

Vibrational Spectroscopy Prospects in Frontline Clinical Diagnosis

Written by **Edward Ian Thomas Duckworth, Bsc. MRes.**

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



Swansea University
Prifysgol Abertawe

2023

Declarations

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

Signed Edward Duckworth

Date **22/02/2023**

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed Edward Duckworth

Date **22/02/2023**

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed Edward Duckworth

Date **22/02/2023**

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed Edward Duckworth

Date **22/02/2023**

Vibrational Spectroscopy Prospects in Frontline Clinical Diagnosis

Contents

1. Abstract	8
2. Introduction	9
3. Literature review	14
3.1 Vibrational Spectroscopic methods	14
3.1 Fourier transform Infra-red spectroscopy	15
3.2 Raman spectroscopy	18
Sample preparation	21
3.2 Cancer and other Malignant Diseases	24
Non-Malignant Disease	26
Buccal mucosa cancer	27
Pancreatic cancer	28
3.3 Sources of uncertainty	29
Transferability difficulties	29
3.4 Digitizing	34
Data pre-processing.....	34

Multivariate analysis	36
Machine learning	39
What constitutes a useful diagnostic accuracy?.....	41
What constitutes a good sample size for a diagnostic study?.....	44
3.5 Biological corroboration	45
Electrochemistry	47
Cell study	48
3.6 Clinical implementation.....	49
Current technology and implementation	49
A model for future advancement	50
4. Utilised Research Methods	51
Introduction	51
4.1 Sample preparation	53
4.2 Centrifugal Filtration for Molecular-weight Windowing	55
Centrifugal filtration.....	55
Molecular windowing	56
4.3 FTIR collection and pre-processing	57
FTIR Instrument Specifications:.....	57
Pre-processing of FTIR spectra:	57
4.4 Post-processing Method using PCA and SVM	58
4.5 Measuring confidence.....	63

Confidence intervals.....	63
DOME compliance	64
5. Results 1: Methodology experimentation and development	66
5.1 Practical methods for uniform deposition investigation.....	66
Methods.....	66
5.2 Normalisation and background correction.....	71
Background correction methods	71
Normalisation methods	73
5.3 Programming and Machine learning model development.....	75
5.4 Methodology experimentation conclusions	82
6. Results 2: Oral (Buccal Mucosa) Cancer	83
Introduction	83
6.1 Raman spectroscopy study to classify buccal mucosa cancer	88
Methods.....	88
Results	90
Discussion.....	91
Conclusions	92
6.2 FTIR spectroscopy study to classify buccal mucosa cancer	94
Results and Discussion.....	94
Conclusions	99
6.3 PCA-SVM and Leave-one-out cross-validation testing	100

Conclusion	102
6.4 Saliva	103
Methodology	104
Results and discussion.....	105
6.5 Oral cancer spectral diagnosis conclusions.....	107
7. Results 3: Pancreatic cancer	108
Introduction	108
7.1 Pancreatic cancer primary case study	111
Methodology	111
Results and discussion.....	113
Conclusion	117
7.2 Pancreatic cancer diagnosis using spray deposition	119
Introduction	119
Methodology	119
Results and discussion.....	120
7.3 ATR-FTIR test on Pancreatic Patient Urine	121
Methodology	121
Results and discussion.....	122
7.4 Cell media and heavy glucose spiking investigation.....	125
Methodology	126
Results and discussion.....	127

7.5	Nuclear Magnetic Resonance testing	131
	Methodology.....	131
	Results and discussion.....	132
7.6	Pancreatic cancer spectral diagnosis conclusions	134
8.	Future work.....	135
8.1	Spectro-electrochemistry investigation	136
	Introduction	136
	Methodology.....	137
	Results and discussion.....	138
9.	Conclusions	141
10.	Acknowledgements.....	145
11.	References	146

1. Abstract

The key experimental results from this research are the viable and cost effective methods of diagnosing oral and pancreatic cancer with accuracies over 90%. Furthermore, development of the molecular windowing method to further narrow down the origins of those cancer biomarkers and further improve accuracy.

Many papers are being published demonstrating how vibrational spectral biomarkers can be used to diagnose a whole variety of diseases, from cancers to colitis. However, much of the research, proposed as discovering a useful tool for clinical diagnosis, has not yet been widely utilised in clinical practice. This is due mainly to the lack or reproducibility of the findings and current lack of relating the spectral observation to a root biological cause. This thesis aims to highlight the inconsistencies between studies and propose an improved process for spectral biomarker identification, including suggestions for follow up studies to discover the foundation of the spectral change. This thesis reassesses, and adds to, ground covered by previous reviews regarding sample preparation, patient selection and multivariate analysis.

Resultantly, this thesis brings light to the need, and suggests solutions, for:

- a method to standardise results between detection devices,
- knowledge of the additional requirements for using biomarkers for disease monitoring/prognosis,
- understanding the biological root cause for the spectral shift.

These promising results and suggestions for combined methodology improvements will provide guidance to enable this burgeoning research field to improve patient outcome in the clinical sphere.

2. Introduction

With many recent medical advances, we have many methods to effectively treat patients if their ailment is known. However, the diseases that still lead to fatality are often preventable if detected early enough. The idea of being able to quickly and effectively screen for diseases is an attractive one, but unfortunately the implementation of such a system is quite complex. Current UK cancer screening processes, for example, have multiple week waiting times for a multitude of symptom tests which may or may not lead to a diagnosis. Radiology and endoscopy are often used in current screening processes, but have a capacity bottleneck which means that 338,000 patients across England have to wait more than a month for radiology results⁴. Often these require use of secondary care investigations from specialists that can be unsustainably expensive, especially with an aging population that is likely to require increasing amounts of screening for multiple conditions. In fact, studies estimate that cancer diagnoses have increased by 2% per annum and that 50% of people born in England since 1960 will receive a cancer diagnosis in their lifetime⁴. Cancer is also a worldwide problem. In 2020, there were over 19 million new cases of cancer and nearly 10 million deaths caused by cancer across the world's population⁵.

According to the World Health Organisation, “Cancer, when identified early, is more likely to respond to effective treatment, resulting in a greater probability of surviving as well as less morbid and less expensive treatment”⁶. This is often because late diagnosis is associated with larger more established cancers, delays in accessing cancer care and therefore a lower likelihood of survival.

This thesis will focus on the potential of spectral biomarkers as a solution to the cancer screening problem. These methods promise a fast, cost effective and minimally invasive

method to diagnose disease from a simple blood sample. There has been much research into these in recent years, but none of the methods have reached the clinical sphere¹.

Of course, if one of these spectral biomarkers could be developed to the point in which it can perform an accurate diagnosis, this would streamline the treatment process even further. However, the accuracy an automatic diagnosis would require would be higher than that which is currently easily attainable. This, such testing is only suitable as additional information to aid with medical decision making.

Another potential use of spectral biomarkers is the ability to monitor an already-diagnosed disease. The process of detecting changes in something you already know to look for can be easier than diagnosing relatively blind. However, one has to know and quantify biomarkers for the disease relevant to the method you are using to monitor it. Of course, if an effective biomarker and associated diagnosis method can be developed, it could be possible to later quantify and develop it into a monitoring method. This would be especially useful when testing those patients that are at risk of cancer re-occurrence. The ability to detect cancers at an early stage has a dramatic effect on the cost of treating them. For example, treatment costs for late stage colon cancer can increase nearly fourfold as compared to treatment started at an early stage⁴.

The effectiveness of screening has been demonstrated with the ‘Supermarket scan’ initiative, in which computerized tomography (CT) scans were offered to 2500 people in Manchester, finding 46 cases of cancer with 80% of the cases being early stage^{7,8}. Of course, strategies like this require the use of high-cost and low-mobility CT equipment, but it demonstrates the effectiveness mass screening can have.

Currently in clinical practice, cancer is diagnosed in a few distinct ways depending on the type. Some tumours can be detected using a physical examination. The patient’s blood can be tested in a few different ways. For example, a complete blood count can

provide evidence of leukaemia. A test for prostate specific antigen (PSA) can indicate prostate cancer⁹. A CT scan or various other types of scan can be taken and a medical professional can examine them to look for cancers in the body. Most commonly, often after one of the other preliminary tests, a biopsy is required to confirm the cancer diagnosis. This is where a sample of cells is visually examined under a microscope¹⁰.

Table 2.1 outlines some currently used cancer diagnostic methods.

Table 2.1: current cancer diagnostic techniques and estimated timescales, costs and accuracies.

Typically, multiple tests will be used in conjunction to assess a patient.

Technique	Cost	Time	Invasiveness	Estimated accuracy
Internal biopsy ¹⁰	~\$10 000 ¹¹	2-10 days	High	~48-99%
Pancreatic ELISA (Table A1.3)	\$150 - \$600	up to 4 hours	Low	~70-85%
Computerized Tomography	~\$2100 ¹¹	up to 1 hour	Medium	~65%
Ultrasonography	~\$2,688 ¹¹	up to 1 hour	Medium	~50-70%
PSA test ⁹	\$50-\$200	up to 1 hour	Low	~60%
Galleri test ¹²	~\$949	10 days	Low	~75%

There are various problems inherent to each of the above approaches. Physical examinations cannot easily detect the many kinds of cancer that are within the body. Even if a lump is found, it may still be benign. Blood and urine tests are often unreliable as they can be influenced by other factors like diet and stress. For example, 75% of men with a raised PSA level don't have prostate cancer, and 1 in 7 with a normal level do have cancer⁹. Scans can be more effective but do not always provide definitive evidence either. Bone scan abnormalities can be caused by other diseases such as arthritis¹³ and CT scans can produce a significant number of false positives, leading to unnecessary

invasive procedures being carried out¹⁴. Furthermore, early detection of cancer in this way is difficult due to the relatively small size of any tumour and minimal other physiological changes. Finally, biopsies can be invasive for the patient and difficult to carry out for the medical staff, this creates medical expense and possibly discouraging patients from getting any symptoms investigated when they arise. To summarise, all of these methods require several steps, long wait times and/or the use of valuable medical staff to be certain of the patient having the correct diagnosis.

Therefore, simple cost-effective methods that can be employed to quickly check the malignancy of a patient's symptoms would be extremely valuable. Of course, the detection process would also have to be as non-invasive as possible. The ideal method for this would be to have a small device that a minimally trained operator could use to examine some low-invasive, minimally prepared, bio-fluids from the patient - saliva, urine or most likely blood.

Like cells and tissues, bio-fluids can be analysed for vibrational spectra. Blood is a particularly useful bio-fluid for inspection due to its high protein concentration and because changes in protein levels are some of the best indicators of disease. Much of the current research is focused on subsets of blood; plasma and serum¹. In whole blood, haemoglobin and other red blood cell associated molecules can interfere with the spectra, so the plasma is preferred as the variable protein concentrations within vary more when there is a disease. Serum is a subset of plasma without the coagulating factors for easier storage and use. Without these natural coagulants present, other de-coagulating chemicals do not need to be added. This is beneficial as some of these added anticoagulants like EDTA and citrate have been demonstrated to produce confounding results in Raman spectra¹⁵. Furthermore, there is even a study demonstrating better quantification of molecules such as glycine in serum by only using the <10kDa

fraction¹⁶, so maybe even subsets of serum will be more useful for the identification of spectral biomarkers for certain diseases.

3. Literature review

3.1 Vibrational Spectroscopic methods

Characterising a sample is key for being able to tell if it indicates disease or not. There are numerous methods by which this can be done, to varying degrees of effectiveness. Of course, the degree of precision required in the characterisation of blood serum for diagnosis purposes typically surpasses the basic use of our unaided senses, and therefore we have to rely on specialised equipment to scan the sample for us. One of the best methods we have devised for this analysis is vibrational spectroscopy – particularly for its non-sample-destructive nature, allowing us to examine it in as close to a natural composition as possible. No other reagents or chemical additions are required to analyse a sample, only requiring light to be directed upon it. The most that some of these methods require is that the sample be dried, though there are some methods that will analyse an aqueous sample as well.

Vibrational spectroscopic methods rely on the principle that molecules, and primarily the bonds binding them together, absorb light. In particular absorption in the region of 400 nm to 1 mm in wavelength are most typically of interest. Specific bond vibrations of a molecule leads to absorption of light at characteristic wavelengths, and by detecting which wavelengths have been absorbed and to what degree, one can make deductions about a particular mixture's composition. With regard to disease detection in blood, this ideally can be used to compare an unknown sample with a premade database of known healthy and unhealthy samples to determine if there is an affliction^{1,17,18}.

Of course, there are inherent difficulties with this, especially in something as complex and individually specific as a human blood sample. All human bio-fluids are inherently

complex, containing a large variety of proteins, lipids and other molecules and therefore the magnitude of shifts in the spectra produced will likely be comparatively small and require significant statistical analysis to discern. However, if a spectral biomarker can be identified comprehensively from these samples, then the benefits could be many. For example, if a particular blood biomarker for a cancer is identified, a patient can be quickly diagnosed from a simple, non-invasive, blood test. Potentially even the stage of cancer can be discerned and prognosis for the patient identified if the biomarker is quantified. Then its progression or recession can, simply and non-invasively, be tracked as the patient is treated - all by use of blood spectroscopy.

Different properties of the chemicals tend to absorb light in particular regions of the spectrum, therefore observing each region has its own specialisation and characterising name. For the infra-red region it is termed simply 'infra-red-spectroscopy' and this focuses on absorption of light as it excites vibrations within the sample, typically in its bonds. Visible light spectroscopy has two distinctions, 'UV-Vis' and 'fluorescence' which deal with light from electronic transitions, the absorption or emission of it respectively. However, there is also 'Raman spectroscopy' which focuses on looking at small fluctuations in the absorption of visible light from inelastic scattering. Both the Infrared and Raman methods are of particular interest as they tend to provide more structural information to better identify a complex molecule.

3.1 Fourier transform Infra-red spectroscopy

Infra-red absorbance spectroscopy looks at the absorbance of a sample in the infrared region of the electromagnetic spectrum. It achieves this by projecting monochromatic infrared light at a sample. At some infra-red wavelengths, features of molecules in the sample will absorb the radiation, transferring the energy into bond vibrations of the same

frequency as the photon absorbed. The absorption pattern is measured over a whole range of wavelengths of infrared light. By knowing the characteristic absorption regions, one can discern the presence of certain structures and even fully characterise the molecules in the sample from the produced spectra.

Fourier transform infra-red spectroscopy (FTIR) is a method for sampling a whole range of wavelengths at once, dramatically speeding up the process of analysing a sample over a large range of the infra-red spectrum. It also allows overall light levels to be higher, improving the signal to noise ratio. It achieves this by guiding the light through an interferometer before or after the sample. This alters the distribution of light and produces a signal called an interferogram. This interferogram can be analysed using a mathematical process known as a 'Fourier transform', which converts it into a readable absorbance spectrum for the sample¹⁹.

A thin polystyrene film is often used as reference standard for FTIR spectrometers. Calibrated polystyrene wavenumber standards are available from the National Institute of Standards and Technology to reference against, though film thickness and surface scattering properties are not as standardised²⁰.

Transmission

Transmission FTIR is the simplest form, in which infrared light is simply shone through a thin layer of sample. A detector is placed behind the sample and the amount of IR light that reaches it is measured, giving a transmission spectrum. The limitation of transmission FTIR on biofluids is that a dried sample is required. Water has a large IR signal so any aqueous sample analysed would be obscured.

ATR

Attenuated total reflectance (ATR) is an accessory commonly used in FTIR spectrophotometry. It purportedly allows for faster sampling and better reproducibility

of results. This means, with less user to user variability, it is easier to produce better databases across laboratories, allowing for more precise material identification and cross referencing²¹.

This is achieved by having a thin layer of sample pressed in close contact with an ATR crystal. The infrared beam then passes through this crystal and undergoes total internal reflectance. The light still penetrates the reflective barrier as evanescent wave, to a limited extent, and the sample's contact with the crystal surface is sufficient to allow this evanescent wave to undergo absorption. Therefore the resultant signal can be detected and interpreted much the same way as normal FTIR. This allows for easier analysis of solid materials, though it comes with the limitation that only a small number of micrometres into the sample will be analysed. The penetration depth (d), defined as the depth where the electric field decays to e^{-1} , is typically between 0.2 and 5 μm is dependent on wavelength (λ), angle of reflectance (θ) and the refractive index between the crystal (n_1) and the sample (n_2) as shown in equation 1²²⁻²⁴.

$$d = \frac{\lambda}{2\pi n_1 \sqrt{\sin^2\theta - \left(\frac{n_1}{n_2}\right)^2}} \quad (3.1)$$

The limitation of only focusing one the immediate interface of 10-100nm into the sample is not always a good indication of the sample as a whole, especially when there is a high likelihood of variable deposition, as would be in a biofluid mixture. Also, absorbance intensity in ATR tends to be skewed toward lower wavenumbers where penetration is higher²⁵. Furthermore, one would have to account for the crystal used in the ATR measurement, when comparing with other measurements using other crystals, as they may have different penetration depths.

Fibre optic FTIR, or Mid-Infrared Fibre Evanescent Wave Spectroscopy (MIR-FEWS) is a method that operates similarly to ATR, in that the absorbance from a surface

interaction of IR light is measured²⁶. In this case the light is bouncing through a fibre optic cable. The advantage of this is that the testing can be done in liquid samples. This can also reduce the impact of the limited surface interaction as a liquid sample would not be prone to variable deposition and therefore the method can better assess the whole sample.

Focal plane arrays are a newer advancement in FTIR and can allow rapid IR spectral imaging by taking thousands of concurrent spectra over a large sample area²⁷. This would allow certain features in a non-uniform sample to be examined. For example a dried sample droplet can exhibit noticeable surface differences like cracks or the coffee ring effect. These differences could be examined and potentially an optimal region of interest could be established for a droplet.

3.2 Raman spectroscopy

Raman spectroscopy focuses on the visible light region. When light is incident upon a transparent sample, its oscillating electric field can act on the charges of a particle, the photon being absorbed and causing the charges to move at the same frequency. This energy is emitted as another photon. Typically these emitted photons will have the same frequency as the original, but a small fraction (~1 in 10million) will be different^{28,29}. This occurrence is known as inelastic scattering as the light is scattered, but the energy is changed and is termed Raman scattering after its discoverer Sir C V Raman. Typically the scattered light is of lower energy, termed ‘Stokes’ scattering, but can also be higher energy, termed ‘anti-Stokes’.

In a Raman spectrometer, monochromatic visible lasers are used to probe a sample and the inelastic scattering it exhibits is measured in order to discern features in its molecular structure³⁰. The data produced is a spectra of ‘Raman shift’, which is calculated as the

difference between the wavenumbers of the laser and the detected light. Like in infrared spectroscopy, the Raman shift is presented in units of wavenumber, and the peaks sometimes occur at similar wavenumber range as in the Infrared spectrum. However, different regions will be highlighted to different magnitudes and in symmetrical molecules with a centre of inversion a signal will only be present in one. The methods are often considered complimentary and therefore, using both methods can give an even more comprehensive characterisation of the sample of interest³¹.

There are several variations of the Raman spectroscopy method. For example, variations on the positioning of the laser and detector. In 'reflectance mode' the light, the laser and the detector are placed on the same side of the sample and the scattered light that is reflected off the sample is measured. In transmittance mode, the laser passes through the sample to the detector on the other side. In reflectance mode there are potentially differences due to interaction with an uneven surface. In transmission mode, the sample surface is less crucial, but a suitable substrate has to be chosen to place your sample on - so not to produce a confounding signal^{32,33}.

There are also more recently developed methods for improving the Raman signal from a sample³⁴. Fibre optic probes are an option when using Raman³⁵, these are especially used in tissue studies.

TERS

Tip enhanced Raman spectroscopy (TERS) is a promising method to improve the precision of the location of molecules contributing to the Raman spectra to allow chemical imaging at nanoscale resolutions³⁶. It achieves this by using a scanning probe microscopy tip, a very fine point, and uses the electromagnetic field at the tip to enhance the Raman signal from the molecules in the close vicinity, allowing them to dominate the measured spectra. This allows for potentially sub 20nm spatial resolutions.

Raman spectra can be achieved using a variety of wavelengths of laser, though each will have its own advantages and disadvantages. In biological materials, there is the potential for fluorescence of a sample to interfere with the spectra, and therefore the near infrared (NIR) region at wavelengths around $730\pm 100\text{nm}$ is preferred to reduce this component. Higher wavelengths are sometimes used, though commonly used silicon CCD detectors have a quantum efficiency that decreases with wavelength meaning that the signal will be reduced to 15% by 1000nm ³⁵. Therefore the NIR region is preferred as a compromise between the two factors.

SERS

Surface Enhanced Raman Spectroscopy (SERS) is a method that can enhance the signal strength of certain peaks in a Raman spectrum. It achieves this by providing a metal surface for the molecules of interest to adsorb onto which causes the observed enhancement via a mechanism that is still in debate. The surface can be anything from a rough flat material to the outer layer of certain nanoparticles. For diagnostic purposes there has been some investigation to see if this enhancement can help. One study compares traditional Raman and SERS, looking at whole blood, plasma and red blood cells. The signal from SERS had a notably higher signal to noise ratio. It was also noticed that the plasma samples degraded over 24h of storage time³⁷. A diagnostic study on lung cancer managed to use silver nanoparticle SERS to detect the cancer with a high specificity/sensitivity of 100/90%. Their method involved improving the distribution of nanoparticles in the blood serum by using an array of pyramidal silicon³⁸.

Once one has taken spectra of the sample, it is important to use a spectral reference to investigate likely candidates for the peaks you observe in the IR or Raman spectra³¹.

Sample preparation

Figure 3.1 presents a suggested workflow for use in biofluid spectroscopy from a recent review. The second category mentioned is sample preparation. When preparing blood serum for investigation, it is important to consider which method is chosen to characterise it first. For example, when using transmission FTIR on a liquid sample, the peak on the spectra due to the water is large and obscuring of more pertinent peaks. Therefore, dried samples are much preferred for this method. However, when using Raman, both aqueous and dried samples can be perfectly acceptable.

When using a dried serum droplet sample, the procedure used for drying is very important. This is due to the variable deposition of constituents, typically due to their differing masses. If a serum sample is left to dry in ambient conditions, it will exhibit a ‘coffee ring’ deposition pattern, with the heavier components distributed in higher concentrations around the edge of the droplet. Often, crack like drying patterns will also be observable in the dry sample^{1,39,40}. It is important to try to minimise these effects without compromising the sample itself, one proposed solution⁴⁰ is diluting the sample before drying, this tends to reduce the deposition variability and reduce the likelihood of cracks forming. A three-fold dilution was recommended as an optimal compromise between variable deposition and keeping small concentrations high enough to be detectable. A recent review on bio-fluid drying tentatively suggests using a diluted 1µl droplet for analysis would be optimal⁴¹.

A recent paper has highlighted the potential of laser induced differential evaporation to dry sample, leaving >95% of the dried components within centre of drop, allowing for better analysis of the complete sample⁴².

A paper compared dried and liquid serum with Raman and ATR-FTIR, in the context of distinguishing between healthy and depressed patients. Though there will be small changes in protein structure from drying, the study concluded that it had no significant effect on the ability to diagnose patients with Raman spectra and only a minor effect on the FTIR spectra⁴³. Therefore one could suggest that comparing between liquid and dried Raman studies should be valid. However, a more comprehensive comparative study should be attempted before the result is applied to all spectra. The minor deviations in the FTIR spectra indicate that a standardised process is still required to

avoid confounding results in its case, as deviations between characteristic disease spectra can also be quite minor.

ATR-FTIR is simpler to set up for blood serum, allowing the user to place a droplet on the sensor for immediate analysis. However, the limited penetration of the ATR method into a sample may give a result that doesn't reflect the sample as a whole²⁵.

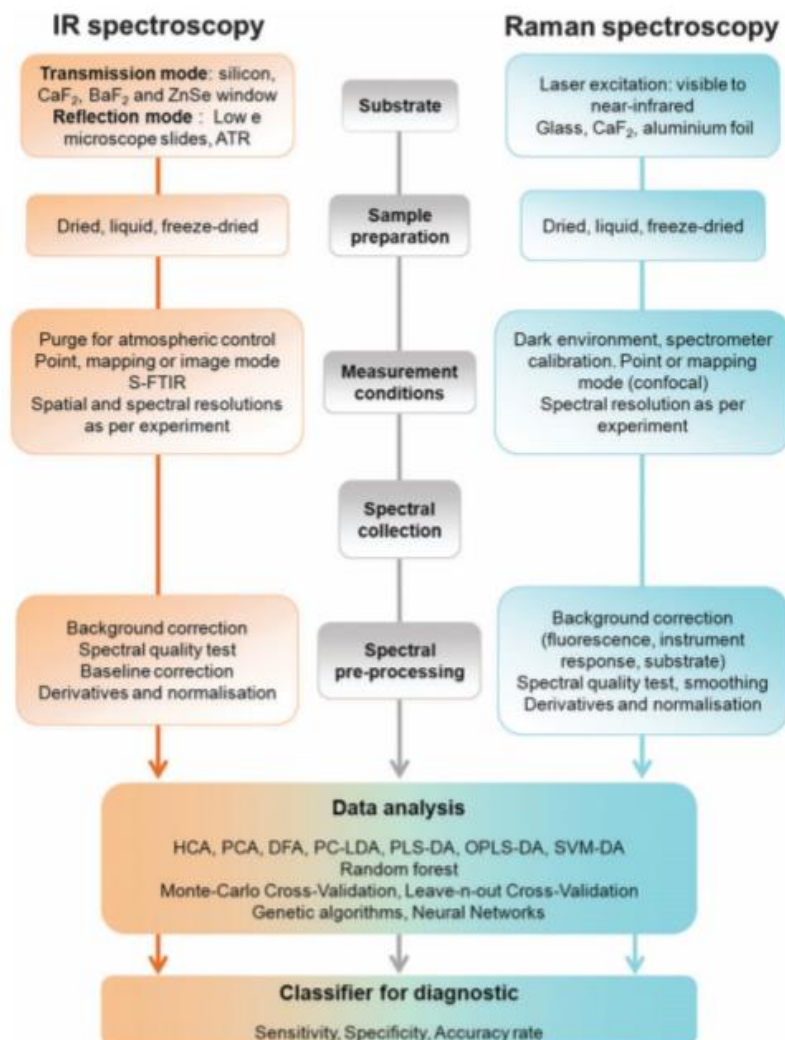


Figure 3.1: Proposed workflow of biofluid spectroscopy from a recent review article¹.

In Raman spectroscopy, if you are using transmission mode then one has to choose a suitable substrate to hold your sample. Glass tends to have a high signal in Raman, so Raman grade calcium fluoride or zinc selenide is recommended as they only exhibit significant Raman peaks at $<600\text{cm}^{-1}$ for lasers ranging between 473-830nm^{32,33}. Reflectance mode is more forgiving, though placing your sample upon a substrate without a confounding signal is recommended, non-oxidised aluminium is one often used as it has a stable, low signal over commonly used laser wavelengths, except 532nm³².

Much of the investigation of the vibrational spectroscopy of blood plasma and serum for the purposes of diagnosing and screening of disease are proof-of-principle studies. These typically demonstrate the potential of FTIR or Raman spectroscopy to distinguish between diseased healthy samples on a relatively small sample set. Of these studies, many of them are investigating cancers in the attempt to distinguish characteristic spectral biomarkers for them, though others will look into non-malignant diseases with a similar goal in mind.

3.2 Cancer and other Malignant Diseases

Studies into malignant diseases like cancer are not only satisfied with the potential for diagnosis, but also the usefulness of finding a biomarker that could be used to monitor the disease in known patients is greater than for non-malignant diseases.

It is not always clear why FTIR or Raman is chosen for the study of a particular disease, it could be as simple as a certain laboratories' access to equipment, or more significant in that the opposite method did not yield as promising results and was therefore not mentioned.

FTIR is quite popular in these studies. It has been used on human blood serum and plasma in one paper looking at colorectal cancer. The study identified deviations in peaks between healthy and cancerous samples; at 1141 and 1105 cm^{-1} in peripheral blood mononuclear cells and 1082, 1050 and 1032 cm^{-1} in plasma⁴⁴. Another FTIR study looked at non-small cell lung carcinoma using Random Forest (RF) and Maximum Relevance, Minimum Redundancy (MRMR) statistical methods to select features of interest from the spectra⁴⁵. The novelty in this study is that it attempted to distinguish between patients with cancerous and non-cancerous lung diseases, not cancerous and healthy as this had been looked at in a previous study⁴⁶. They managed to differentiate cancer from other disease with a 79% accuracy and also went on to differentiate the specific type – distinguishing between squamous cell and adenocarcinoma with 80% accuracy using random forest classification⁴⁵.

Other studies use Raman spectroscopy when looking for spectral biomarkers. A review on gastric cancer detection⁴⁷ highlighted a series of papers that used silver nanoparticle assisted SERS to look at proteins purified from plasma samples and distinguished between the healthy and unhealthy samples with 100% accuracy⁴⁸, then looked at just the serum with RNA, achieving 100/94% sensitivity/specificity⁴⁹. Furthermore another

study looked deeper, reporting that the SERS peak heights were distinguishably different, peak height increasing between benign disease, early stage and late stage gastric cancer⁵⁰. They didn't quantify this observation with a sensitivity/specificity model however.

A recent Raman study on colorectal cancer serum yielded promising results, achieving 83% sensitivity and specificity using cross-validated partial least squares discriminant analysis⁵¹. This study was notable as it examined the ability of their method to be high-throughput, looking towards improving applicability to clinical practice.

Another review highlighted when Raman spectroscopy was used to trace prostate cancer metabolism in a living cell⁵².

There is even a study evaluating the usefulness of two variations of Raman spectroscopy for the application of identifying spectral biomarkers. Both spontaneous Raman and SERS on were used on blood plasma in order to distinguish between benign gynaecological patients and ovarian cancer sufferers⁵³. They identified 5 peaks of interest to be their spectral biomarkers and achieved 94/96% sensitivity/specificity with spontaneous Raman as opposed to 87/89% with SERS. Furthermore, early ovarian cancer cases were diagnosed with 93/97% for spontaneous (traditional) Raman and 80/94% for SERS. Notably, they also evaluated the effect of patients' age on their results and, though there was some decrease with age, the overall accuracy remained high over all age ranges.

Less frequently, other malignant diseases like Alzheimer's are investigated, one proof-of-concept Raman study focused on blood serum in these patients. They managed to differentiate Alzheimer's from other dementia types with an accuracy of 95%⁵⁴.

Occasionally, studies will make use of both the complimentary FTIR and Raman spectra in their investigation. For example, both were used to look for cancer spectral

biomarkers in extracellular vesicles in blood serum, with prostate cancer as the chosen example⁵⁵. They observed differences between several peaks in both of the collected spectra, though no sensitivity/specificity model was used to quantify the observations.

Non-Malignant Disease

Typically investigations into non-malignant disease are more focused on the applicability of the biomarker to be used for direct diagnosis.

FTIR has been used to look into colitis in one study on mice blood serum⁵⁶, where significant differences were shown between colitic and healthy or treated-colitic mice in the amide I and carbohydrates peaks at 1660-1620 and 1100-1000 cm^{-1} respectively.

Depression is a little atypical for a non-malignant disease, where the usefulness of monitoring its levels, rather than just a simple diagnosis, is greater. One FTIR study on blood serum and plasma in humans looked at both depression and bipolar disorder⁵⁷. They observed some reduction in bands in the protein and phospholipid region of the spectra when compared to healthy patients. They also looked into the gender differences and noted that males exhibited a lesser decrease overall and also noted some differences between age groups, though no model was constructed from the data. The same group also looked zinc deficiency induced depression in rats, analysing the serum with FTIR and Raman, and found divergences from healthy samples in the both the observed spectras⁵⁸.

A pair of papers by Khan et al. used Raman to distinguish patients with Dengue fever from healthy patients using different models with corresponding 73%/93% and 91%/91% sensitivity/specificity^{59,60}. This group have also used it to look at hepatitis B and achieved 97%/100% sensitivity/specificity⁶¹. Another study from this group has assessed the effectiveness of many models for classifying Hepatitis C infected blood serum using its Raman spectra⁶². The highest sensitivity/specificity produced in this

study was 97%/94%. These papers seem to have suffered from the similar pitfalls that have been discussed earlier as the only control used appears to be healthy vs diseased samples and little to no discussion of any biological reasons for the signal's origins are suggested.

Buccal mucosa cancer

Oral cancer is one of the most predominant cancers in India due to the predominance of the tobacco chewing habit. It is also one of the most expensive cancers to treat in the United States⁶³. This cancer is of interest to test a diagnostic method due to the ease of access of the cancer for traditional diagnosis, resulting in easy corroboration of results and the large number of patients that can be investigated.

Recent reviews on this cancer have highlighted the promise of FTIR for diagnosis of this cancer, especially highlighting the ability to detect pre-cancerous changes – catching it at early stages⁶³. Though it has not yet reached clinical applications, it is expected to play a significant role in the near future.

Buccal mucosa has had recent research into its potential for Raman screening by Sahu et al. where the feasibility of classification was explored⁶⁴ before being followed up by a larger and more comprehensive study⁶⁵. The latter study contained suitable premalignant and related disease controls and produced sensitivity and specificity values of 64 and 80% in determining the presence of an abnormality, and higher values for determining the correct abnormality from the glioma, premalignant and oral cancer options used in the model. It was noted that these values are comparable to current screening techniques.

Pancreatic cancer

Pancreatic cancer is one of the most deadly cancers in the UK⁴, the 5 years survival rate remaining below 10%⁶⁶. This is mostly due to that it is hard to detect until it has reached a later stage.

Reviews on the recent literature highlighted that promising research had already been done on the disease, concluding that FTIR was of most use of clinical applications due to its speed. Raman was deemed useful to allow deeper investigation into subcellular mechanisms^{67,66}. However, it was also concluded that more research was still necessary with larger sample cohorts and better ability to handle irregular cases and exceptions.

This cancer is very suitable for investigation as this type of hidden cancer requires most diagnostic aid, as it is unlikely to be easily diagnosed by any other means. The best way to improve the survival rate is to have a viable method of effectively screening for the condition.

3.3 Sources of uncertainty

Transferability difficulties

Demonstrably, many studies have been able to identify spectral biomarkers of certain diseases, sometimes with very high specificity. However, very few of these have yet been used as an aid for practical medicine¹. Unfortunately, several difficulties arise when attempting to translate any promising studies into viable clinical procedures. A major factor in this is the reproducibility of the results obtained in these studies, which is often low. This can be due to a multitude of reasons, inconsistency with sample preparation is common as no standard has yet been decided upon. Many papers on FTIR and Raman for diagnostic purposes reference the inconsistency of sample preparation in previous studies and the poor relevancy of the outcomes^{1,16,32,39,68}. Fortunately, most studies describe their particular sample preparation steps, so there are many to select from. More than this, machine to machine variability can completely distort results, and with the high precision required to differentiate the spectra, any minor shift will make the statistical comparison to the original database inaccurate⁶⁸.

Controlling for sample preparation and spectrometer specifics, further issues present themselves in the form of sample deposition patterns. In FTIR, or otherwise when a sample has to be dried, very specific drying conditions are required. This is to reduce the deposition variability, else cracking or ‘coffee ring’ style deposition patterns will occur, with heavier components being deposited first at the faster drying edges or cracks – creating variation in concentrations across the sample³⁹. Even when the drying is controlled as much as possible, ATR-FTIR measurements of dried samples are only

describing the surface properties up to around 100nm into a sample. This means that if heavier components deposit first, or similar uneven vertical deposition patterning, then this method will produce erroneous concentrations for the sample²⁵.

When looking for biomarkers of disease, it is also important to ensure that your spectral shift is due to your specific disease and not just, for example, a body's natural immune response. During a variety of infections the body's albumin to globulin ratio is decreased, leading to a significant shift in the spectra observed. These differences must be considered, especially as these chemicals are at relatively high concentrations in comparison to many potential biomarkers⁶⁸. Unfortunately some studies where this would be relevant forget to control for this, and therefore it is difficult to confirm if their results are of use. In fact, the selection groups should be matched for age and sex as well as atypical conditions, like the hormonal status of pregnancy, as well as additional pathologies. Biomarkers should ideally be identified within these groups before they are compared, otherwise it is likely any discerning feature in the spectrogram data will potentially be due to the confounding factors in the serum composition, and not to the disease of interest¹.

After a marker had been analytically validated, accounting for the above pitfalls, the next stage is to push for clinical validation. To achieve this, a sample has to undergo an extensive diagnostic performance review on a large, independent sample of patients. This involves large, randomised control trials at a number of medical centres, where the sensitivity and specificity of the potential biomarker will be evaluated to the gold standard procedure for its specific purpose. The studies will be carefully designed to test for this purpose, be it screening, monitoring, differential diagnosis or other specific niche use¹. It is important not to make the mistake of extrapolating the performance of a biomarker from the original study to this screening context. Typically, the original

study will be on small numbers of pre-diagnosed patients to better identify and characterise the marker. When applied to the clinical screening setting however it is likely that the sensitivity and specificity of the results will be much lower than in the pilot study. False positive results often mean a patient is referred for further diagnostic procedures, which are usually costly and potentially invasive for the subject, and therefore a high specificity is of paramount importance. The specificity can be increased by selecting case patients and controls for the specific clinical purpose the biomarker is intended for, highlighting the importance of carefully designing these trials¹.

