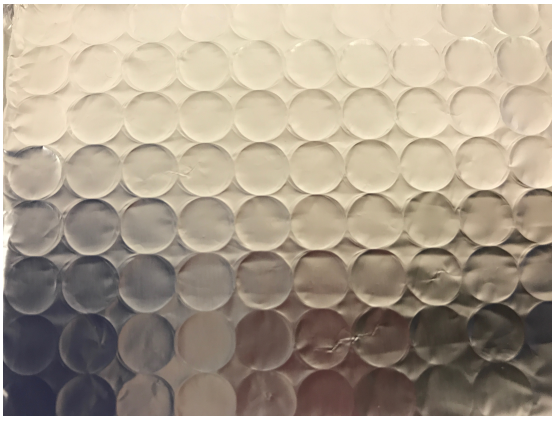
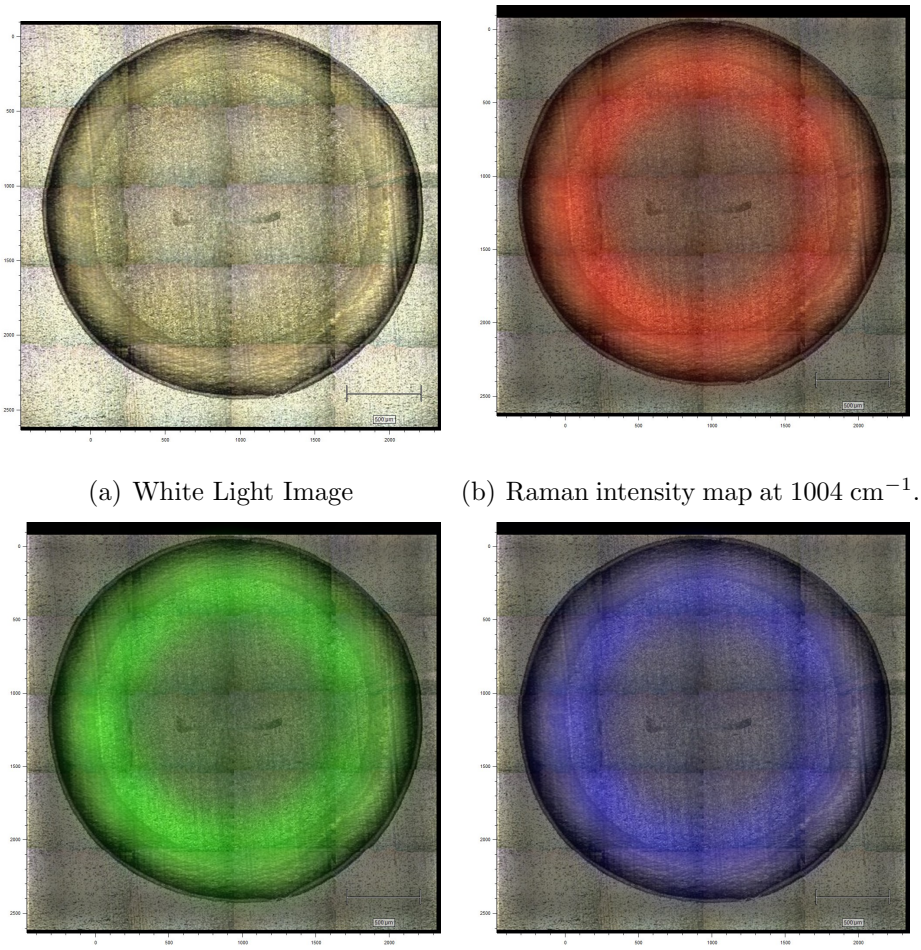


Figure 5.6: Example of the multi-well aluminium plate used for all of the dried spectra in this work.

5.3.6 The Vroman effect

Serum samples that have been dried onto flat surfaces have a characteristic 'coffee ring' appearance with a dark concentrated edge around the outside of the droplet with an inner ring and a flat centre. The process of forming this 'coffee ring' is explained by the Vroman effect wherein molecules of different molecular weights and different hydrophobicities 'fall out' of the serum solution at different rates during the drying process and become absorbed onto the aluminium surface [17]. The rate at which proteins are adsorbed onto a surface depends on the affinity of the proteins in the serum for that surface. Generally, when liquid serum is dropped onto a surface and left to dry, water molecules are the first to reach the surface. The structure of the surface then becomes important to the method in which proteins are deposited. For example, hydrophobic surfaces tend to deposit proteins that are denatured. When considering a spectroscopic substrate or different substrates it is therefore important to characterise the drying pattern for

that particular substrate.

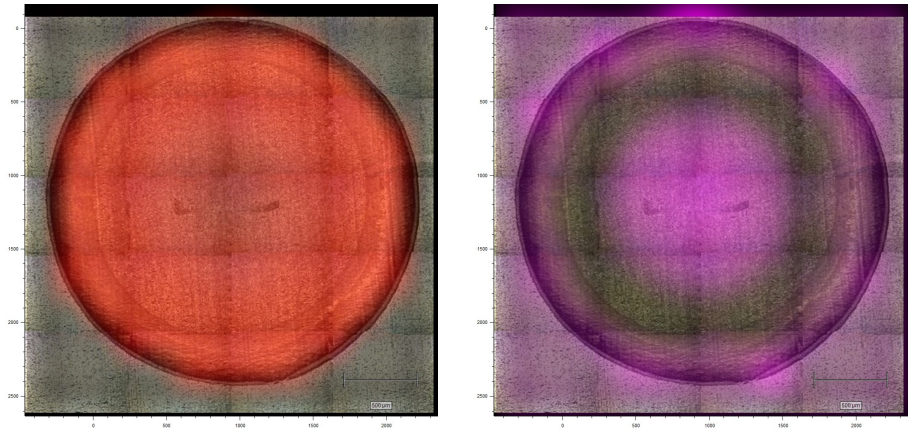


(a) White Light Image (b) Raman intensity map at 1004 cm^{-1}.

(c) Raman intensity map at 1450 cm^{-1}. (d) Raman intensity map at 1658 cm^{-1}.

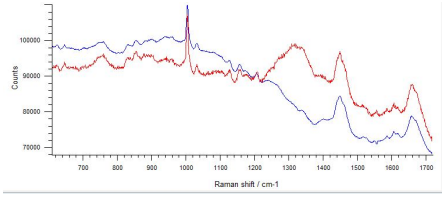
Figure 5.7: Representative Raman Intensity Map over a dry serum droplet. The regions of the highest spectral intensity correspond to the brightest coloured regions in the map.

Figure 5.7 shows a Raman intensity map over a representative example of a dried serum droplet dried onto a flat well as in Figure 5.6. The droplet was mapped over the phenylalanine peak at 1004 cm^{-1}, the CH_2/CH_3 stretching peak at 1450 cm^{-1} and the Amide I peak at 1657 cm^{-1}. Figure 5.7b-d shows the Raman maps over these peaks superimposed on top of the white light image of the droplet (Figure 5.7 a). All of the characteristic peaks were most intense in the inner ring region of the droplet. This suggests that the optimal position for the dry spectral measurements is across this region. When considering application to a clinical setting, the variability of the Raman intensity is also important.

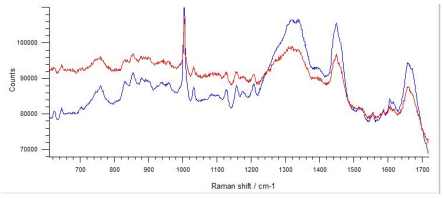


(a) PC 1

(b) PC2



(c) Raman loading for PC1 (blue) and a spectrum from the droplet(red).



(d) Raman loading for PC2 (blue) and a spectrum from the droplet(red).

Figure 5.8: PCA map image over the dried droplet showing PC1 (a), PC2 (b) and the loading vectors from the wire software (c,d).

The variability of intensity over the serum droplets was investigated by generating a PCA map from the dried drop spectral dataset. The PCA algorithm was selected to transform the data according to largest spectral variance. A PCA map was then generated with the most variable areas having brighter mapped colours. Figure 5.8 provides an example of the map for PC1 and PC2 [1] over the dried droplet. Figure 5.8(a) shows that there is an even variance across the sample in PC1 which is expected given that a non-mean centred algorithm for PCA was used. Figure 5.8(b) gives more insight into the optimum position of the droplet to take a spectrum from. The loading corresponds to the overall PCA spectrum from the droplet, the regions of highest intensity on the PCA map correspond with the areas of highest variance across the droplet. Therefore the darkest region

[1]The algorithm used was not mean-centred within the Wire software, therefore the PCA loading plot 1 is approximate to the mean of the dataset.

in the centre of the ring indicated that this would be the optimum position in which to take spectral measurements. The darkest and therefore least variable region of the droplet coincides with the regions of largest spectral intensity in Figure 5.7. This means the optimal region of the droplet in which to take point spectra across the drop is the region highlighted in Figure 5.9. All dry spectral datasets within this work were then taken as point spectra along this region of the dried droplets.

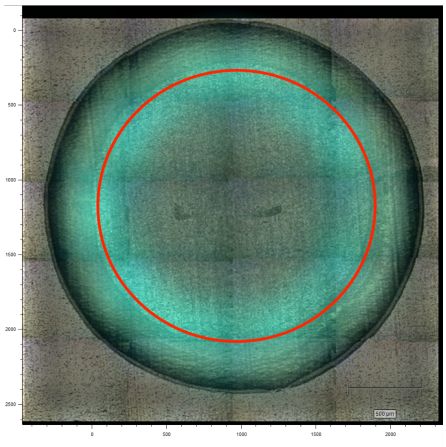
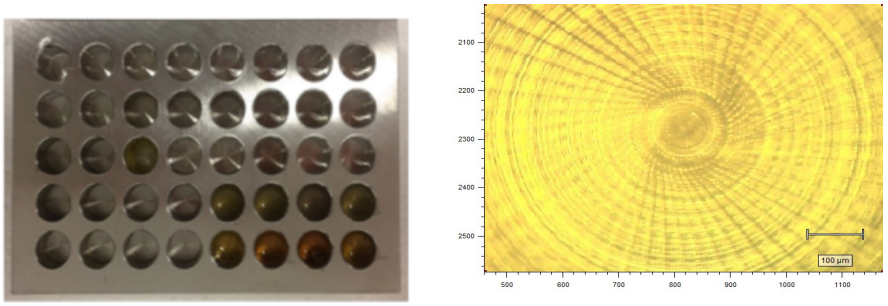


Figure 5.9: Optimal sampling position for dry serum droplets (red).

5.3.7 Development of a measurement platform for liquid data acquisition

Difficulties regarding the drying process are avoided when samples are investigated whilst liquid. A stainless steel 40-well plate was developed for liquid data acquisition during this work. A multi-well plate design gives the ability to analyse multiple samples in one sitting. However, as in Table 2.4, plastic well-plates have a large spectral contribution. An aluminium well plate was developed as a liquid spectral substrate to combat this problem. The aluminium plate was designed with the idea that the well plate could be re-used over time making it more economical. Unfortunately, it was subject to tarnishing and surface oxidisation after cleaning. Stainless steel was therefore considered as a re-usable substrate that has minimal background contribution effects and isn't subject to degradation after cleaning for re-use. This also suggests that it would be a more

economical option for translation.Stainless steel offers a re-usable substrate that has minimal background contribution effects and isn't subject to degradation after cleaning for re-use. This also suggests that it would be a more economical option for translation. Figure 5.10 shows the stainless well plate used in this work. Each well is machined to hold 200 μl of serum and the base machined to have a notch at the centre. This aids with focusing the microscope to the base of the well allowing for measurements to start from the same point in each well. Figure 5.10 (b) shows the notch design at the base of the well plate when filled with liquid serum.



(a) Stainless steel well substrate

(b) The bottom of a filled well through the 10x objective.

Figure 5.10: (a) An example of a stainless steel well substrate with a plurality of wells, each well is designed to hold 200 μl of serum.(b) View of the 'notch' at the base of the stainless steel well plate viewed through the 10x objective.

The working distance required to focus into the well meant that for liquid samples the 10x objective was required. The microscope was focused to the base of each well for spectral measurements, the position of the spectral acquisitions through the well was optimised using a depth profile measurement. The Renishaw encoded stage in the Renishaw system allows precise movement in the z direction to ensure that the sample is in the same position each time. Figure 5.11 is a representative example of a depth profile taken through a sample.

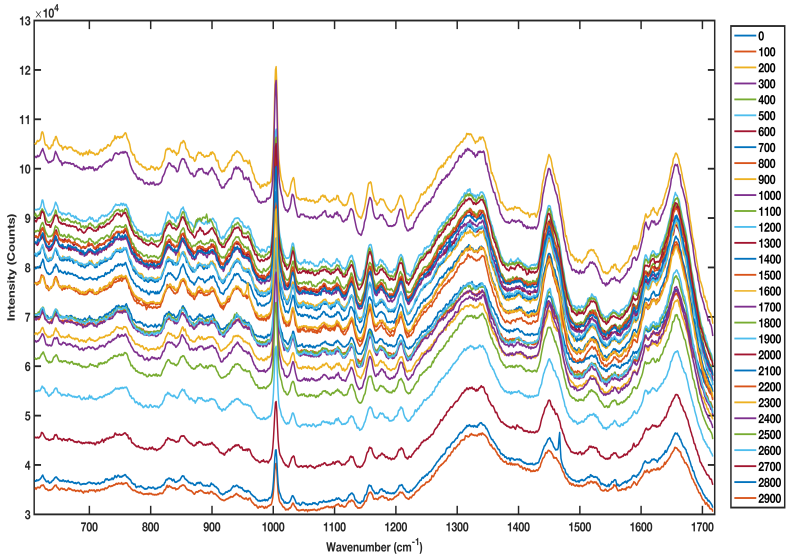


Figure 5.11: Representative depth profile through the sample, step sizes given in μm. Note that raw data are shown therefore cosmic rays have not been removed.

The depth was taken from the base of the well to minimise the effect of loss of focus from the top of the meniscus as some of the sample evaporates. The optimum position was found to be 1200 μm above the base of the well. This position repeatedly gave high intensity values with good SNR, the spectra at this depth were within the grouping of spectra that were most similar.

5.3.8 Cooling the well plate

Generally, when using a multi-well plate system for data collection you would expect to load the well plate in a single step with all samples then have batch data collection. With an open well plate this could be problematic as liquid serum at room temperature will evaporate. To minimise this affect the well plate during this work was temperature stabilised to (18-20°C) using a Peltier cooling system. The cooling of the plate allows the samples to be in the Raman system for longer so will allow batch processing without the sample evaporating before measurement. Figure 5.12 is a schematic of the Peltier cooling system. It consists of a simple USB based Peltier plate that is attached to a customised heat sink and base plate for the Renishaw system. The dimensions of the Peltier cooling system are such that the plate covers the whole base of the well plate. Figure

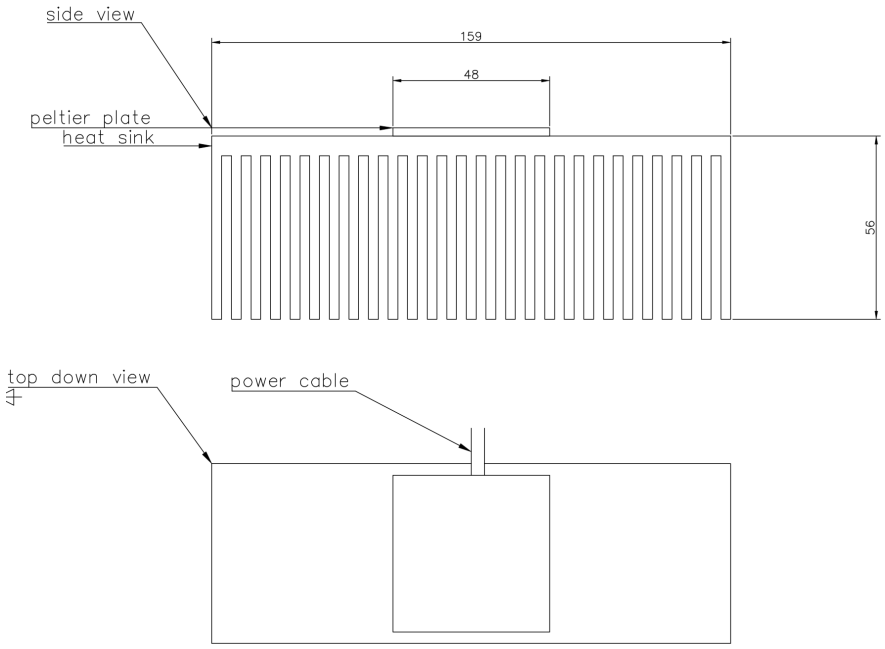


Figure 5.12: USB powered Peltier cooling system for the stainless steel well plate. The base plate schematic can be found in Appendix D.3.

5.13 shows the mean and standard deviation for 5 repeat spectra from the same patient in the same well with the 785 nm laser. During the first measurement set the well plate was used alone, for the second set of measurements the well plate was cooled using the Peltier system. Overall the spectral response from the room temperature data had a larger standard deviation from the mean spectrum, the background fluorescence shape of the spectra taken at room temperature was generally higher and had a steeper sloping baseline from the lower to the higher wavenumbers on the spectrum. The data from the cooled well plate showed that the cooling process helps minimise the fluorescence background contribution from the sample and also allows for more reproducible data collection. The overall spectral intensity was lower for the samples that had been cooled. This can be attributed to there being a lower fluorescence response in the cooled samples and there being a slightly larger volume of serum within the cooled wells due to the slowed rate of evaporation. The focus of the light will therefore be 'deeper' into the sample. This could cause a drop in the number of scattered photons reaching the collection optics. The overall decrease in spectral response is also not so large

that spectra cannot be obtained. Therefore the cooling system was adopted for all liquid spectral collection.

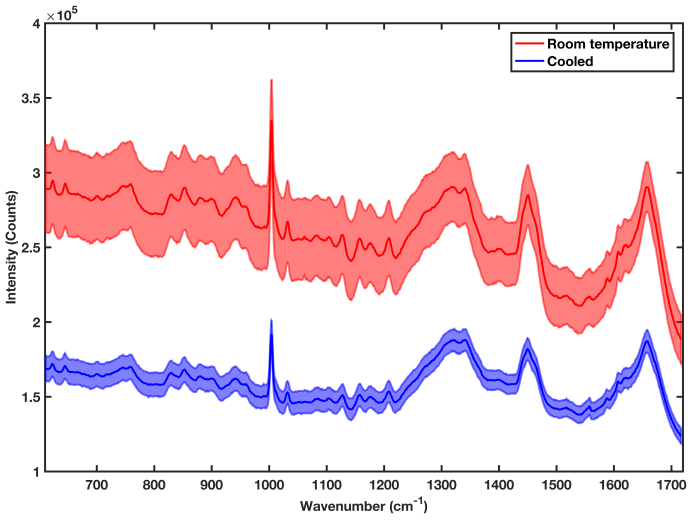
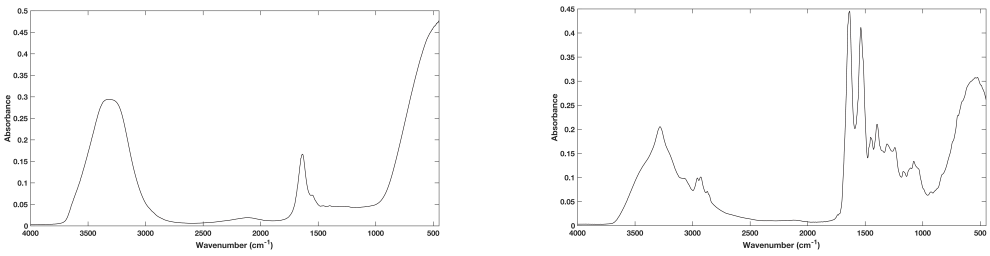


Figure 5.13: Representative example of raw spectral data from the same patient under cooled and room temperature conditions. Spectral data collection was repeated for 5 different wells using control patient samples.

5.3.9 Choosing spectroscopic method

ATR-FTIR was also considered during this work as a potential method of vibrational spectroscopy for the detection of CRC in serum samples. Due to the strong absorbance of O-H bond in water molecules in the infra-red, liquid serum spectra were unable to be used as the spectra were dominated by the water signature within the sample as demonstrated in Figure 5.14 (a).



(a) (b)

Figure 5.14: Example serum FTIR spectra for a liquid sample (a) and a dried sample (b).

To solve this issue samples can be left to dry. Figure 5.14 (b) shows a representative dry serum FTIR spectrum. The dry serum spectra has identifiable peaks that can be attributed to proteins, lipids, carbohydrates etc that are not seen in the liquid spectrum. The dry spectrum was of good SNR and required minimal sample preparation i.e. pipetting onto the diamond ATR substrate. However, one disadvantage to this method is that for each measurement the sample has to be pipetted onto the crystal and when drying samples each sample has to be dried for approx. 20 mins before spectral acquisition can take place. This is very detrimental for any technique that is looking towards translation as the throughput potential is low. Due to the slow spectral acquisition time for the FTIR methods the rest of the work in this thesis was focused upon Raman spectral applications.

5.3.10 Pre-analytical considerations

Serum and plasma samples are commonly used for may different applications in clinical investigations. In some patients with CRC, CEA protein levels are measured in serum samples to monitor progression of the disease [18]. The target analyte within the blood sample usually dictates the pre-analytical treatment of blood samples (i.e. serum or plasma, fresh samples or stored, which collection tubes needed, etc). An example of this is heparin which can bind to ionized calcium in plasma and serum samples causing interference with any tests that measure calcium levels so heparin as an anticoagulant is avoided if calcium levels need to be measured [19]. To ensure that the relevant biological information for the Raman spectroscopy based test is maximised, pre-analytical treatment and other factors that might affect the spectral results need to be considered. The sources of biological variation within the sample set in this work can be split into two groups - sample handling and patient demographics - and will be discussed further in this section.

5.3.11 Sample handling

It was shown above that serum holds the advantage over plasma for vibrational spectroscopy thus eliminating many factors that can contribute to a different

spectrum from different anticoagulants. Local hospitals in which serum samples were collected used only one type of tube for serum collection (Vacutainer SST, BD USA). All samples were collected and processed according to manufacturer gold standards, this eliminated the potential for variations in serum from the collection tubes. Furthermore, any samples that haemolysed during processing were re-collected wherever possible, where repeat collection was not possible the sample was discarded.

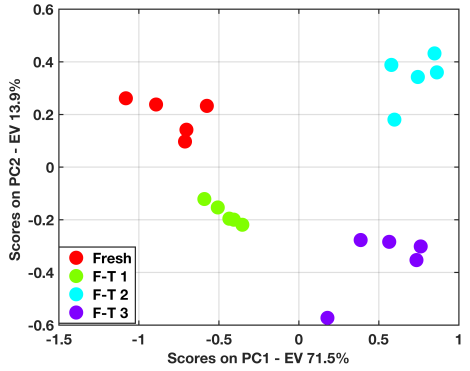
In routine biochemical testing, serum samples can often be used for more than one test analyte, so samples are often stored and frozen at -80°C. Therefore an investigation into the stability of serum samples for Raman spectroscopy was carried out. Furthermore, in a clinical setting there is often more than one member of staff to conduct the same tests so the inter-user robustness of the Raman protocols developed in this chapter was studied.

5.3.12 Freeze-Thaw stability of serum

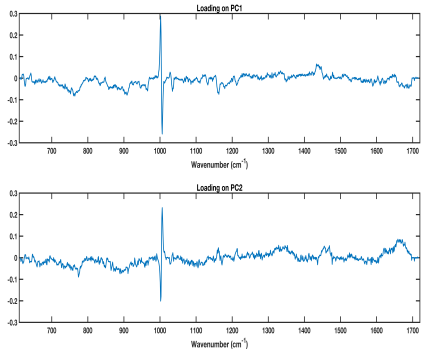
To investigate if serum samples that have been frozen could be used in a Raman based test, five repeat spectra were taken from a liquid serum sample when fresh and then from the same thawed sample that had been frozen at -80°C in three 250 μl aliquots. For three consecutive days after the initial measurement of fresh serum, aliquots were thawed and a measurement was taken from one of the aliquots, the other aliquots were re-frozen for use on other days. Spectra were collected using both the liquid and dry methods developed above. Figure 5.15 shows freeze-thaw analysis for the liquid serum data. PCA analysis confirmed that there was no distinct difference between the fresh sample and samples that had been through freeze-thaw cycles. However, the fresh sample and a sample having gone through a single freeze-thaw cycle were separated across PC1. The loadings for PC1 and PC2 show that the main cause of spectral variation across PC1 can be attributed to a shift of the phenylalanine peak at 1004 cm^{-1}. There were also spectral variations in the peaks at 1032 cm^{-1}, 1082 cm^{-1} and 1127 cm^{-1} attributed to glycogen, DNA and phospholipids, and proteins.

It is expected that during freeze-thaw, water molecules form ice crystals within

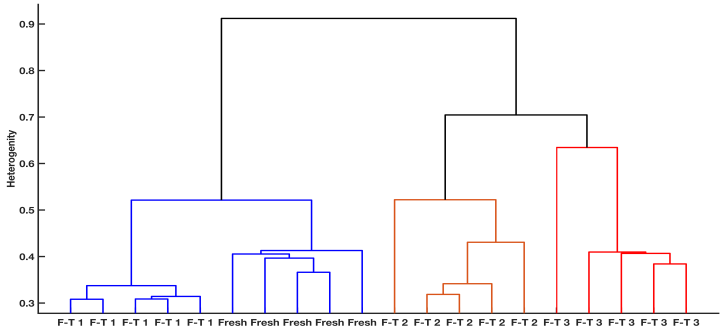
the proteins and when those crystals melt they affect the overall structure of the protein samples. These results are in agreement with previous studies investigating protein and metabolite levels in serum and plasma samples that have been stored at -80°C and then freeze-thawed [20, 21]. The PCA results for the liquid analysis are confirmed by hierarchical cluster analysis of the spectra. The fresh sample and a sample that had been through one freeze-thaw cycle grouped more similarly than the samples that had been through more freeze-thaw cycles.



(a) PC1 vs PC2



(b) Loadings

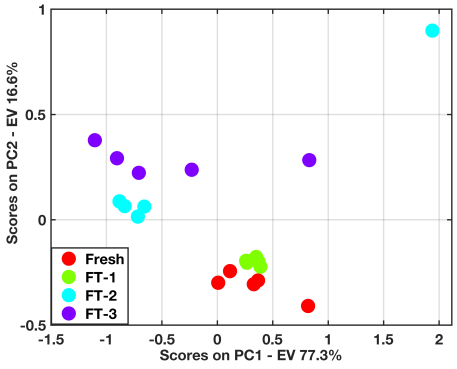


(c) Dendrogram from hierarchical clustering analysis

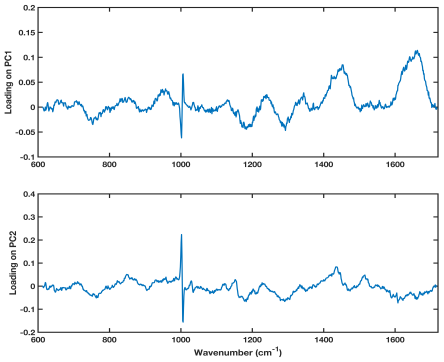
Figure 5.15: PC score, PC loading and hierarchical cluster analysis plots for liquid serum samples that are fresh and those that have undergone freeze-thaw cycles. Samples that were fresh or had gone through fewer freeze-thaw cycles were grouped most similarly. In this case FT-n is the number of freeze-thaw cycles that a sample had been through.

Figure 5.16 shows the PCA score plot, PCA loadings for the first three PCs and also a HC analysis dendrogram for the dry dataset. Unfortunately, within

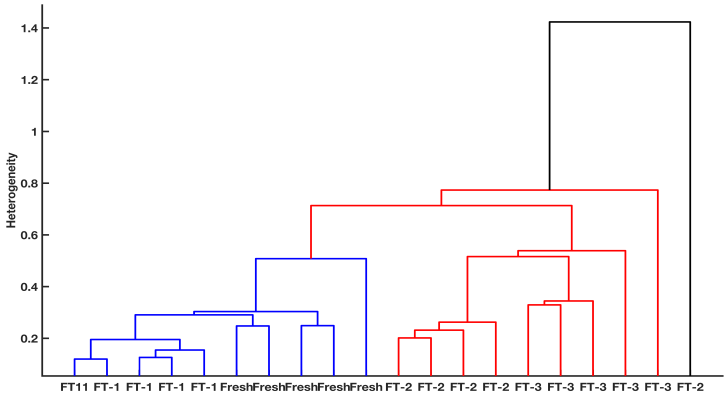
this dataset there was one spectrum that seemed to be an outlier from the main dataset. This was attributed to a large background contribution and low SNR for this data point which can be seen in the loading plot on PC1. The score plot in Figure 5.16 (a) shows separation between the main dataset and the outlier. The scale of the score plot shows a larger general variance in the dry dataset than the liquid dataset. Loadings for PC2 and PC3 show that the causes of variations from the same regions of the spectrum in PC2 to the PC1 on the liquid dataset, and PC3 was similar to PC2, however the magnitude of the variation is higher in the dry samples. This is expected as a dry dataset is taking point spectra from a specific region of the dried droplet whereas the liquid data can be thought of as an overall ensemble average for the sample.



(a) PC1 vs PC2



(b) Loadings



(c) Dendrogram from hierarchical clustering analysis

Figure 5.16: PC score plot for dry samples that are fresh and those that have undergone freeze-thaw cycles.

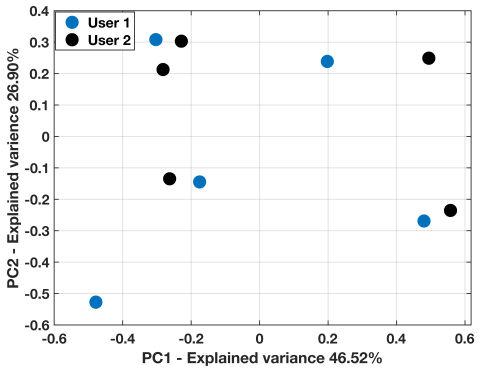
The PC score plot for PC1 vs PC2 also reveals that in the dry dataset the more freeze-thaw cycles that a sample had been through the wider range their PC scores were. Therefore in the dry samples the reproducibility of the data decreases with increasing freeze-thaw cycles.

As with the liquid spectra the fresh samples and the samples that had only been through one freeze thaw cycle (FT-1) grouped more closely than the samples that had been through repeated freeze-thaw cycles. These results were again confirmed by HC analysis. The dendrogram in (c) showed that apart from the rank outlier, the fresh and FT-1 samples grouped together and the FT-2 and FT-3 group were most similar. This suggests that repeated freeze-thaw cycles cause a gradual degradation of the serum samples.

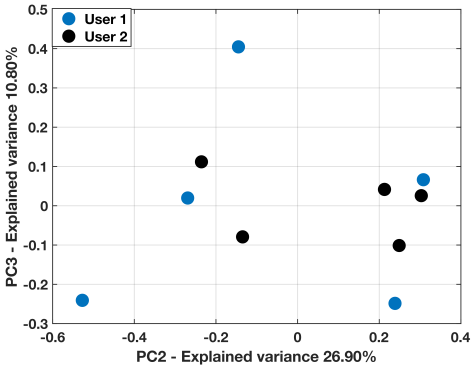
5.3.13 Investigating inter-operator variability

An investigation into inter-operator variability was conducted to test the robustness of the protocols developed for liquid and dry data acquisition. To investigate the robustness of the liquid protocol, spectra were taken by two different users from the same sample. Spectra were taken on the same day from the same well, the well plate was cooled during all spectral data acquisitions and the well used was cleaned between users using the protocol outlined in Chapter 3.3.

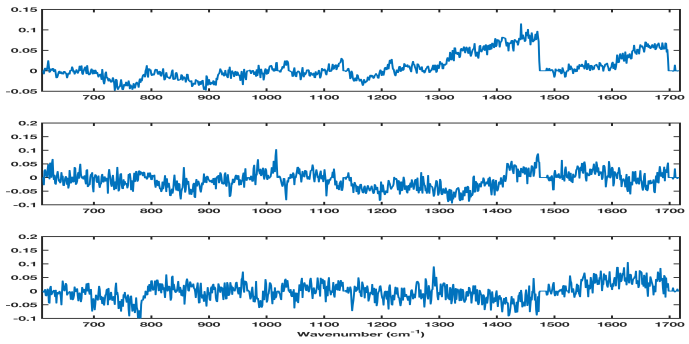
Figure 5.17 (a-b) shows the scatter plots for PC1 vs PC2 and PC2 vs PC3 for the liquid datasets. Figure 5.17 (c) gives the loading plots for the first three PC scores. The PC scores show no correlation between the user that took the spectra and the PC scores. This is confirmed when loadings on the PCs are investigated; the loadings show that the variance across the first three PCs has a very small magnitude and the loadings mostly just show noise. The HC dendrogram as seen in Figure 5.17 (d) also shows that the cluster analysis did not group spectra for liquid samples by the operator. This confirms that there is no correlation between the operator of the equipment and the spectral data acquired for the liquid protocol.



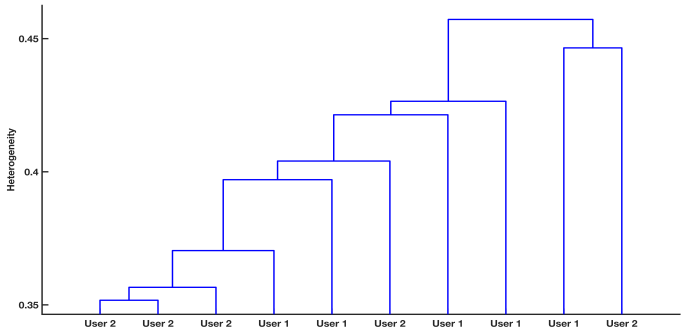
(a) PC1 vs PC2



(b) PC2 vs PC3



(c) PC score loadings for scores 1-3.

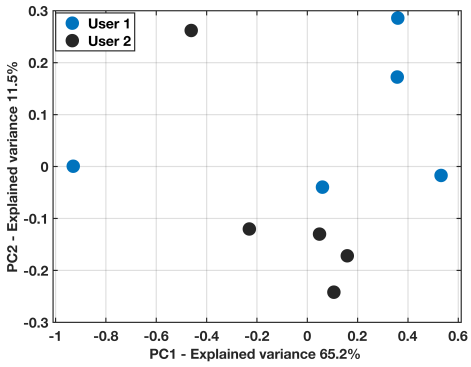


(d) Cluster analysis

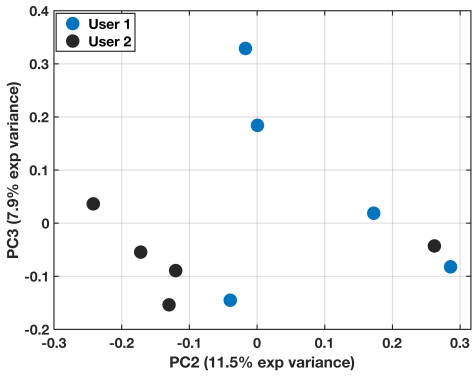
Figure 5.17: PCA scores (a-b) and PC loadings (c) for inter-operator investigation on the liquid dataset. The results were confirmed using HC analysis and plotting a HC dendrogram (d).

The experiment was also repeated for the dry protocol. For this both users took five point spectra across the same dry droplet from the same patient on the same day from the region in Figure 5.9. Figure gives the PC score plots for PC 1-3 and also the PC loadings across these scores as well as a HC analysis dendrogram.

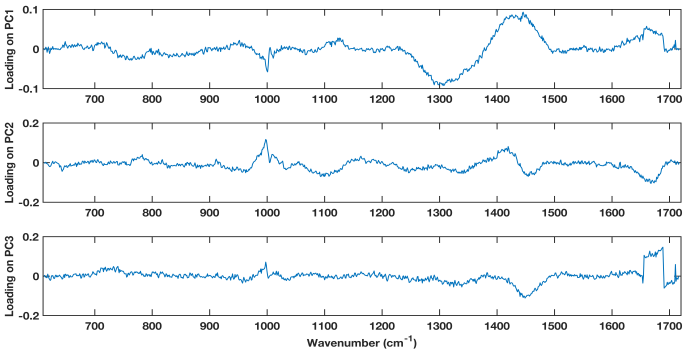
The dry score plot showed a larger PC score variance between the samples with a range of 1 to 0.5 compared to that of the liquid samples. There was some correlation between samples from user 1 and user 2. This is to be expected as the liquid data are an ensemble average of the sample taken from exactly the same point in the well whereas the dry data are taken randomly from the optimum ring on the dry droplet. The loadings show varying contributions in the 1200-1400 cm^{-1} spectral region where the aluminium background has varying contributions. The loading on PC2 which separates the three user 1 data points shows that there are spectral differences in the phenylalanine, Amide I ,CH2/3 stretch and in the Amide III/aluminium background regions. The loading shape suggests that the background for the three separated spectra were slightly different in shape to the other samples and due to the RCF background subtraction this is highlighted by sudden drops in the loading at around 1150 cm^{-1} and 1475 cm^{-1}. Between the other data points there is less variation, this is confirmed with the HC analysis which shows that the three spectra separated by PCA were also separated by HC analysis. However, there is no correlation with the other datapoints. This shows that there is generally more variation between datapoints in the spectra from the dry protocol compared with that from the liquid. Some spectra showed no correlation between which user took the spectra therefore, the region that dry spectra are taken from needs to be chosen very carefully to ensure reproducibility.



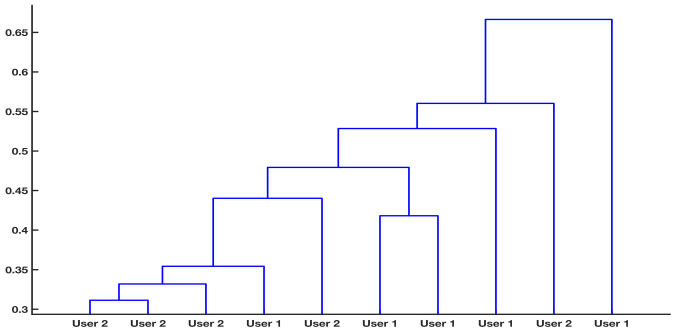
(a) PC1 vs PC2



(b) PC2 vs PC3



(c) PC score loadings for scores 1-3.



(d) Cluster analysis

Figure 5.18: PCA scores (a-b) and PC loadings (c) for inter-operator investigation on the dry dataset. The results were confirmed using HC analysis and plotting a HC dendrogram (d).

