Machine Learning Applied to Raman Spectroscopy to Classify Cancers

Nathan Blake

Academic Supervisors:

Prof. Geraint Thomas

Prof. Lewis Griffin

Industrial Supervisor:

Ian Bell

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Cell and Developmental Biology

University College London

August 2, 2023

I, Nathan Blake, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Cancer diagnosis is notoriously difficult, evident in the inter-rater variability between histopathologists classifying cancerous sub-types. Although there are many cancer pathologies, they have in common that earlier diagnosis would maximise treatment potential. To reduce this variability and expedite diagnosis, there has been a drive to arm histopathologists with additional tools. One such tool is Raman spectroscopy, which has demonstrated potential in distinguishing between various cancer types. However, Raman data has high dimensionality and often contains artefacts and together with challenges inherent to medical data, classification attempts can be frustrated. Deep learning has recently emerged with the promise of unlocking many complex datasets, but it is not clear how this modelling paradigm can best exploit Raman data for cancer diagnosis.

Three Raman oncology datasets (from ovarian, colonic and oesophageal tissue) were used to examine various methodological challenges to machine learning applied to Raman data, in conjunction with a thorough review of the recent literature. The performance of each dataset is assessed with two traditional and one deep learning models. A technique is then applied to the deep learning model to aid interpretability and relate biochemical antecedents to disease classes. In addition, a clinical problem for each dataset was addressed, including the transferability of models developed using multi-centre Raman data taken different on spectrometers of the same make.

Many subtleties of data processing were found to be important to the realistic assessment of a machine learning models. In particular, appropriate cross-validation during hyperparameter selection, splitting data into training and test sets according to the inherent structure of biomedical data and addressing the number of samples

per disease class are all found to be important factors. Additionally, it was found that instrument correction was not needed to ensure system transferability if Raman data is collected with a common protocol on spectrometers of the same make.

Impact Statement

The findings in this thesis can be split into two categories: those related to the methodological rigour of conducting machine learning for cancer diagnosis using Raman spectroscopy, and clinical findings regarding the application of the technique to specific oncology tasks identified by collaborating histopathologists.

Regarding methodological rigour, this thesis has confirmed findings in the literature that without nested cross-validation the accuracy of models can be inflated by 5-10%, by over-fitting model hyperparameters. Similarly, the method of splitting data during model training was found to significantly impact the estimated accuracy of models, with inappropriate methods inflating accuracy by as much as 10-20%. It also finds that baseline correction during pre-processing does not necessarily increase the performance of models, and may even obscure clinically relevant information and complicate cross-validation by introducing more hyperparameters. Together with a systematic review of the related recent literature this thesis contributes to a growing movement within the medical Raman community to improve methodological rigour.

There are three main clinical findings.The first is that the technique can be used to distinguish between ovarian samples with disparate surgical outcomes. Although the results are not outstanding, they are competitive with genetic techniques and data from this thesis will be used in a grant application with the collaborating histopathologist to further investigate the applicability of Raman spectroscopy to this clinical problem.

Other findings in this thesis have demonstrated that, in principle, microsatellite instability can be detected in colon samples by Raman spectroscopy, and that potential biochemical antecedents can be identified. Despite a low sample size, these findings

are competitive with current screening tools. These results are in the process of being published. This paves the way for larger studies to confirm these findings, and explore the potential of the technique to compete with existing diagnostic tools.

Finally, findings in this thesis confirm that a model developed on data collected from one spectrometer can be applied to data taken on another spectrometer of the same make, so long as a common protocol is followed. System transferability is a particular hurdle for any clinical applications, and these findings show one possible pathway to providing clinical consistency. These results are likely to be of interest to spectroscopists and clinicians alike and are being prepared for publication.

Finally, this paper establishes that deep learning models can be competitive with traditional machine learning models in this domain despite low sample sizes.

Author Contribution Statement

This thesis is the confluence of multiple projects, necessarily including a large number of people and institutions. This statement will make explicit my contributions and acknowledge the contributions of others.

For the ovarian tissues, the clinical problem and FFPE samples were provided by Dr Florian Heitz (Kliniken Essen Mitte, Germany). Under the supervision of Riana Gaifulina I cut and mounted these samples and performed Raman spectroscopy upon them. I performed all subsequent data analysis.

For the colon samples Dr Manuel Rodriguez-Justo identified the clinical problem and the FFPE samples were obtained from the UCL biobank by Riana Gaifulina, who also obtained the Raman data. All analyses displayed in this thesis are my own.

The oesophageal samples are from a multi-centre study, involving a large number of collaborators including UCL, University of Exeter, Gloucester Royal Hospital and Renishaw. I had no involvement in protocol development or data acquisition. All pre-processing and data analysis contained within this thesis had been conducted by myself, with the exception of a subset of the SMART data in section 6.4.3, in which Ian Bell performed instrument correction.

The literature reviews of section 1.2 and 1.5 are my own work, including the meta-analysis, although the work received supervisory input during preparation for publication.

Acknowledgements

Many thanks to my academic supervisors Geraint Thomas and Lewis Griffin for their many conversations and insights into this project, science and life in general, as well as my industry supervisor Ian Bell who I am sure has gone far beyond any contractual obligations. You have variously been a compass, a map and voice on the wind. Special thanks to James Nelson, who started me on this journey before sailing on to his own sunset.

Thanks to my fellow lab members, particularly Lewis Tanner and Zhang Liyuan for keeping me company during different stretches of the journey, but mostly Riana Gaifulina without whom I would likely have ended up lost in the wilderness at the first turn. Thanks also to Candice Hermant, Chiara Zambianchi and Kate Ridley for allowing me to guide them along the few paths I know.

Thanks to Manuel Rodriguez-Justo, Chris Barnes and Xue Jinghao for acting as sentinels at different landmarks, ensuring I hadn't drifted too far from a traversable path.

Thanks to Michael Duchen, Yoshiyuki Yamamoto and Gemma Ludbrook who made a particularly jagged precipice easier to climb, allowing me to focus my energies on more immediate priorities at the time. Thanks to my wife, Amy, who survived a nasty fall on the trail in more ways than one. Patches has overcome.

To everyone: it turns out there was a far shorter path to the destination, but then I wouldn't have met you and I would be so much the poorer.

Contents

List of Figures

List of Tables

Chapter 1

Background

> *The history of science is rich in example of the fruitfulness of bringing two sets of techniques, two sets of ideas, developed in separate contexts for the pursuit of new truth, into touch with one another.*

J. Robert Oppenheimer

1.1 Clinical need for improved cancer diagnostics

In 2020 there were an estimated 19.3 million cancer diagnoses and 10 million cancer deaths globally [1]. The number of new cases in 2040 is expected to increase to 28.4 million, led primarily by low income countries increasing life expectancies [1]. The mortality rate will depend upon diagnostic and treatment regimes. Early diagnosis is important as the stage of disease can have a significant impact upon survival, though this is dependent upon the cancer [2]. For instance, in the UK, the one-year survival rate for stage one colorectal cancer (CRC) is 97.7%, falling to 43.9% at stage four. The corresponding rates for prostate cancer are ~100% falling to 87.6% [3].

Notable cancers include ovarian cancer which is particularly sensitive to early detection with a 5-year survival rate of 70% for stage 2 dropping to 20% for stages 3-4, in which the disease has spread beyond the pelvis [4]. Only 20% of such

cancers are diagnosed before stage 3. Unfortunately, the early detection of ovarian cancer remains elusive and a multimodal approach utilising several biomarkers combined with properly calibrated algorithmic models has been advocated [4]. Similarly, oesophageal cancer survival greatly benefits from early detection, with 5-year survival rates of 75-90%. However, most oesophageal cancers are detected at a late stage when the 5-year survival rate is less than 20% [5]. This is compounded by questions regarding the accuracy of endoscopy used for early diagnosis [6]. Improvements in colorectal cancer survival are in large part due to large screening efforts in Western countries, in which adenomatous polyps are identified and removed during endoscopy and are then assessed with standard histological techniques [7]. The success of such screening programs demonstrates the strength of early diagnosis. Indeed, its importance has been recognised by the UK government with targets that by 2028 75% of cancers will be diagnosed at an early stage (stage one or two) and for 55,000 more people to survive cancer for five years or longer [8].

In addition to the imperative to expedite cancer diagnosis, there is also a need to improve the accuracy of diagnoses. An element of subjectivity has been noted in many traditional cancer diagnosis pipelines [9, 10, 11]. To illustrate why cancer diagnosis is so fraught with subjectivity, we first need to understand the diagnosis pathway. This will vary between cancers, but for solid tumours it will be broadly similar. The CRC pathway can be used to illustrate the diagnosis pathway and quantify the degree of variability.

1.1.1 Colorectal cancer diagnosis

The vast majority of CRC is diagnosed by endoscopic biopsy or polypectomy [12]. A provisional diagnosis will be made by the endoscopist who will then remove any suspicious lesions and send them for histopathology. These samples will be processed soon after being removed from the patient. This begins with fixation which prevents autolysis and bacterial attack of the sample. A ubiquitous fixative is formalin which contains formaldehyde and acts by cross-linking proteins. The

sample is then embedded, often in paraffin wax, to preserve tissue morphology and to give the tissue support during sectioning, where thin slices of the sample are taken. This process is known as formalin fixation and paraffin embedding (FFPE). The slices are then mounted onto a slide and stained with appropriate dyes, typically haematoxylin and eosin (H&E), ready for inspection under a microscope. The histopathologist will distinguish various morphological features to determine whether any changes are benign or malignant [13]. For instance, the extent of any villous component and the degree of dysplasia will be assessed. A villous component is a leaf-like projection lined by dysplastic glandular epithelium and can be categorised as villous, tubulovillous or tubular depending on the architecture of the sample. Dysplasia refers to the degree of cell differentiation; i.e how identifiable the cells are. For instance the NHS Bowel Cancer Screening Programme (BCSP) defines a well differentiated adenocarcinoma as >95% of a tumour forming a gland, with moderately differentiated samples 50-95% and poorly differentiated <50% [12]. Both have been found to be independent predictors of advanced neoplasms [14]. However, the use of such biomarkers in guiding patient management is contentious with some pathologists arguing that the assessment of villous components and dysplasia is too subjective to be clinically reliable [13]. Protocols have been developed to ameliorate the subjectivity inherent in this method. In particular two-tier systems to categorise dysplasia, in which well and moderately differentiated samples are classified as low grade while poorly differentiated samples are classed as high grade, have been shown to improve inter-rater agreement [15, 16, 17]. Attempts to standardise approaches across international borders have also been made, most notably the revised Vienna Classification system. Despite such attempts subjectivity remains a prominent threat to objective patient management decisions. There are many published studies quantifying the degree of subjectivity, but this evidence has yet to be statistically synthesised to establish a thorough understanding of the inter and intra-rater variability in pathologists' assessment of colorectal samples. To this end a systematic literature review and meta-analysis was performed to establish the current state of knowledge.

1.2 Inter-rater variability in CRC diagnosis

1.2.1 Systematic review and meta-analysis

A meta-analysis conducted on studies collected by a systematic review is considered one of the best forms of evidence available in healthcare [18]. A systematic review itself involves identifying and retrieving published studies in a systematic and transparent manner such that all the available evidence on a subject is collected in a reproducible manner. This literature is then assessed for methodological quality to ensure biased studies are not included. The remaining studies are then assessed for their clinical, methodological and statistical techniques and if they are similar enough can be statistically combined to find a more accurate effect estimate with far narrower confidence intervals than any single study. To this end a systematic search and meta-analysis was performed with the intention of quantifying the variability of histopathological diagnosis of potentially cancerous colon samples. Ideally this would be assessed by comparing diagnoses with the known truth, generating, for instance, sensitivity and specificity statistics. Unfortunately, pathology is the gold standard for diagnosis and so is as close to the ground truth as current practices allow. However, even a cursory search of the literature reveals that histopathologists will disagree with each other and even themselves, which suggests there is some degree of error to this gold standard. This inter and intra-rater variability serves as a proxy marker for the accuracy of histopathological diagnosis.

1.2.2 Methods

Medline and EMBASE were searched for English language articles from 1980 to 12th December 2017. Identifying diagnostic accuracy studies proves more difficult than identifying randomised trials due to the inconsistent use of keyterms [19]. Therefore broad keywords were selected based on clinical expertise, known literature

and published recommendations for searching medical databases [20, 21, 22]. The following keywords were combined with boolean the operators AND/OR: colorectal$ OR colon$ OR CRC AND inter$ OR intra$ OR variability OR reliability, where "$" allows for any subsequent character. An initial yield of 48701 articles was refined by restricting keyword searches to the title. 75 abstracts were retrieved for closer scrutiny of which 27 warranted full text screening. The reference lists of these studies were searched for additional studies that were potentially relevent. The retrieved studies were included in the review if they sought to evaluate the reliability of any histological outcome of potentially cancerous or pre-cancerous colorectal samples as determined by two or more histopathologists. The studies were then assessed for quality using the The Quality Appraisal of Reliability Studies (QAREL) guidelines [23]; explicitly developed to assess the rigour of diagnostic reliability reports. Studies focusing exclusively on assessing serrated polyps were excluded for two reasons. At the time of conducting this review, there was a lack of consensus on their definition, nomenclature and pathogenesis amongst pathologists [24] which may skew results. Also, serrated polyps were thought a relatively rare occurrence, with a prevalence of approximately 0.1% [25]. Kappa statistics, used to quantify inter and intra-rater variability, are particularly sensitive to rare occurrences and may return low values even when there is a high proportion of agreement among observers. However, evidence emerging after the completion of this review suggests that serrated polyps are under-diagnosed and their role in CRC pathogenesis under-recognised [26].

The data extracted from each identified study included the year and country of the study, the number of samples assessed, the number of pathologists and their expertise, the reported kappa statistics and their standard errors, the histological outcome assessed and what guidelines, if any, were followed. For any meta-analysis it is important that the studies are comparing similar traits, therefore for data synthesis only results of similar outcomes were extracted. These were the inter-observer variability of the determination of hyperplastic vs adenomatous samples, polyp architecture type, degree of dysplasia and completeness of excision (henceforth these will be referred to collectively as 'outcomes'). However, there were insufficient

studies looking at this latter outcome so these too were not included in the synthesis. Many studies reported pairwise kappa statistics between pathologists. These were combined by each study under a fixed-effects model.

Kappa Statistics

Cohen's Kappa statistic, κ, is a common measure of inter-rater variability in healthcare assessments. It measures the degree of agreement between two raters on a particular outcome above that which would be expected by chance alone. It is given by

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1.1}$$

where p_o is the proportion of observed agreement between two raters and p_e is the proportion of agreement expected by chance alone (similar to chi-squared contingency tables). If there is no agreement between raters then $\kappa = 0$ while perfect agreement is given by $\kappa = 1$ and perfect disagreement by $\kappa = -1$.

There are extensions to Cohen's kappa to include the agreement of two or more raters and to give more weight to disagreements separated by more than one category. For instance, the classification of mild vs severe dysplasia could be considered more of a disagreement than mild vs moderate dysplasia.

1.2.3 Results

Appendix A summarises the studies identified for synthesis [17, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36]. Four studies were not included as they did not report or contain sufficient information to construct standard errors and the authors did not reply to requests for this data [37, 38, 39, 40].

An overall kappa of 0.52 (95% CI 0.45 - 0.59) was observed under a fixed-effects model. However, this was with a significant degree of heterogeneity ($Q = 1992.0, p <$

$0.0001, I^2 = 97.4\%$), indicating that the studies were too dissimilar to meaningfully combine their statistical estimates, so a mixed-effects model was preferred with the outcomes, speciality status of pathologists and whether any guidelines were followed as moderators. Only three studies used weighted κ estimates [27, 39, 40]. Pooling weighted and unweighted κ estimates is a potential statistical source of heterogeneity, therefore unweighted estimates were used even when the potentially more realistic weighted estimates were available.

Outcomes

Figure 1.1 shows a forest plot divided by the type of outcome assessed. The subgroups have less heterogeneity than the full model, suggesting some between-study variance is attributable to exactly what outcome pathologists are assessing in a sample. The plot suggests that pathologists often agree with one another when assessing whether a sample is hyperplastic or adenomatous κ =0.80 (95% CI 0.68 - 0.93). However, they are not so consistent when assessing the architectural type or grade of dysplasia of a sample with κ scores of 0.46 (95% CI 0.40 - 0.52) and 0.45 (95% CI 0.38 - 0.52) respectively. Though reduced there remains a considerable amount of heterogeneity between studies, particularly the grade and type. Likely sources include the fact that different numbers of categories were used in some of the studies (e.g. grade ranges from 2-5 categories).

Specialist Assessment

Figure 1.2 shows a forest plot divided by whether an assessment was undertaken by a Gastrointestinal (GI) specialist pathologist or by a non-specialist (or a specialist from another field). The synthesis shows that while there is a slight improvement between agreement among specialists compared to non-specialists with κ values of 0.56 (95% CI 0.47 - 0.65) and 0.47 (95% CI 0.37 - 0.57) respectively, this difference is not statistically significant, evident in overlapping confidence intervals

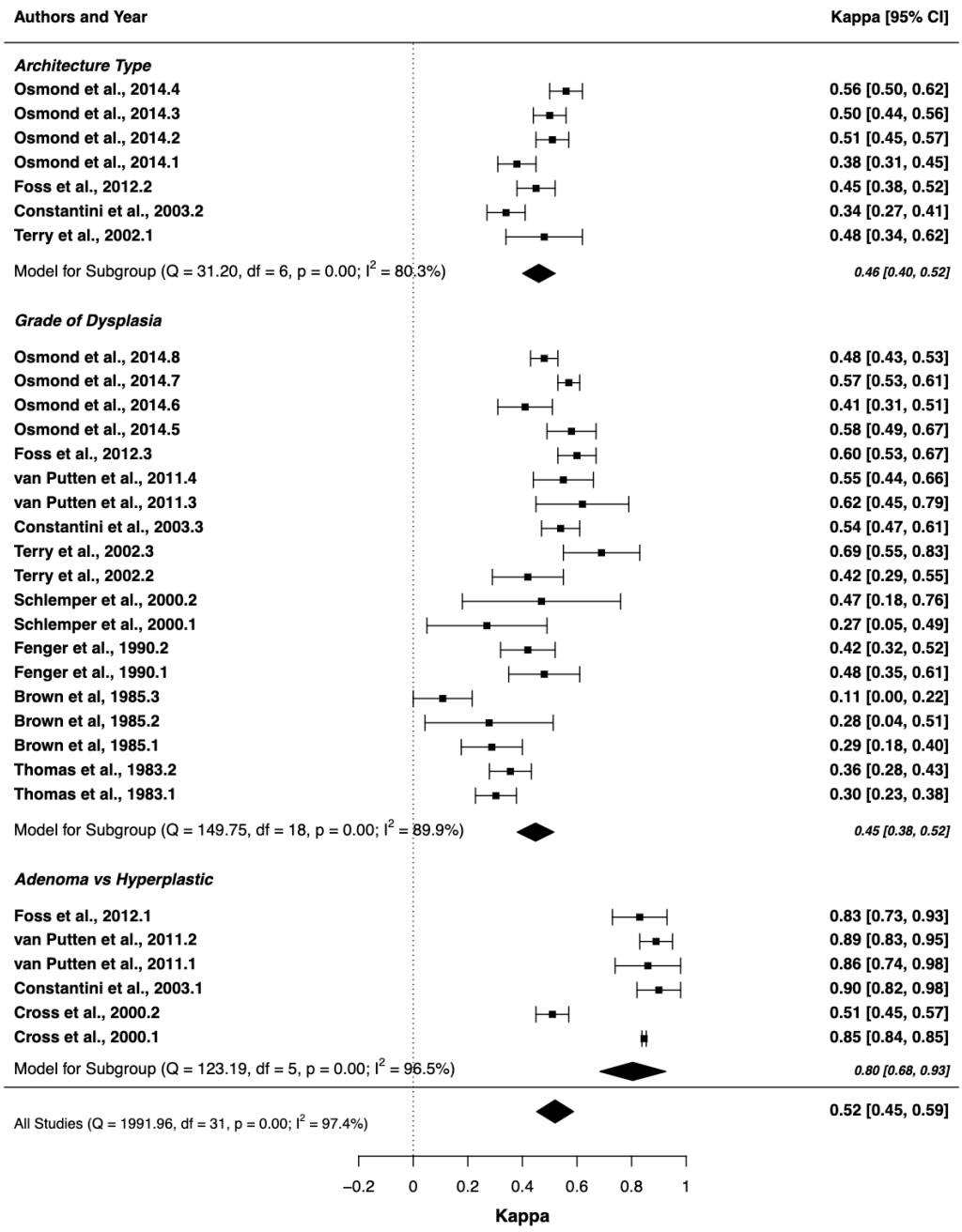


Figure 1.1: Forest plot divided by the type of outcome assessed, distinguishing: architecture type, grade of dysplasia or adenomatous vs hyperplastic samples

of the κ estimates. However, there is still a significant degree of heterogeneity within the reduced models which is obscuring interpretation. Some of this residual heterogeneity may be due to differences between GI specialists: varying years of experience, different definitions of a specialist in different countries and some

specialists being particularly interested with the assessment of colorectal polyps while others may focus on the upper GI tract. While this heterogeneity persists it is not possible to conclude that GI specialists agree with one another more often than non-specialists.

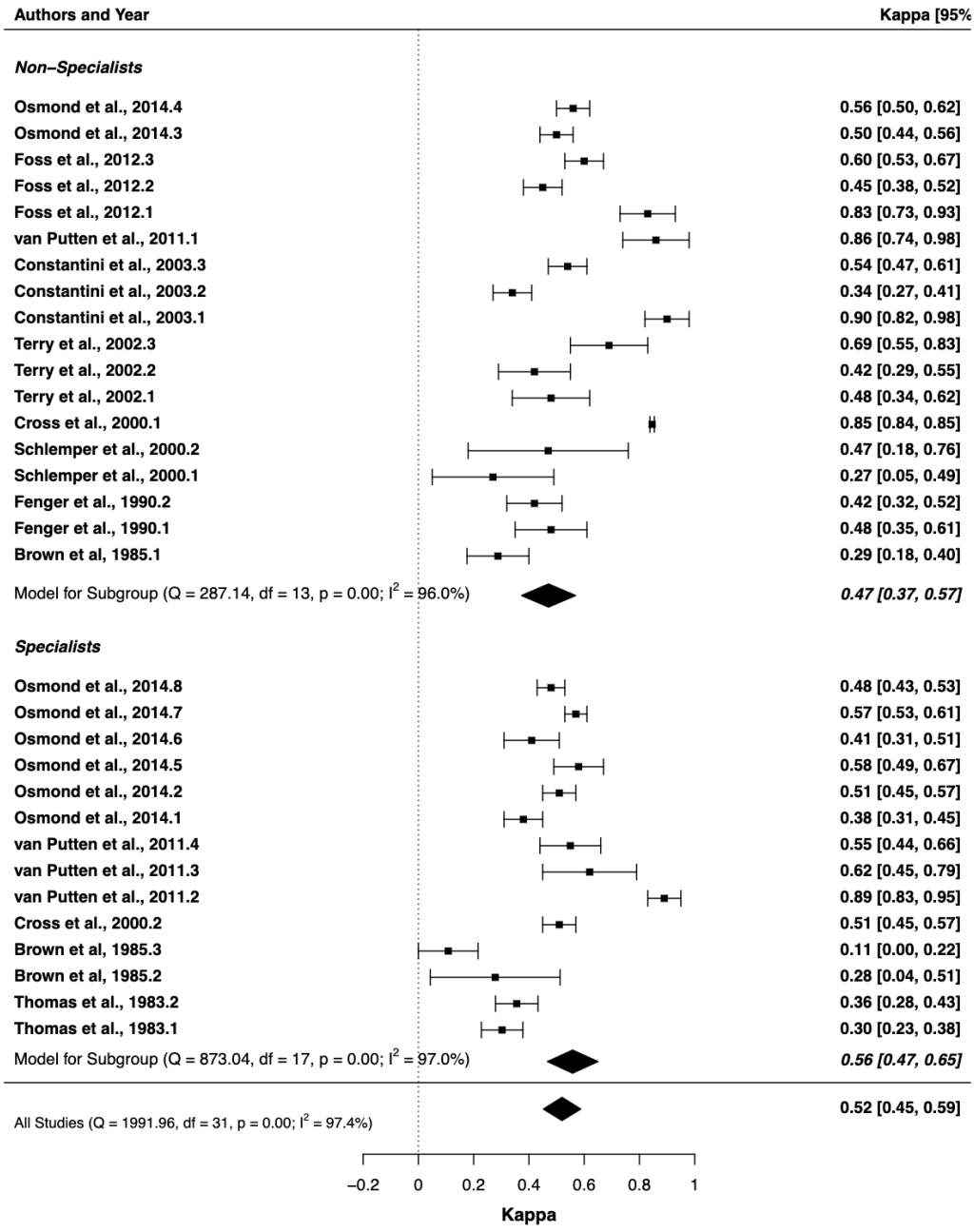


Figure 1.2: Forest plot divided by whether the pathologists specialised in the tissue type being assessed.

Guidelines

Figure 1.3 shows a forest plot divided by whether an assessment was undertaken following formal guidelines. The κ statistic for those following guidelines is 0.53 (0.45 - 0.62) while for those not following any guidelines is 0.49 (0.37 - 0.60). Again, though there is a slight improvement when guidelines are followed it is not statistically significant but this may be obscured by the high degree of heterogeneity. In particular, many of the guidelines which were followed were different, varying from WHO guidelines, Vienna classification guidelines and older classification systems. Also, some studies used one guideline for the assessment of one outcome, and a different guideline for another outcome. Again, with such heterogeneity within the studies, it is not possible to conclude that guidelines in general improve agreement between pathologists.

1.2.3.1 Discussion

Sources of heterogeneity not discussed above are likely to include sample preparation methods, differences in sample composition and regional variations in practice. However, there are insufficient studies to further sub-divide the data to explore these sources. The high degree of heterogeneity is perhaps itself indicative of that which has been noted extensively in the literature: the interpretation of villous features and the grade of dysplasia is not only subjective, but does not have shared demarcations on what is a spectrum, or even the most basic definitions [13]. For instance, villous adenomas have been defined as 'leaf-like projections lined by dysplastic glandular epithelium (which) comprise more than 80% of the luminal surface' [41]: exactly what constitutes a leaf is undefined.

There is, however, little evidence of publication bias. The funnel plot of the mixed-effects model has an even spread of estimates around the residual value, though it does show that large studies, which would have a small standard error, are lacking

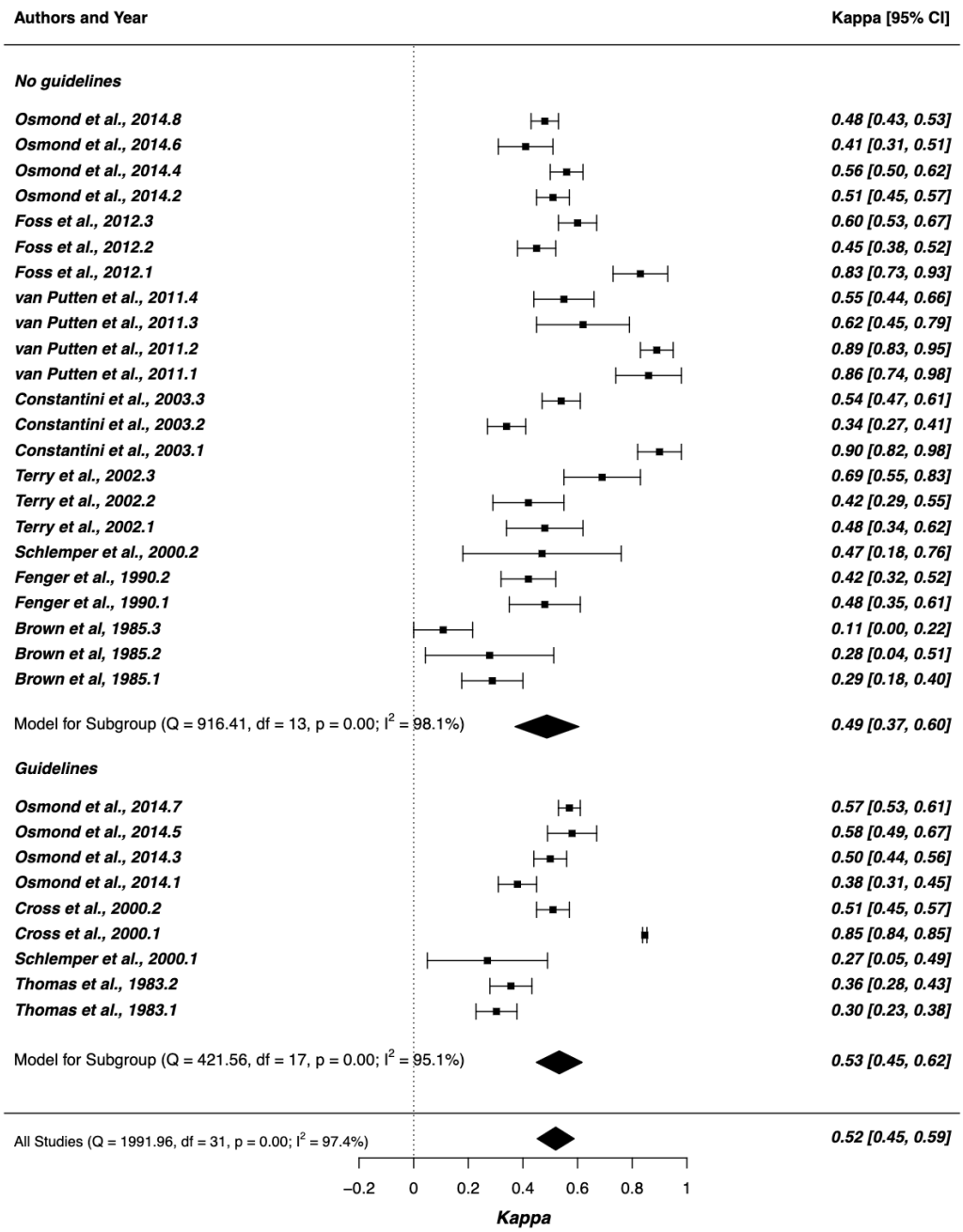


Figure 1.3

(figure 1.4). This demonstrates that the community is engaging with the problem with unbiased data.

Clinical Implications

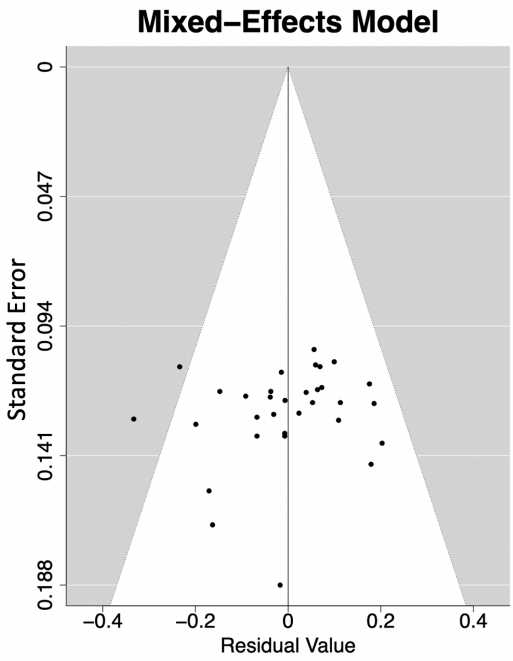


Figure 1.4: Funnel plot of the mixed effects model: an even spread indicates no publication bias

A kappa value > 0.8 is usually interpreted as meaning that the raters are in 'substantial' agreement [42]. We may then conclude that in discerning hyperplastic from adenomatous samples, pathologists often agree with one another. Patient management for these two findings differs significantly with the former requiring no special follow-up and the latter at least a follow-up colonoscopy in 3 years in the most benign case, depending on other findings such as grade or architectural type. It is difficult to gauge the impact of variations in pathology diagnosis on patient management in terms of the proportion that incorrectly miss treatment or surveillance, or how many unnecessarily undergo such management. However, interpreting the κ statistic as a form of correlation coefficient we may approximate the coefficient of determination by squaring κ, which in this context is a measure of the variation in the observations due to the raters. Taking the overall estimate for pathologists' agreement, $\kappa = 0.52$, then we may approximate that 27% of the variation in pathology diagnosis is due to the pathologist.

In addition to potential differences in clinical management there are also implications when assessing new diagnostic procedures. Any novel diagnostic test

will have to be compared to the current gold-standard of clinical pathology: it will be prudent to bear in mind that the gold-standard itself is only moderately reliable.

Histopathologists enthusiastically acknowledge the problem of subjectivity and it is a frequent topic of debate amongst themselves and their journal editorials. It is unsurprising then that many seek additional means to make their decisions more objective. Spectral histopathology (SHP) is one such attempt, and refers to any methodology that employs spectroscopy to determine the disease status, or class, of a sample. One type of SHP which is particularly popular is Raman spectroscopy.

1.3 Raman spectroscopy

1.3.1 Introduction

RS is one of several types of spectroscopy which may augment medical diagnostics. RS shares some advantages with techniques such as infrared spectroscopy (IRS), both being means of acquiring label-free biochemical information of tissue. The non-destructive nature of these techniques make them amenable to *in vivo* applications. RS has far greater spatial resolution than IRS, which could lead to more accurate predictions, although whether this is an advantage depends on the application. For instance, the high spatial resolution maps RS can acquire take a considerable amount of time, making infrared more suitable for certain time critical scenarios. This is compounded by the Raman signal being weak compared to the absorption signals of IRS. A significant advantage of RS over IRS is its relative transparency to water, which complicates the analysis of IRS [43]. These techniques are sometimes contrasted with mass spectroscopy (MS). Although this technique has the advantage of providing quantitative information and has already found clinical use [44], its destructive nature and requirement for particular sample preparation limits its applicability for clinical tasks [45]. This very brief overview serves to illustrate that no one technique is better than another, but rather the strengths and weaknesses of each must be weighed in the context of a particular clinical need.

RS has long been considered a technology that could help transform biomedical imaging. Since Feld [46] in 1995, the potential of RS to diagnose various diseases has been explored. There are a number of factors which make RS suitable for clinical use. It does not require the 'labelling' of samples, which in this context means no contrast agents or fluorescent tags are required, allowing for a smooth integration into the clinical workflow and fewer potential regulatory barriers. It can be used on either *in vitro* or *in vivo* tissue. Fast acquisition times are essential for the latter application which have been demonstrated, for instance, by teams looking at breast tumour surgical margins [47] and brain tumours [48]. It is hoped that it could also detect sub-clinical changes in a tissue, providing clinical teams with information hitherto absent in patient management decisions and perhaps even provide another platform for personalised medicine. However, the technology has yet to become an established adjunct to the histopathologist's arsenal. Barriers include the Raman signal being overwhelmed by autofluorescence, reducing the signal-to-noise ratio (SNR) below useful levels and diagnostic spectral information being packed inside a dense data set [49]. Technological solutions to this problem are often obstructive to clinical translation, such as increasing acquisition times (which could burn biological samples) and more sophisticated, and hence expensive, instruments. Computational solutions are being explored to optimise the extraction of information either for the purposes of classification or to identify specific biomarkers of disease. Bearing in mind the promise of RS for clinical applications and the inherent limitations of the technique, it is prudent to have an understanding of the physical basis of Raman scattering and the instrumentation used to acquire such data.