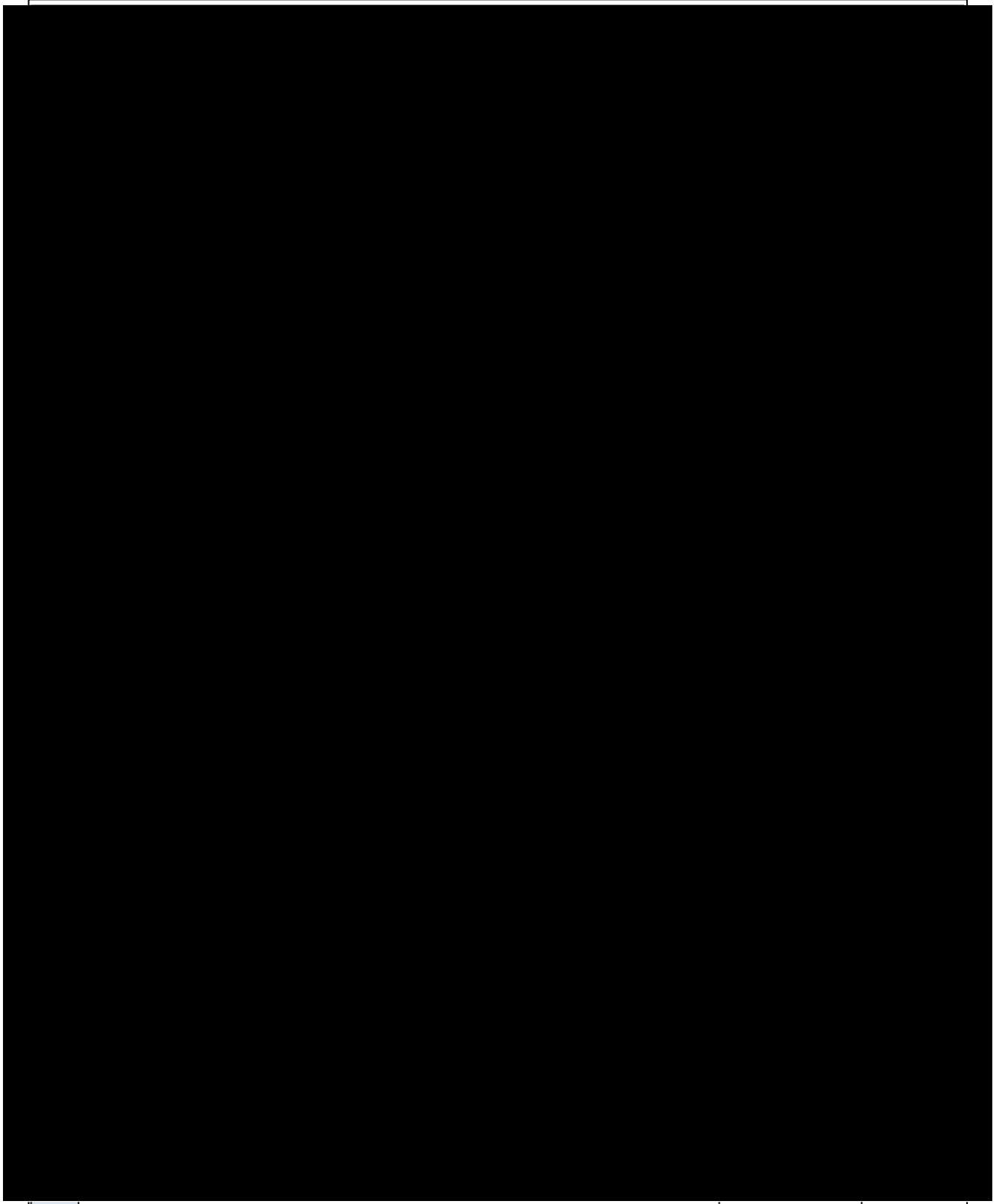
Recently, there has been a study demonstrating better quantification of molecules such as glycine in serum by only using the <10kDa fraction as it removes large obscuring signals from globulin (>80kDa) and albumin (>60kDa)¹⁶, so maybe even subsets of serum will be more useful for the identification of spectral biomarkers for certain diseases. Further research using ATR-FTIR has been done in this area, looking at using ultra-filtration on samples to get better detection of the low molecular weight components^{69,70}.

Recent reviews outlined some of these reproducibility issues^{1,68} though little direction was suggested. Subsequent research has still fallen into these same pitfalls. A few examples of how studies have been limited are outlined here. Several studies exhibit a lack of appropriate control groups in their classification, only classifying between 2 healthy/unhealthy categories for their disease of interest^{60-62,71-76}. Several do not produce a sensitivity/specificity based accuracy model^{55,56,58,74}, merely looking at visual spectral shifts. Occasionally, some papers account for some, but not all, appropriate controls and produce cross validated sensitivity/specificity^{72,77}. A very comprehensive study on liver disease produced diagnosis accuracies and cross validated them, but then didn't discern sensitivity/specificity values from their results⁷⁸. Occasionally, some produce solid

work using the principles of accounting for appropriate controls and produce cross validated sensitivity/specificity results^{38,53}. Even then, some important factors, like gender and age, are not specified or controlled for. Some papers directly cited the recent biofluid spectroscopy review(s) mentioned earlier. These were often better, using appropriate, though not comprehensive, group classification and producing a sensitivity/specificity model^{26,79-82}. Unfortunately, some still didn't have enough grouping and didn't attempt a sensitivity/specificity model⁸³. Even fewer tentatively discussed any biological backing for their observed signals^{26,61,83}.

It is clear that, despite solid reasoning and general good science being used in many of these papers, the lack of any form of standardisation for these studies allows some of the persistent follies of spectral classification to be repeated. It seems necessary for a universal baseline to be established for these studies so that they can be more valid, reproducible and better cross-compared with other similar studies. Table 3.1 outlines the main pitfalls and preliminary suggestions for improvements or standardisations of practice. Throughout this thesis, these will be refined with the aim to have solid solutions to all the issues raised that can be utilised in future research.

Table 3.1: Sources of uncertainty in diagnostic spectroscopy and proposed solutions. Grey highlighted sections are issues mentioned in previous reviews and their proposed solutions. The rest are the author's additions and initial suggestions.



3.4 Digitizing

Data pre-processing

After collecting spectral data from a sample it is important to pre-process it before analysing to get rid of potentially confounding effects. There are a range of pre-processing options, the most commonly used one is background noise removal, typically achieved by acquiring a spectra without the sample present and taking this spectra from the sample spectra directly⁸⁴. 'Substrate compensation' is used mainly in transmission measurements to remove confounding peaks from the presence of the substrate the sample was placed on. This can be done by including the substrate in the background correction, however the scale of the contribution to the spectra from the substrate alone can be different from its contribution to the sample spectra. Therefore a correction scaled to the sample spectra can be employed to compensate for this⁸⁵.

Another common process is baseline removal. This attempts to remove broad inflating curves, due to spurious or constant physical effects, from the spectra so that the peaks can be more easily distinguished from a more uniform, or zeroed, baseline. There are several methods employed to do this, and Raman software packages usually have an option to do this mathematically, often by a process called 'rubberband' baseline removal which stretches a baseline along local minima - setting those to be the new zeroes⁸⁵. One can avoid baseline effects by looking only at the derivative spectrum, though this can be harder to interpret.

In FTIR, a lot of the background curves exhibited in the spectra are due to the Mie effect, in particular resonant Mie scattering. This is when light is scattered instead of transmitting, affecting transmission readings as it doesn't reach the detector. This effect

is greater in tissue samples with larger proteins, but is also present in blood plasma samples and will still therefore have an impact on blood serum⁸⁶. There is an algorithm, termed ‘extended multiplicative signal correction’, that works to remove the effect of this scattering with results indicating better baseline removal than the ‘rubberband’ method. It has the added ability to remove additional confounding peaks, or peak shifts. This method functions by iteratively improving on a spectra from a reference, using each correction as the reference in the next iteration^{87,88}. The ‘rubberband’ or another background correction algorithm^{89,90} could be used for the initial reference if there is not better option.

Intensity calibration is used to help compare between spectra taken on different instruments and laser sources. It is done by taking a spectra of a standard reference material (SRM) on the instrument of interest, processing it to relative spectral intensity using the corresponding standard polynomial and then correcting it to the corresponding standard reference spectra. Then one applies the same correction to the sample spectra to standardise it^{85,91,92}. Typically, spectral SRMs used are manganese doped borate matrix glass from the National Institute of Standards in Colorado, USA.

‘Linearization’ is used in Raman spectrometers that use CCD sensors to normalise the data point spacing. The spacing can be irregular from acquisition because of rounding errors, or simply the use of different gratings and calibration settings. Spacing can be irregular in general, changing as the wavenumber increases. In order to ensure comparability of the spectra produced, the point spacing should be normalised to a defined set of x coordinates via interpolation⁸⁵.

Offset correction is sometimes needed as laser drift over the measurement time period can cause a varying shift in the recorded spectra. This can be corrected by cross correlation where two spectra of the same sample are taken and are systematically offset

and the product of the two spectra taken, the highest value product corresponding to the location of minimal offset⁸⁵. Ideally, this sort of correction should be unnecessary if proper and regular calibration is used.

Noise reduction is often employed to remove unnecessary fluctuations from the spectra so that the more notable patterns can be identified. Of course, there is the risk that over reduction can remove actual features of interest. A common method for this is taking moving averages over a spectra.

In Raman, cosmic ray removal is sometimes necessary as these events produce particle showers on contact with the atmosphere that can randomly impact the detector and cause spurious peaks. These peaks are rare and generally easily identifiable as they are narrower than typical spectral bands. They can be identified and removed by taking the median value from repeat measurements, or smoothed out by either manually removing the data and interpolating or using a, similarly functioning, removal feature present in most acquisition software. Alternatively, W. Pych developed an algorithm to remove the effect from spectra⁹³.

Multivariate analysis

In one review on ‘ratio-metric analysis’, the comparison of the ratio of the two peaks of interest between healthy and diseased samples, highlights its use for quantifying the differences between sample groups and thereby potentially allowing better use of an observed biomarker for medical monitoring and prognosis⁹⁴. The paper also has convenient table listing identified ratios of note from previous studies.

Multivariate analysis is a commonly used tool to look at spectrogram data and ascertain differences between sets of samples with, hopefully, a high statistical accuracy. The

most commonly seen form of this is Principal Component Analysis (PCA) and variants thereof.

PCA

PCA is a statistical procedure that will take a set of spectrograms and use orthogonal transformation to convert them into a set of linearly uncorrelated variables termed ‘principal components’. The process is designed to create the variable with the highest possible variance first, followed by the next highest possible that is orthogonal to the previous component(s) and so on. Typically, one would only look at the first few components, provided that they accounted for a high enough percentage of the variance – which they usually do. Often the data set is plotted as a scatter graph on a 2D map of two of these components, typically the first two components with the highest variance. Often this will result in distinct clustering of the data points, with spectrograms exhibiting similarities being placed together. The hope is that these clusters will match the classes - for example, one cluster being serum from diseased patients, and the other being from healthy.

If the clustering is observed, one can then separate the clusters into classes with a line and make the assumption that any new data point that falls on one side of the line will be a member of that class. The measure of how effective a predictor one’s method is, is termed ‘accuracy’ and is a combination of the ‘specificity’ and ‘sensitivity’ of the method. Using the example of determining if a sample is positive for a disease, ‘sensitivity’ is a percentage measure of how many true positives the method identifies, and ‘specificity’ is a similar measure of the proportion of true negatives⁹⁵.

Receiver Operator Characteristic (ROC) curves are often used on classification methods. They work by modifying the parameter(s) separating two groups in a

classification, producing a graph of how the sensitivity/specificity of the classification would shift as the parameter does. This allows one to bias the classification toward maximising sensitivity or specificity if one is more important for the purpose than the other⁹⁵.

Linear Discriminant Analysis (LDA) is a method similar to PCA, but classification of the datapoints is also included in the calculation. The variance is maximised based on distinguishing the classes within the data⁹⁶.

Regression.

Partial Least Squares (PLS) regression is a method which one can use to produce a model from a dataset of spectra. It is useful when a set of spectra varies depending on several parameters, for example wavelength and aperture size. It finds a linear regression model by projecting an ideal set of your parameters and the spectra onto a new space⁹⁷. It uses the same measure of variance for the spectra/parameters as PCA would, but rotates the spaces to have the maximum linear variation possible, as opposed to just maximum separation of data points. This is particularly good if the data has multiple parameters which affect the spectra linearly. One can then use this model to transform other data to better compare it with results using different parameters. However, the parameters will not always affect the spectra linearly and therefore other methods for modelling will then be more appropriate⁹⁸. Nonlinear PLS regression is also an option⁹⁹. PLS Discriminant Analysis (PLS-DA) is a classification method based on PLS⁹⁷, potentially allowing a diagnostic model to be made via this method.

Once a model has been produced, it is important to test it out on some example data. However, in order to produce the best model possible, one typically uses all the available data to produce it. Several cross validation methods for testing the validity of a model

are available. The most simple is ‘leave-one-out’ cross validation in which a percentage of the sample, often only one spectra, is left out, and the rest is used to calculate the model. Then, the resultant model is tested on the left out data to see if it is classified correctly. Often this is repeated until all of the data has its turn at being left out of the model. If the specificity/sensitivity of the model is demonstrated to be high then the model can be considered valid. Monte-Carlo cross validation is also often used, but the left out subsets are randomly generated and accuracy results are averaged. The disadvantage of this is that the random sampling may lead to random bias⁹⁵.

Machine learning

Machine learning (ML) is a useful and burgeoning method to help automate disease detection. The principle is that you can feed datasets to a program and it would learn to differentiate between their specific features and sort them into categories. Ideally these categories would be, essentially, diseased and healthy samples - and when the program is fed new data it would be able to correctly classify it. The algorithms that do this are mostly termed ‘pattern recognition algorithms’ as they aim to identify distinct patterns in the data.

There are two main approaches to machine learning, supervised and unsupervised. Supervised learning is similar to how one would use PCA and involves taking known datasets and teaching the machine learning model to differentiate between the different subsets using the operator’s defined variables. The process would usually be iterative, to refine the difference between the machine’s result and the expected one. Then one can present it with new data, which it will then be able to classify with, hopefully, a high degree of accuracy. Unsupervised ML removes the operator’s defined variables, only giving the program the data and allowing it to sort the differences itself. The hope with

this method is that potentially new structural motifs in the data will be unearthed by this analysis that the operator didn't know to look for.¹⁰⁰⁻¹⁰²

Again, similar to how PCA and other statistical categorisation methods work, the performance of a particular algorithm can be measured by its accuracy, sensitivity and specificity.

A support vector machine (SVM) is one of the more popular machine learning techniques and is used for devising the optimal line for separating groups in order to maximise the sensitivity/specificity of the grouping^{97,100}. It achieves this by identifying a small portion of the samples, the 'support vectors' and aims to maximise the separation of parallel lines leading from them, the resulting line of separation being the average of the support vectors. Typically this produces a linear separator, but the algorithm can also be modified to work on non-linearly separable data by projecting the data into a higher dimensional feature space and using a plane to separate the two groups. It can also be modelled non-linearly. One of the drawbacks of the SVM method is it not intended for multi-class classification. It cannot be directly used to distinguish between more than two different groups and therefore trying to classify between additional groups involves breaking down the multi-classification into several binary classifications.

A pair of papers by Khan et al. serve as an example of SVM used on Raman data to attempt to diagnose Dengue fever and hepatitis B with corresponding 73%/93% and 97%/100% sensitivity/specificity^{59,61}. This group have also looked at using 'Random Forest' as an alternative to SVM for the Dengue fever case and achieved a higher average diagnostic accuracy of 91% sensitivity and specificity using it⁶⁰. These papers seem to have suffered from the similar pitfalls that have been discussed earlier as the

only control used appears to be healthy vs diseased samples. However, the result that machine learning was able to help produce these high results from the data is still valid. Another study from this group has looked at using machine learning for the study of Hepatitis C infected blood serum in a more comprehensive manner⁶². They compared three methods for examining data variance - PCA, factor analysis and large margin nearest neighbour – and four methods for modelling: SVM, RF, LDA, and k-neural network (kNN). By using fivefold cross validation, they found PCA to be the most useful in all cases. Pairing PCA with kNN yielded the highest average accuracy for their dataset, LDA and SVM were both less than 1% away however and can therefore be considered valid alternatives.

A recent review that found better results using convolutional neural networks (CNN) for analysing vibrational spectroscopy than PLS-LDA and ‘logistic regression’ - in both unprocessed and preprocessed data¹⁰³. The success of CNN, and kNN in the previous study, highlights the use of deep learning in analysing vibrational spectroscopy. Deep learning is a machine learning method inspired by biological neural networks and designed to learn data representations, instead of using task-specific algorithms¹⁰⁴. These artificial neural networks (ANN) produced in deep learning methods are valuable for their fairly universal applicability, combined with sometimes-superior results. They do however require careful training and understanding of the network to avoid the potential of overfitting noisy features^{84,105}.

Genetic algorithm seems useful for feature selection and discerning healthy/diseased spectra, though less useful for making a model⁸⁴.

What constitutes a useful diagnostic accuracy?

When you produce a predictive model, it come with a corresponding accuracy in that prediction. Typically for this field the sensitivity/specificity measures are used.

Sensitivity is the measure of true positives from a classification, whereas specificity is a measure of true negatives. They can be combined to produce an overall classification accuracy.

But what accuracy should one aim for. When is a model 'good enough'?

There are several factors to take into account here. From a statistical perspective, using Bayesian probability a model can have a predictive accuracy of 90%, evenly split between sensitivity and specificity. But if this model is used on a population where only 1% of them are likely to have the disease of interest, then the actual chance of a positive diagnosis being correct for that individual is only 16%. This 16% 'positive predictive value' is based on the 90% sensitivity but also dependant on the disease prevalence, being equivalent at 50% prevalence but lower as the prevalence decreases. The reverse is also true with a 'negative predictive value' diverging from specificity based on prevalence. These probabilities are examined in Figure 3.2. From these curves a few key conclusions can be made for practical diagnosis methods:

1. Of course, increasing the overall accuracy of the model will reduce the likelihood of false positives/negatives.
2. Increasing the likelihood of people within a sample group having the disease of interest massively decrease the chance of false positives (for example, selecting only patients with relatives that have the affliction).
3. Specificity is the more important to bias toward if one is expecting <50% of a population to have the disease of interest. At likely incidence proportions (prevalence) of <10%, a small increase in specificity has a major effect on the positive predictive value, but reducing sensitivity has only a minor effect on the negative. ROC curves can be used to help bias a diagnostic method toward specificity without compromising the sensitivity too much.

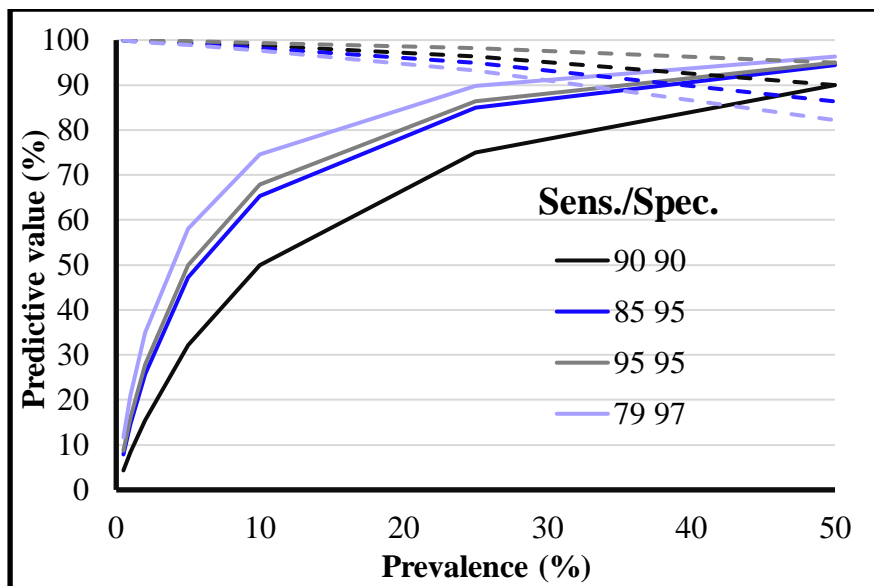


Figure 3.2: The positive (solid) and negative (dashed) predictive values for a diagnosis method with the sensitivities/specificities depicted depending on prevalence of the disease within the population being tested.

Practically, the useful accuracy for a model will also depend on how it compares to other models for the same issue. For example, is the cancer diagnosis model more accurate than current clinical methods? If it is not, it is unlikely to be a useful model, unless it provides some other benefit like cost reduction.

In preliminary studies, it is likely that one will produce a higher accuracy than would be produced when evaluating clinically. This will be due to many factors, for example: the aforementioned instrument to instrument variability or model selection bias. Model selection bias¹⁰⁶ will mean that a researcher is likely to tout the model that they produced the best result with for their dataset. Therefore, it will have been selected with bias for what is optimal for that particular dataset. When introducing further data, the optimisation bias is removed, most likely resulting in a lower accuracy. Cross validation steps should mitigate this, but it is likely there will be a decrease in accuracy to account for.

What constitutes a good sample size for a diagnostic study?

When designing a study, it is key to keep in mind what exactly is required for the study to be useful. More than just accounting for suitable controls and other aspects of good experiment design, one must always take into account the sample number used. If the number of patients is insufficient to produce results with any confidence, then there is little point in running the experiment in the first place. If your 75% accuracy classifier has an error of 25%, that could be the difference between being correct half the time and 100% of the time.

Confidence intervals are a suitable method for assessing the confidence in a particular classification method's accuracy¹⁰⁷. The calculation for this is:

$$Confidence = p \pm c \sqrt{\frac{p(1-p)}{n}} \quad (3.2)$$

Where p is the predictive value of the model used, n is the number of observations and c is a constant based on what confidence interval you are using. These are determined statistically from the distribution function, for example, 90% would use 1.64, 95% uses 1.96 and 99% uses 2.58. For practical purposes 95% confidence is typically recommended, though there may be situations where higher confidence is required to avoid potential extreme negative outcomes.

An example using spectral diagnosis would be: We have a model with a prediction accuracy of 81% ($p=0.81$), the test was done on 20 patients ($n=20$). A 95% ($c=1.96$) confidence interval would result in 0.81 ± 0.17 . This is a little high for a predictive model, as the prediction could actually be as low as 64%. If we tripled the number of patients we would produce a more reasonable ± 0.099 , though this could have been achieved by simply making 3 repeats per patient. 60 patients with 3 repeats each is of course even better with ± 0.057 .

3.5 Biological corroboration

In spectral diagnosis studies, most emphasis goes to the wavenumber values of peaks that shift between classifications and the accuracy that can be achieved in distinguishing them. However, there is little further investigation into the origins of these divergent signals.

As has been mentioned earlier, a few studies so tentatively look at a biological backing for their observed signals^{26,61,83}. These investigations mostly centre around the peak shifts and their corresponding likely chemical origins³¹.

The ideal outcome is that the shifts in the protein composition of the blood samples can be matched to the shifts in signal. This would not only lead to better understanding of the mechanisms by which the spectral biomarker was obtained, but is also useful for quantification of the biomarker. For example, if a varying concentration of particular molecule(s) in a sample can be linked to a shifting classification from its spectra, these specific shifts could be better modelled in future diagnostic methods and better quantified. There are even further implications for research and many potential useful repercussions. For example, the progression or risk or resurgence of a cancer could be modelled and quantified. Another implication is that the causal molecule(s) can be marked and their pathway can be backtracked to a cellular synthesis level. Then marked components could be added into a patient so that a more definite signal can be used to identify/localise cancer expression. Linking spectral biomarkers to real chemical shifts in our bodies is key for many life-saving advancements, yet little is done to link these in current research.

A few promising papers have emerged. One study used SERS and hierarchical cluster analysis of spectral shifts to help discriminate protein biomarkers¹⁰⁸. Another SERS based study aimed to detect tagged pathogens¹⁰⁹. There is even a review on applying

Raman spectroscopy to find molecule specific Raman signatures¹¹⁰. In fact, Raman leads the way in realising the potential of spectral methods with linked biological backing³⁴. One particularly promising Raman based study used a simple approach in which significant peaks were compared to metabolic studies on relevant molecule concentrations during the disease⁶¹. This seems to be the best universal method to use as standard in these studies.

Less has been investigated in Infrared. One study looked at ascites and cirrhosis serum to estimate prognosis²⁶. This used mid infrared fibre-optic evanescent wave spectroscopy to look at various molecule concentrations and certain peak heights and their differences depending on if they were from patients that survived longer than 6 months. The peaks were linked to the molecule concentration via correlation. The resulting method used was demonstrated to be more effective at prognosis estimations than conventional methods.

Distinguishing the origins of the spectral shifts in these spectral biomarkers is not an easy task. Linking the key peaks to likely chemical shifts and relating to metabolic studies has been mentioned as one option. Still, there are many other potential methods for identifying what causes a particular spectral peak, for example one can try labelling specific proteins with heavy atoms and see if the characteristic signals vary between diseased and healthy samples. However, this limits the scope of an investigation to cell or small organism level controlled studies where the metabolism of the cells or organisms can be controlled, e.g. feeding a cell culture heavy glucose to identify increased metabolism in diseased cells¹¹¹.

One simpler method of probing the signals in a bio-fluid solution is to manipulate them by varying conditions, like temperature, charge or pH, and measuring any spectral shifts.

With knowledge about what molecules are likely to be affected by your variable, the molecule of interest can potentially be discerned.

Electrochemistry

Cyclic Voltammetry is a valuable electrochemistry method for evaluating the redox properties of molecules¹¹². It is typically undertaken by varying the potential difference between a working and counter electrode in comparison to a reference electrode and measuring the current. The amount of current produced and therefore the shape of the curve can be affected by the rate at which the voltage is shifted. The curves produced can elucidate important electrochemical information about the molecules involved, for example its stoichiometry or the reversibility of the reaction.

It has been demonstrated that Raman spectroscopy can discern redox changes in biological systems in an in situ study by Brazhe et al.¹¹³. Rat hearts were studied in normal and hypoxic conditions and several peaks were unique to the oxygenated heart. Some of these peaks were attributed to c and b type cytochromes by applying an ‘uncoupler’ (carbonyl cyanide 4-(trifluoromethoxy)phenylhydrazone) to affect their redox state and causing a resultant reduction in their peak intensity. Other variable peaks were assigned to other cytochromes, and to oxymyoglobin using other methods. This ability to detect redox changes in cytochrome c is not only promising for the potential ability to discern more information from redox changes in proteins in serum samples, but will also serve as a suitable method to test a Spectro-electrochemistry setup in future experiments.

Cell study

Cell based studies are another option to probe biological origins of certain signals as they allow for better sample access and the ability to make use of more sample-destructive methods. Any investigation of these would require a healthy and an equivalent cancer cell line, preferably as relatable to blood as possible. Bone marrow has been suggested as one option, though other cell lines are available¹¹⁴.

One study in this vein was comparing metabolism and other pathways in cell lines to normal tissue and cancer tissue¹¹⁵. Cell lines, tumour, and normal tissue cells from 6 different tissue types were compared. Very different expression was found in each type. This included upregulation of nucleotide metabolism, oxidative phosphorylation and downregulation of signals for adhesion and communication. This was mostly attributed to abundance of metabolites in the culture medium. However, the gene expression for cancer emergence and progression was found to be similar between tumour and cancer cell lines.

Further cell-based studies can be a key first step into finding suitable methods to transfer into human based studies that can finally impact the clinical sphere.

3.6 Clinical implementation

The key aim for identification of spectral biomarkers and spectral diagnosis is for the research to eventually impact the clinical sphere. Being used to practically help diagnose disease and ultimately save patients' lives. Yet, little of this has been seen thus far.

A review on the clinical implementation of Raman-based systems highlights the difficulties and occasional successes of establishing these techniques in to clinical practice³⁵. Though much of its commentary is based on tissue analysis, some of the major principles still apply. SERS and TERS are options for Raman, having some success in diagnostic studies. However these methods are more both more costly and difficult to perform, requiring highly skilled researchers and are therefore less reproducible in a general clinical setting. For diagnostic and screening purposes, where ease of use, cost and reproducibility are most key, these methods are currently unfeasible.

Current technology and implementation

There are a few examples of methods and devices that have aimed to enter practical diagnostic use.

There are 2 hand held FTIR studies¹ Glyconics in UK on sputum for COPD, and Malaria screening in Thailand. Though both are still in trial with little evidence of extending beyond that.

A paper outlined a method of using a handheld SERS system for tuberculosis diagnosis from serum¹¹⁶, yet this too has resulted in no further progress.

One system by 'River Diagnostics' has successfully sold Raman instruments for the assessment of skin in the cosmetic industry and another by 'Verisante Technology' has released Raman based devices for oral and skin cancer diagnosis³⁵.

These devices, though promising, remain unaccepted in hospitals. Most likely, they are either too expensive or impractical for actual medical use. With the volume of studies into spectral biomarkers being released, it is disappointing that so little has been successfully linked to implementation in the clinical sphere.

A model for future advancement

Figure 3.3 outlines the pipeline of how a method should be developed, each qualifier having to be met before the next stage can be achieved. First experimentation and the development of a diagnostic model, then expanding it to a viable multi-site process. Next the biological origins and nature of the biomarker can be investigated and quantification of it can be attempted. Beyond that, this quantification can be utilised in the monitoring and prognosis of the disease. Finally the method can be established sufficiently for universal use, helping to screen for and control the disease on a mass scale.

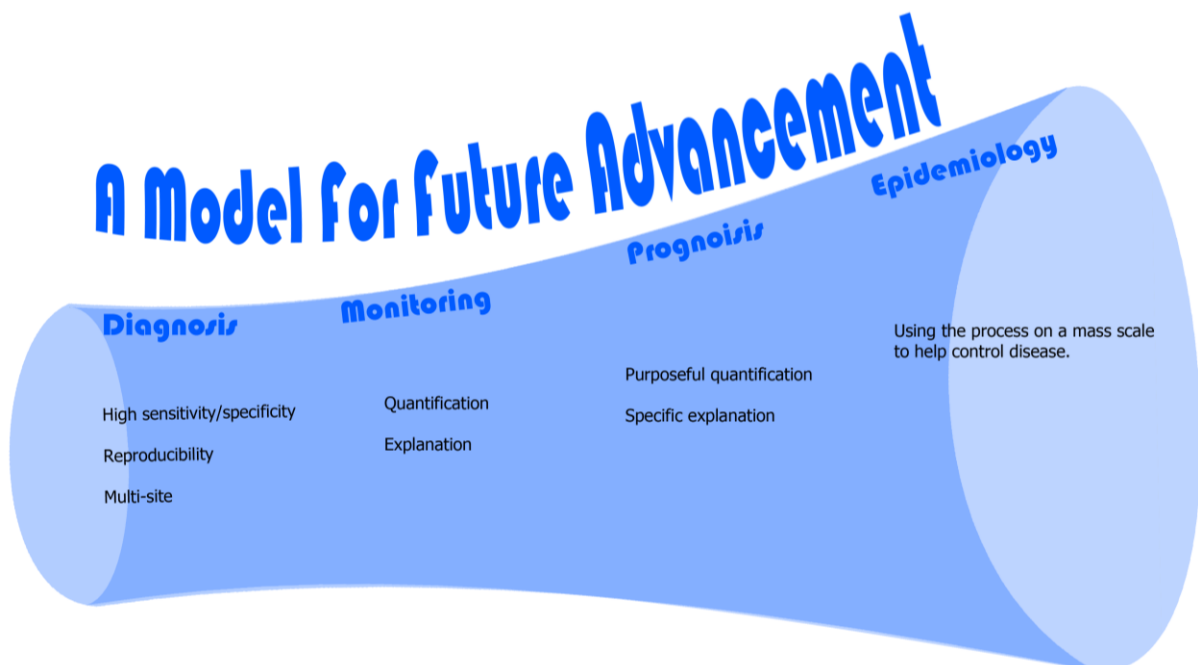


Figure 3.3: diagram of potential future steps to reach clinical viability

In order to achieve these goals, the method first has to be developed to be as valid and reproducible as possible, else it will never achieve even the first stage of this pipeline.

4. Utilised Research Methods

Introduction

The following materials contain additional information not present in the main manuscripts for additional clarity and confidence in the results and greater replicability of the methods. For replicability, sections 4.2-4 are of particular note, the sample preparation and instrument specification.

Sample preparation

- Centrifugal filtration into molecular weight fractions (Figure S3)
- Deposition of serum fractions onto CaF₂ discs (Figure S2)
- Drying of the serum on the disks (Figure S2)

Spectral acquisition

- Clean disc background subtraction.
- Transmission FTIR measurement

Pre-processing

- ALSS Baselineing (Figure S-4)
- Average normalisation

Post-processing

- PCA on spectra (Figure S-7)
- Leave-1-Out cross validation segmentation
- SVM classification on each segment (Figure S-8)
- Optimal PC number obtained for each segment (Figure S-5)

Model development

- Model produced using optimal PC number and the SVM classification.
- New patient samples can now be classified by the produced model

Clinical validation

Clinical implementation in Morriston Hospital (Wales UK) as a diagnosis tool

Figure 4.1: Outline of the process followed for classification of patients from sample collection to clinical implementation in a hospital.

4.1 Sample preparation

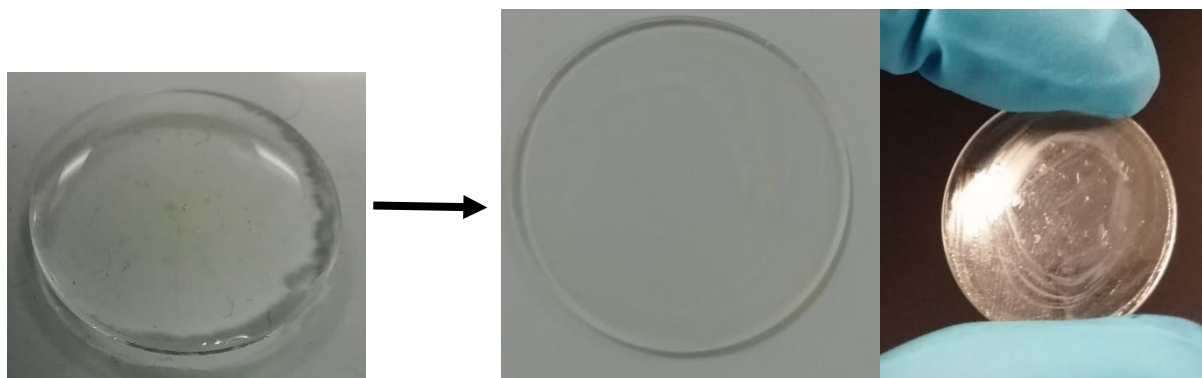


Figure 4.2: images of the droplet deposition before and after drying. Serum fraction was diluted 24:1 and deposited on the disc, ensuring coverage to the edges. The layer is more even than without these precautions, but there is still evidence of the ‘coffee ring effect’ producing rings of higher weight components at the edge of the drying zone.

Droplet deposition

For much to the experimentation, this deposition method was considered sufficient and was used. You can see an example of the drying process used in figure 4.2. 1:24 ratio was determined experimentally. It is selected to give sufficient volume to coat the whole disc while keeping the amount of serum material deposited after drying in acceptable absorbance ranges for the FTIR measurement. Potential sources of error are present from the dilution steps not being sufficient to completely eliminate variable deposition, i.e. the ‘coffee ring effect’. In the HMW sample, there was also evident cracking from the drying process on some samples. Although the location of acquisition was controlled, it is possible some key spectral shifts may have been reduced as a result of these uncontrolled deposition patterns. The consistency of the innermost area (Figure 5.1 from the next section) made it the obvious choice for the controlled detection location, having the lowest average error deviation of 0.45 % absorption (the middle and outer having 0.75 and 8.1 % absorption respectively).

Spray coating

This method was developed later than most of our testing, but had a number of advantages including far quicker drying time of <30minutes and a greater overall spectral consistency across the disc. A blood plasma diagnosis test using it was performed in section 11.2.

4.2 Centrifugal Filtration for Molecular-weight Windowing

Centrifugal filtration

- Place samples in the top compartment of the centrifugal filtration unit (Figure 4.3)
- Centrifuge for recommended time based on kDa rating (10-30 mins) at 14,000 G.
- Dilute sample into 500 μl of ultrapure water before deposition. If >100 kDa concentrate, use only 10 μl . For the rest, use 10-20 μl .
- Let samples dry in ambient conditions (~ 12 hours) before recording IR spectra.