5.3.14 Investigating patient demographic effects on Raman spectra

When considering potential sources of variation to spectral data it is important to consider potentially unwanted biological variations that are not indicators of disease. Recent metabolomics studies have shown that patient demographic factors such as age, fasting status and sex can affect levels of these blood components and has lead to models that discriminate between patient sex and age rather than disease state [22]. The Raman spectrum of serum contains spectral information from a variety of components within the blood such as proteins, metabolites, lipids and DNA therefore the effects of patient fasting status, medication and sex on the variation of spectra were investigated.

To study the affects of patient fasting status and sex on Raman spectra data collected from a cancer cohort and a control cohort were compared. Data were compared using the dry protocol (785 nm) and the liquid protocol (785 nm and 532 nm). Participants involved in the studies for this thesis were also asked about their smoking status. It is expected that the smoking status may change a Raman spectrum but it would be inappropriate to discriminate towards a test for non-smokers and it would not be possible to control this variable. Therefore, participants who were smokers and non-smokers were included in all of the studies and are recorded in the relevant cohort information tables for each study.

5.3.15 Fasted vs non fasted samples

Cohort details

The cohort details for investigating the effect of fasting status are in Table 5.5; including 19 fasted patients vs 19 non-fasted patients with a mixture of control and cancer cases. Fasting status was determined by the patient upon recruitment into the study. Table 5.5 also contains any other data that was included in the patient records received from the patient Case Report Form (CRF).

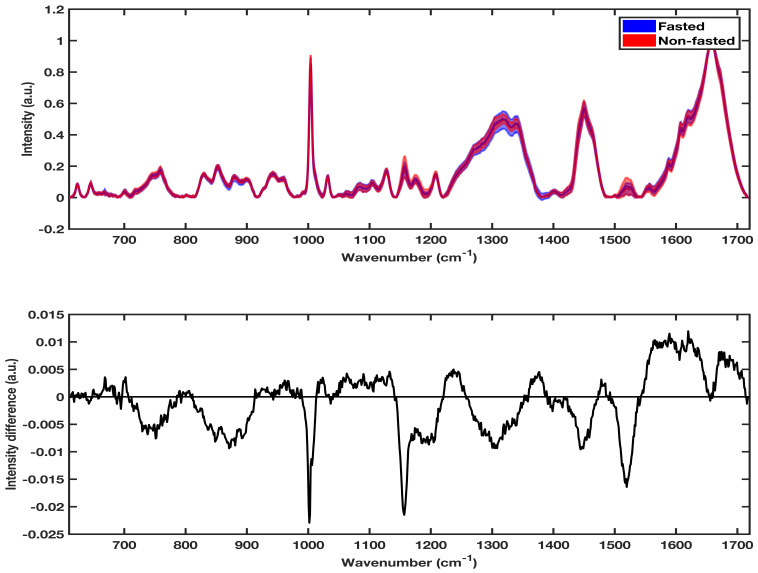
Table 5.5: Cohort details for fasting vs non-fasting patients to investigate effects on Raman spectra. Where Vasc. is vascular and CRT is undergoing chemo-radiotherapy.

Fasted				Non-fasted			
Group	**Sex**	**Age**	**Other**	**Group**	**Sex**	**Age**	**Other**
Cancer	Male	84	none	Cancer	Female	70	none
Cancer	Male	77	none	Control	Female	66	Ovarian cyst
Cancer	Female	71	CRT	Cancer	Male	65	none
Cancer	Male	83	none	Cancer	Female	93	none
Cancer	Female	80	none	Cancer	Female	81	none
Cancer	Female	60	none	Cancer	Female	80	Vasc. Demen-tia
Cancer	Male	68	none	Cancer	Female	73	none
Control	Female	79	IBS	Control	Female	71	none
Control	Female	63	none	Control	Male	60	none
Control	Female	58	none	Control	Female	22	none
Control	Female	50	Prev cancer	Cancer	Male	66	none
Cancer	Male	49	none	Control	Male	53	none
Cancer	Male	46	none	Cancer	Male	86	none
Cancer	Male	71	none	Control	Male	53	none
Control	Male	60	Prev cancer	Control	Male	63	pre cancer
Control	Female	77	none	Control	Female	75	none
Control	Female	42	none	Cancer	Female	83	none
Control	Male	77	none	Control	Female	73	none
Control	Female	82	none	Cancer	Male	61	none

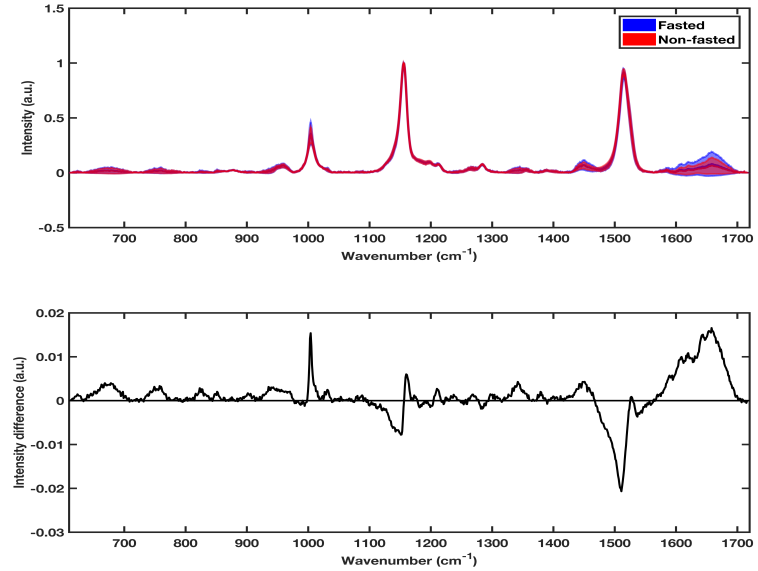
Average spectral comparison

Figure 5.19 shows a comparison between average spectra for fasted and non-fasted samples for 19 cancer and 19 control patients with 785 nm (a) and 532 nm (b) excitation with difference spectra to highlight the regions of spectral differences. The 785 nm (NIR) spectra show that the fasted samples have a higher standard deviation in the 1200-1400 cm^{-1} spectral region but higher variation in the non-fasted samples in spectral regions attributed to carotenoids, lipids, glycoproteins and amino acids. This is to be expected as these levels vary with the time a patient last ate. The difference spectra show that the non-fasted samples have relatively higher levels of amino acids, lipids, glycoproteins and fatty acid chains in the 725-760 cm^{-1} and then 830-904 cm^{-1} regions. They also show higher levels of carotenoids at 1157 cm^{-1} and 1520 cm^{-1} in the non-fasted spectra and differences in the shoulder of the phenylalanine peak at 1002 cm^{-1}. Previous studies have shown that the regions that vary between the fasted and non-fasted spectra can also be attributed to differences between cancer and controls in serum samples, therefore it was decided in future studies with 785 nm excitation that fasted patients should be used [23–25].

The 532 nm (Vis) mean spectra show very few general differences apart from a higher level of carotenoids in the non-fasted samples and slightly higher phenylalanine levels in the fasted patients. The standard deviation of the fasted samples is seemingly higher than that of the non-fasted samples. This is contrary to logic as one would expect samples that have been fasted to have more constant levels of lipids, lipoproteins, carotenoids and mono-saccharides compared to a non-fasted cohort which will have variations due to the time differences caused by the last time a person ate. The main region of spectral difference and variation in both the 785 nm and the 532 nm spectra are in regions associated to proteins. Therefore, further to investigating the overall differences between fasted and non-fasted samples it was considered that the effect of fasting may affect the spectra of patients with cancer and control patients differently. Therefore PCA was used to investigate the spectral variance by fasting status within patient groups.



(a) Mean NIR spectra and difference.



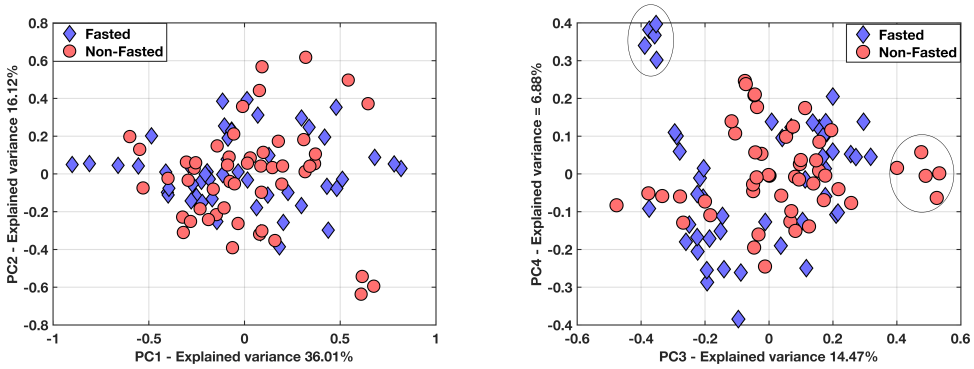
(b) Average 532 nm spectra and difference.

Figure 5.19: Average serum spectra comparing fasted vs non-fasted patients for 785 nm (a) and 532 nm(b).

785 nm PCA analysis

Figure 5.20 (a-b) shows the effect of fasting status on the principal component scores for patients who have confirmed colorectal cancer. The variance within

the scores on Figure 5.20 (a) shows that in general the variance within the cancer cohort is large. There is no correlation/grouping between fasted vs non-fasted patients. When looking at components that contribute to smaller spectral variances such as PC3 vs PC4 (Figure5.20 (b)) there still does not appear to be a dependence on fasting status. However, the PC scores did reveal outliers; the circular shape encasing blue triangles are patients that had undergone CRT treatment prior to sample collection. The red circles that also appear to be outliers and the spectra are from a patient who also has vascular dementia. The overall variance in the PC scores for cancer patients and the outliers show that that a cancer patient's medication, treatment or other conditions have a greater effect on the spectral variance than fasting status in the NIR cancer spectra.



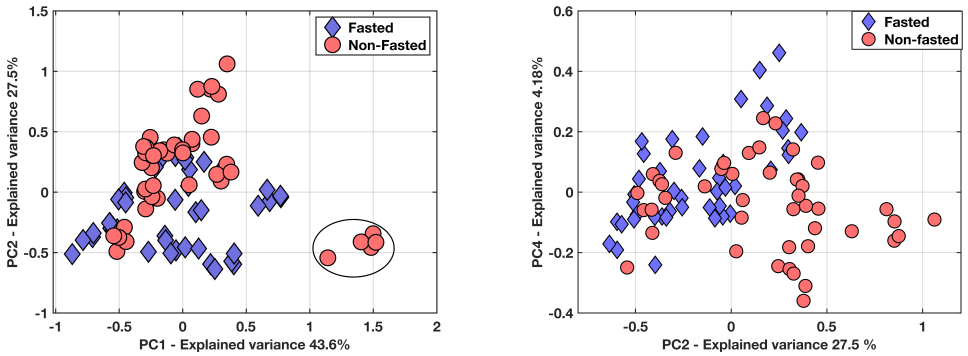
(a) PC1 vs PC2

(b) PC3 vs PC4

Figure 5.20: PC Score plots for 785 nm data from fasted vs non-fasted patients with confirmed CRC.

PCA analysis was repeated for the control patient cohort. Figure 5.21 shows the PC score plots for the control fasted vs non-fasted samples. It is clear that the PC scores are more sensitive to fasting status in control patients than in the confirmed cancer patients. Furthermore, in the control patients the spread of non-fasted PC scores is higher than that of the fasted patients. This is likely due to the control patients having smaller variations in medication leading to higher sensitivity of Raman to the fasting status. The outlier circled in Figure 5.21 (a) is a patient who had an anal fistula. The loadings on PC1 show that PC1 is dominated by an increase in the height of the peak at 1440 cm^{-1} attributed to

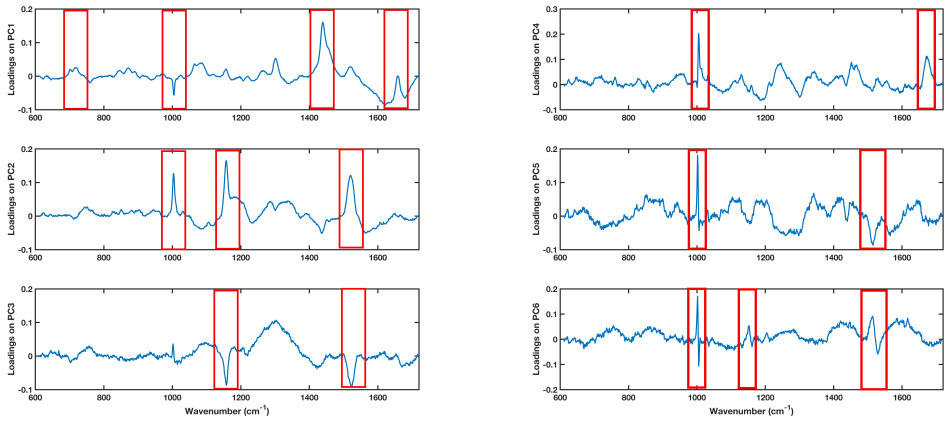
the CH_2/CH_3 stretching and is present in lipids, fatty acids and carbohydrates which is attributed to some clustering between the fasted and non-fasted patients as well as the non-fasted patient with an anal fistula. The loadings on PC2 show differences similar to that in the difference spectra in Figure 5.19 (a) and also shown in previous work within the group to be attributed to fasting status of the patient [9].



(a) PC1 vs PC2

(b) PC2 vs PC4.

Figure 5.21: PC score plots for fasted vs non fasted control patients. With PC1 vs PC2 (a) and PC2 vs PC4 (b). PC3 was not shown due to its dependence on the same wavenumbers as PC2.



(a) Loadings on PCs 1-3

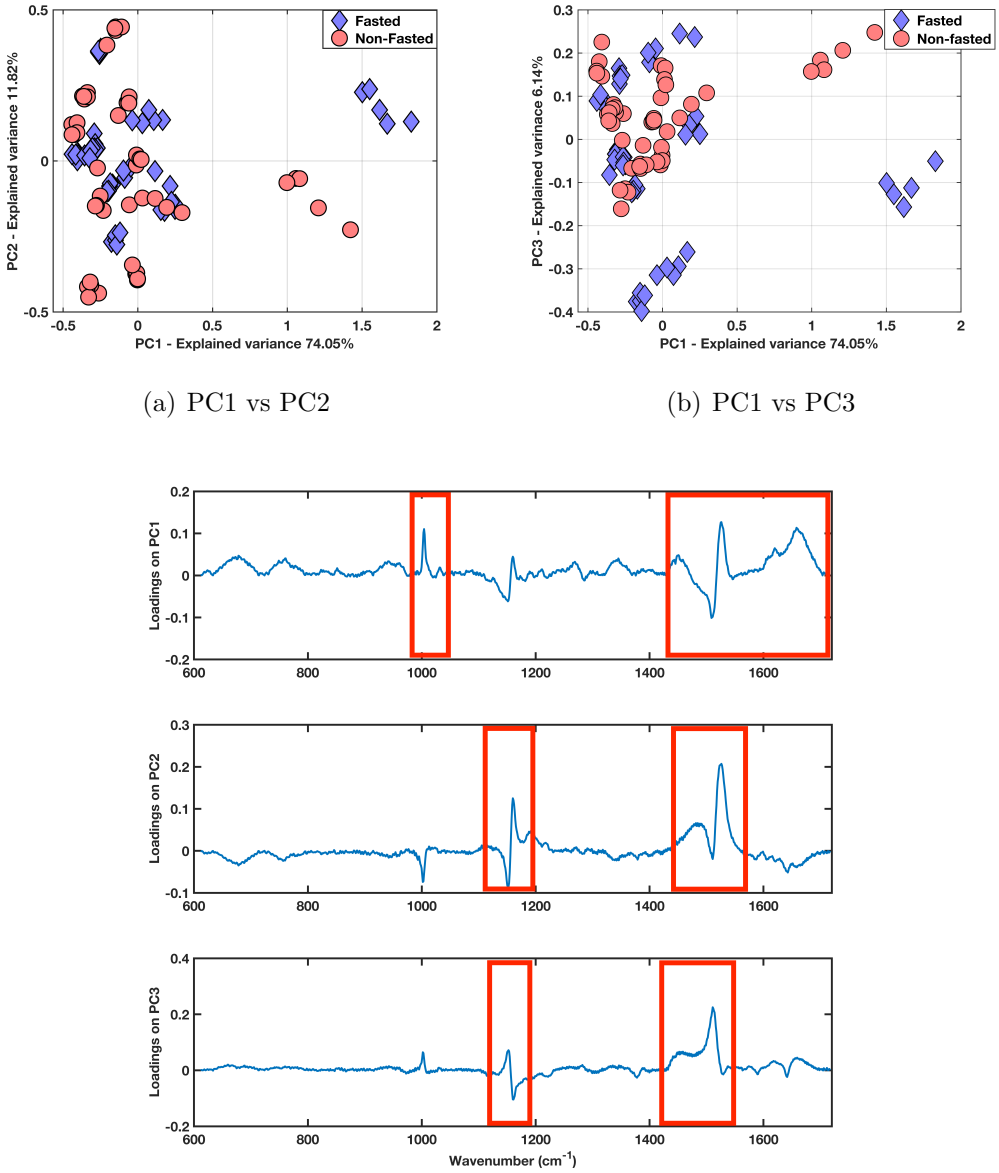
(b) Loadings on PCs 4-6

Figure 5.22: Loadings on PC 1-6 for fasted vs non-fasted patients. Loadings on PC 1,2 and 3 (a) and PCs 4,5 and 6 (b). Features associated with separation between fasting and non-fasting are highlighted (boxes).

The loadings on PCs 4-6 are dominated by a shift at the phenylalanine peak and also an increase in the lipid (1447 cm^{-1}) Amide I regions of the spectra (1658 cm^{-1}) for fasted patients. This indicates that the sensitivity of serum Raman spectroscopy to different analytes within the blood on fasting status. This, along with both the difference spectra and the loadings highlighting spectral regions used previously for discrimination, lead to fasted samples being used.

532 nm excitation PCA analysis

Figure 5.23 shows PC scores and loadings for fasted vs non-fasted patients for the cancer cohort for Vis spectra. There is no correlation between clustering and fasting status for the cancer patients in the PC score plots. The score plot for PC1 vs PC2 show two clear outliers from the rest of the patient dataset. On further investigation of the patient demographics it was found that the fasted outlier had late stage cancer in the rectosigmoid area and the non-fasted outlier had finished CRT for rectal cancer three months previously and was now clear of CRC. The loadings on PC1 show a mixture of features characteristic of NIR and Vis spectra including shifts of the carotenoid peaks as well as peaks in the Amide I region not normally seen in Vis spectra. This shows that for the outlying patients along PC1 there is a loss of resonance in the 532 nm spectra. This is possibly due to the site of the cancer and also the medication/treatment that the patients are undergoing. Figure 5.23 (b) shows the PC score plot of PC1 vs PC3. This plot shows a further two patients that are separated from the main group of spectra. These patients were also rectal cancer patients where the main group of patients were colon cancers. The loading on PC3 shows a shift in the carotenoid peak position at 1156 cm^{-1} and a decrease in the lipids and carotenoids in the rectal cancer outliers; the features are also shared in the PC1 and PC2 loadings. This indicates a sensitivity of serum excited at 532 nm to cancer site and patient medication. The effect of the outliers is not seen in the 785 nm spectra which indicated that 785 nm and 532 nm could potentially be used as complimentary wavelengths to gain different information about cancer patients, such as cancer site within a model.



(a) PC1 vs PC2

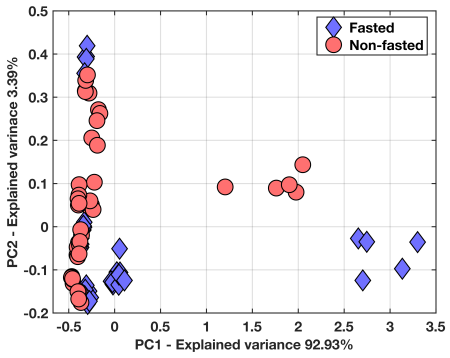
(b) PC1 vs PC3

(c) Loadings

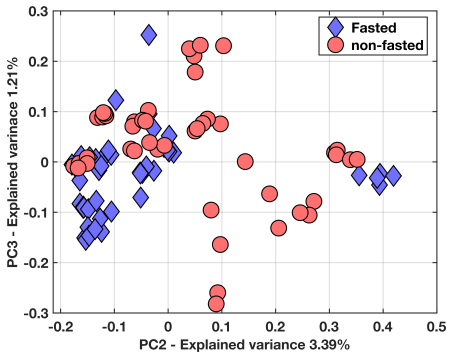
Figure 5.23: PC scores and loadings plots for 532 nm spectra of fasted vs non-fasted cancer patients.

The effect of medication and site of cancers is also visible within the control cohort that included patients with previous rectal cancer. Figure 5.24 shows the PC scores and loadings for the fasted vs non-fasted control patients using 532 nm excitation. Again, there is no correlation between fasting status and the PC clustering in the PC1 vs PC2 plot (Figure 5.24) but there are two outlying patients within the dataset. The outlying patients were both patients with previous rectal

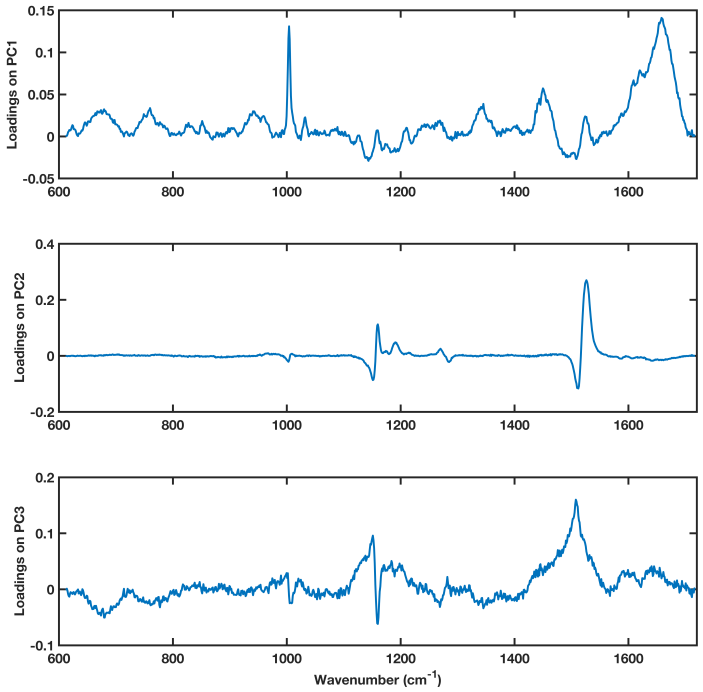
cancer who had been treated for the disease within 6 months previous to the blood samples being taken but were now cancer-free.



(a) PC1 vs PC2



(b) PC2 vs PC3.



(c) Loadings on PCs 1-3

Figure 5.24: PC score plot for fasting status of control patients.

The 532 nm spectra are therefore more susceptible to variance due to the medication a patient has been given and any previous conditions (cancer) of the disease. One of the patients was also found to be an outlier in the NIR data but the loss of resonance in the Vis spectra show that it is more sensitive than 785 nm to the spectral changes. The loading on PC1 for the control data shows a loading

that looks almost identical to a typical NIR serum spectrum. This again shows that the CRT has had an effect of lessening the effects of the resonance of the 532 nm data. This is either due to a change in structure or a decrease of complex conjugated molecules e.g. carotenoids in the patients with previous or current rectal cancer and CRT. The score plot for PC2 vs PC3 also show the non-fasted patients to have a larger overall variance than the fasted patients and separates out the same two patients. The loadings on PC2 and PC3 also show a shift in the carotenoid peaks similar to in the cancer case.

Therefore, the 785 nm and 532 nm spectra are susceptible to different aspects of the participant details. 785 nm spectra are more sensitive to the fasting status of a patient whereas 532 nm spectra are more sensitive to the treatment a patient has received or the position of the patients cancer.

5.3.16 Medication

The effects seen in the PCA loadings of the fasting vs non-fasting spectra were further investigated by looking at representative examples of raw spectral data. Figure 5.25 shows a comparison between NIR and visible spectra from three different control patients. Figure 5.25 (a-b) is that of a representative NIR (a) and Vis (b) spectrum for a control patient. Part (c-d) of the figure show the raw spectra of a control patient with previous cancer from the previous example and (e-f) show the raw spectra of a control patient who has diabetes. Effects seen in the PC loadings in the above section are also seen in the raw spectra. There are very small differences in the 785 nm spectra between all three control patients, however, there are large differences in the raw 532 nm spectra. The previous study showed cancer patient showing different spectral features and a decrease in the resonance effect in 532 nm spectra. The diabetic patient also showed a change in fluorescence baseline direction. Therefore, the 532 nm spectra are clearly more sensitive to medication and disease state than the 785 nm spectra and show some potential for the monitoring of disease post-treatment. The 532 nm spectra are more sensitive to fasting status than 532 nm spectra. In terms of a diagnostic model the decision was made to limit the training of a diagnostic model

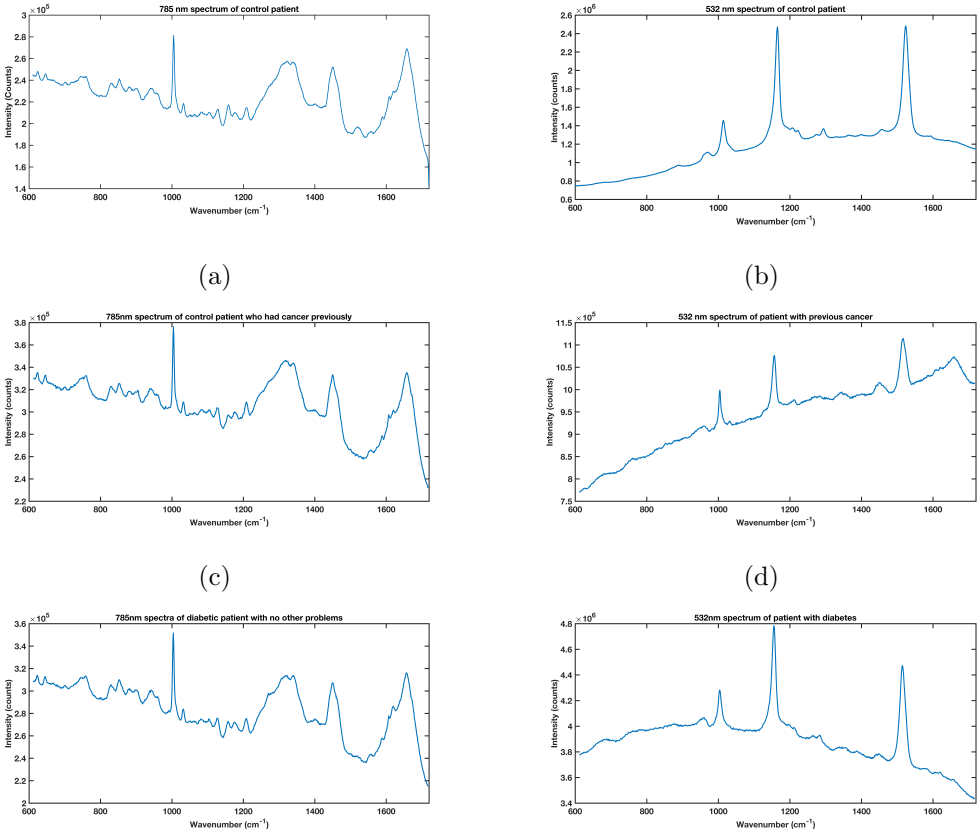


Figure 5.25: Spectra of a control patient with no other health issues with the 785 nm (a) and 532 nm (b); Spectra of control patient who had previous cancer and had CRT with 785 nm (c) and 532 nm (d); Spectra of a control patient with diabetes with 785 nm (e) and 532 nm (f) lasers.

to patients who had previously fasted for more than 6 hours to ensure spectral differences were due to disease. The sensitivity of 532 nm spectra to medication and potentially position of cancer within the body highlights the potential for a multi-modal approach to diagnostic encompassing both types of spectral data. A comparison of 785 nm spectra and 532 nm spectra for diagnostic capability and also the potential of adding the data together is explored more in Chapter 7.3.7.

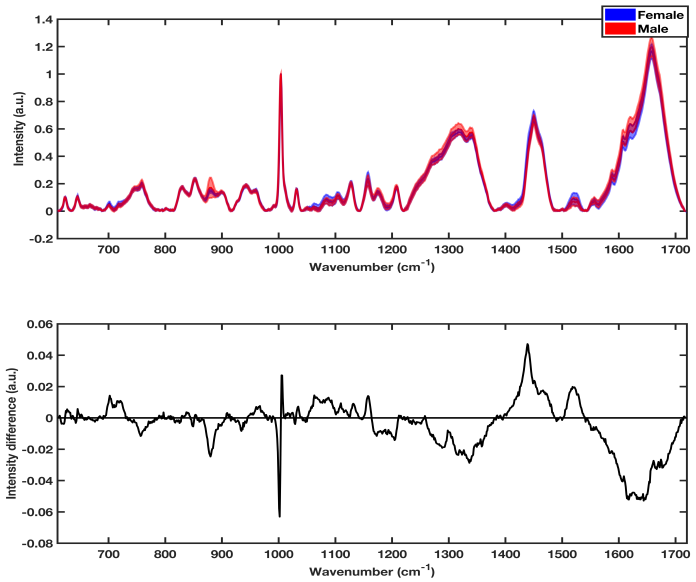
5.3.17 Sex

The sex of a patient can also determine affect the metabolites within serum and there have been mass spectral and chromatography studies profiling the differences between circulating metabolites in serum samples from a UK population [26]. Dunn et al reported differences in metabolites that have strong Raman spectral signatures such as tryptophan and tyrosine, therefore the effect of patient sex on the Raman spectral signature was investigated. Liquid data from both 785 nm and 532 nm laser lines were collected from a cohort of 10 males and 10 females who were confirmed to not have cancer, the patients demographics are summarised in Table 5.6.

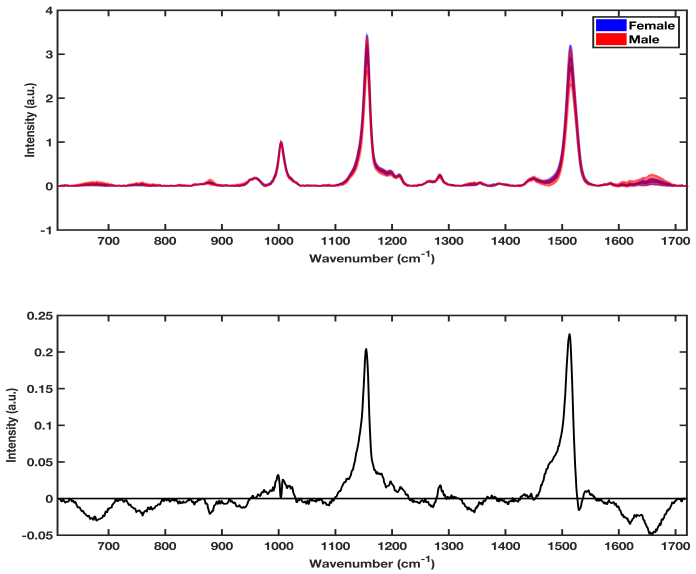
Table 5.6: Age and study number of the control patients used for investigation to the effect of sex on spectra.

Female		**Male**	
Study number	Age	Study Number	Age
80	78	55	78
102	69	67	65
114	63	88	54
115	58	91	86
128	77	104	60
132	82	139	69
153	67	145	78
154	54	147	56
179	59	159	73
185	73	184	76
Mean Age (yrs)	**68 ± 9.5**		**69.5 ± 10.6**

Figure 5.26 shows mean and difference spectra taken with the 785 nm and 532 nm for fasted females and males. The 785 nm difference spectra shows that female patients have higher levels of lipids (700 cm^{-1}), DNA ($\approx$ 1080 cm^{-1}) and CH2/CH3 lipids and carotenoid regions.



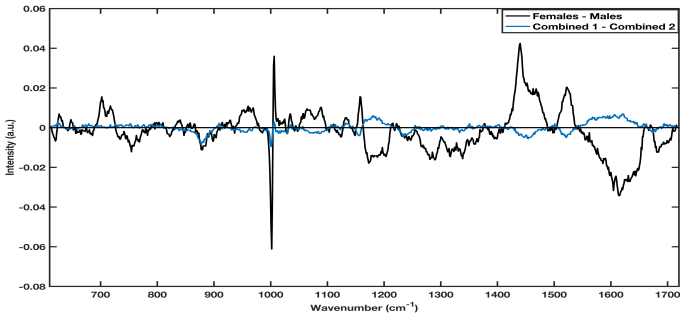
(a)



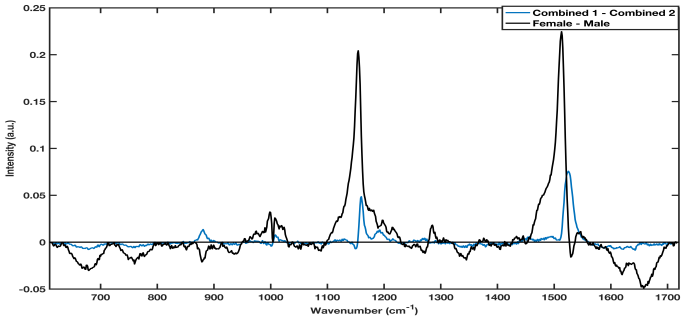
(b)

Figure 5.26: Mean and difference spectra for male and female patients (female - male) from 785 nm (a) and 532 nm lasers (b).

In both the 785 nm and the 532 nm spectra the Amide I region and phenylalanine and tyrosine peaks at 1600-1660 cm^{-1} have a larger standard deviation as well relatively higher levels in the male patients compared to the female patients. Both spectra also show higher levels of phenylalanine in the male patients. As a control measure to ensure differences in spectra were due to sex of the patient and not general variance between samples, an extra investigation was conducted. The spectra from the 10 male and 10 female control patients were randomly mixed together in even cohorts of 5 male and 5 female and combined to produce 'combined control spectra' cohorts. These were then subtracted from each other to produce difference spectra as in the sex study. Figure 5.27 shows the difference spectra from the female - male spectra for 785 nm and 532 nm spectra with the differences in the combined spectra overlaid in each case. It is clear that the



(a)



(b)

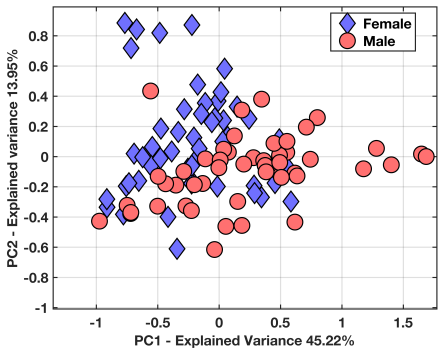
Figure 5.27: Difference in spectra for female - male patients overlaid with combined cohorts from 785 nm (a) and 532 nm lasers (b).

differences when the cohorts are combined are an order of magnitude lower than in the case where female and male spectra are combined. This shows that the

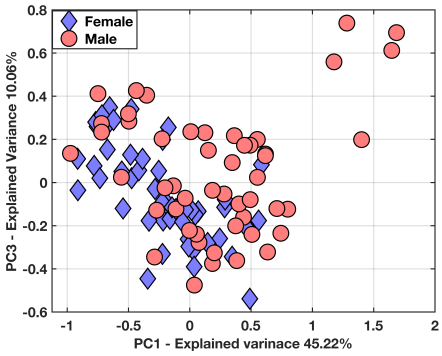
differences within a general dataset are much smaller than those due to the sex of the patients. Furthermore these plots double as a 'control patient - control patient' test showing that there are small differences between control - control spectra. Therefore reinforcing that the differences seen between control and cancer patient spectra are real differences.

To highlight differences in the spectra that may affect diagnostic/classification models, PCA was used to investigate the differences in 785 nm and 532 nm spectra. Figure 5.28 shows the PC scores and associated loadings for liquid 785 nm spectral data. The PC1 vs PC2 score shows clustering within male and females. This is also shown in the PC1 vs PC3 plot. The loadings on PC1 show differences in a shift in the phenylalanine and increased levels in the amide region of the spectra in agreement with the difference spectra for the 785 data. The loadings on PC2 and PC3 also agree with the difference spectra showing shifts in the phenylalanine peak between the sexes and in the loadings of the CH2/CH3 lipid region. As with fasting status, these regions, including shifts in the phenylalanine peaks and changes in carotenoid levels, have been associated with spectral changes due to malignant diseases [4]. It is clear that at 785 nm serum data are sensitive to the sex of the patient, therefore it is possible that there may be improvements in diagnostic capability if the diagnostic models are split for 785 nm data by sex of the patient. This will be discussed further in Chapter 6.4.5.

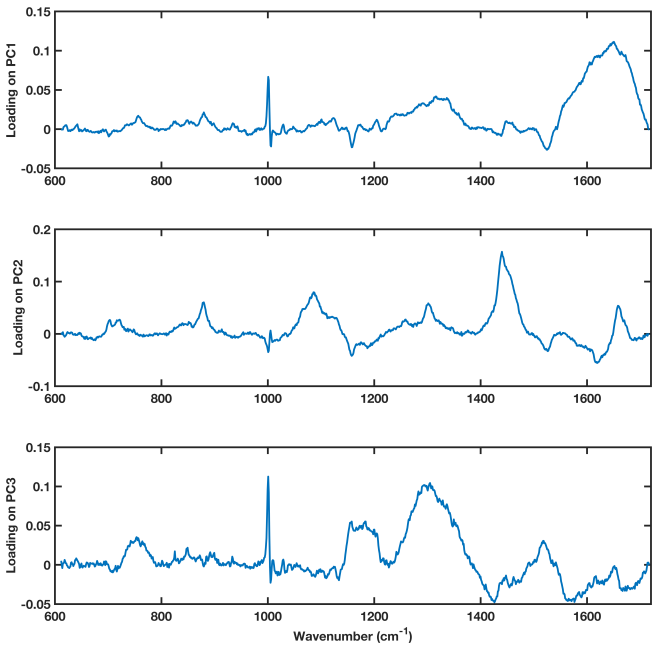
When considering the 532 nm spectra (Figure 5.29) there is a smaller dependence on male vs female patients. As before, there is a clear outlier in the 532 nm spectra along PC1 associated with a decrease in the resonance effects. The patient demographics associated with outlying spectra show that the patient had diverticular disease. This result again suggests that as this patient was not an outlier in the 785 nm data, the 532 nm data are more sensitive to co-morbidities, medication and site and type of disease.