1.3.2 Physical basis of Raman scattering

Blue Seas Thinking

In 1921, during a sea voyage from Europe to India, C.V. Raman noted 'the wonderful blue opalescence of the Mediterranean Sea' and asked why it was this colour [50]. The received wisdom at the time, still found in some textbooks and

online resources, was that the sea simply reflected the blue of the sky. In a series of follow up experiments, inelascatic light scattering was first observed, later named the Raman effect. Ultimately the colour of the sea is not due to the Raman effect, but rather the preferential absorption of visible wavelengths other than blue. However, the 'blue seas thinking' of C.V. Raman led him to become the first Indian to win a Nobel prize in physics (in 1930). The Raman effect can be understood in terms of classical or of quantum physics, both of which have advantages and disadvantages.

Classical Theory

The classical description of Raman scattering proceeds from the polarisation of a molecule by the oscillating electric field of incident light. This induces a dipole in a molecule which then scatters the incident light. Rayleigh scattering occurs when no energy is exchanged from the vibrations in the molecule: hence the scattered light is the same colour as when incident. Raman scattering occurs when energy is either lost or gained from a molecular vibration and transferred to the outgoing photon, thus creating a shift in wavelength. To understand this quantitatively consider the relationship between the induced dipole, P, the polarisability of the molecule, α, and the electric field, E:

$$P = \alpha E \tag{1.2}$$

where the electric field is given by

$$E = E_0 cos 2\pi \nu_0 t \tag{1.3}$$

Here, E_0 refers to the amplitude (or intensity) of the light, ν_0 its frequency and t time. Combining these two equations:

$$P = \alpha E_0 cos 2\pi \nu_0 t \tag{1.4}$$

α is dependent upon the positions of the nuclei in the molecule. A molecule

with N atoms has $3N$ degrees of freedom (i.e. 3 spatial dimensions), of which $3N-6$ result in vibrations (or $3N-5$ for a linear molecule). A thorough understanding of this requires group theory: however, in the simplest case of a diatomic molecule with a single normal coordinate, Q_1 we can express α as the series expansion:

$$\alpha = \alpha_0 + \left(\frac{\partial \alpha}{\partial Q_1}\right)_0 + \ldots \tag{1.5}$$

The position of the nucleus is time dependent as the molecule vibrates with a frequency of ν_1:

$$Q_1 = Q_1^0 cos 2\pi \nu_1 t \tag{1.6}$$

where Q_1^0 is the maximum vibrational amplitude. Substituting this into equation 1.4 and taking the first order approximation of the expansion this becomes:

$$P = \left(\alpha_0 + \left(\frac{\partial \alpha}{\partial Q_1}\right)_0 Q_1\right) E_0 cos 2\pi \nu_0 t = \left(\alpha_0 + \left(\frac{\partial \alpha}{\partial Q_1}\right)_0 Q_1^0 cos 2\pi \nu_1 t\right) E_0 cos 2\pi \nu_0 t \tag{1.7}$$

Multiplying out the terms and using the trigonometric identity $cos\theta cos\phi = \frac{1}{2}(cos(\theta+\phi) + cos(\theta-\phi))$, and using colours to track parts of the equation which will later be pertinent:

$$P = \alpha_0 E_0 cos 2\pi \nu_0 t + \frac{1}{2} E_0 Q_1^0 \left(\frac{\partial \alpha}{\partial Q_1}\right)_0 \left(cos 2\pi t(\nu_0 + \nu_1) + cos 2\pi t(\nu_0 - \nu_1)\right) \tag{1.8}$$

From this derivation we see that classical theory predicts three basic types of light scattering depending on the induced dipole moment, P, oscillating at frequency ν_1. The α_0 term given in *green* is light scattered at an unshifted frequency. This is Rayleigh scattering. If $\left(\frac{\partial \alpha}{\partial Q_1}\right)_0 \neq 0$, that is if the polarisability of the molecule changes with respect to the normal mode Q_1 then we get two additional terms, in *red* and *blue*, both of which describe Raman scattering. Hence we see that there are two forms of Raman scattering. Anti-Stokes Raman scattering occurs when the frequency

of the scattered light is shifted higher ($\nu_0 + \nu_1$). Stokes Raman scattering occurs when the frequency is shifted lower ($\nu_0 - \nu_1$). Anti-Stokes scattering can only occur when the molecule has sufficient energy to transfer to the scattered light by already being in a higher state of excitation. The population of molecules in such higher states is dependent upon the temperature of the molecule and follows a Boltzmann distribution. In samples at room temperature there are typically not many excited molecules and so Stokes scattering usually dominates anti-Stokes scattering [51].

The above demonstrates that Raman scattering is dependent upon the vibrational frequency, ν_1, of the molecule. This in turn is dependent upon the mass of the atoms in the molecule and the bond strength between them. Again, taking the simplified diatomic molecule we may model the vibrational frequency by Hooke's law, whence:

$$\nu_1 = \frac{1}{2\pi C}\sqrt{\frac{K}{\mu}} \tag{1.9}$$

where C is the speed of light, K is the force constant between the atoms and μ is the reduced mass of the two atoms a and b:

$$\mu = \frac{m_a m_b}{m_a + m_b} \tag{1.10}$$

Hence we see that, all else being equal, molecules with a lower reduced mass will confer a greater change in vibrational frequency to a Raman scattered photon. Thus each unique diatomic molecule will have a unique Raman signature making it theoretically possible to identify the molecule from its scattered light. However, as we shall see, there are a number of impediments to interpreting the nature of Raman scattered light.

Quantum Theory

If we consider light in its particle form then a source of light can be thought of as creating a stream of photons, each of which has energy of $h\nu_0$, where h is Planck's constant. When a photon collides with a molecule it may scatter with the same

energy: an elastic collision or Rayleigh scattering. Occasionally though an inelastic collision occurs where the scattered photon has gained or lost energy (see figure 1.5.): Raman scattering. This occurs in accordance with quantum selection rules, where the incident photon causes the molecule to briefly jump to a higher energy state, known as a virtual state, before relaxing to a permissible lower energy state. This differs from infrared spectroscopy, another type of vibrational spectroscopy, which is dependent upon the absorption, rather than the scattering, of photons.

The induced electric dipole moment from initial state i to the final state j can be given by:

$$P_{ij} = \int \Psi_i^* \mu \Psi_j d\tau \tag{1.11}$$

where Ψ_i^* and Ψ_j are the respective time-dependent wave functions. By considering the transition probabilities from various initial and final states it should be possible to show why Raman scattering is so much weaker than Rayleigh scattering, in which the initial and final state are the same: $i = j$. However, an initial search of the Raman physics literature failed to find such a theoretical derivation, though it has certainly been empirically demonstrated [51].

1.3.3 General Raman spectrometer description

Medical applications were made possible in the 1970's when Raman spectroscopy was coupled with microscopy, sometimes referred to as Raman biospectroscopy or microspectroscopy. A schematic of a typical Raman instrument is shown in figure 1.6. Briefly, laser light is focused upon a sample of interest. This light then scatters, as described above. Rayleigh scattered light is filtered out and the remaining light is separated into its composite frequencies. When considering Raman instrumentation (and its outputs) it is conventional to speak in terms of wavelengths, λ , the inverse of frequencies ($\lambda = \frac{c}{\nu}$). These different wavelength photons are resolved into a spectrum using a grating and the number of photons within a narrow band of wavelengths

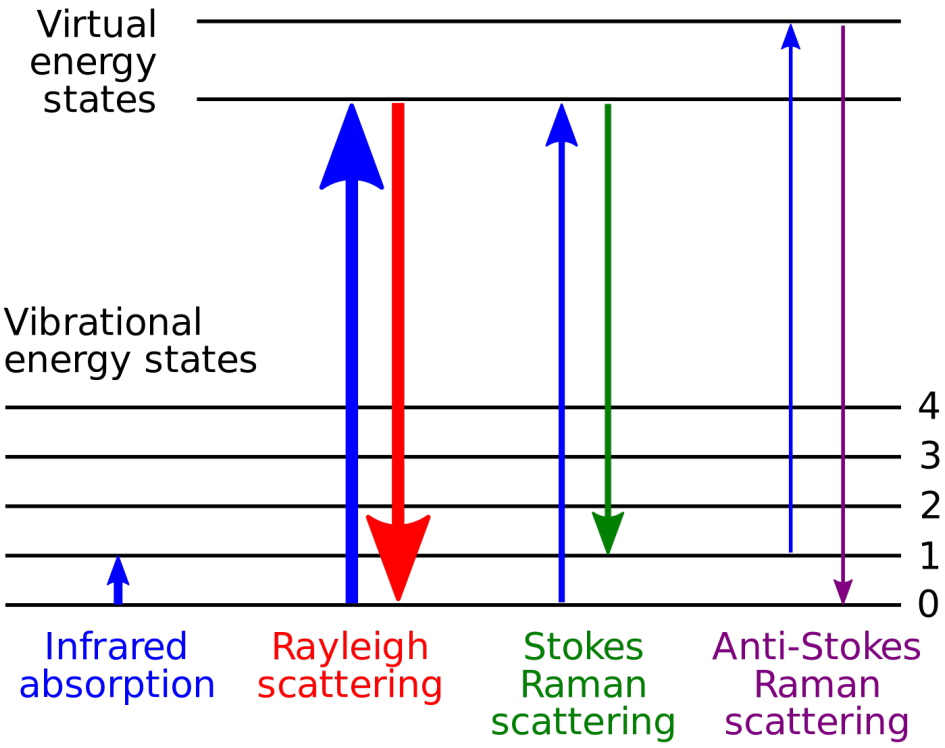


Figure 1.5: Jablonski diagram of Raman scattering: Image from https://commons.wikimedia.orgwikiFile:Raman_energy_levels.svg

counted, resulting in a spectrum such as figure 1.7. It will be important later to note that the instrument typically counts an electric charge induced by the photons, so electrons are what are actually counted. There are four major components to a Raman spectrometer: an excitation source, a sample illumination and collection system, a wavelength selector and a detection and processing system.

Excitation Source

A laser provides an intense and monochromatic source of light. This ensures that the Raman effect is maximised and Rayleigh scattered photons can be distinguished and removed. A typical Raman spectrometer for biomedical applications will deploy a diode laser. Although the magnitude of the Raman wavelength shift is independent of the incident wavelength, there are a number of other considerations to take into account when determining the optimum wavelength of the laser. The intensity

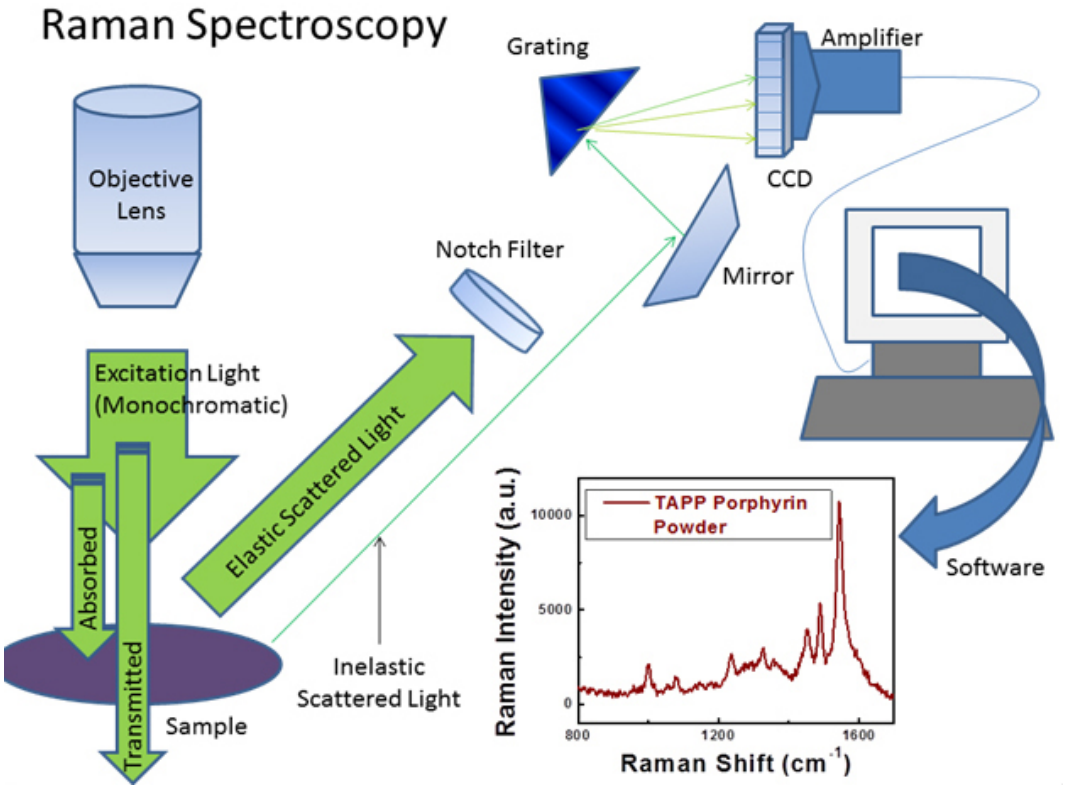


Figure 1.6: Schematic of a Raman micro-spectrometer. Note that scattered light is typically collected back through the objective lens but is shown here as a separate geometry for clarity. Image from http://www.chem.umd.edu/wp-content/uploads/2014/01/RamanSpectroscopy.jpg

of Raman scattering is dependent upon the incident wavelength by λ^{-4}, hence an infrared (IR) laser gives less intense scattering than visible light lasers. Unfortunately, it is not simply the case that a low wavelength laser gives the best Raman spectra; there are additional factors which must be taken into account as we shall later see.

Spatial resolution is also a consideration in micro-spectroscopy. The laser spot diameter (i.e. the area irradiated by a laser) is related to the laser wavelength and the Numerical Aperture (NA) of the objective lens by $1.22\frac{\lambda}{NA}$. Hence for a given NA, the greater the laser wavelength the greater the spot diameter and the less the spatial resolution. However, in biological materials one needs to also consider the simultaneous generation of a fluorescence signal which can overwhelm any Raman signal. Typically longer excitation wavelengths suppress this fluorescence signal. Near IR (NIR) lasers are currently popular as they provide a reasonable trade-off between these factors, but as we shall later see, the fluorescence problem persists.

Sample Illumination

As Raman scattering is a weak process, the laser should be properly focused and the scattered light efficiently collected to maximise the Raman signal. After the laser light has interacted with the sample, light is scattered in all directions. There are several optical configurations used to collect this scattered light: with 90° and 180° geometries between excitation and collection commonly used.

Wavelength Selector

Holographic notch filters are now replacing monochromators as the method by which to suppress Rayleigh scattering. This ensures that only the light of a different wavelength to the laser light passes through the system. A major advantage of these filters compared to older systems is that they can measure both Stokes and anti-Stokes scattering. The scattered light is then collected using a focusing lens or mirrors. It then passes through a diffraction grating or prism in order to disperse the various wavelengths of scattered light, ready for detection.

Detection and Processing

Modern Raman spectrometers use charge-coupled devices (CCDs) to convert photons to an electrical signal which can be read. A CCD is a silicon-based semiconductor arranged as an array of photosensitive pixels, each of which produces photoelectrons which are stored as a small electrical charge. This analogue signal is then converted to a digital signal which is interpreted as the number of photons of a particular wavelength. One problem with CCDs is that the pixels have a limit to the amount of charge that they can hold. If there are too many photons the pixel will become saturated and lose its charge, consequently reading as detecting no photons.

The Raman system described above will hereafter be referred to as a spectrometer

or the instrument and by the above processes will produce a Raman spectrum.

1.3.4 Data description and biochemical interpretations

A Raman spectrum is represented as a graph in which the x axis shows the Raman induced change in wavelength and the y axis the photon count at the different wavelength positions (figure 1.7). In the bio-Raman literature the wavenumber shift, $\Delta\omega$, is the preferred unit of measurement of wavelength shift. It is simply the reciprocal of the wavelength, $\omega = \frac{1}{\lambda} cm^{-1}$ and so has the units of reciprocal length, given in reciprocal centimetres. A peak associated with a wavenumber is often called a Raman band, or simply a peak, and represents the vibrational energy of a particular molecular bond in a sample. The photon count, also called the Raman intensity, is then proportional to the amount of corresponding vibrations in the sample. If a single molecule has distinct molecular bonds, as many do, a Raman spectrum will similarly have distinct Raman peaks.

Figure 1.7 is an example of a 'pure' Raman spectrum - the results of taking a spectrum from a single chemical species. In biological tissues there are complex mixtures of biochemical species, resulting in an equally complex Raman spectrum which is a non-linear composite of many pure signals. As can be seen in figure 1.8, the resulting spectra are more complicated than spectra from pure sources. It is possible to construct some of the biochemical signatures in a spectrum from theoretical quantum principles, though this is extremely difficult and not practical for research and medical applications. Instead, an empirical approach has accrued over the decades, in which certain Raman bands have become associated with various biochemical antecedents [52]. Figure 1.8 highlights some regions which have been associated with broad biochemical species.

A single Raman spectrum taken from a complex sample, such as tissue, will therefore be a combination of many biochemical constituents. The task is to 'unmix' these so that the underlying biochemistry can be revealed and, ultimately, different disease processes can be identified. Unfortunately, the Raman effect is a relatively

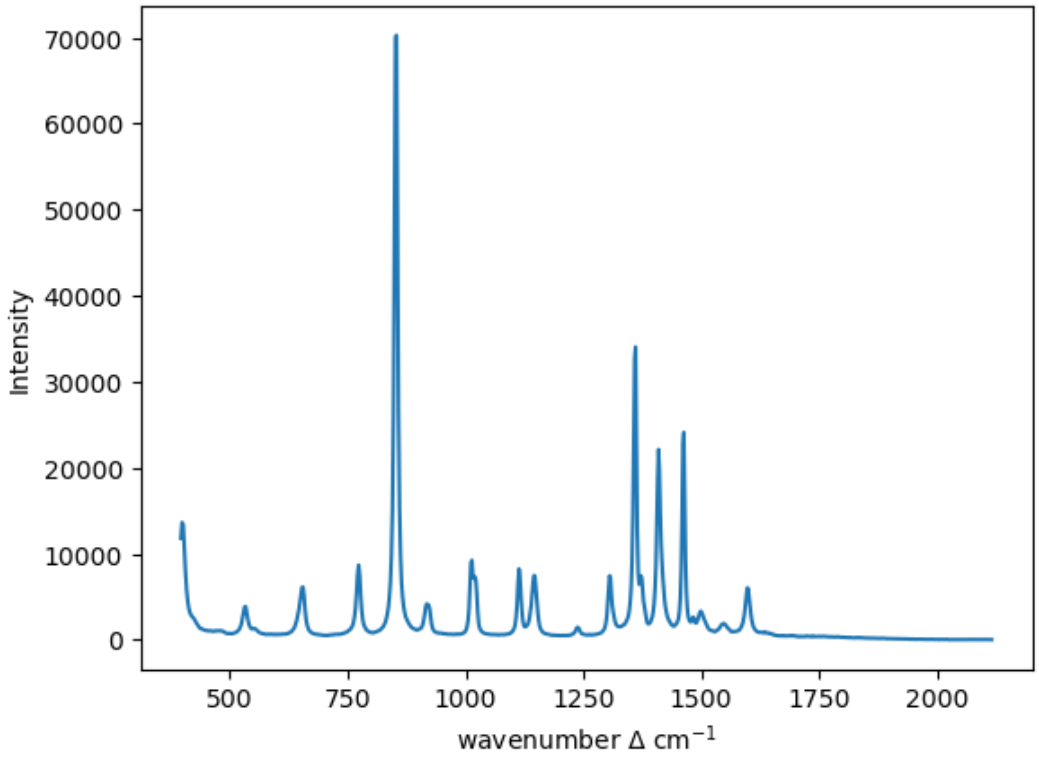


Figure 1.7: Raman spectrum of the amino acid Alanine

weak phenomenon, with approximately 1 in 10^6 photons being inelastically scattered. This means that the Raman signal can easily be obscured by numerous sources of noise, making analysis more difficult. Figure 1.9 shows how these various factors combine to create a measured Raman spectrum.

1.3.5 Anatomy of noise in a Raman spectrum

Each component of a spectrometer can introduce noise into a spectrum. This is in addition to noise that is intrinsic to the light scattering phenomenon. All such contributions can generally be referred to as noise, and are often modelled as a single entity, but here we consider each of these contributions. The sum of these independent sources of noise determine the overall variance of the signal, σ_p^2:

$$\sigma_p^2 = \sigma_x^2 + \sigma_d^2 + \sigma_f^2 + \sigma_r^2 + \sigma_b^2 \tag{1.12}$$

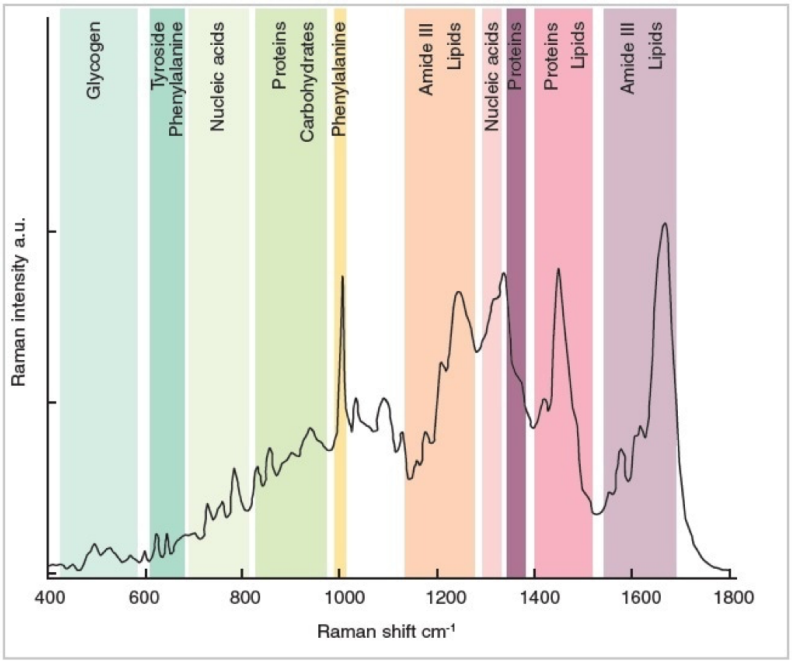


Figure 1.8: Biological markers in a Raman spectrum. Certain regions and Raman bands are known to be associated with some biochemical species. The Raman intensity is given in arbitrary units.

Shot Noise

Shot noise, σ_x^2, manifests due to the discrete nature of photon detection. For a given laser power, p, and frequency, f, there will be a number of photons emitted per second, n. As the power is given by $p = nhf$, where h is Planck's constant, we can calculate the number of photons emitted per second by $n = \frac{p}{hf}$. However, due to the stochastic nature of stimulated photon emission in a laser, this number of photons will not be constant. Rather, it is a Poisson process in which the photon count, N, is given by$P(N = k) = \frac{e^{\lambda t}\lambda t^k}{k!}$, where t is the expected number of photons in a unit of time. This has expectation and variance $E[N] = var(N) = t$. Hence, shot noise is signal dependent: the more signal (i.e. a greater expected number of emitted photons) the greater will be the variance. Shot noise represents a lower limit to the amount of noise present in a Raman signal as it is an irreducible manifestation of quantum stochasticity.

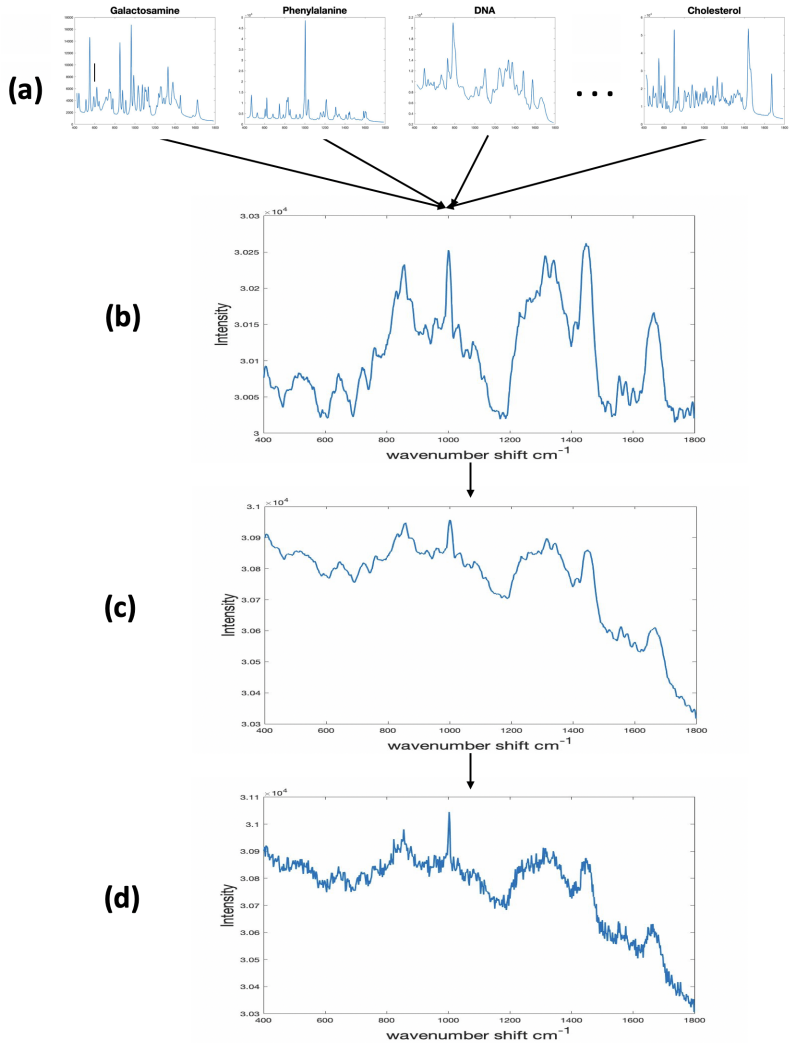


Figure 1.9: Anatomy of a Raman spectrum. (a) Numerous 'pure' biochemicals are in any single tissue sample. (b) These combine in a non-linear fashion to make a spectrum which is a composite of the pure biochemicals. (c) A florescent baseline is often present in tissue samples. (d) Various sources of instrumental noise are also present in the measured spectrum.

Dark Current Noise

Dark noise, σ_d^2, is caused by a random stream of electrons generated thermally within the silicon structure of the CCD. This process is independent of light intensity, but dependent on time. It can be modelled by $\sigma_d^2 = a\beta(T)t$ where a is the conversion efficiency from electrons to counts, $\beta(T)$ is the temperature, T, dependent dark current and t the integration time. Thus dark noise can be diminished by reducing

the integration time. $\beta(T)$ also follows Poisson statistics. As it is strongly sensitive to the CCD temperature the most effective means to reduce this noise is to reduce the temperature. For example, the CCD in the Renishaw RA816 instrument, used extensively in this investigation, is cooled to $-70°C$: at this temperature any contribution from dark noise should be negligible. This cooling also improves the dynamic range (DNR) of the measurement $DNR = \frac{MaxSignal}{\sigma_d^2}$. However, to keep a stable temperature a control unit is required which can add to system flicker noise.

Flicker Noise

Flicker noise, σ_f^2, is caused by variations in laser intensity which cause proportional variations in photon detection. It is frequency dependent. Multichannel spectrometers, such as the RA816, are robust against this noise as all wavelengths are monitored in parallel.

Readout Noise

Readout noise, σ_r^2, occurs during the process of transforming CCD charge carriers into a signal that is stored in digital form on a computer. Unlike other sources of noise, it does not depend on the signal intensity or measurement time.

Background Noise

Perhaps the most pernicious of all, background noise, σ_b^2, is a general term referring to any photons detected other than Raman scattered photons. This includes background light not shielded by the instrument, Rayleigh scattered light not removed by filters and luminescence of the sample in the form of fluorescence or thermal emission. It is dependent upon the intensity of the laser and the presence of fluorophores in the sample and also follows poisson statistics.

The presence of fluorescence can be so severe as to render a Raman signal

illegible. Various methods are used to ameliorate this problem, including the selection of an excitation wavelength that is less likely to cause fluorescence. This is possible because higher wavelength lasers, which have lower energy, are not able to elevate an electron in a molecule to a higher electronic state, but only to the virtual energy state required for Raman (and Rayleigh) scattering to occur. However, this must be balanced with the intensity of Raman scattering which is also dependent upon the wavelength of the laser. The Raman intensity decreases as the wavelength of incident light increases. Hence, for many biomedical applications of RS, a laser of 785nm is chosen as a reasonable compromise between these competing interests. How to model with this noise will be considered in section 5.3.

Cosmic Ray Noise

An additional source of noise, not stated in equation 1.12 as it is considered uncontroversial to remove, comes from extra-terrestrial sources. High intensity spikes in a Raman spectrum are seen when high energy cosmic rays are detected by the CCD. These are spurious spikes that could be misconstrued as Raman peaks, so some care is needed in their detection and removal. Fortunately, such is their distinct nature from Raman peaks that their detection and removal is considered uncontroversial in the Raman literature, even with fully automated algorithms.

1.3.6 Measuring noise

Signal to Noise

The signal to noise ratio (SNR) is a measure of the amount of noise in a signal as a ratio to the amount of true signal present (i.e. Raman scattering). In RS this is defined at a given spectral peak rather than the entire spectrum as shot noise is dependent upon the signal intensity and thus varies by wavenumber. It can be measured by a number of methods, each with different strengths and weaknesses. The main choice is whether to define it for each dataset (or a subset thereof) or

for each spectrum [53]. In the ideal scenario, the same peak would be repeatedly measured from several spectra taken from the same location on a sample, and their mean and variance taken. However, a repeated measurement will never be identical in RS due to several effects. Photobleaching is a phenomenon whereby fluorescence is reduced after consecutive measurements. This occurs as the relaxation time for a fluorescent photon is orders of magnitude longer than for virtual photon absorption and emission. Hence background noise is reduced with consecutive measurements. Despite photobleaching being a useful technique to reduce fluorescence, it has been found that this effect increases calibration model error in biological samples [54]. There is also a risk of burning, or otherwise thermally altering, biological samples with repeated measurements, depending upon the acquisition time and laser intensity. These are in addition to the practical limitations of taking repeated spectra during the acquisition of a Raman map, which may already contain many thousands of spectra.

Hence, for this project I define the SNR for each spectrum and take an average over these for the SNR for a dataset. This is done by taking the height of a selected peak as the signal, then selecting a region near the peak which is assumed to be noise. The SNR is then the ratio of the peak height, S_p, to the standard deviation of the noise, σ_p:

$$SNR = \frac{S_p}{\sigma_p} \tag{1.13}$$

A feature of the SNR in RS is that if shot noise dominates, as the signal intensity increases the noise also increases, but the SNR decreases. This is because the noise grows at the square root to the signal. This can be seen by remembering that for a Poisson variable the mean and variance are equal. Noting that the standard deviation is the square root of the variance, then $\bar{S} = \sqrt{\sigma}$. Then, following from equation 1.13:

$$SNR = \frac{S_p}{\sigma_p} = \frac{\sigma_p}{\sqrt{\sigma}_p} = \sqrt{\sigma}_p \tag{1.14}$$

Hence, all else being equal, having a stronger intensity diminishes the amount of noise. This could be achieved either by increasing the power of the incident laser

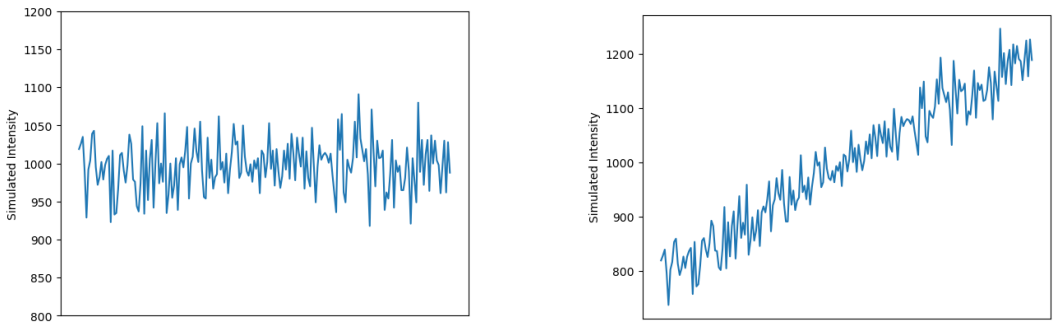
or by exposing the sample to the laser for a longer period of time: both methods increase the number of photons detected by the CCD. However, CCD pixels have a limit to the number of photons they can detect, after which the pixel becomes saturated and reads as zero.

Point to Point Noise

The above derivation of SNR assumes that the region selected to measure the standard deviation is flat. As discussed, Raman spectra often sit atop background noise which creates slopes. Such slopes mean that the standard deviation is not a reliable measure of variation. This is illustrated in figure 1.10 showing a simulated Poisson process with and without a sloping baseline. In both cases, the average is 1000. As this is simulated noise we know the standard deviation is exactly $\sqrt{1000} = 31.6$. The measured standard deviation in figure 1.10a is close to this at 32.1. In figure 1.10b the standard deviation is measured at 120.0, a gross over-estimation.

To account for the effect of a sloping baseline I use another metric of noise; point to point (PP) noise:

$$PP = \frac{1}{\sqrt{2}}\sqrt{(\frac{\sum_2^n (x_i - x_{i-1})^2}{n-1}} \tag{1.15}$$



(a) **(b)**

Figure 1.10: **a**.) Simulated Poisson noise. **a**.) An average intensity of 1000 **b**.) also an average intensity of 1000, but with a slope

The factor of $\frac{1}{\sqrt{2}}$ is a correction factor which is exact for Gaussian noise, and

approximate for Poisson noise, though it is very precise at the intensities typical of RS. Using PP noise, figure 1.10b has noise of 31.3: much closer to the true noise of 31.6. Hence, when measuring noise, I will use the PP noise rather than the standard deviation, thus equation 1.13 becomes:

$$SNR = \frac{S_p}{PP} \tag{1.16}$$

The phenylalanine peak, at approximately $1003cm^{-1}$, is a remarkably consistent peak within biomedical RS and so will be used to measure noise. It will be measured as the maximum value in the region $995cm^{-1} - 1015cm^{-1}$, to account for any wavenumber shifting. The phenylalanine peak is often swiftly followed by another peak at $1033cm^{-1}$, hence to measure the PP noise the region just adjacent to the phenylalanine peak at $960cm^{-1} - 990cm^{-1}$ will be used (figure 1.11).