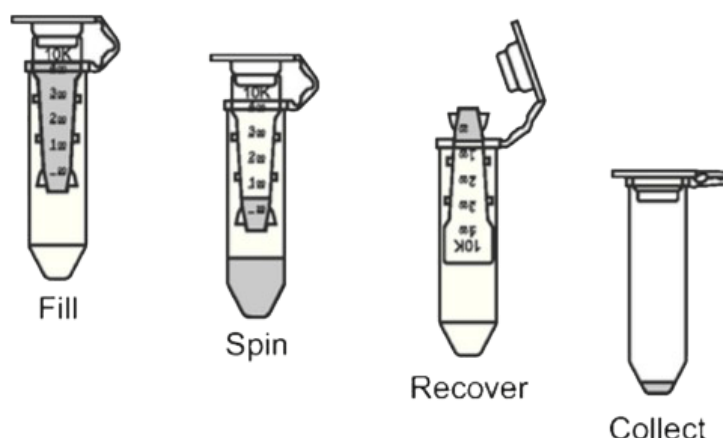


Figure 4.3: Schematic view of concentrating samples using a Merck-Millipore centrifugal filtration unit

The key molecules in blood serum and plasma that could produce large obscuring signals are globulin (>80 kDa) and albumin (>60 kDa)¹⁶, therefore using a 50kDa weight filter will separate out these components from the lower molecular weight ones, allowing their signals to be better discerned.

Molecular windowing

This is when samples are first filtered through a series of filters to produce a set of molecular weight bands that can be analysed individually. Typically this starts using 100 kDa filters, both filtrate and concentrate are collected, and the filtrate was moved on to further filtering using 50, 30, 10 and 3 kDa filters until 6 subsets of sample are produced. Each of the fractions i.e., 0-3, 3-10, 10, 30, 30-50, 50-100, >100 kDa and whole serum are all analysed for comparison.

4.3 FTIR collection and pre-processing

FTIR Instrument Specifications:

A Perkin Elmer Spectrum 2 FTIR was used for the analysis and the spectral data was acquired using the integrated Perkin Elmer Spectrum software. Diagram in Appendix, Figure A1.2. The resolution was 4cm^{-1} , wavenumber range was $750\text{-}4000\text{cm}^{-1}$ and 4 5s acquisitions were taken. Each reading was repeated 3 times for each sample, adjusting the beam location on the disc by 2mm for each repetition. Background of a clean disc was taken beforehand for subtraction. Instrument specifications:

- 8,300 – 350 cm^{-1} long-life IR source,
- Dynascan™ Michelson interferometer with proprietary extended range KBr beam splitter,
- LiTaO₃ detector
- 12mm diameter spot size.

The instrument uses an over-sampling delta-sigma converter, and a small amount of zero filing and cubic spline interpolation.

Pre-processing of FTIR spectra:

Spectra were pre-processed with a background correction using the asymmetric least squares smoothing (ALSS) method. The method uses a smoothing algorithm with an asymmetric weighting of deviations to get a baseline estimator. This allows a corrective baseline to be quickly obtained while retaining the signal peak information. Baselineing was followed by average normalisation by dividing by the average intensity for each spectra. Spectra were also trimmed to $800\text{-}1800\text{ cm}^{-1}$ to focus on the fingerprint region (internal testing produced more efficient classification when doing so) and this also removed some obstructive noise between $750\text{-}800\text{ cm}^{-1}$.

4.4 Post-processing Method using PCA and SVM

It is often observed in chemometric analysis of spectroscopy data that when the markers or biochemical species contributing to a classification are few, commonly fewer than 5 PCs are adequate to explain most of the variance. Therefore, a larger number of PCs are considered to add to overfitting and result in a poor cross-validation accuracy. However, in the case of our FTIR data, the number of cancer biomarkers contributing to classifications are unclear, potentially have multiple components/peaks and are also likely to contribute weaker signals in a spectrum due to their very low concentrations. It is our observation that higher PCs can contribute to increased accuracy, provided the overfitting issue can be handled with utmost care. More recent studies have used higher PC numbers, to good success⁷⁹. We deal with this by rigorous cross-validation to measure any overfitting and to identify a suitable number of PCs to use for the classification.

The complete leave-one-out cross validation removes all the samples from one patient (3 repeats), produces a model of the remaining patients' data and tests it on the 3 left-out samples. This is repeated until all the patients have been left out of the model once and the % of correct classifications then provides the cross-validated sensitivity/specificity/accuracy. This method is repeated for each number of principle components, up to 50, and the accuracies are plotted. With SVM the graph tends to plateau after a certain number of components, mirroring a plot of explained variance (Figure 5.4). The accuracy after the graph has plateaued, and the corresponding number of principle components, were then chosen to produce the ideal model.

It is important to use statistics to discern the spectra. The eye test can pick out certain trends. However, these observable trends can only account for so much of a classification. To reach the higher %, more minor and non-universal spectral differences

must be considered. The use of machine-learning based classification can discern that which our eyes cannot, finding the key minor peak differences and ratios that bring the classifications from the 70% to the 80s and 90s.

PCA

Principal component Analysis (PCA) is a statistical procedure that will take a set of spectrograms and use orthogonal transformation to convert them into a set of linearly uncorrelated variables termed ‘principal components’. The process is designed to create the variable with the highest possible variance first, followed by the next highest possible that is orthogonal to the previous component(s) and so on (Figure 4.4). Typically, one would only look at the first few principal components, provided that they accounted for a high enough percentage of the variance.

$$n \times p = X \qquad t_i = x_i \cdot w_k$$

Figure 4.4: Given our spectral dataset of n patient samples and p intensities at each wavenumber, we can make an $n \times p$ matrix \mathbf{X} . Principle component scores \mathbf{t} can be found for the i -th sample from a transformation \mathbf{w} defined as a set of p dimensional vectors. The first of these has maximum variance and therefore is defined $\mathit{arg\ max}\{\|\mathbf{xw}\|^2$ where \mathbf{w} is a unit vector. As the components are orthogonal eigenvectors of \mathbf{X} , subsequent components can be found by subtracting the previous components from \mathbf{X} and repeating.

Often the dataset is plotted as a scatter graph on a 2D map of two of these components, typically the first two with the highest variance. This can result in distinct clustering of

the data points, with spectrograms exhibiting similarities being placed together. The hope is that these clusters will match the classes - for example, one cluster being serum from diseased patients and the other being healthy samples.

If the clustering is observed, one can then separate the clusters into classes with a line and make the assumption that any new data point that falls on one side of the line will be a member of that class. The accuracy is the measure of how effective a predictor one's method is. It is a combination both of the 'specificity' and 'sensitivity' of the method. Using the example of determining if a sample is positive for a disease, 'sensitivity' is a percentage measure of how many true positives the method identifies, and 'specificity' is a similar measure of the proportion of true negatives. More than 2 components are typically needed for effective classification and thus we need an effective method to draw the optimal separation 'line' for use in a higher (>2) dimensional space.

SVM

Support Vector Machine (SVM) classification is highly, and sometimes the most, effective method used in studies^{62,100}. It also consistently performed well on data collected for this thesis.

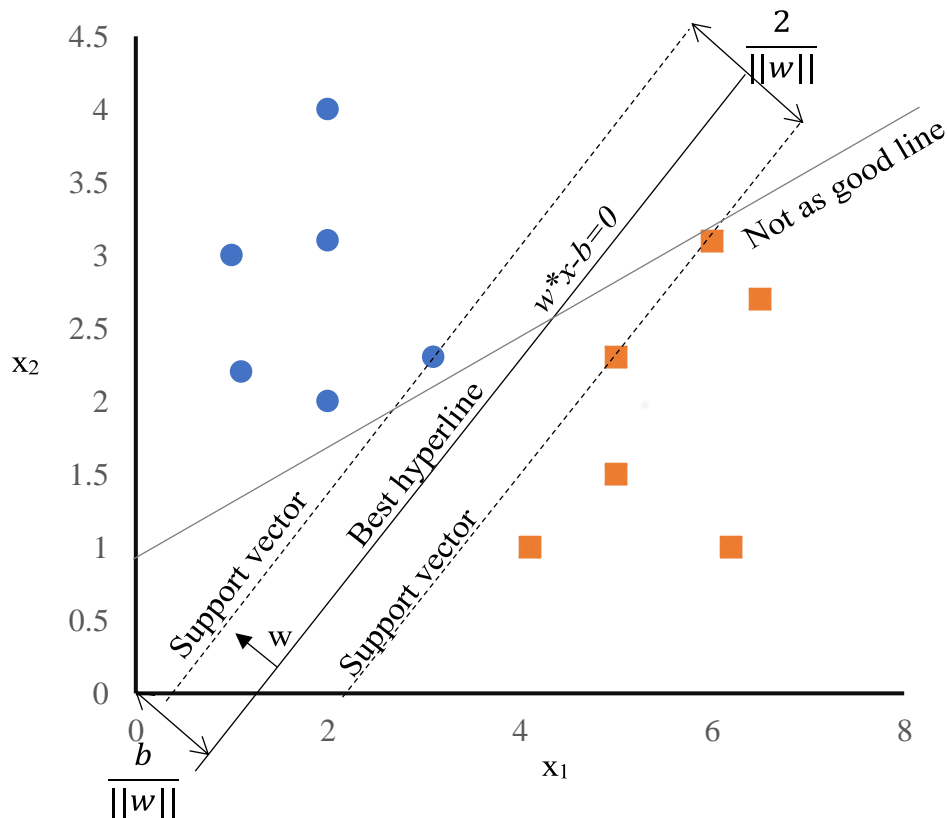


Figure 4.5: Support vector machine (SVM) is a machine learning algorithm that aims to select the best hyperplane between classes. It chooses this by selecting the hyperplane whose distance to the nearest element of each class is the largest, after prioritising correctly classifying each member. It is also robust to outlier data, by opting to ignore these points through a ‘hinge-loss’ function. Through trial and error, it was discerned that linear SVM was most optimal for the data, though it is classifying in many more than 2 dimensions.

Mathematically, the goal of the SVM function is to minimize:

$$\lambda \|w\|^2 + \left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i (w^T x_i - b)) \right] \quad (4.3)$$

Where w is the normal vector to the hyperplane, λ is a tolerance variable and b is defined as $b/\|w\|$ being the offset of the hyperplane from the origin along the normal vector w .

This optimisation of the SVM classifier utilises a hinge loss function where the loss (l) is given by:

$$l = \begin{cases} 0 & \text{if } y \cdot (w \cdot x) \geq 1 \\ 1 - y \cdot (w \cdot x) & \text{otherwise} \end{cases} \quad (4.4)$$

Meaning anything on the incorrect side of the margin's support vector, not just those incorrectly classified, contributes to the loss.

4.5 Measuring confidence

Confidence intervals

Confidence intervals are a suitable method for assessing the confidence in a particular classification method's accuracy. The 'Wald' confidence interval for a Binomial distribution can be given as:

$$\textit{Confidence} = p \pm c \sqrt{\frac{p(1-p)}{n}} \quad (4.2)$$

Where p is the predictive value of the model, e.g. accuracy, used, n is the number of patient spectra and c is a constant based on which confidence interval you are using. These are determined statistically from the standard distribution function. For our purposes, 95% confidence is typically recommended, though there may be situations where either less confidence is sufficient or others when higher confidence is required – e.g. to avoid potential extreme negative outcomes. For 95%, the constant c is 1.96.

Unfortunately, this method has a disadvantage of being skewed at the extremes of predictive value. With a predictive value of 100%, for example, $p=1$ and therefore the expression tends to 0. This is incorrect as even a 100% classification should have a >0 confidence interval. As we are hoping to approach 100% accuracy, having an accurate assessment of confidence at these extreme values is vital.

The 'Clopper Pearson' interval, sometimes termed exact interval, is a method that helps address this inconsistency at extreme predictive values¹¹⁷. It uses the inverse of the cumulative beta probability density function, to produce a similar confidence values to the Wald interval, but with more consistency at the extremes. The interval produced also

will typically have different positive and negative values. For example, a 100% accurate classifier for 50 cancer and 50 healthy patients would have a confidence interval between 96.4 and 100%.

DOME compliance

Table 4.1: DOME compliance summary

Data	The data size is low for a machine-learning based model. However, this is accounted for by the precision of measurement and overall data quality. A larger scale study would be ideal for further validation. There may be a slight bias between healthy/cancer groups based on age.
Optimization	As relatively quick SVM classification is used, full Leave-One-Out (LOO) cross validation is possible. Each model being produced independently of the left-out patient's test samples. Overfitting is accounted for by using an 'early stop' when the number of PCs used starts to reduce cross-validation accuracy.
Model	SVC used from Python module SKlearn.svm, parameters: linear kernel, tolerance 1×10^{-5} Execution time was typically under 5 mins without excessive optimisation.
Evaluation	Statistical method PCA-LDA was also performed on several of the classifications, and produces similar but typically slightly inferior cross validated results. Performance is consistent.

Data Optimisation Model Evaluation (DOME) is a set of controls to use in data science and machine learning to ensure validity and reproducibility of the results obtained¹¹⁸.

Compliance with these guidelines is outlined in Table 4.1.

Spectra were analysed over a range of principle components followed by linear support vector machine (PCA-SVM) classification, and sensitivity and specificity values were obtained by complete LOO cross-validation. All repeats from one sample were left out of each segment and the ability to classify them correctly was assessed. All accuracy values quoted are these cross-validation results. This was done as, if given enough PCs to use, a model will eventually correctly classify close to 100% of samples. Unfortunately, much of this will be due to overfitting the particular sample set's random noise or random confounding factors (due to an individual's diet etc.). Therefore, using the cross-validation values ensures that we are quoting transferable results that could potentially be used to diagnose a new patient sample. These mathematical functions were implemented in Python using in-built functions from the SciPy and Sklearn modules.

5. Results 1: Methodology experimentation and development

5.1 Practical methods for uniform deposition investigation

To ensure reproducibility, methods to reduce or eliminate the ‘coffee ring effect’ when drying serum, and other aqueous samples, for transmission FTIR measurement were investigated. Such a method would need to produce uniform lateral deposition of molecules such that an FTIR measurement at any position in the sample would give the same result and do so without potentially altering the sample or introducing contaminants. It would also need to be being easily replicable and inexpensive for a distribution setting. For example, a method of laser induced drying was found in the literature³⁶, but deemed less suitable for purpose, due to the high cost of the equipment. Potentially, if all experimentation was standardised to take spectra from only the centre or the edge of the droplet, as is currently not the case, this could be less of an issue. However the degree of deposition variability would still be in question, and any difference potentially disruptive to the sensitive art of identifying spectral biomarkers for disease.

Methods

Dilution and small droplet

In the literature, the best solution proposed was to dilute the sample and use as small a droplet as possible. This was demonstrated to reduce, but not eliminate the non-uniform

deposition¹. From our own experimentation, it is clear that the signal from the edge and centre of the droplet was distinguishably different.

Total coating

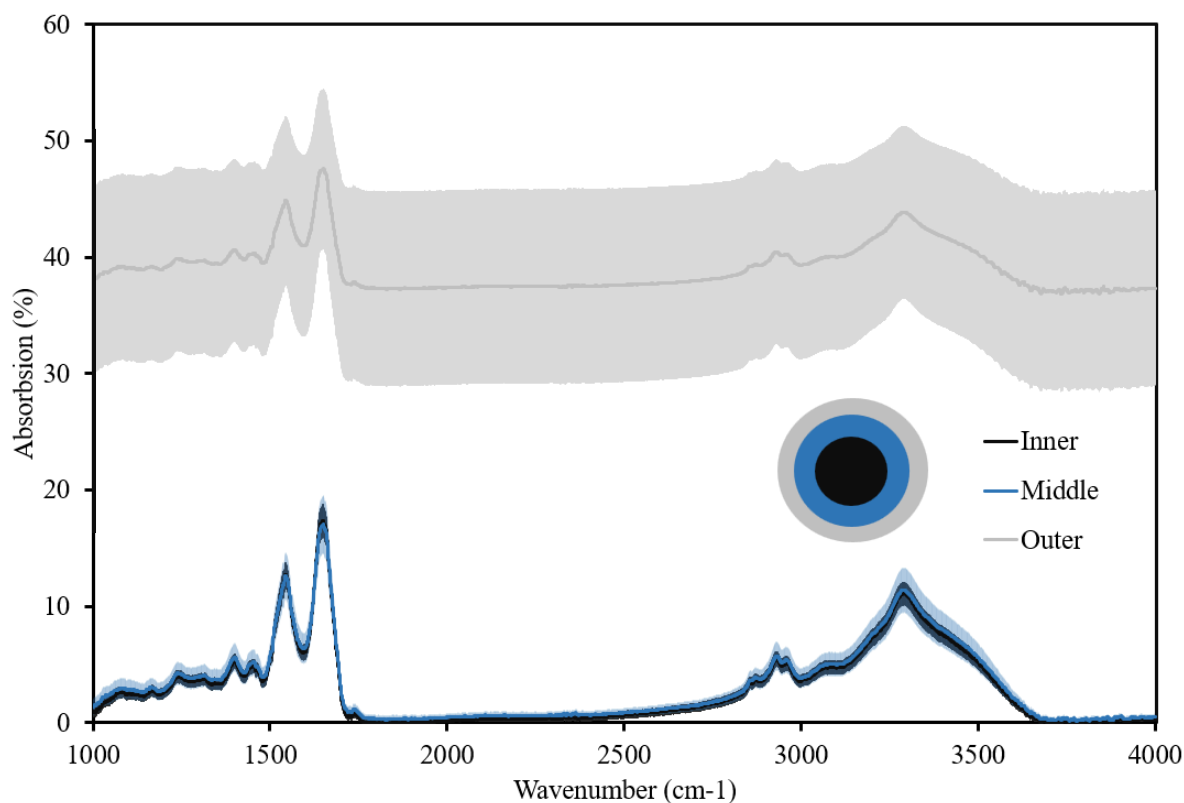


Figure 5.1: Raw absorption spectra averages taken from each disc section by transmission FTIR microscopy. The inner most ~8 mm, the middle ~8-16 mm and the outer ~16-25 mm diameter. One can observe the error increasing as the spectra approach the edge of the disc. The absorption is far higher at the edges due to both the coffee ring effect and contributions of the edge of the disc being within the ~3mm detection aperture.

Total coating was an idea where highly diluted serum would be spread over the entire surface of a CaF₂ disk before drying, the high dilution coupled with surface-edge interaction helping to reduce the deposition differences. However, this was not

completely effective experimentally, as shown in figure 5.1. Additional graph in appendix (A1.1). The innermost area had the lowest average error deviation of 0.45 % absorption, the middle and outer having 0.75 and 8.1 % absorption respectively. The spot size for FTIR spectrometer is 12mm diameter. Therefore, if centred on the disc, the detection area for transmission FTIR would be comfortably in the minimally variable inner and middle sections.

Spin coating

Spin coating is a method where the sample is placed on the substrate, a CaF₂ disk in this case, and the disk is then spun at 250-1000 rpm, leaving only a thin layer of sample which would rapidly dry. However, though the deposition produced is more even, this method is not feasible with aqueous solutions as they dry too slowly. As most human biofluids are aqueous mixtures in nature, replacing the solvent would be an additional time-consuming step and the process would likely alter the key molecules of interest within the sample. Due to these concerns, the method is deemed unsuitable for purpose.

Dip coating

Dip coating is a similar method to leave only a thin layer on the substrate by dipping it in the sample and allowing only the thin layer that adheres to dry onto the substrate. However, the layer formed from aqueous solutions is likely not thin enough to eliminate the coffee ring effect. Additionally, both sides of the disc would be coated, resulting in difficulties with placing the sample for drying and handling the disc into a spectral acquisition instrument. These drawbacks are similar to the ‘droplet deposition’ method, but with additional handling issues. Therefore, it is not deemed a sufficient improvement.

Electrospray

This is a potential avenue for investigation, utilising a high potential to spray a thin layer of tiny droplets of serum over a disc. This would involve developing or purchasing and modifying an electrospray setup for our purposes, which is a high setup cost for a prototype. Additionally, the potential may affect the sensitive biological molecules inside the samples of interest.

Air-spray

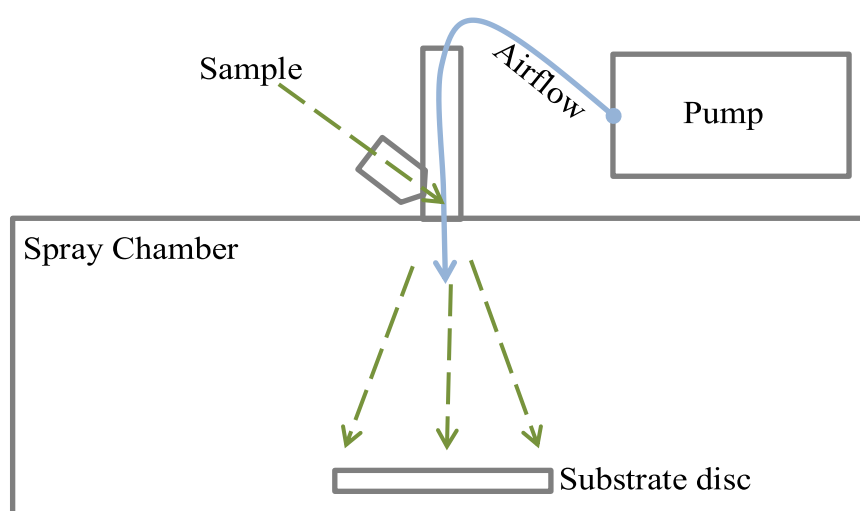


Figure 5.2: Spray deposition setup. The airflow from the pump travels down to the spray gun and jets the sample out as a fine mist onto the disc contained within the sealed spray chamber.

Using an air-pump spraying device is more feasible for prototyping than electrospray, with less of a potential for sample damage. The droplets would likely be larger than one could achieve via an electrospray system. A prototype setup was built using a simple airbrush spray pen. Diluted serum/plasma could be loaded into the pen's input chamber and sprayed onto a CaF_2 substrate (Figure 5.2). In testing, the standard error between points taken along the radius of a disc was 1.40% for this spray deposition method, a decrease from 3.25% from the droplet deposition method. This result suggests it is a

promising method for providing a more uniform deposition. Furthermore, the sample has to be diluted less than droplet deposition, resulting in a much reduced drying time. The discs dried in 10-30 minutes instead of 24h. This is a lot more practical for quick and easy diagnosis as well as reducing the chances of sample errors from contamination or degradation.

5.2 Normalisation and background correction

Background correction methods

Manual

A method used in some papers^{64,65} is to go through each individual spectra obtained and manually plot a baseline curve to flatten out the background. This is functional for their research, but had the capacity to be heavily dependent on the individual performing the baselining and is difficult to reproduce with any consistency.

Automated mathematical transformations

This is a method that is typically encountered implemented as a feature of a software package or can be coded into a baselining program. These types of method typically look at each spectra individually and mathematically automate finding a baseline for each than can be subtracted from them. The method is reproducible, but is only as effective as the maths allows. Typically they have one or more adjustable parameters that will affect tolerances within the method. These will, for example, be used to decide if a spike is characterised as a peak or just some noise in the spectra. Potentially two similar spectra could be transformed substantially differently if one of their parameters just exceeds a particular tolerance of a method but the other spectra does not.

Rubber-band

Rubber band baseline correction looks at each spectra individually and identifies the ‘troughs’ in a spectra and stretches a ‘rubber band’ between them to produce a baseline curve that is then subtracted from the data¹¹⁹.

ALSS

Asymmetric least squares smoothing (ALSS) is another mathematical transformation. It is characterised as being a quick, easy to use baselining method that used a ‘smoother’ to produce a trendline and asymmetric weighting of deviations from it to produce an effective baseline⁸⁹. An example of this being used is shown in figure 5.3.

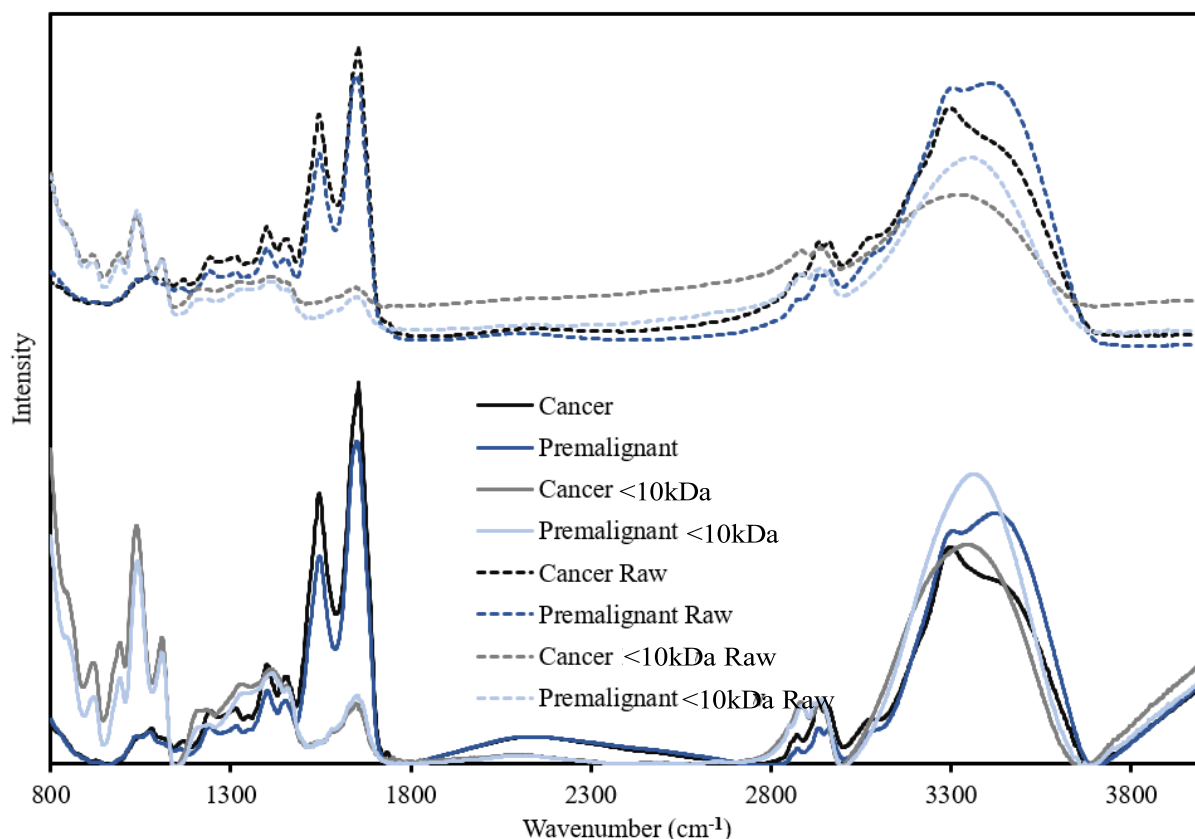


Figure 5.3: Transmission mode FTIR spectra before (‘Raw’, dotted) and after baselining and normalisation (solid). A set of cancerous and premalignant samples were used for this example, showing both whole serum and <10kDa subsets.

Intelligent background correction

The intelligent background correction was developed for Raman spectra in order to effectively remove the fluorescent background without compromising the key peaks of a spectra. It does this with continuous wavelet transforms⁹⁰. Specifically it utilises the

‘mexican hat’ wavelet and the ‘Haar’ wavelet. On top of this it then implements background fitting using least squares and binary marks.

EMSC

The extended multiplicative signal correction (EMSC) is a method employed in near-infrared, Raman and FTIR spectroscopy of biological materials^{120,87}. It is a little different from the other methods as it is model based, requiring an ideal ‘model’ spectra to be input and the other spectra taken are then corrected towards it while, in principle, maintaining their key spectral differences. As it corrects to a model spectra, it also does not require normalisation whilst the other methods outlined typically do.

Normalisation methods

Normalisation allows for greater comparability of spectra of materials that potentially have slightly different overall concentrations. It allows the relative proportions of the constituent materials to be better evaluated. Normalisation should only happen after baseline correction because having non-zero minima will greatly affect the amounts a spectra is normalised.

Maximum normalisation

This normalisation method takes the maximum spectral reading and normalised the spectra to that value. For example, the ‘Amide I’ peak is typically the largest in a FTIR fingerprint-region spectrum of serum. Therefore, normalising around the maximum will typically allow all of the spectra to be compared relative to their concentration of ‘Amide I’. This will encounter issues with noise spikes, especially from things like cosmic rays in Raman spectra. Therefore it is recommended to select a wavenumber for the maximum peak, (for example, 1456cm^{-1} for FTIR of ‘Amide I’) and just normalise around the (absorption) value at that wavenumber.

Average normalisation

Average normalisation is where you average all the y values for a spectra in the range of interest, then dividing all the y values in the spectra by that average. This will produce spectra that can be compared by the relative concentrations of all their constituents. This can be affected greatly by large sample-specific peaks, which would effectively reduce the overall intensity of the other peaks of the spectra, because of the larger average y value compared to those without the additional peak.

5.3 Programming and Machine learning model development

Cross validation selection

There are many options available when it comes to cross validation. Due to the nature of available patient numbers, some of the groups to be classified between would contain relatively few data points. Therefore, creating a separate ‘validation set’ would not always be viable as the reduction in sample number would increase the error by too large a margin. The advantage of lower sample numbers is that iterative cross-validation is less process-intensive and therefore more viable. Therefore, complete cross-validation was chosen as the most suitable method. To limit the sample number reduction, leave-one-out cross validation was chosen. However, for each patient there were 3 repeat measurements taken. If some of those were included in the classification set, it would bias the cross-validation. Therefore, the leave-one-out would practically be a leave-3-out where all the repeats for a patient would be taken out together.

Principle Component number selection

What number of PCA components (PCs) should be used in the classification? LDA and SVM models tend to overfit the data without PCA, resulting in a high disparity between the classification accuracy and the accuracy when cross-validated. This is especially prevalent when lower sample number are used (Table 5.1).

Table 5.1: Example comparison of SVM classification accuracies vs leave-one-out cross-validated (CV) accuracies. Example uses fractions of blood serum.

Fraction (kDa)	No. Cancer	No. non-cancer	Classification Accuracy	CV Sensitivity	CV Specificity	CV Accuracy
< 3	9	9	100	81	88	84.5
3 to 10	9	9	94.4	70	74	72
10 to 30	9	9	98.1	59	70	64.5
30 to 50	9	9	98.1	74	77	75.5
50 to 100	9	9	94.4	62	74	68
> 100	9	9	98.1	66	77	71.5
whole	9	9	98.1	77	74	75.5

Selecting a low number of PCs to use for the classification is a tactic utilised in prior research⁶¹. Using 5 PCs produced more consistent results between the accuracies before and after cross validation, and higher cross-validation accuracies in general when using lower sample numbers (Table 5.2). However, when sample numbers increased, the effectiveness decreased dramatically.

Table 5.2: Example comparison of 5 component PCA-SVM classification accuracies vs leave-one-out cross-validated (CV) accuracies for low/high sample numbers. This example is on fractions of blood serum. Many of the serum fractions are expected not to classify well, whole serum is expected to classify with around 80-90% accuracy.

Fraction (kDa)	No. Cancer	No. non-cancer	Classification Accuracy	CV Sensitivity	CV Specificity	CV Accuracy
< 3	9	9	92.5	92	80	86
3 to 10	9	9	88.8	70	85	77.5
10 to 30	9	9	85.1	59	92	75.5
30 to 50	9	9	79.6	52	80	66
50 to 100	9	9	88.8	77	85	81
> 100	9	9	85.1	76	79	77.5
whole	9	9	83.3	74	85	79.5
whole	31	41	79.6	27	92	64.0
whole	42	82	72.5	20	97	70.9

It is possible to produce a range of cross validation accuracies, plotting them on a graph in order of increasing numbers of principle components used in the classification. By this means, the graph can be investigated to find what might be the optimum number of principle components to use for that classification (Figure 5.4). From the figure, which is typical for a promising human biofluid classification, the accuracy curve with increasing PCs mimics an explained variance graph. There is a plateau after a certain amount of principle components after which the SVM classification only varies minimally (randomly from noise) with increasing PCs. Notably, there appears to be no significant overfitting with the higher numbers of PCs, the SVM classifier managing to

ignore noise contributions. The first 40 PCs are shown in figure 5.4, though the accuracy remains consistent for at least the first 100 components. Thus, the classification model should optimally be produced using the classification with any number of number of PCs after the graph's plateau. For efficiency, the initial value or early peak of the plateau is suggested.

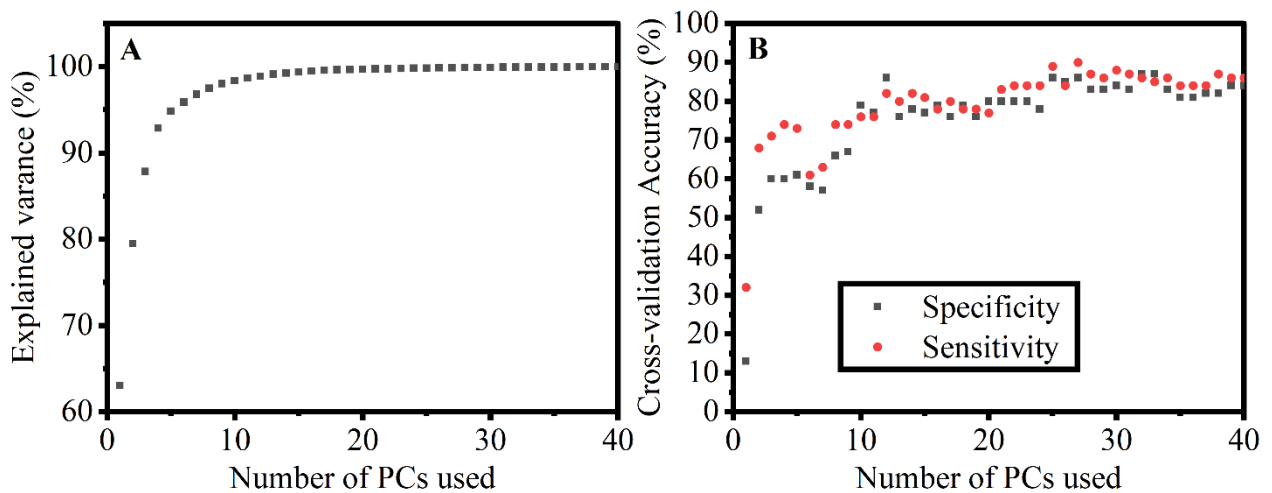


Figure 5.4: A) Explained variance graph depending on the number of principle components (PCs) used. B) Cross validated sensitivity and specificity values dependent on the number of PCs used in the model. The example graph demonstrates how the accuracy plateaus after a certain number of principle components. Before the plateau, the classification is not yet optimised, resulting in lower accuracy. The cross-validated accuracy does not decrease after a point as the SVM algorithm ignores the unnecessary components and minimal or no overfitting occurs. The two graphs mimic one another, the plateau in B starts at 25 principle components where in A there is 99.84% variance explained. This example is from the classification of the ‘Whole plasma’, ‘Cancer v Premalignant’ subset.

Classification method selection

From the earlier review, it was key that using PCA is a suitable foundation for reducing the chances of overfitting. However, the choice of classifier to use on the PCA-transformed data was less obvious. Linear discriminant analysis (LDA), Support Vector Machine (SVM), Random Forest (RF) and Neural Networks (NN) had all been used to good success in earlier work. RF seemed the least effective of those, not managing to surpass the others in any comparative studies^{60,61,102}. NNs were usually comparably effective as the other methods, but their implementation was more complex and would be likely to require longer processing times as a result. LDA and SVM were the clear choices to use, both producing similarly effective results in many studies^{61,62,100}.

Internal testing on our preliminary Buccal mucosa data, classifying between cancerous and non-cancerous patients, yielded similarly effective results for both methods (Table 5.3), with SVM generally equivalent with LDA but using fewer PCs to do so. However, in smaller sample-size tests, occasionally SVM would underperform. LDA was also the faster method, taking anywhere from 1/2 to 1/10th of the time to output.

Table 5.3: Example comparison of PCA-SVM classification accuracies vs PCA-LDA accuracies. This example is of Raman spectroscopy on fractions of blood serum. Many of the serum fractions are expected not to classify well, whole serum is expected to classify with around 80-90% accuracy.