(a) PC1 vs PC2

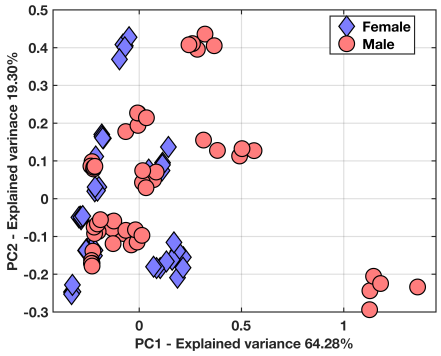


(b) PC1 vs PC3

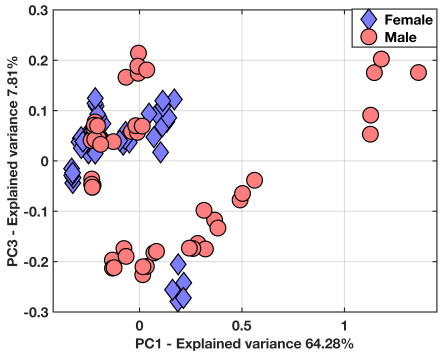


(c) Loadings on PC1 to PC3.

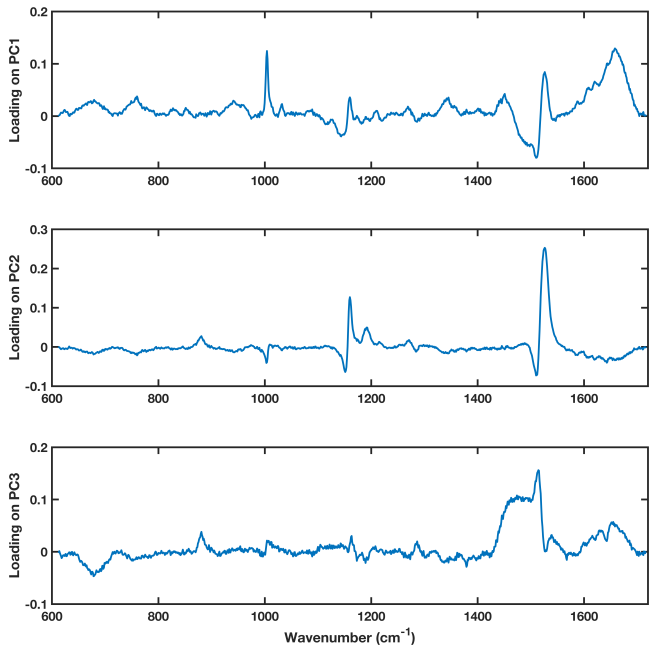
Figure 5.28: Scores plot for PC1 vs PC2 (a) PC1 vs PC3 (b), PC loadings on PC1-3 (c) for sex of patient with 785 nm laser line.



(a) PC1 vs PC2



(b) PC1 vs PC3



(c) Loadings pn PC1 to PC3

Figure 5.29: Scores plot for PC1 vs PC2 (a) PC1 vs PC3 (b) and PC loadings on PC1-3 (c) for sex of patient with 532 nm laser line.

5.3.18 Age and smoking status

The risk of acquiring colorectal cancer increases with age; the cohort of patients recruited during this work has a mean age of 67 years with a standard deviation of 13 years, therefore it is difficult to investigate the changes that are associated with age within the cohort. In future a study of spectral changes would be useful to determine natural differences in age compared to differences caused by natural

ageing and those associated with disease risk. Furthermore, the general cohort of patients who would be within the urgent suspected cancer referral route for CRC patients are generally within the age range of patients that diagnostic models have been trained on in this work.

The diagnostic models in this work were trained on samples from a mixture of patients who were smokers and non-smokers. There have been previous reports showing differences in Raman spectra between smokers and non-smokers however this factor is not controllable in a patient population (as fasting would be) therefore all models were trained on a mixture of smokers, non-smokers and ex-smokers across all of the patient groups to allow for spectral differences caused by smoking.

5.3.19 Discussion and future considerations

The aims of this chapter were three-fold; firstly the optimum biofluid for Raman data acquisition was investigated and characterised. Secondly, protocols were developed and optimised for serum data acquisition using both liquid and dry samples. Finally, pre-analytical considerations such as sample handling and inter-operator reproducibility were investigated to test the robustness of the protocols developed along with potential patient demographic effects on serum spectra.

The results from the first investigation showed that serum was the optimum biofluid for analysis of CRC samples with Raman spectroscopy; serum showed less spectral variability than plasma samples after freeze-thaw cycles. Furthermore, plasma samples are subject to variation from different anticoagulants whereas serum samples are achieved from allowing a blood sample to clot, therefore this negates any possible spectral interference from different anticoagulants or different concentrations of anticoagulants. Serum was then characterised using Raman and FTIR spectroscopy and a literature search was conducted to (wherever possible) assign spectral peaks to biologically relevant molecules such as DNA, amino acids, in the serum.

FTIR spectra had good SNR and required minimal sample preparation i.e. pipetting onto the diamond ATR substrate. However, for each measurement the

sample has to be pipetted onto the crystal and when drying samples each sample has to be dried for approx. 20 mins before spectral acquisition can take place. This is very detrimental for any technique that is looking towards translation as the throughput potential is low. Due to the slow spectral acquisition time for the FTIR methods the rest of the work in this thesis was focused upon Raman spectral applications.

The measurement conditions and protocols for dry and liquid spectral data collection were developed and optimised. Dry spectra could not be acquired using the 532 nm laser however good quality spectra were collected from the 785 nm laser line. The use of chemometric analysis allowed the optimisation of dry sample data acquisition, aluminium foil proved to be a cheap and effective substrate that was able to be used in a high-throughput configuration. It was found that spectra can vary across a serum droplet in accordance with the Vroman drying effect of serum. The optimum region of a dry droplet for Raman spectral data acquisition was found to be a band around the center. This was determined using PCA mapping analysis which maximised spectral intensity and minimised spectral variance.

A stainless steel well plate was designed as a new substrate to allow high throughput analysis of liquid serum samples. The position for spectral measurements in the wells was optimised and found to be 1200 μm from the focus position of a notch as the base of each well. The spectral effects of evaporation and fluorescence contribution were reduced when the well plate was temperature stabilised to between 18-20°C. It was also decided to focus on serum Raman spectroscopy due to it's ability to be used in high throughput in both dry and liquid platforms.

For both liquid and dry serum samples that had been subject to multiple freeze-thaw (FT) cycles, the overall PC score range was small therefore generally the reproducibility of the spectra was good. Spectra could be acquired after at least 3 FT cycles. Fresh serum samples and those that had been through just one FT cycle showed less degradation and PC score variability than those that had been through more FT cycles. Chemometric analysis showed that fresh samples and those that had been through one FT cycle were most similar in both sampling modalities. Therefore, future analysis or any comparisons will

concentrate on samples that are fresh or have gone through only one freeze-thaw. Any spectral comparisons of data from more FT cycles is possible however the analysis must be done with care to ensure that differences due to the degradation of the sample are taken into account.

There is no correlation in the liquid protocol between operators of the equipment and the spectra acquired proving the robustness of the liquid protocol. However, the dry dataset showed some inter-user variation. This was only true for some spectra so to negate this, tighter control and better instructions (including PCA picture in the protocol) of the area of the dried droplet in which spectra should be taken was adopted.

It was found that 785 nm difference spectra of fasted and non-fasted patients and PCA differences in fasted vs non-fasted control patients have differences. The spectral regions with increased variance due to the fasting status have been previously shown in literature to be attributed to the differences between cancer and control samples. Therefore it was decided that fasted patients would be optimal for this work. Furthermore, 532 nm fasted and non-fasted spectra show no correlation based on fasting but in general 532 nm spectra are more sensitive to medication effects of the patients.

Finally, there is clear separation in male vs female patients in 785 nm spectra leading to the potential of separated models for each sex. There is no correlation between sex and 532 nm spectral variation in PCA score plots but again 532 nm spectra show sensitivity to patient comorbidities. As a result of the 532 nm analysis the potential of a multi-model approach is explored further in Chapter 7.3.7.

Future considerations

It has been noted that a further method of x axis calibration and a method of monitoring 'dark noise' within spectra is to use an external calibration standard such as standardised green glass [27]. However, this was discovered after the commencement of recruiting fresh Raman spectra for serum samples. To ensure that all the data had been treated equally during this project an external standard was not used. However, the use of an external standard to calibrate the instrument has been added into subsequent SOPs for Raman data acquisition.

Bibliography

[1] Caryn Sian Hughes. Development of Fourier Transform Infrared Spectroscopy for Drug Response Analysis. pages 30–50, 2011.

[2] Kenny Kong, Catherine Kendall, Nicholas Stone, and Ioan Notingher. Raman spectroscopy for medical diagnostics - From in-vitro biofluid assays to in-vivo cancer detection, 2015.

[3] Matthew J. Baker, Shawn R. Hussain, Lila Lovergne, Valérie Untereiner, Caryn Hughes, Roman A. Lukaszewski, Gérard Thiéfin, and Ganesh D. Sockalingum. Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chem. Soc. Rev.*, 45(7):1803–1818, 2016.

[4] Dinesh K. R. Medipally, Adrian Maguire, Jane Bryant, John Armstrong, Mary Dunne, Marie Finn, Fiona M. Lyng, and Aidan D. Meade. Development of a high throughput (HT) Raman spectroscopy method for rapid screening of liquid blood plasma from prostate cancer patients. *Analyst*, 142(8):1216–1226, 2017.

[5] HA Krebs. Chemical composition of blood plasma and serum. *Annual review of biochemistry*, 19(1):409–430, 1950.

[6] Zhonghao Yu, Gabi Kastenmüller, Ying He, Petra Belcredi, Gabriele Möller, Cornelia Prehn, Joaquim Mendes, Simone Wahl, Werner Roemisch-Margl, Uta Ceglarek, Alexey Polonikov, Norbert Dahmen, Holger Prokisch, Lu Xie, Yixue Li, H. Erich Wichmann, Annette Peters, Florian Kronenberg, Karsten Suhre, Jerzy Adamski, Thomas Illig, and Rui Wang-Sattler. Differences between human plasma and serum metabolite profiles. *PLoS One*, 6(7):1–6, 2011.

[7] Landulfo Silveira, Rita de Cássia Fernandes Borges, Ricardo Scarparo Navarro, Hector Enrique Giana, Renato Amaro Zângaro, Marcos Tadeu Tavares Pacheco, and Adriana Barrinha Fernandes. Quantifying glu-

cose and lipid components in human serum by Raman spectroscopy and multivariate statistics. *Lasers Med. Sci.*, 32(4):787–795, 2017.

[8] J. Duguid, D. F. O'Shaughnessy, C. Atterbury, P. Bolton Maggs, M. Murphy, D. Thomas, S. Yates, and L. M. Williamson. Guidelines for the use of fresh-frozen plasma, cryoprecipitate and cryosupernatant, 2004.

[9] Kathryn A Welsby. *Raman Spectroscopy of Blood Plasma : Immunological Applications in Prenatal Author :.* PhD thesis, Swansea University, 2016.

[10] DF O'shaughnessy, C Atterbury, P Bolton Maggs, M Murphy, D Thomas, S Yates, and LM Williamson. Guidelines for the use of fresh-frozen plasma, cryoprecipitate and cryosupernatant. *British journal of haematology*, 126(1):11–28, 2004.

[11] Miguela Martin, David Perez-Guaita, Dean W. Andrew, Jack S. Richards, Bayden R. Wood, and Philip Heraud. The effect of common anticoagulants in detection and quantification of malaria parasitemia in human red blood cells by ATR-FTIR spectroscopy. *Analyst*, 142(8):1192–1199, 2017.

[12] D. Rohleder, G. Kocherscheidt, K. Gerber, W. Kiefer, W. Kohler, J. Mocks, and W. Petrich. Comparison of mid-infrared and Raman spectroscopy in the quantitative analysis of serum. *J. Biomed. Opt.*, 10(3):031108, 2005.

[13] a J Berger, T W Koo, I Itzkan, G Horowitz, and M S Feld. Multicomponent blood analysis by near-infrared Raman spectroscopy. *Appl. Opt.*, 38:2916–2926, 1999.

[14] Jingwei Shao, Manman Lin, Yongqing Li, Xue Li, Junxian Liu, Jianpin Liang, and Huilu Yao. In Vivo Blood Glucose Quantification Using Raman Spectroscopy. *PLoS One*, 7(10):1–6, 2012.

[15] Norman Tschirner, Matthias Schenderlein, Katharina Brose, Eberhard Schlodder, Maria Andrea Mroginski, Christian Thomsen, and Peter Hildebrandt. Resonance Raman spectra of β-carotene in solution and in photosystems revisited: an experimental and theoretical study. *Phys. Chem. Chem. Phys.*, 11(48):11471, 2009.

[16] Li Cui, Holly J. Butler, Pierre L. Martin-Hirsch, and Francis L. Martin. Aluminium foil as a potential substrate for ATR-FTIR, transflection FTIR or Raman spectrochemical analysis of biological specimens. *Anal. Methods*, 8(3):481–487, 2016.

[17] Stacey L. Hirsh, David R. McKenzie, Neil J. Nosworthy, John A. Denman, Osman U. Sezerman, and Marcela M M Bilek. The Vroman effect: Competitive protein exchange with dynamic multilayer protein aggregates. *Colloids Surfaces B Biointerfaces*, 103:395–404, 2013.

[18] Olga Bajenova, Elena Tolkunova, Sergey Malov, Peter Thomas, Alexey Tomilin, and Stephen O Brien. The Role of the Carcinoembryonic Antigen Receptor in Colorectal Cancer Progression. *J. Integr. Oncol.*, 06(02), 2017.

[19] World Health Organization. Use of anticoagulants in diagnostic laboratory: stability of blood, plasma and serum samples. *Who*, pages 1–62, 2002.

[20] Frida Torell, Kate Bennett, Stefan R?nnar, Katrin Lundstedt-Enkel, Torbj?rn Lundstedt, and Johan Trygg. The effects of thawing on the plasma metabolome: evaluating differences between thawed plasma and multi-organ samples. *Metabolomics*, 13(6):1–10, 2017.

[21] Shunji Takehana, Hiroo Yoshida, Shinichi Ozawa, Junko Yamazaki, Kazutaka Shimbo, Akira Nakayama, Toshimi Mizukoshi, and Hiroshi Miyano. The effects of pre-analysis sample handling on human plasma amino acid concentrations. *Clin. Chim. Acta*, 455:68–74, 2016.

[22] Peiyuan Yin, Rainer Lehmann, and Guowang Xu. Effects of pre-analytical processes on blood samples used in metabolomics studies. *Anal. Bioanal. Chem.*, pages 4879–4892, 2015.

[23] Xiaozhou Li, Tianyue Yang, and Siqi Li. Discrimination of serum Raman spectroscopy between normal and colorectal cancer using selected parameters and regression-discriminant analysis. *Appl. Opt.*, 51(21):5038, 2012.

[24] Philip R T Jess, Daniel D W Smith, Michael Mazilu, Kishan Dholakia, Andrew C. Riches, and C. Simon Herrington. Early detection of cervical neoplasia by Raman spectroscopy. *Int. J. Cancer*, 121(12):2723–2728, 2007.

[25] O. J. Old, L. M. Fullwood, R. Scott, G. R. Lloyd, L. M. Almond, N. a. Shepherd, N. Stone, H. Barr, and C. Kendall. Vibrational spectroscopy for cancer diagnostics. *Anal. Methods*, pages 3901–3917, 2014.

[26] Warwick B. Dunn, Wanchang Lin, David Broadhurst, Paul Begley, Marie Brown, Eva Zelena, Andrew A. Vaughan, Antony Halsall, Nadine Harding, Joshua D. Knowles, Sue Francis-McIntyre, Andy Tseng, David I. Ellis, Steve O'Hagan, Gill Aarons, Boben Benjamin, Stephen Chew-Graham, Carly Moseley, Paula Potter, Catherine L. Winder, Catherine Potts, Paula Thornton, Catriona McWhirter, Mohammed Zubair, Martin Pan, Alistair Burns, J. Kennedy Cruickshank, Gordon C. Jayson, Nitin Purandare, Frederick C.W. Wu, Joe D. Finn, John N. Haselden, Andrew W. Nicholls, Ian D. Wilson, Royston Goodacre, and Douglas B. Kell. Molecular phenotyping of a UK population: defining the human serum metabolome. *Metabolomics*, 11(1):9–26, 2014.

[27] Kelly Marie Curtis. Comparing coherent and spontaneous Raman modalities for the investigation of gastrointestinal cancers . 2017.

Chapter 6

Optimisation of serum Raman spectroscopy for colorectal cancer detection

6.1 Introduction

The previous chapter discussed the optimisation of the experimental methodology for serum Raman spectroscopy (RS). This chapter will investigate the feasibility of using RS to detect CRC using serum samples in both liquid and dry form.

For RS to be successful at detecting CRC in serum samples it is important to consider the optimum conditions for laboratory analysis. However, the primary application of the serum Raman method would be aimed at use as a triage tool within the urgent suspected cancer referral pathway (discussed in Chapter 1.5). It is envisaged that serum RS would be translated for this use in a clinical setting so it should be noted that the ideal laboratory conditions are not always aligned with what is possible or practical within a clinical setting, largely through inter-patient variability. Therefore, as well as considering the optimum laboratory conditions for the technique, the implications of different sampling modalities for translation will also be discussed throughout the results within this chapter.

6.2 Aims and objectives

The aim of this chapter will be to show that serum RS has potential as a diagnostic tool for CRC using a small cohort of patients. It also aims to show the development of the spectral analysis routine and the effect of different patient demographics and sampling modes on diagnostic performance.

The chapter will begin with a feasibility study of an 'ideal' cohort of samples testing dry and liquid methods. It will then go on to optimise the preprocessing routines for maximum diagnostic output for the 785 nm dry spectral dataset and the 785 nm and 532 nm liquid datasets. Finally it will investigate the effects of different sampling modes on the diagnostic capability of the models and discuss the results of these investigations in terms of translation.

6.3 Materials and Methods

Sample Collection

Serum samples used in this chapter were processed as outlined in Chapter 3.1. Briefly, samples were drawn from patients and then span following manufacturer standards to produce serum. Serum samples were then aliquotted. One aliquot was taken for fresh serum analysis and the rest frozen at -80°C for future study. Fresh serum samples were analysed within 12 hours of the sample being drawn from the patient. Fresh samples were kept at 4°C until measurement. Freeze-thawed samples were thawed at room temperature and used immediately after thawing.

Patient Information: Pilot study

To build the diagnostic models within this chapter, 30 control patients with confirmed negative colonoscopies were selected for analysis, the samples were combined with 30 samples from patients with confirmed colorectal malignancies across all cancer stages. Patients were a mixture of smokers and non-smokers, age matched (Table 7.2) and a mixture of male and female.

The same patients were used in the pilot study and the effect of freeze-thaw on diagnostic capability study in this chapter. The spectral data from these patients were also used for the optimisation of the pre-processing parameters for diagnostic models.

Table 6.1: Patient details for pilot study.

Disease State	No patients	Mean age (years)	Smokers	Male	Female
Cancer	30	67.7 ± 9.7	8	15	15
Control	30	65.0 ± 12.8	3	13	17
Totals	**60**	**66.4±11.33**	**11**	**28**	**32**

6.3.1 Raman spectroscopy

Dry serum data collection

Data from dried serum samples were taken using the protocol optimised in Chapter 5.3.6. Briefly, 3 μl samples were pipetted in duplicate onto an aluminium substrate. Five point spectra were acquired from across the two droplets from within the optimised sample area (Chapter 5.3.6). Dry spectra were only collected using the 785 nm laser source using the parameters in Table 7.5.

Table 6.2: Optimised data acquisition conditions for different sampling modes.

	785 nm Laser		**532 nm Laser**
	Liquid	Dry	Liquid
Wavenumber range (cm^{-1})	610-1720	610-1720	610-1720
Grating (l/mm)	1200	1200	2400
Exposure time (s)	5	1	0.6
Accumulations	30	30	120
Laser Power (%)	100	100	100
Pinhole Y/N	N	N	N

Liquid data collection

Data for liquid samples for 785 nm and 532 nm was collected according to the protocol developed in Chapter 5.3.7. Briefly, data from liquid samples were taken using the stainless steel well plate, 200 μl of serum was pipetted into a cleaned well before each measurement and the microscope focused to 1200 μm above the focus point at the base of the well. The well plate was kept between 18-20°C during measurements. Wells were cleaned via the method outlined in Chapter 3.3 after use. The measurement parameters for liquid measurements for both the 785 nm and 532 nm lasers were as in Table 7.5.

6.3.2 Spectral Analysis

All spectra in this section were point spectra, all data were quality checked for minimum intensity and cosmic rays before exporting into the MATLAB environment.

Pilot study

Spectra were pre-processed using a rolling circle filter (RCF) with a radius of 150 for spectra taken with the 785 nm laser line and 300 for the 532 nm spectra and normalised to the phenylalanine peak at 1003-1004 cm^{-1}. Pre-processed spectra were then quality tested to ensure there were no clear outliers. If outliers were detected e.g. spectra that did not fit the general spectral trends the individual spectra were then removed from the model to avoid a skewed result. The data were then used to build discriminatory models using PLS-DA for dry and liquid datasets as described in Chapter 4.3.2. All PLS-DA models were cross validated using 5-fold k-fold cross validation. The PLS-DA model parameters used for each model minimised the cross-validation error of the models whilst keeping true spectral features in the number of loadings. The details of the number of loadings used will be stated in the corresponding results sections.

Pre-processing parameters

The parameters used for different background subtraction methods of dry serum data were as in Table 6.3. The parameters used for liquid data background subtraction are in Table 6.4.

Table 6.3: Optimised parameters for different pre-processing methods for dry serum Raman spectral data taken with 50x objective and the 785 nm laser.

Method	Parameters
RCF	R = 150; IO = 1
Polynomial	Order = 9; bin = 1; NI = 100
Derivative	Order=1; SP=9;PO=4

Where R is radius, IO is interpolation order, NI is number of iterations (max), PO is polynomial order, Derivative order = 1 is first derivative and SP is number of smoothing points.

Table 6.4: Optimised parameters for different pre-processing methods for liquid serum Raman spectral data taken with 10x objective.

785 nm Laser		**532 nm Laser**	
Method	Parameters	Method	Parameters
RCF	R = 200; IO = 1	RCF	r = 300; IO = 1
Polynomial	Order = 9; bin = 1; NI = 100	Polynomial	Order = 9; bin = 1; NI = 100
Derivative	Order=1; SP=9; PO=4	Derivative	Order=1; SP=9; PO=4

Where R is radius, IO is interpolation order, NI is number of iterations(max), PO is polynomial order, Derivative order = 1 is first derivative and SP is number of smoothing points.

Optimising pre-processing methodology

To optimise data pre-processing for PLS-DA models different combinations of pre-processing methods (discussed in detail in Chapter 4.3.2) were tested on 785 nm liquid and dry datasets and a 532 nm liquid dataset. The different combinations of pre-processing methods were investigated using cross-validated PLS-DA results as a measure of performance. Spectral data from participants in the pilot study data were used in this study.

To avoid carrying through fluorescence background contributions, background subtraction methods were employed on all datasets before normalisation [1]. Figure 6.1 shows the methodology for the optimisation of the pre-processing routing for the serum RS data. Pre-processed spectra from the nine different combinations

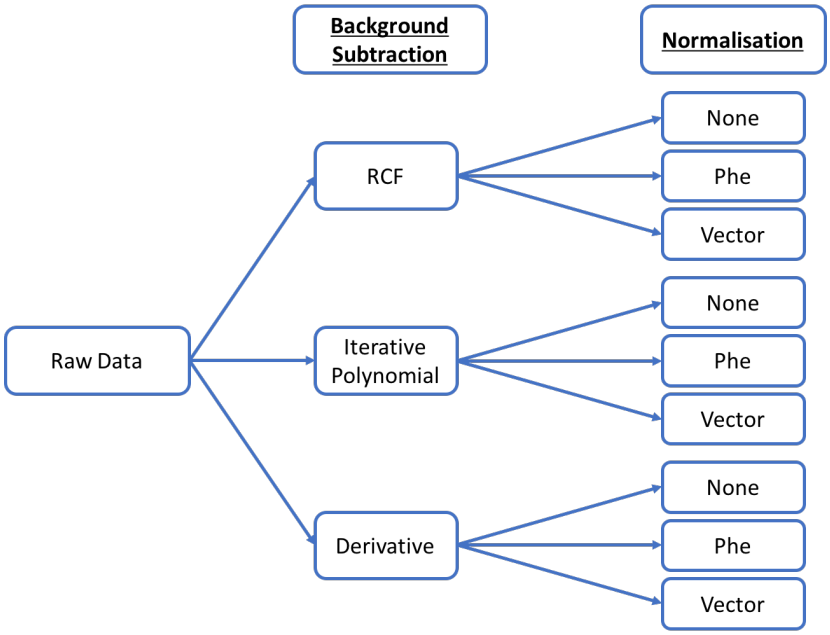


Figure 6.1: Summary of the different methods and combinations of background subtraction and normalisation for pre-processing spectral data where RCF is rolling circle filter, Phe is normalisation to the maximum of the Phenylalanine peak (1003 - 1004 cm^{-1}.)

were then used to create nine PLS-DA models. Each model was cross-validated using 5-fold k-fold cross validation and the optimum method was chosen from comparing the sensitivity and specificity outputs from the nine models.

As well as using the diagnostic outputs as a measure of performance, all spectra were inspected visually after each pre-processing method combination to ensure that no artefacts were being introduced into the spectra through the pre-processing steps. Where this was the case that combination or method was discounted for further use within the diagnostic models.

Investigating effects of different sampling modes and sex of the patient on diagnostic capability

To measure the effect of different sampling modalities on spectra, PLS-DA was used. The libPLS freeware software package was used as a basis for the code (link to material can be found in Appendix C). PLS-DA models were built using data from patients in the feasibility study. PLS-DA model parameters were optimised according to minimising the cross validation error within the models. All PLS-DA models were cross validated using k-fold cross validation with 5 folds. Sensitivities and specificities were then calculated from the resultant confusion matrix from each model. Receiver operating curves (ROC) were generated by considering the training and test results for the diagnostic accuracy for the control and cancer patient populations. If the results from the populations are plotted as a distribution based on the test result the distributions overlap. The ROC curve is plotted by moving the 'cut off' value for sensitivity and specificity of the test along this distribution plot. The curve is then a plotted function of the sensitivity of the test vs the false positive rate (1-specificity) for different cut-off points. The ROC curves were plotted for predicted values from the 5-fold k-fold cross validation and the area under the curve (AUC) plotted for both the training set and the cross validated models. Please see Chapter 4.3.3 for an in depth explanation of the interpretation of these methods. PLS-DA models during this chapter were not tested against a blind testing dataset as PLS-DA was only used as a method of presenting the feasibility study and optimising the datasets.

6.4 Results and discussion

6.4.1 Investigating feasibility of liquid and dry serum Raman spectroscopy for CRC detection.

To evaluate the feasibility of Raman spectroscopy as a potential triage tool for CRC a preliminary study was undertaken involving a 'gold standard' subset of 60 patients out of the 300 recruited during this work. The study used 30 patients with confirmed CRC (adenocarcinoma of varying stages) and 30 that were confirmed controls through negative finding colonoscopies. Furthermore, the control patients were not suffering from any other types of inflammatory diseases and all patients were fasted for at least 6 hours prior to sample collection. The study groups were age matched, they were also a mixture of male and female patients and were a mixture of smokers and non-smokers. Patient information is summarised in Table 7.2.

6.4.2 Pilot study for fresh dry serum Raman spectroscopy

To investigate the feasibility of using dried serum for CRC detection, serum Raman spectra were collected from fresh serum samples that had been dried on the day of collection from the patient. Serum droplets were pipetted onto the high throughput aluminium substrate in duplicate. Five spectra were then collected across the droplets. Figure 6.2 shows the mean and difference spectra for 785 nm data for cancer vs control patients. The difference spectra show that there are a few key regions showing differences between the two groups. Control spectra show higher levels of carotenoids, CH_2/CH_3 stretching (lipids), higher shoulder of phenylalanine (Phe) and higher peaks attributed to Amide I and III bands. The cancer spectra show higher levels in bands attributed to tyrosine (Tyr) as well as phospholipids and lipoproteins. A PLS-DA model was constructed from a 785 nm dataset taken using the dry protocol developed in Chapter 5.3.6 with fresh serum samples to investigate if the differences between the spectra could be used within a discriminatory model.

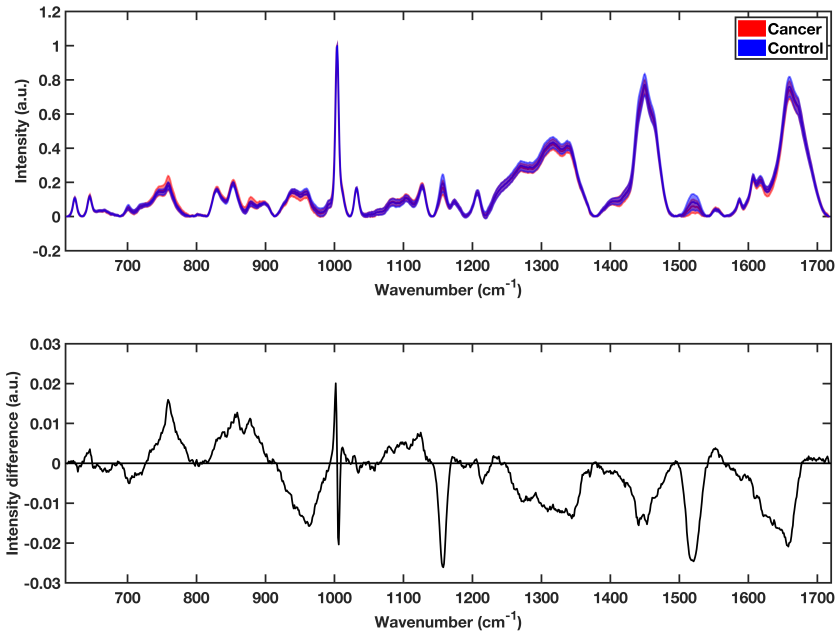
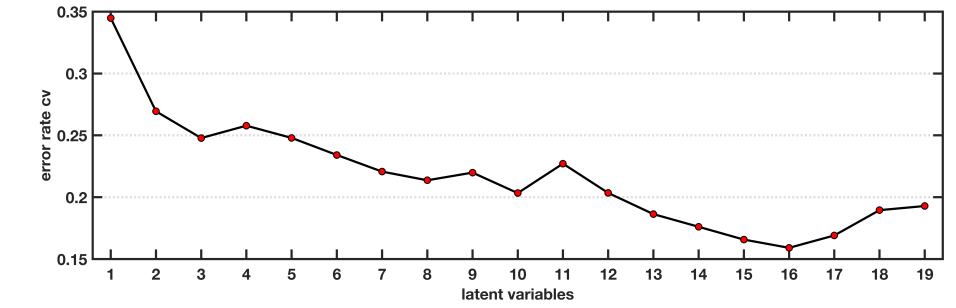
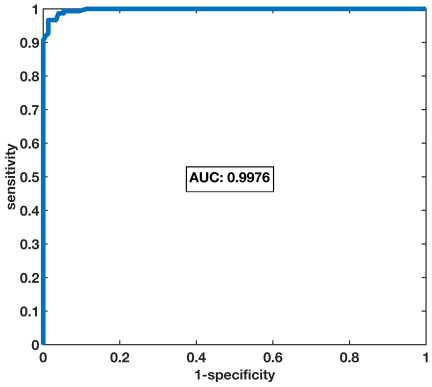


Figure 6.2: Mean and standard deviation 785 nm spectra of cancer vs control (upper) with difference spectra between cancer and control for data from the dry protocol (lower).

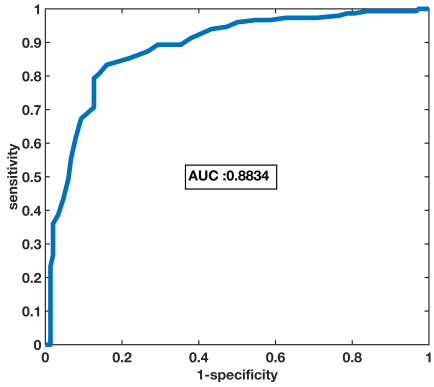
Clear outlying spectra were removed after pre-processing from the dataset prior to PLS-DA analysis (n=5 spectra). A PLS-DA model was then constructed with 30 cancer patients and 30 control patients on a spectrum-wise basis. The PLS-DA model was constructed with cancer as the 'positive result'. The model was first optimised to find the optimal number of latent variables. Figure 6.3(a) shows the cross validation (CV) error versus latent variables (LVs) for the PLS-DA model. The minimum error uses 16 LVs therefore 16 LVs were used within the model. The model was then built and cross validated using 5 fold k-fold cross validation. Figure 6.3 parts (b) and (c) show the ROC curves for the PLS-DA model training and CV respectively. Despite the area under the curves (AUC) dropping from 0.99 to 0.88, the AUC in both instances is above 0.75 which indicated a 'good' learner. Similar to PCA analysis, PLS scores and loadings can also be plotted to visualise the causes of discrimination within the dataset. Figure 6.4 (a) shows the calculated groups vs sample number for cancer vs control samples. Each circle or diamond represents a single spectrum. The cross validated



(a)



(b)



(c)

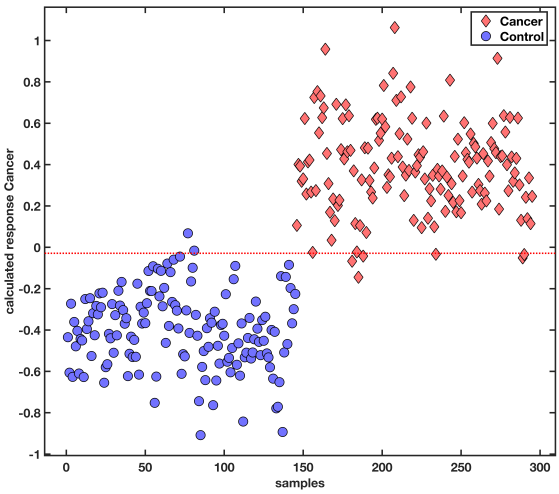
Figure 6.3: CV error minimisation plot for PLS-DA model (a), ROC curve for model training (b), and post CV (c)

response shows that the majority of the single spectra were identified correctly and allows us to see incorrectly identified spectra.

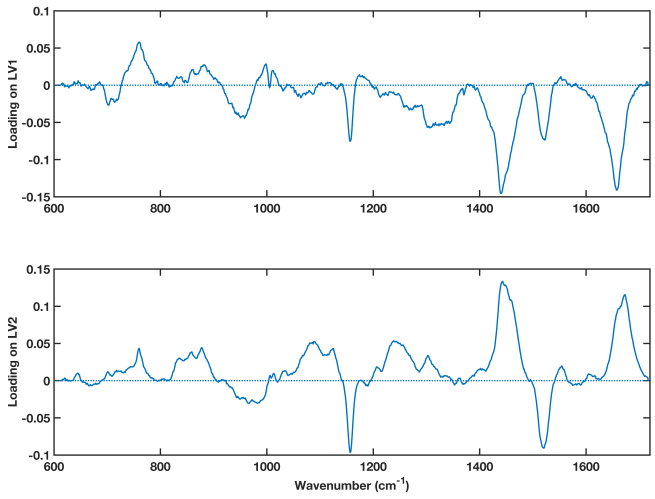
The CV model for the 785 nm dry fresh dataset yielded a sensitivity of 85% and and specificity of 81%. Comparing these values to the sensitivity and specificity of using symptoms alone (Chapter 1, Figure 1.10) for colorectal referrals the serum Raman method is better. The serum Raman result also compares well to other blood based colorectal diagnostic tests such as mSept9 with sensitivities ranging from 50-90% and specificities of 88-91% [2].

Figure 6.4 (b) shows the latent variables loadings associated with the discrimination. The PLS loadings on PLS-1 (LV1) show agreement with the different spectra in Figure 6.2. The control spectra show higher intensities in the carotenoid peak regions as well as differences in the CH2/CH3 stretching spectral region and the Amide I peak (denoted by negative loadings). The loadings on

PLS-1 (LV1) and PLS-2 (LV2) show that cancer patients in general have higher levels of Tyr, lipids and phospholipids. Following analysis and positive results on the dry protocol, this analysis was repeated for the high throughput liquid protocol.



(a)



(b)

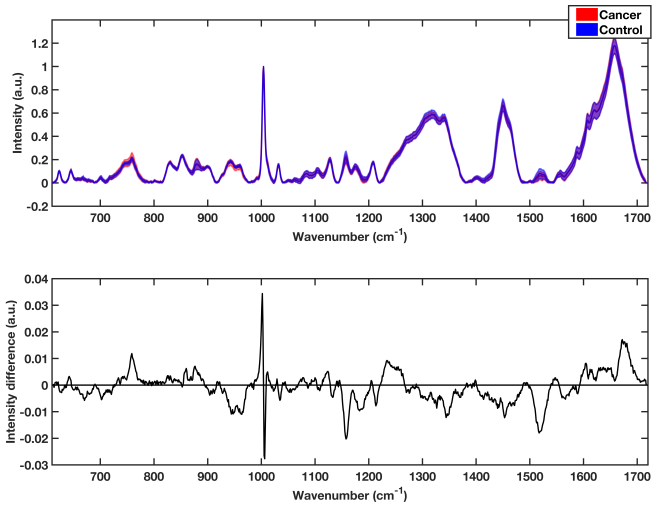
Figure 6.4: Calculated response for cancer vs control samples for dry fresh PLS-DA model (a), loadings on LV1 and LV2 (b).

6.4.3 Pilot study of fresh liquid serum Raman spectroscopy for CRC detection.

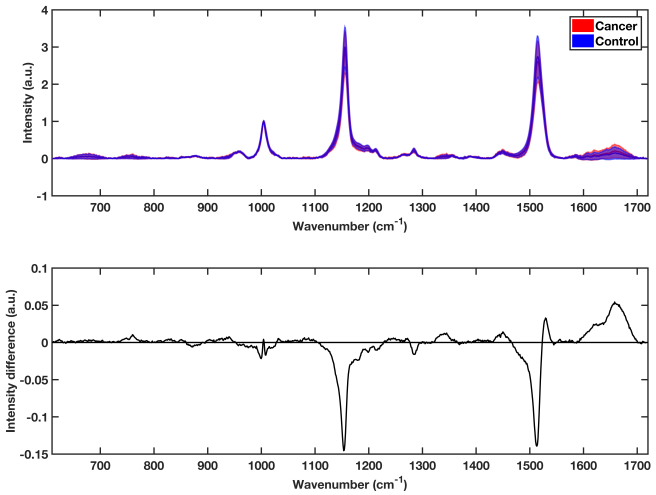
To investigate the feasibility of using the liquid serum protocol for CRC detection, Raman spectra were collected from fresh serum samples on the day of collection from the patient. As with the dry dataset the mean and standard deviation of each dataset were plotted with difference spectra to gain some intuition of the expected differences between the datasets.