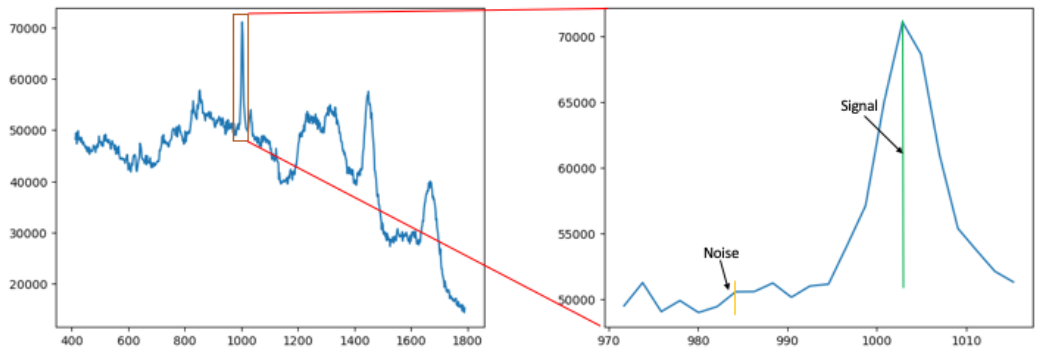


Figure 1.11: Phenylalanine peak for noise measurement

The noise inherent in RS together with the complex nature of the underlying cancer biology we seek to detect mean that the technique has yet to become routinely used in the clinical setting. In order to decipher these dense datasets, we next turn our attention to their analysis and previous work in this domain.

1.4 Introduction to chemometrics and machine learning

1.4.1 Chemometrics

Chemometrics refers to various analytical techniques which are used to detect and extract information from Raman, or other, spectra. This can be divided into two components: data preparation (or pre-processing) and downstream analysis. The former is concerned with improving the quality of Raman spectra by attempting to mitigate the various sources of noise discussed above. The latter is concerned with determining the spectral, and biochemical, characteristics that distinguish different samples. Section 5.3 will outline some of the many pre-processing steps which constitute the data preparation component required to make the data amenable to downstream analyses.

A very common statistical chemometric technique used in biomedical RS is Principal Component Analysis - Linear Discriminant Analysis (PCA-LDA) [55]. PCA reduces the dimensionality of the data and removes some noise; LDA then learns a criterion by which to separate data as belonging to one of several classes (i.e. different diseases), based on labelled examples. This model will be explored in more detail in section 4.3.

One way of analysing data is to build models based on known physical principles. An example is the use of the Beer-Lambert law which relates the attenuation of light through a substance. It has been used in RS to mitigate Raman self-absorption [56]. However, light scattering phenomena, especially when interacting with biological samples, are often too complex to usefully model in this way. A much more common approach is to use machine learning (ML).

1.4.2 Machine learning in healthcare

ML is any model which learns from the data, rather than being based on explicit rules or physical principles; a bottom-up rather than top-down approach [57]. It can therefore be described as a data driven approach to modelling. There are many variations. A common example used in biomedical RS is the Support Vector

Machine (SVM). This, and other traditional ML models, are held in contrast to deep learning (DL) models, which are large and complex models based on neural network architecture. DL models could revolutionise the digital healthcare space [58], including biophotonics [59]. In particular, their ability to capture non-linear complexities in a dataset allow them to exploit patterns too subtle for traditional methods, making them an ideal candidate to realise the full potential of RS. An area that has already benefitted from deep learning is digital pathology, particularly applied to oncology [60].

However, ML, traditional or DL, is not without its limitations. Just as medical researchers need to understand something of the statistical science of hypothesis testing, and the debate and misunderstandings regarding p-values, it is becoming increasingly important to become literate in ML [57]. One of the barriers to transferring promising ML results to clinical settings is the reproducibility of results [61]. Indeed, a recent review of ML applications to diagnose COVID-19 using chest radiographs or CT scans found that of sixty two studies, none were of sufficient quality to be clinically relevant [62]. Prominent among the given reasons were methodological issues that compromise the generalisability of a model to the target population. More generally, a recent high profile and influential editorial has warned of an impending reproducibility crisis in ML science [63], based on the finding that 329 published papers across 17 disciplines, from histopathology to satellite imaging, contained methodological issues sufficient to render findings unreproducible [64].

In the next section we thoroughly investigate the current state of ML in the field of RS applied to oncology problems, both to understand where current research is focussed and to determine to what extent the field suffers from overly-optimistic results which could threaten the generalisability of findings.

1.5 Machine learning with Raman spectroscopy to distinguish cancers

1.5.1 Literature review: methods

This literature review follows the principles set out in the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-analyses) guidelines [65]. The databases PubMed and Web of Science were extensively searched by combining the search terms 'Raman Spectroscopy' and 'Learning' with the AND Boolean operator. Recovered titles and abstracts in the databases were searched as illustrated in figure 1.12, and any oncology studies were identified. Publications were limited to the English language and being published from January 2018 to the date of the search (October 2021). Potentially relevant studies were selected for a full text review. Additional studies were identified among the references of identified studies. Studies were excluded if they did not explicitly classify data or were not peer reviewed. As the ML methodology was itself the focus of this review, no attempt to exclude studies based on methodological quality was made and so the PRISMA quality checklist was not applied.

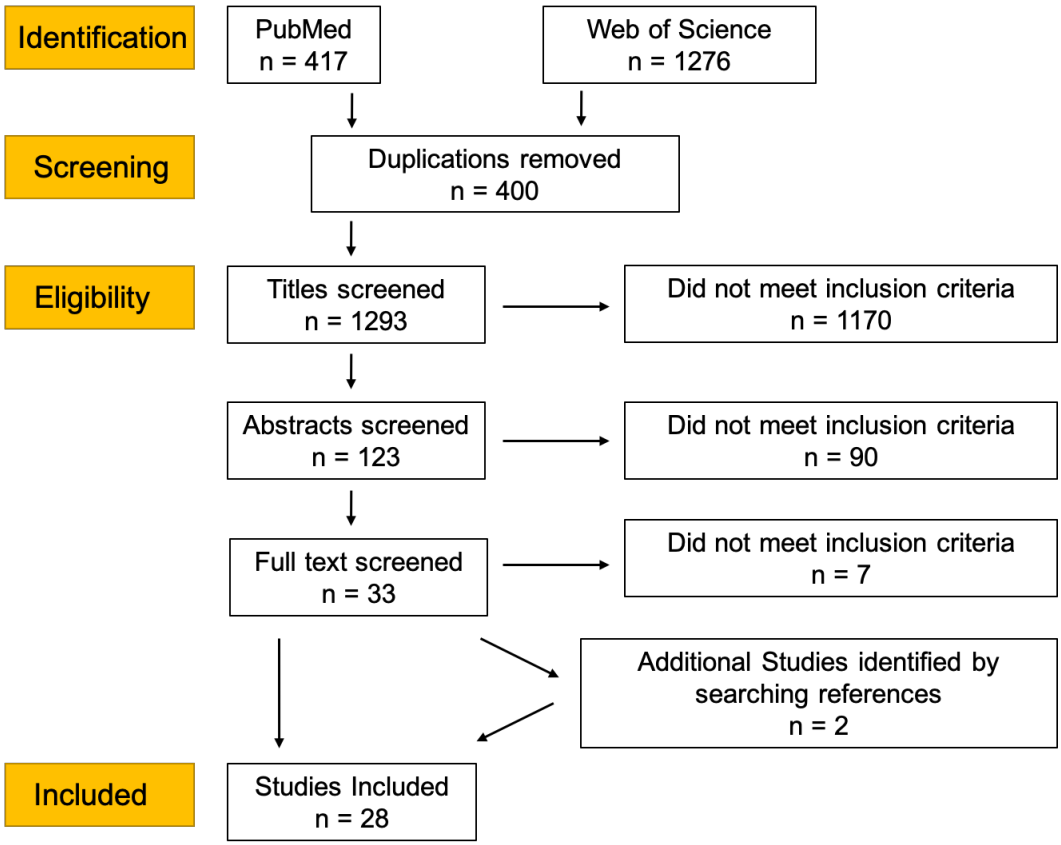


Figure 1.12: Literature search strategy

1.5.2 Literature review: results

A total of 28 studies were identified (table 1.1), 18 of which interrogated tissues, 4 studied cell lines, 5 blood serum and one urine. All of these studies classified Raman spectra into at least two groups, usually healthy and cancerous. The tissues explored in the literature (regardless of the sample substrate) included brain (5), tongue (3), prostate (3), breast (3), lung (3), skin (3), nasopharyngeal (2), colon (2), oral (1), cervical (1), ovarian (1) and kidney (1).

Many studies used several ML models, conducted analyses on different subsets of their data and/or compared several pre-processing techniques, producing a multitude of disparate results. For instance, several studies compared the performance of different machine learning models, often traditional ML models like LDA against DL models, such as convolutional neural networks (CNNs). For ease of comparison,

Authors/year	Pathology (sample type)	Model	Validation Strategy	Number of patients/ samples	Number of spectra	Level of Split	Number of Classes	Accuracy (sensitivity/specificity)
Aubertin *et al.* 2018 [66]	Prostate Cancer (tissue)	ANN	LOOCV	32 subjects/samples	928	Not Stated	2	86% (87%/86%)
Baria *et al.* 2020 [67]	Skin Cancer (cell lines)	PCA-ANN	5-fold CV	Not Stated	150	Not Stated	3	96.7%
Bury *et al.* 2019 [68]	Brain Metastases (tissue)	PCA-LDA	Not Stated	21 subjects	525	Not Stated	2	80.2%
Chen *et al.* 2022 [69]	Ovarian Cancer (plasma)	ANN ensemble	Outer fold - single 66/33 Inner fold - 5-fold CV	174 subjects	870	Spectra	2	94.8% (95.3%/95.3%)
Chen *et al.* 2021 [70]	Prostate Cancer (urine)	CNN	5-fold CV	84 subjects/samples	501	Subject	2	74.95% (77.32%/72.46%)
Chen *et al.* 2021 [71]	Lung cancer & glioma (tissue)	CNN	5-fold CV	104 subjects/samples	520 2700 after augmentation	Subject	2	99% (all pairwise comparisons > 95%)
Daniel *et al.* 2019 [72]	Cervical Cancer (tissue)	PCA-ANN	Single 70/30	245 samples	Not Stated	Not Stated	3	99.0% (87%/86%)
Fang *et al.* 2021 [73]	Various Cancers (cell lines)	CNN (ResNet)	10-fold CV 500 repeats	33 subjects	510 6600 after augmentation	Not Stated	11	100%
He *et al.* 2021 [74]	Renal Cancer (tissue)	SVM	LOOCV	77 subjects/samples	4860	Subject	3	92.89%
Ito *et al.* 2020 [75]	Colon Cancer (serum)	Boosted Tree	Not Stated	184 subjects/samples	3 spectra per subject. Average used.	N/A	2	100%
Jeng *et al.* 2019 [76]	Oral Cancer (tissue)	PCA-QDA	k-fold CV and LOOCV	80 subjects/samples	400	Sample	2	81.75% (83.63%/79.44%)
Koya *et al.* 2020 [77]	Breast Cancer (tissue)	CNN	Single split 60/20/20	88 subjects/samples	34505	Spectra	2	90% (89% - precision 89% - recall)
Lee *et al.* 2020 [78]	Prostate Cancer (extracellular vesicles from cell lines)	CNN	Single split 70/15/15	1 sample per class, 4 classes	300 1200 after augmentation	Spectra	4	96.56%
Ma *et al.* 2021 [79]	Breast Cancer (tissue)	CNN	10-fold CV	20 subjects 40 samples	600 5000 after augmentation	Not Stated	2	92.00% (98.00%/86.00%)
Mehta *et al.* 2018 [80]	Brain Meningioma (serum)	PCA-LDA	LOOCV + independent test set	20 subjects 70 subjects	~8 spectra per subject. Average used.	N/A	2	86%
Qi *et al.* 2022 [81]	Lung Cancer (tissue)	CNN	10-fold CV	77 subjects/samples	15 spectra per sample	Spectra	2	97.7% (96.7%/98.8%)
Riva *et al.* 2021 [82]	Glioma (tissue)	Gradient Boost	LOOCV	63 subjects/samples	3450	Subject	2	83% (82% - precision 82% - recall
Santos *et al.* 2018 [83]	Skin (tissue)	PCA-LDA	Single split 60/40	128 samples	9-19 spectra per sample	Sample	2	62.5%
Sciortino *et al.* 2021 [84]	Glioma (tissue)	SVM	LOOCV	38 subjects/samples	2073	Subject	2	87%
Serzhantov *et al.* 2020 [85]	Skin (tissue)	Gradient with soft voting	Single split 50/50 1000 repeats	139 subjects	556	Not Stated	2	90.5% (93%/88%)
Shin *et al.* 2020 [86]	Lung Cancer (SERS of plasma)	CNN (ResNet)	5-fold CV	63 subjects/samples	2150	Spectra	2	94.8%
Shu *et al.* 2021 [87]	Nasopharyngeal Cancer (*in vivo* tissue)	CNN	10-fold CV Venetian blind	418 subjects 888 samples	15354 Augmented - quantity not specified	Sample	2	84.43% (99.15%/65.77%)
Wu *et al.* 2021 [88]	Colon Cancer (tissue)	CNN	LOOCV	45 subjects/samples	233 2420 after augmentation	Spectra AND Subject	3	93.8% - by spectra 81.3% - by subject
Xia *et al.* 2021 [89]	Tongue Cancer (tissue)	CNN-SVM	5-fold CV	12 subjects 24 samples	At least 216	Not Stated	2	99.54% (99.54%/99.54%)
Yan *et al.* 2021 [90]	Tongue Cancer (tissue)	CNN ensemble	5-fold CV	22 subjects 44 samples	2004	Not Stated	2	98.75% (99.10%/98.29%)
Yu *et al.* 2021 [91]	Tongue Cancer (tissue)	CNN	5-fold CV	12 subjects 24 samples	1440	Not Stated	2	96.90% (99.31%/94.44%)
Zhang *et al.* 2021 [92]	Breast Cancer (cell lines)	PCA-SVM	Single split	6 cell line 900 cells	4500	Not Stated	2	99.0% (99.9%/96.2%)
Zuvela *et al.* 2019 [93]	Nasopharyngeal Cancer (*in vivo* tissue)	GA-PLS-LDA	LOOCV	62 subjects 113 samples	2126	Sample	2	98.23% (93.33%/100%)

Table 1.1: Literature review results table

and to mitigate against selection bias, the best performing model and/or dataset is presented in table 1.1: other results are included when pertinent to a particular discussion. In the vast majority of cases the accuracy of a model was the primary reported performance metric: the number of correct classifications divided by the total number of classification attempts. Although its suitability to prediction tasks has been questioned, because of its ubiquity in the reviewed literature and its intuitive interpretation I report this metric unless otherwise stated.

1.5.2.1 Overview of the studies

Oral and Nasopharangeal Cancers

Xia *et al.* [94] probed tongue squamous cell tissues using a fibre optic Raman spectrometer and developed a CNN-SVM (Support Vector Machine) for binary classification. This model replaces the final dense layer of a typical CNN with a SVM, combining the feature selection prowess of the former with the classification abilities of the latter. SVMs can utilise a number of kernel functions to better model non-linearities in the data; in this paper a radial basis function (RBF) was used. They compared this model to a standard CNN as well as PCA-LDA and PCA-SVM (RBF). The CNN-SVM performed best (accuracy = 99.54%, sensitivity = 99.54%, specificity = 99.54%), as determined by accuracy, though trade-offs between sensitivity and specificity may change this interpretation according to clinical needs.

The same team used a similar set-up to collect two datasets taken under conditions of 'illumination' and 'no light' [90]. These datasets underwent a further division of pre-processing or no pre-processing, to make a total of four datasets. These were used to classify spectra into binary classes using an ensemble CNN, in which several CNN models are trained and the outputs integrated to give a consensus result. They found that the best performance was attained under the no ambient light conditions with pre-processing applied (accuracy = 98.75%, sensitivity = 99.10%, specificity = 98.29%), although the difference in accuracy to the worst performing dataset (illumination and no pre-processing) was only 4.75%.

The last publication from this team used a similar set-up and dataset to compare the performance of a custom built CNN against PCA-LDA and PCA-SVM (with a radial basis and a polynomial kernel) [91]. They found that the CNN outperformed the other ML models (accuracy = 96.90%, sensitivity = 91.67%, specificity = 94.44%).

It is not clear if the tissues used in these three studies are the same. However, in

all cases the diseased and healthy samples were obtained from the same subjects.

Also investigating oral cancers, Jeng *et al.* interrogated cryopreserved samples, seeking to discriminate between healthy and cancerous tissues [76]. They further performed a sub-group analysis, dividing their dataset into tongue, buccal and gingiva tissues to perform three pairwise cancerous versus healthy binary classifications. They additionally performed a 'point-wise' approach in which five spectra were taken per sample and a 'patient-wise' approach in which the average of these five spectra was taken. They explored two cross-validation (CV) techniques, comparing a k-fold versus a leave-one-out-CV (LOOCV) strategy. CV is a vital aspect of dividing data for ML and is considered in detail in section 5.2. Finally, they compared a PCA-LDA and a PCA-QDA (Quadratic Discriminant Analysis) classifier. Using these methods they found that taking the average spectrum of a sample yielded better performance than a point-wise approach and with PCA-QDA typically performing better than PCA-LDA, though not across all sub-group analyses. LOOCV resulted in lower error rates compared to k-fold CV for an 'all cancer' versus 'healthy' analysis, but this was reversed for a sub-group analysis, which consisted of smaller sample sizes.

Two studies focused on nasopharyngeal cancers. Zuvela *et al.* used an *in vivo* set-up to collect data during endoscopy [93]. They employed a genetic algorithm (GA) to perform feature selection for a PLS (Partial Least Squares)-LDA binary classifier, comparing its performance to a PLS-LDA model without this selection. They also compared performance when utilising either the fingerprint region, the high wavenumber region, or both combined. Not only did the GA-PLS-LDA outperform the generic model (accuracy: 98.23% versus 95.58%), but this feature selection was also used to find candidate Raman peaks responsible for this discrimination. Also, though combining fingerprint and high wavenumber regions may actually confuse ML models by including irrelevant data, the GA feature selection was able to mitigate against this potential danger, improving accuracy (fingerprint = 92.04%, high wavenumber = 94.69%, both = 98.23%).

The same team expanded with a similar study which included many more subjects, samples and spectra with a similar recruitment protocol [95]. Of all the

studies reviewed, this is the largest in terms of the number of subjects recruited. This time, however, they used a CNN to classify three classes (cancerous, post-treatment and healthy) in a pairwise fashion. This consistently performed better than a PLS-LDA classifier. The CNNs superior performance was maintained even when the data was down sampled by a factor of two and four, contrary to the idea that CNNs require a plethora of data from which to learn. Indeed, the CNN trained on data down-sampled by a factor of two produced the best performance.

Lung Cancers

Qi *et al.* adopted a novel approach to classify Raman spectra of lung tissue as adenocarcinoma, squamous cell carcinoma or normal in a pairwise fashion [81], They transformed the data into 2D spectrograms in a similar process used to classify audio data. This spectrogram data was used in a CNN accepting 2D inputs, akin to typical image classifiers and different from all the 1D inputs thus far discussed, and compared performance to a PCA-LDA model. For both pairwise comparisons the CNN returned an accuracy over 96% while neither PCA-LDA model broached 90%.

Shin *et al.* used surface enhanced Raman spectroscopy (SERS) of exosomes (a potential oncology biomarker) derived from cell lines to classify early stage lung cancer [86]. They used a well know CNN architecture called ResNet. This CNN includes 'skip connections' which allow the network to learn the identity function during training which allows for a much deeper (i.e. more convolutional layers) network, which should allow it to learn even more subtle features in the data [96]. It compared favourably to the traditional ML models PCA-LDA, SVM and PLS-DA, as well as to another large CNN architecture called VGG-16.

Chen *et al.* also discriminated between lung cancer, as well as glioma, a common brain cancer, using spectra taken from blood serum [71]. They compared both classes against healthy controls in a pairwise manner. Several deep learning architectures were compared: an ANN (Artificial Neural Network), a RNN (Recurrent Neural network), an LSTM (Long Short-Term Memory) and AlexNet. The dimension

reduction techniques of PCA and PLS were also compared to no pre-treatment. The use of data augmentation was also explored by increasing the number of spectra 5-fold. Across all analyses, this augmentation increased performance. This was most pronounced when PLS was first performed on the data. AlexNet, the largest model used, together with PLS and data augmentation, was the best performing model, although the difference amongst all the models, except the ANN, was minimal. However, when a three class model was constructed, the best performance had an accuracy of 85.1%.

Brain Cancers

Two studies from the above team also explored gliomas. Riva *et al.* took fresh tissue biopsies and classified healthy versus cancerous tissue using the traditional models of Random Forest (RF) and Gradient Boost Tree (GB) [82]. The latter model performed best and due to its feature selection allowed for the detection of novel Raman peaks to be implicated in Gliomas. In the team's second study, Sciotino *et al.* explored the potential to discriminate between the mutational status of gliomas, essentially attempting to genotype using RS [84]. They used GB and SVM (RBF) to successfully classify between the two disease genotypes.

Bury *et al.* also analysed brain tissue, attempting to discriminate the primary source of metastatic brain cancers [68]. Seven samples each with primary sources of lung adenocarcinoma, colorectal carcinoma and melanomas were obtained and 25 spectra collected per section. RS was compared to attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectroscopy. An overall accuracy of 69.7% was achieved compared to just 55.3% using similar PCA-LDA modelling on ATR-FTIR data. These improved when the two adenocarcinoma categories were merged into a single group to 80.2% and 84.0% respectively.

Mehta *et al.* used 35 serum samples from meningioma patients and compared them to 35 samples from healthy controls in an attempt to develop an approach to diagnose brain tumours using minimally invasive techniques [80]. Approximately

eight spectra were taken per sample, and the average of these was used for analysis. Employing PCA-LDA they achieved an accuracy of 86% for discriminating meningioma from healthy samples, which fell to 70% when the model was tested against an independent held out test set.

Breast Cancers

Koya *et al.* created Raman maps from *ex vivo* breast tissue and classified spectra as cancerous or healthy [77]. It is the largest study in terms of Raman spectra, though not samples; the Raman mapping methodology allowing them to take many spectra per sample. A CNN was used to classify the spectra. They used a technique called 'permutation importance' to interpret the CNN outputs and find which Raman bands were biologically significant.

Ma *et al.* also classified breast tissue using a CNN [79]. They compared its performance against four SVMs (each with a different kernel) and Fisher's Discriminant Analysis (FDA). Data augmentation was required to improve the CNN from the worst to the best performing model.

Zhang *et al.* interrogated five breast cancer and one healthy breast cell lines with RS [92], using PCA-DFA (Discriminant Factor Analysis) and PCA-SVM to classify spectra. The latter technique in particular was well able to separate healthy from cancerous cell lines with an accuracy of 99.0%. The team also performed a number of clinically relevant sub-group analyses and still achieved an accuracy of 93.9% with a four class model. Performance deteriorated as the sub-class divisions became more nuanced, representing comparisons between ever more biochemically homogenous samples.

Prostate Cancers

Lee *et al.* explored Raman spectra of extracellular vesicles derived from blood serum samples as a biomarker for prostate cancer in combination with a CNN [97].

This was compared to a PCA-LDA and PCA-QDA model. Additionally, analyses were performed on three wavenumber regions (full spectrum, fingerprint and high wavenumber regions), and with the data in its raw form as well as pre-processed. The CNN outperformed the traditional ML models across all subsets. The fingerprint region generally lent to better performance, though that was not ubiquitous across all subset analyses.

Chen *et al.* also used a CNN to classify prostate cancer, using SERS spectra taken from urine samples [70]. The model they used was LeNet-5, one of the pioneer CNN architectures [98], with promising results for such a non-invasive technique.

Gastrointestinal cancers

Wu *et al.* interogated biopsy samples taken during endoscopy, classifying spectra as normal, adenomatous polyps or adenocarcinomas [99]. They found a CNN comprehensively outperformed several traditional ML models. They also explored the difference that conducting analysis on pre-processed versus just normalisation data. Finally, the team performed CV via two methods, one splitting at the level of spectra, the other splitting at the level of subject/sample (there was one sample per subject so these coincide). The former method achieved an accuracy of 93.8%, falling to 81.3% with the latter split.

Ito *et al.* developed a boosted tree model from serum samples taken from suspected colorectal cancer patients, classifying into four categories; colorectal cancer, adenoma, hyper-plastic polyps and neuro-endocrine tumours, in a pairwise fashion [75]. They achieved 100% accuracy in all tasks, although they used the R^2 value as their assessment metric which gives a more nuanced idea of performance by accounting for how certain the model was in its classification, punishing predictions further from the class label. By this metric the boosted trees still performed exceptionally well. It is, however, unclear whether there was any validation/test set, and so these results may reflect the training performance.

Skin cancers

Serzhantov *et al.* used an ensemble of traditional ML models to classify skin tissue as cancerous or normal [85]. The models included were a classification and regression tree, SVM, k-nearest neighbours and logistic regression. Instead of selecting the best single model, the outputs of all models were used to create a soft voting classifier, allowing each model to 'vote' on an outcome, and the consensus across all models was taken. Splitting the data 50/50 into train/test sets, this was repeated 1000 times to build a spread of estimates. This method achieved an accuracy of 90.5%

Baria *et al.* compared PCA-LDA and PCA-ANN for the task of classifying spectra taken from cultured cell lines to distinguish the skin melanoma genotypes NRAS, BRAF or neither mutation [67]. The LDA produced an accuracy of 92.7%, the ANN 96.7%.

Santos *et al.* classified skin samples with spectra from the high wavenumber region using PCA-LDA, distinguishing between melanoma and not-melanoma [83]. They achieved an accuracy of 62.5%. The classification model was used in a unique way: they took the LDA score outputs and, instead of setting a typical limit of 0.5 as the delineation score between melanoma or not, chose a criteria of any two spectra from a single sample having a score greater than 0.35, or any single spectrum having a score greater than 0.8.

Gynaecological cancers

Daniel *et al.* compared a PCA-LDA to a PCA-ANN model in classifying cervical tissue as healthy, neoplastic or malignant [72]. In addition, those samples determined to be malignant were then subject to another LDA model to determine whether the samples were well, moderately or poorly differentiated. The PCA-LDA model achieved an accuracy of 95.3% compared to 99.0% for the PCA-ANN model. To help determine the biochemistry that characterised the three classes, non-negative

least squares (NNLS) was used to fit eleven known biochemical signatures to the spectra. This provides a multivariate method of determining sample biochemistry compared to the usual univariate peak assignment method.

Chen *et al.* used RS on serum to classify ovarian samples as normal, cystic or cancerous in a two step binary classification regime [69]. The first step used an ANN to determine abnormal from healthy samples, using an ensemble method to select the best model architecture, with an accuracy of 94.8%. Abnormal samples were then entered into another ANN to determine whether they were cystic or cancerous, achieving an overall accuracy across the three classes of 86.2%.

Other Cancers

He *et al.* interrogated *ex vivo* renal tissue seeking to identify cancerous tissue and demarc surgical boundaries as well as classify those tissues [74]. Although 100 spectra were obtained per sample, only 30 were used for classification after saturated spectra were removed. They used a suite of ML models, with a SVM (RBF) model marginally outperforming an ANN, while distinguishing between cancerous, normal and fat tissues with 92.89% accuracy, with only slightly lower performance when classifying cancerous sub-types.

Fang *et al.* used SERS to classify numerous cancerous and healthy cell lines [100], comparing ResNet to a custom CNN architecture and PCA-KNN. ResNet achieved 100% accuracy on an 11 class classification problem and was able to perform equally well on pre-processed and just standardised data.

1.5.3 Literature review: discussion

There currently exists no explicit standard for conducting and reporting clinical applications of machine learning, meaning the reviewed literature was not consistent regarding what was reported. This reinforces the concerns underpinning the afore mentioned reproducibility crisis. Due to the heterogeneity between these studies

this has been a qualitative review, making it difficult to draw definitive conclusions. There is some evidence suggesting deep learning can advance the field. As discussed by Blake *et al.* in a more concise exploration of the very same literature review [101], there are many reasons to suspect that the generally high accuracies found here will not translate to the clinical setting including small sample sizes, optimistic sampling and validation strategies, as well as differences in the data generating process itself, including variations in local practice for obtaining and treating samples, of preparing the sample for RS, and the instrumentation itself. I shall frequently return to this literature review throughout this thesis, under the moniker 'Lessons from the Literature', exploring certain topics in detail as they become relevant. This will help motivate and justify certain methodological decisions that need to be made at various points in the development of a robust ML model and training regime to meet the objectives of this thesis.

1.6 Thesis objectives and structure

The aim of this project is to develop a modern ML model and training regime that can be deployed towards the unique constraints and considerations of classifying Raman spectra of potentially cancerous tissues. Three RS datasets taken from human tissue will be used during the development of a learning pipeline. This includes two smaller datasets of colorectal and ovarian tissue, as well as a larger multi-centre dataset of oesophageal tissue taken with the same model of spectrometer across three separate sites. Results from each dataset will be explored, but given the small sample sizes of the first two, these will be regarded as pilot studies in their own right and will also be used to help guide the process of learning on the larger oesophageal dataset. Great detail is given to the numerous methodological considerations which are required to make any results robust enough to be clinically relevant.

To this end, a number of specific objectives manifest:

1. thorough consideration and selection of medically relevant performance metrics

2. appropriate measures of variance of those metrics

3. the selection of hyperparameters, including RS specific pre-processing techniques
4. how to account for the hierarchical structure common to medical datasets
5. explore cross-validation techniques
6. dealing with imbalanced datasets
7. data augmentation for deep learning
8. predict the presence of various cancers classes
9. post analysis interpretation of the results

In chapter 2, the datasets will be described, introducing each particular clinical challenge, the experimental set-up and a description of the data. Chapter 3 will consider the determinants of a medically relevant performance metric in ML, considering these particular datasets. Chapter 4 then describes the main ML models used in this thesis: PCA-LDA, SVM and CNN. Chapter 5 will then explore some of the numerous data preparations required before a thorough analysis can begin. Although presented in a linear fashion for ease of reading, this has necessarily been an iterative process. The results are given in chapter 6 and the post classification analysis in chapter 7 will attempt to relate biochemical changes to the results.

Chapter 2

Description of Datasets

> *“Above all else, show the data”*
>
> Edward R. Tufte

2.1 Introduction

It is impossible to consider ML models as distinct from the data on which they are trained. This is what it means to be a data driven process. Therefore, understanding the downstream analyses starts with understanding the data itself.

This chapter describes the 3 datasets explored in this thesis, including their biomedical rationale, the experimental conditions for data collection - including Raman spectrometer parameters - and an exploratory look at the data. This exploration involves Principal Component Analysis, which is explained in section 4.3. Before considering the individual datasets, there are a number of shared conditions that will first be explained.

2.1.1 Raman system

The same Raman system was used to collect all three datasets: the Renishaw prototype RA816 series biological analyser benchtop Raman system (Renishaw plc, Wotton-under-edge, UK) . This is configured for pathology use with a 785nm laser excitation, a 50x NA 0.8 objective, a 1500 $lines/mm$ grating and a motorised XYZ stage. These systems were configured to have a spectral range of $100-3100cm^{-1}$,

spectral resolution of $2cm^{-1}$ and a step size of $10\mu m$. The system is equipped with transmitted and reflected white light imaging at a variety of magnification levels down to a field view of $30x20mm$ for sample region of interest location. The system performs an automated calibration and optimisation sequence prior to performing measurements, including:

- Automatic adjustment of stage height to ensure the test samples are in focus.
- Slit lateral offset adjusted to maximise signal (silicon standard).
- Calibration of the spectrum x-axis in absolute wavenumber (using internal neon source) and in Raman shift (silicon standard).
- Main spectral properties (signal, bandwidth, asymmetry) are tested on silicon standard.
- Repeatability and reproducibility of response and wavenumber calibrations can be performance qualification (PQ) tested using a standard internal sample of polystyrene.

The spot diameter (or laser line) and step size together give an indication of the spatial resolution of acquired spectra. The Renishaw 816 spectrometer has a laser line measuring approximately $1\mu m$ by $80\mu m$. Data was collected in $StreamLine^{TM}$ mode with a binning of 10 and a step size of $10\mu m$. This results in an approximate spatial resolution of $10\mu m$ in the direction along the laser line, fully sampled, and near to $1\mu m$ orthogonal to the laser line, which means that data is undersampled in this direction. This undersampling was justified by virtue of the regions of interest identified by the histopathologist being homogenous in disease class. The typical human cell is approximately between $10\mu m$ and $100\mu m$, depending on the tissue type, indicating that the instrument gives spatial information at approximately the level of the cell. The optimal degree of sample coverage is necessarily a balance between high spatial fidelity and practical acquisition times. With the above parameters, a 15 by 15 spectra map took approximately 1.5 hours to acquire.

2.1.2 Slide substrates

For all datasets, tissue undergoing Raman spectroscopy was mounted onto 304L super mirror stainless steel slides. For tissues, these have been shown to improve Raman signal acquisition by up to a factor of four and to reduce the background signal compared to calcium fluoride (CaF2), the standard substrate often used in RS [102, 103]. This improvement is attributed to a double pass effect of the laser as it passes through the tissue twice, once back-reflected from the incident laser, and again retro-reflected from the mirrored steel. In addition steel slides are far cheaper.

Before any tissue was mounted onto the steel slides, they were cleaned by sonication in trichloroethylene for 30 minutes, followed by acetone for 30 minutes then isopropanol for a further 30 minutes and were then dried under a stream of nitrogen and stored at room temperature.

2.1.3 Tissue processing

All tissues were received as FFPE blocks. These were were manually sectioned using a Leica RM 2235 microtome (Leica Biosystems Ltd., UK) producing paraffin ribbons of adjacent sections of 8µm or 3µm thickness. These were floated onto a 45°C water bath and the 8µm ribbons mounted onto 304L super mirror stainless steel slides (Renishaw PLC, UK) and the 3µm ribbons onto conventional glass microscopy slides for those samples destined for RS and H&E staining respectively. The mounted steel slides were incubated at 37°C for 24 hours. The H&E slides were subject to standard automated staining and cover-slipping.

Prior to Raman data collection the paraffin in which the mounted tissues are embedded needs removing. This was done by immersing the mounted steel slides in four successive ten minute baths in xylene (VWR International Ltd., UK) with gentle agitation. A series of rehydration steps in graded ethanol absolute (VWR International Ltd., UK) took place via two sequential immersions in each of 100%, 90%, 70% and 50% ethanol baths for five minutes each, followed by a final immersion in distilled water for ten minutes.

2.2 Colorectal cancer, microsatellite instability and Lynch Syndrome

2.2.1 Lynch data: background

Colorectal cancer (CRC) encompasses cancers of the large colon and rectum. It is one of the major malignancies of the world, being the third most commonly occurring and the second most deadly cancer, with an estimated 1.8 million new cases and 881,000 deaths worldwide in 2018 [104]. With a few exceptions, CRC incidence is increasing globally, particularly in developing countries where shifting dietary and lifestyle factors are likely driving an increase in early onset CRC [105].

There are several known pathological pathways leading to CRC, resulting in heterogeneous presentations, therapies and outcomes. One such pathway is DNA mismatch repair deficient (dMMR) CRC, in which there are pathological alterations to any of a number of MMR genes (MLH1, MSH2, MSH6 or PMS2). This loss of MMR function causes high level microsatellite instability (MSI-H), characterised by mononucleotide, dinucleotide and trinucleotide tandem repeats. This can occur sporadically or as an inherited trait, as in Lynch Syndrome (LS). Hence, the detection of MSI-H is recommended in every case of CRC to screen for LS [106]. The high mutational burden seen in MSI-H tumours also has implications for treatment, providing potential targets for immunotherapy such as immune checkpoint inhibitors [107].