Comparison/ Fraction (kDa)	No. Cancer	No. Non- cancer	PCA- SVM No. of PCs	PCA-SVM Classification Accuracy	PCA- LDA No. of PCs	PCA-LDA Classification Accuracy
PvC/Whole	30	34	7	81	7	80
PvC />50	30	34	43	52	25	52
PvC /<50	30	34	15	76	19	76
CvH/Whole	30	27	6	82	8	82
CvH />50	30	27	43	62	37	63
CvH /<50	30	27	9	80	9	81
CvP/<3	9	9	3	55	46	77
CvP/3-10	9	9	7	53	47	82

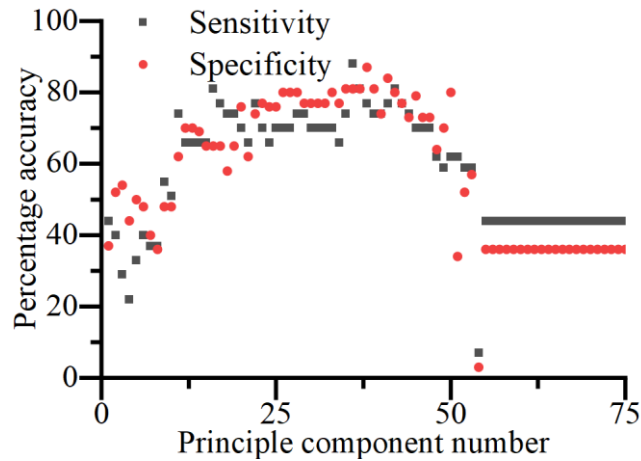


Figure 5.5: Cross validated sensitivity and specificity values dependent on the number of PCs used in the model for PCA-LDA. One can see that, after a certain point, there is evident overfitting with the higher numbers of components.

From the plateau of the accuracies after the initial increase in Figure 5.4, PCA-SVM is quite resistant to overfitting at higher sample numbers. This is not the case for PCA-LDA. One can see in Figure 5.5 that the cross-validation accuracies sharply decline after an initial peak. This suggests that the noise present in the later principle components is being ignored in SVM, but being incorporated into the LDA classification. Therefore, one should be careful using PCA-LDA on higher principle component numbers.

Using SVM, it is possible to make use of non-linear kernels, essentially ‘lines’ of separation that are not straight, based on polynomials or various other options. It might be expected that these would be useful for improving the classification. Practically, this did not prove to be the case.

Therefore, linear SVM was chosen as the classification method of choice due to its high accuracy and greater resistance to overfitting, with LDA occasionally used for preliminary testing and with small sample sets due to its speed and easier handling of fewer data points.

5.4 Methodology experimentation conclusions

The various methods that were considered viable were tested to find out which were most fit for purpose. Ultimately, a narrow selection of these were included and compiled into the utilised methods outlined in Chapter 4.

These methods that were deemed most viable are:

- Sample preparation: diluted droplet and spray deposition onto a CaF₂ disc
- Baselineing: ALSS baselineing
- Normalisation: Average normalisation
- Classification method: PCA-SVM Classification assessed over a range of principle components, and PCA-LDA for low sample number testing.
- Cross-validation: Complete LOO cross-validation

6. Results 2: Oral (Buccal Mucosa) Cancer

This is a result chapter, using data collected on two research trips in 2019 and 2020 to collect oral patient data from Mumbai, India. The FTIR data from both trips were later collated a published research article¹²¹. The Raman data from the original trip was less effective and therefore the methodology was altered in the second one. See appendix §2.1 for the initial concept and Raman and FTIR results from the first trip alone.

Introduction

Blood is a particularly useful bio-fluid for inspection due to its high protein and lipid concentration - as changes in these levels are the some of the best indicators of disease. Much of the current research is focused on subsets of blood, the plasma and serum. In whole blood, haemoglobin and other red blood cell associated molecules can interfere with the spectra, so the plasma is preferred as the variable protein concentrations within are more sensitive to disease. Serum is a subset of plasma, without the coagulating factors, which enables easier storage and use. Without these natural coagulants present, other de-coagulating chemicals do not need to be added. This is beneficial as some of these added anticoagulants have been demonstrated to also produce confounding results¹⁵.

Characterising a serum sample to quantify the minute quantity of markers is key for being able to tell if it is diseased or not. There are numerous methods by which this can be done, to varying degrees of accuracy. One of the best methods we have devised for this analysis is vibrational spectroscopy, particularly for it being a non-destructive procedure, allowing us to examine a sample in as close to a natural composition as possible with any labelling. Some of these methods require that the sample be dried, though there are some methods that will analyse an aqueous sample as well.

Much of the investigation of the vibrational spectroscopy of blood plasma and serum for the purposes of diagnosing and screening of disease are proof-of-principle studies. These typically demonstrate the potential of FTIR or Raman spectroscopy to distinguish between diseased healthy samples on a relatively small sample set. Of these studies, many of them are investigating cancers in the attempt to distinguish characteristic spectral biomarkers for them.

In one exemplar study, FTIR was used on human blood serum and plasma looking at colorectal cancer⁴⁴. The study identified deviations in peaks at 1141 and 1105 cm^{-1} in plasma and 1082, 1050 and 1302 cm^{-1} in serum between healthy and cancerous samples. Another study looked at non-small cell lung carcinoma using Random Forest (RF) and maximum relevance, minimum redundancy (MRMR) to select features of interest⁴⁶. The novelty in this study is that it attempted to distinguish between patients with cancerous and non-cancerous lung diseases, not cancerous and healthy as this had been looked at in a previous study. It was possible to differentiate cancer from other diseases with a 79% accuracy and also extended to differentiate the specific type – distinguishing between squamous cell and adenocarcinoma with 80% accuracy⁴⁵.

Other studies used Raman spectroscopy when looking for spectral biomarkers. A review on gastric cancer detection⁴⁷ highlighted a series of papers that used silver nanoparticle-based surface enhanced Raman spectroscopy (SERS) to look at proteins purified from plasma samples and distinguished between the healthy and unhealthy samples with 100% accuracy⁴⁸. It then looked at just the serum with RNA SERS, achieving 100/94% sensitivity/specificity⁴⁹. Furthermore, another study looked deeper, reporting that the SERS peak heights were distinguishably different, peak heights increasing between benign disease, early stage and late stage gastric cancer⁵⁰. They didn't quantify this observation with a sensitivity/specificity model however.

There is even a study evaluating the usefulness of two variations of Raman spectroscopy for the application of identifying spectral biomarkers. Both spontaneous Raman and SERS were used on blood plasma in order to distinguish between benign gynaecological conditions and ovarian cancer sufferers⁵³. They identified 5 peaks of interest to be their spectral biomarkers and achieved 94% and 96% sensitivity and specificity respectively with spontaneous Raman as opposed to 87% and 89% with SERS. Furthermore, early ovarian cancer cases were diagnosed with 93/97% for Spontaneous and 80/94% for SERS. Notably, they also evaluated the effect of age on their results and, though there was some decrease with age, the overall accuracy remained high over all age ranges. These results suggest that SERS, though a powerful tool in certain situations, does not improve sensitivity/specificity in cancer detection from serum – the simpler, spontaneous Raman producing more discernible classification.

FTIR or Raman methods are often chosen individually for the study of a particular disease, most likely due to access to certain equipment at particular laboratories. Occasionally, there are studies that make use of both the complimentary FTIR and Raman spectra in their investigation. For example, Kraft et al. looked for cancer spectral biomarkers in extracellular vesicles in blood serum using both methods, with prostate cancer as the chosen example⁵⁵. They observed differences between several peaks in both of the collected spectra, though no sensitivity/specificity model was used to quantify the observation. However, the proposed benefit of using both methods in tandem for improved detection seems obvious, although a single more accurate method would be more economically viable in a practical disease-screening scenario.

Recently, there has been a study demonstrating better quantification of molecules such as glycine in serum by only using the <10 kDa fraction as it removes large obscuring signals from globulin (>80 kDa) and albumin (>60 kDa)¹⁶, so maybe even subsets of

serum will be more accurate for the identification of spectral biomarkers for certain diseases. Further research using ATR-FTIR has been done in this area, looking at using ultra-filtration on samples to get better detection of the low molecular weights^{69,70}. Therefore, it is necessary to see if this effect can be transferred both to transmission FTIR and it could perhaps even improve with Raman detection.

It is apparent that by finding a biomarker for cancer within a subset of the serum, the ability to discern the molecules originating the signal would be improved without the unnecessary obscuring molecules and signals. Thereby the spectral biomarker can be connected to the real change in blood molecular concentration caused by the cancer - or the body's reaction to it. Finding this connection would be a major step in the field of spectral diagnosis.

Buccal mucosa was chosen as a suitable cancer to test the potential of fractionating serum before analysis due to recent research into its potential for screening by Sahu et al.⁶⁴ where the feasibility of classification was explored before being followed up by a larger and more comprehensive study⁶⁵. The latter study contained suitable premalignant and related disease controls and produced sensitivity and specificity values of 64 and 80% respectively in determining the presence of an abnormality, and higher values for determining the correct abnormality from the glioma, premalignant and oral cancer options used in the model. It was noted that these values are comparable to current screening techniques.

In this research several research questions were investigated:

- To investigate how diagnosis accuracy obtained using FTIR and Raman spectroscopy compare to one another.
- To investigate how the results using the suggested methodology on whole serum compares to those reported in previous literature for Buccal mucosa cancer.

- Serum will be additionally be filtered to separate out its molecular weight components to see if, by removing obscuring molecules, detection accuracy will be affected.

6.1 Raman spectroscopy study to classify buccal mucosa cancer

Methods

Sample Preparation

In this study, certain factors that could influence the serum spectra such as age, sex, diet, certain habits e.g. smoking, pre-existing conditions or other diseases were controlled. Though little could be done to control diet in this particular experiment, efforts were made to eliminate or control for the other potentially obscuring factors.

The blood serum was collected from male patients from the Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Navi Mumbai, India.

Patients: 28 oral cancer and 28 premalignant oral condition (leucoplakia) patients as well as 17 healthy and 17 healthy tobacco users as an additional control. It should be noted that all the oral cancer and premalignant patients were also tobacco users in both parts of this study. Samples were stored at -80°C until being thawed for analysis.

The serum was separated into two fractions using Millipore 500 μl 50 kDa centrifugal filters. The centrifuge was run for 20 minutes at 14000 g. Whole serum, <50 kDa low molecular weight (LMW) and >50 kDa high molecular weight (HMW) fractions were analysed. Molecular windowing, as outlined in §4.2 was also utilised on 9 cancer and 9 premalignant patients, using 50, 30, 10 and 3 kDa filters until 6 additional subsets of serum were produced.

In Raman spectroscopy, if thin or even slightly transparent samples are investigated then one has to choose a suitable substrate to hold the sample. Glass tends to have a high background fluorescence signal in Raman measurements, therefore Raman grade

calcium fluoride was used as it only exhibits significant Raman peaks at $<600\text{cm}^{-1}$ for lasers excitations ranging between 473-830nm^{32,33}.

For the Raman measurement each fraction was diluted in a 1:3 ratio before 1 μl was deposited on a Crystran Raman grade CaF_2 slide and left to dry for 30 minutes.

Spectral acquisition

Raman spectra were taken using a WiTec alpha 300 spectrometer with a 532 nm laser over the relative wavenumber range 0 to 3500 cm^{-1} . The Resolution was 2 cm^{-1} , an objective lens with a 10x magnification and 0.25 numerical aperture was used along with a 1200 g/mm grating. Spectra were taken at a laser power of 27 mW for 10 seconds with 3 accumulations. Scans were taken over a wavenumber range of -100 to 3500 cm^{-1} to investigate the entire spectra for useful signal regions, this resulted in approximately 24 minute acquisition times. Prior to acquisition the spectrometer was calibrated using a silicon reference at 520 cm^{-1} . Spectra were acquired from the centre of the small, dried droplet.

Pre-processing of spectra

Raman spectra were pre-processed by ALSS and average normalisation as outlined in §4.3.

Post-processing of spectra

Spectra were analysed by PCA-SVM as outlined in §4.4, with LOO cross validation, as described in §4.5.

Results

Table 6.1: Raman cross-validation accuracy results for classifying between Buccal Mucosa Cancer samples from healthy and premalignant.

Sample	Cancer/Healthy			Cancer/Premalignant			Cancer\Non-cancer			Average Accuracy
	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	
LMW	61	62	62	62	41	52	44	48	46	53.3
HMW	80	80	80	73	78	76	74	54	64	73.3
Whole	87	77	82	83	78	81	78	59	69	77.3

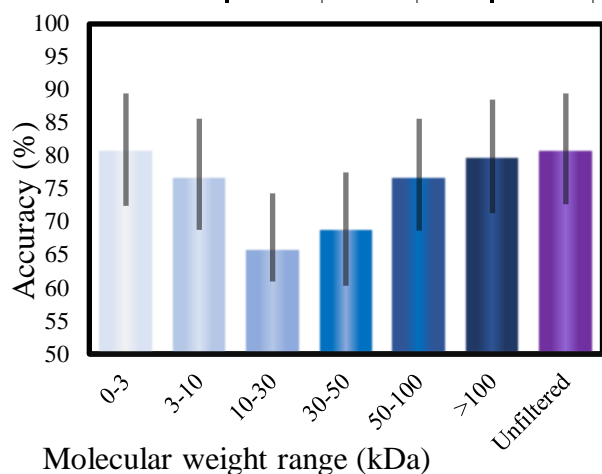


Figure 6.1: Raman accuracies for each molecular weight subset. 95% confidence interval shown by the grey lines atop the bars.

The results in Table 6.1 and Figure 6.1 were similar to the Raman results obtained in the initial research trip outlined in Appendix §2.1. Though the new PCA-SVM method did improve the accuracy, none surpassed 82%.

The molecular windowing accuracy was similarly low, though does follow the same pattern of accuracies as seen in the FTIR results in §6.2.

Discussion

It is valuable to compare to a spectral reference to begin to try to discern the root biological cause of the spectral shifts observed. Knowing which features are present in higher concentrations can help to discover the molecules that contain them. Furthermore, if there is relevant literature on blood composition in the disease of interest, any parallels between the data will provide more clues. This information can then be used to suggest or even perform a follow up study into the root cause of the spectral biomarker identified.

The method of PCA-SVM did show improvements for the Raman accuracies from the results obtained in Appendix §2. The Whole serum was especially promising reaching accuracies as high as 82%. Though improved, they are still not sufficient to compete with those obtained by FTIR. This continued in to the windowing study. The trend of the Raman accuracies mirrored the FTIR study (See manuscript 1), with the <10kDa and whole serum being the most promising subsets, but ultimately still were lower than the corresponding FTIR accuracies.

The accuracy in the Raman study, though lower than corresponding FTIR results, is still slightly above the accuracies reported by Sahu et al.^{64,65}, though the Cancer v Non-cancer group was lower.

Though exposure times were sometimes long for Raman, the laser intensity was confirmed to not be powerful enough to burn the sample, as clear sample damage was evident from a black burn mark as well as severe deviations in the spectra produced when tested using higher power and magnification.

Other potential sources of error come, again, from the drying of the sample droplets. The dilution steps were not sufficient to eliminate variable deposition. In the HMW sample there was also sometimes evident cracking from the drying process. Although

the location of acquisition was controlled, it is possible some key spectral shifts may have been reduced as a result of these uncontrolled deposition patterns.

Use of a 532 nm laser over a 785 nm seems to have produced a similar spectra to the literature²², however it is possible that the higher excitation wavelength would have been more suitable for biological samples as it offers a better compromise between high CCD quantum efficiency and reduced fluorescence^{25,30}.

Conclusions

If the signal to noise ratio and sampling inconsistencies from the Raman spectroscopic investigation can be improved, then more of an effect could be observed. Furthermore, the molecular windowing showed even greater promise from its even higher classification accuracy, though the sample size was comparatively small.

The study would benefit from an additional related, but non tumorous, disease control to examine further the cancer specificity of the spectral biomarkers observed³⁵.

The usefulness in making use of both Raman and FTIR together for screening cannot be realised until the Raman side is improved, though methods like SERS to improve sensitivity would be less viable for screening due to increased complexity and cost. It is likely that in large scale, multi-site testing prediction accuracies would decrease due to detection instrument discrepancies and the increased potential for errors and site-specific discrepancies that comes with a larger scale project. Other spectroscopy and general serum analysis methods, like nuclear magnetic resonance or mass spectrometry, for screening have been attempted with some success^{17,122}, adding one of those to the methodology could be considered instead of Raman to help improve detection.

However, understanding the chemistry behind these spectral biomarkers will bring more confidence in the machine learning outcome than just adding more abstract data. It is imperative to discern the root shifts in protein or other biological molecule

concentrations to validate the spectral diagnosis method. The molecular windowing portion of this study will help with this, as it narrows down the scope of investigation to a key band of molecular weights.

6.2 FTIR spectroscopy study to classify buccal mucosa cancer

The FTIR data from both research trips was collected using the same methodology and was therefore collated into this analysis, resulting in the Analytical Chemistry journal article Duckworth et al. (2022)¹²¹.

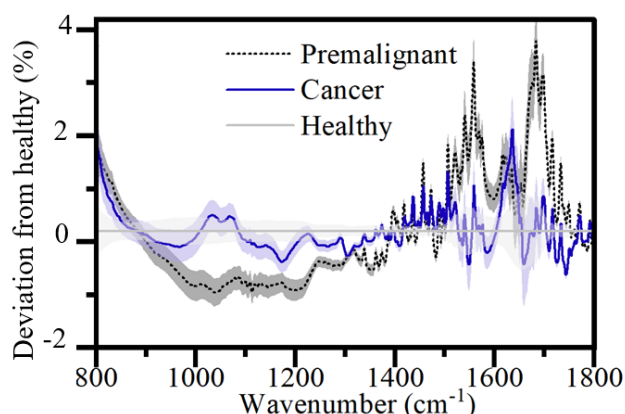


Figure 6.2: Difference in the average spectra of cancer and premalignant patient serum from healthy for whole serum. Error in faded colour around each line to show level of distinction for each spectrum.

Beyond the PCA-SVM used before, further investigation was carried out to compare the cross-validated accuracies obtained from SVM and LDA alone i.e. using the FTIR absorption spectra data directly, as opposed to using the relevant number of PCs for PCA-SVM analysis (see Table 6.2). It can be clearly observed that the PCA-SVM analysis produced higher results on average compared to SVM or LDA analysis alone.

Results and Discussion

Comparison of whole sera with LMW, HMW fractions:

The average spectral differences for whole serum can be seen in Figure 6.2.

The background subtracted spectra (pre-processed) were used to calculate the principal components (PCs). An example of the variation of accuracy and specificity with the number of PCs is shown in Figure 4.4. The highest accuracy and specificity

Table 6.2: FTIR cross-validation sensitivity (Sen), specificity (Spc) and principal components (PCs) results for classifying between buccal mucosa cancer samples from healthy and premalignant using PCA-SVM. Post cross-validation results using LDA or SVM alone are also included for comparison, demonstrating a similar accuracy trend but with lower accuracies overall.

Fraction	Classification of cancer and healthy			Classification of cancer and Premalignant			Classification of cancer and all other			Average Cross-Validation Accuracies (%)		
	Sen (%)	Spc (%)	PCs	Sen (%)	Spc (%)	PCs	Sen (%)	Spc (%)	PCs	PCA-SVM	LDA	SVM
LMW	88	88	29	83	84	46	65	81	30	82.3	76.5	77.2
HMW	94	82	10	83	83	15	81	89	24	86.1	83.9	83.1
Whole	89	86	29	90	84	27	84	90	43	87	82.7	79.7

combinations were chosen for analysis of spectra from low molecular weight (<50 kDa), high molecular weight segments (>50kDa) and the whole sera. The cross-validated sensitivity and specificity results for classifying the spectra are summarised in Table 6.2. The separability of the groups is high all round with >80% accuracy. There is a 95% confidence interval of approximately 4% for classifications on the cohort size used. The FTIR results for whole serum demonstrates the ability to effectively distinguish between healthy, premalignant, and cancerous serum samples with high (>85%) accuracy. Additionally, the ability to classify using the spectra from low and high molecular weight subsets of the serum was demonstrated although no obvious benefit was evident. Therefore, we zoomed-in to narrower molecular windows to investigate further.

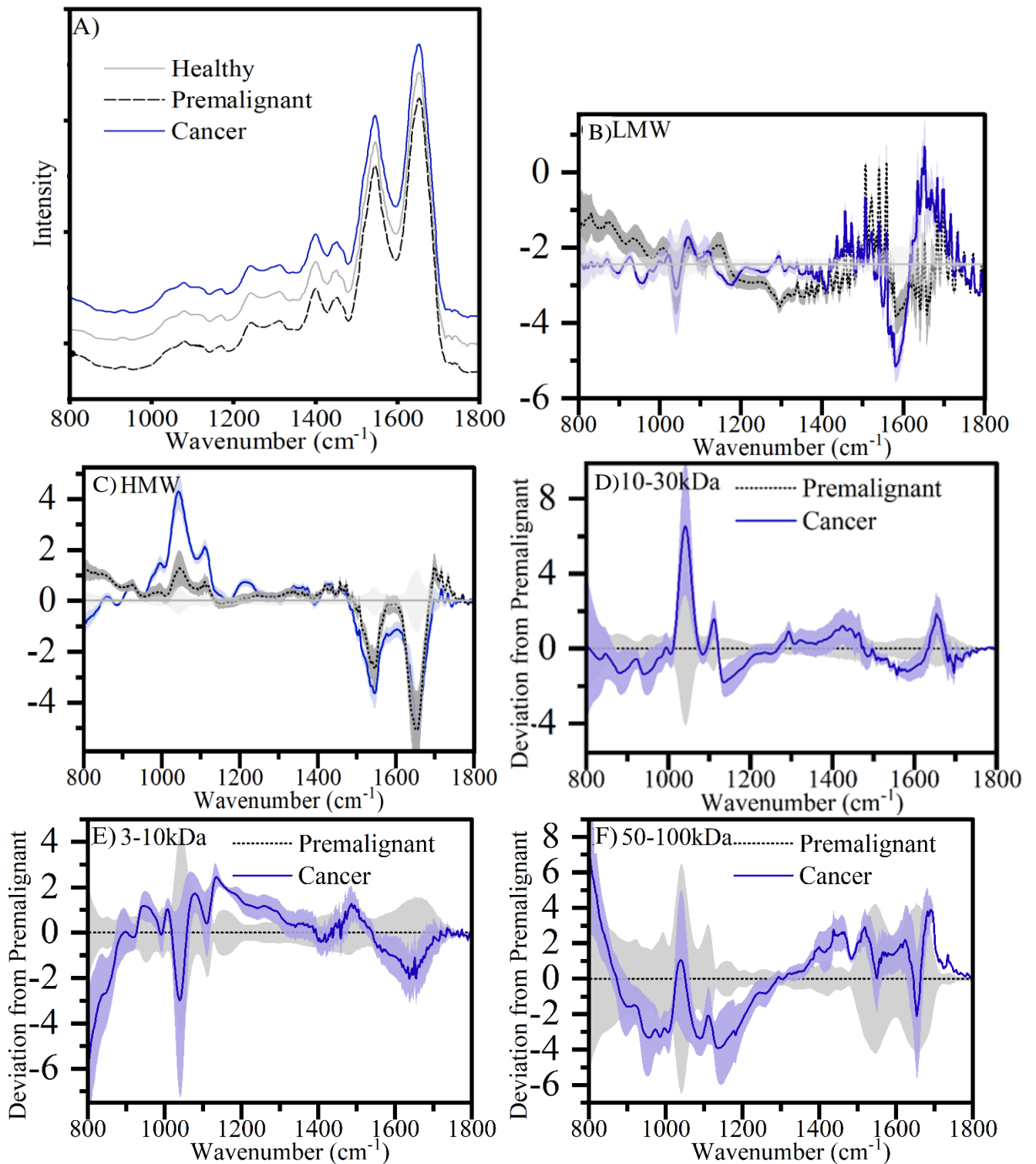


Figure 6.3: A) Average FTIR Spectra of the healthy, premalignant and cancer patients from whole serum. Each spectrum is offset for clarity in this graph only. 1σ error is minimal so is indiscernible from the graph at this scale. Difference in the average spectra of cancer patient serum from healthy : B) LMW, C) HMW, or from premalignant: D) 10-30, E) 3-10 kDa and F) 50-100 kDa windows. Error in faded colour around each line to show level of distinction for each spectrum.

It is obvious from Figure 6.3 (A) that the FTIR spectra acquired from the whole serum are hard to visually discern the cancer patients from the premalignant and healthy patients. To bring out the differences, the healthy patient's spectra are subtracted from the cancer and premalignant spectra the difference spectra are shown in Figure 6.3 (B-F). These plots highlight the clear differences between the sample groups, especially demonstrating how discernible both premalignant and cancer patients are from healthy patients.

Therefore, there are still valuable information to be gleaned in the data from this subset. For example, the observable peak shifts in Figure-1C can be referenced against known chemical signatures^{19,31,123}. An out-of-error shift can be observed in the cancer patient's signal at 900-950, 1537 and 1635 cm^{-1} . The $\sim 900 \text{ cm}^{-1}$ shifts could be attributed to alkene carbon double bond bending or C-C bond stretching. The 1557 cm^{-1} and 1635 cm^{-1} peaks can be related to the Amide II and I peak respectively.

One can also examine the commonalities between the peak patterns for cancer and premalignant samples, the regions with shared peak patterns emphasising the need for the premalignant control so that the cancer specific signals can be discerned, lest those patterns be erroneously incorporated into a spectral biomarker.

Narrower molecular windowing: We continued to search for the narrow molecular weight windows, of cancer and premalignant patient blood serum, where the accuracy is the highest. In this experiment (shown in Figure 6.4) the 10-30 kDa subset performed significantly better than whole serum, producing highly accurate cross-validated classification where all the patients were classified correctly. The sample size for this experiment is small, resulting in the higher confidence intervals depicted in Figure 6.4.

However, the results indicate a valuable 10-30 kDa window of interest for further investigation.

It is worth noting that although our hypothesis that the key signalling molecules were being obscured by the larger proteins in the serum¹² was not disproved, it did not result in a higher classification accuracy. In this regard, the results were similar to the hepatitis study by Roy et al¹⁵. However, our spectra are majorly different after the reduction of the contribution of albumin, globulin and other high weight components.

It is valuable to discern the root biological cause of the spectral shifts observed. Knowing what molecular weight fraction the key information is present in as well as the key peaks of interest can be used together to help identify potential biomarker molecules.

The identification of the 10-30 kDa region as providing the best overall classification accuracy indicates that the molecular weight splitting method can potentially have significant value, especially if this specific region can be examined in further studies.

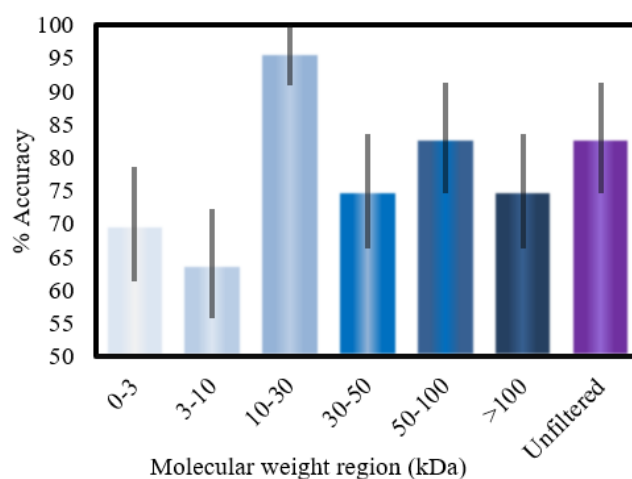


Figure 6.4: Classification accuracies between FTIR spectra of premalignant and cancer patients for different serum molecular weight subsets (molecular windows). The 95% confidence interval is shown by the grey lines over the bars. The 10-30 kDa window performed significantly better than the whole serum.

Conclusions

The potential of FTIR for screening this disease is demonstrated, with classification accuracy of 87% for whole serum. The additional use of ultra-filtration provided more information about the signal's origins, with contributing factors present in both high and low molecular weight regions. Furthermore, the molecular windowing showed even greater promise from its even higher classification accuracy for the 10-30kDa window. This can inform a follow up study into the root cause of the spectral biomarker identified. Other benefits of a narrower molecular window include suppressing external factors, such as alteration of a serum composition due to difference in diet and food culture, and internal factors such as hormonal differences between genders, stage of menstruation, age of patients, co-presence of other diseases, infections, and inflammations in a patient. Narrowing down the molecular window establishes a foundation to minimise numerous possible influences that can deteriorate the accuracy of cancer diagnosis. Further study focused to these factors will be required in future to verify the degree of the individual influences.

6.3 PCA-SVM and Leave-one-out cross-validation testing

A confusion matrix was produced for the whole serum classification in results section §6.4 (Table 6.3). This is produced from complete Leave-One-Out (LOO) cross-validation results of the classification between the cancer and premalignant patients. In this case, the ‘one’ is all three repeat measurements from one patient.

Table 6.3: confusion matrix for whole serum, Cancer v Premalignant.

Number of spectra = 240		True clinical diagnosis		
		Cancer	Premalignant	Total:
Model predicted diagnosis	Cancer	True Positive: 113	False Positive: 18	131
	Premalignant	False Negative: 13	True Negative: 96	109
Total:		126	114	Accuracy: 87.1% Confidence : 4%

To further validate this approach, 20% of the patients were left out for testing the model and the data from the confusion matrix was compared with the accuracy.

The results of a validation experiment of the complete LOO methodology to find a suitable number of principle components can be seen in Table 6.4. Here the PCA-SVM and leave-one-out cross-validation process is used to produce a model on 80%

of the patients. The accuracy is slightly lower than the full cohort due to the reduced sample size. The produced cross-validation accuracy is proven to correspond to the accuracy for using the model to classify the 20% left-out data. The 83% accuracy validation is comfortably within confidence of the 86% accurate model.

Table 6.4: confusion matrix for whole serum, Cancer v Premalignant, with 20% of data removed for validation. Figures are displayed model/validation.

Model/Validation: Number of spectra = 192/48 (80/20%)		True clinical diagnosis		
		Cancer	Premalignant	Total:
Model predicted diagnosis	Cancer	True Positive: 89/22	False Positive: 13/6	102/28
	Premalignant	False Negative: 13/2	True Negative: 77/18	90/20
Total:		102/24	90/24	Accuracy: 86/83% Confidence : 5/10%

Additionally, the efficacy of other classification methods was tested, the results of these depicted in Table 6.5. It was concluded that PCA-SVM was the most valuable of the methods used due to the higher overall accuracy and greater consistency after cross-validation.

Table 6.5: Additional comparative classifications. Using SVM and LDA alone can result in high classification accuracy but fails in the cross validation, due to overfitting.

Method:	Fraction	SVM only		LDA only		PCA-SVM		
		Classification	Original accuracy	Cross validated Accuracy	Original accuracy	Cross validated Accuracy	Original accuracy	Cross validated Accuracy
Whole	P v C		80.8	74	100	84.5	100	87.1
	C v All		86.0	78.1	100	84.3	91.6	87.7
	C v H		84.0	62.3	100	79.3	99.0	87.8
LMW	P v C		97.9	77.4	99.5	71.6	99.5	83.4
	C v All		96.2	73.9	99.1	74.3	97.3	75.6
	C v H		99.0	79.8	100	83.7	100	88.0
HMW	P v C		84.8	71.6	100	83.4	93.6	83.0
	C v All		83.6	76.4	100	87.7	95.6	85.9
	C v H		89.3	86.0	99.5	80.7	89.3	89.1
	Average		89.1	75.5	99.8	81.0	96.2	85.3

Conclusion

Overall, the PCA-SVM method was demonstrated to be the most effective and was also successfully validated during these tests. Therefore, it is recommended to make use of this method over LDA and/or SVM alone.

6.4 Saliva

Another low-invasiveness biofluid is saliva. For oral cancer, due to the close proximity, is possible that this could contain suitable biomarkers to aid with the diagnosis. Saliva is less protein dense than blood based biofluids, being comprised of over 99% fluids and therefore less than 1% proteins, electrolytes and other potential marker molecules¹²⁴. There can also be metabolites, DNA, RNA and other molecules that are key for blood detection.

It is produced by the salivary glands of which there are three pairs of major glands (the parotid, submandibular, and sublingual glands) as well as hundreds of minor ones located around the oral cavity. There will likely be small amounts of gingival crevicular fluid present in an obtained saliva sample, which is an inflammatory exudate and will contain serum, antibodies, inflammatory mediators and tissue breakdown products. Saliva is a key aid for lubrication and digestion of food but may also be a source of clues about local disease.

The relatively low concentrations of biomarkers present in saliva would likely inhibit detection, but the key advantage is the extreme lack of invasiveness involved in taking a saliva sample. Patients can even gather their own samples to be handed in for diagnosis which all could increase the cost-effectiveness and widespread suitability of detection methods based on this biofluid.

Due to the low signal strength of these samples, FTIR is a good option as its sensitivity is comparatively high while still giving key information about the potential biomarker molecules present.

Methodology

Study Design and Sample Preparation

The saliva was collected under ethical guidelines and is approved by the ethical committees in India. 413 patients were used, 189 with oral cancer, Samples were collected from the Advanced Centre for Treatment, Research and Education in Cancer (ACTREC) and the D.Y. Patil University Navi Mumbai, India. Written informed consent was obtained from all the subjects as well. This study was conducted from 2020. Samples were stored at -80°C until being thawed for analysis.

For the FTIR measurement, droplet deposition onto a CaF_2 disc, as in §4.1, was used.

Spectral acquisition

FTIR spectra were acquired with a Perkin Elmer ‘Spectrum Two’ FTIR spectrometer, as outlined in §4.3.

Pre-processing

Spectra were pre-processed with a background correction using the asymmetric least squares smoothing (ALSS) method¹²⁵ followed by average normalisation as outlined in §4.3.

Model development

Spectra were analysed over a range of principle components followed by linear support vector machine (PCA-SVM) classification, as in §4.4. Due to the high sample number and lack of repeats per patients, sensitivity and specificity values were obtained by complete leave-10-out cross-validation. The model is still compliant with the Data Optimisation Model Evaluation (DOME) standard¹¹⁸.

Results and discussion

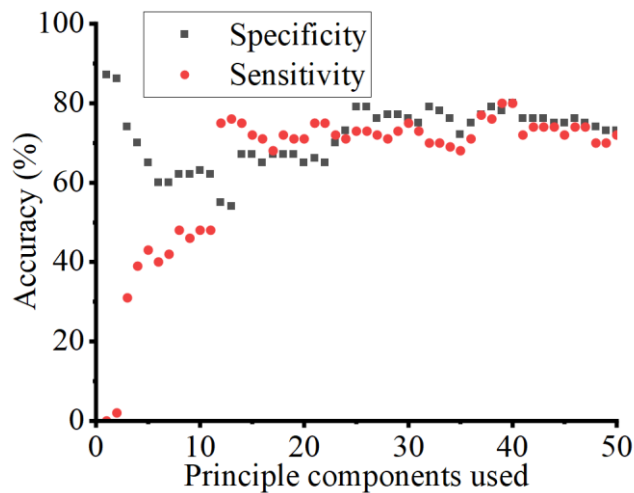


Figure 6.5: Schematic of how the accuracies changed depending on the number of PCs added, produced in the same way as figure 5.4b. The optimal number of PCs was 40, resulting in 80% for sensitivity and specificity.

The classification resulted in a cross validated sensitivity and specificity of 80% using 40 principle components (figure 6.5). From the sample number used, there is a 95% confidence interval of 3% for this classification. In Figure 6.6 the average spectral differences can be observed. One can see clear differences in these graphs, therefore it is surprising the classification accuracy only reached 80%. One could hypothesise that though the vast majority (80%) of the cancer patients are exhibiting the spectral differences, the others are significantly more in line with the healthy patients and therefore become misclassified.

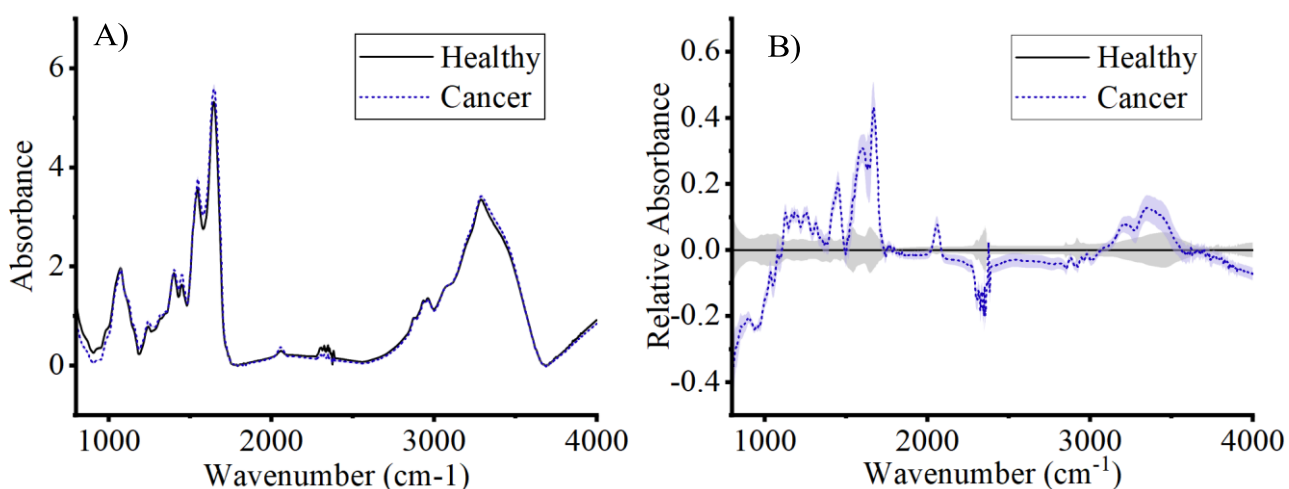


Figure 6.6: Difference in the average absorbance graph for the cancer patients compared to the healthy patients. Both the full spectra A) and the relative spectra B) are depicted for easier identification of differences. The 2-sigma error can be seen in faded colour around each graph to give an indication of the variability in each peak and section, though this is too minimal to be evident in A.