Figure 6.5 shows the mean, standard deviation and difference spectra between 785 nm and 532 nm liquid data. The 785 nm difference spectra shows similarities in the dry and liquid examples. In general, the main peak magnitude differences follow the same trend as within the dry difference spectra. In contrast to the dry difference spectra, the magnitude of the peaks in the 1600-1720 cm^{-1} region are opposite to the dry spectra with higher peaks in the cancer spectra.

The 532 nm spectra supported the differences in the 785 nm spectra with higher peaks at 1157 cm^{-1} and 1513-1520 cm^{-1} in the control croup. The 532 nm difference spectrum also shows higher magnitude in cancer spectra in the 1600-1700 cm^{-1} spectral region.



(a)

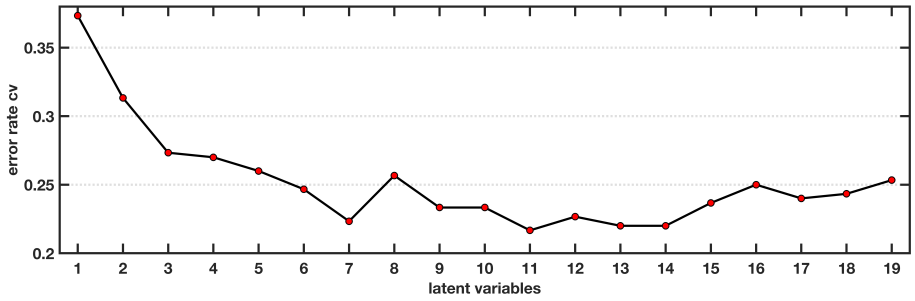


(b)

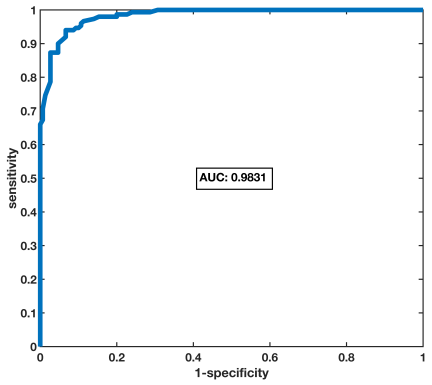
Figure 6.5: Mean, standard deviation and difference spectra for cancer vs control patients for fresh liquid 785 nm excitation (a) and 532 nm excitation (b).

Following univariate analysis PLS-DA models were calculated for samples excited with both 785 nm and 532 nm laser excitation. Figure 6.6 (a) shows the CV error minimisation for the fresh liquid 785 nm model. The minimum CV error was with a model based on 11 LVs. The PLS-DA model was then trained and cross validated using 11 LVs and 5 fold k-fold CV. Figure 6.6 (b-c) show the ROC curves for the training and CV model. The AUC for the trained model was 0.9831 with a sensitivity of 94% and specificity of 91.33%. After CV,

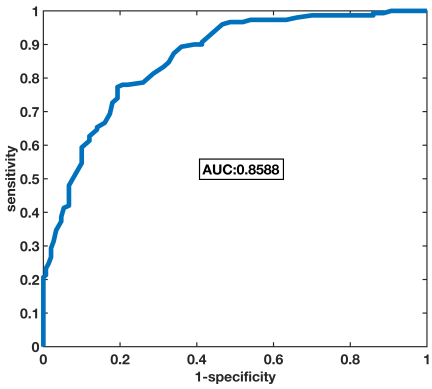
the model had an AUC of 0.8588, a sensitivity of 77.33% and a specificity of 80.67%. Despite having lower values than the dry protocol with 785 nm laser, the liquid methodology holds the advantage of being less susceptible to inter-user variability and has the potential for higher-throughput of samples with minimal sample destruction. Furthermore, the values are still improved compared to those of symptoms and age as a referral decision tool (Chapter 1.5).



(a)



(b)

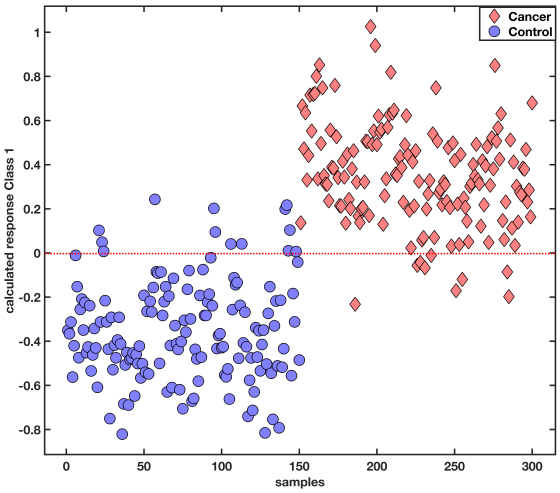


(c)

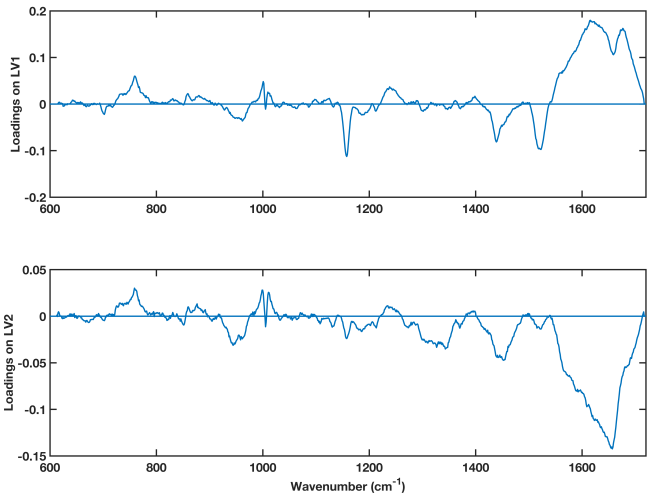
Figure 6.6: CV error minimisation plot for PLS-DA model (a), ROC curve for model training (b), and post CV (c).

The trained model vs samples plot for the fresh liquid model is shown in Figure 6.7 (a) with the loadings on latent variables in Figure 6.7 (b). The CV calculated response for the liquid mode shows more mis-classified spectra compared to the dry PLS-DA model. The loadings on the LVs show that again the spectral regions attributed to carotenoids and the Amide regions are significant for discrimination between cancer and control samples. The lower sensitivity and specificity of the dry and liquid cohort is possibly due to the differences in liquid and dry spectra

in the 1550-1720 cm^{-1} region of the spectrum.



(a)



(b)

Figure 6.7: Calculated response for cancer vs control samples for liquid fresh 785 nm PLS-DA model (a) and loadings on LV1 and LV2 (b).

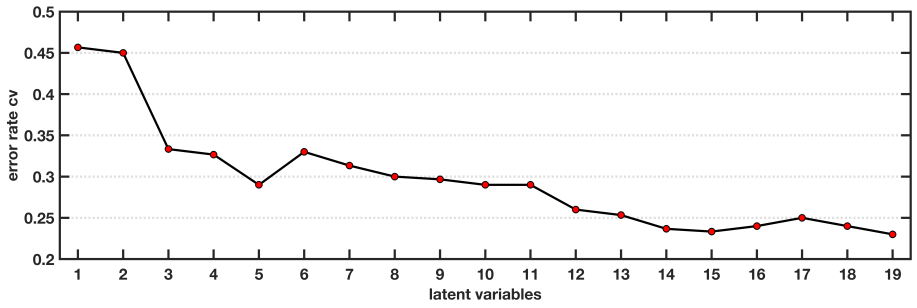
The liquid serum spectra show peaks at 1557 cm^{-1}, 1594 cm^{-1}, 1606 cm^{-1} and 1622 cm^{-1} are convoluted within the largest peak in that cluster of Amide I at 1659 cm^{-1}, this is due to the broad OH Raman active peak at 1650 cm^{-1} that is present in the liquid samples. Therefore the dry spectra have more defined peaks in that region compared to the liquid spectra as the OH contribution reduces as the sample dries. This trend translates to the spectral discrimination

and the loadings where the dry spectrum shows the region to be important for discrimination. The difference between the liquid and dry model performance could be due to the spectral differences within this region.

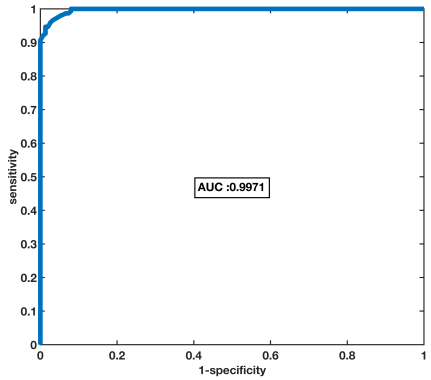
532 nm liquid PLS-DA analysis

A PLS-DA model was also produced for the liquid fresh spectra. Figure 6.8 shows the PLS-DA error min and the ROC curves. The error minimisation for the 532 nm model showed that a model with 15 LVs carried the lowest within model error. The calculated model for the 532 nm fresh model had an AUC for both training and CV of over 0.8. The sensitivity and specificity of the CV model were 76.67% and 78.00%, respectively. This is less than both of the 785 nm based models, however the 532 nm discrimination would still give an improved sensitivity and specificity compared to symptoms alone [3]. Furthermore, as introduced in Chapter 5 (5.3.16), the 532 nm spectra are more sensitive to patient demographics such as medication so the 532 nm spectra could be used in conjunction to the 785 nm model to give adjunct spectral information. This is explored further in Chapter 7 (7.3.8).

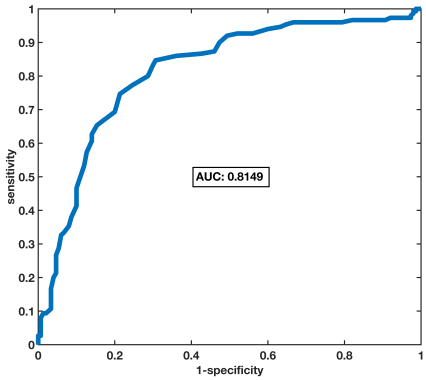
The loadings for the 532 nm model for LV1 (Figure 6.9) matches almost exactly to the 532 nm difference spectra. The loadings on LV2 show that there are some spectral shifts including the phenylalanine peak showing a higher shoulder on the peak in the cancer spectra which was not visible in the difference spectra. There are also peak shifts at 1157 cm^{-1} and 1520 cm^{-1} and a difference in the peak that shoulders the peak at 1157 cm^{-1} which contribute to the classification in the 532 nm spectra but are not present in the 785 nm spectra.



(a) CV error minimisation

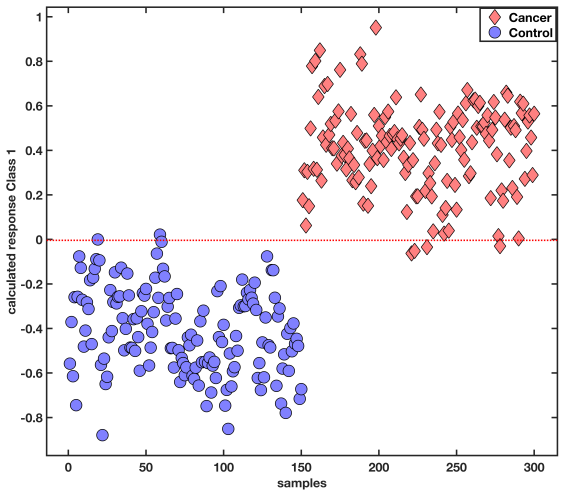


(b) Training ROC for fresh liquid 532 nm data.

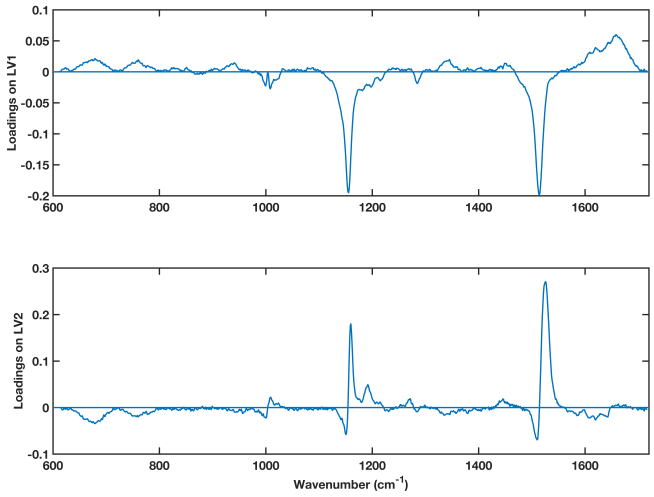


(c) Cross validated ROC for fresh liquid 532 nm data.

Figure 6.8: CV error minimisation plot for PLS-DA model (a), ROC curve for model training (b), and post CV (c).



(a) Samples vs calculated value and loadings



(b) Latent variable loadings 1-2

Figure 6.9: Calculated response for cancer vs control samples for liquid fresh 532 nm PLS-DA model (a) and loadings on LV1 and LV2 (b).

6.4.4 Optimisation of pre-processing methods for diagnostics

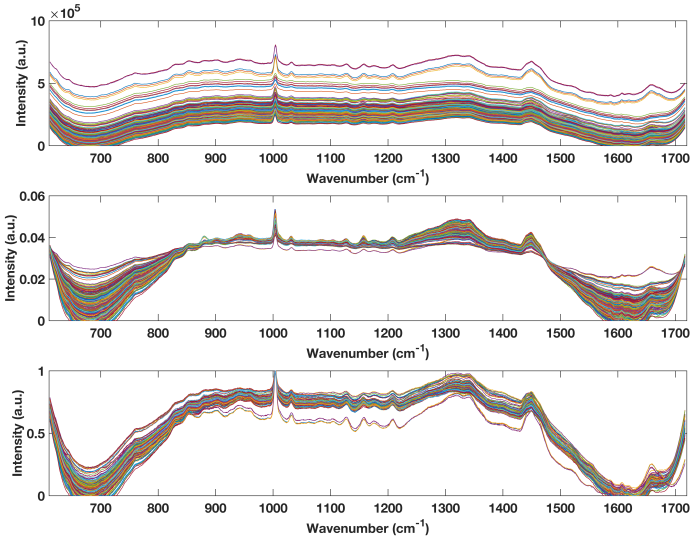
To ensure maximum diagnostic capability for serum Raman diagnostic tests, different combinations of pre-processing methods were investigated using cross validated PLS-DA results as a measure of performance. Spectral data from the pilot study participants were used. For spectral background subtraction the RCF, iterative polynomial and first derivative were used; these were combined with no normalisation, normalisation to the maximum of the Phe peak and vector normalisation. The combinations tested are summarised in Figure 6.1. The nine possible combinations were tested on liquid data (785 nm and 532 nm excitation) and dry data (785 nm excitation) and the results are as in Table 6.5.

Table 6.5: Comparison of sensitivities (Sens) and specificities (Spec) for different combinations of pre-processing methods for PLS-DA models.

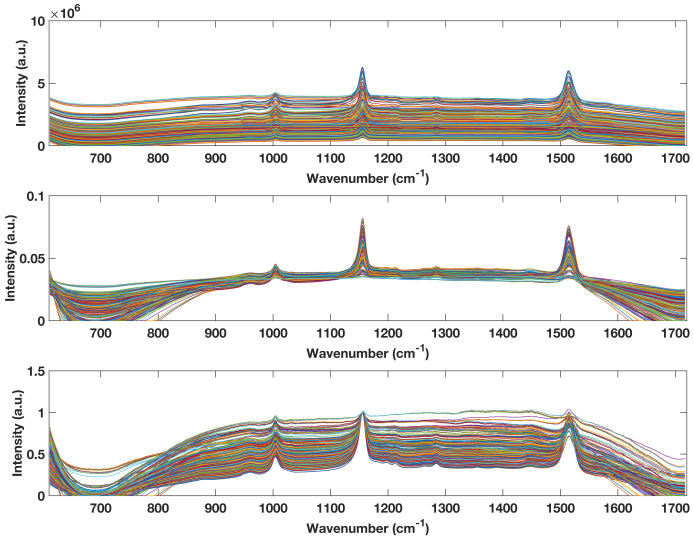
		785 nm dry		**785 nm Liq**		**532 nm liq**	
Background	**Normalise**	**Sens**	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**
RCF	Vector	0.83	0.83	0.78	0.81	0.77	0.77
	Phenylalanine	0.85	0.81	0.77	0.81	0.77	0.78
	None	0.81	0.83	0.77	0.81	0.84	0.72
Derivative	Vector	0.77	0.77	0.73	0.8	0.72	0.74
	Phenylalanine	0.77	0.77	0.72	0.78	0.75	0.73
	None	0.75	0.79	0.75	0.79	0.8	0.67
Polynomial	Vector	0.84	0.85	0.88	0.84	0.82	0.91
	Phenylalanine	0.83	0.86	0.88	0.87	0.83	0.84
	None	0.8	0.88	0.86	0.86	0.85	0.83

The results show that the polynomial background substitution method produced the highest sensitivities and specificities for all of the data. When visually inspecting the data during the processing of all three datasets it was found that the polynomial background subtraction seemed to produce unphysical effects in

the spectra such as negative or zero values for large portions of the spectra (Figure 6.10) which was not improved via normalisation.



(a)



(b)

Figure 6.10: Representative examples of polynomial baseline correction method (upper) combined with vector normalisation (middle) and normalisation to the phenylalanine peak (bottom) for 785 nm (a) and 532 nm (b) spectra.

Furthermore, the polynomial baseline code used requires large datasets as the polynomial baseline is estimated via an iterative approach. This performs well in

training large datasets, however for individual spectra being tested this approach would not fit the same baseline for the testing and training sets meaning that each time any spectra were to be tested the whole dataset would need to be trained together. This is not a practical translatable approach as it would add much more analysis time compared to just testing new spectra against a stored model.

The polynomial background subtraction for future processing of both 785 nm and 532 nm spectra was therefore disregarded for further analysis. The remaining combinations of preprocessing methods showed that the rolling circle filter outperformed the derivative spectra. Within the RCF results for all three datasets the vector normalisation showed on average very similar to results with the vector normalisation. Average sensitivity was 79% and specificity was 80% and the phenylalanine peak average sensitivities and specificities of 80% and 80%, respectively, therefore it was concluded that for future PLS-DA models the RCF background subtraction with phenylalanine normalisation method would be used for the purposes of comparing and optimising the diagnostic models. However, when constructing larger models in future both methods would be considered.

6.4.5 The effects of sample modality and patient demographics on diagnostic models

Preliminary work shown in the previous chapter (5.3.12) showed that different sampling methods such as having fresh or freeze-thawed samples can have different effects on the variance of spectra within a dataset. The following section will discuss how this type of variance can affect the capability of a diagnostic model. It will therefore compare different sampling methods to determine the optimum for diagnostic capability. The effects of sex of the patient on diagnostic models will also be discussed.

A comparison between the 785 nm and 532 nm data was conducted for dry vs liquid and fresh vs frozen samples to find the ideal sampling method. The samples used in this study were the same as that in the pilot study therefore the cohort details remain the same. Table 6.6 shows the results from a comparison of

cross validated PLS-DA models constructed for both 785 nm and 532 nm datasets with fresh serum data compared against data from samples that had undergone one freeze-thaw cycle (FT) after at least one month of storage (-80°C).

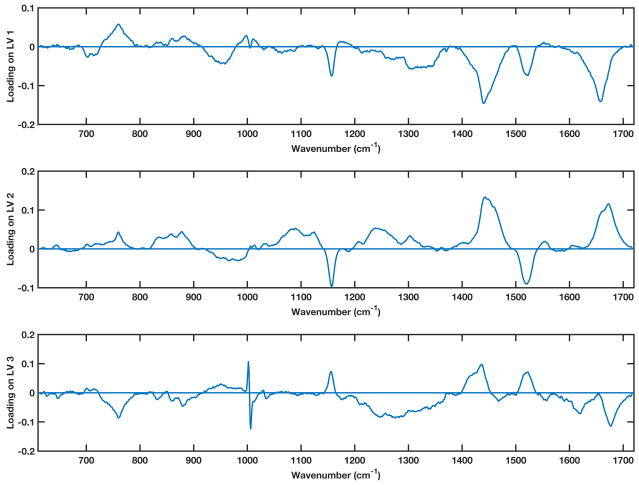
Table 6.6: Comparison of the sensitivity and specificity values for different sampling methods and different laser excitations.

	785 nm dry		**785 nm Liq**		**532 nm liq**	
Sample	**Sens**	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**
Fresh	0.83	0.83	0.77	0.81	0.77	0.78
FT = 1	0.72	0.78	0.79	0.77	0.77	0.77

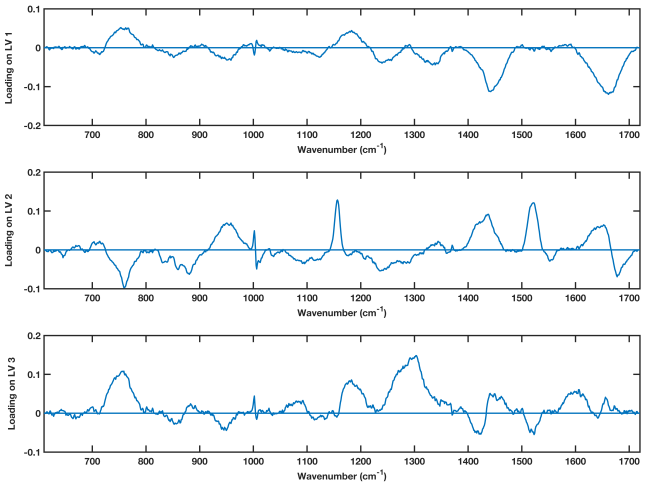
It is clear that fresh dry samples yield the highest sensitivities and specificities and is therefore the optimum method within the laboratory for discriminating between cancer and control serum samples. However, when samples had been through a freeze-thaw cycle before drying the sensitivity and specificity both dropped by at least 5%. This is possibly due to the freeze-thaw affecting the proteins within the sample. However, the dramatic drop in the sensitivity and specificity is not seen within the liquid datasets for either laser wavelength. So it is likely that the sampling method itself causes this. This agrees with the preliminary work showing that there are some spectral differences as samples undergo freeze-thaw cycles. To try and understand better what differences were affecting diagnostic capability between the different sampling methods, the loadings on the latent variables for the PLS-DA models for each method were investigated.

Investigating PLS-DA loadings for different sampling modalities

Figure 6.11 shows a comparison between the latent variable loadings for the fresh vs the frozen dry PLS-DA models. The loadings show similarities in latent LV 1. However, within LV 2 there are some key differences.



(a) 785 nm fresh dry LV loadings



(b) 785 nm FT dry LV loadings

Figure 6.11: Comparison between PLS-DA latent variable loadings for a 785 nm dataset with the fresh LV loadings (a) and the loadings from freeze-thawed spectra (b).

The relationships between the loading magnitudes are opposite for the peaks at 1157 cm^{-1}, 1520 cm^{-1} as well as for the peaks in the region between 700-

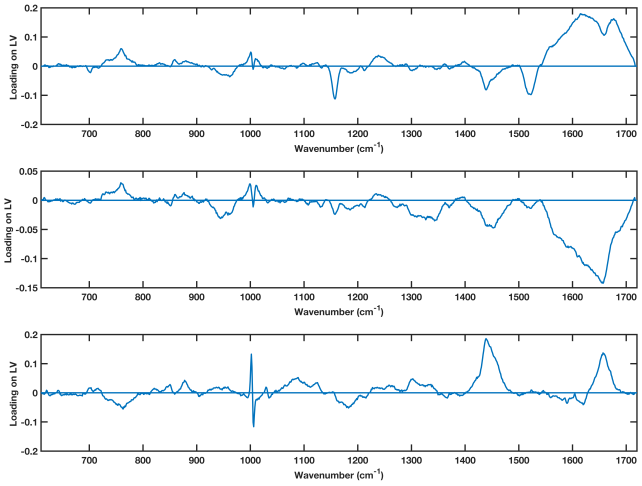
900 cm^{-1}. There is also a marked shift in the phenylalanine peak that appears in LV2 of the frozen dataset that isn't in the fresh loading as well as a shift in the Amide I peak at 1655 cm^{-1}. These differences can potentially be due to the FT process affecting the conformational changes to the proteins and their subsequent binding to metabolites as these are where the main spectral differences between the fresh and frozen dataset lie.

The differences within the loadings tie in with previous findings showing spectral variances in PCA plots between fresh samples and those that had been through a freeze-thaw cycle (Chapter 5.3.12). The drop in sensitivity and specificity of the frozen model can therefore be attributed to the overall variances in the dataset being higher within the frozen dataset and therefore the relationships between different peaks causes a large difference in the PLS-DA performance.

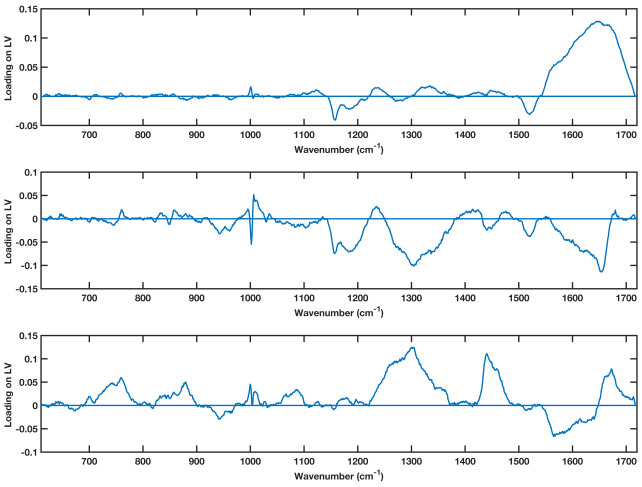
A comparison of the LV loadings from fresh and freeze-thawed samples using the liquid protocol with the 785 nm and 532 nm datasets was also conducted. The sensitivities and specificities of the fresh and freeze-thawed spectral models in the 785 nm dataset were very similar with the fresh samples having a slightly higher specificity than frozen samples.

Figure 6.12 shows that in the 785 nm dataset both the fresh and frozen loadings on LV1 from the PLS-DA models showed similarities across the spectral loadings apart from a loss of definition at the peak at 1619 cm^{-1} and 1675 cm^{-1}. The 1675 cm^{-1} peak is attributed to the amide bonds within proteins of the serum. The loss of definition could be due to a change in the protein conformation in relation to the water (OH) within the liquid samples through the freeze-thaw process. This is supported by the other spectral regions varying such as the protein regions, whereas constant peaks such as at 1004 cm^{-1} remain unaffected. The lack of variation in the principal LV loading agrees with the smaller variation between the fresh and FT samples seen in the previous chapter (5.3.12). Across the loadings on LV2 and LV3 the general trend also stays the same. Therefore despite a lower overall accuracy than the dry protocol the liquid protocol shows less variation between the fresh and FT spectra and also has faster acquisition time for repeat measurements with the potential to be optimised further. This suggests that in a situation where samples would need to be compared between

fresh and frozen or if serum samples needed to be frozen for storage logistics then the liquid protocol would be more suitable as a tool for detecting CRC using serum samples.



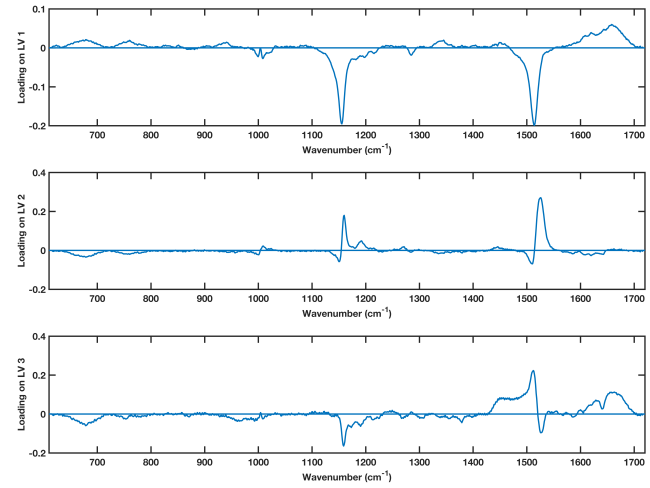
(a)



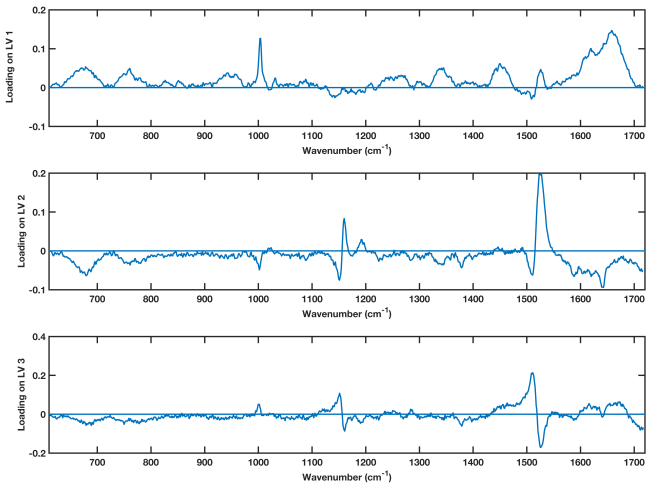
(b)

Figure 6.12: PLS-DA loadings comparison between fresh and FT liquid samples, with 785 nm fresh liquid LV loadings (a) and 785 nm FT liquid LV loadings (b).

The loading analysis was also repeated for the 532 nm dataset (Figure 6.13), there was very little difference in sensitivity or specificity of the 532 nm dataset (Table 6.6).



(a) 532 nm fresh liquid spectra



(b) 532 nm FT liquid spectra

Figure 6.13: Comparison between 532 nm loadings on LV1-3 for PLS-DA models constructed from fresh and FT datasets with the fresh loadings on LV1-3 (a) and the FT loadings (b).

The loadings show that as for patients who had undergone CRT treatments (Chapter 5.3.14) there is a loss of the characteristic 532 nm peaks within the spectra from the FT samples. However, the loadings on LV2 are very similar and there are some contributions within LV1 of the fresh spectra. This is a potential area of further investigation as to why this occurs but the differences in loadings seem to counteract each other resulting in a similar diagnostic result.

To conclude, the dry fresh protocol produced the highest sensitivity and specificity for a 'gold standard' dataset and was the optimal process within the laboratory setting for detecting CRC using fresh serum from patients. However, samples analysed with the dry methodology that had undergone a FT cycle were susceptible to spectral variation and had much lower diagnostic capability and less inter-operator variation than the same samples analysed via the liquid protocol. The 785 nm liquid protocol showed higher diagnostic capability than 532 nm protocol and the most consistency within the contributions to spectra that were attributing to the diagnostic capability. Therefore, the optimal methodology to maximise performance is using fresh dried spectra whereas the most consistent method proved to be the 785 nm liquid protocol for either fresh or frozen serum samples.

In relation to a clinical setting, discussion within the local laboratory medicine team revealed that logistically transporting fresh serum samples from multiple sites to one hospital would not be possible. The dry fresh method would therefore not be practical for analysis within their laboratory or within that setting. However, the distribution of frozen samples was possible between at least three hospital sites within the health board (Abertawe Bro Morgannwg University Local Health Board - ABMU). The 785 nm liquid protocol with freeze-thawed samples was therefore decided to be the best route for application to a clinical setting.

Sex of patient

This chapter has shown that for a cohort of patients that are age-matched, but a mixture of both male and female and smokers and non-smokers, that RS has the ability to distinguish between serum samples from cancer and control patients. The preliminary work for this thesis discussed differences in patient demographic information and how this affects the variance of a Raman spectral dataset (Chapter 5, section 5.3.16). The sex of a patient showed some differences in spectra between male and female patients for dry 785 nm spectra such as different carotenoid levels that have been shown above and in literature to also contribute to the discrimination between cancer and control patients [4–6]. In a clinical application when a patient is being considered for referral to secondary

care the sex of the patient is known. Therefore it is feasible to consider that for a 785 nm methodology it would be possible to build a diagnostic for both male and female patients. The dataset used throughout this chapter included an almost equal number of male and female patients. Therefore, to test the ability of a split model the dataset was divided into male and female patients. PLS-DA models were constructed on male only and female only datasets as well as a smaller combined dataset allowing for a direct comparison without number bias issues. Table 6.7 shows the comparison between the combined dataset vs individual PLS-DA models for a freeze-thawed liquid 785 nm dataset. Splitting the dataset shows that there is some improvement versus the combined dataset for male patients with the sensitivity rising from 83% to 89%. However, despite a small rise in the sensitivity for a female-only model the specificity decreased which is the factor that would need to be maximised for translation.

The differences within the model performance could be due to a number of biological factors such as female patients experiencing menopause at different times [7]. From an analytical point of view there were also only small numbers of patients within this study with slightly higher number of females. This may be something to re-consider in future when larger numbers of patients have been recruited. Lower sensitivity in a female only model compared to a combined model would not be acceptable. So despite the increase in sensitivity and specificity for samples from males it was decided to keep the data from the serum Raman CRC model data combined until greater numbers of patients for both sexes could be collected to provide a larger training group for a diagnostic model.

Table 6.7: A comparison between liquid models from data from male and females participants.

	785 nm Liq FT=1	
Sex	**Sens**	**Spec**
Male	0.89	0.89
Female	0.84	0.76
Combined	0.83	0.83

6.5 Conclusions and future work

This chapter has shown that for a cohort of 60 patients, 30 cancers vs 30 controls - it is possible to distinguish between CRC and control serum samples with sensitivity and specificity towards the current available tests to diagnose CRC (detailed table comparing current tests in development can be seen in Chapter 1, Figure 1.12) [8].

The optimal data pre-processing routine was found as a combination of the RCF background subtraction with either a vector normalisation or normalisation to the phenylalanine peak within the spectra.

The optimal pre-processing data method was then applied to datasets from different sampling modes to find that dry fresh 785 nm serum dataset had the highest diagnostic performance within the laboratory setting. However, on considering translation to a clinical setting the 785 nm liquid methodology was actually the best performing for clinical application due to the requirement of clinical samples being frozen and thawed before transportation to the laboratory for analysis.

It was also found that a patient's sex will affect diagnostic capability in a model with small sample numbers. It was decided that to properly assess the application of two separate models without losing sensitivity or specificity within one of the sexes, that a further study conducted on a larger cohort of patients spanning all possible inclusion criteria for the study e.g. early and late stage CRC, controls and inflammatory non-malignant diseases is needed. It is also noted that the difference spectra between fasted and non fasted patients and the loadings from the PCA analysis in Chapter 5.3.15 are similar to regions of the spectra that loadings within the liquid 785 nm PLS-DA loadings that are important for discrimination between cancer and control samples. Therefore selecting fasted patients for cancer vs control studies is essential.

Although PLS-DA based analysis is a useful tool for optimising different parameters within a discriminatory model, the analysis within this chapter was conducted on a 'gold standard' dataset with matched cancer/controls,also matched for sex ad age of the patient. One disadvantage of PLS-DA is that if training groups are unequal the models can be subject bias and prediction error can in-

crease [9]. Furthermore, by choosing the number of LVs to minimise the error it is noted that models requiring large numbers of LVs may use LVs for the model that are not showing true spectral features. This leads to a possibility of including regression lines that potentially find trends within spectral noise rather than true features.

Finally, it has also been shown previously that for a non-binary model PLS-DA algorithms can struggle to discriminate between groups of patients. Whilst the loading plots are good for showing overall peaks used for discrimination within the PLS-DA models they don't provide a rank order of the most important peaks. This makes further investigation into assigning the critical biomolecules for CRC detection more difficult.

Therefore, PLS-DA models were good for investigating smaller binary models and are useful in optimising the pre-processing because it is not an ensemble method so produce comparable results. The PLS-DA method is also good for smaller groups as ensemble methods such as Random Forest are not recommended for sample sets where the number of observables i.e. wavenumbers is greater than the number of samples. However, when moving to larger datasets with potentially non-binary groups the literature suggests that ensemble methods such as Random Forest may be better suited algorithm for the spectral data. Random Forest could potentially protect the models against over fitting and are also able to rank and isolate the key spectral features for discrimination.

Bibliography

[1] Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, Michael J Walsh, Martin R McAinsh, Nicholas Stone, and Francis L Martin. Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.*, 11(4):664–687, 2016.

[2] Hye Seung Lee, Sang Mee Hwang, Taek Soo Kim, Duck-Woo Kim, Do Joong Park, Sung-Bum Kang, Hyung-Ho Kim, and Kyoung Un Park. Circulating methylated septin 9 nucleic Acid in the plasma of patients with gastrointestinal cancer in the stomach and colon. *Transl. Oncol.*, 6(3):290–6, 2013.

[3] BH Willis. The diagnostic value of symptoms for colorectal cancer in primary care. *Br. J. Gen. Pract.*, (May):231–243, 2011.

[4] Dinesh K. R. Medipally, Adrian Maguire, Jane Bryant, John Armstrong, Mary Dunne, Marie Finn, Fiona M. Lyng, and Aidan D. Meade. Development of a high throughput (HT) Raman spectroscopy method for rapid screening of liquid blood plasma from prostate cancer patients. *Analyst*, 142(8):1216–1226, 2017.

[5] T H E Analyst, Aditi Sahu, Tata Memorial, Hitesh Mamgain, Witec Wissenschaftliche, Murali Krishna, and Chilakapati Advanced. Raman spectroscopy of serum : An exploratory study for detection of oral cancers. (January 2016), 2013.

[6] Andrew T. Harris, Anxhela Lungari, Christopher J. Needham, Stephen L. Smith, Michael A. Lones, Sheila E. Fisher, Xuebin B. Yang, Nicola Cooper, Jennifer Kirkham, D. Alastair Smith, Dominic P. Martin-Hirsch, and Alec S. High. Potential for Raman spectroscopy to provide cancer screening using a peripheral blood sample. *Head Neck Oncol.*, 1:34, 2009.

[7] Kirsi Auro, Anni Joensuu, Krista Fischer, Johannes Kettunen, Perttu Salo, Hannele Mattsson, Marjo Niironen, Jaakko Kaprio, Johan G. Eriksson, Terho

Lehtimäki, Olli Raitakari, Antti Jula, Aila Tiitinen, Matti Jauhiainen, Pasi Soininen, Antti J. Kangas, Mika Kähönen, Aki S. Havulinna, Mika Ala-Korpela, Veikko Salomaa, Andres Metspalu, and Markus Perola. A metabolic view on menopause and ageing. *Nat. Commun.*, 5:1–11, 2014.

[8] Frank Rinaldi. Global market assessment for Raman spectroscopy and colorectal cancer., 2017.

[9] Richard G. Brereton and Gavin R. Lloyd. Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.*, 28(4):213–225, 2014.