Despite recommendations for universal testing for MSI-H in all CRC cases, resource limitations mean that this cannot always happen and is particularly poor for young adults [108]. Testing for dMMR/MSI-H typically involves either immunohistochemistry (IHC) of the mismatch repair proteins or PCR amplification of consensus microsatellite repeats. Recent developments in ML have led to the possibility of exploiting morphological information in standard H&E slides [109, 107]. Such digital pathology techniques require few additional resources and have proven highly accurate in high quality, curated datasets. However, consistent with other domains using modern ML, these promising results do not generalise well when applied

to settings or cohorts outside of the narrow context in which they were developed [109]. The literature has thus far focused on H&E stained slides but ML applied to other histological stains, such as IHC, or digital staining via Raman spectroscopy may yield further improvements by exploiting molecular level information. Such techniques need not outperform the current diagnostic gold standards of MMR-IHC or MSI testing to be clinically useful as the screening process could benefit from even marginal improvements, particularly with regard to detecting LS.

There are no histological features specific to LS samples which distinguish them from adenocarcinoma (AC), but they are often poorly differentiated with excess mucus, signet-cell features, medullary growth patterns, an inflammatory reaction and high lymphatic infiltration [106]. Because of the strong genetic component of the disease screening has traditionally focused on family history. There are several guidelines to screen for LS in CRC patients. The earliest is the Amsterdam I criteria which has a sensitivity and specificity of 61% and 67%, respectively. The Amsterdam II criteria improved this to 72% and 78%. The more recent Bethesda guidelines have a sensitivity of 94% and specificity of 25% [110]. If a case meets the criteria outlined in these guidelines, then the tumour will be assessed for the presence of LS. This is done either by MMR immunohistochemistry (IHC) testing or micro-satellite instability (MSI) testing, which have sensitivities of 88-100% and 73-100% respectively and specificities of 68-84% and 78-98% [106].

A dataset was collected to explore the potential of RS to discriminate between normal, microsatellite stable adenocarcinoma (MSS AC) and MSI-H AC in human tissue, particularly the latter two disease classes which present the more pressing clinical challenge. This data will be referred to as the Lynch dataset.

2.2.2 Lynch data: sample collection

FFPE human colonic tissue blocks were obtained from the UCL/UCLH Biobank for Health and Disease under REC:15/YH/0311. 10 FFPE samples of resection margins of normal colonic mucosa from sporadic CRC cases were obtained along with 10 MSS/MMR proficient samples from the same patients. 10 archival MSI-H samples were also obtained and matched to the sporadic AC samples by TNM

(Tumour/Node/Metastasis) stage, making for a total of 30 samples across 3 classes taken from 20 patients. All these blocks have an undetermined fixation time but standard practice requires that all biopsies undergo 4-6 hours of fixation, while resections undergo 24-48 hours of fixation. Tissue was mounted as described in section 2.1.3. The H&E slides were re-analysed by a registered consultant pathologist (Dr Manuel Rodriguez-Justo) to confirm sample pathology. The sample characteristics can be found in appendix B.

Point spectra were acquired using the Renishaw RA816 system (Renishaw plc, Wotton-under-edge, UK) described in section 2.1.1. A total laser intensity of approximately 158 mW was focused onto samples through a 50x NA 0.8 objective. A 1500 l/mm grating was used to disperse the light providing a spectral range of 0 to 2100 cm^{-1} in the low wavenumber range. An integration time of 20 seconds was used for all measurements. A total of 50 individual spectra were collected from each tissue sample, except for one sample with only 40 spectra, resulting in a total of 1490 spectra across the 3 classes. All spectra were acquired from the glandular mucosal region in normal samples and from confirmed cancerous regions in all cancer samples, located by the resident pathologist prior to Raman measurement.

2.2.3 Lynch data: description

For analysis, the spectral range was truncated to 400-1800 cm^{-1}. Figure 2.1 shows the average spectrum for the three classes: normal tissue, MSI-H and MSS/MMR proficient. These show some differences between classes. However, the side panels show the variation of the classes, which obscures these subtle differences. A PCA score plot (figure 2.2) also suggests that the three classes significantly overlap, although the normal class is easier to distinguish from the others. This is true for all pairwise plots of other PCs, though only PC1 and PC2 are shown, which account for 77.8% and 10.5% of the variation in the data respectively.

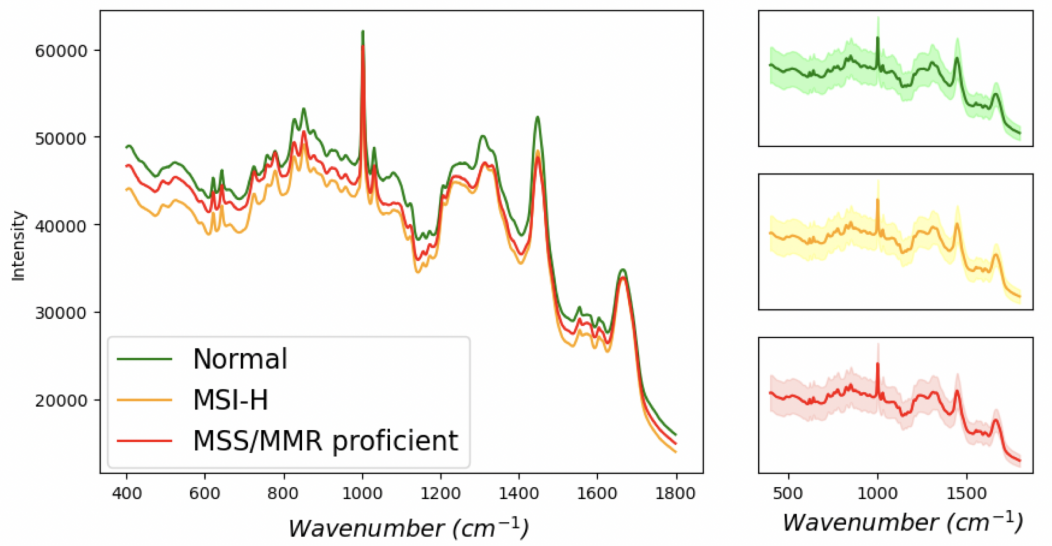


Figure 2.1: Lynch data: average spectrum by class. Right panels indicate 1 standard deviation. Spectra unprocessed other than cosmic ray removal.

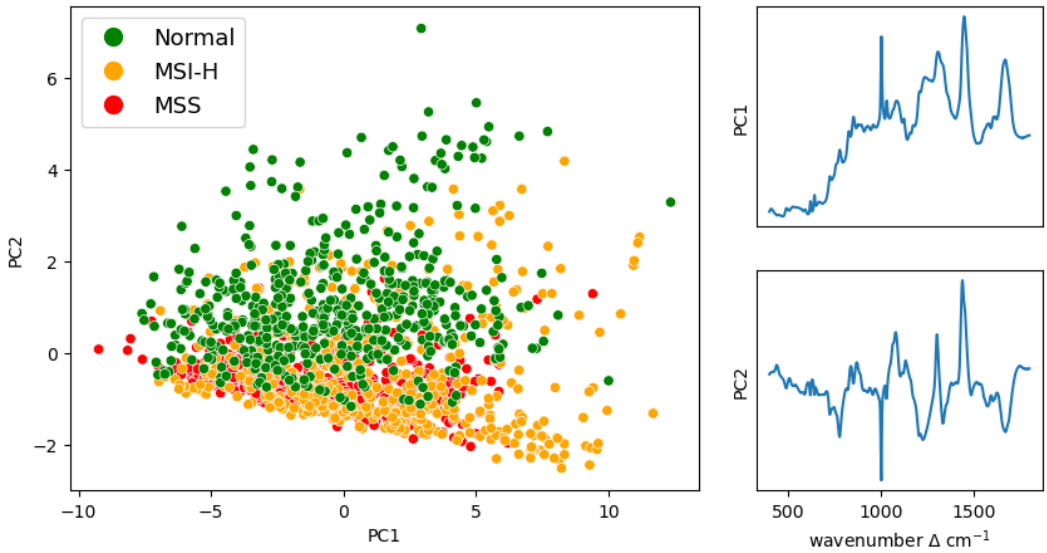


Figure 2.2: Lynch data: Score plot of the first two PCs with corresponding loading plots measured in arbitrary units

The SNR of this dataset was measured as described in section 1.3.6, and found to have a mean (+/- 1 SD) $SNR = 86.6 + / - 18.6$, indicating a good overall quality of the acquired spectra. This dataset will hereafter be referred to as the Lynch dataset.

2.3 Detecting post surgical debulking of ovarian tissue

2.3.1 Ovarian data: background

Ovarian cancer is the seventh most common malignancy in the world and the most lethal of the gynaecological cancers. It is predicted to increase in many areas of the world [111]. With no effective screening strategy most malignancies present with widespread intra-abdominal disease. Primary debulking surgery followed by chemotherapy is the standard treatment. One of the most important prognostic factors for this regimen is the presence of residual disease after surgery; patients with more than 0.1 cm of residual disease have substantially lower survival rates [112]. The reasons why a complete resection of all cancerous tissue is not always possible are complex. Among them are tumour-biology factors. If these could be identified then more appropriate treatments options could be explored and inefficacious surgery avoided. Many early attempts were made to this end including radiological imaging, laparoscopic inspection, physical examination and serum CA-125 levels but none have yet proved reliable enough for clinical applications [113]. More recently genetic markers have been explored, but these too have failed to yield reliable results [114]. Of those patients with residual disease, in about 20% of cases this is due to medical reasons, for instance comorbidities significant enough to limit time in surgery, while for the remaining 80% it is due to tumour-biology [114].

Ovarian cancers are known to be morphologically heterogenous, with features such as polyploidy, increased frequency of genome duplication, immune cell infiltration and subclonality all being identified as discriminative of tumour sub-types [115]. Such differences have biochemical manifestations that may be visible to RS.

The purpose of this dataset is an exploratory analysis to assess the potential of RS to help determine which samples are predictive of leaving residual disease post surgical debulking versus those which leave no residual disease. Such a determination would be of great clinical benefit if it could reliably identify which patients would be unlikely to benefit from surgery. Due to the presence of the afore mentioned medical

reasons surgery may be unsuccessful, the best performance of a classifier is lower than it would otherwise be if tumour-biology alone was a factor. For this clinical problem a theoretical upper limit of an oracle is AUC = 0.83 (this metric is explained in section 3.3).

2.3.2 Ovarian data: sample collection

The samples collected were a subset of AGO-OVAR11 trial in which subjects gave informed consent for their tissues to be used in future research [114]. This trial is the German contribution to the ICON 7 multi-centre trial, a phase three clinical trial of ovarian cancer treatment regimes [116]. Eighteen samples were collected from eighteen subjects. All subjects had stage III or IV high-grade serous ovarian cancer. The presence of residual disease was determined by the treating surgeon at the end of surgery using International Federation of Gynecology and Obstetrics guidelines [117].

Data was collected with the system described in section 2.1.1. Raman maps were obtained of regions of interest identified by the collaborating histopathologist (Florian Heintz) using the spectrometers $StreamLine^{TM}$ mode over the 'fingerprint' region of 400-1800 cm^{-1} (so called as this region contains many biochemically pertinent Raman peaks). A 1.0 mm step size was used for all maps with an integration time of 15 seconds. Within the identified region maps of 15 x 15 pixels were taken, giving 225 spectra per map. This was supplemented with 30 point spectra taken randomly from the region of interest, but not the mapped region, for a total of 255 spectra per subject and 4590 spectra overall. Saturated spectra were removed from the dataset, defined as any spectrum with 20 or more contiguous wavenumbers reading zero, leaving 4342 for analysis, 2122 with no residual disease and 2220 with residual disease.

Data was acquired from each patient alternating between the two classes. This is pertinent as the high precision nature of Raman spectrometers means that its various components can 'drift' over time, potentially introducing a time dependent component to the acquired spectra. If this time dependence is systematically introduced into the data acquisition process, then subsequent ML analysis could pick up this artefact

rather than learn real clinical feature. This could happen, for instance, if all samples of one disease class were first acquired, followed by all samples of another class, effectively subjecting the different classes to different treatments. This was mitigated by alternating samples between classes during acquistion.

2.3.3 Ovarian data: description

Figure 2.3 shows the average spectra for the two classes, residual disease and no residual disease, showing considerable variance within the classes.

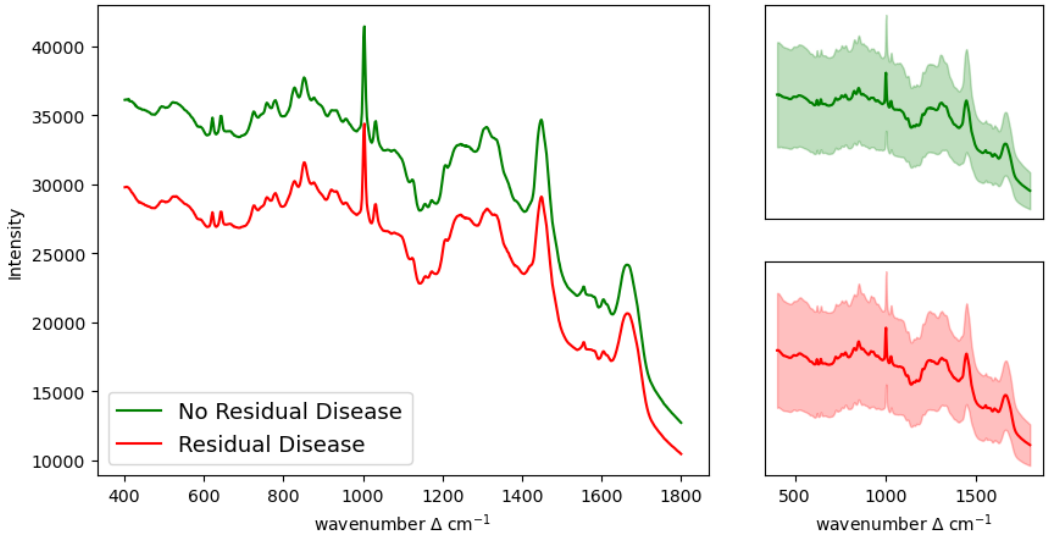


Figure 2.3: Ovarian data: average spectrum by class. Right panels indicate 1 standard deviation. Spectra unprocessed other than cosmic ray removal and saturated spectra removal.

This is further corroborated by considering the PCA score plots (figure 2.4) which shows considerable overlap between the classes in the first 2 PCs. This lack of obvious separability was consistent for all pairwise comparisons of PCs up to PC15.

This dataset has a mean (+/- 1 SD) $SNR = 54.4 + / - 19.8$, and will hereafter be referred to as the Ovarian dataset.

2.4 Oesophageal cancer and system transferability

2.4.1 SMART data: background

Oesophageal cancer is the sixth most common cause of cancer deaths globally. It has a five-year survival rate of less than 20% [118]. There are two distinct

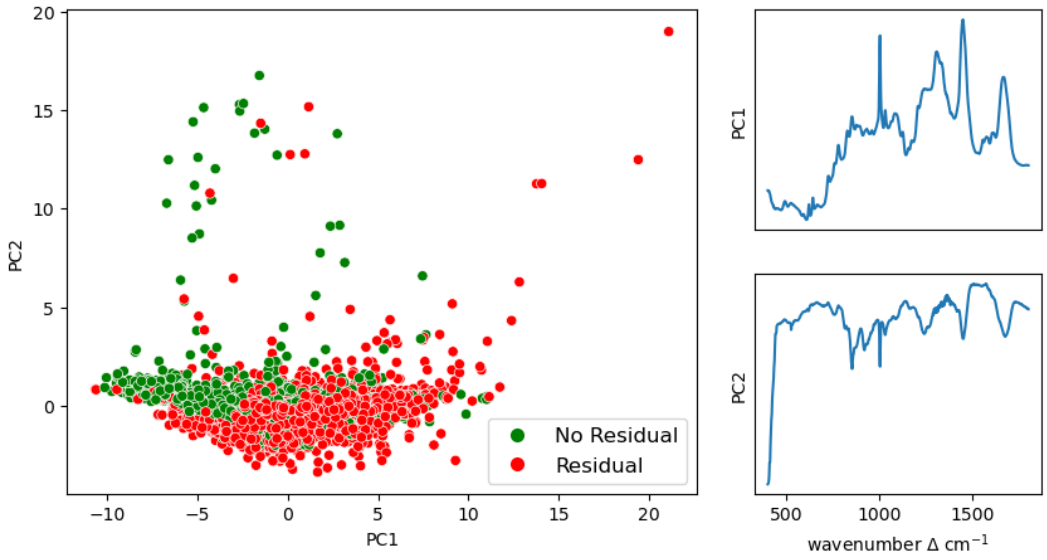


Figure 2.4: Ovarian data: score plot of PC1 and PC2 with corresponding loading plots in right sided panels.

pathologies: squamous cell carcinoma and adenocarcinoma (AC). While the former is more prevalent worldwide, it is declining. The latter is more common in Western countries and is increasing globally. Treatment regimes are often extensive, including chemotherapy, chemo-radiotherapy and/or surgical resection. The high mortality rate is in part due to late diagnosis, and methods which can expedite detection are a high priority [118]. To this end, the applicability of RS to the oesophageal cancer setting has been explored, establishing that RS can distinguish differences in clinical samples [119]. But despite reported sensitivities and specificities of 0.91 (95% CI, 0.89-0.93) and 0.92 (95% CI, 0.91-0.94), the technology has yet to translate to the clinical setting.

One unanswered issue that is impeding adoption is that of system transferability: the ability of diagnostic ML models built on one spectrometer to be reliably transferred to data taken on a different spectrometer. Instrumental artefacts, environmental differences and workplace practices can all systematically interfere with Raman signals, rendering models built on one spectrometer inapplicable to data taken on another spectrometer. As a result a centre developing diagnostic RS would need to build its own model from data collected on its own instrument. Individual spectrometers will produce a spectrum which contains a true Raman signal and an

instrument response function (IRF). This latter can be affected by differences in parts, ageing of parts and sources, and a gradual change in quantum efficiency. These factors can be ameliorated by pre-calibrating prior to sample measurement using calibration standards. Additionally, temperature fluctuations can cause thermal expansion, leading to the misalignment of optical components and the shifting of spectral peaks along the wavelength axis [120]. Such factors have been shown to deteriorate the SNR, wavenumber axis shifting, peak width and peak ratios over a number of common substances with well characterised Raman spectra [121]. A number of methods attempt to correct these artefacts: single wavelength standardisation, direct standardisation and piecewise direct standardisation. These attempt to remove instrumental affects by using a subset of samples measured on one spectrometer to regress the same data on another spectrometer. Other options include wavenumber offset correction, instrument response correction and baseline fluorescence correction. Isabelle *et al.* assessed these latter factors on an oesophageal RS dataset taken across 3 sites, each with a Renishaw RA816 series benchtop spectrometer [122]. A binary classifier was trained on one spectrometer to distinguish between two oesophageal pathologies and performance tested on the other two sites, both with and without corrections. These methods improved sensitivity from 86% and 73% to 96% and 79% across the two test sites respectively, but the corresponding specificity dropped from 84% and 85% to 77% and 81%. This was despite showing that the corrections had the desired effect of aligning wavenumber axes and reducing class variance. The problem of system transferability has proven stubborn, leading to calls for default spectrometer calibration and extensive data sharing between researchers to create large RS datasets [121].

Another possible solution has recently developed with the advent of deep learning architectures such as CNNs. These models can be robust against many sources of noise, provided they are appropriately trained with a sufficient amount of data. For instance, MRI data also suffers from system transferability [123], but multi-site studies have trained deep learning models to improve the generalisability of the diagnosis of schizophrenia [124].

The work of Isabelle *et al.* was an early analysis of the Stratified Medicine through Advanced Raman Technologies (SMART) project. Since the projects inception many more samples have been collected with the purpose of exploring the transferability of RS as applied to oesophageal cancer. With this SMART dataset I seek to discriminate between five oesophageal pathologies, while investigating the ability of various models to reliably transfer results across multiple sites.

2.4.2 SMART data: sample collection

Samples were obtained from patients with a scheduled endoscopy for Barrett's surveillance or from patients who had surgery for oesophageal cancer. 66 FFPE samples were taken from the histopathology archive at Gloucester Hospital NHS Foundation Trust. These procedures were performed under local (endoscopic resection) or general (oesophageal resection) anaesthetic in accordance with an approved ethical proposal [Gloucestershire Local Research Ethics Committee]. At all times the General Medical Council (GMC) guidelines on good clinical practice were followed. Routine histopathology reports were used to assist with sample selection, identifying those with one of five clear histological pathologies: normal squamous (NSQ), intestinal metaplasia (IM), low and high grade dysplasia (LGD and HGD) and adenocarcinoma (AC). Tissue scetions were cut at $8\mu m$ and mounted onto stainless steel slides as described in section 2.1.3. Three such samples were collected from each of the 66 subjects, one sent to each of the three participating centres. Standard H&E slides of adjacent tissues were also taken to identify regions of interest and to confirm pathologies. One consultant and one registrar histopathologist used sections at one centre (GRH) to outline regions of interest and agree on a diagnosis, providing a two person consensus (although the robustness of such consensus could be called into question where there is one senior and one junior partner involved).

Each centre used the same make of spectrometer, described in detail in section 2.1.1. Data were collected in $StreamLine^{TM}$ mode over the fingerprint spectral region 400-1800 cm^{-1}, using a $10\mu m$ grid and an integrated exposure time of 6 seconds per point. Data were collected over pixel regions exhibiting a homogenous pathology, with map sizes varying from 11 x 18 to 75 x 93 pixels. All three centres

followed the same protocol for taking the Raman maps, as detailed in section 2.1.3 and the same system parameters were used. Regions of interest were matched between all centres to that identified by the histopathologists so that all three centres mapped approximately the same regions - though this will not be exact as the slides were taken from adjacent samples of the same tissue.

Unfortunately, the practicalities of organising a large multi-centre study meant that the data was not always collected as per protocol, potentially leaving a portion of questionable quality. This took the form of data which was not taken from the same identified region of interest. Such spectra may not have the same disease class as those identified in the region of interest. Additionally, as the focus of this study was to assess the transferability of models across instruments, it is important to isolate all other variables as much as possible, including the regions of samples used across centres. Hence, a mask was applied to remove such extraneous spectra and ensure all Raman mapped areas matched across centres. This was a manual process, and so was prone to human error. An exploratory analysis revealed that though most matched maps were of a similar size, as expected, a few maps remained significantly larger than the corresponding maps taken at the other centres. This was measured by taking the ratio of the number of spectra for the same map across the 3 centres in a pairwise fashion. Using a criterion of the same sample being no more than 25% different between centres (figure 2.5), a total of 15 samples were excluded from subsequent analysis, leaving a total of 61 samples from 51 patients per centre for the study. A preliminary analysis using a 5 x 3-fold CV procedure, defined in section 5.2, showed that this procedure increased PCA-LDA classification accuracy (+/- 1 SD) to 56.7% +/- 3.6% , from 52.6% +/- 3.7%, lending some credence to the process.

Having to process the data this way is not ideal, but neither is retaining unmatched data, thus the more conservative approach of excluding some samples was taken. Despite these efforts to only map areas with a homogenous pathology, the assignment of a single label to multiple spectra taken from the same map could induce label noise into this dataset. This is true of all Raman datasets, but is particularly pernicious in this dataset due to the cross centre comparison. This can occur because we

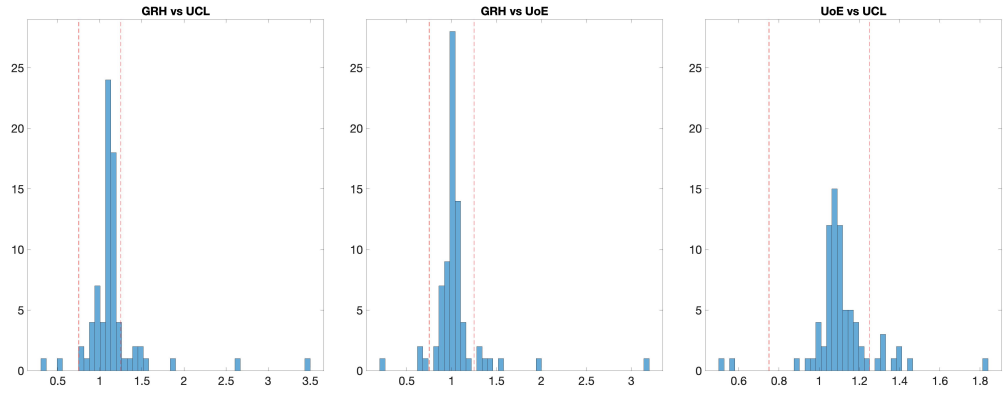


Figure 2.5: Pairwise ratio of spectra from the same samples. x-axis indicates the ratio, y axis the count. Beyond either of the red dashed lines indicates samples which differ between centres by more than 25%

have assumed that all spectra taken from a sample represent the same disease class, but it is possible that a single map could contain multiple pathologies. If multiple pathologies are present, the histopathologist will assign a single label based on the clinically worst present pathology. If we are classifying the whole image, this is an appropriate labelling method as the worst present pathology is the clinically relevant feature. However, if we are classifying individual spectra, assigning a single label to potentially heterogeneous spectra can lead to incorrect labels. This is a common problem in digital pathology in general, where many 'patches' of single whole slide images are segmented and used to train classifiers on the assumption that each patch will inherit the class label of the whole slide. This is referred to as weakly supervised classification, and has been shown to be surprisingly effective and far more practical compared to pixel-wise manual annotations [125].

2.4.3 SMART data: description

A total of 560819 spectra were taken from 61 samples across 51 patients. Figure 2.8 shows that these are not evenly distributed across the pathology classes, but form an imbalanced dataset. Such class imbalance requires careful considerations which are addressed in section 5.6.

2.4.3.1 Instrument correction

Guo *et al.* suggest that data taken from different instruments require some treatment for any meaningful comparison, if the inter-instrument variations are larger than the

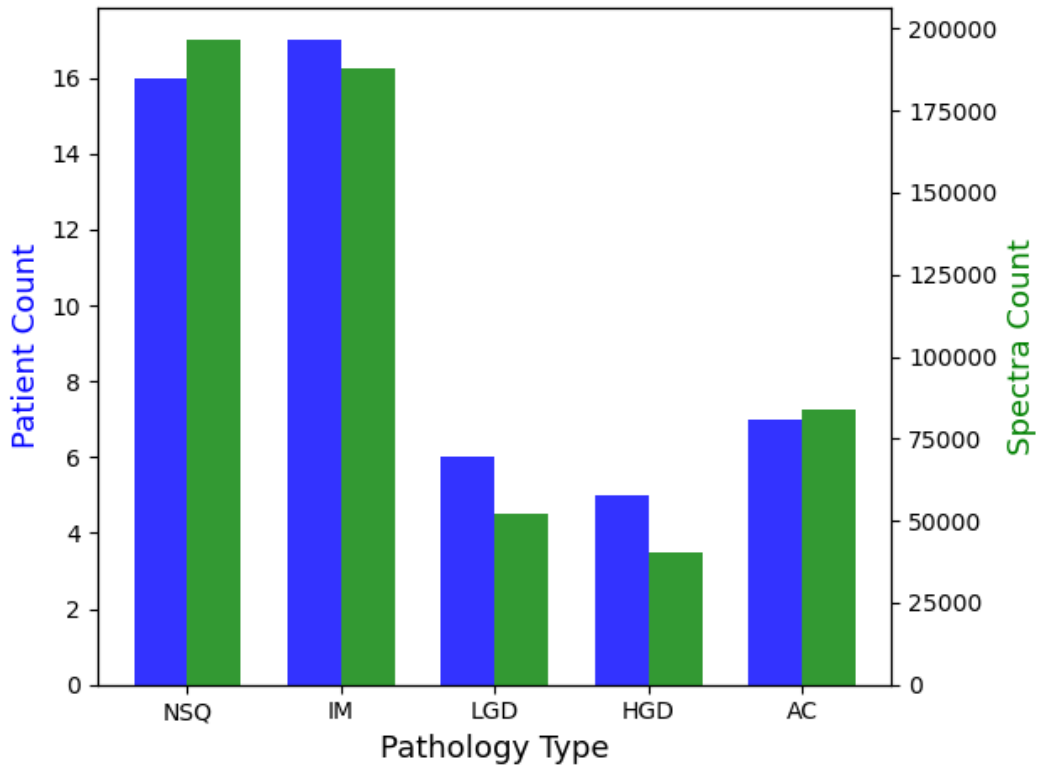
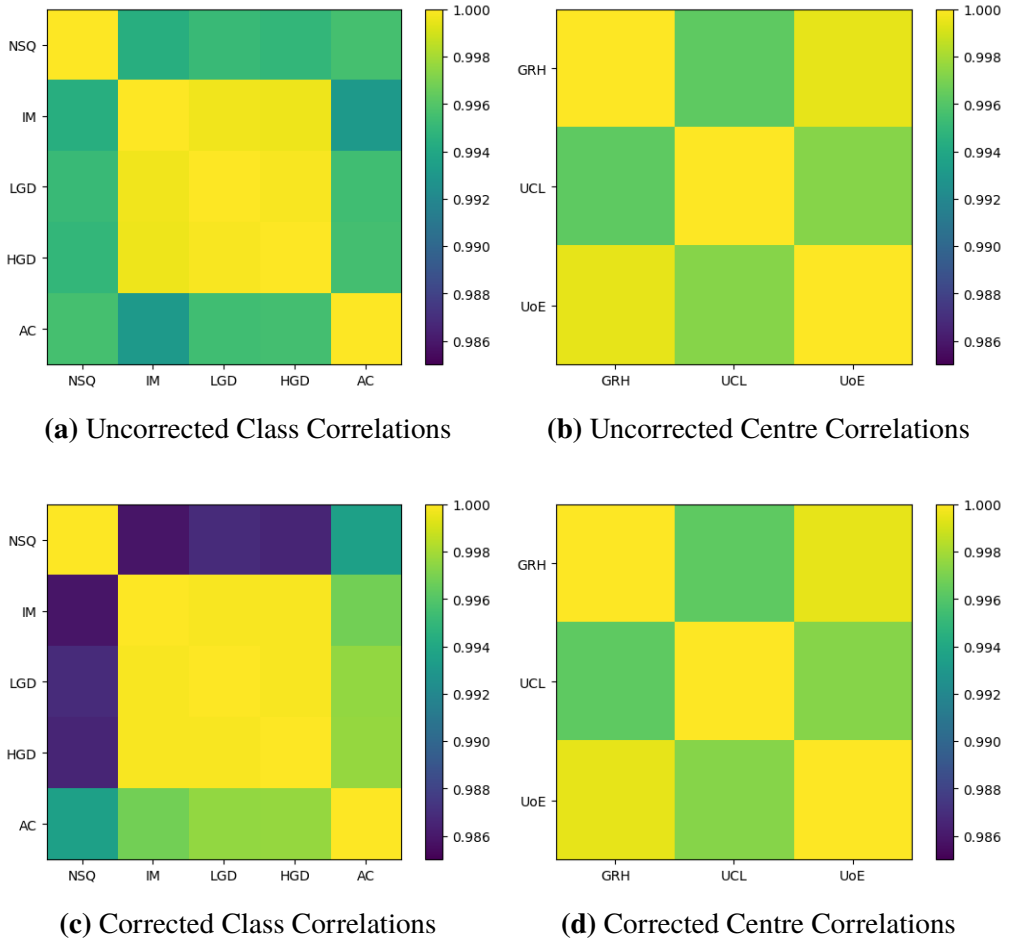


Figure 2.6: SMART data: Count of spectra and patients per pathology class

inter-group variations [126]. They name this process model transfer. Figures 2.7a and 2.7b show the correlations between the mean spectra of the five classes and the three centres respectively. Although the correlations between all mean spectra is high, the correlations between centres are generally higher than for classes, except for the intermediate classes of IM, LGD and HGD. The corresponding mean spectra are shown in figures 2.8a and 2.8b.



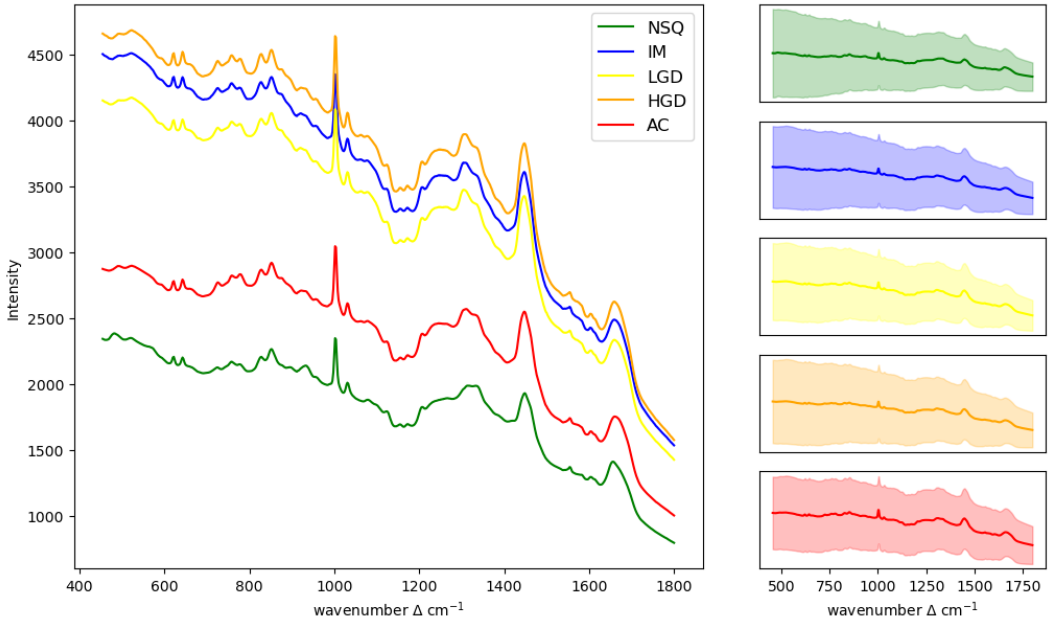
(a) Uncorrected Class Correlations **(b)** Uncorrected Centre Correlations

(c) Corrected Class Correlations **(d)** Corrected Centre Correlations

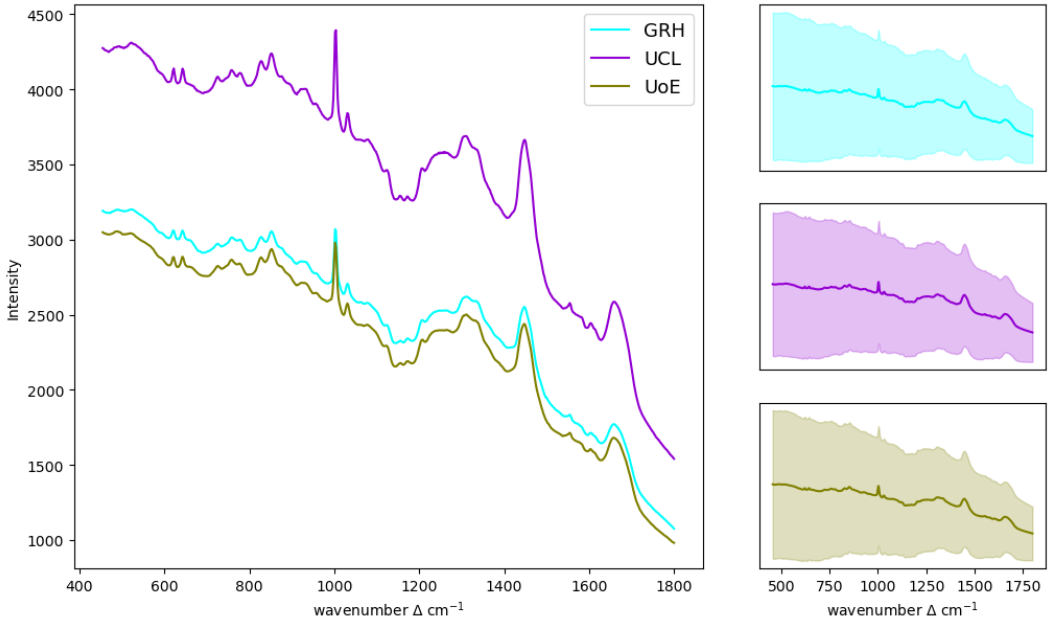
Figure 2.7: SMART data correlations of the mean spectra of the five classes and three centres for the original (uncorrected) and instrument corrected data

Instrument correction was applied to the dataset in order to assess whether this improved the separability of the data. In particular, extended multiplicative scatter correction (EMSC) and instrument response correction was performed. EMSC accounts for scaling differences and artefacts by using a non-tissue measured instrument spectrum for each centre, including the measured objective spectrum, and a 3^{rd} order polynomial baseline using a least-squares modelling procedure. Instrument response correction involved comparing the ratio of daily measured and calibrated spectra as measured by the NIST SRM 2241 standard reference material to provide an instrument response profile which was used to correct spectra. This correction was conducted using Renishaw's WiRE software. More details can be found in Isabelle *et al.* [122]. During this process the wavenumber axis was truncated

to $450-1800cm^{-1}$. The uncorrected data was similarly truncated to ensure a like for like comparison between the datasets.