A predictive model in this accuracy range could be useful as a diagnosis aid, especially with the simplicity of preparation and lack of invasiveness of acquisition. Oral cancer is relatively easy to diagnose clinically however, due to the ease of access to the affected area. A model of this sort would reduce the need for medically trained staff to perform such a clinical examination, allowing valuable medical staff to be utilised elsewhere. Additional group controls, for example an oral infection control, would be required to ensure the model is differentiating based on cancer biomarkers rather than just inflammation signals.

6.5 Oral cancer spectral diagnosis conclusions

The methodology developed for both Raman and FTIR was proven to effectively classify oral cancer with an above 80% accuracy, corroborating with previous research and equalling or surpassing their accuracies.

The data from FTIR investigation was more promising than the Raman in accuracy, reaching >90%. FTIR also has the benefit of having greater signal to noise ratio, ease of implementation and cost effectiveness.

Using FTIR on saliva samples was additionally demonstrated to be effective, producing an accuracy of 80% on an extremely easy to access and low invasiveness biofluid.

The effect of ultrafiltration, to remove confounding high molecular weight signals, in Fourier transform infra-red (FTIR) and Raman spectroscopic diagnostic scans of human serum, was also investigated. The method yielded promising results as high as 97% accuracy in diagnosing Buccal Mucosa with FTIR. Furthermore, an additional smaller-scale molecular windowing experiment aimed at yielding greater precision on the discerning signal's origins also produced promising results of 100% differentiation accuracy in the 10-30kDa region for the serum between the two hardest to discern classes: premalignant and cancer patients.

7. Results 3: Pancreatic cancer

Introduction

The ability to detect cancers at an early stage has a dramatic effect on the cost of treating the disease, for example, early-stage colon cancer treatment costs can increase nearly fourfold when having to treat at a late stage⁷. In practice, the effectiveness of screening has been demonstrated with the ‘supermarket scan’ initiative, in which CT scans were offered to 2500 people in Manchester (UK), finding 46 cases of cancer with 80% of the cases being early stage¹²⁶. The idea of being able to quickly and effectively screen oneself for a disease is a tempting one, as current UK cancer detection processes have multiple weeks waiting times for a multitude of symptom tests that may or may not lead to a diagnosis. Radiology and endoscopy are often used but have a capacity bottleneck which means that 338,000 patients across England (UK) have to wait more than a month for radiology results⁷. Often these require the use of secondary care investigations from specialists that can be unsustainably expensive, especially for an ageing population that will require more and more screening. In fact, studies estimate that cancer diagnoses have increased by 2% per annum and that 50% of people born in England since 1960 will receive a cancer diagnosis in their lifetime⁷. All these point to a need for fast, affordable, non-invasive, and easy methods for cancer screening and raises the question: which bio-fluid spectroscopy technique can provide the solution by comparing an unknown sample with an organised database of known healthy and cancerous samples to determine if there is an affliction^{1,17,18}.

Late diagnosis of cancer inevitably leads low survival rate especially when the organ is as small as a romano pepper and the symptoms are nonspecific, allowing a tumour to

disguise and invade the pancreas in few months. Only an affordable and accurate diagnostic test can improve this abysmal 5% survival rate of pancreatic cancer.

Blood is a particularly useful bio-fluid for inspection due to its high protein and lipid concentration - as changes in these levels are some of the best indicators of disease. Much of the current research is focused on subsets of blood, plasma and serum. In whole blood, haemoglobin and other red blood cell-associated molecules can interfere with the spectra, so the plasma is preferred as the variable protein concentrations within are more sensitive to disease.

Vibrational spectroscopy as a potential method to diagnose cancerous patients has been frequently explored over the past decades, utilising many varied methodologies^{1,15,127,128}. Figure 7.2 compares the laboratory based vibrational spectroscopy methods with commonly used mammogram and magnetic resonance imaging (MRI). There is excellent potential in this field of research to produce an accurate, non-invasive detection method when used to analyse key human biofluids like blood, saliva or urine. Amongst the vibrational spectroscopy instruments, the Fourier Transform Infrared (FTIR) spectroscopy offers economically viable opportunity due to its set up cost being as low as \$15000. If measurement costs could be minimised, this would then allow for more readily available screening for these diseases, leading to earlier detection overall.

Much of the investigations found are proof-of-principle studies, which demonstrate the potential of a particular method to produce a spectral biomarker between diseased and healthy samples on a relatively small sample set. In one exemplar study, FTIR was used on human blood serum and plasma looking at colorectal cancer⁴⁴. The study identified deviations in certain peaks in plasma serum between healthy and cancerous samples. Due to presence of a large variety of molecules, it is extremely difficult to assign a

specific peak to the responsible biomarkers. With the advancement of machine learning tools, it is now possible to analyse the spectra with significantly higher accuracy. One study deployed statistical methods to analyse FTIR spectra of non-small cell lung carcinoma serum to distinguish between patients with cancerous and non-cancerous lung diseases and healthy volunteers⁴⁶ and achieved 80% accuracy⁴⁵.

The goals of the research in this section are:

- To replicate the oral cancer study in §6.2 in the case of pancreatic cancer in the UK
- To test the efficacy of a potentially improved deposition method
- To investigate if urine is a suitable biofluid for spectral diagnosis using this method
- To further explore the origins of any spectral biomarker found

7.1 Pancreatic cancer primary case study

Recently, there has been a study using attenuated total reflectance FTIR (ATR-FTIR) spectroscopy demonstrating better quantification of small molecules, such as amino acids in serum. By only using the <10 kDa fraction, it removes large obscuring signals from globulin (>80 kDa) and albumin (>60 kDa)¹⁶. Other groups have attempted improving detection of the low molecular weight molecules using ultra-filtration and ATR-FTIR^{69,70}. Yet the cancer diagnosis technology has not reached the clinics due to several limitations such as poor accuracy arising from confounding effects of molecules that can vary significantly from patient to patient as well as other physiological changes. This study successfully overcomes the current limitations by selectively probing the signalling molecules within a *molecular-weight window* using transmission mode FTIR and significantly increases the accuracy of diagnosis.

Henceforth, a *molecular-weight window* would refer to the set of molecules whose molecular-weights lie between an upper and lower cut off limits. This novel methodology identifies a *sweet-spot* or the molecular-weight window with high classification accuracy for a group of patients' diseases and discards molecules outside the window. Focusing on a narrow molecular-weight window provides the precision and enables detection of the spectral changes due to a specific disease.

Methodology

Study Design and Sample Preparation

Instead of a large cohort patient which requires significant resources, this study was designed on the principle of achieving “statistical precision” from “measurement precision” from a small cohort size to develop the underpinning method. Blood plasma and urine samples were collected with ethical approval from the same cohort of patients

from Morriston hospital, Swansea, UK. IRAS ID: 252525. A full list of patients is given in Supplementary Information (Appendix, Table A1.1). For the full cohort plasma experiment, 17 had late-stage pancreatic cancer (C), 14 had early-stage (EC), 10 were healthy (H) and 31 had premalignant pancreatic conditions (P). For the molecular-weight windowing experiment, 9 patients had advanced pancreatic cancer and were compared to 9 other patients who had premalignant pancreatic conditions. Samples were stored frozen until being thawed for analysis emulating transporting samples from hospitals and triages to elsewhere for analysis.

For the FTIR measurement, droplet deposition was used as outlined in §4.1. Centrifugal filtration of the plasma samples was performed as outlined in §4.2. Molecular windowing was performed on the smaller 18 patient cohort using 100, 50, 30, 10 and 3kDa filters. For the full cohort, only a 10 kDa filter was used, both whole and <10 kDa plasma being analysed.

Spectral acquisition

FTIR spectra were acquired with a Perkin Elmer ‘Spectrum Two’ FTIR spectrometer as outlined in §4.3.

Pre-processing

Spectra were trimmed to the 1000 datapoints in the 800-1800 cm^{-1} fingerprint region of most interest, then pre-processed with a background correction using the asymmetric least squares smoothing (ALSS) method¹²⁵ and followed by average normalisation as is outlined in §4.3.

Model development

We collected patient samples with 1000 dimensions (one intensity value per wavenumber in the 800-1800 cm^{-1} range) and classified these samples with Support Vector Machine (SVM) classifiers. We used a linear SVM classifier with and without

PCA. As we have more features than samples, which can sometimes lead to overfitting, we used PCA to reduce the number of dimensions in the original data as is described in §4.4. The results were validated with complete Leave-One-Out (LOO) cross validation in §4.5. Confidence values for each classification were produced using 95% Clopper-Pearson confidence intervals¹²⁹, as described in §4.5.

Results and discussion

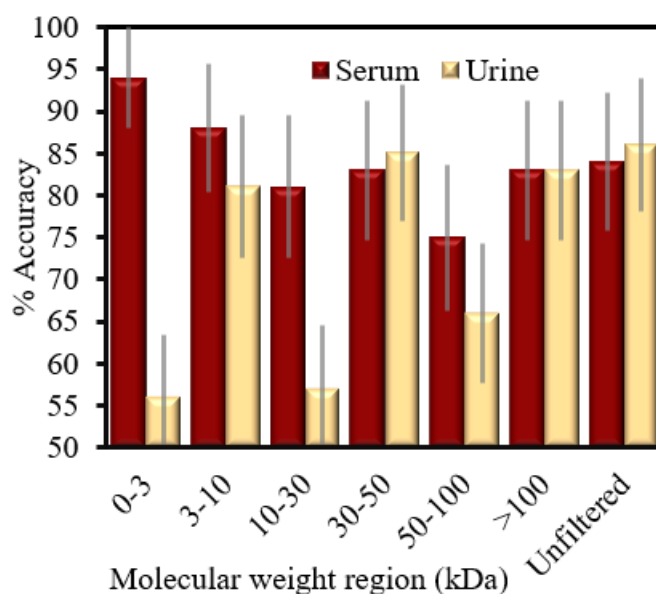


Figure 7.1: Classification accuracies between pancreatic-cancer and other-pancreatic-disease patients using FTIR spectra of their plasma and urine. The each of the 6 molecular weight molecular weight subsets is depicted as well as whole, unfiltered plasma/urine. Errors are standard binomial error calculations.

Classifying cancer and healthy patients is relatively straightforward. We believe that the main challenge lies in eliminating the standard inflammatory and other general disease markers from the premalignant conditions that deteriorate the diagnosis accuracy of cancer as both cancer as well as premalignant patients reach out to clinicians with similar symptoms. Therefore, we focused on classifying the hardest to discern i.e. cancer and

pre-malignant conditions. For screening and potential quantification purposes in the full cohort study, a healthy control set was also used.

In the blood plasma “molecular-weight windowing” study, as shown in Figure-7.1, the lower molecular weight regions produced the highest cross-validated classification accuracies. The most accurate was the <3 kDa region, with 94% accuracy, followed by

Table 7.1: Comparison of serum diagnostic accuracies after PCA-SVM and LOO-cross-validation on the subsets in the study. C: Cancer, P: Premalignant, H: Healthy, EC: Early-stage Cancer. Cross-validated SVM accuracy included for comparison. Appendix Table A1.2 contains additional information.

Fraction	Subsets compared	Sens. (%)	Spec. (%)	Acc. before CV (%)	PCA-SVM Acc. (%)	95% Confidence interval(%)	SVM only Acc. (%)
Whole serum	C v P	74	90	91.6	84.3	77.3-90.4	75.3
	C v H	100	100	100	100.0	95.6-100.0	92.4
	C+EC v H+P	75	86	100	81.3	75.4-86.2	75.7
	C v EC	88	78	100	83.5	74.4-90.4	76.9
< 10 kDa window	C v P	90	90	100	90.0	84.5-95.1	86.5
	C v H	97	93	100	95.3	87.8-99.0	87.2
	C+EC v H+P	90	91	100	90.6	85.7-94.3	89.2
	C v EC	90	90	100	90.0	80.6-95.8	58.6

the 3-10 kDa at 88%. Both were higher than the classification for whole pancreatic cancer plasma, which scored 84%. The urine data (Figure 7.1) performed relatively

worse overall. Unfiltered urine, >100 kDa filtrate and 30-50 kDa filtrate scored around of 86%. The cohort size for this preliminary experiment was quite small, resulting in the larger 95% confidence of 5% for the most effective groupings, up to 10% for those with less accuracy. These can be seen on Figure 7.1.

As the two highest scoring regions were the <3kDa and the 30-10kDa, for practical purpose and ease of implementation, we designed our study to probe the <10 kDa plasma filtrate, with a comparison to whole plasma. From Table 7.1, we can see that this <10kDa filtrate performed better than whole plasma for most comparisons made, each has 90% or higher accuracy when diagnosing late-stage patients against healthy and premalignant pancreatic patients (See Table 7.2 for an example confusion matrix of this classification). Furthermore, when early-stage cancer patients were included, 90% accuracy was still achieved. The ability to distinguish between early and late stage cancer patients with a 90% accuracy is also demonstrated.

Table 7.2: Confusion matrix for <10kDa plasma, Cancer v Premalignant.

Model: Number of spectra = 152		True clinical diagnosis		Total:
		Cancer	Premalignant	
Model predicted diagnosis	Cancer	True Positive: 46	False Positive: 10	56
	Premalignant	False Negative: 5	True Negative: 81	86
Total:		51	91	Accuracy: 90% Confidence : 5%

The one case where whole plasma performed better than <10kDa was with the healthy subset. The values for classifying Cancer v Healthy were both high, and within confidence of one another. Furthermore, this is the least important classification to achieve, as it is likely to be affected by non-cancer factors.

The results can be compared to currently used ELISA methods using the known pancreatic cancer biomarker Carbohydrate Antigen 19-9 (CA19-9, molecular weight 820 Da), which produced 70-80% accuracy on the same patient samples (Figure 7.2). This biomarker is in the <10 kDa region used. However, the patients misclassified by each method were different. Attempts were made to establish if ELISA and the FTIR method are probing the same biomarkers by measuring the <10 kDa filtrate using both

methods. In our study, ELISA consistently reported lower level of CA19-9 in the <10 kDa filtrate than the whole serum. This clearly indicates that the spectral biomarker is unrelated to CA19-9, and CA 19-9 is either not contributing significantly to the IR signal, or it is attached as glycoprotein and is filtered out with high molecular weight molecules.

Conclusion

From these results, one can conclude that using the <10kDa molecular weight region can provide a practical and superior classifier model to using unfiltered plasma alone. Furthermore, urine could be used as a diagnostic biofluid, but the accuracy would not be as high as when plasma is used. The additional ability to diagnose early-stage patients only increases the potential for the method to be used for screening cancers before they can progress to late-stage cancer. Furthermore, being able to discern early from late stage cancers with a high accuracy is also invaluable information for a patient's treatment.

We also deployed the molecular weight windowing method for diagnosis of oral cancer (§6.1) and observed significant improvement in diagnosis accuracy. One can see a summary of the pancreatic and these accuracies, along with the ones from practice and other promising research, in figure 7.2. From the combined results of the pancreatic and oral studies, it is plausible that narrowing down the molecular weight window can allow us to probe the sweet spots to diagnose different types of cancer with high accuracy at an early stage.

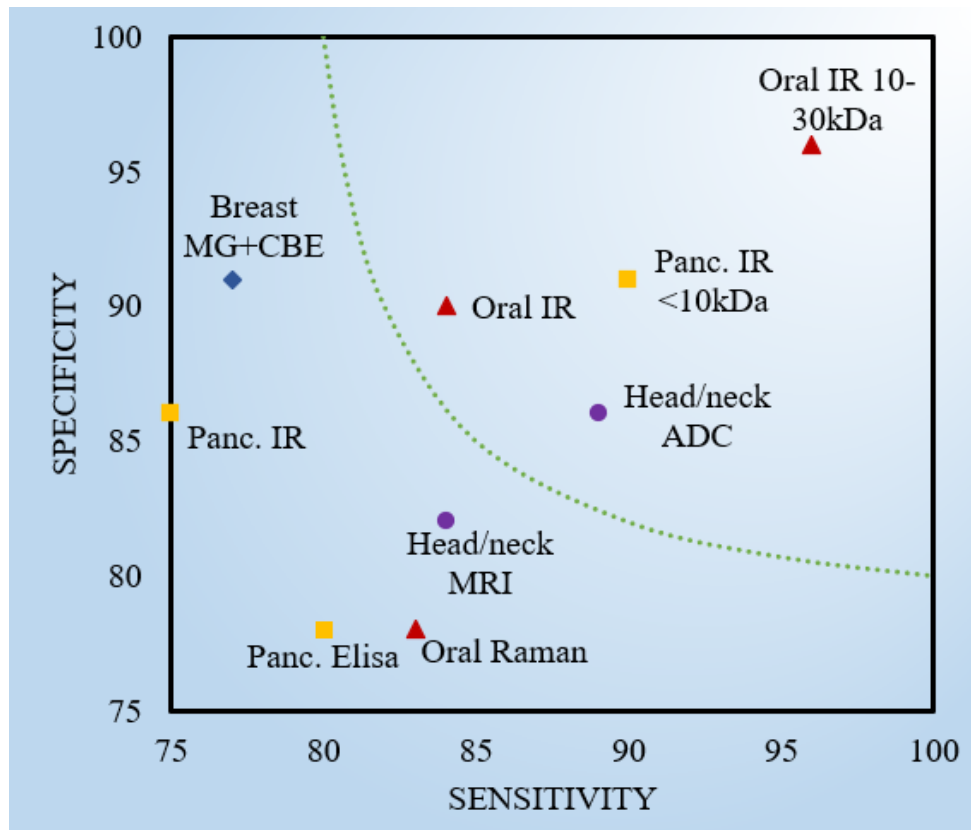


Figure 7.2: Summary of cancer diagnostic techniques. Pancreatic cancer work (shown as ■) and oral cancer work (shown as ▲) on plasma is from our research group on 74 pancreatic and 120 oral cancer patients. Raman data is also from our own research on 90 patients' serum, though the 10-30kDa variation is on an 18-patient subset. Head/Neck cancer data from a recent study by Van Der Hoorn et al.² using 854 patients for the MRI and 287 for the apparent diffusion coefficient (ADC) imaging metric (shown as ●). Mammogram (MG) combined with clinical examination (CBE) were from a breast cancer (shown as ◆) study on 32,080 patients by Noriaki et al.³ The dotted curve represents a qualitative boundary of acceptability followed by us with a minimum 85% accuracy.

7.2 Pancreatic cancer diagnosis using spray deposition

Introduction

In order to counteract the coffee ring effect we developed an air-spray system to deposit a biofluid sample onto a disc with relative uniformity. From our testing, the standard error between points taken along the radius of a disc was 3.25 for droplet deposition and 1.40 for spray deposition. In this experiment, the pancreatic cancer investigation was replicated with as many duplicate samples as was possible to see if using the spray system had any, positive or negative, effect on the FTIR-based classification accuracies obtained.

Methodology

Study Design and Sample Preparation

Blood plasma samples were collected with ethical approval from the same cohort of patients from Morriston hospital, Swansea, UK. IRAS ID: 252525 (see Table A1.3). For this experiment, 10 had late-stage pancreatic cancer (C), 14 had early-stage (EC), 10 were healthy (H) and 27 had premalignant pancreatic conditions (P). Samples were stored frozen until being thawed for analysis, emulating transporting samples from hospitals and triages to elsewhere for analysis.

Most of the experiment was consistent with §7.1, except the deposition method. The deposition method was spray deposition as outlined at the end of §4.1. Each fraction was diluted to a volume of 50µl, apart from 'whole', which had 30µl added to 20µl of sample and the <3kDa/<10kDa fractions which didn't require dilution. The samples were placed into the air-spray system before being deposited uniformly on a 25 mm diameter Crystran CaF₂ slide and left to dry for 20 minutes for analysis.

The spectral processing and data analysis was consistent with §7.1, though only the best classification method identified there, PCA-SVM, was used.

Results and discussion

Table 7.3: Comparison of serum diagnostic accuracies from subsets in the study.

C: Cancer, P: Premalignant, H: Healthy, EC: Early-stage Cancer

Fraction	Subsets compared	Sens.	Spec.	% Acc.	Confidence
Whole plasma	C v P	78	85	82.4	76.2-87.6
	C+EC v H+P	72	82	77.6	72.0-82.6
	C v EC	79	76	77.2	66.3-85.9
< 10 kDa window	C v P	80	93	89.5	82.2-94.5
	C+EC v H+P	87	90	88.8	83.4-92.9
	C v EC	90	91	90.6	81.8-96.1

From Table 7.3, it is clear the <10kDa fraction with the spray system produced accuracies within confidence of the droplet deposited samples. However, the whole serum performed significantly worse for the Cancer v Premalignant and The Cancer v Non-cancer comparisons. These results suggest the spray system can be used to increase the drying speed and deposition uniformity of the plasma samples without compromising accuracy, especially with the <10kDa fraction of interest. The whole serum will need to be examined to see if an increase/decrease in concentration of deposited sample will help bring the accuracy closer toward the previous findings.

7.3 ATR-FTIR test on Pancreatic Patient Urine

In our earlier testing (manuscript 2), urine was shown to be less useful for pancreatic cancer diagnosis than blood. The molecular windowing did not increase its accuracy either. However, it is another biofluid with very low invasiveness and the potential for patients to submit their own samples without the need for trained medical staff. Taking this into account, the 80% accuracy could still be sufficient for a diagnostic aid. Furthermore, ATR FTIR offers the ease-of-use option of being able to analyse samples while they are still aqueous. Therefore, it would be useful to establish if this quick method with a high ease-of use could replicate, or even improve, the accuracies from manuscript 2.

Methodology

Study Design and Sample Preparation

The saliva was collected and urine samples were collected with ethical approval from the same cohort of patients used in §7.1 from Morriston hospital, Swansea, UK. IRAS ID: 252525.

For the FTIR measurement, 10µl of each sample was deposited on the ATR crystal and was then immediately scanned.

Spectral acquisition

FTIR spectra were acquired with a Perkin Elmer ‘Spectrum Two’ FTIR spectrometer as outlined in §4.3.

Pre-processing

Spectra were trimmed to the 1000 datapoints in the 800-1800 cm⁻¹ fingerprint region of most interest, then pre-processed with a background correction using the asymmetric

least squares smoothing (ALSS) method¹²⁵ and followed by average normalisation as is outlined in §4.3.

Model development

Spectra were analysed over a range of principle components followed by linear support vector machine (PCA-SVM) classification as is described in §4.4, though sensitivity and specificity values were obtained by complete leave-5-out cross-validation. Confidence values for each classification were produced using 95% Clopper-Pearson confidence intervals¹²⁹, as described in §4.5.

Results and discussion

Table 7.4: PCA-SVM results in wavenumber range 800-1800cm⁻¹

Comparison	Principle components	Sensitivity	Specificity	Accuracy	confidence
Cancer + Early cancer v Healthy + Benign	6	80	90	85.4	78.4-90.9
Cancer v Healthy	2	80	92	86.8	76.5-93.7
Early cancer v Late stage cancer	9	70	60	64.8	51.7-76.4

Table 7.5: PCA-SVM results in wavenumber range 800-4000cm⁻¹

Comparison	Principle components	Sensitivity	Specificity	Accuracy	confidence
Cancer + Early cancer v Healthy + Benign	6	93	90	91.4	85.4-95.5
Cancer v Healthy	2	90	82	85.5	74.9-92.8
Early cancer v Late stage cancer	2	76	87	81.8	70.0-90.3

Table 7.6: PCA-LDA results in wavenumber range 800-4000cm⁻¹

Comparison	Principle components	Sensitivity	Specificity	Accuracy	confidence
Cancer + Early cancer v Healthy + Benign	10	95	92	93.4	87.8-96.9
Cancer v Healthy	2	90	92	91.1	81.8-96.6
Early cancer v Late stage cancer	8	90	93	91.6	81.8-97.1

For the typical 800-1800 cm^{-1} fingerprint region, using PCA-SVM, the results were similar to the transmission FTIR trial from paper 2, the classification of cancer v non-cancer once again having an 85% accuracy (Table 7.4). The Cancer v healthy performed slightly better, as can be expected. Unfortunately, the early v late stage cancer classification was far less distinct. However, when increasing the wavenumber range to 800-4000, there is a major increase in the accuracy (Table 7.5). This is even more pronounced when using PCA-LDA, which can often handle smaller sample numbers better (Table 7.6).

This would be promising for the potential of the higher wavenumbers being required for improved classification. However, from examining the spectra, it is possible that this difference is due to a systematic acquisition error. The broad water peak at 300-3500 seems inconsistently flattened and there are 2 inverse peaks at 2850 and 2920 that are far more prominent in healthy spectra acquisitions. Both of these suggest errors with background correction. The water difference is likely due to imprecise droplet positioning on the ATR crystal, or potentially volume variation from the small amounts pipetted, meaning a variable amount of liquid being directly above the crystal. I suspect the inverse peaks are due to contaminants present in the background correction spectra on a particular day of acquisitions. Therefore, it is deemed safer to use the relatively unaffected 800-1800 region to develop any conclusions. The PCA-LDA also corroborated with the PCA-SVM values for this range (Appendix, Table A1.4).

The PCA-SVM accuracies do suggest some viability of this method of practical diagnosis. 85% is sufficient for how relatively quick and simple this method is to perform. If the background consistency issues could be mitigated, and the method went through another trial with a larger cohort, then it could be converted into a viable, quick and easy screening method.

7.4 Cell media and heavy glucose spiking investigation

Cancer cell lines are often used in biomedical research as model systems for the study of the cancer they are associated with. They can provide key insight into understanding the mechanisms of cancer development without having to rely on a human patient. They also offer the advantage of being easy to access, manipulate and maintain in vitro and there is no risk to a patient when carrying out more potentially cell-destructive tests. Typically, the cells are obtained from tumours and subsequently cultured. Though other animal cell lines are studied, of most relevance are the human ones¹³⁰.

Just as cancer cells release signalling molecules, debris and metabolites into the bloodstream, cell culture will release potential biomarkers into the cell media in which they are cultured. Therefore, it would be revealing to see if cancer cell lines could be differentiated from healthy cell lines by their spent media alone.

FTIR is a suitable choice to investigate the cell media, due to its high sensitivity and ability to detect a large variety of metabolite signatures. An early test was performed with this, with some mixed results (Appendix, Table A1.5).

C-13 glucose is a 'heavy' variety of glucose, containing a carbon atom with an additional neutron. This can be distinguished in an FTIR spectrum¹³¹. It was sought to see if spiking some cell samples with this 'heavy' glucose can affect their spectra and thereby affect the accuracy of the cancer diagnosis.

CA 19-9, also known as 3-Sialyl Lewis A, is a carbohydrate antigen that has a positive correlation with most tumour cell lines of gastrointestinal carcinoma; including adenocarcinoma(s) of the stomach, intestine, and pancreas. Healthy patients typically have up to 37 U/ml of the biomarker in their blood, while patients with cancer can have levels greater than 2500 U/ml¹³². It would be interesting to see if spiking a healthy

sample with approximately cancer levels of this antigen would result in a cancer diagnosis via our spectral diagnosis method.

Methodology

Sample preparation

The HPDE/H6C7 (Human Pancreatic Duct Epithelial Cell Line)¹³³ healthy cell line was compared to HS766T, PANC1, and MIA PACA2 pancreatic cancer cell lines¹³⁴. 3 collections of media were made per cell line with 3 repeat spectra taken per media. This resulted in a comparison of 9 healthy and 27 cancer spectra. The cells were cultured in the media then removed by centrifugation. The spiked cells were cultured in a media containing 20mM concentration of c13 D-Glucose, the control set containing standard glucose. The spent media was then collected and used for the FTIR experimentation.

For the CA19-9 spiking experiment, the 80µl healthy media was spiked with 0.2µg of 3-Sial Lewis A to see if this would influence the classification. Measurements for the CA19-9 spiking section were performed by a student collaborator Rosi Toncheva, under my supervision.

For the FTIR measurement, droplet deposition was used as outlined in §4.1. Centrifugal filtration of the plasma samples was performed as outlined in §4.2. Only a 3 kDa filter was used, both whole and <3 kDa media being analysed.

Spectral acquisition

FTIR spectra were acquired with a Perkin Elmer ‘Spectrum Two’ FTIR spectrometer as outlined in §4.3.

Pre-processing

Spectra were trimmed to the 1000 datapoints in the 800-1800 cm⁻¹ fingerprint region of most interest, then pre-processed with a background correction using the asymmetric

least squares smoothing (ALSS) method¹²⁵ and followed by average normalisation as is outlined in §4.3.

Model development

A linear PCA-SVM classifier was used as outlined in §4.4, a PCA-LDA classifier was also used. The results were validated with complete Leave-One-Out (LOO) cross validation, described in §4.5.

Results and discussion

D13

The sample size was insufficient for SVM to classify well, so PCA-LDA was used.

Table 7.7: FTIR results for cell media, cancer v healthy classifications, using PCA-LDA in wavenumber range 800-4000cm⁻¹.

Sample	Principle components	Sensitivity	Specificity	Accuracy	confidence
Whole	25	96	88	94	80.7-99.2
Whole C13 glucose	24	77	66	74.3	57.0-87.3
<3kDa	12	88	83	86.8	70.8-96.2
<3kDa C13 glucose	12	96	100	97	83.9-99.9
Both Whole	9	98	77	92.8	84.1-97.5
Both <3kDa	24	91	75	88.1	77.8-94.8

From the results in Table 7.7, one can generally see that it is possible to classify healthy from cancer cell media with a high accuracy. One can see that the heavy (C13) glucose

spiked whole media performed poorly when compared to standard glucose. However, the spiked <3kDa subset performed better, having the highest overall accuracy.

This would suggest that the C13 signal is lost in the whole media, but filtering it to just the lowest weight components, of which small sugars are one, allows the divergent signal between the spiked healthy and unhealthy to be discerned. This will be explored further in the NMR section (§7.5), where we scan these same sample set with NMR as well.

CA19-9

For this subset of cell media (Figure 7.3a), which were analysed by 3-way PCA-LDA due to the small sample number, the spectra from the media of the healthy cell lines were distinct from the cancer lines and blank media, classifying with a high sensitivity/specificity of 80/100%. Like with the D-Glucose data, this is similarly promising for the hypothesis that the cells release signals into the media similarly to how they would release signals into the human body, resulting in a similarly distinct classification for the different cell types.

This did not change with the spiking, however the spiked healthy cells can be seen to have shifted toward the cancer classification region in the PCA plot (Figure 7.3b, c). This would suggest that CA19-9 is a component in the classification, but not the sole thing that distinguishes the signals from the cancer and healthy cells. Additionally, the spiking of the blank media appeared to have no distinct effect on their classification.

Further tests and a larger sample cohort was planned for this experiment, but it was unfortunately curtailed by the sudden arrival of the Covid pandemic. In the context of the comparison with the CA19-9 ELISA test in manuscript 2, it seems that the carbohydrate antigen is not a major component of the distinguishing classification

between cancer and healthy patients using FITR spectroscopy. Therefore, it was unnecessary to pursue this line of investigation further.

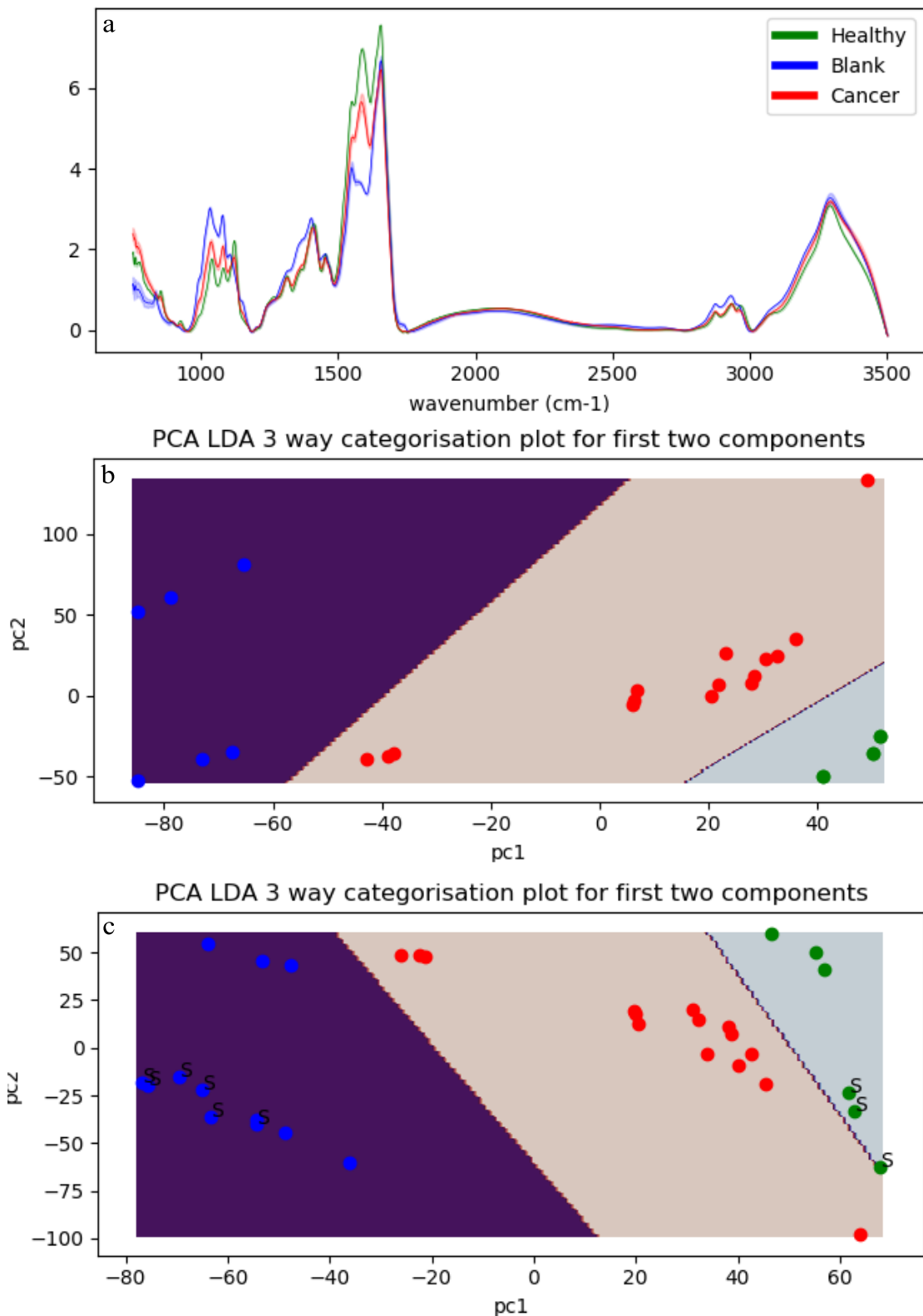


Figure 7.3: a) spectra of the healthy cell media, blank-media and cancer cell media
 b) PC plot of the first to components without the CA-19.9 spiking c) PC plot of the first to components with the CA-19.9 spiking, denoted by the S character.

7.5 Nuclear Magnetic Resonance testing

Nuclear Magnetic Resonance (NMR) spectroscopy is another method that can be used to obtain chemical and structural information about molecules within a sample. It operates by detecting the chemical shift of the resonance frequencies of the nuclear spins within the sample¹²². Prior research had indicated the potential for this method in biofluid diagnostic spectroscopy.

Preliminary NMR scans were performed on some pancreatic plasma samples. Unfortunately, no significant signal was raised in the C-13 NMR. The proton NMR spectra were differentiable with an 83% accuracy by our cancer distinction method (Appendix Figure A1.3). This suggested that further investigation would be valuable. Furthermore, it was decided to further investigate the potential of spiking a sample with heavy (C13) glucose. C13 spiked samples should be NMR active, so using the spiked cell media samples from the cell media experimentation, it was decided that it could be valuable to also use NMR to test them.

Methodology

Sample preparation

The same cell line samples were used as in the cell media experiment in §7.5. The HPDE/H6C7 (Human Pancreatic Duct Epithelial Cell Line)¹³³ healthy cell line was compared to HS766T, PANC1, and MIA PACA2 pancreatic cancer cell lines¹³⁴. 3 collections of media were made per cell line with 3 repeat spectra taken per media. This resulted in a comparison of 9 healthy and 27 cancer spectra. The cells were cultured in the media then removed by centrifugation. The spiked cells were cultured in a media containing 20mM concentration of c13 D-Glucose, the control set containing standard

glucose. The spent media was then collected and put into a 1:1 ratio mix with deuterium oxide and used for the NMR experimentation.