Chapter 7

Establishing the limits of diagnostic capability for colorectal cancer with serum Raman spectroscopy

7.1 Introduction

The previous chapters outlined the development of robust protocols required for clinical translation of serum Raman spectroscopy (RS) for the detection of CRC. The results from preliminary studies showed that a fresh dry protocol maximised diagnostic capability but is not readily translatable to a clinical setting when considering logistic effects. Therefore, this chapter focuses on the high-throughput liquid 785 nm platform for serum Raman analysis. If successful, the liquid methodology could be implemented as a triage tool for USC referral pathways that were discussed in Chapter 1 (Section 1.5). This chapter will use data from cancer patients, control patients, patients with inflammatory diseases and patient with other cancers from different sources to establish the diagnostic limits of the platforms developed throughout this thesis.

In addition to samples that were originally recruited for this study, an additional recruitment programme was started for testing the diagnostic accuracy of this work. The aim was to validate the pilot study findings and test GP and patient acceptance of the technique. Recruitment involved input from the local laboratory medicine team to Swansea University as a part of a NHS portfolio study (RAMAN-CRC). The portfolio study was an extension under the original study ethics (IRAS 146942, REC: 14/WA/0028). The patients from the RAMAN CRC study would be the envisaged target for a Raman based triage tool for colorectal USC pathway patients. Patients were recruited using the same inclusion

criteria as the original study however, patients were from a primary care population. The samples from this line of recruitment would be frozen before analysis due to the logistics of transporting blood samples from GP surgeries to the laboratory for analysis without degradation. The samples were spun, aliquotted and frozen via the same methodology as outlined in Chapter 3.1. However, the sample handling was done entirely by the automated laboratory medicine equipment rather than by hand as with the previous samples [1].

To maximise the flexibility of the diagnostic models within this chapter and allow maximum number of spectra to be included into diagnostic model training, diagnostic models in this chapter were trained on a mixture of data from fresh samples and those that had been through one freeze-thaw cycle. The diagnostic models trained during this chapter were then tested using samples recruited from the original source (secondary care clinician) and the frozen methodology from the larger scale recruitment.

7.1.1 Aims and objectives

This chapter aims to establish the limits of serum RS for use as a triage tool for patients who would normally be referred via the urgent CRC referral pathway. Random forest (RF) based analysis or partial least squares discriminant analysis (PLS-DA) has been utilised within the chapter to establish the following:

- What is the earliest stage that the liquid RS platform can detect i.e. is there a difference between control patients and those with precursor lesions (polyps)?
- What is the effect of a larger patient cohort on the ability of RS to detect CRC in liquid serum samples from cancer patients and control patients?
- What is the effect of introducing non-cancer patients with inflammatory conditions to the diagnostic capability of serum RS?
- Is it possible to enhance the ability of liquid serum RS using both 785 nm and 532 nm data combined.

[1] Please see Appendix for PCA analysis to show there is no separation between control patients from the two methods.

- Is it possible to tell the difference between colorectal cancer and other cancer types using the liquid serum RS platform?

The chapter will conclude with a discussion of the limits of the methodologies developed within this thesis. This will include the potential for translation that this technique could have.

7.2 Materials and methods

7.2.1 Cohort information

The following section describes the summary information for participants during this chapter. For the purposes of this chapter all patients labelled as control patients have confirmed negative colonoscopies and no other inflammatory conditions. Patients labelled as polyp patients had at least one colorectal polyp confirmed by colonoscopy and investigated by histology. Cancer patients have had cancer confirmed by their clinical care team. Patients referred to as inflammatory controls are patients who had colonoscopies negative for colorectal cancer but may have benign colorectal disease or other co-morbidities e.g. colitis.

Polyp vs control patient study

Table 7.1 is a summary of the patient metadata breakdown for training and testing spectra for the diagnostic model of polyp vs controls. In general, the number of male participants was greater than female participants in the polyp population, therefore this is reflected in the patients included in both the training and testing datasets. The average polyp size was calculated from the maximum size polyp in patients with multiple polyps. The majority of the patients with polyps were found to have tubular adenomas with low grade dysplasia. However, some were found to be hyper-plastic in post analysis of the training set. The breakdown of the individual patient polyp sizes for the testing and training sets can be found in Appendix E.1.

Table 7.1: Patient details for the binary polyp vs control study.

Study group	No patients	Mean Age (years)	Number smokers	Number male	Number female	Average polyp size (mm)
Polyp (train)	44	68.7±11.0	6	30	14	15±18.7
Control (train)	50	64.6±12.1	3	23	27	N/A
Polyp (test)	11	70.9±10.0	2	8	3	4.7±2.8
Control (test)	23	66±10.2	4	10	13	N/A
Totals	**128**	**67±11.23**	**15**	**61**	**37**	**13.2±17.3**

Large binary cancer vs control study

Table 7.2 is a summary of the patient demographics used for the binary cancer vs healthy control cohort model training and testing. As with the previous studies, cancer patients were confirmed by histology and control patients were confirmed to have normal colonoscopies. The RAMAN CRC recruited patients were also confirmed via colonoscopy. A larger number of control patients were included in the testing set than cancer patients to reflect a theoretical prevalence rate of roughly 1/3 of patients. Participants were age and gender matched as closely as possible. The number of smokers were calculated from the patients responses to questions, therefore there is the potential caveat to the numbers of smokers as to if patients have been truthful in their smoking declarations. This is true for all studies within this chapter.

Table 7.2: Patient details for RF binary patient study with cancer and healthy control patients

Study group	No patients	Mean Age (years)	Number smokers	Number male	Number female
Cancer (training)	52	67.3 ± 9.7	7	25	27
Control (training)	51	64.5 ± 11.4	5	24	27
Cancer (testing)	16	75.3±11.5	1	8	8
Control (testing)	33	66.0±10.0	5	16	17
Totals	**152**	**66.9±11.2**	**18**	**73**	**79**

Non-binary and inflammatory study

Two models were also constructed during the study to test the detection limits of the models and to investigate if having non-binary systems affected the diagnostic capabilities. A model for polyp, cancer and control patients and a model including non-malignant disease patients (for the purposes of this study these were labelled Inflam). To investigate the effect of non-malignant gastrointestinal diseases, data from the polyp-cancer-control dataset were combined with data from the control patient samples who were cancer-free but had other diseases such as diverticular disease, colitis etc. The full cohort details for the non-binary models can be found in Table 7.3.

Multi-modal approach

Preliminary work in Chapter 5.3 showed that there are differences in serum spectra excited with different wavelengths. The 785 nm models built throughout this

Table 7.3: Patient details for the non-binary model including healthy control, cancer, inflammatory, and polyp patients.

Study group	No patients	Mean Age (years)	Number smokers	Number male	Number female
Cancer (train)	35	68.8±10.1	6	15	20
Control (train)	38	64.6±12.2	2	18	20
Polyp (train)	44	69.1±10.5	5	29	15
Inflam (train)	47	70.1±12.0	4	20	27
Cancer (test)	16	75.3±11.5	1	8	8
Control (test)	47	66.7±9.3	8	24	23
Polyp (test)	11	70.9±10.0	2	8	3
Inflam (test)	10	66.7±11.8	1	7	3
Totals	**248**	**68.5±11.1**	**29**	**129**	**119**

work have consistently shown a higher diagnostic capability than models built on data from a 532 nm model.

PCA analysis showed the sensitivity of the 532 nm data to other comorbidities such as medication could be a factor in the diagnostic models. Therefore, in an attempt to improve the diagnostic capability of a binary model of cancer vs control patients a multi-modal model was constructed. Data from both laser sources were collected from a cohort of 145 participants for model training and testing. A summary of the details for the multi-modal patient cohort model are in Table 7.4.

Table 7.4: Patient details for the multi-laser excitation cancer vs control model

Study group	No patients	Mean Age (years)	Number smokers	Number male	Number female
Cancer (train)	50	67.1±11.0	7	24	26
Control (train)	48	64.5±11.0	4	23	25
Cancer (test)	16	71.6±12.8	1	8	8
Control (test)	31	66.0±10	5	15	16
Totals	**145**	**66.5±11.1**	**17**	**70**	**75**

Disease monitoring

To show the potential for serum RS for disease monitoring, samples were collected and spectra were taken from 2 patients throughout their CRT and surgical treatment. One patient was a 68 year old male with rectal cancer and one patient who was female, 72 years old and also had rectal cancer.

7.2.2 Raman microspectroscopy

The Renishaw InVia Raman spectrometer was used for all Raman data collection in this chapter. A liquid serum approach was used for building and testing the serum Raman CRC methodology in this chapter.

Liquid Spectral collection

The method for sample preparation for both 785 nm and 532 nm lasers was equivalent for liquid samples. Data from liquid samples were taken using the stainless steel well plate; 200 μl of serum was pipetted into a cleaned well before each measurement and the microscope focused to 1200 μm above the focus point at the base of the well. The well plate was kept between 18-20°C during measurements using the cooling platform developed in Chapter 5.3.7. Wells were cleaned via the method outlined in Chapter 3.3 between measurements. The measurement

parameters for liquid measurements for both the 785 nm and 532 nm lasers were as in Table 7.5. Samples were interrogated with approx. 165-185 mW of power from the 532 nm laser and 45-55 mW from the 785 nm laser. The total time for data collection was approximately 6 mins and 12.5 mins for the 532 nm and 785 nm laser sources including 3 repeat spectra.

Table 7.5: Optimised data acquisition conditions for 785 nm and 532 nm liquid excitation.

	785nm Laser	**532nm Laser**
	Liquid	Liquid
Wavenumber range (cm^{-1})	610-1720	610-1720
Grating (l/mm)	1200	2400
Exposure time (s)	5	0.6
Accumulations	30	120
Laser Power (%)	100	100
Pinhole Y/N	N	N

7.2.3 Data pre-processing

Raman spectral Pre-processing

Raw Raman spectral data were exported into the MATLAB GUI for spectral pre-processing. The data were pre-processed according to the optimised analysis routines developed in Chapter 6 (6.4.4). Briefly, both 785 nm and 532 nm data were imported into the GUI where they are automatically shifted and interpolated to a singular wavenumber axis. The 785 nm spectra were baseline corrected using a rolling circle filter (RCF) background subtraction with circles of radius 150, and 532 nm spectra were corrected with circles of radius 300. Spectra were then normalised to the phenylalanine peak at 1003-1004 cm^{-1}. Label vectors denoting

which class each spectra belonged to were also generated within the GUI ready for either use as training or testing sets for analysis.

7.2.4 Data analysis

Preprocessed data were used for spectral comparisons and for computing the mean and difference spectra to demonstrate spectral differences and make spectral comparisons. To highlight spectral variation due to different diseases, preprocessed spectra were subject to PCA analysis.

The number of patients recruited during this study across their CRT treatment pathway was small (n=2), therefore it was not possible to create RF or PLS-DA models. However, preliminary results are shown from average spectra from 2 patients at different stages; baseline, pre-op and post-op. Spectra were preprocessed within the GUI outlined in Chapter 4.4. The parameters for preprocessing equivalent to the preprocessing methods for the RF models above. Average spectra from the patients at each stage were then plotted to visually investigate the treatment-related changes within the spectra.

PLS-DA

To measure the potential of serum RS for other disease types PLS-DA was used due to the number of patients being smaller than needed for RF. PLS-DA models were built using data from patients in the feasibility study using the lib-PLS software package. PLS-DA model parameters were optimised according to minimising the cross validation error within the models. All PLS-DA models were cross validated using k-fold cross validation with 5 folds. Sensitivities and specificities were then calculated from the resultant confusion matrix from each model. Receiver operating curves (ROC) were also generated from the predicted values from the 5 fold cross validation and the area under the curve (AUC) plotted for both the training set and the cross validated models.

Random Forest

The fitensemble function within the MATLAB classification toolbox was used to fit a RF classifier to data (example of an RF model code can be seen in Appendix C). All RF models were trained on a spectrum-wise basis on a combination of data from fresh and freeze-thawed samples to maximise the number of training spectra. All random forest classifier models were trained with 499 trees to minimise the out of bag (OOB) error rate (discussed in detail in Chapter 4.3.2). Trees within the RF model were left 'unpruned' therefore the maximum splitting of the branches was number of spectra - 1, i.e. for a model with 100 training spectra the max splits were set to 99.

Each training dataset used was run through the fitensemble function 10 times and the model with the highest cross validated error was selected as the final classifier model. A 'predict' or fit function was included within the code so a testing set could be evaluated easily. The test datasets were tested on a spectrum-wise basis. To evaluate sensitivity and specificity for each patient, three spectra[2] from each patient were tested, the majority vote for each each patient was then taken as the final diagnosis (Table 7.6).

Table 7.6: Example model result decision key. The 'positive' class for models, i.e. cancer or polyp is 1 and a negative result is 2.

Raw result	Overall decision
0 0 0	Control
0 0 1	Control
1 1 0	Cancer
1 1 1	Cancer

[2]Three spectra were used for testing instead of five in this case. This was due to the instrument time for the larger cohort of patients. Also to maximise resources and instrument time.

Feature selection

Within the training models for each RF model, a predictor importance was also calculated using the 'predictorImportance' function within MATLAB for the overall classification ensemble for each model. This allowed the selection of the wavenumbers (predictors) that were most important for correctly classifying a spectrum as cancer if the spectrum was from a cancer patient. The results of the feature selection function were plotted in the form of a predictor importance bar chart across each wavenumber. The top 50 wavenumbers were also taken from the predictor importance to create tables with the most important regions for correct group classification.

7.3 Results and discussion

Currently, the gold standard test for patients referred under the USC pathway in the UK is colonoscopy [1]. Colonoscopy currently has a sensitivity and specificity for detecting colorectal cancer in the range of 95% and 90% [2]. As well as sensitivities and specificities, negative predictive values (NPV) and the positive predictive values (PPV) for diagnostic tests are also often used to gauge diagnostic efficiency. For colorectal cancer, colonoscopy currently has a NPV of 99.4% and a PPV of 2-11% depending on which symptoms are used to refer the patient for colonoscopy. The NPV and PPV for a given test are defined as;

$$PPV = \frac{TP}{TP + FP}, \tag{7.3.1}$$

$$NPV = \frac{TN}{TN + FN}, \tag{7.3.2}$$

where TP is true positives, FP are false positives, TN are true negatives and FN are false negatives. Furthermore, the True positive rate (TPR) and the true negative rate (TNR) which are also commonly referred to are defined as

$$TPR = sensitivity, \tag{7.3.3}$$

$$TNR = 1 - specificity = 1 - FP. \tag{7.3.4}$$

The detection limit for precursor lesions (polyps) using colonoscopies is dependant on polyp size. Small adenomas (<10 mm) in size have a significantly higher miss rate compared to larger adenomas (>10 mm) [3]. Despite improved detection for polyps of greater than 10 mm in size, endoscopists have been found to miss up to 6% of adenomas larger than 10 mm in size and 30% of all adenomas [4].

The following results aim to compare the performance of the liquid serum Raman platform to other currently available diagnostic tests to evaluate the limits of the technique as a translatable tool to triage colorectal referrals. The effects on diagnostic accuracy including precursor lesions (polyps) and inflammatory

conditions into the RF model training and testing sets are then evaluated. Finally, preliminary studies show the potential for RS to detect polyps and also the potential of RS for application to disease monitoring and other cancer types.

7.3.1 Investigating precursor lesions and the limit of sensitivity

Establishing a limit of detection is important for a diagnostic tool. The precursor lesions to colorectal adenocarcinoma are colorectal polyps. To investigate spectral differences between polyp and control spectra, mean, standard deviation and difference spectra were produced for the polyp and control participant cohorts. Figure 7.1 shows the mean, standard deviation and difference spectra. The differ-

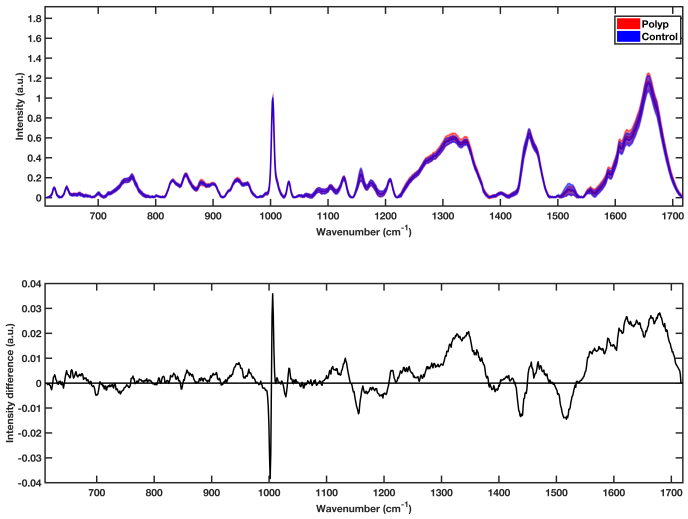


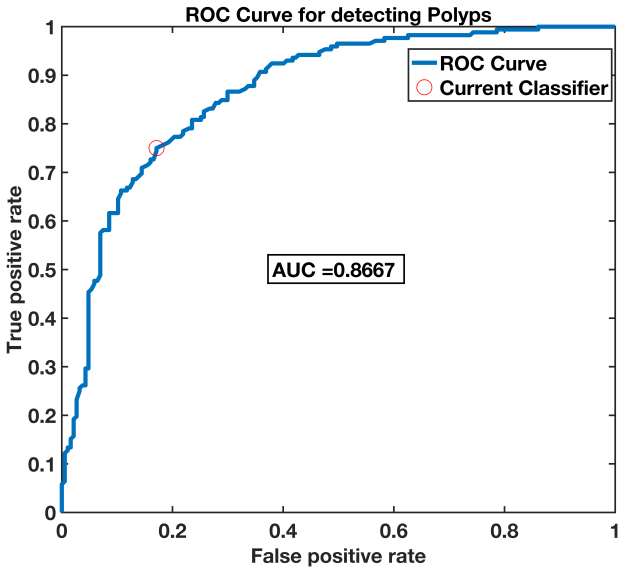
Figure 7.1: Mean and standard deviation for polyp vs control spectra for 785 nm excitation (upper) and difference spectra for the polyp vs control spectra (lower).

ence spectra shows a large difference in the phenylalanine peak between control and polyp spectra. There are also differences with polyps showing higher spectral peaks in the Amide III region and controls showing higher peaks attributed to carotenoids. Differences in carotenoid spectral regions of cancer patients vs control has been reported previously [5], however the differences in the precursor lesions and controls had not been studied in the study. These differences shown between polyp and control spetra are similar to those seen between cancer and

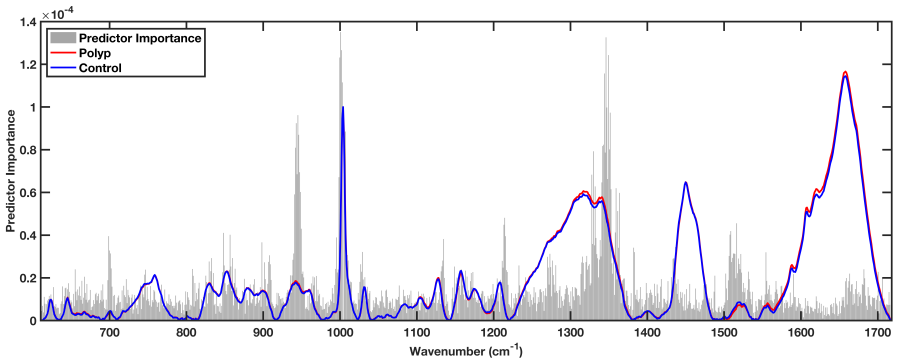
control patients in Chapter 6.4. Moreover, there is less of a difference in the 700-800 cm^{-1} region than seen in the previous analysis of differences spectra between cancer and control patients. The 700-800 cm^{-1} spectral region is attributed mainly to cholesterol, amino acids (AA) and nucleic acid (NA) differences. Passarelli et al have discussed previously that the serum lipid levels and cholesterol levels are elevated from healthy control patients in patients with colorectal polyps in GCMS studies [6]. It might be expected that during the progression from polyp to cancer there would be more cell proliferation in the cells within the bowel and therefore one would expect higher levels of circulating NA and AA to be found in the serum.

To establish the effect these differences have on a diagnostic model, a RF model for polyp vs control spectra was constructed with a total of 359 spectra. The model was trained with 172 spectra from 44 patient samples with colorectal polyp/s and 187 control spectra from 38 patients with negative colonoscopies (for breakdown see Appendix E.1). The RF model was calculated and cross validated via k fold cross validation. The ROC curve for the cross validated (CV) diagnostic model was plotted with the position of the RF classifier marked as seen in Figure 7.2 (a). The AUC for the ROC was calculated to be 0.8667 which classifies the model as a 'good' learner, however there is a trade off in the position of the classifier. For example, if the sensitivity was selected for the model in Figure 7.2 (a) to be higher than 90% the position of the classifier would move along the curve. Therefore, the false positive rate based on this model would be more than 30%. The sensitivity of the test being lower also lowers the false positive rate.

The RF classifier here was chosen to maximise the true positive rate whist keeping the false positive rate below 20%. The calculated sensitivity and specificity per spectra for polyps in the CV training model was found to be 74.42% and 82.89% respectively.



(a)



(b)

Figure 7.2: ROC and classifier position for polyp vs control samples (a) and Gini importance for the overall predictor importance for the large RF binary model (b).

The RF model predictor importance values were also calculated for the RF polyp detection model and plotted against the mean spectra for polyps and controls and the top 50 wavenumber regions most important for discrimination were calculated from the RF importance (Table 7.7).

For better intuition of the importance plot, the top 50 most important wavenumbers were ranked from the importance output and grouped into their spectral regions as seen in Table 7.7. The spectral regions found to be most important to correctly classifying between polyp and control samples was around the phenylalanine peak centred around 1004 cm^{-1} and amide III from 1340-1360 cm^{-1}.

There was an overall dominance of bands associated to protein differences within the importance table. This, along with the phenylalanine differences correspond to the difference regions in Figure 7.1.

Table 7.7: Spectral band assignments for regions within the top 50 most important wavenumbers from the RF predictor importance, [7].

Band position (cm^{-1})	**Molecular assignment**	**Biological component**
998-1007	Aromatic breathing	Phenylalanine
1340-1360	(C=O), (C-N),$N-H$	Amide III
940-950	not assigned	Proteins
1327-1333	(C=O), (C-N),$N-H$	Amide III
1214-1215	β sheet	Amide III
1507-1516	$(C{=}C)_n$	Carotenoids
848-856	$\nu_s(CH_2)$	Lipids
698	ν(C-S)	Lipids

The diagnostic ability of the trained model was tested using an independent test set of 102 spectra from 34 patients who were a mixture of control and patients with colorectal polyps at a ratio of 2:1. The testing spectra were predicted on a spectrum wise basis. The result per patient was then calculated using the decision criteria in Table 7.6. The result per patient showed a sensitivity of 90.91% but a specificity of 34.78%. So despite the technique missing only one patient with a polyp there were a large number of false positive results. This is reflected in the NPV and PPV values of 88.88% and 40% respectively. The average size of the polyps detected was 4.6±2.8 mm. This is smaller than is currently possible with methods such as CT colonography with a limit of 10 mm [8]. This however comes at the cost of a high number of false positive results.

Furthermore, on inspection of histology results post analysis the miss-classified polyp patient was found to have a terminal ileal polyp. This was the only polyp of its kind in the study and therefore may have been misclassified due to the lack of model training on this type of polyp.

Nevertheless, the initial RF results show good promise for clinical translation as a triage tool for colorectal referral for colorectal polyps in patients with sus-

Table 7.8: Patient-wise confusion matrix for polyp spectra vs control spectra.

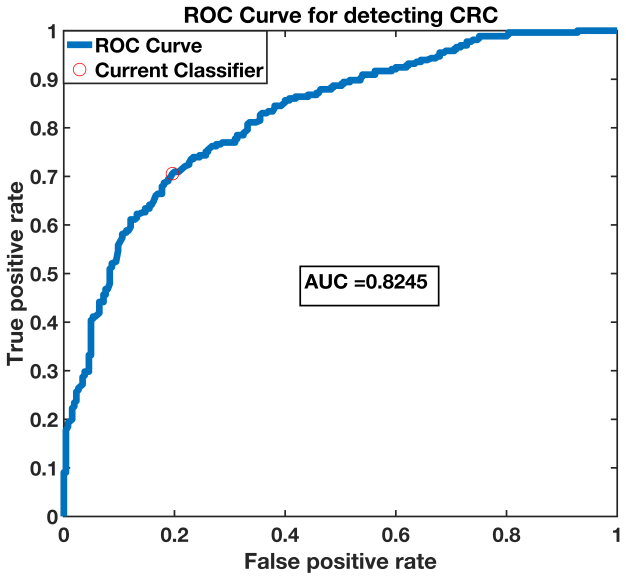
		Actual			
	n=34	**polyp**	**control**		
Predicted	**polyp**	10	15	**NPV**	**88.89%**
	control	1	8	**PPV**	**40.00%**
		Sensitivity	**Specificity**		
		90.91%	**34.78%**		

pected cancer due to its ability to detect polyps at a lower size than currently possible with CT colonography techniques. It also shows potential as a screening tool for polyp patients as it had a high NPV of 88.8% for polyp vs control patients.

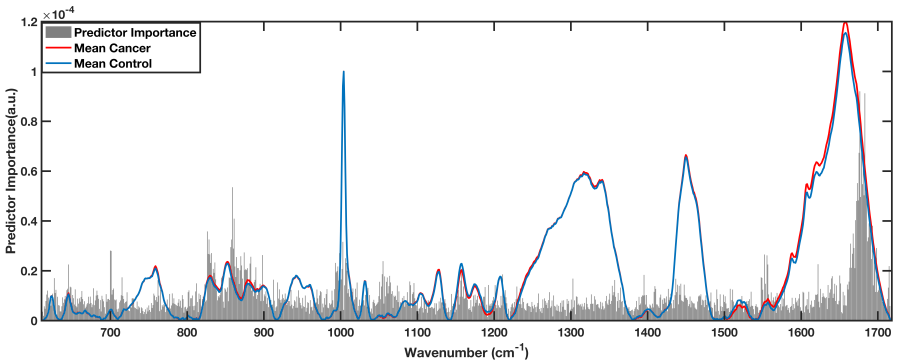
7.3.2 Application of liquid serum Raman to a binary cancer vs control cohort

For an evaluation of cancer vs control, a cohort of 103 patients was used to train a diagnostic model. The idea of this cohort being that it would show the maximum diagnostic capability of the 'ideal' cohort to establish the maximum sensitivity and specificity.

To enable the translation of liquid serum RS as a triage tool for CRC the diagnostic model must be evaluated using samples from the desired patients. A RF model based on serum excited with the 785 nm liquid data methodology was trained on 530 spectra from 52 cancer and 51 control patients. The best performing training model after 5-fold cross-validation was chosen from 10 repeat model training processes. Figure 7.3 (a) shows the calculated ROC curve for the final cross-validated trained dataset along with the position along the ROC for the diagnostic. It shows that the model has an AUC = 0.8245.



(a)



(b)

Figure 7.3: ROC for larger RF diagnostic model (a) and Gini importance for the overall predictor importance for the large RF binary model (b).

Figure 7.3 (b) shows the predictor importance for the each predictor (wavenumber) in the model. This is a measure of which wavenumbers are most important for correct classification. The top 50 most important wavenumber regions were taken from the importance plot and constructed into Table 7.9.

Table 7.9: Spectral band assignments for regions within the top 50 most important wavenumbers from the RF predictor importance, [7].

Band position (cm^{-1})	**Molecular assignment**	**Biological component**
1667-1678	(C=O), (C-N),NH_2	Amide I
1680-1695	(C=O), (C-N),NH_2	Amide I
856-874	$\nu(COO^-)$	Amino acids
998-1006	Aromatic ring breathing	Phenylalanine
826-830	ν(C=O)	Protein (Amide I)
1664-1668	ν(C=O)	Fatty acids
883-898	$\nu_s(CH_2)$	Lipids
1552-1556	$\nu_s(CH_3)$	Lipids
1055	not assigned	not assigned

The Amide I spectral region between 1667-1678 cm^{-1} are most important for correct diagnosis in the cancer vs control. This indicates changes in the protein structures within the serum between cancer and control patients. Another important region for diagnosis is assigned tentatively to phenylalanine. Finally there are also regions showing that Tyr levels and fatty acid chain differences between the spectra are also responsible for diagnostic capability. Watanabe et al reported significant differences in levels of Tyr between healthy controls and hepatocellular carcinoma patients with some differences in Phe levels [9]. The study also showed that there were differences in the levels of Phe and Tyr in serum profiles of CRC patients versus control patients.

The CV confusion matrix for the diagnostic training model (Table 7.10) yields a model with 74% sensitivity and 77% specificity on a spectrum-wise basis. The NPV value and PPV values were calculated for the training set, however, the NPV and PPV values are also dependent on prevalence of the disease in a given population. To gain more realistic values for the NPV and PPV for the serum liquid RS method, data from a 49 patient cohort were used to test the binary RF model on a spectrum-wise basis. The patients were from a GP patient cohort on the USC referral pathway and their final diagnosis was confirmed via colonoscopy and pathology. The testing cohort contained a prevalence of cancer patients

Table 7.10: Large RF model cross-validation result on a spectrum-wise basis.

		Actual result			
	n=530	**Cancer**	**Control**		
Predicted	**Cancer**	**196**	**62**	**NPV**	**74.63%**
	Control	**69**	**203**	**PPV**	**75.97%**
		Sensitivity	**Specificity**		
		73.96%	**76.60%**		

in roughly 1 in 3 cases with 33 control patients and 16 cancer patients. The imbalance in the 49 patient cohort between control and cancer patients is to represent a more realistic prevalence in the clinical population.[3] The spectrum-wise results were then converted to patient wise results using the model decision key (Table 7.6).

The confusion matrix for the blind testing set shows a sensitivity of 94% sensitivity and a specificity of 52% for colorectal cancer (Table 7.11). Despite a high number of false positive results the model missed one cancer patient. Given the proposed function of liquid serum RS as a triage tool these results show promise towards translation given the NPV of 94%. Colonoscopy which is the gold standard test has a NPV of up to 99.4% [2]. Colonoscopy has a relatively low completion rate at 57% [10], the RS method takes a standard blood test therefore the completion rate was much higher 99%. Despite the PPV of the RS method being at just 48% for the method this is still higher than the current NICE guidelines which are at just 3% [1]. Comparing the Raman diagnostic model to FIT testing for screening, FIT has a slightly higher NPV at 99.7% but a lower PPV at just 11.2%. Furthermore, FIT is unable to be performed on

[3]It is appreciated that in reality the ratio is more like only 8% have CRC vs control patients in a USC referral cohort in the UK. However, the numbers of control patients collected was not enough to have this large a ratio. 1 in 3 was therefore used as the maximum ratio possible with the patient cohort.

patients with the symptom of rectal bleeding. Whereas, for the RS method this is not a limitation.

Table 7.11: Patient-wise confusion matrix for the detection of colorectal cancer vs control patients in a 49 patient blind model testing set.

		Actual			
	n = 49	**Cancer**	**Control**		
Predicted	**Cancer**	**15**	**16**	**NPV**	**94.44%**
	Control	**1**	**17**	**PPV**	**48.39%**
		Sensitivity	**Specificity**		
		93.75%	**51.52%**		

7.3.3 Feature selection comparisons between polyp and colorectal cancer RF models

The importance selection from the RF models for both polyp vs control and cancer vs control are useful to map the spectral changes back to biologically relevant differences between the serum samples of the patients. Table 7.7 and Table 7.9 show the top 50 wavenumbers split into spectral regions from the RF importance plots from the polyp and cancer diagnostic models.

The model for the detection of polyps highlighted 8 main regions from the top 50 highest ranked wavenumbers. The cancer based model highlighted 8 main spectral regions. The control vs polyp cohort and the cancer vs control models only shared two importance regions around the phenylalanine band between 997-1007 cm^{-1} and at 856 cm^{-1} assigned to other amino acids. The protein changes in the spectra migrate from differences in the Amide III regions to the Amide I region differences. There are some lipid differences in both spectra but these are assigned to different spectral regions. The differences in the most important spectral regions indicate that it potentially would be possible to use the selected features for a downstream selection in the diagnostic modelling for each stage

of progression. To further this idea a non-binary RF model was constructed including inflammatory diseases and also polyp patients to test the capabilities of RF for a non-binary dataset.

7.3.4 Comparison to MS characterisation of CRC from blood samples

This work originally set out to verify the spectral finding with GCMS, however the instrument suffered a catastrophic failure of the GC column. Therefore the peaks attributed to diagnostic capability were compared to literature for serum samples of colorectal cancer patients versus controls to gain some literature based verification of the Raman based results. Farshidfar et al recently published a GC-MS study using orthogonal PLS-DA (OPLS-DA) to establish a serum signature of colorectal cancer [11]. The study included a cohort of 605 patients in three study groups of adenoma, CRC and controls. Figure 7.4 shows a summary of the findings from the study in terms of metabolites with increased and decreased levels in CRC compared to matched controls.

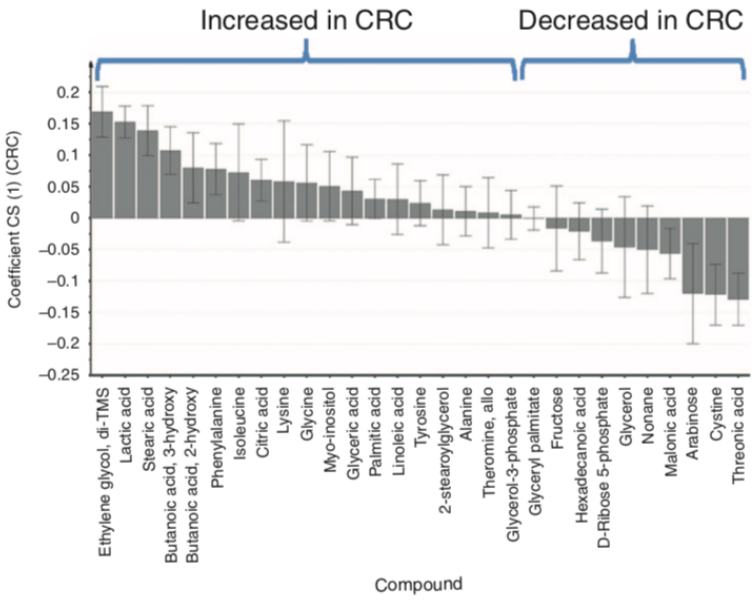


Figure 7.4: Literature study summary of metabolite changes between CRC and controls, taken from [12]

When comparing to the Raman serum results from the importance wavenumbers it is interesting to note that both the Raman and the published study found that levels of Phe and Tyr increase in CRC. Furthermore, one of the contributing peaks not assigned in Table 7.9 is at 1055 cm^{-1}. When comparing to a lactic acid spectrum this peak appears in the spectrum and was also found in the literature values [13]. Furthermore, The Raman importance shows dependence on fatty acids and proteins. Within the literature based results fatty acids such as stearic acid and amino acids responsible for protein structure such as glycerol were also found to be different. In future a GC-MS and LC-MS study would be useful to validate the spectral findings and potentially discover the metabolites that are being detected via RS. To link the results, a network based analysis is proposed.

7.3.5 Investigating a non-binary random forest diagnostic model including polyp, control and cancer patients

The differences in importance distributions between control-polyp models and control-cancer models indicate that it may be possible to combine the training data from both models into a non-binary classification model. The number of polyp spectra collected was smaller than cancer and healthy control spectra due to the total number of patients recruited. To avoid any RF bias towards any particular group the number of control and cancer patients was reduced to keep training groups numbers as even as possible. A training set was constructed from data from 187 control spectra, 172 polyp spectra and 180 cancer spectra from a total of 117 participants. The distribution of the patients and the patient information is summarised in Table 7.3. Spectral data for both training and testing were pre-processed for the model using the GUI via the process in section 7.2.3 and used to construct a RF based model.

A cross-validated confusion matrix was used to calculate the training sensitivities and specificities as seen in Appendix E.3. The calculated sensitivity and specificity for control patients was 71% and 81%, respectively. The polyp sensitivity and specificity were 59% and 85% and the cancer patients 64% and 82%. As with the individual models, the predictor importance for the non-binary

model was also plotted as seen in Figure 7.5.

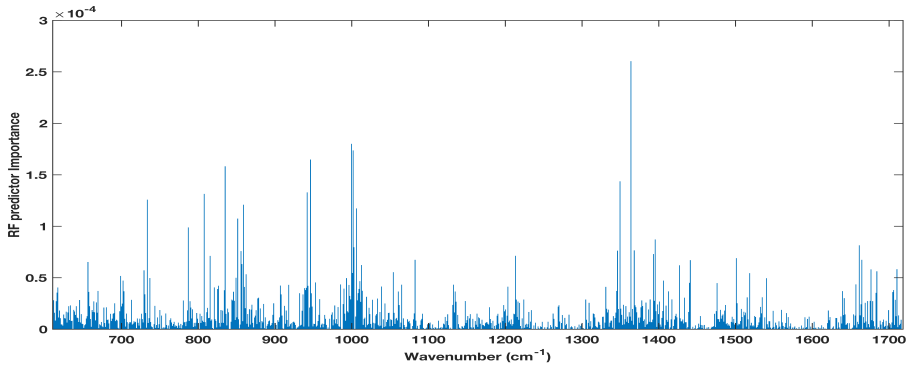


Figure 7.5: Predictor importance plot for the polyp, control and cancer RF model.

As expected, the regions of highest importance for correct classification that were shared in the individual models such as the region around 997-1007 cm^{-1} and at 856 cm^{-1} were also large features of importance in the combined model. The largest region of importance was at 1346 cm^{-1} which was mostly prominent in the polyp model. However, there was also contributions in the 1667-1668 cm^{-1} region in the combined model. As with previous models within this work the trained RF model was tested against an independent testing dataset.The testing set consisted of 225 spectra from 75 patients. The set contained 47 control patients, 11 polyp patients and 17 cancer patients. Table 7.12 shows the calculated confusion matrix from the predicted responses for the polyp vs control calculated model. The per-spectrum sensitivities and specificities for each group were calculated from the confusion matrix and can be seen in Table 7.13.

Table 7.12: Confusion matrix for spectrum wise-analysis from an independent test set of 225 spectra.

		Actual		
	n=225	**Control**	**Polyp**	**Cancer**
Predicted	**Control**	**56**	**1**	**8**
	Polyp	**50**	**26**	**12**
	Cancer	**35**	**6**	**31**

Table 7.13: Table of sensitivities, specificities, NPV and PPV calculated from the spectrum-wise independent test set confusion matrix in 7.12

	Sensitivity	Specificity	PPV	NPV
Control	**0.40**	**0.89**	**0.86**	**0.47**
Polyp	**0.79**	**0.68**	**0.30**	**0.95**
Cancer	**0.61**	**0.76**	**0.43**	**0.87**

As with the individual models, the NPV for polyps and CRC was high at 95% and 87% respectively. In comparison with the gold standard of colonoscopy with an NPV of 98% for CRC, this is a good comparative result for a serum based test. This is potentially at the cost of a large number of false positive results. However, the cohort of testing patients that came through this study would all undergo colonoscopies in the current treatment pathway. The Raman based triage tool could potentially have prevented a number of 'needless' colonoscopies. To quantify the number of patients that would have been referred correctly or incorrectly with the Raman CRC triage tool the results per patient were calculated. Table 7.14 shows the confusion matrix for the independent testing set per patient for the non-binary model. The confusion matrix shows that although

Table 7.14: Per patient results from non-binary model testing polyp, control and cancer patients.