(a) Mean spectra by class

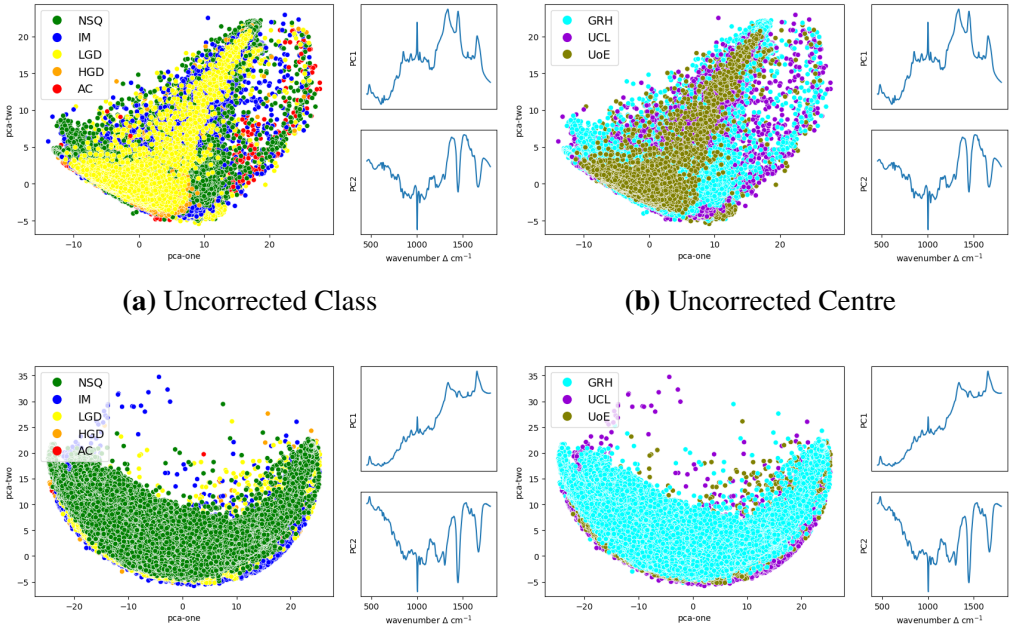


(b) Mean spectra by centre

Figure 2.8: Uncorrected SMART data: Mean spectra by class and by centre with plots showing 1 standard deviation on right hand side.

The correlation plots for the corrected data (figures 2.7c and 2.7d) show a marginal decrease in the correlation between NSQ and the three intermediary classes

IM, LGD and HGD, compared to the same plots for the uncorrected data. There is no appreciable change between the centres. Neither did correction result in a noticeable difference in separability in the PCA plots (figure 2.9).



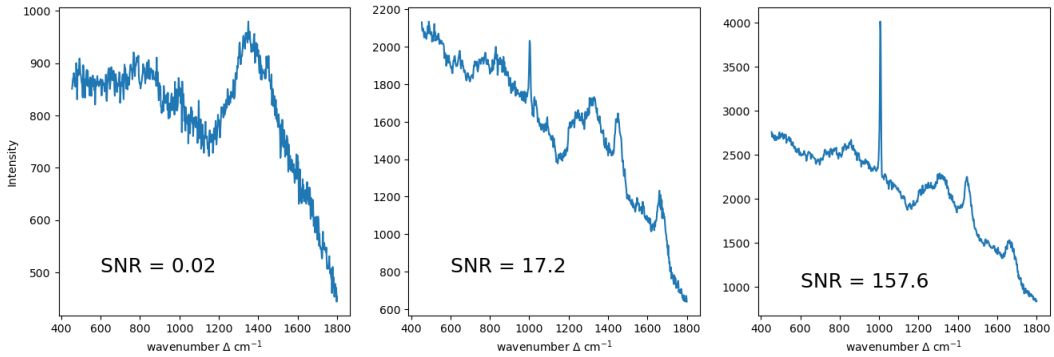
(a) Uncorrected Class **(b)** Uncorrected Centre

(c) Corrected Class **(d)** Corrected Centre

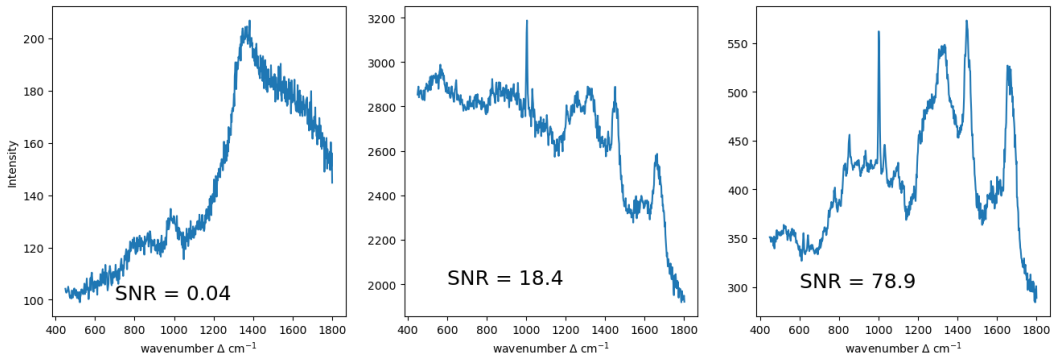
Figure 2.9: SMART data: PCA plots showing separation by Class and by Centre

These plots all serve to illustrate that any distinction between pathology classes is subtle. So too is the distinction between centres, hence it may still be a pertinent factor to consider during model building. Though there seems to be no improvement in discernment between the uncorrected and instrument corrected data, it may simply be too subtle to detect in these plots. Therefore, both datasets will still be explored and compared during learning.

One of the strengths of CNNs is that the convolutional layer encourages the model to be invariant to certain translations in the input data. For instance, in the well known case of an animal image classifier, the model should be invariant to the location, and orientation, of any animals in the images. A particular point of interest in examining these two SMART datasets is whether the CNN will be robust to any centre level artefacts due to this invariance property when compared to the more traditional ML models.



(a) SMART uncorrected data



(b) SMART corrected data

Figure 2.10: SMART data: example spectra at varying noise levels, top row uncorrected spectra, bottom row the same spectra corrected

2.5 Extracting lessons from the data

The Ovarian and Lynch datasets are extremely small in the context of deep learning. However, there are a number of techniques which can be employed which may make them amenable to deep learning. These techniques are thoroughly explored in section 5. Although the SMART dataset is small in the context of deep learning, it is large enough that the same rigorous optimisation that will be performed on the Lynch and Ovarian datasets is impractical due to computational limitations. Therefore, while exploring the two smaller datasets, attempts will be made to find principles that can be generalised across all three datasets, leading to a much smaller optimisation space to search for the SMART dataset. This optimisation involves the selection of appropriate model hyperparameters, which are explored in detail in section 4.

The relative SNR values of the datasets are listed in table 2.1. The SNR of the Ovarian dataset is lower than that for the Lynch dataset. This is likely due to the

spectral acquisition methods: the former being acquired principally as maps, the latter as point spectra. Maps typically have a lower SNR due to variations of the sample topology which can bring the microspectrometer slightly out of focus. As a consequence, the Ovarian dataset presents a more difficult classification task.

The SNR is particularly poor for the SMART dataset. This is barely improved by instrument correction. There is a great deal of variation within the SMART dataset, which accounts for the low mean SNR. It is possible to remove spectra below a certain SNR threshold in order to ensure the model only sees data of a certain quality. However, this would necessitate removing spectra below this threshold when deployed in clinical practice. As discussed in section 3, the most important factor is that the models are trained in a manner that reflects how they will be used in clinical practice. It may well be preferable to exclude spectra below a certain SNR, but this would then need to be embedded into clinical practice and what that threshold would be warrants a thorough investigation itself. As these quantities are not yet established I proceed without applying a SNR threshold.

Dataset	Number of Patients	Number of Samples	Number of Spectra	Number of Classes	SNR
Ovarian	18	18	4590	2	54.4 +/- 19.8
Lynch	20	30	1490	3	86.6 +/- 18.6
SMART (Original)	51	61	560819	5	18.3 +/- 8.1
SMART (Instrument Corrected)	51	61	560819	5	19.5 +/- 7.9

Table 2.1: Comparison of the datasets

In addition to being used as pathfinder datasets, the Lynch and Ovarian datasets will be assessed in their own right. While they are small datasets, this is not unusual in the context of ML in RS and medicine. They can be considered proof-of-concept studies, to be used to justify and inform larger studies. In addition to seeking the best performance for each, they will be used to guide decisions regarding the SMART data.

I will use the hyperparameters found in the Lynch and Ovarian datasets to guide the choice for the SMART dataset. The SMART dataset is two orders of magnitude larger than the Lynch or Ovarian datasets, meaning it is far more computationally expensive to explore the hyperparameter space. In addition, I do not want to overfit

to any particular dataset: even if we could perform an exhaustive hyperparameter search of the SMART dataset, it does not guarantee a good performance on the general population. It is possible to pick hyperparameters which result in the best performance for our particular dataset, but which would fail to generalise to the broader population of interest: the topic of the next section.

Chapter 3

Assessing Model Performance

> *“Measure what is measurable, and make measurable what is not so”*
>
> Galileo Galilei

3.1 Introduction

ML is a data driven process. It follows that the technique is sensitive to the data upon which it is trained. Any medical dataset is a sample from a greater population. The degree to which the sample is representative of the population of interest is pivotal if results derived from the sample are to generalise to the entire population. This is egregiously demonstrated with the case of racially biased healthcare ML models. This occurs when the datasets on which they were trained were racially homogenous compared to the population to which they were applied [127], or due to the poor selection of proxy biomarkers [128]. For every model, the pertinent question is how well does the model generalise to the population of interest. In this thesis, the population to which we wish to generalise our findings are Raman spectra acquired from tissues of patients who have had a particular tissue biopsy due to suspicious symptoms. For the dataset sets used in this thesis, no demographic data is available due to anonymisation and so I am unable to assess the extent to which the data is representative of the population. I will therefore proceed under the assumption that the samples are representative with the caveat that any optimistic results will not

necessarily generalise, and that any ML model developed for clinical practice must be subject to strict post-deployment surveillance.

In this section I discuss important components in measuring the generalisability of a model, factors which determine the suitability of any given metric and explain the particular metrics used to assess model performance which will hereafter be used.

3.2 Generalisation error

The generalisation error is any metric which seeks to measure how well a model will perform on new data (i.e. its generalisibility). For instance, in a medical context this could be an estimate of the expected accuracy of a technique when deployed on new patients. This estimate should be derived from data a ML model has not previously seen (i.e. not used to train the model). In the ideal case this will be an entirely new dataset. However, the process of collecting and annotating new data can be prohibitively expensive, particularly in medical settings. Therefore, to estimate the generalisation error it is common to split a dataset into segments, retaining a portion for training the model (the training set) and holding out a smaller portion to assess the performance of the model (the test set or hold-out set). This held-out data should not be used to construct the model in any way, as its role is to emulate the acquisition of new data. We hope that the performance on the test data is then indicative of performance on newly collected data.

However, there are many reasons why a model may not generalise. Overfitting is a particular risk in ML, where a model is so thoroughly trained on a subset of the data that it learns random fluctuations that just so happen to distinguish one class from another. These fluctuations are due entirely to chance, but the model will use these irrelevant features when classifying new data, thus inhibiting performance. In sections 4 and 5 many methods to reduce the possibility of overfitting are considered. The potential of over-fitting also needs to be considered when choosing how the data is split into training and test sets and is carefully analysed in sections 5.2 and 5.8. Medical data is often also hierarchical in nature, having several layers to the data's structure. For instance, many Raman spectra could be taken from a single

sample and many samples could be taken from a single patient. This structure necessitates the careful consideration of the correct level at which to split data, and is discussed in section 5.4. The correct preprocessing of training and test sets is also explored in section 5.3. Others reasons for erroneous generalisation errors fall into the category of the dataset itself being unrepresentative of the general population, usually occurring during sample collection.

Much deep learning research has been developed in the domain of image recognition - it was for this application that CNNs were developed. Due to the popularity of this task there are a number of large and thoroughly annotated benchmark datasets by which to train and test a model. However, the generalisability of even the best performing models trained on these large and well curated datasets has been called into question, when new datasets were created to which the models should be able to generalise. Accuracy rates dropped by 3-15% on the cifar-10 dataset and 11-14% on the imagenet dataset [129].

There does not seem to be any studies explicitly investigating if this persists in deep learning medical applications, perhaps as it is substantially harder to collate and annotate medical data. In lieu of research to the contrary, it is prudent to assume that medical data will at least be similarly afflicted - i.e. it is a feature of contemporary deep learning methods (from data collection through to analysis) rather than being unique to particular datasets.

3.2.1 Bayes irreducible error

Even a hypothetical perfect model, called an oracle, which knows the true probability distribution that generates the data will still incur some error in its predictions. This is because the generating process may be inherently stochastic (like Raman scattering, as discussed in section 1.3.5), and the generating process may involve variables not present in the data. Thus, the true distributions of classes in a population very likely overlap. The extent to which they overlap determines the best theoretical performance; the greater the degree of overlapping, the greater the error. The error incurred by the overlapping of true class distributions is known as the Bayes irreducible error.

An oracle can be understood in terms of a Bayes classifier. Consider a set of p

features: $\boldsymbol{x} = [x_1, ..., x_p]$, which we wish to map to one of C classes (e.g. assign a spectrum with p wavenumber values to one of C diseases). To construct a Bayes classifier we need to know the prior probabilities of the classes: $\mathbb{P}(1), ..., \mathbb{P}(C)$, and the class conditional probability density functions (pdfs, $p(\cdot)$): $p(\boldsymbol{x}|1), ..., p(\boldsymbol{x}|C)$. Specifically to medical RS, these can respectively be interpreted as the prior probabilities, or underlying proportion, of a disease class in a population and the probability distribution of $\boldsymbol{x}$, Raman spectra, given membership to a particular disease class. As the latter are distributions, there may well be overlap between classes. The Bayes classifier, α, is derived from Bayes rule and defined as $\alpha(\boldsymbol{x}) = arg \max_{c=1,...,C} p(\boldsymbol{x}|c)\mathbb{P}(c)$. The error associated with this classifier is the Bayes irreducible error. More details on the Bayes classifier and its error can be found in Tohka and Gils [130]. Of particular note is that to achieve this theoretical optimal performance, the prior class distributions need to be known, or estimated from the proportions of subjects with the disease classes in the setting of interest. This will become relevant when discussing class imbalance in section 5.6. However, in practice we never know the class conditional pdfs, $p(\boldsymbol{x}|c)$. Therefore the Bayes classifier is usually an unattainable gold standard, and the irreducible error an unknown quantity.

In lieu of this knowledge, it is common to assume that human domain expert performance represents the best possible performance. As was demonstrated in the literature review of inter-rater variability in CRC diagnosis in section 1.2 this can be problematic. Difficulties when constructing gold standards can occur when tasks involve ambiguous intermediary classes, such as is found in some cancer taxonomies, where certain morphological features, and sometimes biochemical and genetic features, have been found to correlate with clinically relevant outcomes. The correlations of these various features to clinical outcomes may be weaker or stronger, their causal pathways unclear, and their relevance regularly discussed amongst experts in the light of new evidence with consensus occasionally changing to exclude certain features or to include new ones.

As shown in section 2, all three datasets show significant overlap in their classes, at least when projected into low dimensional spaces (i.e. PCA space), suggesting a

degree of Bayes irreducible error. As discussed in section 2.3, this is particularly so for the Ovarian dataset due to contamination of tumour-biology specific and overall medical health related outcomes.

ML is a data-centric method of modelling. By taking different random subsets of a dataset we would achieve varying results. Thus we see that any estimate of model performance is itself a random variable (RV). In *frequentist* interpretations of probability, we may imagine that there is a true generalisation error associated with a particular model and population (not necessarily the same as the Bayes irreducible error as any real ML model is unlikely to be an oracle). Let that fixed but unknown error be θ. Our estimate of it is given by $\hat{\theta}$. We may then define two properties of our estimate: its bias and its variance.

3.2.2 Bias

The expectation, $\mathbb{E}$, of a RV is its asymptotic mean. It is usually estimated by a sample mean. The bias of a RV is defined as:

$$bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \tag{3.1}$$

If the expected generalisation error is equal to the true generalisation error, then $bias(\hat{\theta}) = 0$, and the estimate is said to be unbiased. This can be understood as a consistent difference between a parameter estimate and its true value. Thus, bias can be understood as a systematic error introduced by the model. This most commonly manifests when a simple ML model is unable to sufficiently follow a particularly complex data generating process (i.e. nature), as illustrated in figure 3.1a. All else being equal, we would prefer the bias to be as low as possible.

3.2.3 Variance

The variance of an estimate is a measure of how much, on average, it varies. In the context of the generalisation error this is expressed as:

$$Var(\hat{\theta}) \tag{3.2}$$

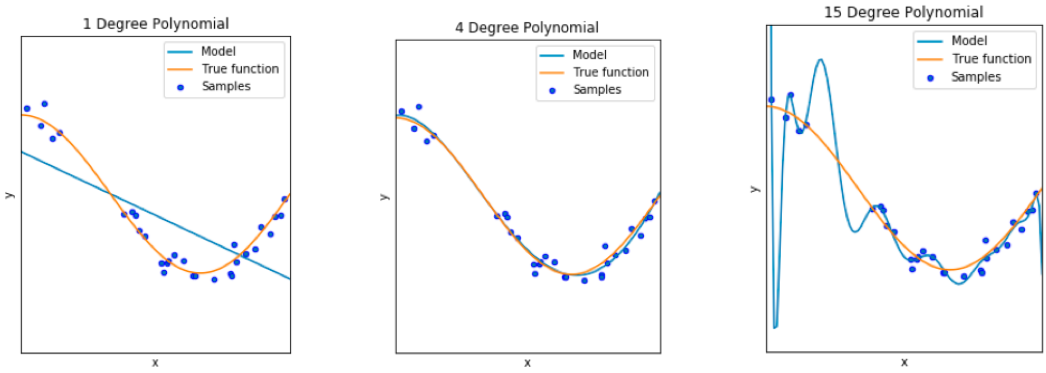
which measures how much we expect the error estimate to vary as a function of the training data: that is variations in different training sets will result in variations in the error estimation. All else being equal, models which are more complex (i.e. more parameters and/or non-linear) are more susceptible to having high variance as such models are more able to overfit to the training data (illustrated in figure 3.1c). The true variance is usually unknown, hence itself is estimated from the dataset. In order to derive an estimate of this variance, at least two measures of the generalisation error need to be made (i.e. the same model trained twice on two different training sets): the more such measures the more accurate will be the variance estimate.

The generalisation error balances bias against variance. Bias is a measure of how well the model fits the data: a high bias indicates under-fitting; it is not learning enough features to sufficiently discriminate the data (figure 3.1a). High variance indicates that over-fitting has occurred: the model has fit so well to the training data that it has also learned to utilise random features that just so happen to be present in the training set (figure 3.1c). Such overfitting leads to very high accuracies for the training set, but does not generalise well and the generalisation error will vary to reflect these random perturbations.

3.2.4 Learning curves

A learning curve is a plot showing the error rate (or sometimes the accuracy) as a function of training, particularly used when training DL models. Learning curves also sometimes represent the error rate as a function of model complexity. Figure 3.2 shows an idealised learning curve, displaying the trade off between bias and variance in terms of the influence of under-fitting and over-fitting. The model starts overly simplistic, unable to distinguish relevant features. At this point the model has high bias and low variance. As training proceeds, the model learns to identify relevant features, thus reducing the error until the gap between the training error and the test error is at its smallest. The model then begins to perform worse on the test data even as the training error improves as it begins to identify common random perturbations in the training data. This represents a model with low bias and high variance.

In figure 3.2 the x-axis shows 'capacity'. Informally, capacity is synonymous



(a) 1^{st} degree polynomial **(b)** 4^{th} degree polynomial **(c)** 15^{th} degree ploynomial

Figure 3.1: A representation of model fitting. Random data (blue points) were generated around a sinusoid (orange lines) and three different polynomial models (blue lines) were fitted to the data. 3.1a: An under-fitted model in which a straight line fails to capture the general features of the underlying function. Such a model has high bias. 3.1b: A model which well captures the data by approximating the general shape of the data while not trying to account for every nuance. 3.1c: An over-fit model: though it fits the training data extremely well it would give erroneous predictions on data drawn from the true function. Such a model has high variance.

with the complexity of a model. The figure shows that it is not simply the case that a model with more capacity will necessarily have the lower generalisation error. Section 4 will detail the models used in this thesis: for now it is sufficient to note they represent increasingly complex models. DL models are by definition extremely complex, and medical Raman datasets would theoretically benefit from such nuance. However, though DL has been heralded as the future of medical modelling, this section serves to illustrate that such models are prone to high variance and overfitting. Hence, it may be that simpler models may still outperform more complex models even though they may not capture the nuances of the data generating process. By Occam's razor, we may then prefer the simplest model for the task. The discernment between these two types of generalisation error highlights a central tension within this thesis. The total generalisation error is the sum of bias and variance and it is unclear which source of error dominates in the oncology datasets under study. Without knowing the true data generating process it is not possible to decompose these sources of noise in real data, hence an empirical approach is taken to compare model performance bearing in mind these theoretical considerations.

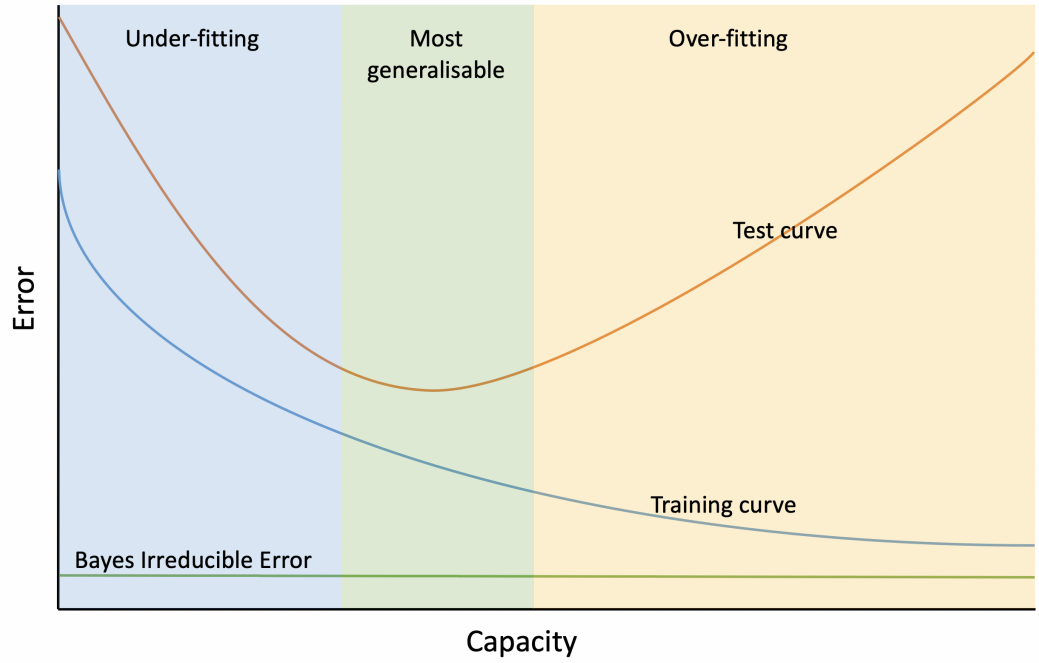


Figure 3.2: Learning Curve: bias and variance trade-off. Green line - Bayes Irreducible Error. Blue line - training error. Orange line - test error

This has necessarily been a brief exploration of bias-variance trade-off. More detail can be found in James *et al.* [131] which has formed the basis of this section. Armed with an understanding of what we wish to measure, we now turn our attention to how best to measure it for the given tasks.

3.3 Performance metrics

The generalisation error can be quantified by many performance metrics. Here, we focus on those most common in the medical literature, and consider some less popular metrics which may have some desirable qualities. All have in common that they compare a models classification, or prediction, of a sample against its true label, thus estimating the generalisation error. However, not all mistakes are equal and each metric balances certain strengths and weaknesses.

3.3.1 Accuracy

Perhaps the most intuitive, and very common, performance metric is the accuracy. It is simply the proportion of correct classifications over all classifications.

As will be discussed in section 5.6, accuracy becomes an increasingly poor metric as class imbalance increases. Even with balanced datasets (i.e. the same number of samples per class), accuracy hides a number of subtleties in the performance of a model which are revealed when we consider other performance metrics.

3.3.2 Confusion matrix and related metrics

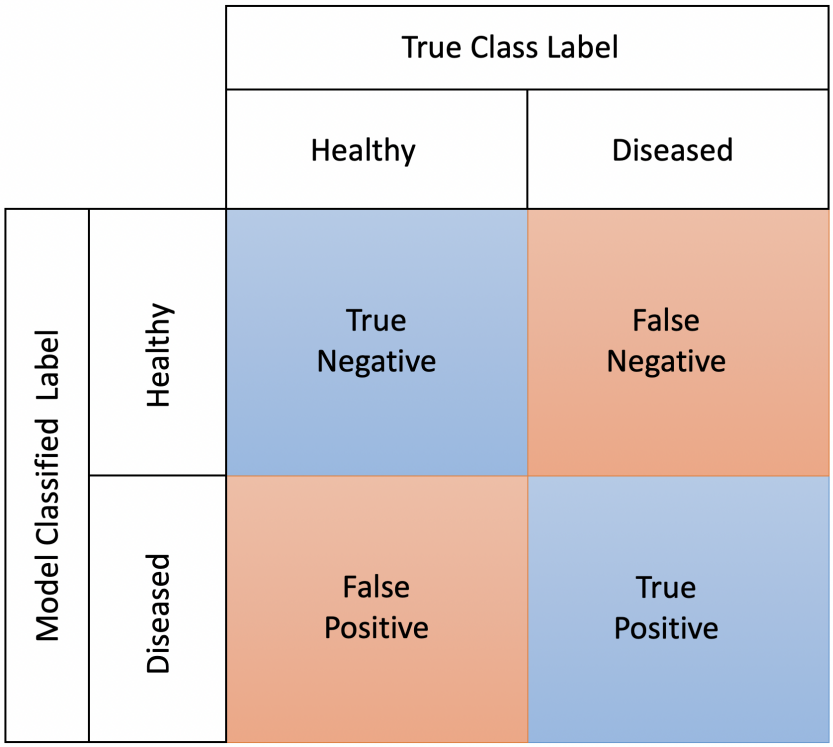


Figure 3.3: Schematic of a (binary) confusion matrix: the matrix decomposes results into true negatives (TN), false negatives (FP), true positives (TP) and false negatives (FN). This can be extended to any number of classes.

Many metrics, particularly in medicine, can be derived from the confusion matrix. This is a matrix specifying the classifications made by a model against the true class labels (figure 3.3). This allows a more nuanced understanding of the types of errors a model may be making. For instance, of those samples classified as diseased, how many truly were diseased, and likewise, of those classified as healthy, how many were truly healthy. These are known as the sensitivity and specificity, and are formally defined as:

$$sensitivity = \frac{TP}{TP + FN} \tag{3.3}$$

$$specificity = \frac{TN}{TN + FP} \tag{3.4}$$

These are commonly used metrics in healthcare, particularly for multi-class problems, though they have come under some scrutiny [132], as discussed later. They are related to the concepts of type 1 and type 2 errors in statistical hypothesis testing, in that the probability of a type 1 error is measured as the false positive rate and type two errors as the false negatives. The strength of these metrics is that they are able to distinguish between types of error which may be qualitatively different. In medicine a false positive would lead to the over-treatment of a healthy person, and a false negative would lead to disease being missed. If all else is held constant (such as the sample size and nature of the phenomenon under investigation), it is the case that the false positives can only be improved at the expense of false negatives. Which of these errors should be minimised will depend upon the application; a screening test may be able to tolerate a greater degree of false positives than a gold-standard diagnostic test. These considerations broaden to become complex multi-faceted decisions including medical ethics and health economics, which are beyond the scope of this thesis.

Accuracy can also be defined in terms of the elements of the confusion matrix:

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{3.5}$$

though we can now see that the accuracy provides no information into *how* a model is failing to classify.

A number of other metrics can also be defined by arithmetical combinations of these terms, such as the positive and negative predictive value, precision, recall and the related F1 score. Their use is less common is medical applications and so will not be described here; more information about them can be found in Tohka and Gils [130].

3.3.3 ROC

A common medical metric is the receiver operator curve (ROC), and the associated area-under-ROC (AUROC). These are derived from the sensitivity and specificity. Most ML models do not output a simple class, but rather a number. We would prefer the model to be trained to output, say, 1 for diseased and 0 for healthy but usually a number between these is given. Therefore a threshold needs to be set, above which the class 1 will be assigned and below which to the class 0. A common default is 0.5, but this is an arbitrary choice. The ROC takes numerous threshold values and plots the associated sensitivity against 1 - specificity (figure 3.4). Thus we are able to explore the effect that the threshold choice has upon performance. The optimal choice will be a balance between the sensitivity and specificity, which will be application specific. For instance in screening programmes, which may anticipate a very large number of samples, a high specificity may be preferred to a high sensitivity to avoid a large number of false positives and the incumbent mental and physical distress and financial burden of subsequent confirmatory diagnostics. Those very same confirmatory tests will prefer high sensitivities in order to avoid false negatives, where disease is missed. We would prefer both to be high, but when only adjusting the decision threshold for classification, it is necessarily a trade-off between the two. The AUROC measures the normalised area under the ROC, with scores between 0 and 1. A perfect classifier has $AUROC = 1.0$, whereas $AUROC = 0.5$ is often regarded as performing at random. What constitutes a good AUROC is application dependent; 0.7 might be considered very good when human level performance is little better than guessing, whereas as score of 0.95 might not be good enough when humans very seldom make classification errors, or the cost of an error is life-threatening.

For classification models to be clinically meaningful, a calibration step must also be undertaken: a thorough and clinically relevant exploration of where to draw the line between classes, rather than relying, usually unknowingly, on the default of 0.5 specified by most ML packages/libraries. This is most pertinent for models which are to be deployed in real settings, as opposed to proof-of-concept type studies.

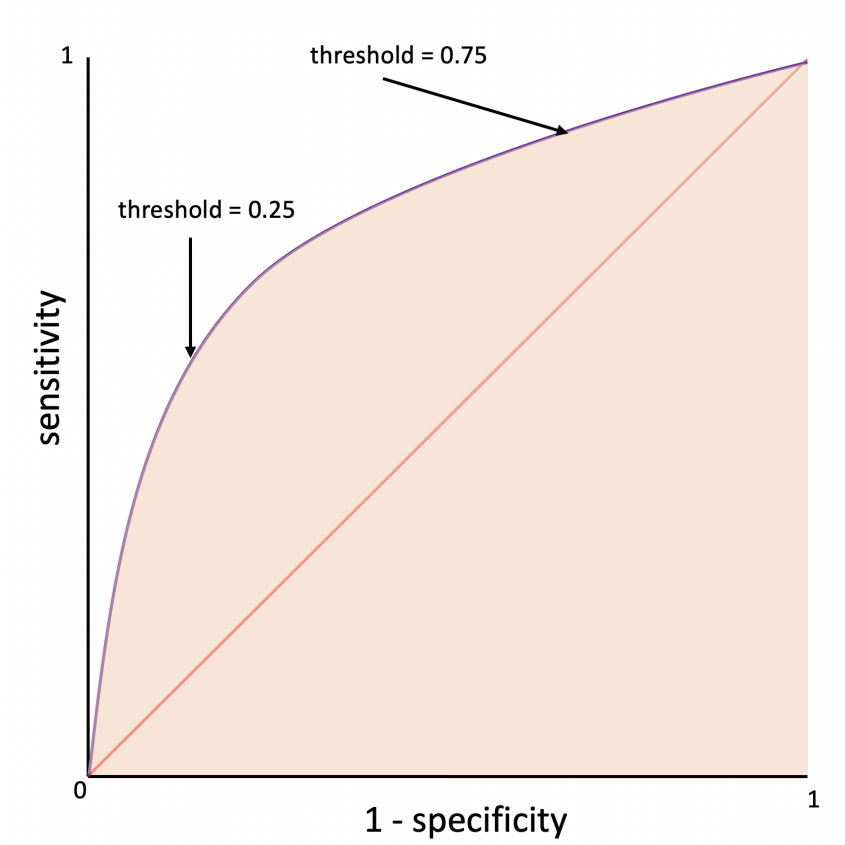


Figure 3.4: Schematic representation of a ROC. Each point on the purple line indicates a different threshold which balances the sensitivity and 1 - specificity. Red line indicates a model classifying at random. Shaded region indicates the AUROC

3.3.4 One class versus all others

The above metrics have been defined only for binary classifiers. They can be extended to multiclass classifiers by considering one class as the positive class and all others as the negative class. However, the usefulness and interpretation of such metrics is debatable. If the real world application is explicitly multi-class then binarised metrics will not reflect the nuances of actual disease classification. It is common to see a model able to well classify healthy against combined diseased categories, but struggle when classifying between the disease subtypes [130]. Additionally, the binary case is often also clear to health professionals, and it is precisely the intermediary disease classes for which a model is built to distinguish between, therefore the reporting of appropriate metrics is particularly important. For instance, in the Lynch dataset, which has 3 classes, we could make 3 pairwise comparisons: normal vs MSI-H and

MSS, MSI-H vs normal and MSS vs normal and MSI-H. If the clinical problem was to distinguish MSI-H from either healthy or MSS samples this may be a reasonable measure of performance. However, the clinical problem, as discussed in section 2.2, explicitly seeks a model that could distinguish between the three classes. This is also true of the SMART dataset which has 5 classes, in which the intermediary classes are of more clinical interest than the normal and AC classes.