Spectral acquisition

The proton NMR spectra were acquired with a Bruker NMR 500 MHz for 64 scans with an operating frequency of 500MHz.

Pre-processing

Spectra were pre-processed with a background correction using the asymmetric least squares smoothing (ALSS) method¹²⁵ and followed by average normalisation as is outlined in §4.3.

Model development

A linear PCA-SVM classifier was used as outlined in §4.4, a PCA-LDA classifier was also used. The results were validated with complete Leave-One-Out (LOO) cross validation, described in §4.5.

Results and discussion

Table 7.8: NMR whole cell media, cancer v healthy, PCA-LDA 800-4000

Sample	Sensitivity	Specificity	Accuracy	confidence
Normal glucose	100	46	78.4	50.2-94.9
C13 glucose	100	84	93.6	68.4-99.9
Both	75	76	75.4	56.3-89.2

The results (Table 7.8) indicate that the C13 glucose spiked samples produce the highest accuracy. This could indicate that the spiking does alter the signal in a disproportionate way, making the cancer cell media more distinguishable, as was in the <3kDa cell media FTIR experiment. If this is the case, the obscuring effect of the higher weight components from the FTIR experiment does not seem to affect the NMR test.

Combining the spiked and un-spiked samples does not improve the classification. This could be due to the novel input from the spiking being obscured by the non-spiked samples.

The conclusions from this data are tenuous, as the confidence in the results are low. This is due to insufficient repeats. Unfortunately, the student acquiring the results did not repeat the measurement in triplicate for each sample and there was insufficient time to correct this. If this was performed, the confidence interval for the C13 glucose would have been reduced to 82.1-98.7%, assuming the accuracy was retained. These repeats coupled with adding in additional samples to this and the FTIR experiment would be required for significant conclusions to be drawn from this data.

Additionally it would be useful to compare a NMR scan of a <3kDa (or <10kDa for both NMR and FTIR) plasma sets. Unfortunately, this was not feasible on this occasion due to the relatively high amounts of sample required for NMR and there being insufficient of the filtered sample remaining.

7.6 Pancreatic cancer spectral diagnosis conclusions

Diagnosis accuracy as high as 90% has been achieved using a novel, affordable, non-invasive diagnostic method by combining measurement precision of infra-red spectroscopy with classification using machine learning tools. The study in §7.1 investigated urine and blood from pancreas cancer patients and healthy volunteers, and significantly improved accuracy by focusing on sweet-spots within blood plasma fractions containing molecules within a narrow range (<10kDa) of molecular-weights. Furthermore, the plasma results were corroborated with a replication using a different deposition method, with the advantage of more even sample distribution and a faster drying time, producing similarly high accuracy from the identified <10kDa region of interest.

The urine results, though less accurate than plasma, were still of note as they come from a high ease of access and low invasiveness biofluid.

The additional investigation into pancreatic cancer using cell media has been less conclusive due to their limited sample sizes and resultantly low confidence, but promising overall. These should this be expanded further in future research, using FTIR in conjunction with NMR.

8. Future work

For future progression from the work in this thesis, there are several routes that are worthwhile being explored. One of those is the potential for implementation of these methods into a usable technology. The prospects of this have been investigated in Appendix §2.3, along with preliminary investigation into correcting between different FTIR instruments.

Another consideration is the root biological cause of these spectral biomarkers. The cell media investigation is promising, though needs more iterations to become conclusive. Additionally, it is worthwhile probing the <10kDa section of blood plasma to identify the constituent molecules present.

A final avenue identified is using electrochemistry to try and elucidate extra biochemical information. This is preliminarily explored in the following section.

8.1 Spectro-electrochemistry investigation

Introduction

One of the key factors that is often neglected in the field of identifying spectral biomarkers is distinguishing the origins of the spectral shifts. There are many potential methods for identifying what causes a particular spectral peak, for example one can try labelling specific proteins with heavy atoms and see if the characteristic signals vary between diseased and healthy samples. However, this limits the scope of an investigation to cell or small organism level controlled studies where the metabolism of the cells or organisms can be controlled, e.g. feeding a cell culture heavy glucose to identify increased metabolism in diseased cells³⁸. One simpler method of probing the signals in a bio-fluid solution is to manipulate them by varying conditions, like temperature, charge or pH, and measuring any spectral shifts. With knowledge about what molecules are likely to be affected by your variable, the molecule of interest can potentially be discerned.

Cyclic Voltammetry (CV) is a valuable electrochemistry method for evaluating the redox properties of molecules³⁹. It is typically undertaken by varying the potential difference between a working and counter electrode, in comparison to a reference electrode, and measuring the current. The amount of current produced and therefore the shape of the curve can be affected by the rate at which the voltage is shifted. The curves produced can elucidate important electrochemical information about the species, for example its stoichiometry or the reversibility of the reaction.

It has been demonstrated that Raman spectroscopy can discern redox changes in biological systems in an in-situ study by Brazhe et al.⁴⁰. Rat hearts were studied at

normal and hypoxic conditions and several peaks were found to be unique to the oxygenated heart. Some of these peaks were attributed to c and b type cytochromes by applying an uncoupler (carbonyl cyanide 4(trifluoromethoxy)phenylhydrazone) to affect their redox state and causing a resultant reduction in their peak intensity. Additional variable peaks were assigned to other cytochromes and oxymyoglobin using other methods, but this ability to detect redox changes in cytochrome c is not only promising for the potential ability to discern more information from redox changes in proteins in serum samples, but will also serve as a suitable method to test a Spectro-electrochemistry setup in future experiments.

The spectro-electrochemistry setup implemented currently consists of an Autolab PGSTAT204 potentiostat coupled to a Renishaw In-via Raman spectrometer. Using the Autolab Nova software as a basis, the Raman has been set up to be triggered remotely. This was achieved by sending a .net signal from the Nova software to Renishaw's example triggering app which signals the Renishaw Wire software via Windows event messages – and triggers spectral acquisition almost instantaneously. This allows for simultaneous acquisition of spectra and electrochemical information as well as allowing for a setup where spectra can be automatically taken at key electrochemical events. This is particularly valuable as higher scan rates required to see greater fluctuations in electrochemical signals may only allow for a few key spectra to be taken in the same time frame.

Methodology

The methodology is tested with cytochrome-c⁴¹.

Prepare a solution of 75 μ M cytochrome-c in standard phosphate buffered saline solution. This requires 0.9g cytochrome-c, 8g NaCl, 0.2g KCl, 1.44g Na₂HPO₄ · 2H₂O and 0.24g KH₂PO₄ per litre of water⁴².

Set up glass-bottomed-plastic cuvette with solution, platinum gauss working electrode, platinum counter electrode and Ag/AgCl reference electrode.

Set up experiment to record Raman spectra and current after every 50mV jump over a range of -600 to 600mV cyclically for 3 cycles.

Raman spectra to be taken with 20mW 532nm laser at 5x magnification, for 1 second with 3 accumulations in the spectral range of 700-1900 cm^{-1}

Results and discussion

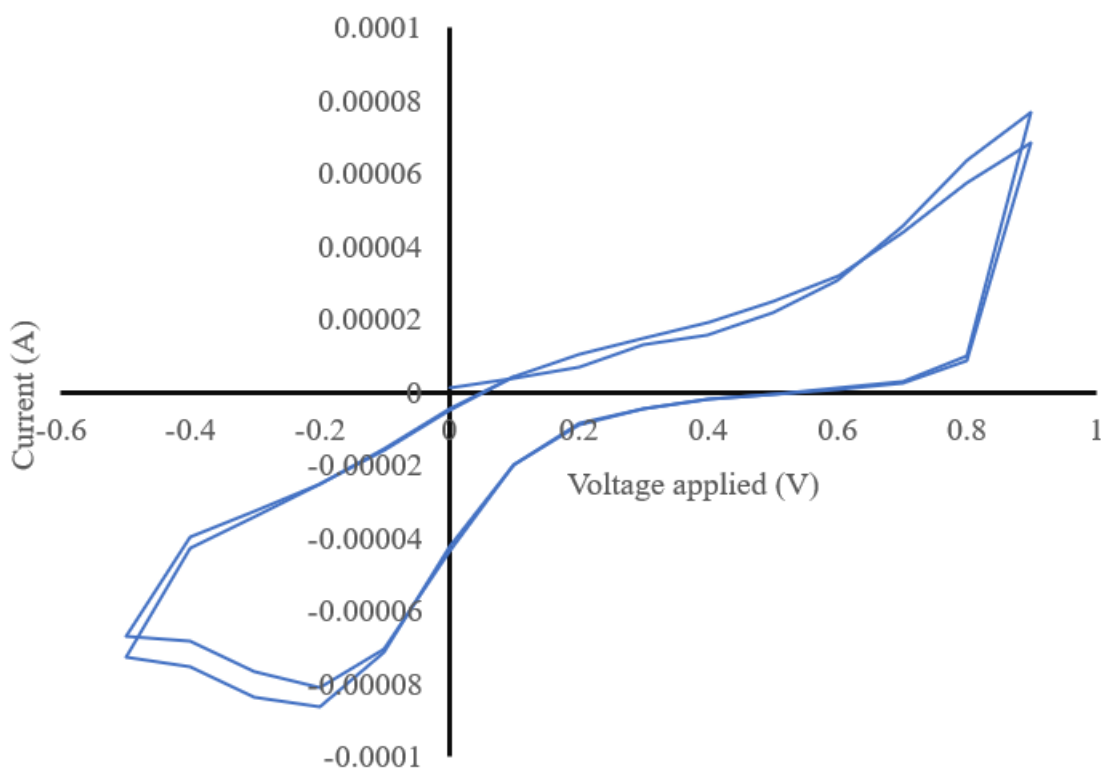


Figure 8.1: Current/Voltage CV curve for Cytochrome-C

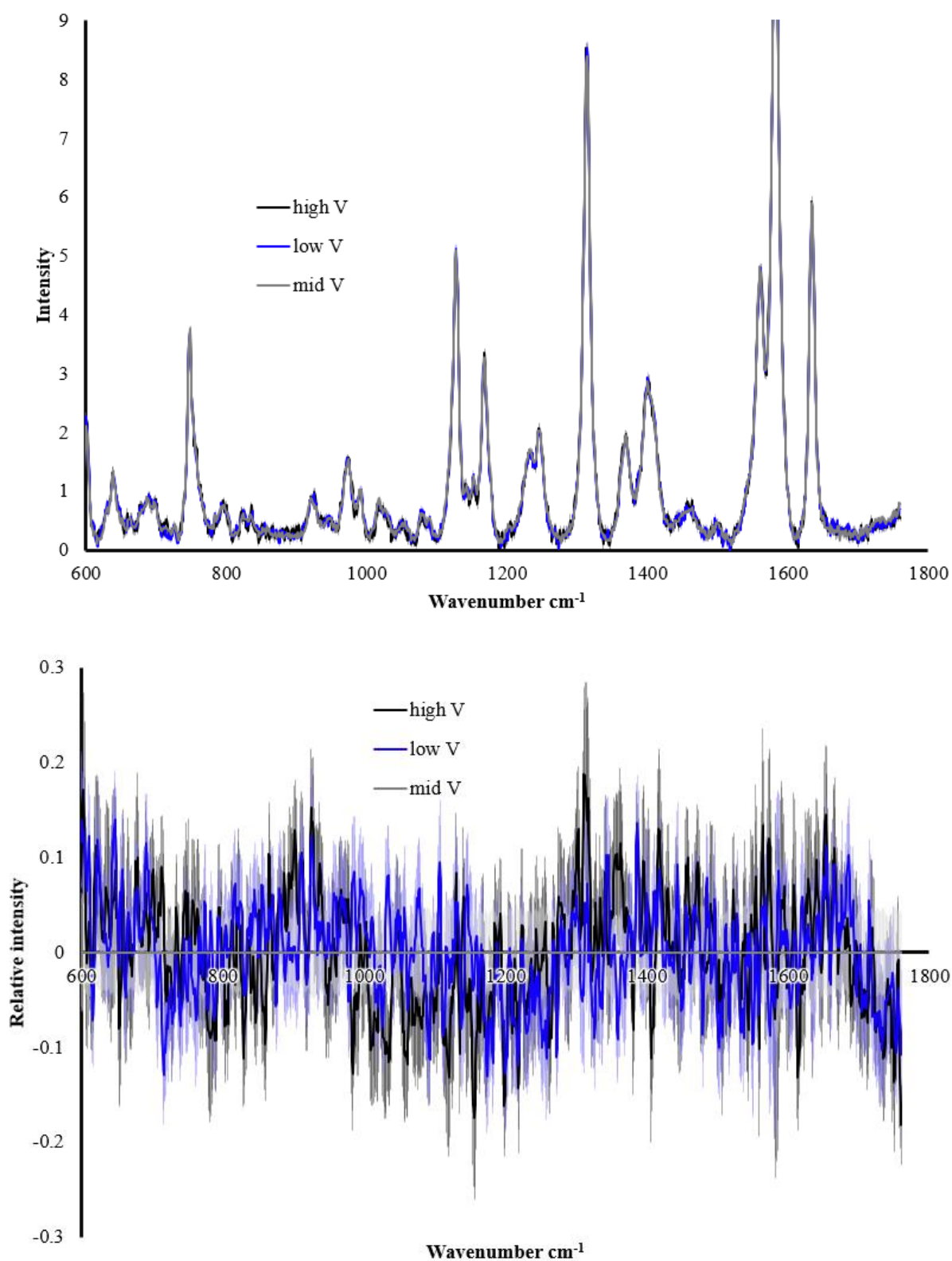


Figure 8.2: The cytochrome c spectra at low (<40 μ A) medium and high (>30 μ A) currents/voltages in the CV curve. Error in faded colour around the graph lines. a) normal spectra b) spectra relative to medium voltage.

From Figures 8.1 and 8.2 one can see that, though a limited CV curve is observed, there doesn't appear to be a significant shift in the spectra at different current/voltage levels. Though there may be some shifts, the precision achieved in the Raman measurement was not sufficient to discern it. Ultimately it was decided to delegate another doctoral student from our group to investigate this further using a more advanced custom setup.

9. Conclusions

Before starting any investigation, a proposed methodology was developed from extensive review of the literature (Appendix Table A2.1). Developing upon those standard recommendations throughout experimentation and analysis, a more refined version of this has been developed. FTIR was most used in this thesis, so Table 9.1 outlines the standard FTIR recommendations for future biomarker investigation.

Table 9.1: Standard FTIR Biomarker identification recommendations

Patient selection
Have a healthy control Have a related infection(s) control Define patients also based on biological gender and age groups. Each group should have >5 members If prognosis findings required compare between disease stages
Sample preparation
Use blood serum/plasma Use CaF ₂ substrate Droplet: dilute in ratio 1:25 serum to ultrapure water, deposit 500µl disc, Leave to dry for 24h in contained chamber Spray: Air-spray 50µl of diluted sample onto disc, dilution with ultrapure water if sample is too concentrated (e.g. whole plasma 2:3 ratio of plasma:water), leave to dry for 30minutes.
Spectral acquisition
Use transmission mode FTIR

Resolution of 4cm^{-1}

Take spectra from $75\text{-}4000\text{cm}^{-1}$ to capture both the fingerprint ($<1800\text{cm}^{-1}$) and high wavenumber regions ($2700\text{-}3500\text{cm}^{-1}$)

Spectral pre-processing

Trim to key $800\text{-}1800\text{ cm}^{-1}$ region

Use ALSS baseline correction

Use average normalisation

Statistical analysis

Compare 2 groups at a time

Compare all groups (including comparisons between age and gender groups)

Use PCA to maximise variance

Use SVM on the PCs to separate the two groups and produce sensitivity and specificity values

Use complete LOO cross validation to validate sensitivity and specificity values for each pairing

For prognosis, attempt to quantify any differences between disease stages

Post analysis

Examine which peaks are of greatest difference between groups by examining the average graphs (with errors) and compare these with a spectral reference

Examine literature for disease of interest, if there are blood composition studies compare to see if the spectral reference findings relate to related cancer metabolic studies.

Otherwise suggest or perform a follow up study to ascertain serum compositional causes for the spectral shifts, identifying causal proteins etc.

Conclude the usefulness of the findings for the purpose(s) proposed

The studies within this thesis promise at a future where spectral diagnosis will be usable clinically. The methods in manuscript 2 and the subsequent spray-system variation have been effectively tested in a clinical environment. This, coupled with the ~90% accuracy results from these studies indicate that this method is viable for pancreatic cancer testing. The identification of the key <10kDa region of interest can assist greatly with an ease-of use diagnostic system. Furthermore, biological and chemical investigation of this region is far more achievable than investigating a patient's blood as a whole. Though the cell and NMR based studies to investigate this further require further iteration, more investigation into the causal molecules in this region are planned.

For the promise of this thesis to be realised, the method must reach true clinical validation. Having a multi-site diagnostic study on pancreatic patients is the next step towards this. In order to achieve that, the spray system and overall prototype must be refined further into a unified system or maybe even a single device. The current system requires a cost in the region of £10-20,000, mostly from the cost of the FTIR instrument and a refrigerated centrifuge. This is sufficiently low for a viable diagnostic system, but any reductions to the required components are essential for more widespread use. Once this has been achieved, there is little reason that this methodology cannot be expanded for pancreatic cancer screening and diagnostics. The potential for other cancer types to be expanded into is huge as well, the same blood test could even be sufficient for investigating other cancers. The promise of an all-in-one spectral test for cancer is a lofty goal, though achieving success on 2 different cancers suggests it is not impossible. Recent innovations like the Galleri test look at circulating DNA and promise valuable multi-cancer detection from a single blood test, though it currently will only pick up a cancer 50% of the time¹². It is possible spectral biomarkers could be added to such a

test, improving its detection rate. With sufficient further testing and development, this FTIR-based diagnostic method will soon have the ability to impact the clinical sphere.

10. Acknowledgements

I would like to thank EPSRC and Swansea University for funding and facilitating this work.

This endeavour would not have been possible without the guidance from my supervisor Dr Debdulal Roy. I am always grateful for his helpful suggestions and inspirations for my work. Furthermore, Dr Murali Chilakapati and Professor Bilal Al-Sarireh were key in establishing connections to hospitals and patient samples. Without their support, little achieved in this thesis would have been possible.

I'd also like to extend my sincerest thanks to Matt Mortimer and Arti Hole for their support and help accessing patient samples throughout my thesis work. The additional collaboration of Professor Venkateswarlu Kanamarlapudi and Dr Benjamin Mora was invaluable.

It is important to acknowledge the work of my examiners, Professor Owen Guy and Professor Hugh Barr, and chair, Professor Gil Alexandrowicz who helped my viva be an enjoyable experience.

Lastly, I'd like to thank my partner, Meghan Hosch, and my parents, Karen and Ian Duckworth, for helping support me throughout my academic career. My father has always encouraged my work, helping proofread much of my writing up to and including this thesis itself.

11. References

- 1 Baker, M. J. *et al.* Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chem Soc Rev* **45**, 1803-1818 (2016). <https://doi.org:10.1039/c5cs00585j>
- 2 Van der Hoorn A, v. L. P., Holtman GA, Westerlaan HE. Vol. 12 (PLOS ONE, 2017).
- 3 Noriaki Ohuchi, A. S., Tomotaka Sobue, Masaaki Kawai, Seiichiro Yamamoto, Ying-Fang Zheng, Yoko Narikawa Shiono, Hiroshi Saito, Shinichi Kuriyama, Eriko Tohno, Tokiko Endo, Akira Fukao, Ichiro Tsuji, Takuhiro Yamaguchi, Yasuo Ohashi, Mamoru Fukuda, Takanori Ishida,. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial,. *The Lancet*, **387**, 341-348 (2016,).
- 4 NHS. Achieving world-class cancer outcomes: a strategy for England 2015-2020. (2015).
- 5 Organisation, W. H. *All cancer's fact sheet*, (2020).
- 6 Organization, W. H. *Guide to cancer early diagnosis*, (2017).
- 7 NHS. Achieving World-Class Cancer Outcomes: A strategy for England 2015-2010. Progress report 2016-17. (2017).
- 8 Wise, J. Mobile lung cancer testing in supermarket car parks is to be expanded. *BMJ*, j5450 (2017). <https://doi.org:10.1136/bmj.j5450>
- 9 Ilic, D. *et al.* Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. *BMJ* **362**, k3519 (2018). <https://doi.org:10.1136/bmj.k3519>
- 10 Clinic, M. , (2021).
- 11 www.costevaluation.com/.
- 12 Nadauld, L. & Goldman, D. P. Considerations in the implementation of multicancer early detection tests. *Future Oncology* (2022). <https://doi.org:10.2217/fon-2022-0120>
- 13 Tokuda, O., Harada, Y., Ohishi, Y., Matsunaga, N. & Edenbrandt, L. Investigation of computer-aided diagnosis system for bone scans: a retrospective analysis in 406 patients. *Annals of Nuclear Medicine* **28**, 329-339 (2014). <https://doi.org:10.1007/s12149-014-0819-8>
- 14 Cancer.gov, (2022).
- 15 Bonifacio, A. *et al.* Surface-enhanced Raman spectroscopy of blood plasma and serum using Ag and Au nanoparticles: a systematic study. *Anal Bioanal Chem* **406**, 2355-2365 (2014). <https://doi.org:10.1007/s00216-014-7622-1>
- 16 Bonnier, F. *et al.* Screening the low molecular weight fraction of human serum using ATR-IR spectroscopy. *J Biophotonics* **9**, 1085-1097 (2016). <https://doi.org:10.1002/jbio.201600015>
- 17 Krafft, C. Modern trends in biophotonics for clinical diagnosis and therapy to solve unmet clinical needs. *Journal of Biophotonics* **9**, 1362-1375 (2016). <https://doi.org:10.1002/jbio.201600290>
- 18 Mitchell, A. L., Gajjar, K. B., Theophilou, G., Martin, F. L. & Martin-Hirsch, P. L. Vibrational spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory to a clinical setting: Vibrational spectroscopy of biofluids: laboratory to clinical setting. *Journal of Biophotonics* **7**, 153-165 (2014). <https://doi.org:10.1002/jbio.201400018>
- 19 Barth, A. Infrared spectroscopy of proteins. *Biochim Biophys Acta* **1767**, 1073-1101 (2007). <https://doi.org:10.1016/j.bbabi.2007.06.004>

- 20 R.A. Hoult, B. Perston & Spragg, R. A. Polystyrene Film as a Standard for Testing FT-IR Spectrometers. **28** (2013). <<http://www.spectroscopyonline.com/polystyrene-film-standard-testing-ft-ir-spectrometers?id=&sk=&date=&pageID=3>>.
- 21 Elmer, P. FTIR Spectroscopy: Attenuated Total Reflectance (ATR). 5 (2005).
- 22 Kazarian, S. G. & Chan, K. L. ATR-FTIR spectroscopic imaging: recent advances and applications to biological systems. *Analyst* **138**, 1940-1951 (2013). <https://doi.org:10.1039/c3an36865c>
- 23 Kazarian, S. G. & Chan, K. L. Applications of ATR-FTIR spectroscopic imaging to biomedical samples. *Biochim Biophys Acta* **1758**, 858-867 (2006). <https://doi.org:10.1016/j.bbamem.2006.02.011>
- 24 Grdadolnik, J. ATR-FTIR spectroscopy: Its advantages and limitations. *Acta Chim Slov* **49**, 631-642 (2002).
- 25 Shimadzu. Q: How deep does the infrared light penetrate at the position of contact between the prism and sample during ATR measurements?, <<https://www.shimadzu.com/an/ftir/support/faq/2.html>> (2018).
- 26 Le Corvec, M. et al. Mid-infrared spectroscopy of serum, a promising non-invasive method to assess prognosis in patients with ascites and cirrhosis. *PLoS One* **12**, e0185997 (2017). <https://doi.org:10.1371/journal.pone.0185997>
- 27 Dorling, K. M. & Baker, M. J. Rapid FTIR chemical imaging: highlighting FPA detectors. *Trends in Biotechnology* **31**, 437-438 (2013). <https://doi.org:10.1016/j.tibtech.2013.05.008>
- 28 Barth, A. & Zscherp, C. What vibrations tell us about proteins. *Q Rev Biophys* **35**, 369-430 (2002). <https://doi.org:10.1017/S0033583502003815>
- 29 Goodacre, R. M. J. a. R. Discrimination of Bacteria Using Surface-Enhanced Raman Spectroscopy. *Anal. Chem.* **76**, 40-47 (2004).
- 30 Thomas, G. J., Jr. Raman spectroscopy of protein and nucleic acid assemblies. *Annu Rev Biophys Biomol Struct* **28**, 1-27 (1999). <https://doi.org:10.1146/annurev.biophys.28.1.1>
- 31 Socrates, G. *Infrared and Raman Characteristic Group Frequencies*. 3rd edn, (JOHN WILEY & SONS, LTD, 2001).
- 32 Kerr, L. T., Byrne, H. J. & Hennelly, B. M. Optimal choice of sample substrate and laser wavelength for Raman spectroscopic analysis of biological specimen. *Analytical Methods* **7**, 5041-5052 (2015). <https://doi.org:10.1039/c5ay00327j>
- 33 Crystran. (2013).
- 34 Rae, A., Stosch, R., Klapetek, P., Hight Walker, A. R. & Roy, D. State of the art Raman techniques for biological applications. *Methods* **68**, 338-347 (2014). <https://doi.org:10.1016/j.ymeth.2014.02.035>
- 35 Pence, I. & Mahadevan-Jansen, A. Clinical instrumentation and applications of Raman spectroscopy. *Chem Soc Rev* **45**, 1958-1979 (2016). <https://doi.org:10.1039/c5cs00581g>
- 36 Kumar, N., Mignuzzi, S., Su, W. & Roy, D. Tip-enhanced Raman spectroscopy: principles and applications. *EPJ Techniques and Instrumentation* **2** (2015). <https://doi.org:10.1140/epjti/s40485-015-0019-5>
- 37 Premasiri, W. R., Lee, J. C. & Ziegler, L. D. Surface-enhanced Raman scattering of whole human blood, blood plasma, and red blood cells: cellular processes and bioanalytical sensing. *J Phys Chem B* **116**, 9376-9386 (2012). <https://doi.org:10.1021/jp304932g>
- 38 Zhang, K. et al. Label-free and stable serum analysis based on Ag-NPs/PSi surface-enhanced Raman scattering for noninvasive lung cancer detection. *Biomedical Optics Express* **9** (2018). <https://doi.org:10.1364/boe.9.004345>
- 39 Hughes, C. et al. Assessing the challenges of Fourier transform infrared spectroscopic analysis of blood serum. *Journal of Biophotonics* **7**, 180-188 (2014). <https://doi.org:10.1002/jbio.201300167>

- 40 Lovergne, L. *et al.* Investigating optimum sample preparation for infrared spectroscopic serum diagnostics. *Analytical Methods* **7**, 7140-7149 (2015). <https://doi.org/10.1039/c5ay00502g>
- 41 Cameron, J. M., Butler, H. J., Palmer, D. S. & Baker, M. J. Biofluid spectroscopic disease diagnostics: A review on the processes and spectral impact of drying. *J Biophotonics* **11**, e201700299 (2018). <https://doi.org/10.1002/jbio.201700299>
- 42 Yen, T. M. *et al.* Reversing Coffee-Ring Effect by Laser-Induced Differential Evaporation. *Sci Rep* **8**, 3157 (2018). <https://doi.org/10.1038/s41598-018-20581-0>
- 43 Depciuch, J. & Parlinska-Wojtan, M. Comparing dried and liquid blood serum samples of depressed patients: An analysis by Raman and infrared spectroscopy methods. *J Pharm Biomed Anal* **150**, 80-86 (2018). <https://doi.org/10.1016/j.jpba.2017.11.074>
- 44 Barlev, E. *et al.* A novel method for screening colorectal cancer by infrared spectroscopy of peripheral blood mononuclear cells and plasma. *Journal of Gastroenterology* **51**, 214-221 (2016). <https://doi.org/10.1007/s00535-015-1095-7>
- 45 Ollesch, J. *et al.* An infrared spectroscopic blood test for non-small cell lung carcinoma and subtyping into pulmonary squamous cell carcinoma or adenocarcinoma. *Biomedical Spectroscopy and Imaging* **5**, 129-144 (2016). <https://doi.org/10.3233/BSI-160144>
- 46 Wang, X., Shen, X., Sheng, D., Chen, X. & Liu, X. FTIR spectroscopic comparison of serum from lung cancer patients and healthy persons. *Spectrochim Acta A Mol Biomol Spectrosc* **122**, 193-197 (2014). <https://doi.org/10.1016/j.saa.2013.11.049>
- 47 Wan, Q. S. & Zhang, K. H. Noninvasive detection of gastric cancer. *Tumour Biol* **37**, 11633-11643 (2016). <https://doi.org/10.1007/s13277-016-5129-4>
- 48 Lin, J. *et al.* A novel blood plasma analysis technique combining membrane electrophoresis with silver nanoparticle-based SERS spectroscopy for potential applications in noninvasive cancer detection. *Nanomedicine* **7**, 655-663 (2011). <https://doi.org/10.1016/j.nano.2011.01.012>
- 49 Chen, Y. *et al.* Discrimination of gastric cancer from normal by serum RNA based on surface-enhanced Raman spectroscopy (SERS) and multivariate analysis. *Med Phys* **39**, 5664-5668 (2012). <https://doi.org/10.1118/1.4747269>
- 50 Ito, H. *et al.* Use of surface-enhanced Raman scattering for detection of cancer-related serum-constituents in gastrointestinal cancer patients. *Nanomedicine* **10**, 599-608 (2014). <https://doi.org/10.1016/j.nano.2013.09.006>
- 51 Jenkins, C. A. *et al.* A high-throughput serum Raman spectroscopy platform and methodology for colorectal cancer diagnostics. *Analyst* **143**, 6014-6024 (2018). <https://doi.org/10.1039/c8an01323c>
- 52 Cacciatore, S. & Loda, M. Innovation in metabolomics to improve personalized healthcare. *Ann N Y Acad Sci* **1346**, 57-62 (2015). <https://doi.org/10.1111/nyas.12775>
- 53 Paraskevaidi, M. *et al.* Raman spectroscopic techniques to detect ovarian cancer biomarkers in blood plasma. *Talanta* **189**, 281-288 (2018). <https://doi.org/10.1016/j.talanta.2018.06.084>
- 54 Ryzhikova, E. *et al.* Raman spectroscopy of blood serum for Alzheimer's disease diagnostics: specificity relative to other types of dementia. *J Biophotonics* **8**, 584-596 (2015). <https://doi.org/10.1002/jbio.201400060>
- 55 Krafft, C. *et al.* A specific spectral signature of serum and plasma-derived extracellular vesicles for cancer screening. *Nanomedicine* **13**, 835-841 (2017). <https://doi.org/10.1016/j.nano.2016.11.016>
- 56 Titus, J., Ghimire, H., Viennois, E., Merlin, D. & Unil Perera, A. G. Protein secondary structure analysis of dried blood serum using infrared spectroscopy to identify markers for colitis screening. *Journal of Biophotonics* **11**, e201700057 (2018). <https://doi.org/10.1002/jbio.201700057>

- 57 Depciuch, J. *et al.* Phospholipid-protein balance in affective disorders: Analysis of human blood serum using Raman and FTIR spectroscopy. A pilot study. *Journal of Pharmaceutical and Biomedical Analysis* **131**, 287-296 (2016). <https://doi.org/10.1016/j.jpba.2016.08.037>
- 58 Depciuch, J. *et al.* The role of zinc deficiency-induced changes in the phospholipid-protein balance of blood serum in animal depression model by Raman, FTIR and UV-vis spectroscopy. *Biomedicine & Pharmacotherapy* **89**, 549-558 (2017). <https://doi.org/10.1016/j.biopha.2017.01.180>
- 59 Khan, S. *et al.* Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM). *Biomed Opt Express* **7**, 2249-2256 (2016). <https://doi.org/10.1364/BOE.7.002249>
- 60 Khan, S. *et al.* Random Forest-Based Evaluation of Raman Spectroscopy for Dengue Fever Analysis. *Appl Spectrosc* **71**, 2111-2117 (2017). <https://doi.org/10.1177/0003702817695571>
- 61 Khan, S. *et al.* Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning. *Photodiagnosis Photodyn Ther* **23**, 89-93 (2018). <https://doi.org/10.1016/j.pdpdt.2018.05.010>
- 62 Sohail, A. *et al.* Analysis of hepatitis C infection using Raman spectroscopy and proximity based classification in the transformed domain. *Biomed Opt Express* **9**, 2041-2055 (2018). <https://doi.org/10.1364/BOE.9.002041>
- 63 Wang, R. & Wang, Y. Fourier Transform Infrared Spectroscopy in Oral Cancer Diagnosis. *International Journal of Molecular Sciences* **22** (2021). <https://doi.org/10.3390/ijms22031206>
- 64 Mahadevan-Jansen, A. *et al.* in *Biomedical Vibrational Spectroscopy VI: Advances in Research and Industry* (2014).
- 65 Sahu, A. K. *et al.* Oral cancer screening: serum Raman spectroscopic approach. *J Biomed Opt* **20**, 115006 (2015). <https://doi.org/10.1117/1.JBO.20.11.115006>
- 66 Szymoński, K. *et al.* Spectroscopic screening of pancreatic cancer. *Clinical Spectroscopy* **3**, 100016 (2021). <https://doi.org/https://doi.org/10.1016/j.clispe.2021.100016>
- 67 Auner, G. W. *et al.* Applications of Raman spectroscopy in cancer diagnosis. *Cancer and Metastasis Reviews* **37**, 691-717 (2018). <https://doi.org/10.1007/s10555-018-9770-9>
- 68 Diem, M. Comments on recent reports on infrared spectral detection of disease markers in blood components. *Journal of Biophotonics* **11**, e201800064 (2018). <https://doi.org/10.1002/jbio.201800064>
- 69 Spalding, K. *et al.* Enabling quantification of protein concentration in human serum biopsies using attenuated total reflectance – Fourier transform infrared (ATR-FTIR) spectroscopy. *Vibrational Spectroscopy* **99**, 50-58 (2018). <https://doi.org/10.1016/j.vibspec.2018.08.019>
- 70 Bonnier, F. *et al.* Ultra-filtration of human serum for improved quantitative analysis of low molecular weight biomarkers using ATR-IR spectroscopy. *Analyst* **142**, 1285-1298 (2017). <https://doi.org/10.1039/c6an01888b>
- 71 Naseer, K., Amin, A., Saleem, M. & Qazi, J. Raman spectroscopy based differentiation of typhoid and dengue fever in infected human sera. *Spectrochim Acta A Mol Biomol Spectrosc* **206**, 197-201 (2019). <https://doi.org/10.1016/j.saa.2018.08.008>
- 72 Wang, H. *et al.* Screening and staging for non-small cell lung cancer by serum laser Raman spectroscopy. *Spectrochim Acta A Mol Biomol Spectrosc* **201**, 34-38 (2018). <https://doi.org/10.1016/j.saa.2018.04.002>
- 73 Lu, Y. *et al.* Label free hepatitis B detection based on serum derivative surface enhanced Raman spectroscopy combined with multivariate analysis. *Biomed Opt Express* **9**, 4755-4766 (2018). <https://doi.org/10.1364/BOE.9.004755>