		Predicted			
	n=75	**Control**	**Polyp**	**Cancer**	**equivocal**
Actual	**Control**	**15**	**15**	**7**	**10**
	Polyp	**0**	**9**	**2**	**0**
	Cancer	**2**	**4**	**11**	**0**

there were mis-classifications between the polyp and cancer spectra there were only 2 patients that were predicted to be control patients (i.e. would not be referred) from both the polyp and the cancer patients.

It is clear that despite the high NPV values for cancer and polyp patients on a spectrum-wise basis when considering the per patient results there was a higher number of false positive results from the controls than there were true positive results. Nevertheless in a cohort of 75 patients this could have prevented 20% of the colonoscopies. The per patient sensitivity, specificity, NPV and PPV values were not calculated from this table due to the presence of equivocal results wherein there was no majority to decide the class decision. The equivocal results were caused by each of the three repeat spectra being assigned to different classes. For example, if the result was 0, 1, 2 this would be equivocal. Whereas a result that was 0, 1, 0 or 0, 0, 2 would give a definitive answer of control. This was only the case for control patient spectra and is expected from the high number of false positive results. This could be changed in future to add equivocal as an extra class decision or by increasing the number of spectra per test, e.g. 5 instead of 3 or by repeating spectral collection and analysis for these patients.

7.3.6 The effects of inflammatory diseases on diagnostic capability

Thus far a patient group that has not been considered within this analysis are patients with non-malignant diseases. Within a GP cohort of patients, there would be patients needing referral to secondary care that would not necessarily need to be referred under the USC pathway if they had benign diseases. Therefore, some patients with inflammatory bowel diseases such as diverticular disease were included to investigate the effect that adding these patients had into a diagnostic model would have on the sensitivity and specificity of the RF model. As patients with polyps and CRC would need to be referred the control vs polyp vs cancer model trained in the previous section was extended to include control patients with inflammatory diseases. However, to keep the outputs of the test as simple as possible, the data were grouped into control and inflammatory control versus polyp and cancer in a binary model. The final aim of this project would be a diagnostic model able to truly triage colorectal referrals and discriminate between patients who need to be referred under the USC pathway for CRC and

patients who do not. The spectral data were preprocessed using the GUI via the methodology in Chapter 4.4 and preprocessed as per section 7.2.3.

The binary RF model was trained via the same protocol as in section 7.2.4 using the preprocessed spectral data. The spectra used were the same as the non-binary model in section (patient demographic information in Table 7.3) with the addition of a group of patients with diverticular, colitis and Crohn's disease. The testing cohort was also the same as in the previous polyp, control, cancer model with the addition of spectral data collected from 47 patients with non-malignant colorectal diseases. The details of these patients can be found in Table 7.3.

Figure 7.6 shows the calculated CV ROC curve for the diagnostic model including inflammatory diseases. It is clear that the AUC is less than previously seen at 0.7583, the classifier sensitivity of the CV training model was 64.29% with the specificity at 59.65%. The NPV value for a combined group model of patients that needed to be referred under USC (i.e. polyp+cancer) was 77.27% which is a 10% reduction compared to the worst NPV value from the previously calculated models in this chapter. This indicated that the inclusion of inflammatory patients reduces the capability of the serum RS platform to distinguish between patients who would need to be referred under the USC pathway and patients that would not.

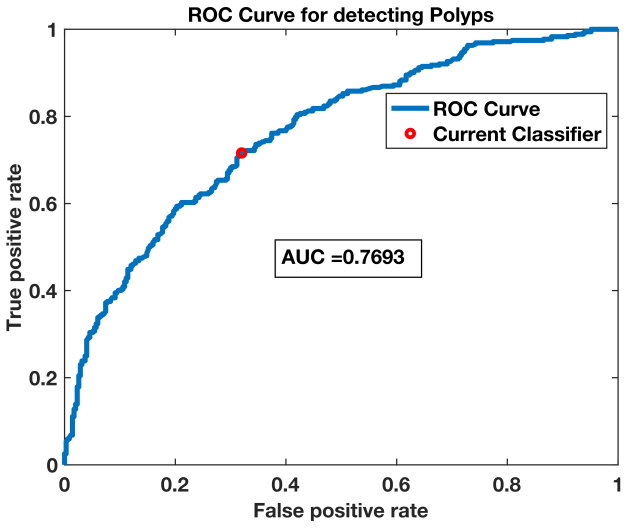


Figure 7.6: Predictor importance plot for the polyp, control and cancer RF model.

The per patient refer or non-refer decisions for the testing cohort for the in-

flammatory model were calculated as in Table 8.3. As with the ROC and the AUC, the overall sensitivity and specificity of the model was lower than previous studies. The sensitivity and specificity were calculated to be 64.29% and 59.65% respectively. The NPV value for a combined group of polyp and cancer patients was 77.27% which was considerably less than in the individual case-control models and the combined polyp, control and cancer models which had a minimum NPV of 87% for cancer patients. The individual number of cancer and polyps that were missed by the test increased to 10 patients in total. Therefore, the inclusion of inflammatory control patients such as patients with colitis significantly reduced the diagnostic capability of the model. This is potentially due to there being similar biochemical changes within serum for patients with cancer and inflammatory controls. When considering the predictor importance for this model the overall dominating spectral feature is in the region of the phenylalanine peak at 1002-1003 cm^{-1}. Future work could include characterising the differences in this peak throughout disease progression for patients in the hope to characterise disease related changes. However, the results of this model suggest that more work would need to be conducted into the spectral differences between inflammatory patient spectra and cancer spectra.

Table 7.15: Confusion matrix and calculated sensitivity, specificity, NPV and PPV values for a binary classification including inflammatory control patients.

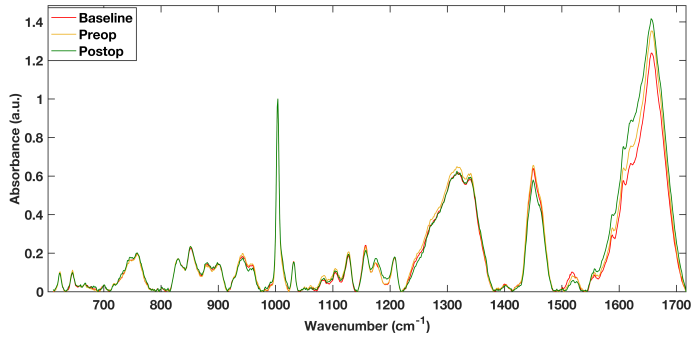
		Predicted			
	n=85	**no refer**	**refer**		
Actual	**control**	**27**	**20**	**NPV**	**77.27%**
	inflam	**7**	**3**		
	polyp	**7**	**4**	**PPV**	**43.90%**
	cancer	**3**	**14**		
		Sensitivity	**Specificity**		
		64.29%	**59.65%**		

7.3.7 Disease monitoring

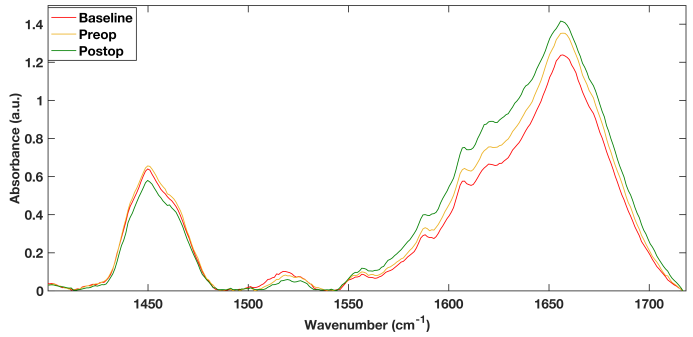
Within this study two rectal cancer patients were able to be sampled sequentially before treatment, during treatment and post surgery. Therefore the total spectral changes were averaged across two patients for their individual path and plotted. The first patient was T2N0 (from a CT) at baseline. The patient was treated via chemo-radiotherapy (CRT), pre- treatment the patient was confirmed to be at stage T4N0. The patient then has a complete resection of the cancer. The samples for this patient were collected at baseline/admittance, post CRT and post surgery.

Figure 7.7 (a) shows the average spectral differences for the first patient across their treatment stages. Many regions across the region have little to no differences as the treatment progresses. Figure 7.7 (b-d) shows zoomed spectral regions between that show the most significant spectral changes. As the treatment progresses for the patient there are differences in the region between 1400-1720 cm^{-1} as seen in Figure 7.7 (b). In this region the peaks attributed to CH2/CH3 and OH bonds within the samples show a drop post surgery. There are also differences across the treatment in the carotenoids region (between 1500-1550 cm^{-1}). There is a small drop in levels between the baseline (basl) and preoperative (preop) stages. Postoperatively, (postop) this peak reduces again. The baseline, preop and postop peak ratio changes between the carotenoid region and the Amide 1, Phe and Trp region of the spectra. The baseline had a much lower response in this region between 1550-1720 cm^{-1} whereas the preop has higher levels with the highest at postop.

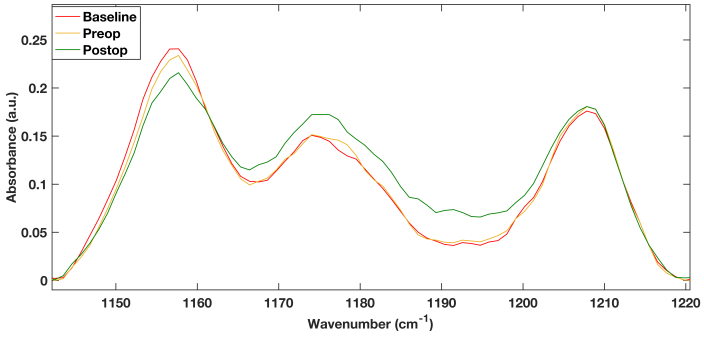
Figure 7.7 (c) also shows a ratio change between the baseline and preop samples compared to the postop samples at 1157 cm^{-1} and 1174 cm^{-1} attributed to carotenoids and DNA (Tyr) respectively. Furthermore, previous work has shown inflammatory response to effect the phenylalanine peak shoulders. This is associated to an inflammatory response [14]. In the postop sample the shoulders around the peak were increased in line with previous work as seen in Figure 7.7 (d). One potential explanation for this trend may be the response of the patient to the surgery.



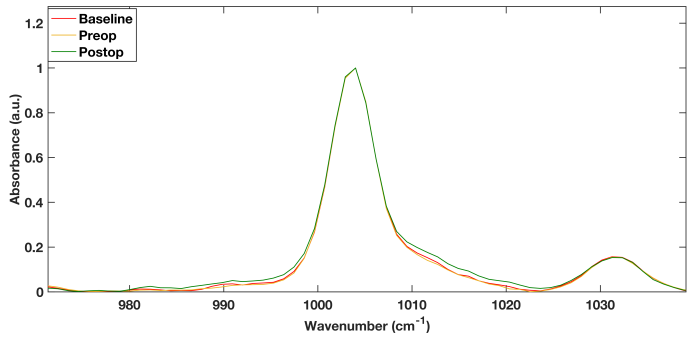
(a)



(b)



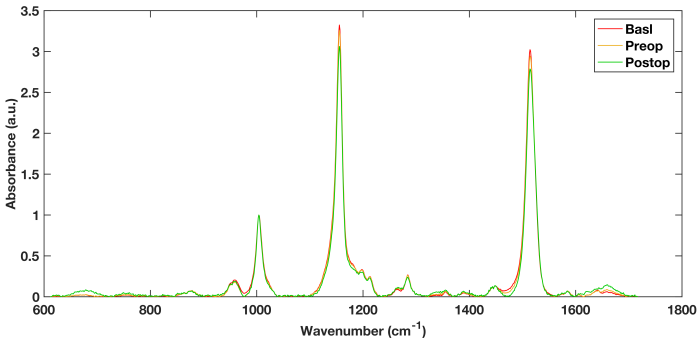
(c)



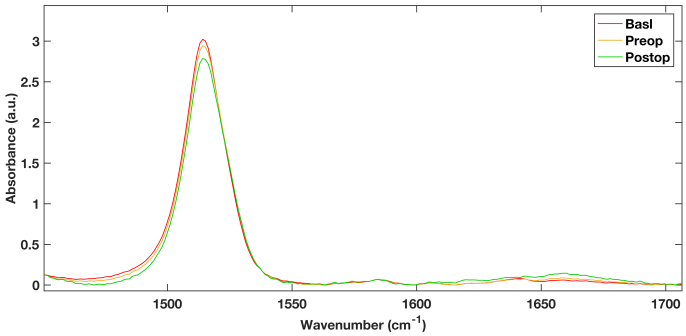
(d)

Figure 7.7: Average 785 nm patient spectra for a patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. Spectra for (a) (b) (c) and (d) .

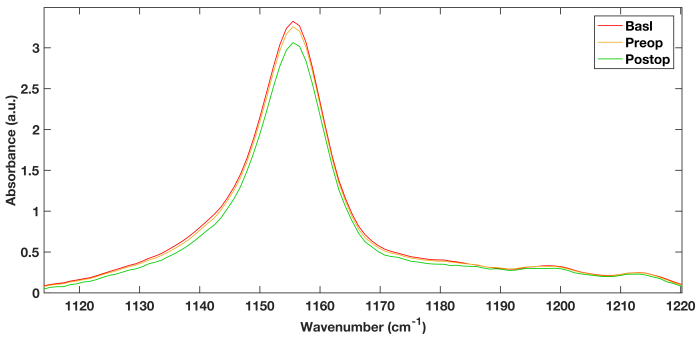
The spectral differences shown in the longitudinal samples are consistent with regions of the spectra that are important for correct disease classification as seen in Tables 7.7 and 7.9. Specifically, it appears to be sensitive to the carotenoid, amino acid and amide peaks. This analysis was repeated with 532 nm excitation. Figure 7.8 (a) shows the average changes in the 532 nm spectra for the first patient. The carotenoid peaks at 1157 cm^{-1} and 1520 cm^{-1} show matching trends in the peak heights to the 785 nm which is expected. The spectra show that postop the levels of carotenoids are at their lowest. From the previous analysis within this thesis, this seems contradictory, throughout this work non-cancer patients have shown higher carotenoid levels. However, the post op sample was within 24h of the patient undergoing surgery. Ideally an extra sample e.g. 6 months after could be taken to monitor if the levels move towards levels higher than at the patient baseline level. The trend in the Amide I / amino acid region is also seen in the 532 nm spectra. The baseline sample having the lowest spectral response in that region. The trend with the Phe peak was not seen in the 532 nm spectra as can be seen from Figure 7.8 (d). However, in 532 nm excitation this peak is broader than in 785 nm so this could contribute to this effect.



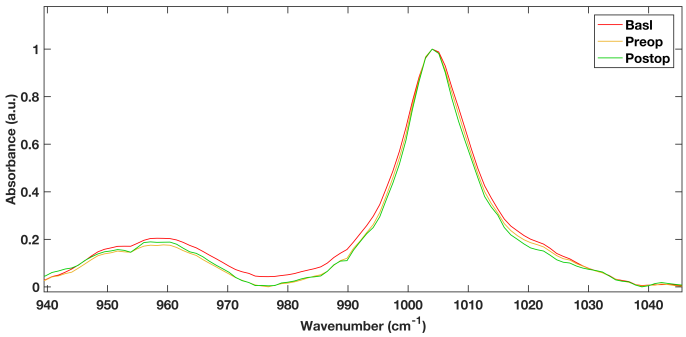
(a)



(b)



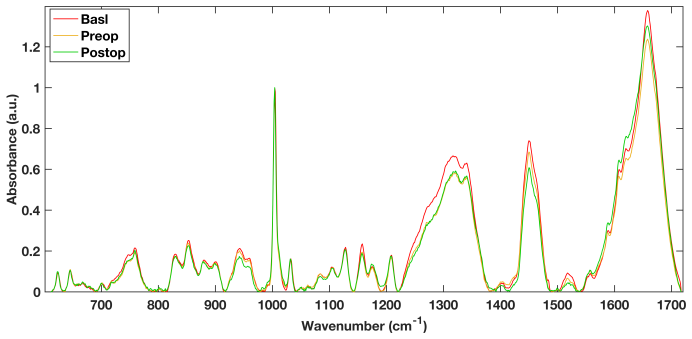
(c)



(d)

Figure 7.8: Average 532 nm patient spectra for a patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. Spectra for (a) (b) (c) and (d) .

Another patient was also monitored who was a female with T3N0 cancer at baseline. This patient showed similar trends to the first patient. The full figures for this patient can be seen in Appendix E.4. In the 785 nm spectra the carotenoid peaks followed the same trend as the first patient with decreasing levels across each treatment stage. The region containing the Amide I and amino acid peaks (Phe, Tyr, Trp) shows that the postop has the highest spectral response whereas the baseline and preop are lower (but not in the same order as patient 1). This is can be seen in the average 785 nm plot (Figure 7.9). The shoulder on the right of the Phe peak at 1004 cm^{-1} also follows the same trend as patient 1. There were additional differences within the spectra from the second patient with a much higher spectral difference in the Amide III/NH stretch region between 1225 cm^{-1} and 1378 cm^{-1}. There were also larger differences in the cholesterol, amino and nucleic acid region at 740-778 cm^{-1} than in the first patient's analysis.



(a)

Figure 7.9: Average 785 nm patient spectra for the second patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative.

The prospect of disease monitoring would be useful for less invasive treatment monitoring within a clinical setting. The trends within the spectra show that there are potentially traceable differences as patients go through treatment.

One weakness of this study is the lack of a control patient tracked across a similar time period. In future this would be conducted to show that spectral changes aren't only due to different sampling times. Furthermore, modelling the trends between patients will require further sequential sampling across treatments. The response of each patient will need to be considered within a model for disease

monitoring.

7.3.8 Investigating a multi laser 785 nm and 532 nm diagnostic model

The sensitivity of the serum RS platform within a multi-model approach was also investigated. Data from 785 nm and 532 nm excited samples were combined to see if the spectral differences highlighted by different excitation wavelengths made any improvement to the diagnostic capability of serum RS. To achieve this pre-processed 785 nm and 532 nm data were stitched together for each patient creating a combined spectrum (Figure 7.10).

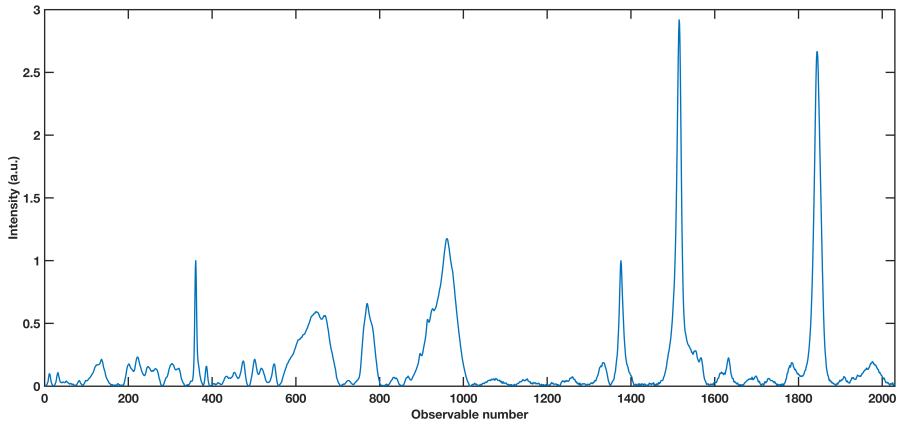


Figure 7.10: Example combined 785 nm and 532 nm spectrum. Both spectra have been normalised to the Phe peak at 1004 cm^{-1}.

A RF classifier was then constructed from a combined dataset with 490 individual spectra from 98 patients (cohort details in Table 7.4). As with the 785 nm model the training process for the RF classifier was repeated 10 times and the model producing the best CV sensitivities and specificities was chosen as the final trained model. Table 7.16 shows the confusion matrix from the multi-modal model. The sensitivity and specificity of the training model was 82% and a sensitivity of 77.92% on a spectrum-wise basis.

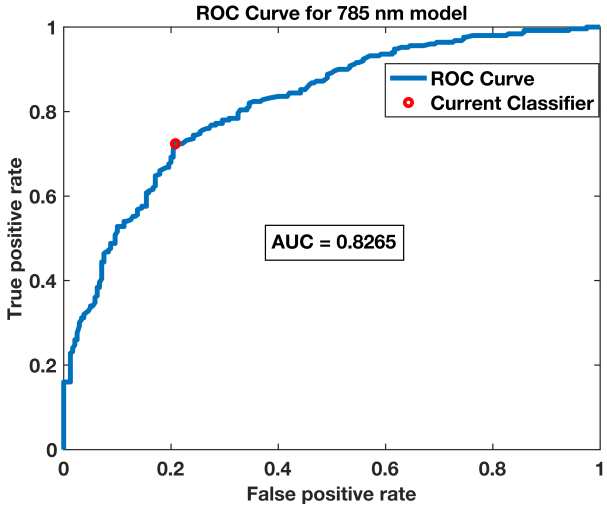
The NPV and PPV values for the training dataset were not considered in an independent test set because NPV and PPV values depend on prevalence of the

disease within a tested population. Therefore, to gain a better sense of these values a blind test cohort was constructed from data from 49 GP referral patients totalling 147 spectra. The cohort contained almost 2 control patients for each cancer patient which is more indicative of the real life situation. The testing spectra were tested blindly against the combined model and then final results per spectrum and per patient were calculated as well as NPV and PPV values for the model on a spectrum-wise and patient wise basis.

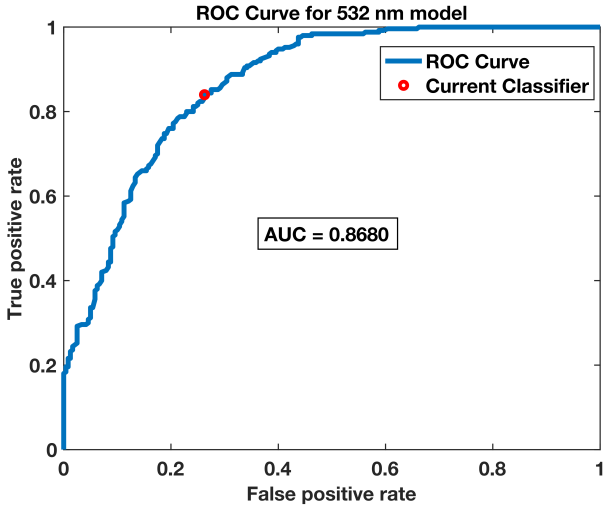
Table 7.16: Spectrum-wise confusion matrix for a combined laser excitation model.

		Actual			
	n= 490	**Cancer**	**Control**		
Predicted	**Cancer**	**205**	**53**	**NPV**	**80.60%**
	Control	**45**	**187**	**PPV**	**79.46%**
		Sensitivity	**Specificity**		
		82.00%	**77.92%**		

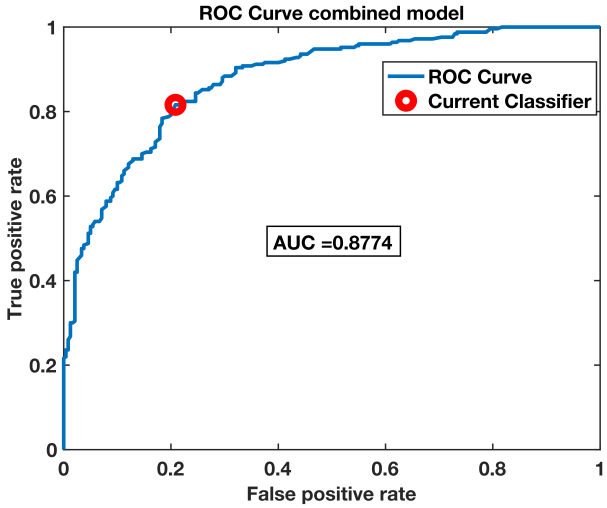
As the spectra entered into this model were slightly different to the large cohort study reported above, a direct comparison was conducted between this multi-mode model and models constructed from the individual 785 nm and 532 nm spectra. Figure 7.11 shows a comparison between the CV-ROC for the 785 nm, 532 nm and combined models including their calculated AUCs. The ROC curves show that the multi-modal cross validation showed higher AUC than either of the individually trained models.



(a)



(b)



(c)

Figure 7.11: Cross validated ROC curve comparison between a 785 nm model (a), a 532 nm model (b) and a combined model (c).

The diagnostic ability within a test set was also investigated. Table 7.17 shows the comparative sensitivities, specificities, NPV and PPV values for individual laser excitation diagnostic models versus a combined approach for a training dataset and an independent test set. Within the model training (CV) values the combined dataset performs the best across the board compared to the individual datasets. Furthermore, contrary to the earlier work the 532 nm model performs better during the CV than the 785 nm model.

Comparing the independent test sets of the combined model to the individual models, the combined model gained sensitivity and NPV over 785 nm and 532 nm models on a per-spectrum basis. The combined model had improved sensitivity over the 532 nm model on a patient-wise basis and equivalent sensitivity to the 785 nm model. On a spectrum-wise basis and patient-wise basis the combined model had a lower specificity than both of the individual models. The NPV and PPV values were higher in the combined model than in the individual models on a per spectrum basis. On a per patient basis, the combined model had NPV and PPV higher than the 532 nm model. The NPV was 0.37% lower in the combined model than in the 785 nm model and the PPV was 1.52% lower in the combined model than the 785 nm model. This suggests that the 785 nm was dominant in the diagnostic model.

The dominance of the 785 nm data can also be seen when investigating the predictor importance plot for the combined RF model. Figure 7.12 shows that the majority of the most important peaks are similar to those found in Figure 7.10. The 532 nm spectrum does have some peaks that show some importance but these are not as high as the 785 nm spectrum. Therefore, the 532 nm addition to the end of the spectrum marginally improves the diagnostic capability but there would be a practical trade-off given the extra acquisition time of the data. This would limit the technique in terms of acquisition time. The acquisition time for 3 spectra with the 785 nm spectra is at 12 mins per patient. However this time would double with the 532 nm acquired spectra which may affect the high-throughput capabilities of the diagnostic.

Analysis in the previous chapter showed that vector normalised data could produce similar results to data normalised to the Phe peak. This was also tested

Table 7.17: Comparison of the sensitivities, specificities, NPV and PPV values of individual excitation based models and a combined model.

		Sens (%)	**Spec** (%)	**NPV** (%)	**PPV** (%)
Training (CV)	785 nm	76	73.33	74.58	74.8
	532 nm	81.2	75.42	79.39	77.48
	Combined	82	77.92	80.6	79.46
Test (per spec)	785 nm	88.24	46.88	88.24	46.88
	532 nm	74.51	55.21	80.3	46.91
	Combined	92.16	44.79	93.75	48.48
Test (per pt)	785 nm	94.12	50	94.12	50
	532 nm	76.47	56.25	81.82	48.15
	Combined	94.12	46.88	93.75	48.48

and the confusion matrices can be found in Appendix E.2. It was found that the models with data normalised to the phenylalanine peak out performed the vector normalised model in both training and testing datasets with a maximum sensitivity of 85% and specificity of 56.24% and an NPV and PPV of 86% and 50% respectively

The multi modal model tested here would need to be extended past a binary model and tested in the future to see if the combination would improve the effects of inflammatory diseases on the diagnostic capability. Another method of improving this technique could be to feature select the spectral wavenumbers using the RF importance from the combined model to tune the diagnostic between smaller differences in the spectra.

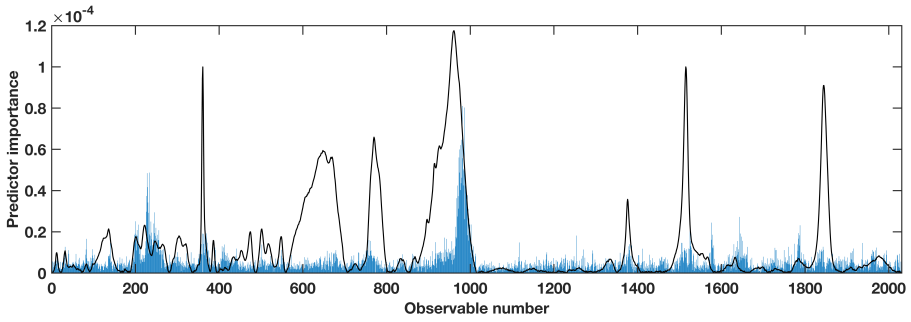


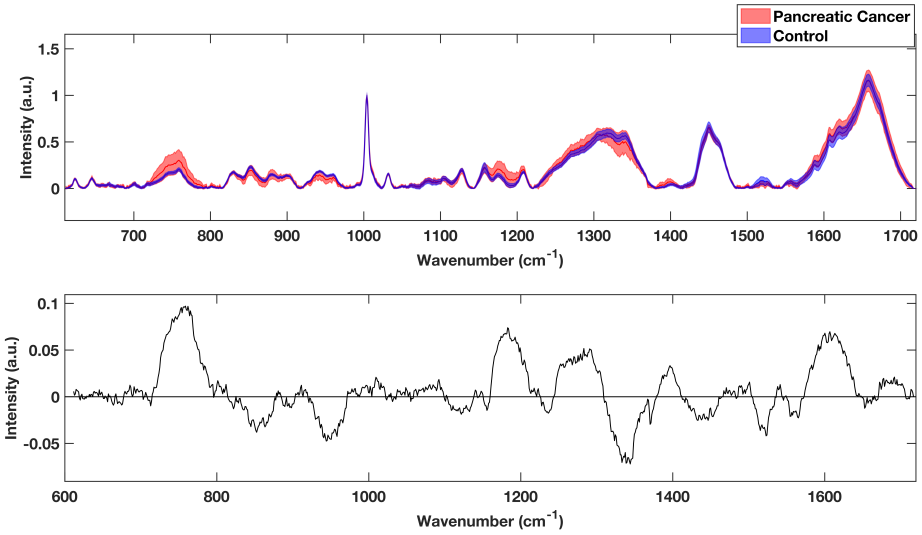
Figure 7.12: Predictor importance for a combined multi-modal model for colorectal cancer.

7.3.9 Application of methodologies to other cancer types

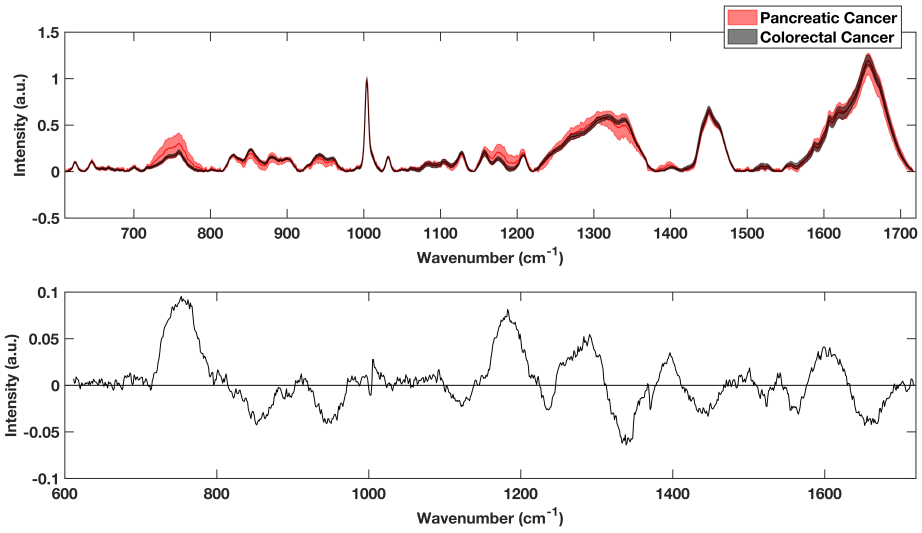
The data acquired during this work has a main focus on the detection of CRC in serum samples versus controls. The maximum NPV and PPV values from the models discussed above show that detection of CRC is possible to a standard above that of just the USC pathway [15]. It is therefore important for a clinical tool and the serum RS techniques to be as specific to CRC as possible. Spectra from patients with pancreatic cancer were tested to assess if the liquid serum RS platform would have the ability to distinguish between control and cancer and then also be able to distinguish which type of cancer is being detected. Pancreatic cancer is the fourth largest cause of cancer related deaths worldwide [16, 17]. As with CRC, early diagnosis correlates to better clinical outcomes. There are currently no accepted biofluid biomarkers for the early diagnosis of pancreatic cancer. Therefore, there is also a potential for HT analysis of serum to also be applied to pancreatic cancer detection in serum samples.

To investigate the potential of serum RS for pancreatic cancer detection, a small number of patients with pancreatic ductal adenocarcinoma (PDAC, n=15) were recruited and their serum samples were analysed using the HT liquid 785 nm platforms developed in (Chapter 5, section 5.3.7). Data from the samples were processed in the standardised GUI methods (Chapter 4, section 4.4). Data were then compared to matched control samples (n=15) and colorectal cancer samples (n=15). The mean, standard deviation and difference spectra were calculated

between pancreatic data and control data and between pancreatic cancer and CRC for the 785 nm liquid process as seen in Figure 7.13. The pancreatic data has clear visual differences and a larger spectral standard deviation compared to the control samples (Figure 7.13 a). The main spectral differences highlighted between pancreatic cancer and control samples are that 758 cm^{-1}, 1179 cm^{-1}, 1252-1290 cm^{-1}, 1394 cm^{-1} and 1590-1610 cm^{-1} are higher in the cancer spectra and peaks at 853 cm^{-1}, 944 cm^{-1}, 1343 cm^{-1}, 1518-1525 cm^{-1}, 1566cm^{-1} are higher in the control spectra. The differences are similar in both the pancreatic cancer vs control samples and the differences spectra between pancreatic cancer and colorectal cancer as seen in Figure 7.13 b. There are a few extra differences shown in the inter-cancer difference at 1004 cm^{-1} attributed to phenylalanine and slight changes in the 1600-1700 cm^{-1} region. The differences shown indicate that the control spectra and colorectal cancer spectra are more similar than the pancreatic cancer samples.



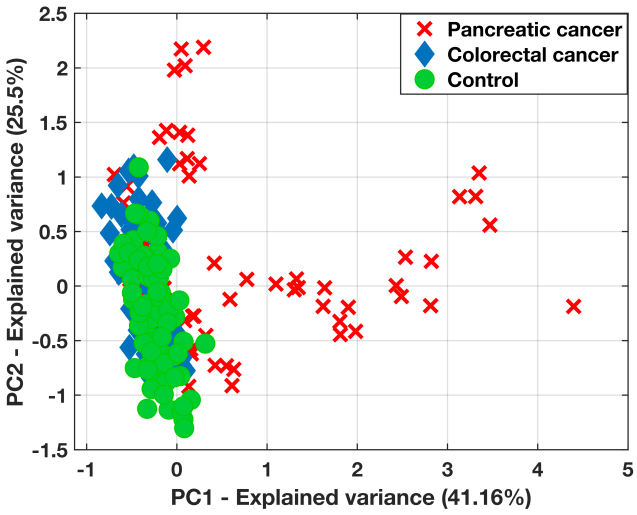
(a)



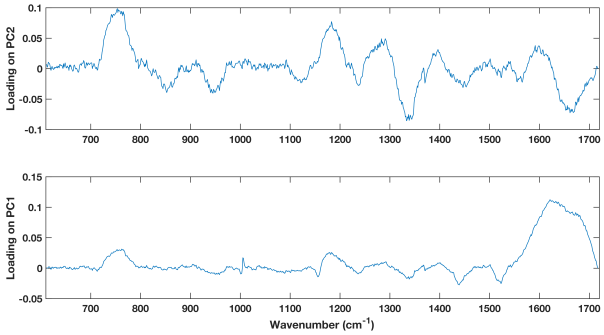
(b)

Figure 7.13: Mean, standard deviation (upper) and calculated difference (lower) spectra between pancreatic cancer and control samples (a) and between pancreatic and colorectal cancer (b).

To further investigate the magnitude of the spectral variances between pancreatic cancer, colorectal cancer and control spectra the spectra were subject to PCA analysis. Figure 7.14 shows the PC1 vs PC2 score plot for pancreatic cancer, colorectal cancer and control patients.



(a)



(b)

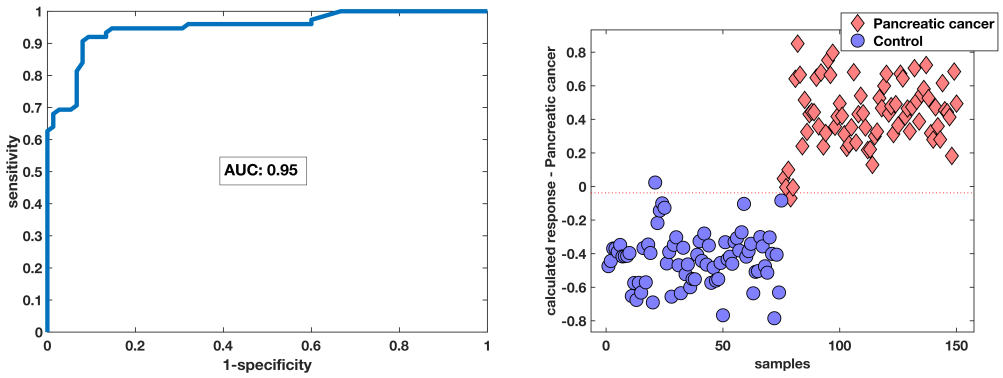
Figure 7.14: PCA score plot (PC1 vs PC2) (a) and associated loadings on PC1 and PC2 (b) for pancreatic cancer, colorectal cancer and control data.

The pancreatic cancer spectra show a much larger spread and variance than the colorectal and control sample spectra, however the pancreatic spectra are separated from the others along PC1. The explained variance of the dataset explained by PC1 was 41.16%. When considering that the pancreatic cancer spectra are separated by PC1, the loading on PC1 shows good agreement with the calculated difference spectra in Figure 7.13. The differences in the phenylalanine peak are also evident in the loading on PC2 where, despite overlapping, there is some slight separation between the colorectal and control spectra and the pancreatic vs colorectal cancer spectra. The separation of the pancreatic spectra using unsupervised PCA analysis shows that there is great potential for pancreatic cancer

to be detected using supervised discriminant analysis. The number of patients involved within the pancreatic cancer study were relatively small, therefore PLS-DA models were used to investigate pancreatic cancer vs control samples and pancreatic cancer vs colorectal cancer for basic detection and the potential for specific cancer detection.