Therefore, contrary to most medical applications, the sensitivity and specificity will not be reported, and where AUROC and ROC are used, all pairwise comparisons will be reported, not just those most favourable (i.e. normal vs all diseased). The confusion matrix will be reported for every classification problem, allowing the construction of all the above metrics. Although the Ovarian dataset is truly a binary problem, the same metrics will be reported to allow for more direct comparisons across all three datasets.

3.3.5 Log-loss

3.3.5.1 Prediction vs classification

In lieu of these performance metrics, we consider an alternative metric. Many from the more traditional statistics community advocate for the use of 'proper scoring' metrics, as opposed to those above which are described as 'improper scoring'. A proper scoring metric imposes no threshold upon the model outputs, but rather regards them as continuous variables. In particular, the model outputs are put through a 'softmax' layer, which requires that the outputs are non-negative and sum to one, allowing them to be interpreted as probabilities. This gives a prediction, with a certain confidence, rather than a definitive classification. The putative benefit of prediction is that it sits easily within a Bayesian paradigm of medical diagnostics, ready to be integrated with previous and subsequent clinical findings to make diagnoses increasingly confident. Thus, a medical tests error rates explicitly become part of the medical decision making process, something lamentably missing in current clinical practice. It also means that no information regarding the 'certainty' of the model is being discarded. Consider the case where a model outputs 0.51 for one input and 0.99 for another. In an improper scoring regime, using the typical threshold of 0.5,

both will be classified as 1 (diseased). However, the model is in some sense more certain of the latter classification compared to the former. By imposing any threshold this information is lost.

The log-loss (ll) is a proper scoring metric. In the programming language Python (version 3.10), the Scikit learn library defines it as [133]:

$$ll = -lnI\!P(Y|\hat{Y}) = \frac{1}{N}\sum_{i=0}^{N-1}\sum_{c=0}^{C-1} y_{i,c} ln(\hat{y}_{i,c}) \tag{3.6}$$

where Y is a 1by C binary indicator matrix, column Y_i being the i^{th} sample. Then an element of the matrix, $y_{i,c} = 1$, if sample i has label c from the set of C labels, otherwise it is 0. Similarly, $\hat{Y}$ is a matrix where $\hat{y}_{i,c}$ is the prediction of the model for the i^{th} sample and c^{th} class label. Its columns, $\hat{Y}_i$, represent the probability distribution of sample i belonging to the C classes, and are non-negative and sum to one.

For instance, the true class vector for a three class model, such as the with the Lynch data, which has classes normal, MSI-H and MSS, can be represented as $Y_i = [1,0,0]^T$, meaning it has the label 'normal'. An example of the output of a model from a single input (i.e. a spectrum) might be $\hat{Y}_i = [0.5,0.4,0.1]^T$. Note with improper scoring this latter case would simply be classified as normal, as that has the highest value and the other values are ignored. With the log-loss, this 'uncertainty' is directly quantified.

Regardless of which performance metrics are used, alone they are only one factor in determining the suitability of a technique. Other factors include usability for end-users, ease of integration into existing workflows, ease of integration into existing IT infrastructures and regulations, robustness to missing data, interpretability, cost-effectiveness and post-deployment surveillance on actual patient outcomes (i.e. mortality and morbidity) [130], some of which we consider next.

3.4 Reproducibility

Parallel to the concept of the generalisation error is the idea of reproducibility. ML in general is facing what has been described as a "reproducibility crisis" [134], due

to a lack of transparency and reporting, from data collection and curation to model selection and training. This is exacerbated in medical applications as reproducibility is fundamental to establishing clinical veracity [135].

Reproducibility can be deconstructed into three facets [135]. *Technical replicability* refers to the ability of others to replicate results given the dataset and code (i.e. the model and any pre-processing steps). Dataset availability is often lacking in medical studies due to the sensitive nature of medical data. Technical replicability also relies upon sufficient detail being given about all the steps of a study, including modelling. *Statistical replicability* is principally concerned with a robust estimation of the generalisation error as described above. This includes using appropriate pre-processing steps, hyperparameter optimisation strategies and validation strategies, as described in section 5.2. It is the thorough reporting of these strategies, sufficient to allow others to replicate a study, which facilitates technical replicability. *Conceptual replicability* refers to the ability to generalise across multiple centres, and ultimately to the entire population of interest. Arguably this could be collapsed into statistical replicability, as it is concerned with generalising to the entire population of interest. However, this is such a persistent problem in healthcare applications it is often highlighted separately. It also includes the tendency of healthcare models to 'drift' over time due to updating clinical protocols and training regimes, servicing and replacing of equipment and other time dependent factors [136]. It is related to the concept of external validity used in medical research parlance concerned with the degree to which causal relationships can be generalised to various measures such as people, settings and time.

Much effort in this thesis has been placed into statistical replicability. Conceptual replicability is also explored with the SMART dataset, which uses data from multiple centres. It may not be possible to satisfy technical replication due to the medical nature of the datasets, but enough details to reproduce the work if the dataset were available will still be given as a testament to best practice.

3.5 Small sample sizes

Throughout this thesis reference to problems incurred from small sample sizes are discussed. This is related to the discussion of model complexity and overfitting in section 3.2.4: the smaller the dataset the greater the risk of overfitting. This is exacerbated with increasingly complex models, as characterised in figure 3.5. However, as will be discussed in section 5 there are techniques which may change the shape of these curves to make DL more amenable to smaller datasets.

These problems ultimately stem from samples being unable to represent the entire population of interest. Even though it is usual to treat individual Raman spectra as the base unit of prediction/classification, it is the patients' samples which should determine the sample size. This is because biological variation will exist at the level of the patient, and a sufficiently sized sample is required in order to represent the full breadth of all patients. Little is currently known about how biological variation manifests in Raman spectroscopy, although it has been suggested that if natural healthy variation can be incorporated into a RS model it will perform better even on data from diseased samples [137].

3.6 Statistical comparisons of model performance

One of the requirements for statistical replicability is to report some measure of variance [135]. This is possible when repeated measures of a performance metric are taken. In this study, the standard deviation (SD) is often given. However, care is needed in interpreting such interval estimates in the context of ML. This is because the various samples used to repeatedly train a model are usually not independent - they will include some of the same samples (details in section 5.2). This breaks the identically and independent distributed (IID) assumption required to construct confidence intervals and statistical hypothesis tests (for instance to statistically compare the performance of two ML models). Only if all the repeated measures are truly disjoint, would standard statistical tests be valid. Although there do exist methods to compensate for the lack of independence between samples [138, 139], they are based on assumptions which have not been examined in the context of RS

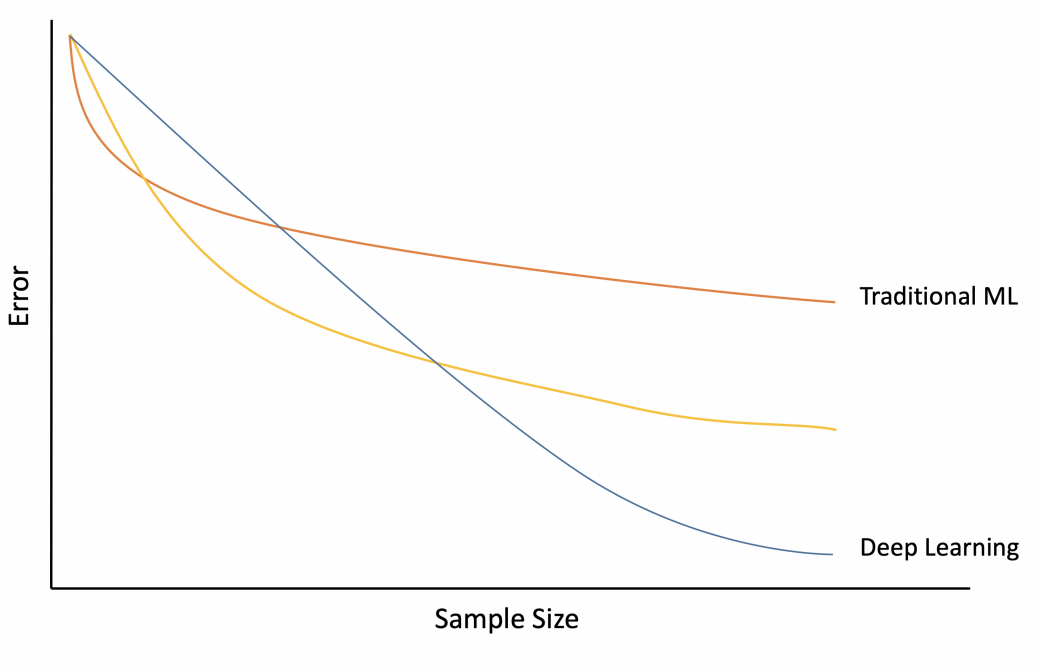


Figure 3.5: Schematic representation of error as a function of sample size with models of increasing capacity. Red line: traditional ML models such as LDA. Yellow line: Model with intermediary capacity, such as a shallow NN. Blue line: deep learning models

and their use is not common in the medical literature. As the datasets examined in this study are exploratory rather than for clinical implementation, I will not apply hypothesis testing and instead simply report mean performance +/- 1 SD as a measure of variance. In subsequent chapters there will be a number of exploratory results, generating a plethora of results. To remain concise yet informative, results in section 5, which explores a number of pre-processing and data preparation steps, will focus primarily on accuracy and log loss, unless otherwise stated. Though accuracy may be an imprecise measure, it is intuitive and widely understood, so is well suited to tasks generating copious amounts of results. In section 6, in which the final models are implemented, the confusion matrix, ROC and AUROC will additionally be reported, allowing for a more nuanced understanding of the final results.

Chapter 4

Machine Learning Models

> *"All models are wrong, but some are useful"*
>
> George Box

4.1 Introduction

RS, and more generally spectral histopathology (SHP), like many of the *omics* type data (genomics, proteomics, metabolomics etc.), tends to be high dimensional and complex, making it well suited to the data-driven ML techniques. This paradigm is different to the model driven approach of traditional statistical modelling, which assumes that data are caused by some true data generating process (i.e. nature), which is usually unknown and so the task is to find a model that well approximates the unknown process. In contrast, data driven modelling assumes nothing of the data, instead allowing the data itself to generate the model, thus mimicking the data generation process. Model driven approaches find a model that generates the data; data driven approaches allow the data to generate a model [140].

Data driven approaches have become increasingly popular in medicine [58], emboldened by the successes of deep learning in computer vision and natural language processing. This is now extending to SHP [59], with the more recent ML models applied to oncological datasets exploring the potential of deep learning architectures [101].

Despite this promise, the approach is not without its own problems. Barriers to medical applications of data driven approaches include a lack of reproducibility [141, 142], difficulties in model interpretation (particularly important in medicine as an understanding of casual mechanisms is one of the pillars of the Bradford-Hill criteria of assessing medical evidence [143]) and the need for large and well annotated data sets.

RS data is suited to data driven models as the light scattering phenomena which generate it are sufficiently complex to limit model driven approaches. There are many techniques within data driven modelling, and as there are few applications of RS in routine clinical use at least some of the scrutiny must fall onto the models used thus far. Within this remit, I first distinguish between two types of data driven modelling: traditional and deep ML.

4.2 Traditional versus deep machine learning

Deep learning refers explicitly to the depth of a neural network (detailed in section 4.5), which can be understood as the number of parameters, or the capacity, of the model. For the purposes of this thesis deep learning is defined as any model based on a neural network architecture, from artificial neural networks (ANNs) to more sophisticated architectures such as convolutional neural networks (CNNs). Traditional models include all other models, whether they are linear statistical techniques such as LDA, or non-linear, such as support-vector machines (SVMs) with an appropriate kernel function. As discussed in section 3.2.4, the constraints upon more complex models are such that traditional models may be preferable, alluded to in figure 3.5. PCA-LDA is one of the most common methods used in ML [144], though its use in oncological applications of RS is waning in favour of deep learning methods [101]. Therefore, we will explore the use of this model and use it as a baseline of traditional ML performance. SVMs are a more sophisticated traditional ML model which often outperform PCA-LDA, and other traditional models [101]. CNNs are becoming increasingly common in oncological RS literature, and are current exemplars of the medical imaging deep learning paradigm.

This section details these three models, which will subsequently be used to analyse the data used in this thesis in sufficient detail to allow an understanding of the various hyperparameter choices later examined.

4.3 Principal Component Analysis - Linear Discriminant Analysis

Principal Component Analysis (PCA) - Linear Discriminant Analysis (LDA) has been a remarkably popular choice of chemometric model used in biomedical applications of RS. It remains popular even in the most recent literature, as demonstrated by the literature review of section 1.5, though often as a baseline comparator to other models.

Principal Component Analysis

PCA is commonly used to reduce the dimensionality of a dataset. Given the high dimensionality of RS data (the number of discrete wavenumbers, or bands, on the x-axis) PCA can be used as a pre-processing technique which helps to select those features of the data most pertinent for classification. Specifically, PCA is an unsupervised technique that takes in a set of p features $x_1, x_2, ..., x_p$ (the wavenumber bands in the case of RS with p equal to the number of wavenumbers). PCA finds a low-dimensional representation of these p features whilst retaining as much information as possible, based on the assumption that relevant information is contained in those directions which most vary.

Given a mean-centred dataset, $\mathbf{X}$, of size nxp with n observations (Raman spectra), PCA computes the first principal component (PC) as the linear combination of the p features of the form

$$u_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + ... + \phi_{p1} X_{ip} \tag{4.1}$$

which has the largest variance. u_{i1} refers to the PC score. The first PC, U_1, is constructed from these scores $u_{11}, ..., u_{n1}$. It is possible to plot the n scores of a dataset, $U_1, ..., U_n$ against each other. This can be useful when there are multiple

classes within the data as the technique sometimes shows that the classes can be linearly separated on a lower dimensional hyperplane in an unsupervised manner (PCA does not take into account class labels). For ease of understanding this is often done by selecting two PC scores to plot against one another (figure 4.1c).

The above equation is normalised such that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ and these elements, $\phi_{11}...\phi_{p1}$, are referred to as the PC loadings. These can be represented as a single vector, $\phi_1 = (\phi_{11}, \phi_{21}, ..., \phi_{p1})^T$. In the context of biomedical RS, these loadings can be interpreted as a Raman spectrum which is a linear combination of the dominant spectral features in a dataset, from which it may be possible to identify some biochemical species (figure 4.1a).

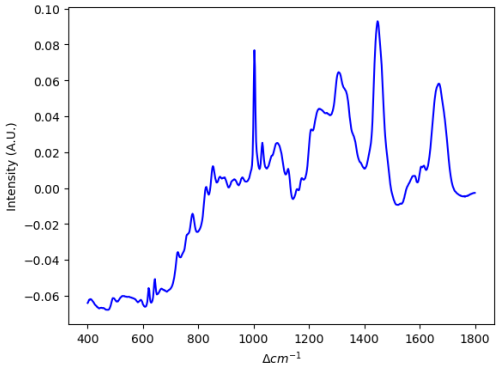
The second PC, U_2, is calculated with the additional constraint that it is orthogonal to the first PC, U_1, and finds the direction which varies the most in the remaining data space. Subsequent PCs are then similarly calculated, constrained to be orthogonal to all proceeding PCs, with subsequent PCs being ranked by decreasing variance in the data space. Thus, an underlying assumption implicit in PCA is that the directions of maximum variance in the data space are the most relevant for a problem, which may be generally, though not always, true.

As variance is the pertinent information for PCA it is common to scale the data. This scaling ensures that features which are orders of magnitude larger than others (such as the phenylalanine peak in biomedical RS) are not disproportionately weighted when calculating the maximum variance. Thus the remaining variance is not due to the scale of the variables, but is relevant information. This is perhaps more clearly demonstrated in medical datasets consisting of disparate data: for instance systolic blood pressure, measured in mmHg and usually in the 100s, and blood glucose levels measured in mmol/L which are typically below 10. Without scaling, blood pressure would very likely dominate the loadings, giving the impression that it is the most important feature of the dataset, which may not be the case.

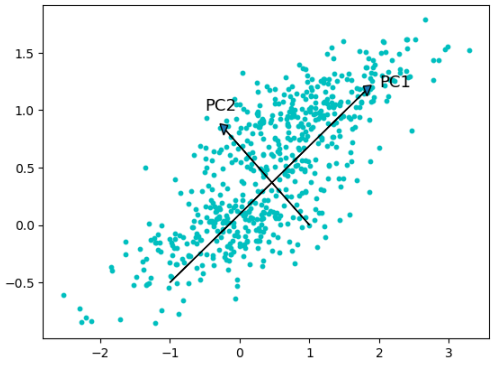
PCA can be geometrically understood as a rotation of axes such that the new principal axis spans the data in the direction of maximum variance, with subsequent axes orthogonal to this (compare figures 4.1b and 4.1c). Thus PCA can be understood

as a rotation and scaling of the data cloud along its direction of maximum variance.

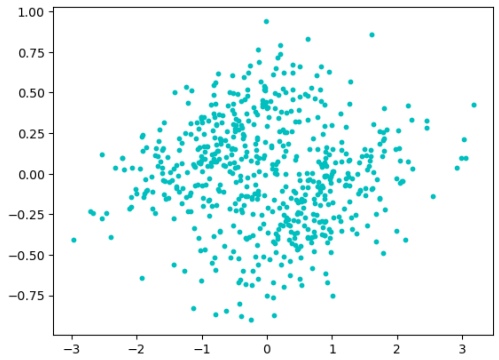
The data clouds in figure 4.1 represent an idealised case of PCA, constructed only for illustrative purposes from a simulated dataset, showing only the first two PCs. Here, PC1 captures 56% of the total variance in the data, PC2 44%.



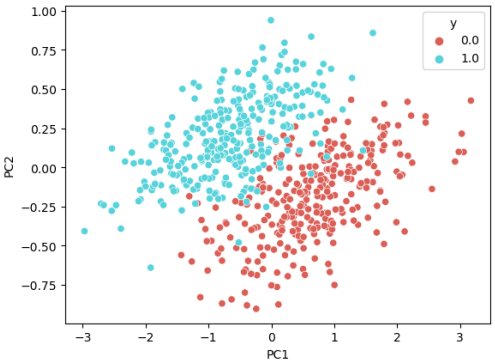
(a) A PC interpreted as a Raman spectrum



(b) Random Data Cloud



(c) PCA transformed Data Cloud



(d) Labelled Data Cloud

Figure 4.1: (**a**) PC1 loading taken from the lynch dataset (**b**) A data cloud of random two dimensional data. The arrows show the new axes created by the transformation which form PC1 and PC2 (**c**) The data cloud after the rotation and scaling of PCA, such that the x and y axis represent PC1 and PC2 (**d**) Points in the data cloud have been labelled, showing the possibility of separability of the data

p PCs will be produced by PCA. To reduce the dimensionality of the data a subset, q, of these p PCs will be selected. There are a number of heuristics to choosing this number, like selecting q such that the cumulative percentage of the variance is accounted for (often 90%); the remaining PCs are assumed to be noise and discarded. Figure 4.1c shows an idealised use of PCA on a randomly generated 2D dataset with a systematic offset added to generate two 'classes'. PC1 captures the

largest amount of variance in the original data space. PC2 could be discarded under the assumption that it is noise, making for a simpler 1D plot. In this illustrative 2D case there is little benefit to PCA, but the technique becomes increasingly attractive with higher dimensional spaces. In section 6, we treat the number of PCs to retain as a hyperparameter of the model, allowing the data to determine the optimum number.

Although it is possible to keep the class labels during PCA, they are not used during the calculation; hence the reference to PCA as an unsupervised technique. The resulting plot, such as in figure 4.1d can be useful to visualise the separability of the data, but critically it does not give a decision boundary with which to classify new samples. For this, we need a supervised learning technique, such as LDA.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a supervised learning technique which takes a number of observations (i.e. spectra) and assigns each to one of K classes. Its derivation is beyond the scope of this thesis (see [131] for a full account), but following from the definition of Bayes rule and using the multivariate Gaussian distribution and taking the log-odds, a decision rule, $\delta_k(z)$, can be constructed in which an observation is classified to one of the K classes for which

$$\delta_k(z) = U^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + ln(\pi_k) \tag{4.2}$$

is largest. μ_k is the k^{th} class specific mean vector, π_k the prior probability of class k and Σ is the covariance matrix, common across all classes. These are all estimated from the training data, unless there is specific domain knowledge which can be used, such as the prior probabilities (i.e. prevalence) of disease classes in the population of interest. This provides a criterion on which to classify new observations as belonging to one of the K classes which forms a linear decision boundary (figure 4.2).

Hence, assumptions of LDA are that the observations are multivariate normally distributed and have a class-specific mean vector and a common covariance matrix. In the Raman oncology literature review performed in section 1.5, none of the studies using LDA tested the assumptions of multivariate normality. However, in other

domains it has been found that this assumption does not need to be strictly adhered to for practical applications of LDA [145]. The $ln(\pi_k)$ in equation 4.2 refers to the K^{th} classes prior probability. As this is usually estimated from the training data, this makes explicit that LDA is sensitive to the distribution of the classes. This sensitivity has also been shown empirically [146]. This will become relevant section 5.6 where imbalance amongst the classes is considered.

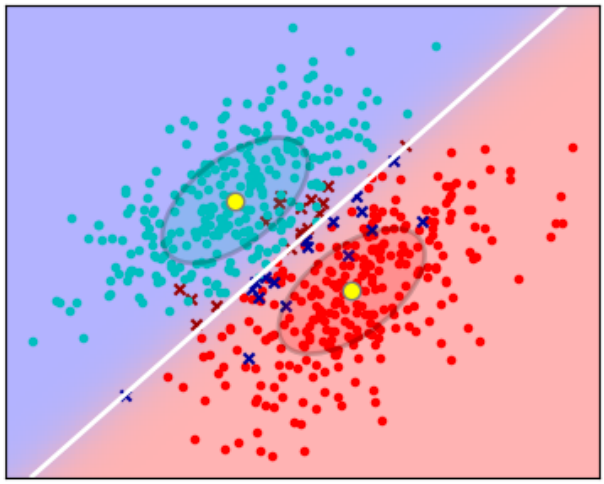


Figure 4.2: Taking the 2D PCA transformed random data cloud of plot 4.1b, LDA was performed to create a decision boundary (white line). Dots represent correctly classified training data, crosses incorrectly classified. Ellipses represent 1SD around the yellow dot of the class centroid mean.

There are many derivations of LDA, although they are ultimately the same. A common understanding stems from the close relation of LDA to analysis of variance (ANOVA) often used in medical hypothesis testing. In this conception LDA can be understood as maximising the separation between the K classes, while minimising the separation within those classes. The resulting groups (centroids) form ellipsoids which can be visualised as seen in figure 4.2.

4.4 Support Vector Machine

A Support Vector Machine (SVM) finds a hyperplane that separates data which can then be used to classify new observations. A hyperplane in p dimensions is a flat affine space of dimension $p-1$: in 2D a hyperplane is a line, in 3D a plane and in 4D a cube. In higher dimensions it is difficult to visualise but the principle remains.

However, for perfectly separable data (i.e. no irreducible error) there are an infinite number of such hyperplanes. Thus the hyperplane with the maximum margin is sought. The orthogonal distance from a separating hyperplane can be measured for each observation; the smallest distance of these is called the margin and the corresponding observation the support vector (as a point can be represented as a vector). The maximal margin hyperplane then is the separating hyperplane which maximises the distance from this point. An interesting feature of SVMs is that they are dependent upon the support vector but not the remainder of the data. However, this is true only of idealised data. Real data will not be perfectly separable and several support vectors will be used to define the separating hyperplane (see figure 4.3a). This would lead to overfitting, as the model is exquisitely sensitive to a subset of the data. Hence, soft margins are used. The 'softness' of the model is determined with the parameter C: the larger is C the more misclassifications are tolerated.

The derivation of SVMs is too complex to be covered here, so it is simply stated here as:

$$\max_{\beta_0,\ldots\beta_p,\epsilon_1,\ldots,\epsilon_p} M \tag{4.3}$$

subject to the constraints,

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\beta_p x_{ip}) \geq M(1-\epsilon_i) \tag{4.4}$$

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{4.5}$$

$$\epsilon_i \geq 0, \sum_{i}^{n} \epsilon_i \leq C \tag{4.6}$$

In equation 4.3 M is the width of the margin which we seek to maximise. Equation 4.4 defines the separating hyperplane (thus creating a decision boundary), with an indicator function, y_i, for the class label which indicates which side of the hyperplane an observation falls, i indicating the i^{th} observation of n total observations.

Equation 4.6 defines the 'softness' of the margins where ϵ_i are slack variables that allow individual observations to be on the wrong side of the margin or hyperplane. If $\epsilon_i = 0$ then it is on the correct side, if $\epsilon_i > 0$ then the observation is on the wrong side of the margin and if $\epsilon_i > 1$ the it is on the wrong side of the hyperplane. Pertinent to parameter searching, we can thus see that C bounds the sum of the ϵ_i's and so determines the extent of misclassifications. Thus it can be understood as a parameter controlling a trade-off between bias and variance as the larger it is, the wider the margin will become as more observations will be tolerated within it. Of note is that all observations within this margin are support vectors, not just those on the separating hyperplane, hence as C increases so a larger subset of the data is involved in determining the decision boundary. The influence of C on the decision boundary can be seen in figures 4.3a and 4.3b. This contrasts with LDA which always uses the entire dataset in order to construct class means and common covariance.

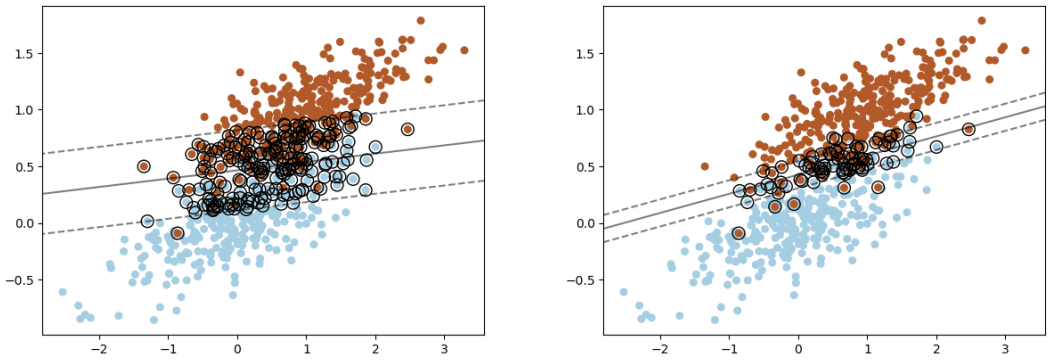
SVMs are not easily amenable to multi-class extensions, hence a one-versus-one approach is often adopted. This constructs $\binom{K}{2}$ SVMs, each of which compares a pair of classes. This is essentially a pairwise comparison and an observations final classification is given by the class to which it was most frequently assigned.

A great strength of SVMs is that they are flexible enough to create non-linear decision boundaries by using kernel functions. The kernel function creates an enlarged feature space, actually increasing the dimensionality of the data. In this space it is possible to find linear decision boundaries, like a separating hyperplane, which are non-linear when projected back down to the original feature space. The equations 4.3 - 4.6 are solved by using a generalisation of the inner products, $K(.)$, which quantifies the similarity of two observations, x and x_i. Again, the derivation is beyond the scope of this thesis, so is simply stated:

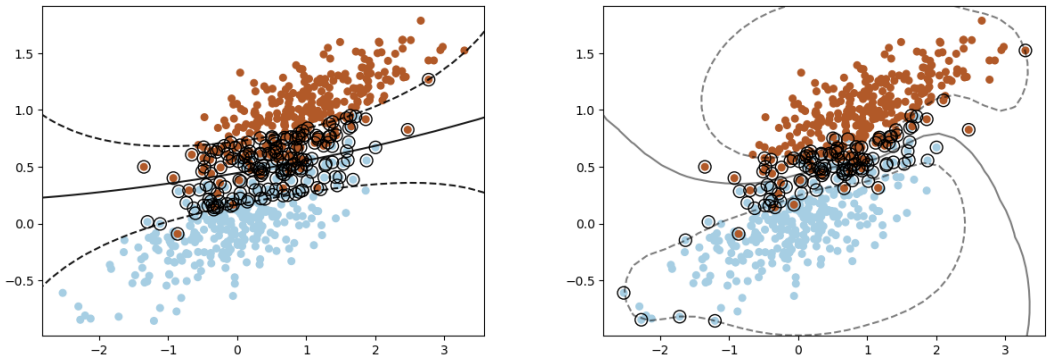
$$f(x) = \beta_0 + \sum_{i \in S} \beta_i K(x, x_i) \tag{4.7}$$

S being the collection of indices of these support points.

However, the computational cost of kernel based SVM scales quadratically with n, thus this technique may be unsuitable for particularly large datasets. There are



(a) Linear SVM: $C = 0.1$ **(b)** Linear SVM: $C = 10000$



(c) SVM RBF: $C = 1, \gamma = 0.1$ **(d)** SVM RBF: $C = 1, \gamma = 1$

Figure 4.3: SVM performed on the same random data cloud produced in figure 4.1. The thick line represents the decision boundary and the dashed line the corresponding margin. Circled observations are those on the decision boundary or on the wrong side: i.e. they are support vectors. Note that in Python the C parameter is inverted, hence a smaller C leads to a larger margin: the principle remains the same.

many choices of kernel function. Through an initial discovery phase I chose to use the popular radial basis function (RBF) which takes the form:

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \tag{4.8}$$

where γ is a non-negative parameter which controls the curvature of the function and thus overly high values could lead to overfitting. The influence of γ on the decision boundary can be seen in figures 4.3c and 4.3d. Thus, for our choice of kernel function, we have two parameters to consider (which become hyperparameters in the context of ML): C and γ.

4.5 Neural Networks

Artificial Neural Networks

Artificial Neural Networks (ANNs) are a useful gateway to understanding deep learning architectures in general. They are a type of feedforward network inspired by the network of neurons in the brain. The network has an input layer, a number of intermediate layers referred to as hidden layers, and an output layer which classifies the input as belonging to one of several possible disease classes in a medical context (figure 4.4). Feedforward refers to each 'neuron' only passing information forward in the network, as opposed to recurrent neural networks which can send information back to earlier layers. Such a neural network is described as being fully connected, alluding to the fact that each feature is connected to each weight in the first layer, which are in turn connected with all weights in subsequent layers and so on until the output layer.

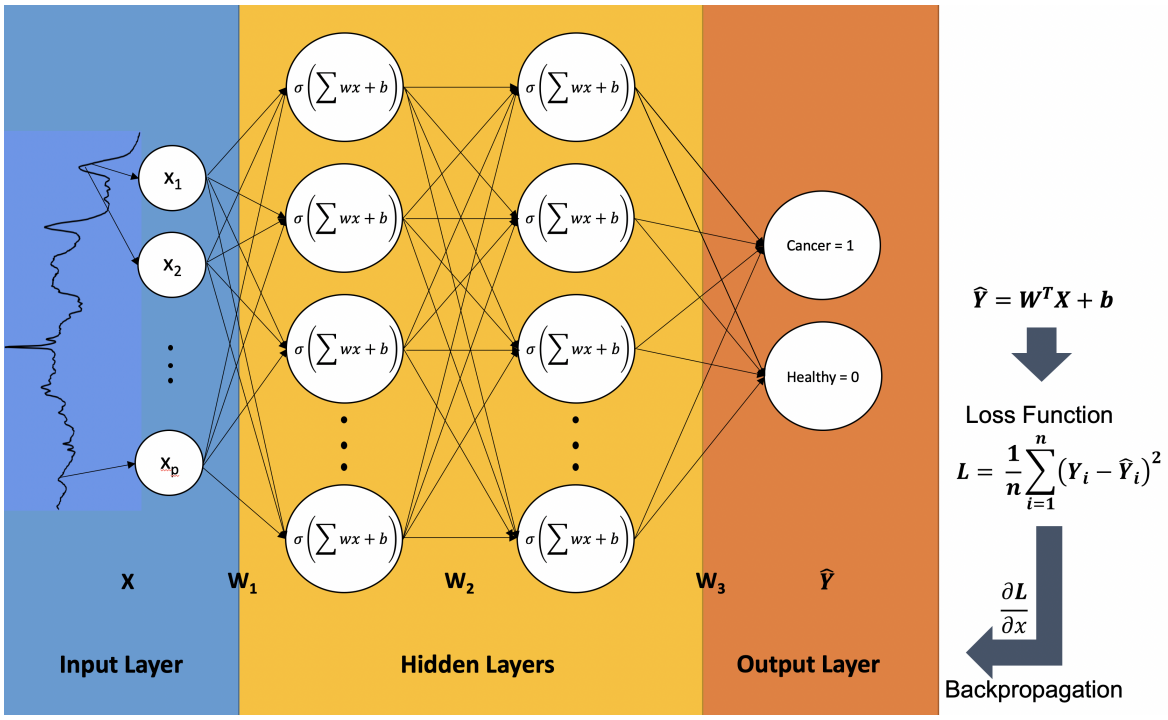


Figure 4.4: A schematic of an ANN: each circle in the hidden layer represents an artificial neuron

In more detail, and specific to RS, the input takes the form of a vector which represents a Raman spectrum with p features $X_i = x_1, x_2, ..., x_p$, where p is the number

of wavenumbers and the subscript i indicating the i^{th} spectrum of the dataset. The task is to estimate the output, $\hat{Y}_i$, which we wish to be equal to the known class label Y_i. The output, $\hat{Y}_i$, is then a function of the various operations performed upon the input, X_i, through the neurons of the intermediate layers. Each neuron, $Z_l^{[m]}$, takes in the output from each neuron in the proceeding hidden layer, where m indicates which hidden layer (and is in brackets to make clear that it is not an exponent) and l indicates which neuron of a given layer. Each neurons outputs are adjusted by their weights, $W^{[m]}$. A receiving neuron then sums all of the incoming weights and adds a bias term, b : $Z^{[m]} = W^{[m]}Z^{[m-1]} + b^{[m]}$. In the case of the 1^{st} hidden layer the spectrum is the input. This sum is then subject to a non-linear activation function, σ. This can take various forms, but all act to constrain the range of a neurons output and creates the non-linearity of the model.

All layers together can be expressed more concisely using matrix notation:

$$\hat{Y} = \boldsymbol{W}^T \boldsymbol{X} + \boldsymbol{b} \tag{4.9}$$

where the activation function, σ, is implied rather than explicitly stated and $\mathbf{X}$ captures the input spectrum in its first column and subsequently the outputs of hidden layers, Z.