- 74 Zheng, X. X. *et al.* Rapid and Low-Cost Detection of Thyroid Dysfunction Using Raman Spectroscopy and an Improved Support Vector Machine. *Ieee Photonics J* **10** (2018). <https://doi.org/Artn3901412> 10.1109/Jphot.2018.2876686
- 75 Guo, J. *et al.* Diagnosis of chronic kidney diseases based on surface-enhanced Raman spectroscopy and multivariate analysis. *Laser Phys* **28** (2018). <https://doi.org:ARTN075603> 10.1088/1555-6611/aabec5
- 76 Khan, S., Ullah, R., Shahzad, S., Javaid, S. & Khan, A. Optical screening of nasopharyngeal cancer using Raman spectroscopy and support vector machine. *Optik* **157**, 565-570 (2018). <https://doi.org:10.1016/j.ijleo.2017.11.97>
- 77 Stefancu, A. *et al.* Combining SERS analysis of serum with PSA levels for improving the detection of prostate cancer. *Nanomedicine (Lond)* **13**, 2455-2467 (2018). <https://doi.org:10.2217/nnm-2018-0127>
- 78 Shao, L. *et al.* Fast and non-invasive serum detection technology based on surface-enhanced Raman spectroscopy and multivariate statistical analysis for liver disease. *Nanomedicine* **14**, 451-459 (2018). <https://doi.org:10.1016/j.nano.2017.11.022>
- 79 Hands, J. R. *et al.* Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy. *J Neurooncol* **127**, 463-472 (2016). <https://doi.org:10.1007/s11060-016-2060-x>
- 80 Paraskevasidi, M. *et al.* Differential diagnosis of Alzheimer's disease using spectrochemical analysis of blood. *Proc Natl Acad Sci U S A* **114**, E7929-E7938 (2017). <https://doi.org:10.1073/pnas.1701517114>
- 81 Lin, H. *et al.* Species identification of bloodstains by ATR-FTIR spectroscopy: the effects of bloodstain age and the deposition environment. *Int J Legal Med* **132**, 667-674 (2018). <https://doi.org:10.1007/s00414-017-1634-2>
- 82 Agbaria, A. H. *et al.* Differential Diagnosis of the Etiologies of Bacterial and Viral Infections Using Infrared Microscopy of Peripheral Human Blood Samples and Multivariate Analysis. *Anal Chem* **90**, 7888-7895 (2018). <https://doi.org:10.1021/acs.analchem.8b00017>
- 83 Ghimire, H., Venkataramani, M., Bian, Z., Liu, Y. & Perera, A. G. U. ATR-FTIR spectral discrimination between normal and tumorous mouse models of lymphoma and melanoma from serum samples. *Sci Rep* **7**, 16993 (2017). <https://doi.org:10.1038/s41598-017-17027-4>
- 84 Byrne, H. J., Knief, P., Keating, M. E. & Bonnier, F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. *Chem Soc Rev* **45**, 1865-1878 (2016). <https://doi.org:10.1039/c5cs00440c>
- 85 Knief, P. *Interactions of Carbon Nanotubes with Human Lung Epithelial Cells in vitro, Assessed by Raman Spectroscopy* PhD thesis, Dublin Institute of Technology, (2010).
- 86 Ivanov, Y. V., Karimov, A. R., Pyatnitsky, L. N., Seryakov, A. P. & Shcheglov, V. A. Light Scattering by Human Blood Plasma. *Journal of Russian Laser Research* **26**, 363-372 (2005). <https://doi.org:10.1007/s10946-005-0039-8>
- 87 Bassan, P. *et al.* RMieS-EMSC correction for infrared spectra of biological cells: extension using full Mie theory and GPU computing. *J Biophotonics* **3**, 609-620 (2010). <https://doi.org:10.1002/jbio.201000036>
- 88 Bassan, P. *Light scattering during infrared spectroscopic measurements of biomedical samples*, (2011).
- 89 Eilers, P. a. B. H. Baseline Correction with Asymmetric Least Squares Smoothing. *Unpubl. Manuscr* (2005).
- 90 Zhang, Z.-M. *et al.* An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *Journal of Raman Spectroscopy* **41**, 659-669 (2009). <https://doi.org:10.1002/jrs.2500>

- 91 Choquette, S. J., Etz, E. S., Hurst, W. S., Blackburn, D. H. & Leigh, S. D. Relative intensity correction of Raman spectrometers: NIST SRMs 2241 through 2243 for 785 nm, 532 nm, and 488 nm/514.5 nm excitation. *Applied Spectroscopy* **61**, 117-129 (2007). <https://doi.org/Doi.10.1366/000370207779947585>
- 92 Anne L. Plant, R. L. W. (NIST, 2013).
- 93 Pych, W. A fast algorithm for cosmic-ray removal from single images. *Publ Astron Soc Pac* **116**, 148-153 (2004). <https://doi.org/Doi.10.1086/381786>
- 94 Kumar, S. *et al.* Raman and infra-red microspectroscopy: towards quantitative evaluation for clinical research by ratiometric analysis. *Chem Soc Rev* **45**, 1879-1900 (2016). <https://doi.org:10.1039/c5cs00540j>
- 95 Xia, J., Broadhurst, D. I., Wilson, M. & Wishart, D. S. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics* **9**, 280-299 (2013). <https://doi.org:10.1007/s11306-012-0482-9>
- 96 (ed Jacob Cohen) (L. Erlbaum Associates, Mahwah, N.J. :, 2003).
- 97 Gromski, P. S. *et al.* A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. *Anal Chim Acta* **879**, 10-23 (2015). <https://doi.org:10.1016/j.aca.2015.02.012>
- 98 Zhao, Q. B., Zhang, L. Q. & Cichocki, A. Multilinear and nonlinear generalizations of partial least squares: an overview of recent advances. *Wires Data Min Knowl* **4**, 104-115 (2014). <https://doi.org:10.1002/widm.1120>
- 99 Rosipal, R. in *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques* Ch. 9, 169-189 (2010).
- 100 Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G. & Mechelli, A. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev* **36**, 1140-1152 (2012). <https://doi.org:10.1016/j.neubiorev.2012.01.004>
- 101 Shen, D., Wu, G. & Suk, H. I. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng* **19**, 221-248 (2017). <https://doi.org:10.1146/annurev-bioeng-071516-044442>
- 102 Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* **13**, 8-17 (2015). <https://doi.org:10.1016/j.csbj.2014.11.005>
- 103 Acquarelli, J. *et al.* Convolutional neural networks for vibrational spectroscopic data analysis. *Anal Chim Acta* **954**, 22-31 (2017). <https://doi.org:10.1016/j.aca.2016.12.010>
- 104 Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov Today* **23**, 1241-1250 (2018). <https://doi.org:10.1016/j.drudis.2018.01.039>
- 105 Hamaguchi, H. K. Y. S. H. Y. N. R. S. H. o. in *Raman, infrared and near infrared chemical imaging* (ed Slobodan Šašić Yukihiro Ozaki) (2011).
- 106 Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. Valid post-selection inference. *The Annals of Statistics* **41**, 802-837, 836 (2013).
- 107 Mercaldo, N. D., Lau, K. F. & Zhou, X. H. Confidence intervals for predictive values with an emphasis to case-control studies. *Stat Med* **26**, 2170-2183 (2007). <https://doi.org:10.1002/sim.2677>
- 108 Ma, H. *et al.* Antibody-Free Discrimination of Protein Biomarkers in Human Serum Based on Surface-Enhanced Raman Spectroscopy. *Anal Chem* **90**, 12342-12346 (2018). <https://doi.org:10.1021/acs.analchem.8b03701>
- 109 Neng, J., Li, Y., Driscoll, A. J., Wilson, W. C. & Johnson, P. A. Detection of Multiple Pathogens in Serum Using Silica-Encapsulated Nanotags in a Surface-Enhanced Raman Scattering-Based Immunoassay. *J Agric Food Chem* **66**, 5707-5712 (2018). <https://doi.org:10.1021/acs.jafc.8b00026>

- 110 Kuhar, N., Sil, S., Verma, T. & Umapathy, S. Challenges in application of Raman spectroscopy to biology and materials. *Rsc Adv* **8**, 25888-25908 (2018). <https://doi.org:10.1039/c8ra04491k>
- 111 Dudek, M. *et al.* Raman Optical Activity and Raman spectroscopy of carbohydrates in solution. *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.* **206**, 597-612 (2019). <https://doi.org:10.1016/j.saa.2018.08.017>
- 112 Elgrishi, N. *et al.* A Practical Beginner's Guide to Cyclic Voltammetry. *Journal of Chemical Education* **95**, 197-206 (2017). <https://doi.org:10.1021/acs.jchemed.7b00361>
- 113 Brazhe, N. A., Treiman, M., Faricelli, B., Vestergaard, J. H. & Sosnovtseva, O. In situ Raman study of redox state changes of mitochondrial cytochromes in a perfused rat heart. *PLoS One* **8**, e70488 (2013). <https://doi.org:10.1371/journal.pone.0070488>
- 114 Smith, R., Wright, K. L. & Ashton, L. Raman spectroscopy: an evolving technique for live cell studies. *Analyst* **141**, 3590-3600 (2016). <https://doi.org:10.1039/c6an00152a>
- 115 Ertel, A., Verghese, A., Byers, S. W., Ochs, M. & Tozeren, A. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Molecular Cancer* **5** (2006). <https://doi.org:10.1186/1476-4598-5-55>
- 116 Owens, N. A. *et al.* Handheld Raman Spectrometer Instrumentation for Quantitative Tuberculosis Biomarker Detection: A Performance Assessment for Point-of-Need Infectious Disease Diagnostics. *Appl Spectrosc* **72**, 1104-1115 (2018). <https://doi.org:10.1177/0003702818770666>
- 117 Borek Puza, T. O. n. Generalised Clopper–Pearson confidence intervals for the binomial proportion. *Journal of Statistical Computation and Simulation* **76**, 489-508 (2006). <https://doi.org:10.1080/10629360500107527>
- 118 Walsh, I. *et al.* DOME: recommendations for supervised machine learning validation in biology. *Nat Methods* **18**, 1122-1127 (2021). <https://doi.org:10.1038/s41592-021-01205-4>
- 119 Shen, X. *et al.* Study on baseline correction methods for the Fourier transform infrared spectra with different signal-to-noise ratios. *Applied Optics* **57**, 5794-5799 (2018). <https://doi.org:10.1364/AO.57.005794>
- 120 Afseth, N. K. & Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems* **117**, 92-99 (2012). <https://doi.org:https://doi.org/10.1016/j.chemolab.2012.03.004>
- 121 Duckworth, E. *et al.* Improving Vibrational Spectroscopy Prospects in Frontline Clinical Diagnosis: Fourier Transform Infrared on Buccal Mucosa Cancer. *Analytical Chemistry* (2022). <https://doi.org:10.1021/acs.analchem.2c02496>
- 122 Song, Z., Wang, H., Yin, X., Deng, P. & Jiang, W. Application of NMR metabolomics to search for human disease biomarkers in blood. *Clin Chem Lab Med* **57**, 417-441 (2019). <https://doi.org:10.1515/cclm-2018-0380>
- 123 Aldrich, S. *IR Spectrum Table*, (2019).
- 124 Wang, X., Kaczor-Urbanowicz, K. E. & Wong, D. T. W. Salivary biomarkers in cancer detection. *Medical Oncology* **34**, 7 (2016). <https://doi.org:10.1007/s12032-016-0863-4>
- 125 Shixuan He, W. Z. *et al.* Baseline correction for Raman spectra using an improved asymmetric least squares method. *Analytical Methods* **6**, 4402-4407 (2014).
- 126 Wise, J. Mobile lung cancer testing in supermarket car parks is to be expanded. *British Medical Journal* **359**, j5450 (2017). <https://doi.org:10.1136/bmj.j5450>
- 127 Leal, L. B., Nogueira, M. S., Canevari, R. A. & Carvalho, L. Vibration spectroscopy and body biofluids: Literature review for clinical applications. *Photodiagnosis Photodyn Ther* **24**, 237-244 (2018). <https://doi.org:10.1016/j.pdpdt.2018.09.008>
- 128 Butler, H. J. *et al.* Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. *Nature Communications* **10**, 4501 (2019). <https://doi.org:10.1038/s41467-019-12527-5>

- 129 Borek Puza and Terence, O. n. Generalised Clopper–Pearson confidence intervals for the binomial proportion. *Journal of Statistical Computation and Simulation* **76**, 489-508 (2006). <https://doi.org:10.1080/10629360500107527>
- 130 Mouche, A. & Pedeux, R. *Cancer cell culture basics handbook*. (2020).
- 131 Muhamadali, H., Chisanga, M., Subaihi, A. & Goodacre, R. Combining Raman and FT-IR Spectroscopy with Quantitative Isotopic Labeling for Differentiation of E. coli Cells at Community and Single Cell Levels. *Analytical Chemistry* **87**, 4578-4586 (2015). <https://doi.org:10.1021/acs.analchem.5b00892>
- 132 Balmaña, M. *et al.* Analysis of sialyl-Lewis x on MUC5AC and MUC1 mucins in pancreatic cancer tissues. *International Journal of Biological Macromolecules* **112**, 33-45 (2018). <https://doi.org:https://doi.org/10.1016/j.ijbiomac.2018.01.148>
- 133 Furukawa, T. *et al.* Long-term culture and immortalization of epithelial cells from normal adult human pancreatic ducts transfected by the E6E7 gene of human papilloma virus 16. *Am J Pathol* **148**, 1763-1770 (1996).
- 134 Deer, E. L. *et al.* Phenotype and genotype of pancreatic cancer cell lines. *Pancreas* **39**, 425-435 (2010). <https://doi.org:10.1097/MPA.0b013e3181c15963>

Vibrational Spectroscopy Prospects in Frontline Clinical Diagnosis: Appendix

Contents

1. Additional Figures	2
Methods	2
Oral Cancer Section	3
Pancreatic section	4
2. Additional investigation.....	10
2.1 2019 Raman and FTIR investigation on Buccal Mucosa Cancer	10
Introduction.....	10
Pre-study design	10
Methods	16
Results and discussion	18
Discussion	23
Conclusions.....	26
2.2 Biological spectral imaging	27
2.3 Spectral differentiation and machine correction app development	29
Spectral differentiation app.....	29
Instrument Difference Correction System	30
Results.....	33
Discussion	34
3. References.....	37

1. Additional Figures

Methods

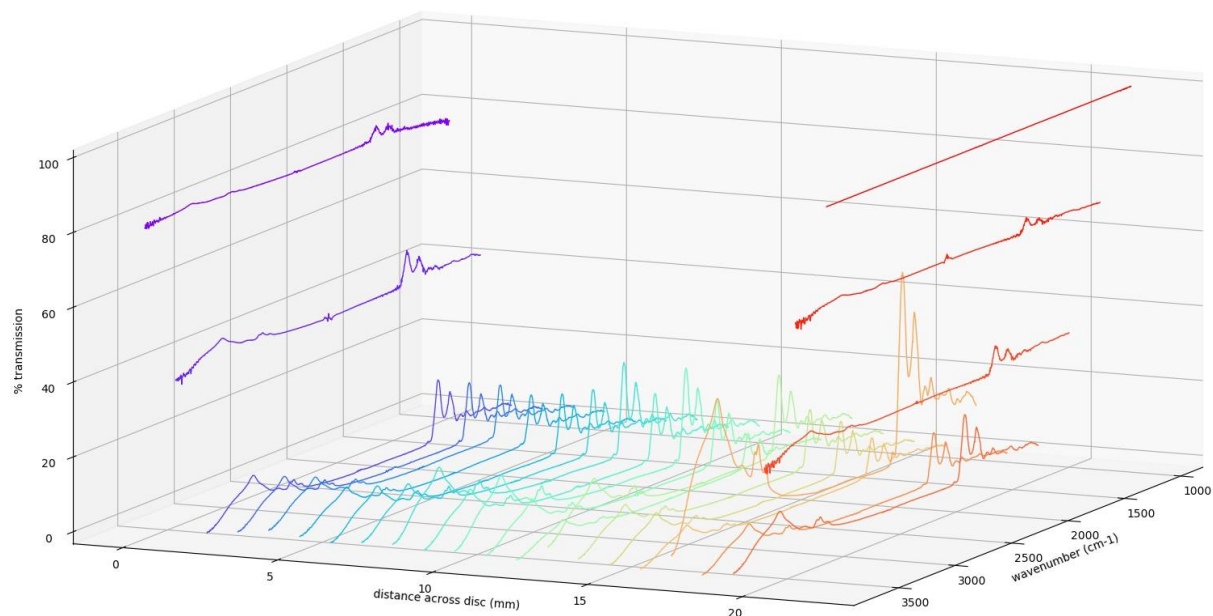


Figure A1.1: Serum FTIR spectra taken at uniformly spaced points across a CaF₂ disc. Serum deposited and dried using the total coating method. High % transmission on edge samples results from when the light spot overlaps the edge of the disc.

Ignoring the first two and last three spectra that have been shifted by the light being transmitted around the edge of the disc, the spectra are measurably different. This difference can be summarised by PCA of the spectra - the third principal component varying most with distance from the edge. The innermost 9 spectra had an average value of -2 ± 1 and the outermost 8 spectra 3.1 ± 0.7 for this component.

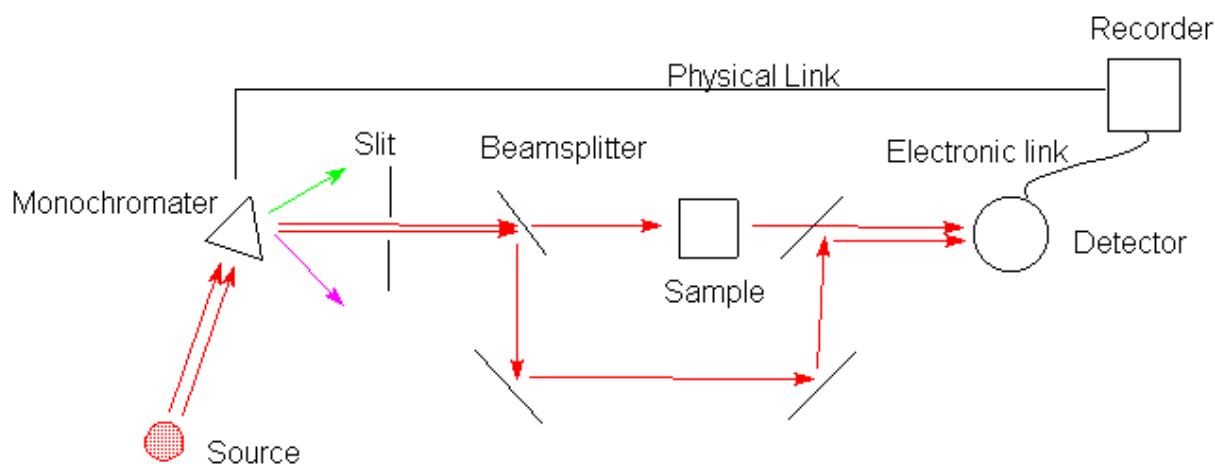


Figure A1.2: FTIR diagram

Oral Cancer Section

Table A1.1: patient age distribution

	Average age	Standard Deviation (SD)	2SD error
Healthy (with and without tobacco habit)	37	10.9	5.5
Cancer	59	15.7	6.7
Premalignant	42	11.2	4.8

Table A1.2: Design of the study and specification of patient numbers and molecular weight regions of serum.		
Patients	Study involving whole serum, low mol weight and high mol weight windows of the serum	Study involving the 10-30 kDa molecular weight window
Buccal mucosa cancers	42	9
Premalignant oral conditions (leucoplakia)	40	9

Healthy tobacco user	17	0
Healthy (not tobacco user)	27	0
Total	126	18

Pancreatic section

Patient information

The table (A1.1) below outlines the patient information for the cohort used in the pancreatic cancer case study in §7.1, as well as corresponding ELISA data for the same sample set.

Table A1.3: Table showing patient profiles (anonymous) with the initial diagnosis, final diagnosis, information from ELISA tests and information from our FTIR tests. The general principle for the colour codes: green – predicted correctly with high probability, yellow – predicted correctly with medium probability and red- predicted wrong. More specifically, for the Elisa the numbers are a reflection of the concentration of Sial Lewis (Ca-19.9) in the patient’s blood, higher concentrations than 59 U/ml suggesting the patient has pancreatic cancer. More specific ranges of CA19-9 for each diagnostic group (both for ELISA and FTIR) were supplied: 2-30 U/ml classifies benign (Group 3), 33-58 U/ml classifies control (Group 4), 59-81 U/ml classifies early (resectable) cancer (Group 1), and >82 U/ml classifies advanced (severe) stage (Group 2). The same classification has been used for FTIR diagnosis column. Green indicates correct diagnosis and patient group. Yellow indicates a sample was correctly diagnosed as cancer/non-cancer but was in the incorrect group. Red indicates incorrect diagnosis. For FTIR the number indicates the certainty of the model’s prediction. Above 0= cancer,

below 0 =non-cancer. Yellow indicates that at least one of the 3 repeats was diagnosing incorrectly, but they averaged out to a correct diagnosis. PDAC: pancreatic ductal adenocarcinoma, IPMN: intraductal papillary mucinous neoplasm.

Initial diagnosis	Final Clinical Diagnosis	Age	Sex	ELISA number	Grouping by ELSIA	FTIR number	Grouping by FTIR
Group 1, Early Cancer				Correct diagnosis: 1, 2			
1.03	Resected PDAC	73	M	110.4	2	11.0	1
1.04	Resected PDAC	62	M	136.4	2	10.1	1
1.05	Resected PDAC	61	M	30.06	3	9.8	1
1.08	Resected PDAC	70	M	143.1	2	9.6	1
1.12	Resected PDAC	68	M	19.7	3	7.5	1
1.14	Resected PDAC	71	M	19.8	3	6.3	1
1.16	Resected PDAC	56	M	88.8	2	3.4	1
1.17	Resected PDAC	69	M	151.0	2	7.0	1
1.18	Resectable PDAC – Patient borderline	76	M	109.0	2	4.9	1
1.19	Resectable PDAC	65	M	129.4	2	5.4	1
1.20	Resectable PDAC	72	M	155.0	2	13.1	1
1.21	Resectable PDAC	75	M	46.4	4	4.8	1
1.22	Resected PDAC	62	M	-	-	5.2	1
Group 2, Late stage Cancer				Correct diagnosis: 1, 2			
1.01	Metastatic PDAC	78	F	126.9	2	6.7	1
1.02	Locally Advanced PDAC	59	F	68.4	1	5.3	1
1.06	Locally Advanced PDAC	70	F	63.9	1	13.7	1
1.07	Locally Advanced PDAC	67	M	125.1	2	4.2	1

1.09	Metastatic PDAC	74	M	111.3	2	5.5	1
1.10	Locally Advanced PDAC	64	M	60.3	1	7.9	1
1.13	Metastatic PDAC	73	F	57.8	4	5.9	1
1.15	Metastatic PDAC	73	F	-	-	-0.2	3
2.01	Locally Advanced PDAC	72	F	114.6	2	9.7	1
2.02	Locally Advanced PDAC	75	M	198.2	2	4.9	1
2.03	Locally Advanced PDAC	72	F	175.2	2	6.0	1
2.04	Metastatic PDAC	61	M	129.3	2	9.7	1
2.05	Metastatic PDAC	70	M	162.9	2	3.1	1
2.06	Locally Advanced PDAC	76	M	55.9	4	5.8	1
2.07	Locally Advanced PDAC	69	F	-		14.8	1
2.08	Locally Advanced PDAC	82	F	-		12.9	1
2.09	Locally Advanced PDAC	50	-	-		13.5	1
Group 3, Benign				Correct diagnosis: 3, 4			
3.01	Chronic Pancreatitis	45	F	94.8	2	-5.2	3
3.02	Chronic Pancreatitis	42	M	74.9	1	-3.8	4
3.03	Acute Pancreatitis	80	F	61.2	1	-8.7	4
3.04	Main Duct IPMN (Cyst)	72	F	4.7	3	-2.9	4
3.05	Branch Duct IPMN (Cyst)	68	F	43.2	4	-6.3	4
3.06	Acute Pancreatitis	68	M	23.3	3	-13.2	4
3.07	Pancreatic Cyst	51	M	73.1	1	-14.6	4

3.08	Pancreatic Cyst	34	F	18.5	3	-2.9	4
3.11	Acute Pancreatitis	80	F	15.9	3	-4.1	4
3.12	IPMN (cyst)	48	F	20.9	3	-1.3	4
3.13	Acute Pancreatitis	64	M	43.7	4	0.1	2
3.14	Acute Pancreatitis	67	F	8.6	3	-7.7	4
3.15	Chronic Pancreatitis	40	F	-	-	-4.8	4
3.16	Chronic Pancreatitis	52	M	17.8	3	-4.5	4
3.17	Acute Pancreatitis	70	F	40.6	4	-6.6	4
3.18	Acute Pancreatitis	69	M	-	-	-5.0	4
3.19	Acute Pancreatitis	59	F	42.9	4	-6.5	4
3.20	Acute Pancreatitis	50	F	34.8	3	-11.9	4
3.21	Acute Pancreatitis	43	M	10.4	3	-10.4	4
3.22	Chronic Pancreatitis	68	M	46.9	4	-10.2	4
3.23	Acute Pancreatitis	42	M	36.1	3	-9.1	4
3.24	Chronic Pancreatitis	60	M	10.6	3	-13.8	4
3.25	Acute Pancreatitis	44	M	43.7	4	-6.6	4
3.26	Pancreatic Cyst	69	F	19.1	3	-8.5	4
3.27	Acute Pancreatitis	55	F	15.4	3	-4.4	4
3.28	Acute Pancreatitis	86	F	15.4	3	-4.1	4
3.29	Cyst IPMN	83	F	10.6	3	-9.0	4
3.30	Chronic Pancreatitis	65	F	40.6	3	-14.0	4
3.32	Branch Duct IPMN	76	F	33.7	3	-5.8	4
3.33	IPMN (Cyst)	64	F	36.6	3	-13.0	4
Group 4, Control				Correct diagnosis: 3, 4			
1.11	Benign Biliary Stricture	74	F	25.7	1	-9.3	4
3.09	CDKN2A mutation	56	F	10.3	3	-12.5	4
3.10	CDKN2A mutation	63	F	32.4	3	-1.7	3
4.01	Gallstones	65	M	23.1	3	-13.6	4
4.02	Gallstones	60	F	55.8	4	-6.5	4
4.03	Gallbladder Adenomyosis	57	F	152.5	2	-7.3	4
4.04	Anal Fistula	47	F	16.4	3	-11.5	4

4.05	Incisional Hernia	72	M	42.5	4	-8.1	4
4.06	Inguinal Hernia	67	M	14.8	3	-7.9	4
4.07	Sleeve Gastroectomy	60	F	Undetec table	-	-10.4	4
4.08	Gallstones	59	F	9.2	3	-9.3	4
4.09	Umbilical Hernia	56	F	8.4	3	-12.5	4

Figure A1.4: Pancreatic patient urine PCA-LDA 800-1800

Comparison	Principle components	Sensitivity	Specificity	Accuracy	confidence
Cancer + Early cancer v Healthy + Benign	5	79	92	86.1	79.1-91.4
Cancer v Healthy	2	80	92	86.8	76.5-93.7
Early cancer v Late stage cancer	48	83	72	77.2	64.9-86.9

Table A1.5: Cell media try 1, pancreatic cancer v healthy, PCA-LDA or SVM 800-4000

Comparison	Principle components	Sensitivity	Specificity	Accuracy	confidence
Whole SVM 15v12	4	95	83	89.7	71.8-98
Whole SVM 24v24	16	54	62	58	42.9-72.1
Whole LDA 24v24	22	66	75	70.5	55.6-82.8

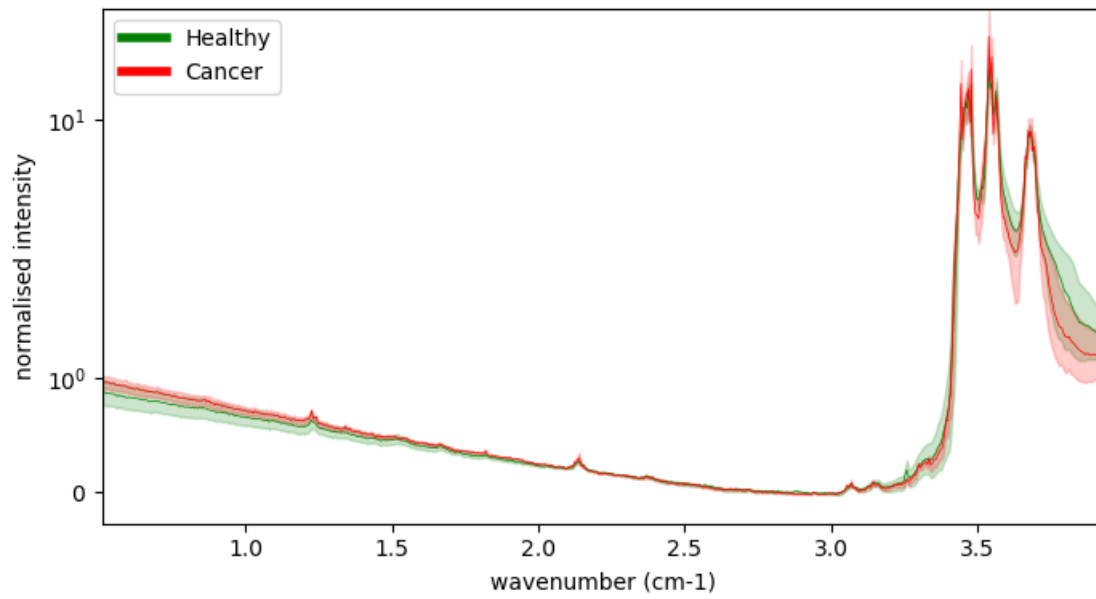


Figure A1.3: average NMR scan of blood plasma from 3 healthy and 3 pancreatic cancer patients. Y axis scale is logarithmic to show the small variations at low intensity values. Error is in faded colour around the line.

2. Additional investigation

2.1 2019 Raman and FTIR investigation on Buccal Mucosa Cancer

Introduction

This is a chapter contains the original plan and the results from the first research trip to gather Oral cancer patient data from the TATA research hospital in Mumbai, India.

Pre-study design

Going in to this study, from the literature search and analysis of the different methods employed, a preliminary standard methodology was developed for use. This is later refined into the 'utilised methods' section. The driving principle of this standard method is that it should be the easiest to transfer to a clinical setting. Striving for parameters and techniques that were simple, repeatable and effective. Once developed, further studies could reference the standards and if it needs to deviate from them, the reasons can be outlined. Adding why the deviation is an improvement for their purpose would enable greater universal understanding of the study's novelty. Indeed, the addition of ultrafiltration of the sample was the additional novelty used in this first study.

Anything highlighted in yellow was to be refined in testing:

Table 6.1: Standard recommendations	
FTIR	Raman
Patient selection	
<ul style="list-style-type: none">• Have a healthy control• Have a related infection(s) control	

<ul style="list-style-type: none"> Define patients based on biological gender and age groups, each group should have at least 5 patients If prognosis findings required compare between disease stages 	
Sample preparation	
<ul style="list-style-type: none"> use blood serum dilute in ratio 1:25 serum to ultrapure water deposit 500µl on CaF slide Leave to dry for 30mins. Anything better (drying conditions etc.)? 	<ul style="list-style-type: none"> use blood serum analyse wet analyse in aluminium well plate 20µl
Spectral acquisition	
<ul style="list-style-type: none"> use transmission mode FTIR Resolution of 4cm⁻¹ take spectra of both the fingerprint (700-1800cm⁻¹) and high wavenumber regions (2700-3500cm⁻¹) 	<ul style="list-style-type: none"> calibrate using a standard crystalline silicon peak at 520.7cm⁻¹ Use reflectance mode spontaneous Raman spectroscopy Use 785nm laser source Resolution of 4cm⁻¹ take spectra of both the fingerprint (700-1800cm⁻¹) and high wavenumber regions (2700-3500cm⁻¹) 20x objective

	<ul style="list-style-type: none"> • Approx. adjusted laser power (25mW?) • Acquisition time 10s
Spectral pre-processing	
<ul style="list-style-type: none"> • Use scaled background subtraction including the substrate • Use RMieS correction algorithm • normalise intensities using maximum normalisation • normalise to $\frac{1}{2}$ wavenumber integer intervals by interpolation 	<ul style="list-style-type: none"> • Eliminate cosmic rays by taking median acquisition from 3 repeat measurements. • Use scaled background subtraction • use rubberband baseline correction • calibrate intensity using a reference standard • normalise to $\frac{1}{2}$ wavenumber integer intervals by interpolation
Statistical analysis	
<ul style="list-style-type: none"> • Compare 2 groups at a time • Compare all groups (including age and gender) • Use PCA to maximise variance • Use SVM to separate the two groups and produce sensitivity and specificity values • Use k(5)fold cross validation to produce validate sensitivity and specificity values for each pairing • For prognosis, attempt to quantify any differences between disease stages 	
Post analysis	

- Examine which peaks are of greatest difference between groups and compare with spectral reference
- Examine literature for disease of interest, if there are blood composition studies compare to see if the spectral reference findings relate
- Suggest or perform a follow up study to ascertain serum compositional causes for the spectral shifts, identifying causal proteins etc. **If quantification of the spectral shifts is obtained, try relating this to causal molecule concentrations.**
- Conclude the usefulness of the findings for the purpose(s) proposed

For patient selection, a related disease control was selected as this would best emulate a practical diagnosis scenario where the disease of interest should be discernible from similar diseases. A study on ovarian cancer performed this control, effectively discerning cancer patients from other benign ovarian patients. An additional healthy control is also recommended as a reference however, and to potentially allow better quantification. Gender and age groups are recommended as often there are group related shifts in blood composition, defining them from the start will allow later comparison between these groups. Stage of the disease should also be taken into account, especially in cancer patients and where quantification is desirable as it will allow disease severity to be estimated at diagnosis and an expected prognosis to also be made.

Serum is recommended for both processes as most of the components of note are contained within and it removes the coagulation issues with plasma¹. Sample preparation for FTIR it is recommended to dilute the serum with 3 parts water to reduce cracking when drying and also to only use 1 μ l total amounts to reduce the coffee ring effect. The sample should be deposited on a CaF₂ slide for analysis as it has minimal spectral interference² and left to dry in ambient conditions, 30 minutes was

chosen as it leaves ample time for full drying³. With Raman, wet samples are recommended to avoid any potential drying issues and because wet analysis appears to have no spectral difference to well controlled dry⁴. An aluminium well plate should be sufficient as it should have minimal spectral interference². The precise volume is not of vital importance, though 20µl or more is recommended to ensure the sample is discerned from the aluminium.

For the spectral acquisition, 4cm⁻¹ resolution is standard among current studies and equipment and should be sufficient for discerning spectral biomarkers and analysis of both the fingerprint (<1800cm⁻¹) and high wavenumber (2700-3500cm⁻¹) region is recommended for the best characterisation of a sample over all useful data points, as both regions have been used to find biomarkers before^{1,5}. Transmission mode FTIR is recommended over ATR to avoid variable vertical deposition issues. In Raman, a 785nm laser is recommended for serum analysis as it is commonly available and offers a compromise between high CCD quantum efficiency and reduced fluorescence^{2,6}. Reflectance mode is recommended to get minimal substrate signal. Calibration should be standardised at 520.7cm⁻¹±0.2 with a silicon crystal⁷, for better comparison between studies.

For spectral pre-processing, an automated baselining method was deemed to be most suitable for reproducibility and 'rubber band' was a simple and effective one from the literature. A National Institute of Standards and Technology Raman reference standard was brought with the idea that the intensities could be normalised from it. FTIR would have to use maximum normalisation instead. The decision to standardise the data analysis to a ½ wavenumber standard via interpolation was to allow greater comparability between studies/acquisitions that had different ranges and/or

wavenumber values. This is especially important in Raman, which uses less discrete wavenumber values.

For the statistical analysis, it is recommended to compare 2 groups at a time using PCA to separate the data points and SVM to define the two groups with appropriate sensitivity and specificity values, and validate this with 5-fold cross validation. These methods were chosen as they are relatively well known and achievable with many analysis toolkits and were often highly, if not the most, effective in studies^{8,9}. It is recommended to compare all the relevant group combinations to discern any particular biases so that that can be accounted for in the produced biomarker. For example, if gender groups were significantly divergent than perhaps a spectral biomarker could be identified for each. Quantification is essential for monitoring and a decent patient prognosis, it is recommended to do this by comparing the shifts between principle peaks between the healthy and diseased spectra.

As for post analysis, it is recommended to compare to a spectral reference to begin to try to discern the root biological cause of the spectral shift. Knowing which features are present in higher concentrations can help to discover the molecules that contain them. Next, if there is relevant literature on blood composition in the disease of interest, see if any parallels can be drawn. Then, this information can be used to suggest or even perform a follow up study into the root cause of the spectral biomarker identified. Finally one can summarise the effectiveness of their biomarker from its sensitivity/specificity values and suggest how it would be best used in clinical practice, to ensure any reader can make the best transfer of the work in to the medical sphere.

Methods

Sample Preparation

In this study, certain factors that could influence the serum spectra such as age, sex, diet, certain habits e.g. smoking, pre-existing conditions or other diseases were controlled. Though little could be done to control diet in this particular experiment, efforts were made to eliminate or control for the other potentially obscuring factors.

The blood serum was collected from male patients from the Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Navi Mumbai, India. This study had two phases, the second used to validate the results of the first on a larger cohort of patients. In the initial phase, 16 had buccal mucosa cancers, 12 had premalignant oral conditions (leucoplakia) and 13 were healthy volunteers. Samples were stored at -80°C until being thawed for analysis.

The second round was performed on a larger cohort of 90 patients. Furthermore, the identification of the 10-30 kDa region as providing the best overall classification accuracy suggests that the molecular weight splitting can have significant value, especially if this particular region is exploited.

Patients: 28 oral cancer and premalignant patients as well as 17 healthy and 17 healthy tobacco users as an additional control. It should be noted that all the oral cancer and premalignant patients were also tobacco users in both parts of this study.

The serum was separated into two fractions using Millipore 500 μl 50 kDa centrifugal filters. The centrifuge was run for 20 minutes at 14000 g. Whole serum, <50 kDa low molecular weight (LMW) and >50 kDa high molecular weight (HMW) fractions were analysed. Molecular windowing, as outlined in 4.2 was also utilised on 9 cancer and 9 premalignant patients, using 50, 30, 10 and 3 kDa filters until 6 additional subsets of serum were produced.