Partial least squares discriminant analysis pancreatic cancer vs control

Initially, it is important to compare the potential diagnostic capability of pancreatic cancer compared against control. Therefore PLS-DA model was calculated from 15 pancreatic vs 15 control samples and was tested with 5 pancreatic cancer and 5 control samples. The PLS-DA model was constructed with 11 latent variables such that the CV error was minimised and cross validated with k-fold 5 fold CV. The training model CV-ROC and confusion matrix were calculated for the PLS-DA model and can be seen in Figure 7.15 and Table 7.18. The CV model AUC was higher than the colorectal PLS-DA models, with the CV sensitivity calculated to be 88% and specificity of 92% and the AUC = 0.95. This indicates that the platform would also be suited to pancreatic cancer detection in serum samples.



(a) (b)

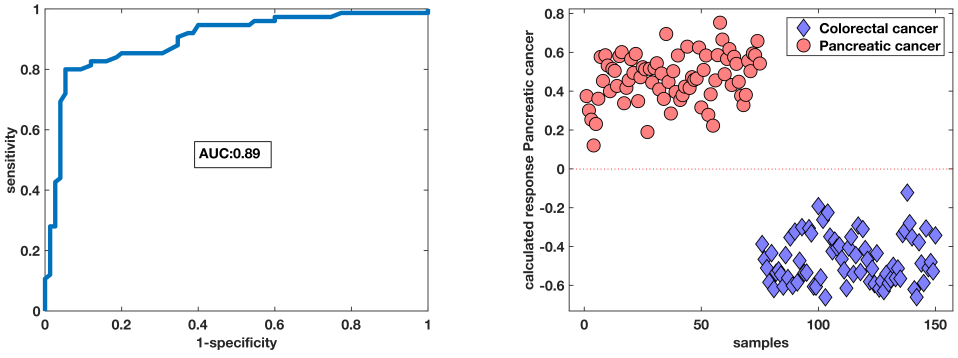
Figure 7.15: ROC (a) and CV calculated response (b) for pancreatic cancer vs control patients.

Table 7.18: Cross-validated confusion matrix for PLS-DA model on a spectrum-wise basis for pancreatic cancer versus control.

		Actual	
	n= 150	**Cancer**	**Control**
Predicted	**Cancer**	**66**	**6**
	Control	**9**	**69**
		Sensitivity	**Specificity**
		88.00%	**92.00%**

Partial least squares discriminant analysis pancreatic cancer vs colorectal cancer

The ability to discriminate between different cancer samples and control samples could be useful for a multi-cancer detection technique. However, for maximum impact it would be advantageous for the serum Raman platform to be able to distinguish between cancer types. To assess the possibility of specific cancer detection a PLS-DA model was calculated for CRC spectra vs PDAC spectra. The PLS-DA model was constructed with 14 LVs such that the CV error was minimised within the model.



(a) (b)

Figure 7.16: Cross validated calculated ROC (a) and calculated responses vs scores for pancreatic versus colorectal cancer (b).

Table 7.19: Cross validation confusion matrix for spectrum-wise PLS-DA model discriminating between pancreatic cancer and colorectal cancer.

		Actual	
	n= 150	**CRC**	**Panc**
Predicted	**CRC**	**60**	**4**
	Panc	**15**	**71**
		Sensitivity	**Specificity**
		80.00%	**94.67%**

The calculated CV-ROC showed an area under the curve of 0.89 showing a good learner. Between cancer types, the sensitivity and specificity calculated from the confusion matrix (Table 7.19) are 80% and 95%. The diagnostic capabilities in a pilot dataset for the discrimination of control and pancreatic cancer samples as well as between cancer types were above 80%. This shows potential for the technique to be used for both pan-cancer screening and potentially for there to be specific spectral biomarkers associated to primary adenocarcinomas at different locations. However, more work would need to be conducted to establish if the technique would be used within one large model or if the final test would be a downstream selection of cancer vs control and then another test for cancer location. Furthermore, more patients will need to be recruited for a range of other GI and other cancers to determine if the type of malignancy (e.g. adenocarcinoma vs melanoma) can be determined via serum RS coupled to machine learning based techniques.

7.4 Conclusions and further work

The goal of this chapter was to establish the diagnostic limits of the Raman platform. The binary cancer vs control model based on 785 nm data showed a good NPV at over 90%. The NPV of the testing set was 94.44% meaning for that cohort of patients 35% of the colonoscopy referrals could have been prevented.

This is a good move towards the gold standard of colonoscopy which has a NPV of 98% and shows good promise as a triage tool.

The results could be improved upon in the future by repeating the fresh dataset after a freeze-thaw cycle to create a new algorithm based on samples that have had the exact same treatment through analysis. As work in previous chapters showed this had little effect on the overall diagnostic capability but did affect the results across different sampling modes.

The limits of the Raman platform were established within binary models showing that the technique could distinguish between polyp and control samples. It still distinguished well between patients that needed referral under USC and those who do not against a healthy control cohort. However,currently the NPV and PPV of the test is reduced to 77% and 43% when inflammatory control samples are introduced into the model. This could be improved by using feature selection for the diagnostic model. Features can be selected from the RF importance plots in the binary cases and also by performing validation work via mass spectrometry to try and isolate spectral regions to target in the study groups.

The effect of adding 785 nm and 532 nm data together was investigated and showed that the 785 nm sensitivity was dominant for a binary cohort model with multi-laser excitation. This approach could also be investigated further to investigate the sensitivity to inflammatory diseases to see if the dependence on the 785 nm spectra remains.

The effects of treatment on spectral response was not able to be investigated via diagnostic models but average spectra for a few patients shows that there is potential to track spectral changes across a patient's treatment towards treatment monitoring. The study shows, along with the RF importance in the larger models that serum RS is very sensitive to Phe and Tyr levels in the serum. Wiggins et al discussed that the ability to track levels of these in serum has potential for a biomarker in GI cancers [18]. Therefore, further work tracking a larger cohort of patients and matched controls will need to be conducted. This will also need to be paired to an external validation method such as GCMS/LCMS.

The liquid serum RS platform is applicable to more than one cancer type with the ability to distinguish between control and pancreatic cancer spectra as

well as between CRC and pancreatic cancer spectra. The CV PLS-DA models constructed showed higher sensitivity and specificity between controls and cancers than the pilot studies for CRC. This shows great promise for the Raman platform to be a pan-cancer platform. In future, more pancreatic cancer samples will be analysed to establish diagnostic limits for a case-control study as well as a pan-cancer study.

Bibliography

[1] National Institute for Health and Care Excellence from 1 April 2013.

[2] Hooi C Ee, James B Semmens, Neville E Hoffman, et al. Complete colonoscopy rarely misses cancer. *Gastrointestinal endoscopy*, 55(2):167–171, 2002.

[3] Jeroen C. Van Rijn, Johannes B. Reitsma, Jaap Stoker, Patrick M. Bossuyt, Sander J. Van Deventer, and Evelien Dekker. Polyp miss rate determined by tandem colonoscopy: A systematic review. *Am. J. Gastroenterol.*, 101(2):343–350, 2006.

[4] D. K. Rex, C. S. Cutler, G. T. Lemmel, E. Y. Rahmani, D. W. Clark, D. J. Helper, G. A. Lehman, and D. G. Mark. Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. *Gastroenterology*, 112(1):24–28, 1997.

[5] Dinesh K. R. Medipally, Adrian Maguire, Jane Bryant, John Armstrong, Mary Dunne, Marie Finn, Fiona M. Lyng, and Aidan D. Meade. Development of a high throughput (HT) Raman spectroscopy method for rapid screening of liquid blood plasma from prostate cancer patients. *Analyst*, 142(8):1216–1226, 2017.

[6] Michael N Passarelli and Polly A Newcomb. Blood lipid concentrations and colorectal adenomas: a systematic review and meta-analysis of colonoscopy studies in asia, 2000–2014. *American journal of epidemiology*, 183(8):691–700, 2016.

[7] Matthew J. Baker, Shawn R. Hussain, Lila Lovergne, Valérie Untereiner, Caryn Hughes, Roman A. Lukaszewski, Gérard Thiéfin, and Ganesh D. Sockalingum. Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chem. Soc. Rev.*, 45(7):1803–1818, 2016.

[8] PFC Lung, D Burling, L Kallarackel, J Muckian, R Ilangovan, A Gupta, M Marshall, P Shorvon, S Halligan, G Bhatnagar, et al. Implementation

of a new ct colonography service: 5 year experience. *Clinical radiology*, 69(6):597–605, 2014.

[9] Akiharu Watanabe, Toshihiro Higashi, Tatsuro Sakata, and Hideo Nagashima. Serum amino acid levels in patients with hepatocellular carcinoma. *Cancer*, 54(9):1875–1882, 1984.

[10] CJA Bowles, R Leicester, C Romaya, E Swarbrick, CB Williams, and O Epstein. A prospective study of colonoscopy practice in the uk today: are we adequately prepared for national colorectal cancer screening tomorrow? *Gut*, 53(2):277–283, 2004.

[11] Farshad Farshidfar, Aalim M. Weljie, Karen A. Kopciuk, Robert Hilsden, S. Elizabeth McGregor, W. Donald Buie, Anthony MacLean, Hans J. Vogel, and Oliver F. Bathe. A validated metabolomic signature for colorectal cancer: Exploration of the clinical value of metabolomics. *Br. J. Cancer*, 115(7):848–857, 2016.

[12] Farshad Farshidfar, Aalim M Weljie, Karen A Kopciuk, Robert Hilsden, S Elizabeth McGregor, W Donald Buie, Anthony MacLean, Hans J Vogel, and Oliver F Bathe. A validated metabolomic signature for colorectal cancer: exploration of the clinical value of metabolomics. *Br. J. Cancer*, 115(7):848–857, 2016.

[13] Sakhamuri Sivakesava, Joseph Irudayaraj, and Demirci Ali. Simultaneous determination of multiple components in lactic acid fermentation using ft-mir, nir, and ft-raman spectroscopic techniques. *Process Biochemistry*, 37(4):371–378, 2001.

[14] Kathryn A Welsby. *Raman Spectroscopy of Blood Plasma : Immunological Applications in Prenatal Author :.* PhD thesis, Swansea University, 2016.

[15] Samuel J Simpkins, MI Pinto-Sanchez, Paul Moayyedi, Premysl Bercik, David G Morgan, Carolina Bolino, and Alexander C Ford. Poor predictive value of lower gastrointestinal alarm features in the diagnosis of colorectal

cancer in 1981 patients in secondary care. *Alimentary pharmacology & therapeutics*, 45(1):91–99, 2017.

[16] Cancer Research UK. Bowel cancer statistics, September 2016.

[17] Deepak Hariharan, A Saied, and HM Kocher. Analysis of mortality rates for pancreatic cancer across the world. *Hpb*, 10(1):58–62, 2008.

[18] Tom Wiggins, Sacheen Kumar, Sheraz R Markar, Stefan Antonowicz, and George B Hanna. Tyrosine, phenylalanine, and tryptophan in gastroesophageal malignancy: a systematic review. *Cancer Epidemiology and Prevention Biomarkers*, 2014.

Chapter 8

Conclusions and future outlook

The main goal of this work was the development and optimise a high-throughput methodology for serum RS for triaging colorectal referrals to the urgent suspected cancer (USC) pathway in primary care. The translation of serum RS has therefore driven the majority of this work towards standardised protocols for sample and data collection and data analysis. This section will summarise the main findings of the work thus far and will discuss the future direction of this project.

8.1 Development of method and apparatus for high throughput data collection and analysis

Simple, high throughput platforms for liquid and dry data acquisition were developed during this work by considering potential sources of experimental variation to spectra. It was found that serum provided a better sample choice than plasma for a triage based tool due to less sample variability upon freeze-thawing.

A high-throughput (HT) aluminium substrate was developed for dry data collection. A cooled stainless steel platform for liquid data collection was also developed. The measurement platforms were tested in a pilot study with a 60 patient cohort of matched cancer and control patients that were also sex and age matched. The platforms were tested for diagnostic capability as well as their practicality and inter-user variability. Table 8.1 shows a comparison of the HT platforms in terms of the calculated CV sensitivities, specificities, analysis times and the effects of inter-operator variability.

The dry 785 nm methodology yielded the most effective diagnostic results with the highest sensitivity, specificity and AUC. Therefore, within a research

Table 8.1: Comparison between calculated sensitivities, specificities, susceptibility to user variability and also total sample measurement time for one sample (including pipetting, drying, etc)

	Sensitivity (%)	Specificity (%)	User variable?	Total time (mins)
785 nm dry	83	83	Yes	80
785 nm Liquid	77	81	No	22.5
532 nm liquid	77	78	No	16

laboratory with one user, this method may be considered optimal. However, when extended to considering aspects of translation the dry methodology exhibited inter-user spectral variability which would potentially cause a large variation in diagnostic results. The liquid serum platform showed higher sensitivity and specificity with 785 nm excitation than with 532 nm excitation. Despite the liquid methodologies having a slightly lower sensitivity and specificity they are not affected by inter-user variability. Moreover the overall analysis time for the liquid methods is also quicker as there is no need to wait for the samples to dry.

Previous work has shown that in vibrational spectroscopic studies the preparation of a sample can affect spectra. For example, Lovergne et al showed that freeze-thaw cycles affect spectral variability in plasma samples within FTIR studies [1]. When using clinical samples, some sample sources may sometimes be available fresh and sometimes available after storage (freezing). To investigate the effect that freezing samples has on diagnostic capability serum samples were compared for both the dry and liquid HT measurement platforms for fresh samples and samples that had undergone a freeze-thaw cycle. One aliquot of each patient sample was used on day of collection for immediate Raman analysis (results presented above in table 8.1) and another was frozen at -80°C. After one month of storage, frozen samples were thawed at room temperature and analysed. Data from the samples that had undergone a freeze thaw cycle were subject to PLS-DA analysis. Diagnostic models were then calculated, cross validated and compared to the models calculated for fresh serum samples. Table 8.2 demon-

Table 8.2: Comparison between calculated sensitivities and specificities of the frozen samples, and the relative change in sensitivity and specificity due to freeze-thawing serum samples prior to analysis.

	Sensitivity (%)	Specificity (%)	Sensitivity change vs fresh (%)	Specificity change vs fresh (%)
785 nm dry	72	78	-11	-5
785 nm Liquid	79	77	+2	-4
532 nm liquid	77	77	0	+1

strates that the dry samples are affected more significantly by the freeze-thaw process and the liquid samples maintain a similarly high sensitivity, specificity and AUC as the fresh samples in Table 8.1. These results, coupled with discussions with local laboratory medicine teams demonstrate the strong motivation for the application of a liquid sample for analysis on the basis of both reproducibility and sample handling flexibility.

As well as the experimental sources of variation, patient demographics were also considered in the development of the platform. It was found that the fasting status of a patient caused significant spectral differences in regions which were diagnostically relevant. Therefore to maximise diagnostic capability only fasted patients should be used in serum RS studies. Furthermore, preliminary studies showed that sex and medications may also have a large baring on diagnostic capability when using a 785 nm system. Preliminary results showed that the 785 nm specta can be separated by patient sex.

8.1.1 Data analysis routines

The optimal methods of pre-processing spectral data for colorectal diagnostics was also considered. It was found that using a rolling circle filter (RCF) combined with a normalisation to the Phe peak provided rapid spectral processing that was applicable to both large and small datasets. RCF background subtraction combined with vector normalisation also produced equivalent results.

An automated application for complete data analysis was also developed towards creating a black box software package. The application was developed within MATLAB and has the ability to pre-process and analyse unknown spectra against a known diagnostic model to produce a standardised result which can then be interpreted. This allowed the complete processing of test spectra in just a few minutes after collection.

8.1.2 Further developments

To progress this work the platforms developed need to be tested on more than one instrument. The instrument calibrations can then be investigated further towards translation. It is envisaged that a triage tool would be based within a centralised laboratory that receive samples for analysis. This would allow maximised control over the system and environmental parameters of the samples. Therefore, as long as the system is transferable across systems within that laboratory this would be a large step towards translation of the technique.

The development of an automated spectral processing app contributed to the quick analysis of data. However, data still need to be transferred from the instrument in a raw format across to the software. Ideally, an integrated software for controlling data collection and analysis in one system needs to be developed for analysis efficiency. This would allow direct results to be generated out of a system to then feed results back to the clinical care team.

8.2 Clinical validation study

The diagnostic limits of the HT liquid platform were tested in a clinical validation study. Binary and non-binary models were tested. The maximum sensitivity and specificity achieved for a model was with a 785 nm binary cancer vs control model which showed a NPV at over $\geq$90%. The NPV of the testing set was 94.44% meaning for that cohort of patients 35% of the colonoscopy referrals could have been prevented if the test had been used to influence a GP's referral decision. Table 8.3 shows a comparison of the performance of the large binary cancer vs

control results to current methods of screening and diagnosis for CRC. The table shows that the maximum performance of the Raman CRC test was higher overall than the current USC pathway. This shows that the Raman test could work as a triage based tool for CRC improving on the current USC pathway. The high NPV of the Raman test is moving towards that of the FIT and colonoscopy. Furthermore, it shows that there is space for a serum based RS method to compensate for patients that are unable to complete or ineligible for other tests. For example a patient with rectal bleeding would not be able to have a FIT test but would be able to have a Raman based test. Additionally, colonoscopy has 5% miss rate for cancer and CTC has a 3.4% miss rate. The diagnostic values for the serum RS test are within these bounds with a miss rate of 5% in the binary cohort study.

Table 8.3: Comparison between the maximum performance of the liquid serum RS diagnostic model compared to current screening and diagnostic methods. Diagnostic vales taken from [2–6].

Test	Sens	Spec	PPV	NPV	Limitations
Liquid RS	93.8%	51.5%	48.4%	94.4%	Fasting required, inflammatory affects results
USC pathway	80.4%	47.2%	3.5%	99.0%	low numbers of confirmed CRC, AUC=0.65
FIT faecal test	93.3%	77.3%	11.2%	99.7%	n/a to patients with rectal bleeding
CT colonogram	89.0%	75.0%	n/a	99.9%	Small polyps (PPV 80%), 30% need colonoscopy too
Colonoscopy	95.0%	90.0%	2-11% (symptom dependent)	99.4%	Capacity, invasive, bowel prep req, completion rate just 57%

Multi-modal and non-binary models were tested in an attempt to improve diagnostic capability. However, the model with more than one laser excitation data showed little improvement for 785 nm models for a considerable increase in data acquisition time. The addition of patients with inflammatory diseases into diagnostic models also decreased the diagnostic capability.

8.2.1 Future work towards translation

To improve the diagnostic performance of the liquid serum RS platform for inflammatory patients there are a few potential work streams. Overall, patient numbers will need to be increased to maximise the performance of the diagnostic models.

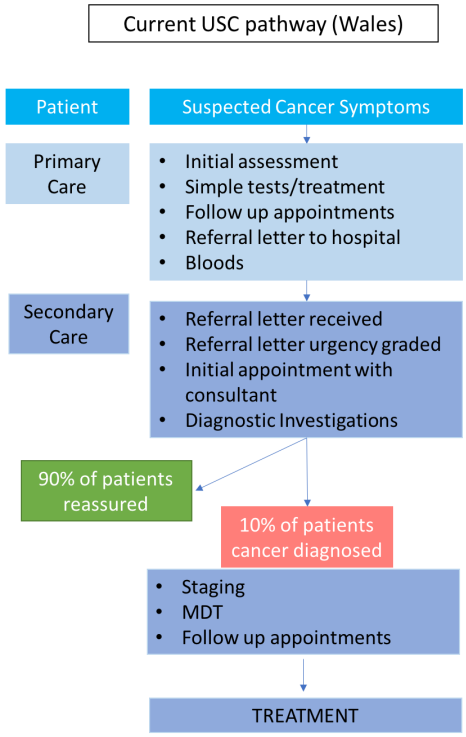
To reduce the number of equivocal results the testing datasets could be increased to five spectra for each patient. This would maximise the chances of getting a majority result from the model without taking too much more time for data collection. Furthermore, more patients can be included in the training model. An investigation into the need for a non-binary model would also need to be included. Another approach could be to separate the models into binary models and create a decision tree based algorithm with downstream selection i.e. the first model would distinguish between healthy and disease. The patients showing as disease would then be passed to a secondary binary model classifying between malignancy and benign diseases.

Figure 8.1 (a) shows the current USC pathway for CRC. Figure 8.1 (b) shows the proposed future use of a serum based RS test to triage colorectal referrals. Provided the Raman test continues to show good NPV in a primary care setting it is envisaged that the number of colonoscopy procedures could be reduced overall. With optimisation it is hoped that the rate of early reassurance could be improved up to 80% with a RS test. This would give patient reassurance sooner for those not in need of further investigations. The decreased steps of the pathway and earlier reassurance of patients will shorten the pathway overall. This has the potential to save money, decrease patient anxiety and also lead to the patients in most need being treated more quickly. However, the endpoint for the optimum performance

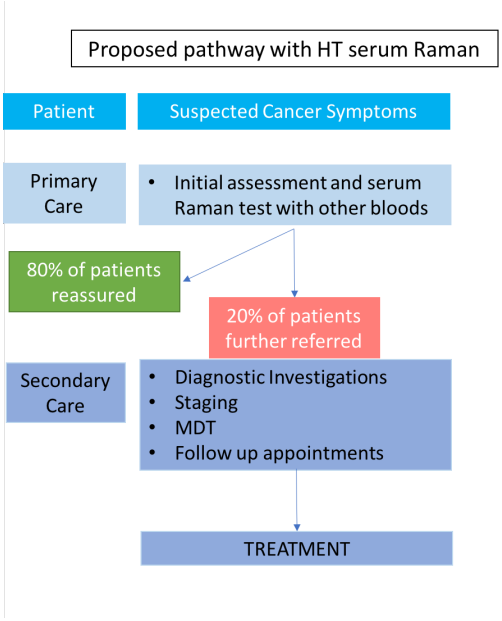
of a RS test may not be set by the maximum sensitivity and specificity of the test. It will be important in future studies to evaluate the point along the ROC curve in which the false negative rate of the test is minimised. This will ensure that patients with cancer are not missed by the test and incorrectly sent home. Therefore, it may be the case that the specificity of the test may have to be lowered to ensure maximum sensitivity. This would most likely result in a larger number false positive results potentially causing more referrals than necessary to secondary care (hence being less cost effective). A balance will have to be found for the RS test to minimise the false negative results to whilst also minimising the false positive results. This will be to maximise patient outcomes but also for the test to be cost effective. Therefore, a health economic evaluation will have to take place to balance the position along the ROC curve which the Raman test should sit. The health economic modelling evaluates the sensitivity and specificity of the RS test (in relation to the cost of the test) to metrics such as patient quality-adjusted life-years (QALY) and NHS cost effectiveness. A recent study has shown the potential of a spectroscopic test for brain cancer to be economically viable for the NHS in primary care at a cost of 50 at a sensitivities and specificities of 80% [7].

The models used throughout this thesis are based in MATLAB. The position of ROC curves in MATLAB are fixed to the output from the model training. However work is now ongoing to convert to an R programming language where it will be possible to move the threshold of the test along the ROC curve as required by the health economic and patient outcomes.

Further to the clinical validation study results shown in this thesis, analysis will need to be completed on a larger cohort of patients across recruitment sites. The aim would be for a patient cohort of many thousands. This would allow further data to be included in RF models.



(a)



(b)

Figure 8.1: Comparison between the current USC referral pathway (a) and a shortened USC pathway (b) by using Raman spectroscopy as a triage tool. Where HT is high throughput and MDT is a multi-disciplinary telecommunication meeting.

Exploration will also need to be made of the acceptability of a Raman based test within the proposed altered pathway. This will include the confidence levels of a GP in the test based on the preliminary work within this thesis.

If the future performance of the Raman spectroscopic test was sufficiently high it could also be considered as a screening test. It would therefore need to be compared to the current accepted screening method (FIT testing). For a true comparison to FIT a study would need to be performed on the same patient population as FIT results are reported on i.e. screening population. This would allow the creation of ROC curves and NPV/PPV values for Raman that would be comparable to FIT values as a screening method. It may also be useful as a combined tool for screening. Further work towards this is currently being planned within a combined Raman and FIT trial at Swansea University (CRaFT) and is due to commence recruitment May 2019.

The effects of treatment and applications to disease monitoring through cancer treatment could also be explored as a further application of the HT serum platform. The technique showed it was able to detect some differences in patients across treatment. Furthermore, there is potential for further work into other diseases. So far, a pilot study for pancreatic cancer shows high sensitivity and specificities of 80% and 95%.

The RF models generated in this work revealed spectral regions of interest that proved important for diagnostic capability. These 'spectral' biomarkers will need to be validated via other analytical methods such as GCMS. A validation of the spectral markers would provide a solid foundation for translation as it would be validated by an external technique. Performing GCMS on samples may also reveal other information not revealed via RS such as the specific metabolites changed between cancer and control samples. This would provide more information about the disease development. Although some preliminary work for this was attempted equipment failure meant this could not be included in this thesis.

Given the positive results in this thesis. The author believes that this work will move towards translation as a triage tool for CRC referrals in primary care. Therefore the final stage of future work would involve investigations into the regulatory approvals, ethics, and a proposed business plan to gain further investment

towards translation.

8.2.2 University spin out company

A spin-out company has been registered from the work produced in this thesis. The author is a founder and director of CanSense LTD. The company was registered in the UK on the 17th May 2018 and is focused on the triage of colorectal referrals from GPs in the UK.

8.3 Contributions and Publications

There have been various outputs from this thesis including conference contributions, publications, patents filed and also planned future publications.

8.3.1 Poster presentations

1. NHS QIPP Day 2015: Surface enhanced Raman spectroscopy and colorectal cancer: towards early diagnosis and personalised medicine.

2. Cancer Research Wales research day (2016): Raman spectroscopy and colorectal cancer.

3. CLIRSPEC summer school (2016) - Surface enhanced Raman spectroscopy for colorectal cancer detection.

4. SPEC (2016) - Raman spectroscopy and colorectal cancer: towards early detection and personalised medicine.

8.3.2 Oral contributions

1. CLIRCON (2017) - Raman Spectroscopy and colorectal cancer: Effect of sampling modality on diagnostic capability.

2. ICAVS9 (2017) - High throughput serum Raman spectroscopy to aid diagnosis of colorectal cancer.

3. SciX (2017) - Invited contribution - Raman spectroscopy to aid diagnosis of colorectal cancer - practical considerations.

8.3.3 Publications

1. Jenkins, C. A., Lewis, P. D., Dunstan, P. R., & Harris, D. A. (2016). Role of Raman spectroscopy and surface enhanced Raman spectroscopy in colorectal cancer. World Journal of Gastrointestinal Oncology, 8(5), 427438. http://doi.org/10.4251/wjgo.v8.i5.427

2. The methodology developed within this thesis for liquid serum Raman spectroscopy was protected via a filing of a worldwide patent. GB Pat App. No. 1704128.6 Method and apparatus for detecting colorectal cancer using Raman spectroscopy 6026P/GB. Filed 13 March 2016.

3. Jenkins, Cerys A., et al. "A high-throughput serum Raman spectroscopy platform and methodology for colorectal cancer diagnostics." Analyst 143.24 (2018): 6014-6024.

The planned publications resulting from this work as follows;

- Transforming the USC referral pathway for colorectal cancer using high throughput serum Raman spectroscopy. Planned submission (April 2019)

Bibliography

[1] L. Lovergne, P. Bouzy, V. Untereiner, R. Garnotel, M. J. Baker, G. Thiéfin, and G. D. Sockalingum. Biofluid infrared spectro-diagnostics: Pre-analytical considerations for clinical applications. *Faraday Discuss.*, 187:521–537, 2016.

[2] CJA Bowles, R Leicester, C Romaya, E Swarbrick, CB Williams, and O Epstein. A prospective study of colonoscopy practice in the uk today: are we adequately prepared for national colorectal cancer screening tomorrow? *Gut*, 53(2):277–283, 2004.

[3] Samuel J Simpkins, MI Pinto-Sanchez, Paul Moayyedi, Premysl Bercik, David G Morgan, Carolina Bolino, and Alexander C Ford. Poor predictive value of lower gastrointestinal alarm features in the diagnosis of colorectal cancer in 1981 patients in secondary care. *Alimentary pharmacology & therapeutics*, 45(1):91–99, 2017.

[4] Hooi C Ee, James B Semmens, Neville E Hoffman, et al. Complete colonoscopy rarely misses cancer. *Gastrointestinal endoscopy*, 55(2):167–171, 2002.

[5] PFC Lung, D Burling, L Kallarackel, J Muckian, R Ilangovan, A Gupta, M Marshall, P Shorvon, S Halligan, G Bhatnagar, et al. Implementation of a new ct colonography service: 5 year experience. *Clinical radiology*, 69(6):597–605, 2014.

[6] Ming Ming Zhu, Xi Tao Xu, NIE Fang, Jin Lu Tong, Shu Dong Xiao, and Zhi Hua Ran. Comparison of immunochemical and guaiac-based fecal occult blood test in screening and surveillance for advanced colorectal neoplasms: A meta-analysis. *Journal of digestive diseases*, 11(3):148–160, 2010.

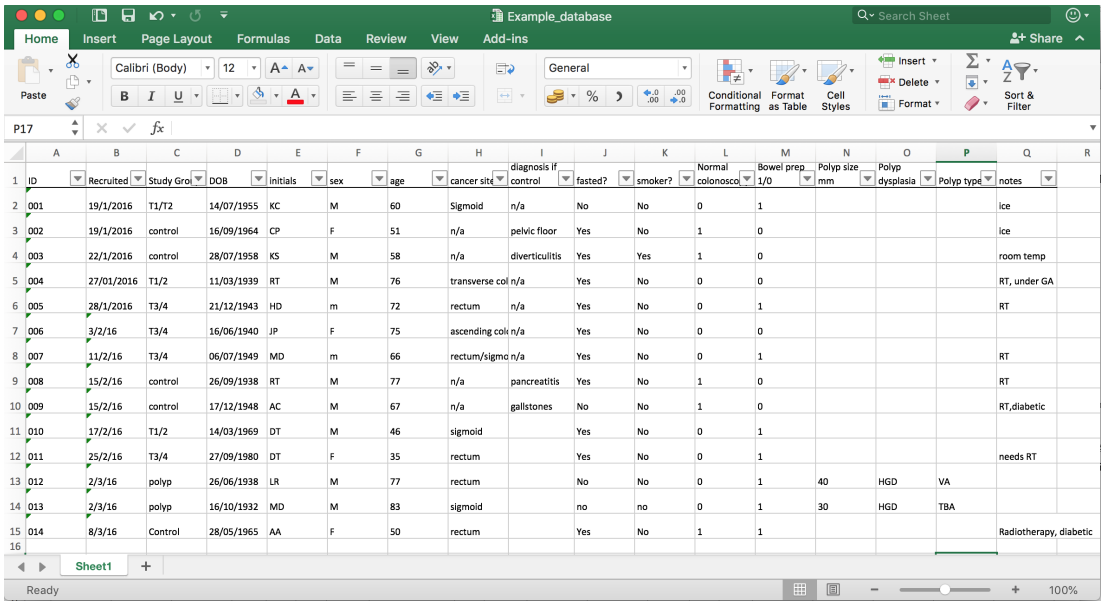
[7] Ewan Gray, Holly J Butler, Ruth Board, Paul M Brennan, Anthony J Chalmers, Timothy Dawson, John Goodden, Willie Hamilton, Mark G Hegarty, Allan James, Michael D Jenkinson, David Kernick, Elvira Lekka, Laurent J Livermore, Samantha J Mills, Kevin O Neill, David S Palmer,

Babar Vaqas, and Matthew J Baker. Health economic evaluation of a serum-based blood test for brain tumour diagnosis : exploration of two clinical scenarios. 2018.

Appendix A

Patient metadata database

During the course of this work serum samples were collected, aliquotted and banked to form a small serum biobank. Patient demographics were built into a database using Microsoft excel. Figure A.1 shows a representative example of the database with the categories of patient metadata that was collected. The final column (notes) was used to collect any extra information that may have been deemed clinically relevant including medication, if the patient was having radio therapy (RT) and if the patient had any other co morbidities e.g. diabetic, asthma, previous cancer etc.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	ID	Recruited	Study Gro	DOB	initials	sex	age	cancer site	diagnosis if control	fasted?	smoker?	Normal colonosco	Bowel prep 1/0	Polyp size mm	Polyp dysplasia	Polyp type	notes
2	001	19/1/2016	T1/T2	14/07/1955	KC	M	60	Sigmoid	n/a	No	No	0	1				ice
3	002	19/1/2016	control	16/09/1964	CP	F	51	n/a	pelvic floor	Yes	No	1	0				ice
4	003	22/1/2016	control	28/07/1958	KS	M	58	n/a	diverticulitis	Yes	Yes	1	0				room temp
5	004	27/01/2016	T1/2	11/03/1939	RT	M	76	transverse col	n/a	Yes	No	0	0				RT, under GA
6	005	28/1/2016	T3/4	21/12/1943	HD	m	72	rectum	n/a	Yes	No	0	1				RT
7	006	3/2/16	T3/4	16/06/1940	JP	F	75	ascending col	n/a	Yes	No	0	0				
8	007	11/2/16	T3/4	06/07/1949	MD	m	66	rectum/sigmo	n/a	Yes	No	0	1				RT
9	008	15/2/16	control	26/09/1938	RT	M	77	n/a	pancreatitis	Yes	No	1	0				RT
10	009	15/2/16	control	17/12/1948	AC	M	67	n/a	gallstones	No	No	1	0				RT,diabetic
11	010	17/2/16	T1/2	14/03/1969	DT	M	46	sigmoid		Yes	No	0	1				
12	011	25/2/16	T3/4	27/09/1980	DT	F	35	rectum		Yes	No	0	1				needs RT
13	012	2/3/16	polyp	26/06/1938	LR	M	77	rectum		No	No	0	1	40	HGD	VA	
14	013	2/3/16	polyp	16/10/1932	MD	M	83	sigmoid		no	no	0	1	30	HGD	TBA	
15	014	8/3/16	Control	28/05/1965	AA	F	50	rectum		Yes	No	1	1				Radiotherapy, diabetic
16																	

Figure A.1: Example of the patient meta data database created during this work.

In total, participants in the studies within this thesis were recruited into 7 research groups. The groups were as follows:

1. CRC - T 3/4
2. CRC - T 1/2

3. Polyps

4. Control - healthy

5. Control - non-disease

6. other cancers

For patient selection for each study e.g. fasted vs non-fasted samples the filter function was used to isolate patients that were eligible for the study. Spectral data from the corresponding patient ID were then used for analysis.

The database at the time of completion of this thesis contained the information of over 300 patients across the study groups within this work. It is envisaged that sample collection will continue to build the sample database.

Appendix B

Spectral normalisation

Consider a d-dimensional raw spectral dataset Y = $\{X1, X2, ..., Xn\}$. The spectral intensities in the data matrix is then a $n \times d$ matrix given by:

$$X1, X2, \ldots, Xn = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nd} \end{bmatrix} \tag{B.0.1}$$

The vector normalisation for a given spectral dataset is given by:

$$X_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \tag{B.0.2}$$

where $\bar{x}_j$ is the mean spectral intensity for a spectrum and σ_j is the standard deviation of the jth wavenumber of the spectrum. The standard deviation is given by:

$$\sigma_j = \sqrt{(x_{ij} - \bar{x}_j)^2}. \tag{B.0.3}$$

The min-max normalisation is defined as:

$$MM(X_{ij}) = \frac{X_{ij} - X_{min}}{X_{max} - X_{min}}. \tag{B.0.4}$$

Appendix C

Matlab code

This appendix includes the detailed codes used for the optimised data analysis routine for spectral data. The code for the front end GUI is not included in this as the underlying processing codes were just packaged by creating an app in the MATLAB environment. The codes only difference between Raman and FTIR spectral data was that FTIR data had different dimensions and that they were not standardised as they all have a consistent x-axis.