During training, the parameters $\mathbf{W}$ and $\mathbf{b}$ are first randomly initialised and an output, $\hat{Y}_i$, is produced. This output is then compared to the true class label, Y_i, via a loss function. This loss function can take many forms, but in figure 4.4 the mean square loss is given as it has an intuitive interpretation. It takes the squared difference between the predicted and true label. This is zero for exact predictions and grows larger the further the prediction is from reality. We seek to minimise the difference. This is achieved by a process called backpropagation in which derivatives are taken via the chain rule. The parameters are then adjusted to fall down the gradient of the loss function (figure 4.5). In this parameter space we seek to find the global minimum by learning the weights and biases.

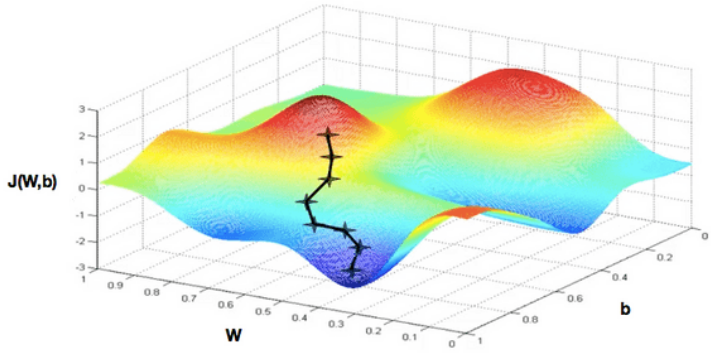


Figure 4.5: Loss Function Schematic. Each black star represents one step, or update, of the model parameters during back-propagation. This 'landscape' is sometimes referred to as the parameter space.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of ANN. CNNs have fewer connections than ANNs as the convolutional layers only connect to a subset of the neurons in the preceding layer, as opposed to all as with a typical ANN. This relative sparsity of connections has been argued to make CNNs more suited to vibrational spectroscopy applications as they are less prone to overfitting and make for more interpretable models [147].

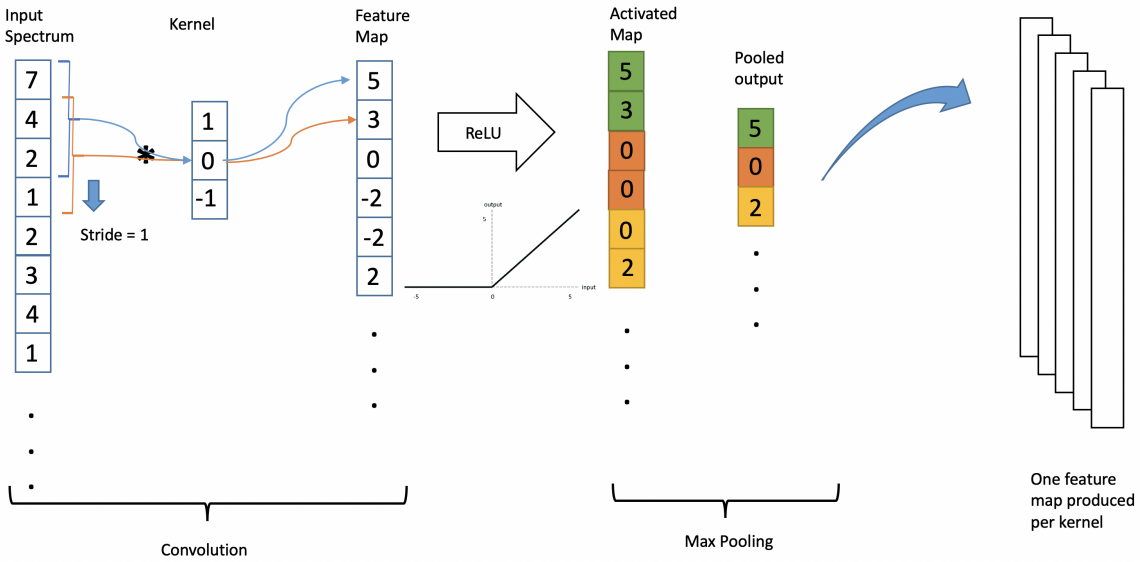


Figure 4.6: A Single Convolutional Layer. * indicates the convolution operation. The numbers used are entirely illustrative.

A kernel of a certain size is passed over the input data and a convolution operation performed to produce a number. The kernel is then moved down the input spectrum, each time the convolution produces a number. The step size, or stride, the kernel takes as it passes over the data can be altered. The output is a feature map, which we seek to train to extract features from the data. Mathematically, this can be expressed $(I * K)(i) = \sum_a I(a)K(i-a)$, where I and K indicate the input and the kernel respectively for the i^{th} input index and the a^{th} kernel index. This is illustrated in figure 4.6. In the case of 2D CNNs for image classification, one feature map may learn to recognise vertical lines, while another may learn to recognise rounded corners and so on. In the context of RS these should learn to extract pertinent spectral features. Each feature map is then subjected to a form of normalisation called batch normalisation to overcome a problem known as internal covariate shift, where there is change in the distribution of an activation due to changes in the network weights during training, meaning the network will take longer to converge to an optimal solution [148]. The convolutional neurons are then subject to an activation function as described in ANNs. Finally, a pooling layer may be applied which reduces the size of the extracted feature maps.

In the convolutional layers of a CNN, it is the values in the kernels which are learned via back-propagation.

Non-linear Activation Functions

σ is an activation function which transforms linear inputs into a non-linear output. This function can take several forms, which can have a significant impact upon the performance of a model and is sometimes treated as a hyperparameter during model development. A particularly common activation function is the rectified linear unit (ReLU), given by $max(0, z_l)$, where z is the linear output of the l^{th} neuron. It allows for sparse activations (some values will be 0), is less prone to vanishing gradients (where gradients become so small as to effectively prevent updates) and is computationally efficient [149]. This will be used for all subsequent DL models

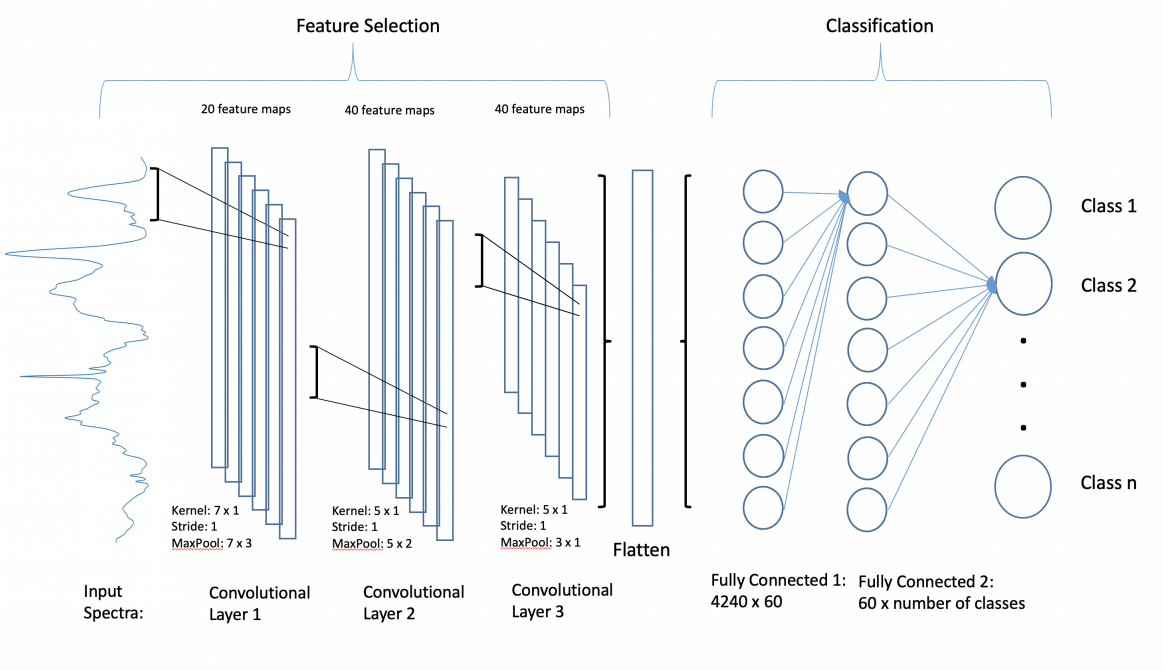


Figure 4.7: Schematic of CNN architecture. The convolutional layers learn to extract pertinent features in the data, while the classification layer, which is fully connected as described in the ANN, classifies according to those features.

developed in this thesis.

Parameter Initialisation

The model weights, W, will be learned during backpropagation, but to start values for W must be initialised. This can be thought of in terms of figure 4.5 as the starting point in the parameter space. The weights learned during training can be sensitive to these initial values, and a process known as seed hacking could be used to artificially inflate the performance of a model [150]. This can be avoided by using an appropriate initialisation procedure and repeating initialisation (for instance by using repeated CV). When using a ReLU activation function, Kaiming He initialisation is a robust choice [151]. This randomly initialises weights according to a Gaussian distribution with 0 mean and $\frac{2}{l^{[m]}}$ variance: $W \sim \mathcal{N}(0, \frac{2}{l^{[m]}})$, where l is the number of inputs into a given layer, m.

Loss Function

The loss function returns a prediction, $\hat{Y}_i$, of the true class, Y_i. Figure 4.4 shows the mean square error (MSE) loss, shown as it has an intuitive interpretation of being the sum of the squared differences between predicted and actual labels:

$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{N} y_i(Y_i - \hat{Y}_i)^2 \tag{4.10}$$

However, it is not necessarily the best choice and there are many possibilities, so long as the function is differentiable. A common choice is the cross-entropy loss, related to the Kullback-Leibler divergence in that it measures the difference between two probability distributions.

To build an intuition of the cross-entropy loss, let us begin with the binary case, which is also called the log loss:

$$\mathcal{L}_{ll} = -\frac{1}{N}\sum_{N} Y_i ln(\hat{Y}) + (1 - Y_i) ln(1 - \hat{Y}) \tag{4.11}$$

where ln is the natural log. In this binary case, the true label Y_i only takes values of 0 or 1, hence only one of the sums will be active. Compared to the MSE, this loss function more strongly penalises predictions, $\hat{Y}$, which can take any value between 0 and 1, that are far from the true values (figure 4.8). This may be preferred in medical applications when we wish to strongly penalise very 'confident' incorrect predictions.

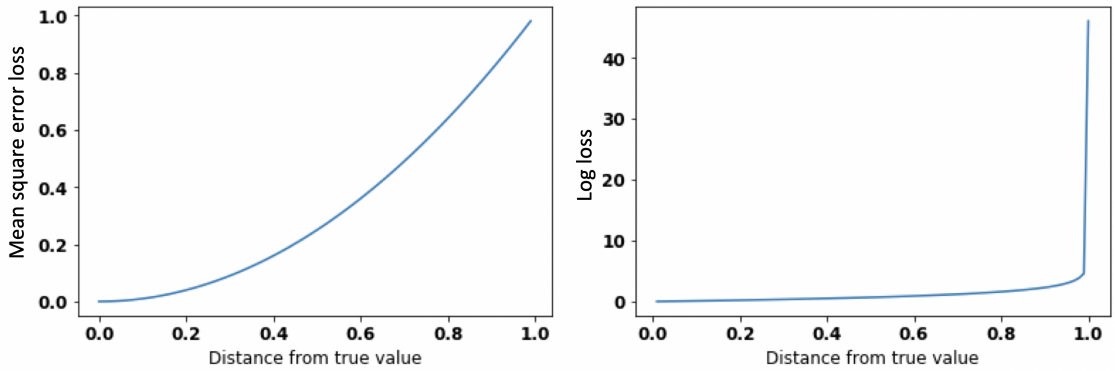


Figure 4.8: Mean Square Error vs Log Loss

The log-loss extends to multiple classes via the cross-entropy loss. In PyTorch [152], this is defined as:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{C} w_C ln \frac{exp(x_i)}{\sum_C exp(x_i)} y_i \tag{4.12}$$

where C is the number of classes and w_C is an optional weighting vector which can be used when dealing with imbalanced dataset (discussed in section 5.6).

Learning rate

The learning rate, α, is a hyperparameter which controls the step size during gradient descent. In figure 4.5 it is the size of the steps between the black stars. Mathematically, this can be understood as a factor of the partial derivative calculated during backprogagation: $\alpha \frac{\partial \mathcal{L}}{\partial x}$. A large α corresponds to a larger step size in the parameter space and could result in constantly over-stepping an optimal solution. Conversely, an overly small step size may lead to the inability to converge to a solution in a reasonable time. The learning rate is generally considered to be one of the most important hyperparameters [153], and so an optimal value is explored for each of the datasets in this thesis.

Batch gradient descent with Adam

Thus far we have considered the case of a single spectrum being entered into the CNN at a time. This search process is known as stochastic gradient descent (SGD). An alternative is to feed several spectra, known as a mini-batch, into the model at a time. The loss function is then calculated as the average over the mini-batch. This has the effect of smoothing the loss function which may improve convergence, although smaller batch sizes may result in a more rigorous exploration of the parameter space and apply a form of regularisation, as a less smooth exploration mimics noise in the gradient estimator [153]. The Adam optimiser is a popular extension of

classical gradient descent [154]. It combines the benefits of the adaptive gradient algorithm (AdaGrad) and root mean square propagations (RMSProp). Adam adjusts the learning rate per-parameter to accelerate learning in the directions with large gradients and to dampen learning in directions with small gradients. This introduces two hyperparameters to the model, β_1 and β_2, though $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are very common default values and are used in this thesis.

Once all the data has been passed through the model during training, it is said that a single epoch of data has been trained. For full training, many epochs will be passed through the model; it is not uncommon to use several hundred epochs.

Early stopping

When a neural network is trained over several hundred epochs it is possible that even though the training performance continues to improve, the test performance deteriorates. This is a classic case of overfitting. One method to prevent overfitting during training is to track the test performance at the end of each training epoch. If the test score does not improve, or worsens, for a given number of epochs, it is assumed that overfitting has begun and training is stopped early.

Drop-out

Another method to mitigate overfitting is drop-out regularisation. A certain proportion of the weights are randomly excluded from updating during backpropagation. This prevents the model from focusing on patterns present only in the training data, forcing the model to find more general patterns. Drop out is only performed on the fully connected layer of a CNN.

4.6 Hyperparameter exploration

This has been a brief treatment of PCA-LDA, SVMs and CNNs sufficient only to understand the choices made during model development in subsequent chapters.

Learning Rate	10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} 10^{-6}
Batch Size	32, 64, 128, 256, 512, 1032
Adam β_1	0.9
Adam β_2	0.999
Drop out rate	0.2
Early Stopping	5 epochs

Table 4.1: CNN hyperparameters

We see that there are a deluge of hyperparameters from which to choose when constructing a CNN. In the best case this facilitates the training of an appropriate model for a particular dataset, and at worst allows for the overfitting of the model at the second-level of inference (i.e. hyperparameter selection). Thus I have used values that have become standard in other domains of deep learning for many of these hyperparameters. Although this means I may not be fully exploiting the parameter space, it also gives a more conservative estimate of performances, mitigating against overfitting. However, two hyperparameters have been found to be particularly important for effective learning: the learning rate and the batch size. Thus these will be considered as hyperparameters to search with the Lynch and Ovarian datasets, and the best performing values will be used for the SMART dataset.

The CNN architecture used for all datasets in given in figure 4.7. Table 4.1 shows the corresponding hyperparameters.

Along with the hyperparameters identified for PCA-LDA and SVM, these can be classed as *model* hyperparameters, but there are a number of additional choices we need to make before analysis begins. Many of these are specific to RS and involve the appropriate pre-processing of spectral data, and can be considered data-specific hyperparameters (as opposed to the model-specific hyperparameters heretofore considered). The following chapter explores many of these choices, as well as appropriate validation strategies to ensure a reasonable estimation of the generalisation error.

Chapter 5

Data Preparation

> *“Preparation is the key to success.”*
>
> Alexander Graham Bell

5.1 Introduction

Before using the models to analyse the datasets there are a number of considerations regarding the nature of the data which need to be taken into account, else they might fall foul of the dictum rubbish in, rubbish out. Some of these considerations are particular to Raman spectra and medical datasets, and how best to split the data for training, others are bound to the ML methods to be applied, and some intersect these areas.

5.2 Model validation

As discussed in section 3 the pertinent point for assessing a ML model is how well it performs with previously unseen data from the clinical setting of interest. This is essential to finding a realistic estimate of the generalisation error. However, there are many practical limitations to collecting new data, particularly in clinical research which can be expensive and time-consuming, and so ways of exploiting existing data must be used, which this section explores.

Single split

A common compromise is to split an existing dataset into a training and a test set, the former being used to train a model, the latter simulating the process of collecting a new dataset and being used to test the performance of a model. This requires holding out a proportion of the data, and so the test set is sometimes called the hold-out set. If hyperparameters also need to be chosen an additional set called the validation set can also be split from the data. This can be used to select the best hyperparameters, and then the generalisation error can be estimated using the test set. The partition of the training and validation sets can be understood as the former being used to learn the model parameters while the latter is used to choose the model hyperparameters.

However, these single splits can compound the problem of small datasets as less data is used to train the model. In particular, though the method is relatively unbiased it suffers a high degree of variance [155]. This means that, even under ideal conditions in which the train, validation and test sets are all drawn from the same underlying distribution, the generalisation error of the test set can vary significantly, especially when the validation/test sets are relatively small. Hence, we seek alternatives.

Cross Validation

An extension of single training/test splitting is k-fold cross-validation (CV). This repeats the training/test split k times so that k models are trained and tested on k sets of disjoint data, each one producing an estimate of model performance. The average performance is reported, sometimes with an accompanying measure of variance. Taken to its logical extreme is leave-one-out CV (LOOCV) in which the test set comprises of a single datum (i.e. a spectrum), leaving the model with the maximum amount of data on which to train. LOOCV theoretically should provide the least biased performance, as it incorporates the largest amount of training data, and with the lowest variance, as it only shifts one sample across folds [156]. However, this assumes that CV is averaging independent estimates, but the reality is that the samples may be highly correlated (which Raman spectra are known to be). LOOCV

could then struggle to detect model instabilities caused by changing the dataset as only one sample is changing at a time. Beleites *et al.* empirically showed that, in the context of spectroscopy, LOOCV becomes pessimistically biased as sample size decreases, in addition to having a high variance [157]. Consequently a k-fold CV strategy may be preferable, though the precise number of k depends on a number of interacting factors such as the sample size, SNR and the model used, which are not trivial to reconcile.

An alternative to CV is bootstrapping. In this method, instead of splitting the data into disjoint folds, the training set is constructed by randomly sampling from the data with replacement (so the same datum could be selected multiple times), with all the data not so selected being put into the test set. It has been argued that this method is more robust than CV as the repeated random sampling is less likely to inadvertently bias the sets compared to a single validation split [158]. However, there are extensions to CV which mitigate this weakness.

A single instantiation of k-fold CV will produce k disjoint test sets to assess the performance of a model. It is possible to repeat this process such that each new instantiation contains different combinations of samples in the training set. This is known as repeated CV and has been shown to reduce the variance of the generalisation error [157, 159]. Repeating k-fold CV is computationally expensive and becomes unfeasible for particularly large datasets or computationally demanding models. For this reason, CV is not as common for CNNs compared to other classification models used in RS, instead relying on more biased single train/validation/test splits. This is computationally tractable, but leads to less stable estimates of the generalisability of a model, hence overfitting may not be detected.

Stratified CV is a process that can be added to the above strategies which ensures that after data splitting an approximately equal number of samples is present from each class. This can be useful for small datasets which, if randomly split, could result in only one class being present during training [160]. It could also be useful for particularly imbalanced datasets, where minority classes could be removed entirely from the training set. This can ensure at least one sample is present in the training

set from each class, though this imposes a limit on the size of k during k-fold CV (and cannot be reconciled with LOOCV), as if k exceeds the size of the minority class it will not be possible to ensure it is present during training.

The above shows that the choice of validation strategy influences the interpretation of results as different strategies have different amounts of bias and variance in their generalisation estimates. In the following section we return to the literature reviewed in section 1.5 to determine if any lessons can be drawn before conducting my own experiments upon the the Lynch and Ovarian datasets.

5.2.1 Lessons from the literature: validation

Nearly all of the reviewed studies conducted some kind split upon the data in order to produce train, test and, in some cases, validation sets. Several partitioning strategies were used (figure 5.1). Most common was k-fold CV, either 5-fold or 10-fold. These are common default values in ML as they have previously been found to produce a good balance between bias, variance and computational cost [159]. However, they are somewhat arbitrary choices and, as discussed above, what would constitute the optimal strategy is a nuanced topic. Most of the reviewed studies had low sample sizes, with an average of 82 subjects. When the sample size is small LOOCV has been shown to have a high bias and variance, while k-fold strategies had a lower bias and their variance can be further reduced by performing repeated splits [157]. Hence the latter might be preferred. Unfortunately due to the heterogeneity of the studies, it is difficult to draw comparisons of CV strategies across them. One study explicitly explored the difference in performance between LOOCV and k-fold CV. Jeng *et al.* found that LOOCV yielded higher accuracies than k-fold CV (the value of k was not specified) for a binary classification task, but this was reversed in a three class problem. No attempt was made to compare the variance or bias of these performances and so little can be inferred from this study alone.

Two of the reviewed studies performed repeated validation: Serzhantov *et al.* repeated a single 50/50 train/test split 1000 times [85], while Fang *et al.* performed 10-fold CV 500 times, giving 5000 model estimates [100]. However, it was not the purpose of either study to investigate the effect this method had on the bias or

variance of the generalisation error and so was not explored. With a large enough dataset this could be estimated by comparing models trained on partitions of the data to a model trained on an entire dataset [159].

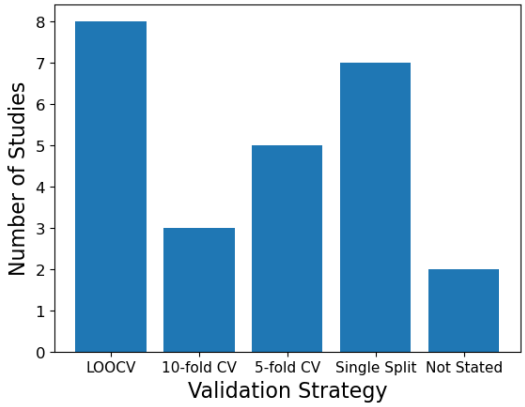


Figure 5.1: Validation Strategy used in the reviewed literature

The recent literature offers little exploration of the unique structure of oncological Raman datasets, and so in the next section I conduct my own experiments upon the Lynch and Ovarian datasets.

5.2.2 Cross validation experiments

Methods

I employ a stratified k-fold CV in these experiments, taking k-fold to its extreme where it becomes LOOCV. Here stratification is not possible as there is only one sample left out. This has been repeated 5 times, balancing the reduction in variance this brings with increased computational cost. With this CV strategy, the three models were all held constant in their hyperparameters so that the only variable changing between experiments was the number of folds.

The outcomes of interest are the variance and bias of the generalisation error, approximated here with accuracy for ease of interpretation. However, measuring bias requires some estimation of the ground truth. With a large enough dataset a large portion of the data could be used to create a model and use its score as a proxy for the 'true' score, then smaller portions used to create train and test splits. Unfortunately,

even the SMART dataset is not large enough for such an experiment. Without such an estimate of the true performance we can only say whether the bias is changing and so we shall note any trends in overall accuracy.

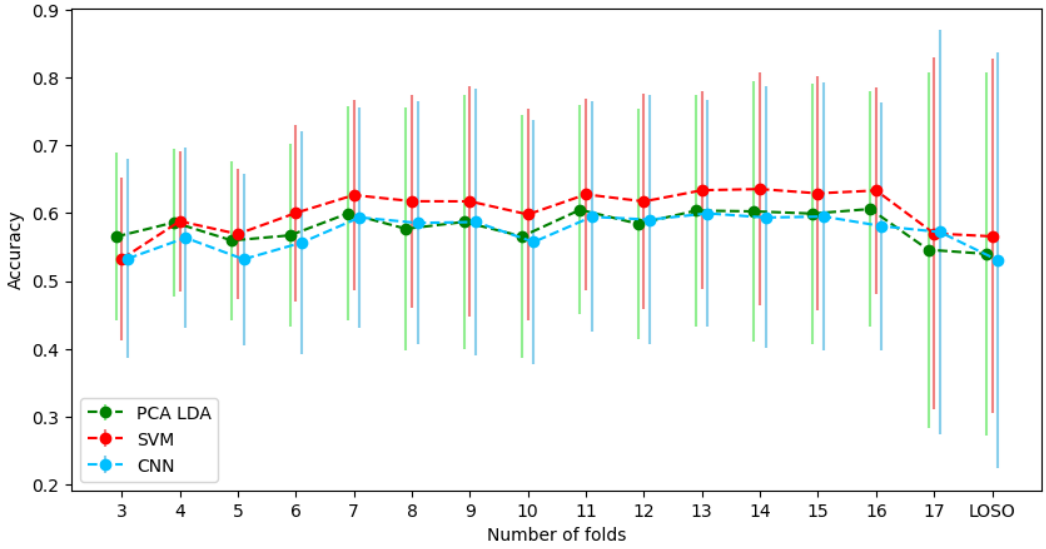
Results and Discussion

As can be seen in figure 5.2, the lower the value of k the less variance in the estimate of performance. This is consistent with evidence which suggests that in highly correlated datasets, variance increases with k [156]. This is due to the dependence of the training data; the different folds used in estimating performance reuse a certain portion of the data, thus violating the assumption of independence which is normally made when estimating variance. The mean accuracies also remain largely consistent. This implies that any bias, which is likely to be present, is not being overly affected, either optimistically or pessimistically, as the number of folds changes. Also of note is that these findings are consistent across all three models trained upon two different datasets, adding credence to the idea that the principles behind CV are invariant across models [157].

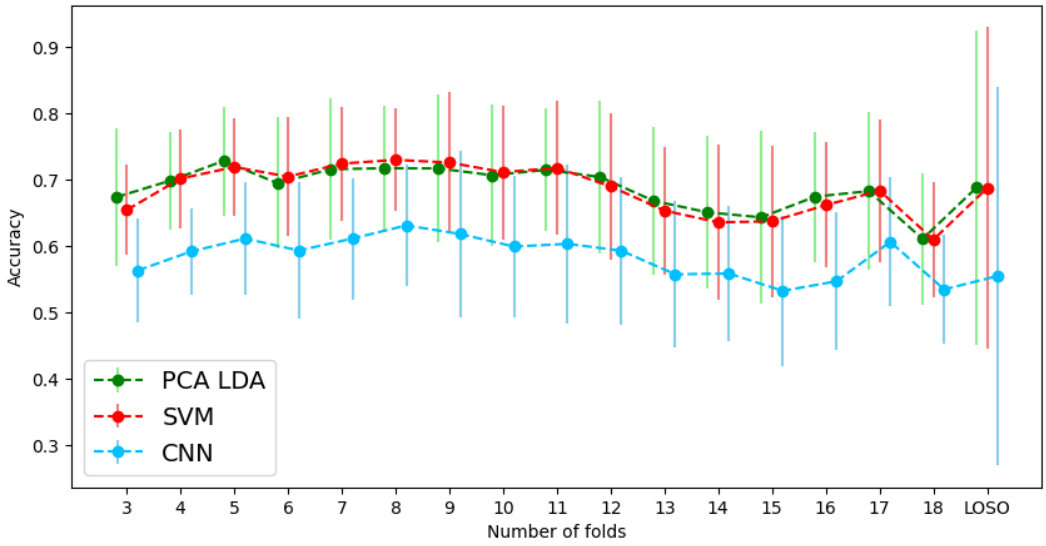
Based on these empirical results, together with theoretical considerations, I proceed with a 3-fold CV strategy, stratified to ensure that minority classes are always present in the training data and that no class is omitted with the smaller datasets, repeated 5 times to reduce the variance of estimates while not demanding too much computational cost. In the rest of this section various *in silico* experiments are conducted. They will use this 5 x 3-fold CV strategy unless otherwise stated.

5.3 Pre-processing

Pre-processing refers to any changes made to the original Raman data. This is usually done to remove artefacts in the data, making it more amenable to downstream analysis.



(a) Ovarian Data



(b) Lynch Data

Figure 5.2: Mean accuracy of PCA-LDA, SVM and CNN by the number of folds used during k-fold CV. Error bars indicate +/- 1SD, smaller bars suggesting less variance. Bias manifests as differences in mean performance, though true bias is unknown. Note that the last value along each x-axis corresponds to the number of samples in the dataset and so represent leave-one-sample-out CV.

Cosmic Ray Removal

Cosmic rays are high energy particles that originate from extraterrestrial sources.

Occasionally such a particle will interact with a spectrometers CCD which manifests as a spike on a spectrum, which could be confused for a real feature. Fortunately, cosmic rays are usually conspicuous, and removing them with good fidelity is relatively easy using one of several algorithms. For all three datasets, this was performed automatically using WiRE softwares 'median filtering' cosmic ray removal.

Smoothing

Smoothing is an optional pre-processing step applied to particularly noisy spectra. It removes noise by passing a low-band filter over the wavenumber axis and aggregates the intensity within the filter by some method such as the mean, median or a polynomial function. The circumstances under which spectral smoothing provides benefits are known to be narrow, and can worsen performance by exacerbating correlations within the dataset [161]. No smoothing was performed on any of the datasets in this thesis.

Saturated Spectra

The pixels of the CCD (not to be confused with image pixels) accumulate charge at a given wavenumber which is counted after a given time. However, there is an upper limit to how much charge each pixel can hold. When this limit is met, the pixel becomes saturated and the count reads as zero.

In the Ovarian datasets these were removed by identifying spectra that had more than 20 contiguous wavenumber regions at zero. In the SMART dataset they were flagged automatically by the WiRE software and removed. The Lynch dataset consists only of point spectra, allowing manual inspection of each spectrum in real time and any saturated spectra were discarded and a new spectrum obtained.

Normalisation

Normalisation is usually applied in RS to ensure that analysed spectra are independent of different Raman scattering collection geometries which would otherwise lead to varying spectral intensities. This can mitigate some of the effects of varying instrument conditions, such as alignment and laser power [162]. In the context of machine learning normalisation ensures that no single variable is orders of magnitude greater than others, which would effectively bias the model towards that variable during training. Some models, such as tree based algorithms are robust to normalisation, but the three models used in this thesis all require normalisation. Figure 5.3 shows the same class average spectra as in figure 2.3 but baseline corrected, as discussed in section 5.3, and Standard Normal Variate (SNV) normalised, as might be performed in a typical Raman study pipeline [162] The comparison shows how subtle spectral differences can be accentuated and spectral variance reduced by pre-processing, which may facilitate downstream analysis.

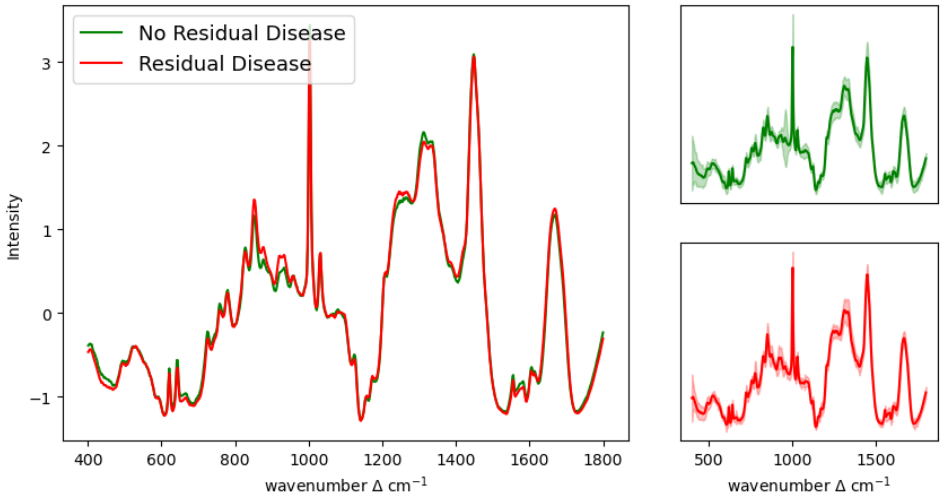


Figure 5.3: An example of processed spectra from the Ovarian dataset, baseline corrected via ModPoly fit (4^{th} order polynomial) and SNV normalised

There are many types of normalisation. Unless otherwise stated, in this thesis I perform SNV normalisation:

$$\frac{X - \mu_X}{\sigma_X} \tag{5.1}$$

where X is a single spectrum, μ_X is its mean and σ_X its standard deviation.

This results in a spectrum with 0 mean and unit variance.

Baseline Correction

The raised baseline often observed in Raman spectra of biological origin is usually attributed to fluorescence and manifests as much broader spectral features compared to Raman peaks [163], though it may also contain other contaminants. There is an abundance of baseline correction techniques used in the Raman literature to correct for this elevated signal including; asymmetric least squares, polynomial fitting with modified, iterative and alternating variations, 'rubberband' algorithms and 'rolling ball' algorithms, to name just a few. Even studies focussed solely on exploring the effect of different baseline correction techniques can only analyse a subset of these methods, not to mention the numerous parameters that each may require [164, 165, 166]. Thus, it is clear that the choice of algorithm and its parameters cannot be exhaustively searched and will involve some degree of arbitrariness. These choices can have a profound impact upon downstream analyses. This has been demonstrated in headline science news with the recent discovery of phosphine in the Venusian atmosphere, indicating the presence of extraterrestrial life [167]. It was shown that the spectroscopic phosphine signal was an error caused by the choice of polynomial baseline with which to correct the data [168].

It has even been suggested that the fluorescent component of a Raman spectrum, hitherto regarded as noise, may actually contain diagnostically valuable information. Gaifulina *et al.* found that the broad fluorescence signal was correlated to degeneration of human cartilage [169]. If this is the case, then baseline subtraction may be removing clinically relevant information. Indeed, there is evidence suggesting that most pre-processing techniques, and their numerous parameters, actually worsen subsequent classification performance [164]. Section 5.3.2 explores experiments with various baseline correction techniques, but first we look at pre-processing in the Raman oncology literature.

5.3.1 Lessons from the literature: pre-processing

The literature review conducted in section 1.5, identified numerous pre-processing techniques for RS oncology data. Finding the best pre-processing method and parameters is often left to trial and error, or sticking with what worked well in the past. Although more rational approaches exist, such as searching with a genetic algorithm [164], removing the need for pre-processing is attractive. Hence, some of the reviewed literature explored the potential of CNNs to automatically perform feature selection and pre-processing at the same time. However, most ML models require normalisation, hence in the subsequent discussion, references to 'raw' data include normalisation. Four studies explored a suite of pre-processing steps against raw data.

Lee *et al.* compared 'baseline corrected' data to raw data and found that for traditional ML models baseline correction improved performance [97]. However, the CNN performed better on the raw data with an accuracy of 96.56% +/- 0.91% compared to 90.22% +/- 0.50%. The best performance on pre-processed data was PCA-QDA with an accuracy of 95.00%. The precise nature of the baseline correction used was omitted.

Fang *et al.* compared baseline correction and spectral smoothing via the Vancouver Raman Algorithm (VRA) to raw data. They show a learning curve in which the former method converges to 100% accuracy fractionally more quickly than the latter, though both ultimately give the same results. In this case pre-processing did not improve the performance of the CNN.