For the FTIR measurement droplet deposition, as outlined in 4.1 was used to prepare the serums samples on CaF₂ discs.

For the Raman measurement each fraction was diluted in a 1:3 ratio before 1µl was deposited on a Crystran Raman grade CaF₂ slide and left to dry for 30 minutes.

Spectral acquisition

FTIR spectra were acquired with a Perkin Elmer 'Spectrum Two' FTIR spectrometer as outlined in 4.3.

Raman spectra were taken using a WiTec alpha 300 spectrometer with a 532 nm laser over the relative wavenumber range 0 to 3500 cm⁻¹. The Resolution was 2 cm⁻¹, an objective lens with a 10x magnification and 0.25 numerical aperture was used along with a 1200 g/mm grating. Spectra were taken at a laser power of 27 mW for 10 seconds with 3 accumulations. Scans were taken over a wavenumber range of -100 to 3500 cm⁻¹ to investigate the entire spectra for useful signal regions, this resulted in approximately 24 minute acquisition times. Prior to acquisition the spectrometer was calibrated using a silicon reference at 520 cm⁻¹. Spectra were acquired from the centre of the small, dried droplet.

Pre-processing of spectra

FTIR spectra were pre-processed by ALSS and average normalisation as outlined in 4.3

Raman spectra were pre-processed with a background correction using the wavelet transformation algorithm devised by Zhang et al.¹⁰, as this is more optimised for Raman.

Post-processing of spectra

Spectra were analysed by PCA-SVM as outlined in 4.4, though only the first 2 principle components were analysed. Leave-one-out cross validation was used, as described in 4.5.

Results and discussion

Phase 1: FTIR spectroscopy

Table 6.1: FTIR cross-validation sensitivity (Sens.) Specificity (Spec.) and accuracy (Acc.) results for classifying between Buccal Mucosa Cancer (C) samples from healthy (H) and premalignant (P), demonstrating increased accuracy when appending spectra together.

Sample	Cancer/Healthy			Cancer/Premalignant			Premalignant/Healthy			Average Accuracy
	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	
LMW	92%	94%	93%	95%	91%	93%	90%	100%	95%	94%
HMW	88%	83%	86%	69%	63%	66%	71%	90%	81%	77%
Whole	92%	83%	88%	90%	77%	84%	95%	93%	94%	88%
LMW and Whole	95%	97%	96%	100%	97%	99%	97%	96%	97%	97%
LMW and HMW	100%	88%	94%	85%	77%	81%	78%	90%	84%	86%
Whole and HMW	100%	83%	92%	92%	91%	92%	100%	100%	100%	94%
All 3	100%	97%	99%	97%	97%	97%	85%	100%	93%	96%

The processed FTIR spectra in Figure 2 show clear differences between the sample groups. The cross validated sensitivity and specificity results are summarised in Table 6.1. The separability of the groups is high all round though the HMW section is less effective. It appears that the concentration was higher than expected for some of the HMW samples and therefore the peak around 1650 cm^{-1} produced such high absorbance (>90%) that its value is distorted and cannot be valid for mathematical comparison.

Appending the spectra together increased the average accuracy in each case, though appending all three showed lower values than the LMW and Whole sets appended. This is likely due to the HMW distortion affecting the categorisation.

Looking at specific regions of deviation within the spectra (Figure 6.2) it appears that the peaks around 1650cm^{-1} and $2840\text{-}2970\text{cm}^{-1}$ show most consistent deviation between the groups of spectra. The 1650 cm^{-1} peak was higher in cancerous and premalignant patients than healthy in the LMW spectra. The peaks in the $2840\text{-}2970\text{cm}^{-1}$ region was lower for cancer patients than the premalignant or healthy patients in both the whole and LMW spectra. Healthy patients also show higher absorbance in the $1030\text{-}1120\text{cm}^{-1}$ region than the cancerous patients in the HMW and LMW spectra, but not the whole.

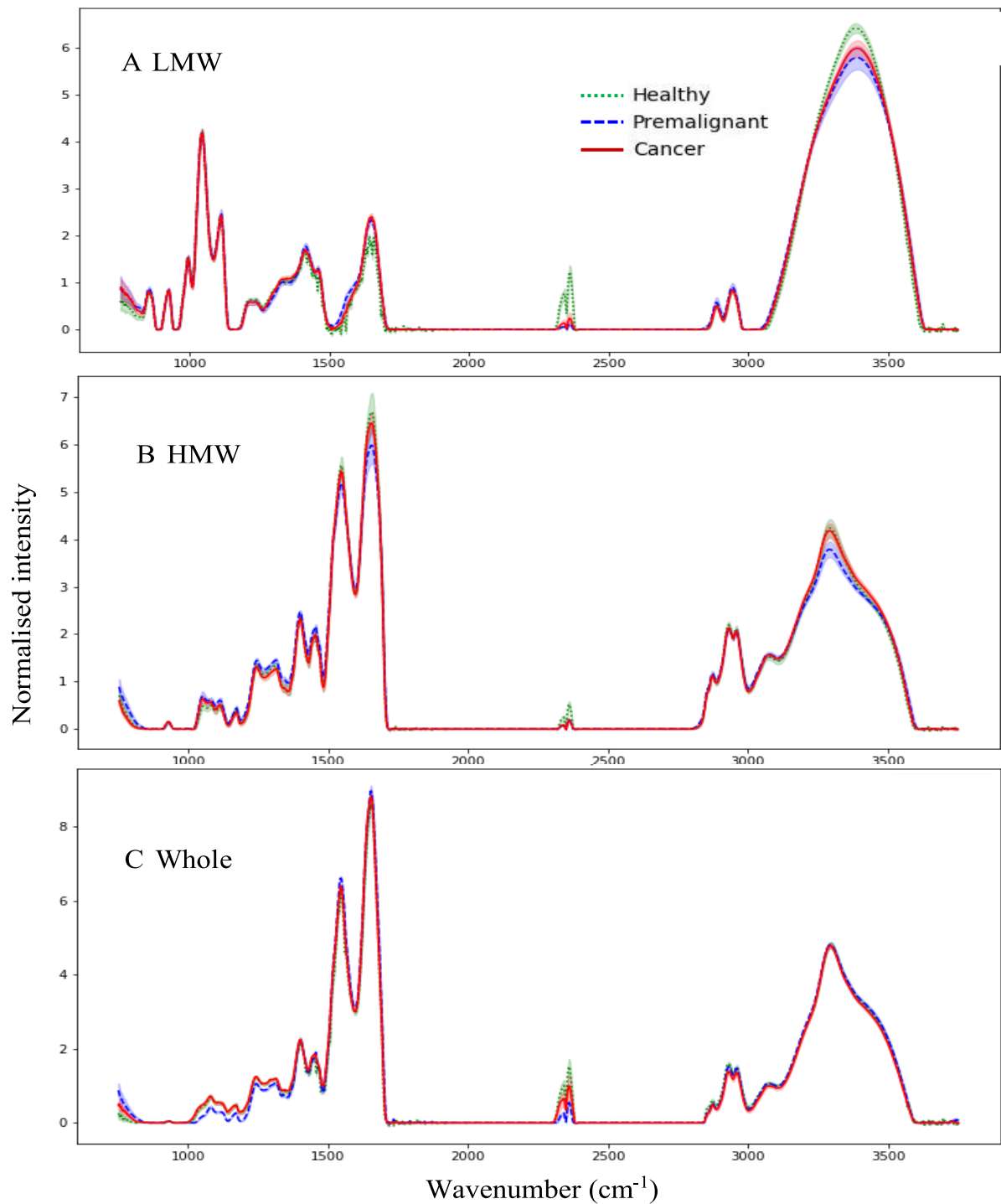


Figure 6.2: Average FTIR Spectra of the healthy, premalignant and cancer patients A) low molecular weight serum fraction. Peak of interest at 1650 and 2840-2970 cm⁻¹. B) High molecular weight. Peaks of interest at 1030-1120 cm⁻¹. C) Whole serum. Peaks of interest at 2840-2970 cm⁻¹. 1 σ error is minimal but can be seen in faded colour around the more varied sections of each spectra.

Table 6.3: Raman cross-validation accuracy results for classifying between Buccal Mucosa Cancer (C) samples from healthy (H) and premalignant (P), demonstrating increased accuracy when appending spectra together.

Sample	Cancer/Healthy			Cancer/Premalignant			Premalignant/Healthy			Average Accuracy
	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.	Acc.	
LMW	56%	38%	47%	62%	50%	56%	30%	41%	36%	46%
HMW	56%	61%	59%	93%	91%	92%	61%	66%	64%	71%
Whole	68%	46%	57%	81%	66%	74%	69%	75%	72%	68%
LMW and Whole	43%	46%	45%	75%	58%	67%	46%	41%	44%	52%
LMW and HMW	68%	61%	65%	87%	75%	81%	46%	50%	48%	65%
Whole and HMW	56%	38%	47%	87%	83%	85%	69%	58%	64%	65%
All 3	56%	53%	55%	81%	66%	74%	53%	41%	47%	58%

As is evident from Table 6.3 and Figure 6.3, the results from the Raman spectroscopic investigation showed less clear classification. The LMW subset providing the worst cross validation results - averaging 46% and meaning the model used is inferior to random selection. The HMW subset and whole serum were marginally better, averaging 71% and 68% respectively. Classification seemed to be highest between cancer and premalignant samples, and lowest between cancer and healthy.

Furthermore, the addition of appending the spectra did not serve to improve the classification in this case, only the LMW and HMW combination producing a higher average accuracy than the average from the individual spectra.

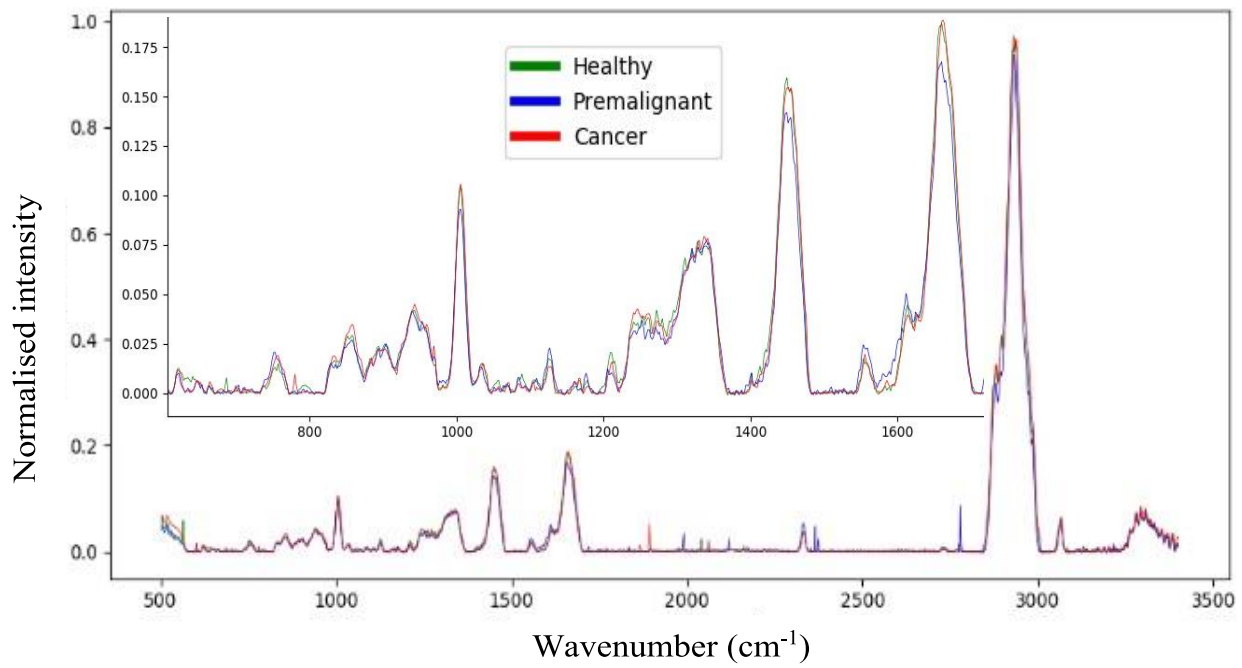


Figure 6.3: Average Raman Spectra of the healthy, premalignant and cancer patients for whole serum. No major visible deviations.

The spectra themselves contained a large amount of background noise. Several spectra had to be retaken to ensure at least a 10x noise to signal ratio of the key peaks, but the noise would still be significant to potentially obscure key minor peak variations or weaker peaks.

Discussion

In Raman spectroscopy, if thin or even slightly transparent samples are investigated then one has to choose a suitable substrate to hold the sample. Glass tends to have a high background fluorescence signal in Raman measurements, therefore Raman grade calcium fluoride was used as it only exhibits significant Raman peaks at $<600\text{cm}^{-1}$ for lasers excitations ranging between 473-830nm^{2,11}.

For patient selection, a premalignant control was selected as this would best emulate a practical diagnosis scenario where the disease of interest should be discernible from similar, non-malignant diseases. A study on ovarian cancer performed similar control, effectively discerning cancer patients from other benign ovarian patients¹⁶. An additional healthy control is also used as a reference and to potentially allow quantification of the cancer severity if patient outcomes are monitored. Gender was controlled by only male patients being selected for this initial study to help reduce unintentional bias.

The intelligent baseline correction algorithm²³ was chosen for its renown and effectiveness.

Average normalisation was used so that variation in all the peaks could be assessed. Linear PCA-SVM and complete 'leave-one-out' cross validation were chosen as the dataset was small enough for complete cross validation, linear SVM produced the best classification for our dataset, and SVM is highly, and sometimes the most, effective in studies^{8,9}. Only 2 PCA factors were used to reduce the risk of overfitting from the small sample set.

It is valuable to compare to a spectral reference to begin to try to discern the root biological cause of the spectral shifts observed. Knowing which features are present in higher concentrations can help to discover the molecules that contain them.

Furthermore, if there is relevant literature on blood composition in the disease of interest, any parallels between the data will provide more clues. This information can then be used to suggest or even perform a follow up study into the root cause of the spectral biomarker identified.

The accuracy in the Raman study, though lower than the FTIR data, is still comparable to the results obtained in the previous studies ^{21,22}.

Though exposure times were sometimes long for Raman, the laser intensity was confirmed to not be powerful enough to burn the sample, as clear sample damage was evident from a black burn mark as well as severe deviations in the spectra produced when tested using higher power and magnification.

Use of a 532 nm laser over a 785 nm seems to have produced a similar spectra to the literature²², however it is possible that the higher excitation wavelength would have been more suitable for biological samples as it offers a better compromise between high CCD quantum efficiency and reduced fluorescence^{25,30}.

The FTIR results demonstrated the ability to effectively distinguish between healthy, premalignant and cancerous serum samples with high accuracy. Additionally, the ability to distinguish the spectra from low and high molecular weight subsets of the serum was demonstrated. Appending the spectra together in analysis managed to improve the average accuracy of the cross validated categorisation only when classifications were similar in quality. This is demonstrated by the improvement of Whole and LMW serum in the first set improving to 97% from 94 and 88% classification accuracies. Though theory proposed that the key small molecules were being obscured by the larger proteins in the serum¹⁸ seemed to be validated by the first round of experiments, the second produced a less conclusive result. However, the classification is still present from a majorly different low molecular weight spectra, as it reduces the contribution of

albumin, globulin and other high weight components. Therefore, there is definitely valuable information to be gleaned from this subset. Furthermore, the identification of the 10-30 kDa region as providing the best overall classification accuracy suggests that the molecular weight splitting can have significant value, especially if this particular region is exploited.

For this Buccal Mucosa case, the peaks at 1030-1120 cm^{-1} , 1650 cm^{-1} , 2840-2970 cm^{-1} were noted as key variations. The 1030-1120 cm^{-1} region varies inconsistently between samples and subsets. It is close to the edge of the FTIRs detection range and consequently close to the edge of the baseline correction and therefore it is possible that some minor, but category consistent, fluctuations may have been amplified by errors in this. It perhaps deserves further investigation, but no valid conclusions can be drawn from this data alone.

The 1650 cm^{-1} peak is likely the amide 1 peak which is very strong in the HMW and whole spectra. The observed shifts between groups in the LMW spectra is likely obscured by the same peak from contributions from large molecules in the other spectra. However, this shift is consistent between the premalignant and cancer patients and therefore is likely to be the result of increases in inflammatory or similar general disease response molecules produced by the body to defend from potentially any illness, and not cancer specific. Further investigation of this peak in other diseases will help discern if it is tumour specific or just a simple inflammatory response.

The 2840-2970 cm^{-1} peak is most of note as it only reduces in the cancerous samples, meaning it may well be a cancer specific biomarker. It is also only seen in the LMW and whole sets, meaning it is likely to be a contribution from the <50 kDa molecules and therefore not present in the HMW set. There are several candidates for the peak's

molecular origins; C-H stretching from aldehydes and several other bonds, as well as N-H stretching from bonded quaternary amine salts^{12,13}

One additional potential source of error comes from the drying of the samples. The dilution steps were not sufficient to eliminate the 'coffee ring effect'. Although the location of acquisition was controlled, it is possible the results may be skewed by this.

Conclusions

The potential of FTIR for screening this disease is demonstrated by this study, as well as the additional use of ultra-filtration to further the categorisation accuracy and provide more information about the signal's origins. Though only the FTIR data showed improvement in classification with the addition of different subset spectra, if the signal to noise ratio and sampling inconsistencies from the Raman spectroscopic investigation can be improved then more of an effect could be observed.

The study would benefit from an additional related, but non tumorous, disease control to examine further the cancer specificity of the spectral biomarkers observed⁶.

The usefulness in making use of both Raman and FTIR together for screening cannot be realised until the Raman side is improved, though methods like SERS to improve sensitivity would be less viable for screening due to increased complexity and cost. It is likely that in large scale, multi-site testing prediction accuracies would decrease due to detection instrument discrepancies and the increased potential for errors and site-specific discrepancies that comes with a larger scale project. Other spectroscopy and general serum analysis methods, like nuclear magnetic resonance or mass spectrometry, for screening have been attempted with some success^{14,15}, adding one of those to the methodology could be considered instead of Raman to help improve detection.

However, understanding the chemistry behind these spectral biomarkers will bring more confidence in the machine learning outcome than just adding more abstract data. It is imperative to discern the root shifts in protein or other biological molecule concentrations to validate the spectral diagnosis method. The molecular windowing portion of this study will help with this, as it narrows down the scope of investigation to a key band of molecular weights.

To conclude, it was decided that more promise was present in the FTIR section of this study. The higher accuracy, coupled with higher signal to noise ratio, lower relative cost of the instrument and higher ease of use made it a clear choice for developing a new cost effective diagnosis method, leading to the work outlined in §6.2.

2.2 Biological spectral imaging

A brief investigation into imaging of the biological organism *D.magna*, with the goal of highlighting the eggs was undertaken. The *D.Magna* were suspended in carbonated water to arrest their movement and spectra were taken at designated x and y positions along the organism's body. The spectra were baselined and analysed by PCA to highlight any consistencies in the data. As demonstrated in figure 11.1, the regions attributed to high protein concentration were highlighted in the first principle component of the spectra, and the eggs were clearly visible. Another principle component highlighted separate regions, attributed to greater lipid concentration.

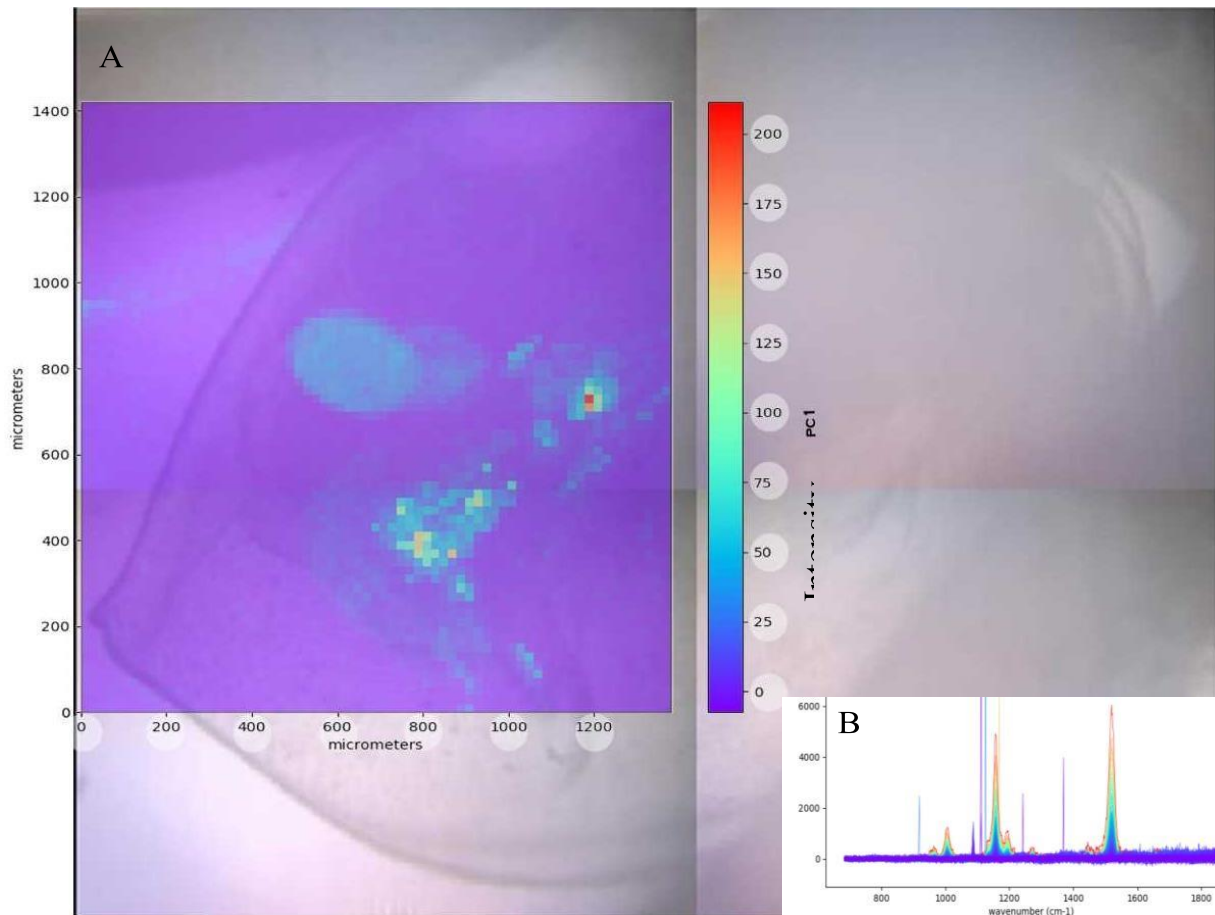


Figure A2.1: A) image of D.Magna egg and internal organs highlighted by PCA, the associated spectra of most interest evident from the corresponding colour. The high density of proteins/lipids in these areas seem to be highlighted by this imaging method B) the corresponding spectra from each pixel of the image, demonstrating what features are being targeted by PC1.

This discerning of different regions from a spectral image may prove a useful technique in later cell studies, and therefore having the framework ready to implement it could prove to be of great value.

2.3 Spectral differentiation and machine correction app development

This section is a collaborative project with Oliver Hart, a masters student from computer science¹⁶. It covers two topics, the first producing a prototype app to demonstrate the functionality of user friendly, cloud-based spectral diagnosis. The second part is an investigation to attempt at correcting for the problem of instrument to instrument variability in spectroscopy, which can significantly impact larger scale spectroscopic diagnosis projects. One piece of research that uses a similar approach of combining an application, cloud computing and machine learning of which we are aware is that of O'Toole¹⁷. This research used wholly simulated data in order to determine the system's feasibility. Our research will develop this by using actual spectroscopic data from patients' blood samples to bring this approach a step closer to reality, as well as to attempt to overcome the spectrometer variability issue.

Spectral differentiation app

A program was produced for baseline and background correcting spectra before differentiating them based on Support vector machine categorisation, a form of which was used in the Buccal Mucosa investigation. This was then adapted into an executable application, for any windows pc with an internet connection, which pulls from an online model made using data from that investigation. One can then select a .csv file in the app and have the classification of the sample (cancerous or healthy) returned, along with a percentage likelihood of the prediction based on the ratio of the sample's healthy or cancer categorisation factors. Its functionality is demonstrated in figure 11.4.

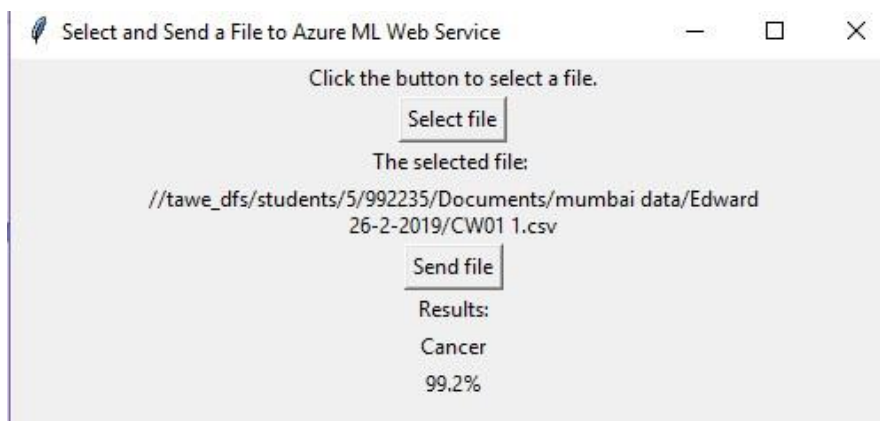


Figure A3.1: Demonstration of the application interface showing the predicted result and accuracy after a spectrum is selected and run automatically through the cloud model. The goal of this is to eventually have a system for quickly, easily and cost-effectively screening cancers in a hospital setting.

Instrument Difference Correction System

The next part of this project was to attempt to address the machine to machine variability of results from the same spectral analysis that had been observed in many studies^{1,18}. For usability of spectral diagnosis it is vital that blood sample data can be measured using any previously unknown spectrometer and accurately classified using the trained machine learning model. The approach to resolving this was tackled in two ways. The first is to control and account for as many physical and chemical factors that could be affecting the instrument's reading - as is reasonably feasible in a reproducible and cost-effective manner. The second was to attempt to account for other spurious factors from each specific machine by producing a model, based on spectra taken from several FTIR machines of several distinct compounds, that can correct for the variations.

In order to correct for these intrinsic spectrometer differences, machine learning algorithms are a necessary tool. Various chemicals are readily available worldwide in pure form and that, when measured using a spectrometer, they should all yield exactly

the same set of measurements. This will act as a set of known standards. Thus, a particular spectrometer's measurements of a chemical can be compared to the ideal measurements and from these discern what intrinsic measurement effects are present in that spectrometer.

Data from seven different chemical molecules was obtained: citric acid, glycine, potassium phosphate, sodium bicarbonate, sodium sulphate and urea. Dummy variables were assigned to each of these molecules as labels. Each of these samples were measured by the three distinct spectrometers to produce our testing dataset. From the spectra obtained, one spectrometer was assigned as the model and attempt to correct the spectra obtained from the other two to match it. Raw data was used for the correction, without the typical pre-processing steps of baseline correction or normalisation.

Before working on the real dataset of molecules first a proof of concept system was built to attempt to correct for the three different possible measurement effects that may be present in spectrometer readings. Simulated data was used for each individual effect, testing our system one them one at a time, before adding multiple effects and investigating if it remained effective.

The first effect is intensity scaling, where there is an increase or decrease in intensity of the peaks. This can be an overall shift, a background curve, or a non-linear shift dependant on the intensity of the signal. For example, an exponential decrease at high intensities. The second is a horizontal offset. This is when the wavenumber values for the spectrum are shifted, resulting in incorrect positions for the peaks. The final potential effect is the most complex. Peak broadening involves certain peaks on the spectrum being wider than they should be. This can be due to lack of resolution of the spectrometer or other errors in the instrument¹⁹.

To start, spectra of the six molecules was taken by one of the spectrometers to use as our target ideal data. These values are used as the labels for the machine learning model. To simulate intensity scaling, ideal data was taken and linear shift was added. When that could be corrected for effectively, then the ideal data with a non-linear parabolic shift added to it was tested. These simulations were achieved using for-loops in Python.

For the horizontal offset, the input data was simulated using a simple for-loop by copying the ideal data and then using NumPy's roll function to add an offset. Such effects can be in the region of a couple of wavenumbers, so an offset of three wavenumbers was used.

The peak broadening effect was simulated by convolving the data with a Gaussian. This was achieved by using SciPy's `gaussian_filter1d` function. An appropriate value for the standard deviation of the Gaussian is ten, based on observations of the graph of the transformed data and the original data.

Data from five of the molecules was used for training our model and one for testing it. The model was supplied with simulated intensity data from the five molecules, along with the corresponding wavenumber for each intensity and a dummy variable distinguishing the molecules. The labels are the pre-transformed data. Then the trained model was used to make predictions on the simulated data for the final molecule and compare these predictions to the actual values in order to assess the model's accuracy. Objective methods, of the root-mean-square (RMS) error and mean absolute (MA) error of the differences between the predictions and the actual values, were to evaluate the effectiveness of our model, with the aim for the differences to be as small as possible. Several different types of model were tested in order to determine which is most effective for the task. These included Support vector regression, AdaBoost regression, Gradient boosting regression, Decision forest regression and neural network regression.

Finally, after correcting for the three simulated effects, similar approach was used to try to correct the real molecule data from different spectrometers, choosing one of the instruments to be the target ideal.

Results

The best results from each different machine learning algorithm on the data simulated with all three possible effects are presented in the table below:

Table A3.1: The best results from the five different machine learning models on the simulated data that were tested on the urea data (simulated with a Gaussian filter applied using a standard deviation of 10, a parabolic shift applied to the intensities and a horizontal offset of 3 wavenumbers)

	Support Vector Regression	AdaBoost Regression	Gradient Boosting Regression	Decision Forest Regression	Neural Network Regression
RMS error	0.0929	1.1470	0.5670	1.1602	0.1702
MA error	0.0662	0.7297	0.4451	0.9047	0.1250

Table A3.1 displays the RMS error and MA error between the baseline corrected measurements of the same molecule by different spectrometers, i.e. the upper bound on the acceptable size of errors from our model, are as follows:

RMS error: 0.0952

MA error: 0.0719

Even though the results are from the most effective machine learning model, the support vector regression, they are only just below the upper bound on the acceptable size of

errors shown above. Finally, the results of the support vector regression are reported using the actual molecule data, despite the data being noisier and hence more challenging to correct than the simulated data:

Table A3.2: The best results from the support vector regression on the actual molecule data, tested on the urea data

	Support Vector Regression
RMS error	0.1148
MA error	0.0768

Table A3.2 displays the RMS error and MA error between the baseline corrected measurements of the same urea molecule by the two spectrometers used to collect the actual data we used. The upper bounds on the acceptable size of errors from our model when using the actual molecule data are as follows:

RMS error: 0.0686

MA error: 0.0545

Discussion

From the results in Table A3.1, it is clear that the support vector regression was the most effective of the different algorithms at making the corrections. The neural network regression was also relatively effective, while the other models can be discarded.

It is also clear that the instrument correction system is not yet ready to be incorporated into the cloud diagnosis application. The simulated data was promising, the best machine learning model made improvements, with prediction errors being slightly below the real errors between the two spectrometers. However, using the actual molecule data produced prediction errors that were above the real errors between

spectrometers. Therefore, no improvement was achieved using the real data in our model.

Though the model is not yet sufficient, this experiment has aided understanding in several ways, providing key insight into how to progress in the future. It is possible that increasing the number of molecules and instruments available for testing could potentially improve these results. Additionally, changing which molecule was left out during the model production could also have affected the findings.

Future research should focus on support vector regression and, especially when more input data is available, neural network regression. Secondly the errors on the data that was simulated with only the peak broadening effect show that the support vector regression was not able to correct for that effect. The neural network regression was also unable to correct for it but fared better than the support vector regression (Table A3.3).

Table A3.3: The errors produced by the support vector regression and the neural network regression on the data that was simulated with only the peak broadening effect applied. Errors using the actual data values provided for comparison.

	Support Vector Regression	Neural Network Regression	Using Actual Values
RMS error	0.0542	0.0192	0.0190
MA error	0.0416	0.0113	0.0113

Furthermore, only one kind of neural network, with limited parameters, was so far attempted. Further research should consider different types of neural networks (such as recurrent neural networks, since the intensity value at a particular wavenumber is related to the value at previous wavenumbers) and assess varying more parameters (such as including more hidden layers). This, combined with more input data, could prove a more effective way of handling the instrument variation problem.

Another area for potential future research, especially if using instruments of varied design, is to provide more input features into the regression models. As all spectrometers used were of uniform design, this has been unnecessary. Variables like path length and resolution should be included when they are not uniform.

3. References

- 1 Baker, M. J. *et al.* Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chem Soc Rev* **45**, 1803-1818 (2016). <https://doi.org:10.1039/c5cs00585j>
- 2 Kerr, L. T., Byrne, H. J. & Hennelly, B. M. Optimal choice of sample substrate and laser wavelength for Raman spectroscopic analysis of biological specimen. *Analytical Methods* **7**, 5041-5052 (2015). <https://doi.org:10.1039/c5ay00327j>
- 3 Hughes, C. *et al.* Assessing the challenges of Fourier transform infrared spectroscopic analysis of blood serum. *Journal of Biophotonics* **7**, 180-188 (2014). <https://doi.org:10.1002/jbio.201300167>
- 4 Depciuch, J. & Parlinska-Wojtan, M. Comparing dried and liquid blood serum samples of depressed patients: An analysis by Raman and infrared spectroscopy methods. *J Pharm Biomed Anal* **150**, 80-86 (2018). <https://doi.org:10.1016/j.jpba.2017.11.074>
- 5 Diem, M. Comments on recent reports on infrared spectral detection of disease markers in blood components. *Journal of Biophotonics* **11**, e201800064 (2018). <https://doi.org:10.1002/jbio.201800064>
- 6 Pence, I. & Mahadevan-Jansen, A. Clinical instrumentation and applications of Raman spectroscopy. *Chem Soc Rev* **45**, 1958-1979 (2016). <https://doi.org:10.1039/c5cs00581g>
- 7 Meng, X. C. & Zhu, L. Q. Fast Testing Methods on the Performance of Laser Confocal Micro-Raman Spectroscopy System for Cell Analysis. *Adv Mater Res-Switz* **884-885**, 570-573 (2014). <https://doi.org:10.4028/www.scientific.net/AMR.884-885.570>
- 8 Sohail, A. *et al.* Analysis of hepatitis C infection using Raman spectroscopy and proximity based classification in the transformed domain. *Biomed Opt Express* **9**, 2041-2055 (2018). <https://doi.org:10.1364/BOE.9.002041>
- 9 Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G. & Mechelli, A. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev* **36**, 1140-1152 (2012). <https://doi.org:10.1016/j.neubiorev.2012.01.004>
- 10 Zhang, Z.-M. *et al.* An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *Journal of Raman Spectroscopy* **41**, 659-669 (2009). <https://doi.org:10.1002/jrs.2500>
- 11 Crystran. (2013).
- 12 Socrates, G. *Infrared and Raman Characteristic Group Frequencies*. 3rd edn, (JOHN WILEY & SONS, LTD, 2001).
- 13 Aldrich, S. *IR Spectrum Table*, 2019).
- 14 Krafft, C. Modern trends in biophotonics for clinical diagnosis and therapy to solve unmet clinical needs. *Journal of Biophotonics* **9**, 1362-1375 (2016). <https://doi.org:10.1002/jbio.201600290>
- 15 Song, Z., Wang, H., Yin, X., Deng, P. & Jiang, W. Application of NMR metabolomics to search for human disease biomarkers in blood. *Clin Chem Lab Med* **57**, 417-441 (2019). <https://doi.org:10.1515/cclm-2018-0380>
- 16 Hart, O. *Improving Early Diagnosis of Cancer Using Cloud Computing* Master of Science thesis, Swansea University, (2019).
- 17 O'Toole, M. *Improving Early Diagnosis of Cancer Using Raman Spectroscopy and Cloud Computing* MSc. thesis, Swansea University, (2018).

- 18 Leal, L. B., Nogueira, M. S., Canevari, R. A. & Carvalho, L. Vibration spectroscopy and body biofluids: Literature review for clinical applications. *Photodiagnosis Photodyn Ther* **24**, 237-244 (2018). <https://doi.org/10.1016/j.pdpdt.2018.09.008>
- 19 Coleman, M. D. & Gardiner, T. D. Sensitivity of model-based quantitative FTIR to instrumental and spectroscopic database error sources. *Vibrational Spectroscopy* **51**, 177-183 (2009). <https://doi.org/https://doi.org/10.1016/j.vibspec.2009.04.005>