C.1 Data importing and preprocessing

C.1.1 Import

```
clear; clc; close all; format longg;


[FileName,PathName,FilterIndex] = uigetfile('*.txt','
    Select the txt file','MultiSelect','on');

%SpecNo = length(FileName);
imputfiles = [];
meandat=[];% Pre-allocate empty matrix.
%%Data = [];% Pre-allocate empty matrix.
%spectrum_sample_mean=zeros(SpecNo,samples);

%% Creates a matrix with all the raw data
for i = 1:length(FileName)
```

```
14        str   = FileName{1,i};     % Create temp string from
              user data.
15      A = dlmread(str,'\t');    % Read said file.\
16      imputfiles = cat(3, imputfiles, A); % Concatate to
          existing matrix.
17  end
18
19  %% Creates a new matrix with the averaged sample data
20  % Prompt user for number of files to be read.
21  disp('Enter number of patients');
22  sample_num = input('patient_num = ');
23  disp('Enter number of repeats');%for dried samples, number
       of repeats
24  rep_num = input('rep_num = ');
25
26  %Code to average spectra
27  M=repmat(rep_num,1,sample_num);
28
29  split = mat2cell(imputfiles,[1015],[2],[M]); %make sure
      the 1015 matches the length of the data
30  for i=1:sample_num
31       means = mean(split{:,:,i},3);
32       meandat = cat(3,meandat,means);
33  end
34
35  %% Flipping to be low-high
36
37  %This checks the data is in low-high format and flips it
      if it's the wrong
38  %way.
39
```

```
w1= meandat(1,1,1);
w2 = meandat(end,1,1);
if (w1>w2)
    Data = flipud(meandat);
else Data = (meandat);

end

disp('Data is now in lo-hi format saved in "Data".')

clear imputfiles means M
clear A
clear FilterIndex
clear PathName
clear w1
clear w2
clear str
clear split
```

C.1.2 Standardisation

```
function [Xq,VqAll] = Shift_Interp3_clean(Data, sample_num
    );
%% Hopefully this function will fit the peaks to the 4th
    order polynomial
% of the phenyalalanine peak
%  Shift_intrp 3

% This function will standardise wavenumber axis to the
    highest peak
%
%  Input -
```

```
% Data : Spectra with intensoty and wavenumvs
% sample_num : numer of spectra
%
% Output -
% Xq : new x axis
% VqAll : wavenumber corrected spectra
% Copywrite Cerys Jenkins 2017

%% Preamble
intensity = Data(:,2,:); %Takes intensity and wavenos
    from imported files
waveno = Data(:,1,:);

V=[]; %% Pre assaigned for the intensities
X=[];  %% Pre assigned for the wavenumbers

for i = sample_num:-1:1 % This loop just fills V with the
    spectral intensities
X=cat(2,waveno(:,1,i),X); % and X with the wavenumber axis
V = cat(2,intensity(:,1,i),V);
end


SignalPeak = [];
Shifted_int = [];
Shifted_W=[];
RP = 1004; %%% This is where Phylalanine should be make
    sure this is correct
Signal_max= [];

for i = 1:sample_num
```

```
37   YI = V(:,i);
38  MZ2 = X(:,i);
39  [val, idx] = max(YI(300:380),[],1); % Find the maximim
        peak in a spectrum and the index
40  idx2 = idx+299;
41  Signal_max = cat(1,Signal_max,idx2);
42  SignalP = MZ2(idx2); % convert this so it's the right
        place
43  SignalPeak = cat(1,SignalPeak,SignalP);
44
45  end
46
47  for k = 1:sample_num
48      SP = SignalPeak(k);
49      MZ  = X(:,k);
50      Y = V(:,k);
51
52      SignalPeak_Points = [Signal_max(k)-2,Signal_max(k)-1,
            Signal_max(k),Signal_max(k)+1,Signal_max(k)+2];
53
54      Signal_Peak_Int = [Y(SignalPeak_Points(1)),Y(
            SignalPeak_Points(2)),Y(SignalPeak_Points(3)),Y(
            SignalPeak_Points(4)),Y(SignalPeak_Points(5))];
55      SignalPeak_wavenum = [MZ(SignalPeak_Points(1)),MZ(
            SignalPeak_Points(2)),MZ(SignalPeak_Points(3)),MZ(
            SignalPeak_Points(4)),MZ(SignalPeak_Points(5))];
56      [p,S,mu] = polyfit(SignalPeak_wavenum,Signal_Peak_Int
            ,4);
57      x1 = SignalPeak_wavenum(1):.1:SignalPeak_wavenum(5);
58      y1 = polyval(p,x1,S,mu);
59      [val3,idx3] = max(y1);
```

```
60
61      bandpos = SignalPeak_wavenum(1) + (0.1*(idx3-1));
62      shift = RP - bandpos;
63      Xshift = MZ-(shift);
64      Yout = interp1(MZ,Y,Xshift,'pchip');
65      Shifted_W = cat(2,Shifted_W,Xshift);
66      Shifted_int = cat(2,Shifted_int,Yout);
67  end
68
69  VqAll = [];
70  %Xq= (611:1.09064039:1717);   %% Making new set of X
        coords
71  %Xq= (611:1:1718);
72  Xq= (611.6:1.09:1717);
73
74  for k = 1:sample_num
75  X2 = Shifted_W(:,k).';      % Transpose matricies so they
        are in the right format
76  V2 = Shifted_int(:,k).';
77  Vq = interp1(X2,V2,Xq,'linear');
78  VqAll = cat(1,VqAll,Vq);
79  end
80  end
```

C.1.3 Rolling circle filter

```
1  function [corrected,baseline] = RCF_multi(wavenumber,
       intensity,radius,sample_num);
2   %% Rolling circle filter for multiple spectra
3  % 1/9/2016
4  % updated 7/8/2017
5  % Copyright Cerys Anne Jenkins
```

```
6
7  %wavenumber - x axis
8  % intensity - matirix of spectra
9  % radius - circle radius size
10 % sample_num - number of spectra
11
12 % Output: baseline corrected spectra
13
14 %% create the circle
15 x= linspace(-radius,radius,(2*radius+1));
16 for i = 1:length(x)
17 y = sqrt((radius.^2)-(x.^2));
18 end
19
20 %% making some fabricated data to put at the start and end
       of the data
21 % So that the circle starts rolling from the start of the
      data
22
23 for i = 1:length(x) %% makes the fake start data depending
       on the size of the radius of the circle
24 fakestartxdata(i) = wavenumber(1)-i;
25 end
26 fakexcoords = fliplr(fakestartxdata);%flips these lr so
      that we start with the lowest value
27 for i=1:length(x)
28     fakeendxdata(i)= wavenumber(:,end)+i;
29 end
30 totalxdata = [fakexcoords,wavenumber,fakeendxdata];
31 fakeycoords = [];
32 for i = 1:sample_num
```

```
33 fakeycoord = max( intensity(i,:) ,[],2); % minds the max of
       each spectrum
34 fakeycoords = cat(1,fakeycoords,fakeycoord);
35 totalfakeycoords(:,i) = repmat(fakeycoords(i),length(
       fakexcoords),1);
36 totalydata(i,:) = horzcat(totalfakeycoords(:,i).',
       intensity(i,:),totalfakeycoords(:,i).');
37 end
38 testdatasets = [];
39 for i = 1:sample_num
40 testdataset = [totalxdata.' totalydata(i,:).'];
41 [testdatasets] = cat(3,testdatasets,testdataset);
42 end
43
44 %% ASPECT RATIO
45 %Calculating the aspect ratio for the circle so that it
       scales to the size of the data
46 for i = 1:sample_num
47 Aspectratio(:,i) = 2*((max(intensity(i,:))/(max(wavenumber
       )-min(wavenumber) )));
48 AspectYcoords(i,:) = y*Aspectratio(i); %Adjust the y
       coords of the circle by * Apsect ratio for each
       spectrum
49 CircleYstartpoint(i,:) = AspectYcoords(i,:) + fakeycoords(
       i) - AspectYcoords(i,ceil((length(x)/2)));
50 end
51 CirclestartpointX = x+totalxdata(1)+radius;
52 %CircleStartY = zeros(length(x),2,sample_num);
53 CircleStartY = [];
54 for i = 1:sample_num
55 Circlestartpos = [CirclestartpointX.'  CircleYstartpoint(i
```

```
    ,:).'];
[CircleStartY] = cat(3,CircleStartY,Circlestartpos);
end
%% DataInCirc
%%% This section of code loops over the spectrum and finds
    all of the
%%% datapoints that are inside each circle radius.

 for j = 1:sample_num
loopsize = length(testdatasets) - length(x);
b = length(x);

for i = 1:loopsize
    goal(:,:,i) = testdatasets(i:i-1+b,:,j);
end
for i = 1: length(goal)
dist(:,i) = goal(:,2,i) - CircleStartY(:,2,j);
end
distance = dist.';
[Min ,Index] = min(distance,[],2);
xbit = [];
for i = 1:length(Min)
xbit(i,:) = goal(Index(i),:,i);

end

XY = xbit;
Unique = unique(XY,'rows');
Xaxis = Unique(:,1);
Yaxis = Unique(:,2);
%Xquery = (610.07:1.09262:1719);
```

```
85
86 baseline = interp1(Xaxis,Yaxis,wavenumber,'pchip'); % this
       way interpolates a background
87
88 %[p,S,mu] = polyfit(Xaxis,Yaxis,2); %this way fits a
       polynomial to the
89 % background points
90 %x1 = wavenumber;
91 %baseline2 = polyval(p,x1,S,mu);
92 %xbits = xbit(:,1);
93 %ybits = ybit(:,2);
94 %baselinefit = csaps(xbits,ybits)% this fits a cubic
       spline smoothing to all of the bg points from the RCF
       this seems to give best fit
95 %fnplt(baseline3) % Plots the function from the spline
       smoothing
96 %points = fnplt(pp); this doesn't plot anything but gives
       you the points
97 % it would have plotted in fnplt
98 %baseline3 = fnval(baselinefit,wavenumber);
99
100
101 corrected(j,:) = intensity(j,:) - baseline;
102 %corrected2(j,:) = intensity(j,:) - baseline2;
103 %corrected3(j,:) = intensity(j,:) - baseline3;
104
105 corrected(corrected<0)=0;  % this forces any negative
       values to zero
106 %corrected2(corrected2<0)=0;
107
108  end
```

```
109 end
```

C.1.4 Derivatives

```
1  function [SG0,SG1,SG2] = cj_deriv_func(wavenumber,
       intensity, smoothing_pts)
2  % Calculates the first derivative of intensity values
3  %
4  %   Usage: output = firstderiv(wavenumber, intensity,
       smoothing_pts);
5  %   Where
6  %       wavenumber - vector of wavenumbers (x values)
7  %       intensity  - matrix of intensities (y values in
       COLUMNS)
8  %       smoothing_pts - number of points to smooth data by
       . This MUST be
9  %           an odd number and should be quite small - 3,
       5, 7 etc
10 %       output - matrix of first derivative data (in
       columns)
11 %
12 % This calculates a first derivative while performing a
       Savitzky-Golay
13 % smooth at the same time. The data should be smoothed to
       prevent noise in
14 % the data from swamping the result. The smoothing_pts
       value determines the
15 % degree of smoothing - a bigger value mean more smoothing
       .
16 %
17 % Note that smoothing means you lose points from either
       end of the data.
```

```
18 % For example; if you have a 7 point smooth you will lose
       3 points from
19 % either end of the data. Here these points have been set
       to zero so that
20 % the length of the output is the same as the length of
       the input.
21 %
22 %   (c) Alex Henderson Dec 2007
23 %
24 % Updated and edited to do second derivatives Cerys
       Jenkins
25 %
26
27
28
29
30 if rem(smoothing_pts,2) == 0
31     error('smoothing_pts must be an odd number');
32 end
33
34 N = 4;                  % Order of polynomial fit
35 F = smoothing_pts;      % Window length
36 [b,g] = sgolay(N,F);    % Calculate S-G coefficients
37
38 x=wavenumber;
39 y = intensity;
40
41 dx = (x(end)-x(1))/length(x);
42
43 HalfWin  = ((F+1)/2) -1;
44 SG0 = zeros(size(y));
```

```
45 SG1=zeros(size(y));
46 SG2=zeros(size(y));
47 [rows,cols]=size(y);
48 for r = 1:rows
49     for n = (F+1)/2 : cols-(F+1)/2,
50         SG0(r,n)= dot(g(:,1), y(r,(n - HalfWin): (n +
               HalfWin)));
51         % 1st differential
52         SG1(r,n) =   dot(g(:,2), y(r, (n-HalfWin) : (n+
               HalfWin)));
53          % 2nd differential
54         SG2(r,n) = 2*dot(g(:,3)', y(r,(n - HalfWin):( n +
               HalfWin))');
55     end
56
57 end
58
59 SG1 = SG1/dx;           % Turn differential into derivative
60 SG2 = SG2/(dx*dx);      % and into 2nd derivative
61
62 % put into output
63 SG1 = SG1;
64 SG2 = SG2;
65 SG0 = SG0;
```

C.1.5 Rubber-band baseline subtraction (FTIR specific, not self produced)

```
1 %> @ingroup maths
2 %>@file
3 %>@brief Convex Polygonal Line baseline correction
4 %>
```

```
5  %> This was inspired on OPUS Rubberband baseline
       correction (RBBC) [1].
6  %>
7  %> Stretches a convex polygonal line whose vertices touch
       troughs of x
8  %> without crossing x (see below).
9  %>
10 %> This one is parameterless, whereas OPUS RBBC asks for a
        number of points.
11 %>
12 %> @image html rubberlike_explain.png
13 %>
14 %> <h3>References</h3>
15 %>    [1] Bruker Optik GmbH, OPUS 5 Reference Manual.
       Ettlingen: Bruker, 2004.
16 %>
17 %> @sa demo_pre_bc_rubber.m
18 %
19 %> @param X [@ref no]x[@ref nf] matrix whose rows will be
       individually baseline-corrected
20 %>
21 %> @return @em [Y] or @em [Y, L] Where @em L are the
       baselines
22 function varargout = bc_rubber(X)
23
24 %msgstring = nargoutchk(1, 2, nargout);
25 %msgstring = nargoutchk(1, 2);
26 %if ~isempty(msgstring)
27 %    error(msgstring);
28 %end
29
```

```
30
31 [no, nf] = size(X);
32
33 Y = zeros(no, nf);
34 L = zeros(no, nf);
35
36 for i = 1:no
37     if nf > 0
38         l = [];
39         x = X(i, :);
40         if length(x) > 1
41             l2 = rubber(x);
42         else
43             l2 = [];
44         end;
45         l = [x(1) l2];
46
47         Y(i, :) = x-l;
48         L(i, :) = l;
49     end;
50 end;
51
52
53
54 if nargout == 1
55     varargout = {Y};
56 elseif nargout == 2
57     varargout = {Y, L};
58 end;
59
60 %> @cond
```

```
61 %
```

```
62 % returns a "rubber" vector with one element less than the
       length of x
63 function y = rubber(x)
64
65 nf = length(x); % number of points
66
67 l = linspace(x(1), x(end), nf);
68
69 xflat = x-l;
70 [val, idx] = min(xflat);
71 if ~isempty(val) && val < 0
72     y = [rubber(x(1:idx)), rubber(x(idx:end))];
73 else
74     y = l(2:end);
75 end;
76 %> @endcond
```

C.1.6 Iterative polynomial baseline subtraction (Written by G.A.Lloyd)

```
1 function [spectra,y_hat]=removeBaseline(x,y,order,pos,
      max_iter,binning)
2 % Function to correct the baseline of a Raman spectrum
3 % by Gavin Rhys Lloyd 30.04.14
4 % REF: Applied Spectroscopy, Volume 57, Issue 11,Pages 320
      A-340A and 1317-1453 (November 2003) , pp. 1363-1367(5)
5 % Modified 07.10.2016 GL - tidied up code / comments
6 %
7 % INPUTS
```

```
8  % x = wavenumbers
9  % y = spectrum
10 % order = integer representing polynomial order of
       baseline
11 % pos = true / [false] to force -ve spectral values to be
       0 in the corrected spectrum
12 % max_iter = [100] maxium number of iterations
13 % binning = [1] wavenumber binning to apply
14 %
15 % OUTPUTS
16 % spectra = corrected spectrum
17 % y_hat = estimated baseline
18 %
19
20 tol=1e-6;
21 cont=true;
22 counter=0;
23 x_orig=x;
24 y_orig=y;
25
26 x=x(1:binning:length(x));
27 y=y(1:binning:length(y));
28
29 while cont
30     if counter>max_iter
31         cont=false;
32         disp('max number of iterations reached');
33     end
34     % fit polynomial to the data
35     [na,y_hat,b]=fitpoly(x,y,order,[]);
36
```

```
37      % replace data with min of poly and data
38      y=min([y_hat(:),y(:)]');
39
40      % check for convergence...
41      if (y-y_hat')*(y-y_hat')'<tol
42          cont=false;
43      end
44      counter=counter+1;
45  end
46  [na,y_hat,b,D]=fitpoly(x,y,order,x_orig);
47  spectra=(y_orig-y_hat');
48  if pos
49      spectra(spectra<0)=0;
50  end
51
52  function [x_hat,y_hat,b,D]=fitpoly(x,y,order,x_new)
53  % function to fit a polynomial to data
54  % by Gavin Rhys Lloyd 14.01.10
55  %
56  %
57  if nargin<4
58      x_new=x;
59  end
60  if isempty(x_new)
61      x_new=x;
62  end
63  m=mean(x);
64  my=mean(y);
65  y=y(:);
66  x=x(:)-m;
67
```

```
68 p=[];
69 for i=0:order
70     p(:,i+1)=x.^i;
71 end
72 D=[ones(size(x)),p];
73 b=pinv(D)*y; % LSQ coefficients
74
75 x_new=x_new(:)-m;
76 p=[];
77 for i=0:order
78     p(:,i+1)=x_new.^i;
79 end
80 D=[ones(size(x_new)),p];
81 y_hat=D*b; % apply coefficients
82 x_hat=x;
```

C.1.7 Vector normalisation

```
1  function output = vecnorm_cj(input)
2  [rows,cols]=size(input);
3
4
5
6  for col = 1:cols
7      input(:,col) = input(:,col);      ([n,m])
8  end
9
10 squares = input.^2;              % square of each absorbance
       ([n,m])
11 sum_of_squares = sum(squares, 2);        % sum of the
      squares along the rows ([n,1])
12
```

```
13 divisor = sqrt(sum_of_squares);            % ([n,1])
14
15 for col = 1:cols
16     output(:,col) = input(:,col)./divisor;    % divide the
       data by the vector length ([n,m])
17 end
```

C.1.8 Normalisation to the phenylalanine peak/min/max (Original code by Royston Goodacre)

```
1  function [basl] = scaleM(M)
2  %[basl] = scaleM(M)
3  %scales so min and max are between 0 and 1
4  %takes matrix M and makes basl
5  %
6  % Copyright (c) 1997, Royston Goodacre
7  %
8
9
10 [rows,cols]=size(M);
11 basl=zeros(rows,cols);
12
13 for i=1:rows
14     Mmin=min(M(i,:));
15     Mmax=max(M(i,:));
16     for j=1:cols
17         basl(i,j)=(M(i,j)-Mmin)/(Mmax-Mmin);
18     end
19 end
```

C.2 Data analysis

C.2.1 RF machine learning code

```
function [trainedClassifier_polyp, validationAccuracy,AUC,
    validationScores] = trainClassifier_polyp(trainingData)
% [trainedClassifier, validationAccuracy] =
    trainClassifier(trainingData)
% returns a trained classifier and its accuracy. Use the
% generated code to automate training the same model with
    new data, or to
% learn how to programmatically train models.
%
%  Input:
%      trainingData: a matrix with the same number of
    columns and data type
%       as imported into the app.
%
%  Output:
%      trainedClassifier: a struct containing the trained
    classifier. The
%       struct contains various fields with information
    about the trained
%       classifier.
%
%      trainedClassifier.predictFcn: a function to make
    predictions on new
%       data.
%
%      validationAccuracy: a double containing the
    accuracy in percent.
```

```
20 %
21 %       AUC: Area under calculated ROC curve for positive
       class
22 %
23 %       validationScores: Used to calculate ROC curve if
       needed to plot
24 %       later
25 % Use the code to train the model with new data. To
       retrain your
26 % classifier, call the function from the command line with
        your original
27 % data or new data as the input argument trainingData.
28 %
29 % For example, to retrain a classifier trained with the
       original data set
30 % T, enter:
31 %   [trainedClassifier, validationAccuracy] =
       trainClassifier(T)
32 %
33 % To make predictions with the returned 'trainedClassifier
       ' on new data T2,
34 % use
35 %   yfit = trainedClassifier.predictFcn(T2)
36 %
37 % T2 must be a matrix containing only the predictor
       columns used for
38 % training.
39
40 % Extract predictors and response
41 % This code processes the data into the right shape for
       training the
```

```
42 % model.
43 % Convert input to table
44 inputTable = array2table(trainingData, 'VariableNames', {'
       column_1', 'column_2', 'column_3', 'column_4', '
       column_5', ...});
45
46 predictorNames = {'column_1', 'column_2', 'column_3', '
       column_4', 'column_5', ...};
47 predictors = inputTable(:, predictorNames);
48 response = inputTable.column_1016;
49 isCategoricalPredictor = [false, false, false, false,
       ...];
50
51 % Train a classifier
52 % This code specifies all the classifier options and
       trains the classifier.
53 template = templateTree(...
54     'MaxNumSplits', 351);
55 classificationEnsemble = fitcensemble(...
56     predictors, ...
57     response, ...
58     'Method', 'Bag', ...
59     'NumLearningCycles', 499, ...
60     'Learners', template, ...
61     'ClassNames', [0; 1]);
62
63 % Create the result struct with predict function
64 predictorExtractionFcn = @(x) array2table(x, '
       VariableNames', predictorNames);
65 ensemblePredictFcn = @(x) predict(classificationEnsemble,
       x);
```

```
66 trainedClassifier_polyp.predictFcn = @(x)
       ensemblePredictFcn(predictorExtractionFcn(x));
67
68 % Add additional fields to the result struct
69 trainedClassifier_polyp.ClassificationEnsemble =
       classificationEnsemble;
70 trainedClassifier_polyp.About = 'This struct is a trained
       model exported from Classification Learner R2017b.';
71 trainedClassifier_polyp.HowToPredict = sprintf('To make
       predictions on a new predictor column matrix, X, use: \
       n  yfit = c.predictFcn(X) \nreplacing ''c'' with the
       name of the variable that is this struct, e.g. ''
       trainedModel''. \n \nX must contain exactly 1015
       columns because this model was trained using 1015
       predictors. \nX must contain only predictor columns in
       exactly the same order and format as your training \
       ndata. Do not include the response column or any
       columns you did not import into the app. \n \nFor more
       information, see <a href="matlab:helpview(fullfile(
       docroot, ''stats'', ''stats.map''), ''
       appclassification_exportmodeltoworkspace'')">How to
       predict using an exported model</a>.');
72 trainedClassifier_polyp.imp = predictorImportance(
       classificationEnsemble);
73 % Extract predictors and response
74 % This code processes the data into the right shape for
       training the
75 % model.
76 % Convert input to table
77 inputTable = array2table(trainingData, 'VariableNames', {'
       column_1', 'column_2', 'column_3', 'column_4', '
```

```
    column_5',...});

predictorNames = {'column_1', 'column_2', 'column_3', '
    column_4', 'column_5', ...};
predictors = inputTable(:, predictorNames);
response = inputTable.column_1016;
isCategoricalPredictor = [false, false, false, false,
    false,...];

% Perform cross-validation
partitionedModel = crossval(trainedClassifier_polyp.
    ClassificationEnsemble, 'KFold', 5);

% Compute validation predictions
[validationPredictions, validationScores] = kfoldPredict(
    partitionedModel);

% Compute validation accuracy
validationAccuracy = 1 - kfoldLoss(partitionedModel, '
    LossFun', 'ClassifError');
% Fit an ROC Curve based on the cross validation
[X,Y,T,AUC,OPTROCPT,suby,subnames] = perfcurve(response
    ,...
validationScores(:,2),1);
%Compute the confusion matrix
conf = confusionmat(response,validationPredictions);

trainedClassifier_polyp.conf = conf;

% To plot the ROC
plot_ROC(X,Y,OPTROCPT(1),OPTROCPT(2))
```

```
function plotROC(X1, Y1, X2, Y2)
%CREATEFIGURE1(X1, Y1, X2, Y2)
%  X1:  X
%  Y1:  Y
%  X2:  OPTROCPT(1)
%  Y2:  OPTROCPT(2)

%  Auto-generated by MATLAB on 16-Aug-2017 16:32:09

% Create figure
figure1 = figure;

% Create axes
axes1 = axes('Parent',figure1);
hold(axes1,'on');

% Create plot
plot(X1,Y1,'DisplayName','ROC Curve','LineWidth',4);

% Create plot
plot(X2,Y2,'DisplayName','Current Classifier','Marker','o'
    ,'LineWidth',3,...
     'LineStyle','none',...
     'Color',[1 0 0]);

% Create xlabel
xlabel('False positive rate');

% Create title
title('ROC Curve for detecting Polyps');

```

```
% Create ylabel
ylabel('True positive rate');

box(axes1,'on');
% Create legend
legend1 = legend(axes1,'show');
set(legend1,...
    'Position',[0.695535714285714 0.809523809523811
      0.197321428571429 0.0630952380952381]);

% Create textbox
annotation(figure1,'textbox',...
    [0.459928571428571 0.483333333333334 0.132928571428571
      0.047619047619048],...
    'String',{'AUC ='},...
    'FitBoxToText','off');
```

C.3 PLS-DA

The Lib-pls 9.2 package was used and edited for PLS-DA analysis. Available at http://www.libpls.net/.

Appendix D

Additional baseline studies

D.1 Whole blood Raman spectrum

Figure D.1 is an example of a liquid whole blood Raman spectrum taken with the 785nm laser line. The blood coagulated during the measurement, after coagulation spectra could not be obtained as the signal intensity off the sample saturated.

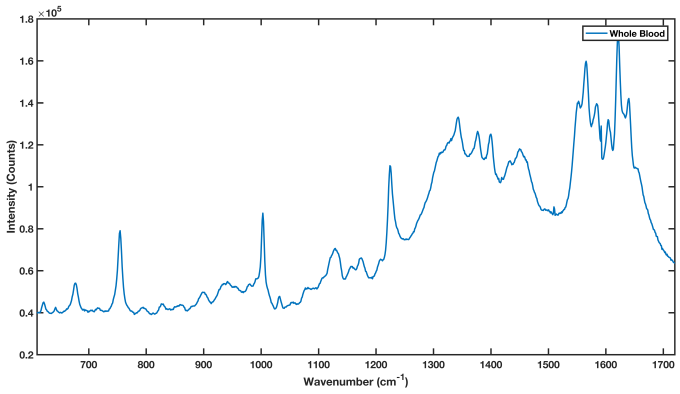


Figure D.1: Example of liquid whole blood Raman spectrum. The blood is said to be 'liquid' however during the measurement time the blood coagulated.

D.2 Dry substrate comparison

A dry substrate comparison was conducted below. The CaF disk response was much higher compared to the expected response. On inspection, the CaF disk tested had been 'cleaned' prior to the measurement by another lab user. This caused degradation of the slide and hence a large spectral response. Despite this it shows that the expensive slides would not be acceptable for translation due to the high cost per slide and lack or cleanability/re-usability.

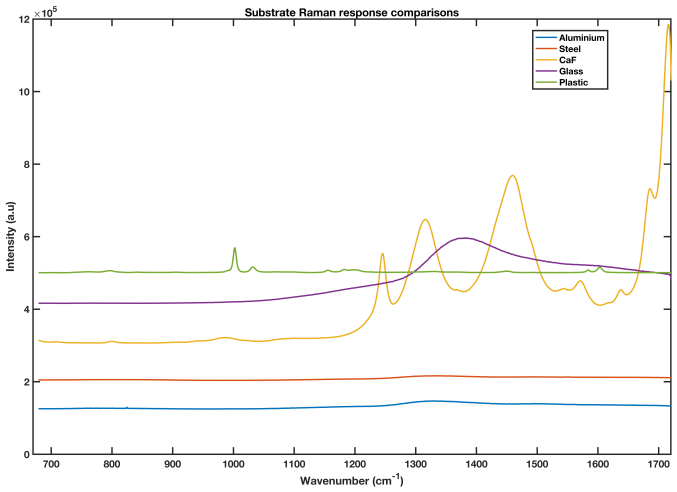


Figure D.2: Comparison of different substrate Raman activity.

D.3 Peltier cooling system base plate

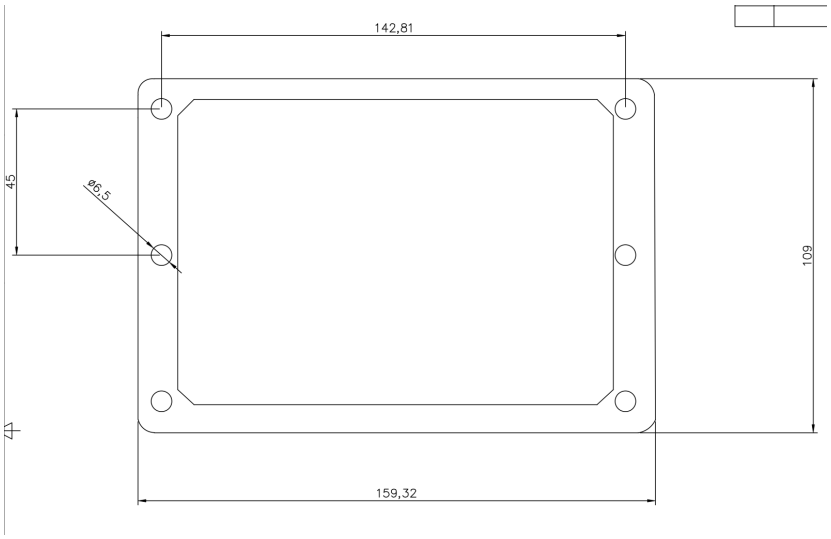


Figure D.3: Schematic of the aluminium base plate design used in the cooling system. This base plate replaces the standard renishaw plate.

Appendix E

Limits of the liquid platform

The limits of the liquid platform were tested by adding in a polyp patient dataset into the binary model. This was in order to establish the limit of detection. Figure 7.1 shows the full patient metadata for the polyp patients involved in the studies.

E.1 Polyp patient metadata

Patient ID	Sex	Age	Smoking status	polyp size (mm)	polyp type	polyp dysplasia	Model Group
SH187	F	63	Non-smoker	6	tubular adenoma	LGD	Train
SH013	M	83	Non-smoker	40	villous adenoma	HGD	Train
SH012	M	77	Non-smoker	30	tubuloviool ous adenoma	HGD	Train
SH198	M	69	Non-smoker	4	tubular adenoma	LGD	Train
SH204	M	48	Non-smoker	4	tubular adenoma	LGD	Train
SH206	F	67	Non-smoker	65	tubuloviool ous adenoma	HGD	Train
SH041	F	47	Non-smoker	10	tubular adenoma	LGD	Train
SH042	F	75	Non-smoker	11	tubular adenoma	LGD	Train
SH059	M	76	Non-smoker	7	hyperplastic	n/a	Train
SH060	M	78	Non-smoker	10	villous adenoma	lGD	Train
SH066	F	64	Non-smoker	28	Inflammatory polyp	n/a	Train
SH094	Male	57	Smoker	35	villous adenoma	HGD	Train
SH095	Male	46	Smoker	4	tubular adenoma	LGD	Train
SH216	Male	87	Non-smoker	50	villous adenoma	HGD	Train
SH106	Female	61	Non-smoker	5	tubuloviool ous adenoma	lGD	Train
SH129	Female	57	Non-smoker	2	tubular adenoma	LGD	Train
SH151	Male	64	Smoker	?	not resected	3 polyps	Train
id015	Male	85	non-smoker	2	tubular adenoma	LGD	Train
id016	Female	73	Non-smoker	<10mm	Sessile adenoma	6 polyps	Train
id020	Male	77	Non-smoker	3	tubular adenoma	LGD	Train
id049	Male	58	Smoker	3	tubular adenoma	LGD	Train
id052	Female	83	Non-smoker	3	tubular adenoma	LGD	Train
id064	Male	51	Smoker	10	tubular adenoma	LGD	Train
id074	Female	71	Non-smoker	3	hyperplastic	?	Train
id094	Female	60	Non-smoker	3	sessile hyperplastic	?	Train
id104	Male	69	Ex-smoker	3	tubular adenoma	LGD	Train
id105	Male	83	Non-smoker	2	sessile hyperplastic	?	Train
id110	Female	79	Non-smoker	4	sessile adenoma	?	Train
id111	Male	74	Ex-smoker	2	tubular adenoma	LGD	Train
id115	Male	70	Non-smoker	<10mm	caecal polyp	?	Train
id140	Male	75	Non-smoker	3	tubular adenoma	LGD	Train
id143	Female	77	Non-smoker	?	Sessile adenoma	LGD	Train
is145	Female	83	Non-smoker	8	tubular adenoma	LGD	Train
id146	Male	72	Smoker	7	tubular adenoma	LGD	Train
id156	Male	78	Ex-smoker	8	Sessile adenoma	?	Train
id157	Male	71	Ex-smoker	3	hyperplastic polyps	?	Train
id159	Male	55	Ex-smoker	10	tubuloviool ous adenoma	LGD	Train
id195	Male	60	Non-smoker	1	hyperplastic polyps	?	Train
id196	Male	54	Non-smoker	2	tubular adenoma	LGD	Train
id197	Male	72	Non-smoker	8	tubular adenoma	LGD	Train
SH260	M	81	ex-smoker	55	villous adenoma	HGD	Train
SH275	M	72	non-smoker	37	villous adenoma	LGD	Train
SH294	M	66	non-smoker	50	tubuloviool ous adenoma	HGD	Train
id048	Male	56	Non-smoker	60	villous adenoma	?	Train
id98	Male	88	Ex-smoker	3	adenoma	?	Testing
id115	Male	70	Non-smoker	?	Caecal polyp	?	Testing
id143	Female	77	Non-smoker	8	sessile adenoma	LGD	Testing
id164	Male	68	Ex-smoker	3	sessile adenoma	?	Testing
id181	Male	65	Non-smoker	2	fibroepithelial polyp	a/w	Testing
id225	Male	80	Non-smoker	3	tubular adenoma	a/w	Testing
id227	Male	54	Smoker	2	adenoma	a/w	Testing
id235	Female	82	Non-smoker	6	sessie adenoma	a/w	Testing
id241	Male	70	Ex-smoker	8	adenoma	a/w	Testing
is268	Male	61	Smoker	5	adenoma	a/w	Testing
id269	Female	65	Ex-smoker	10	terimnal ileal polyp	a/w	Testing
Mean		**69.16**		**13.02**			
Std deviation		**10.80**		**17.26**			

Figure E.1: Full polyp patient dataset

E.2 Vector normalised combined model confusion matrices

Table E.1: Training spectra confusion matricies

	n= 490	Cancer	Control		
Predicted	Cancer	196	44	NPV	78.40%
	Control	54	196	PPV	81.67%
		Sensitivity	Specificty		
		78.40%	81.67%		

Table E.2: Per spectrum testing set confusion matrix

	n= 147	Cancer	Control		
Predicted	Cancer	41	43	NPV	84.13%
	Control	10	53	PPV	48.81%
		Sensitivity	Specificty		
		80.39%	55.21%		

Table E.3: Per patient confusion matrix for the testing set.

	n= 49	Cancer	Control		
Predicted	Cancer	14	14	NPV	85.71%
	Control	3	18	PPV	50.00%
		Sensitivity	Specificty		
		82.35%	56.25%		

E.3 Non-binary model training confusion matrix

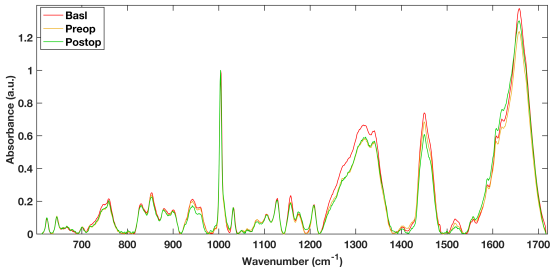
Table E.4: Training confusion matrix for non-binary model including polyp, control and cancer spectra.

		Actual		
	n=539	Control	Polyp	Cancer
Predicted	Control	133	31	36
	Polyp	28	102	28
	Cancer	26	39	116

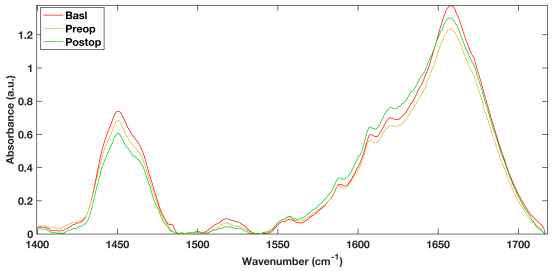
Table E.5: Calculated sensitivity and specificity for each class from the non-binary model

	Sensitivity	Specificity
Control	0.71	0.81
Polyp	0.59	0.85
Cancer	0.64	0.82

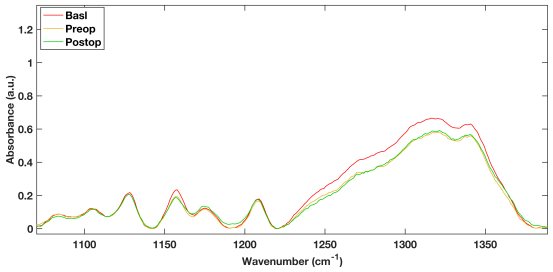
E.4 Disease monitoring



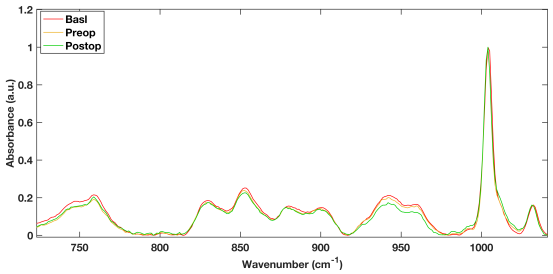
(a)



(b)

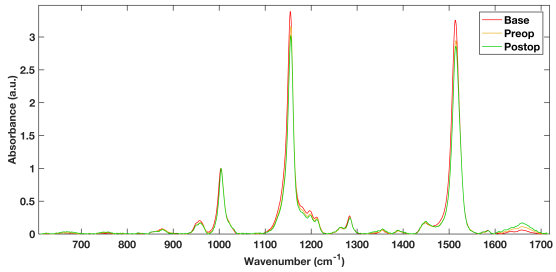


(c)

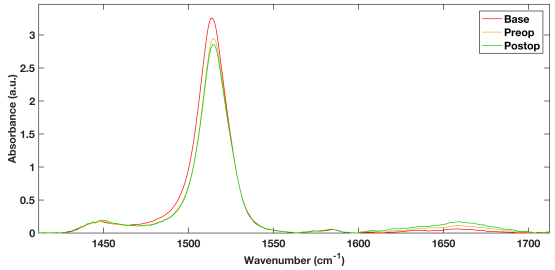


(d)

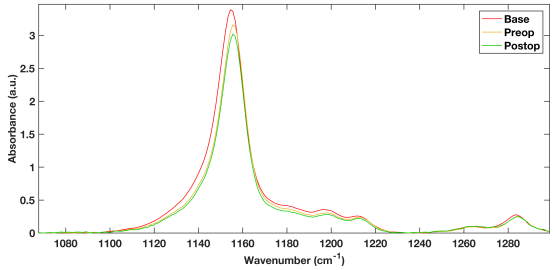
Figure E.2: Average 785 nm patient spectra for the second patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. With full spectra (a), and zoomed in spectral regions (b-d).



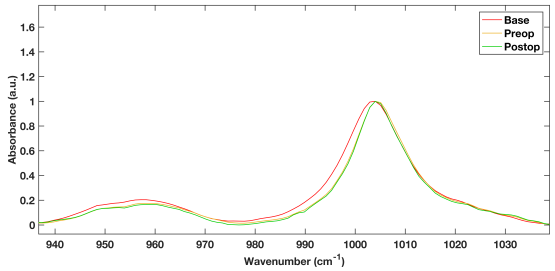
(a)



(b)



(c)



(d)

Figure E.3: Average 532 nm patient spectra for the second patient at baseline diagnosis (rectal cancer), post CRT treatment and post-operative. With full spectra (a), and zoomed in spectral regions (b-d).

E.5 PCA analysis of different sample collection methods

Figure E.4 shows the PC1 vs PC2 score plot for 5 control patients analysed via the original collection methods and 5 samples through the laboratory medicine method. The samples do not seem to cluster according to sample method when the method of sampling is used as a legend. The overall variance between the laboratory medicine samples seems to be slightly higher than the original method, across both PC1 and PC2. However, there is no distinction within the 70% of the explained variance of the dataset indicating that it may be variance expected from different patients.

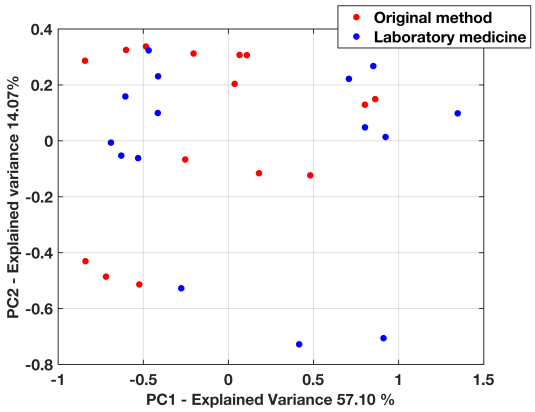


Figure E.4: PC1 vs PC2 score plot for sample processing methods.