These studies suggest that CNNs can classify data without pre-processing. There is even a suggestion that the architecture is able to exploit diagnostic information present in raw data too subtle for traditional models to detect and is usually discarded. However, this finding was not ubiquitous.

Yan *et al.* pre-processed data by smoothing with a Savitsky-Golay filter and baseline removal via asymmetric weighted penalty least squares [90]. Compared to raw data, pre-processing improved CNN accuracy to 98.75% from 96.70%.

Wu *et al.* found similar results. They performed baseline correction and spectral smoothing using the VRA and compared this regimen to raw data [99]. The former regime outperformed the latter across all subset analyses. This includes the traditional models, KNN, RF and SVM, but the largest decrease in performance was observed in the CNN (accuracy: pre-processed 81.3%, raw 75%).

In all the above cases, where traditional ML models were used, they were improved by pre-processing. Although some of these studies suggest that CNNs could preclude an explicit pre-processing stage, this is not clearly established in the literature. There are a plethora of pre-processing techniques and it may be that some techniques were better suited to particular datasets than others, i.e. the pre-processing step has introduced overfitting and CNNs trained on raw data would generalise better when clinically deployed, even though their performance would be worse during training and testing. This is just speculation until it is more thoroughly explored.

Bjerrum *et al.* commented on an apparently contradictory finding, that first removing baseline offsets during pre-processing via Extended Multiplicative Scatter Correction (EMSC), and then randomly adding this back into the data during data augmentation, yielded better results than just augmentation alone [170]. They speculated that the CNN was forced to focus on features invariant to the shortcomings of EMSC while still benefitting from baseline correction. Augmentation is discussed in section 5.7.

The order in which these pre-processing steps occur also needs considering. For instance, if SNV normalisation is first performed followed by baseline correction, then the subsequent spectra will not have zero mean and unit variance as intended. The order of pre-processing was inconsistently reported in the literature. Where explicitly reported this followed; truncation to a given wavenumber region, removal of saturated spectra, removal of noisy spectra, baseline removal, outlier removal and normalisation [82], or; truncation, outlier removal, MSC normalisation, removal of noisy spectra, baseline correction, smoothing and min-max normalisation [84].

These findings are unfortunately inconclusive, thus in the following section I take the Ovarian and Lynch dataset to perform some baseline correction experiments.

5.3.2 Baseline correction experiments

The number of baseline correction techniques is too great to search exhaustively, and there is no *a priori* reason to think that any one technique will work better than another for any given application. Hence I investigate a few techniques, which are well established in RS and are relatively easy to perform. The first is the famous 'Mod Polyfit' method of Lieber and Mahadevan-Jansen in which a modified polynomial is iteratively fit to a spectrum using a least-squares-based polynomial curve fitting function which ignores Raman peaks [171]. This requires selecting a single parameter: the order of the polynomial to fit. An extension of this is 'Improved Modified Multi-Polynomial Fitting' (I Mod Polyfit) [172]. This was designed to improve fluorescence background removal by being able to account for signal noise distortion and the influence of large Raman peaks. This method also has the parameter of polynomial order. Zhang *et al.* developed an adaptive iteratively reweighed Penalised Least Squares method which is similar to the previous methods but has the desirable trait of not needing any parameters [173]. Finally, Extended Multiplicative Scattering Correction (EMSC) was developed to correct for additive baseline effects, multiplicative scaling effects and interference in near infrared (NIR) spectroscopy, but has been found to be useful in Raman applications for fluorescence removal [174].

I took each of these techniques, with a range of parameters, and applied them to the Ovarian and Lynch datasets. The polynomial order ranged from 1-5. Higher orders were not explored due to a mathematical artefact known as Runge's phenomenon which causes increasingly severe distortions to the tails of Raman spectra as the order of the polynomial increases.

Methods

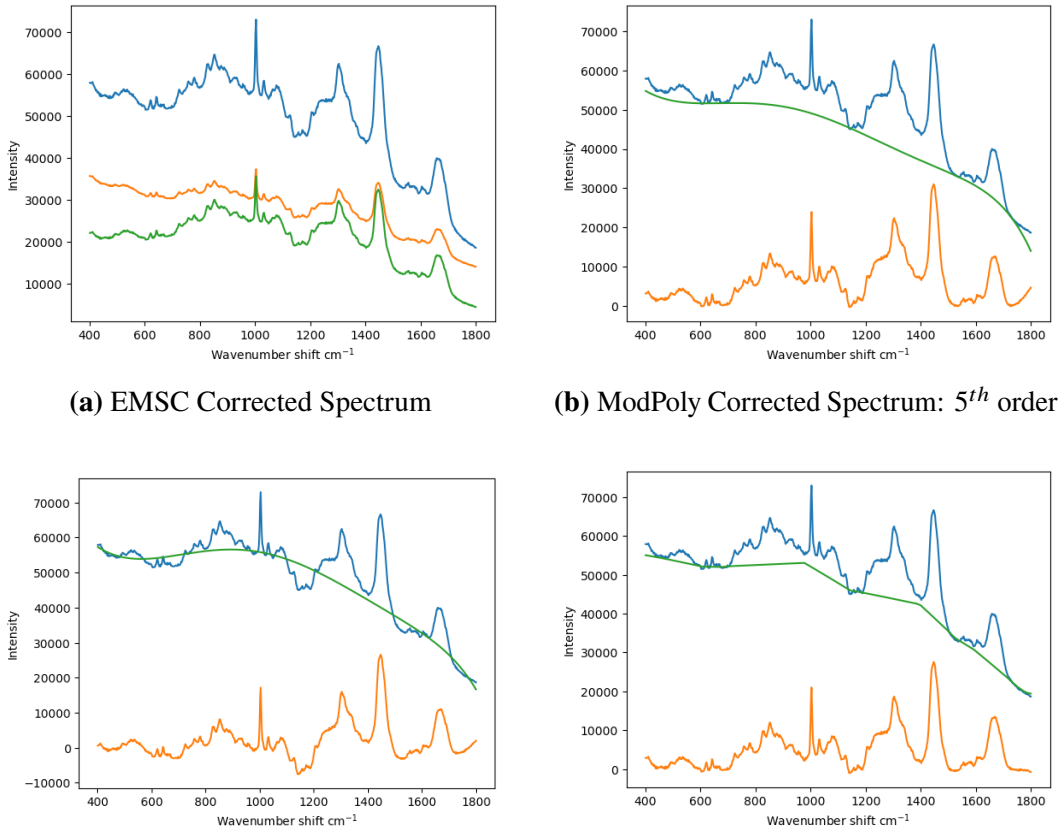
Data preparation for this experiment was identical for the Ovarian and Lynch

datasets. Thirteen baseline correction methods were employed; one with no correction, 5 using Mod Polyfit with order of 1-5, I Mod Polyfit with order of 1-5, Zhang's method and EMSC. These were analysed using all three models (PCA-LDA, SVM and CNN) with all other hyperparameters held constant, allowing only the baseline method to vary. The 5 x 3-fold CV strategy was used for a total of 15 folds per correction method. The mean accuracy and SD of each method was calculated. Default values for the ML hyperparameters were used and so the accuracies reported here are not necessarily representative of final model performance, and it is the change in performance as a function of baseline correction method that it is pertinent.

Results and Discussion

Figure 5.4 shows the same spectrum under the different baseline correction methods. Of note is that the Zhang method performs piecewise linear-fitting over the wavenumber region. The figure also shows how EMSC uses the average spectrum from the dataset to correct a single spectrum. For the polynomial fitting methods, it could be argued that a larger order of polynomial may be better able to fit to the contours of the fluorescence component. However, there is no objective way of knowing when over-fitting has occurred. As the methods are not dynamic, so do not make allowances for any particular spectrum, what may work well for any one spectrum may not work well for all spectra, even in the same dataset. For this reason, it is prudent to consider the effects of these methods upon averaged spectra.

Figure 5.5 shows the results of baseline correction upon the mean Lynch spectrum via four different methods, two of which span five parameters. By inspection it is not clear which method and parameters are most 'correct'. Of note is that some methods will return negative values at certain points in the spectrum. This was one reason why SNV normalisation was performed, which is not affected by negative values (some forms of normalisation are only defined for non-negative values and so would require an additional step of processing). A similar story is apparent in figure 5.6, which shows the mean Ovarian spectrum by the same baseline correction methods.



(a) EMSC Corrected Spectrum

(b) ModPoly Corrected Spectrum: 5^{th} order

(c) IModPolyCorrectedSpectrum: 4^{th} order

(d) Zhang Corrected Spectrum

Figure 5.4: Baseline Corrected Spectra, showing the same original spectrum from the Ovarian dataset (blue), the baseline corrected spectrum by different methods (orange) and the removed component (green).

In this instance the EMSC had a far more 'correcting' effect compared to the Lynch data. EMSC also involves a choice of polynomial order for baseline subtraction. For both the Ovarian and Lynch data this was fixed at 4. The differences in the corrected spectra show how sensitive the technique is to such choices. Often, automated software packages will simply use a default value, obscuring this parameter. For the Lynch data, the choice had little impact.

Figure 5.4 shows the same spectrum under the different baseline correction methods. A larger order of polynomial may be better able to fit to the contours of fluorescence, though there is no objective way of knowing when over-fitting has occurred. Of note is that the Zhang method performs piecewise linear-fitting over the wavenumber region.

One way to tease out the biochemical differences between a collection of spectra

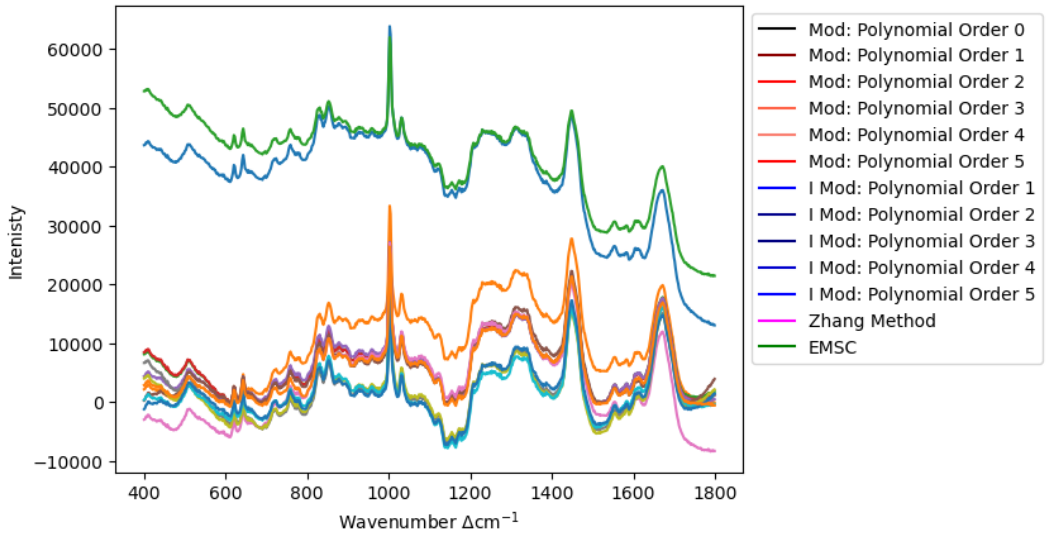


Figure 5.5: The mean spectrum from the Lynch dataset subject to four different baseline correction techniques, two of which also have varying parameters.

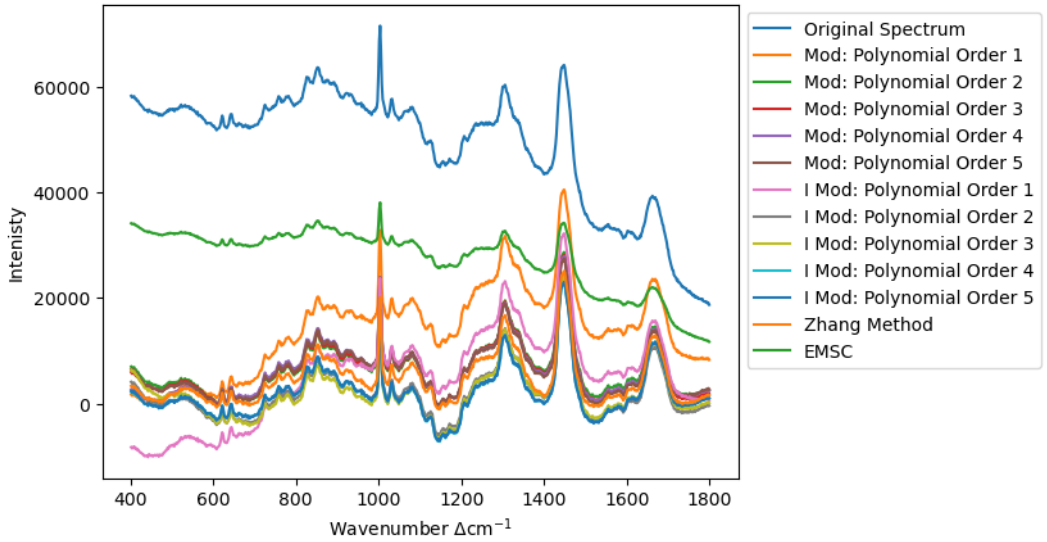
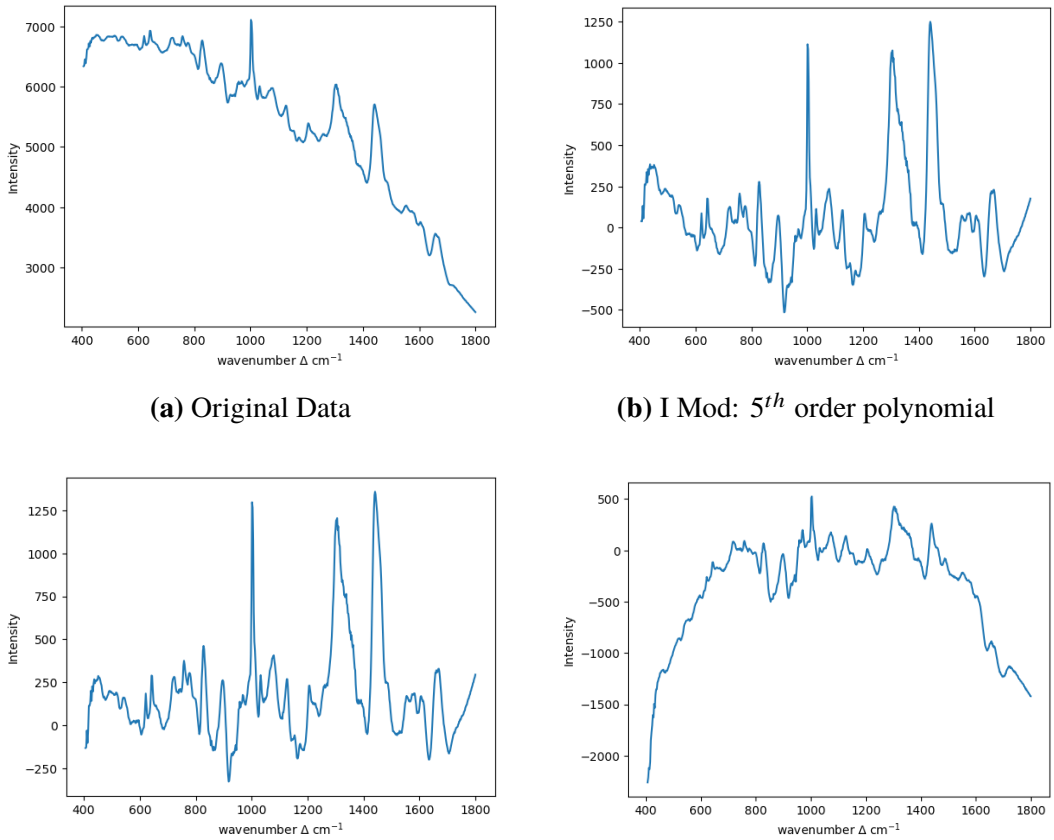


Figure 5.6: The mean spectrum from the Ovarian dataset subject to four different baseline correction techniques, two of which also have varying parameters.

is to take the difference spectrum. This involves taking the average spectrum for each class in a dataset; residual and no-residual disease in the case of the Ovarian dataset. We then subtract from the no residual disease (the null case) the residual disease spectrum. What is left is the difference between the two groups, which highlights which peaks distinguish the groups. Applying this method to four different cases: the original spectra, and one from each baseline correction method, choosing the

best polynomial order (as defined by returning the highest accuracy), we can see the differences in figure 5.7.



(a) Original Data

(b) I Mod: 5^{th} order polynomial

(c) Mod: 4^{th} order polynomial

(d) Zhang Method

Figure 5.7: Difference spectra according to baseline correction method of the residual and no-residual disease groups from the Ovarian dataset

We can see that baseline correction in general has a significant effect upon the visual appearance of the difference spectra. The most striking differences between the methods occur at the tails of the difference spectra, with evidence that Runge's phenomenon is still causing mis-fitting, despite the relatively low order of the polynomials. Aside from the tails, there are important differences between the spectra which would lead to differences of interpretation of the biochemical constituents that distinguish the two groups.

A more quantitative metric is to assess the impact upon model performance. Figures 5.8 and 5.9 shows the accuracy of the three models with varying baseline line correction methods on the Ovarian and Lynch datasets respectively.

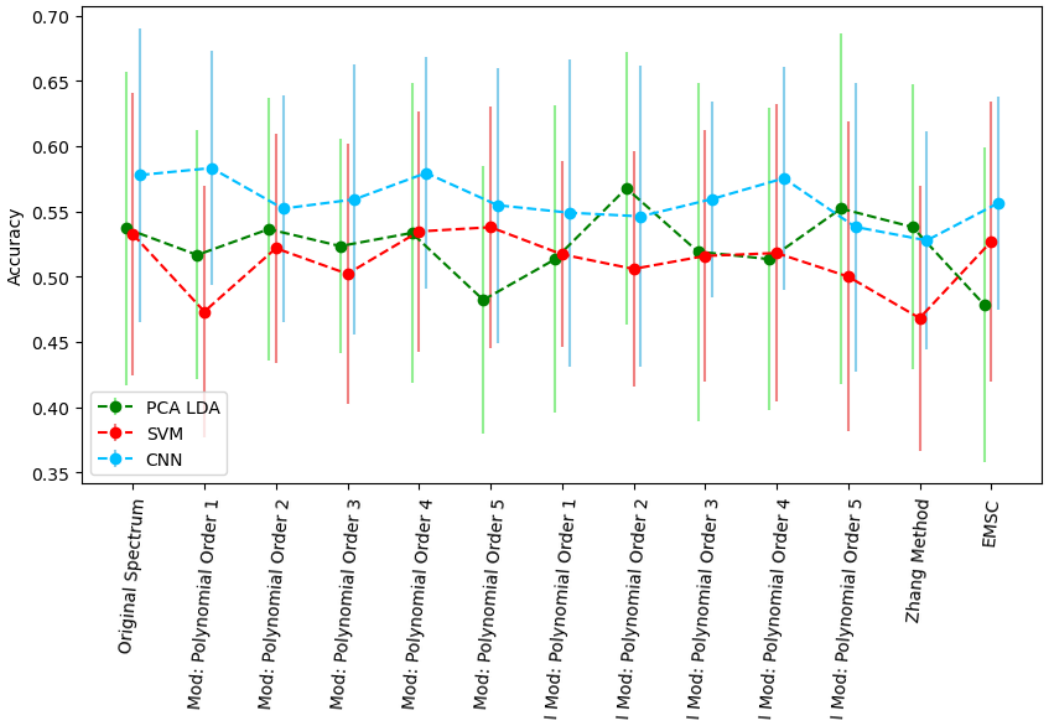


Figure 5.8: Model accuracy by baseline correction method: Ovarian dataset. Average value over 15 folds and +/- 1 SD bars.

These graphs show that different ML models may be better suited to different baseline correction methods and parameters. For instance, with the Lynch data the CNN seemed to improve with a small degree of ModPoly correction, though other models did not likewise improve. However, the Ovarian data did not replicate this trend and any interpretations are tentative, as it is difficult to say that any of the differences observed here are large enough to be considered significant in light of the SD bars.

We may additionally construct scatter plots of the data separated by class to visually represent the degree of separation, and see if this is improved by baseline correction. In figure 5.10 this is done via PCA plots of the first two principle components (representing 87% and 84% of the variance for the uncorrected and the best baseline corrected data respectively). This shows that there is considerable overlap in the classes particularly between the MSI-H and MSS groups - consistent with traditional methods which struggle to distinguish these two groups. This shows that a linear separation of the data would be difficult. But most pertinent in this

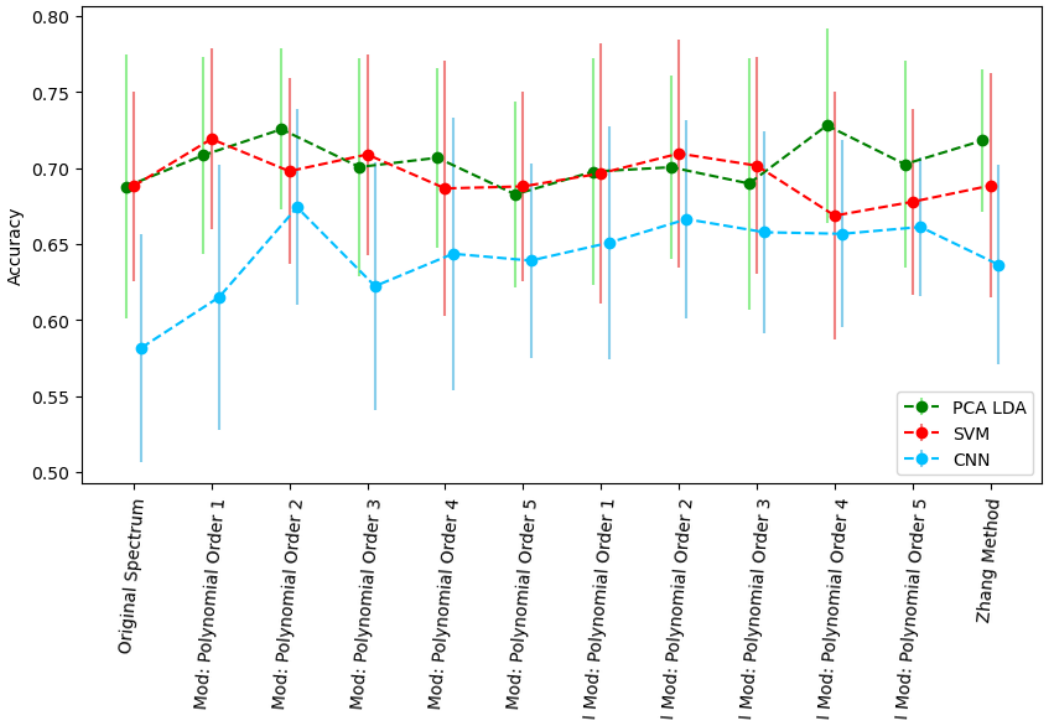
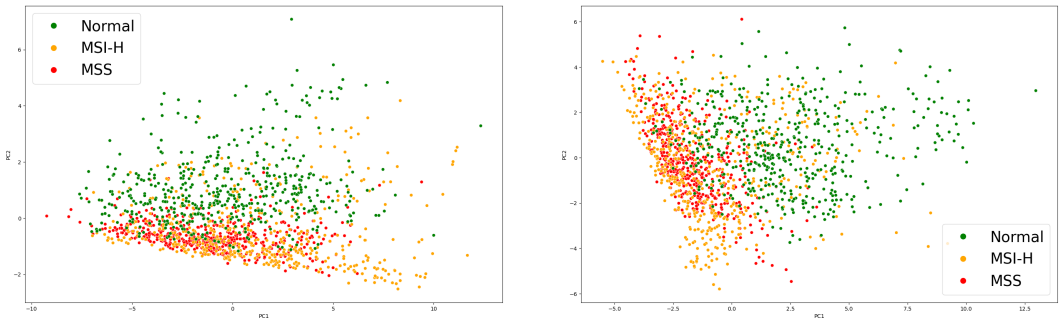


Figure 5.9: Model accuracy by baseline correction method: Lynch dataset. Mean value over 15 folds and +/- 1 SD bars

section is that this is not improved by baseline correction.



(a) Uncorrected data PC1 and PC2 **(b)** Baseline corrected data PC1 and PC2

Figure 5.10: Visualising class separability of uncorrected Lynch data and baseline corrected (Mod poly 2^{nd}) data via PCA plots

Overall, there is little appreciable difference in performance across the various baseline correction methods. Added to the inconclusive findings in the literature and considering that altering Raman spectra can change the interpretation of biochemical differences, it is most conservative to apply the principle of Occam's razor and to cut out this pre-processing step.

Proceeding without baseline correction, the order of pre-processing in subsequent analyses, unless otherwise stated, is cosmic ray removal, removal of saturated spectra (via 3 x 3 median filtering of adjacent spectra), both conducted on the manufacturer's software, followed by truncation to the wavenumber region 400-1800 cm^{-1} and SNV normalisation.

These pre-processing steps are particular to Raman data. There are additional data preparation steps which need to be taken into consideration due to the medical nature of the data. The subsequent sections explore three of these, including the hierarchical structure of the data and whether the datasets are balanced in terms of their class labels. Understanding these is essential in order to properly split the data during model training such that bias is not introduced into the results. These issues are generally exacerbated by small sample sizes. One strategy to ameliorate this problem is data augmentation, which can significantly inflate the training sample size for CNNs by creating synthetic samples. However, this technique comes with its own risks and so is carefully appraised in section 5.7.

5.4 Hierarchical structure of biomedical Raman data

As described in section 2, for all datasets in this study multiple spectra were taken from samples. In addition to this, for the Lynch and SMART datasets, several samples were occasionally taken from the same subject. And unique to the SMART dataset, three clinical sites collected ostensibly the same data. This introduces a hierarchical structure to the data, illustrated in figure 5.11, which must be considered during analysis. In this section I first consult the literature review of section 1.5 to ascertain current practices and then perform my own experiments on the Ovarian and Lynch datasets.

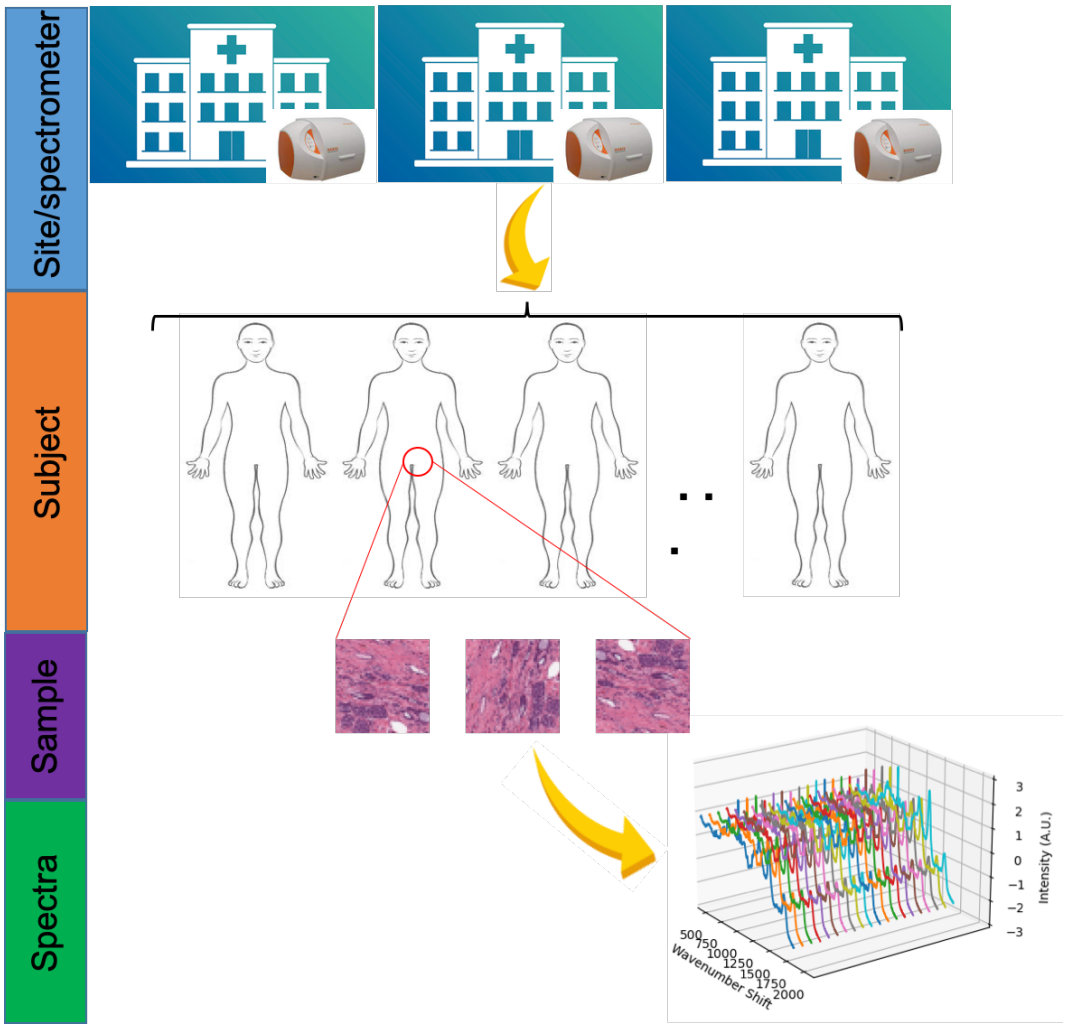


Figure 5.11: Top level: Clinical setting and/or instrumentation. Second level: Patient. Third Level: tissue sample. Fourth level: spectra. Only the SMART dataset contains all four level, the Lynch and Ovarian data being single centre.

5.4.1 Lessons from the literature: hierarchical splitting of the data

Most of the reviewed studies classified individual spectra. Many such spectra were often taken from a sample, and several samples were sometimes taken from the same subject. This raises the question of at which level is it appropriate to split the data during CV: spectra, sample or subject. If split at the level of spectra, this could mean that spectra belonging to the same sample and/or subject are present in both the training and test set (figure 5.12). This could lead to overly optimistic estimates of

the generalisability of the model as it is not a realistic assessment of the model, which in the clinical setting would be classifying spectra obtained from unseen samples.

Of those studies that split the data at the level of spectra the best accuracies were: 90%, 94.8%, 96.56%, 97.7%, 93.8% and 94.8%. Of those studies in which the level of split was not explicitly stated the accuracies were: 99.0%, 98.75%, 96.90%, 99.54%, 92.00%, 100%, 80.2%,99.0%, 90.5% and 96.7%. And of those studies splitting data at the level of subject or sample the accuracies were: 84.43%, 99%, 81.75%, 98.23%, 83%, 87%, 92.89%, 74.95%, 81.3% and 65.5%. No attempt has been made to statistically compare these groups as might be performed in a meta-analysis as the various study aims and methodologies are too heterogeneous to make such comparisons statistically valid. However, it can qualitatively be seen that studies which split data at the level of subject or sample tend to report lower accuracies than those that split data at the level of the spectra, or those that do not explicitly state the level of the split. Splits at the subject or sample level likely reflects more realistic assessments of how well the model would perform in the clinical setting. Of particular note is the study by Wu *et al.*, the only study which explicitly compared methods of splitting upon the same dataset. They found a drop in performance of 12.5% in the overall accuracy when splitting at the sample level [99]. This is consistent with findings from Guo *et al.* who explicitly examined the difference that the level of the split makes during CV with tumour cell lines, concluding that the highest hierarchical level of the dataset should be used when partitioning the data [175]. There is reason to believe that some of the high accuracies reported in the literature are due to not taking into account the structure of the data and splitting accordingly.

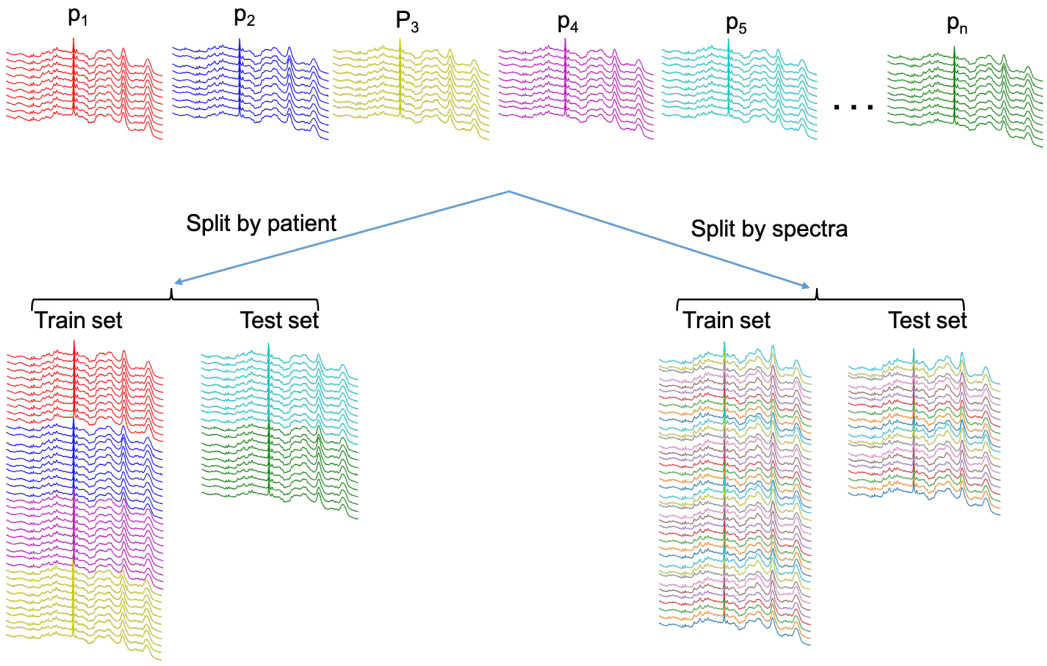


Figure 5.12: Spectra versus patient data splitting. P_n refers to the n^{th} patient. Of note is how the test set split by spectra includes all patients also contained in the train set.

In many of the studies one sample was taken per subject. However, some studies took multiple samples from the same subjects, which introduces an additional strata into the hierarchy. For instance, Zuvela *et al.* took 113 samples from 60 patients [93] and Shu *et al.* sampled 888 sites from 418 subjects [95]. Both studies split the data at the level of samples rather than the highest level, subject, so it is not possible to ascertain what impact this may have had on subsequent analyses.

Two studies classified only the average spectrum from a single sample, thus flattening the hierarchical structure of the data and bypassing this issue [75, 80]. Both studies took spectra from serum samples and analysed the data using traditional chemometric models. Neither study examined how taking the average spectrum per sample compared to using all sample spectra. Jeng *et al.* did compare the performance of using average spectra vs all five spectra of a single sample in their PCA-QDA model, finding the former method had an accuracy of 88.75% vs 83.00% [76]. We shall return to this in section 5.5; in the following sub-section we investigate how taking into account the hierarchical structure of the data influences outcomes.

5.4.2 Hierarchical splitting of the data: experiments

In this section I explore the impact that the level of split during CV has upon the accuracies obtained from the Lynch and Ovarian datasets.

Methods

The three models of PCA-LDA, SVM and CNN were again used for this experiment, as described above. The only difference made during training was that in one instance the data were split randomly during CV as would typically be done for data with a flat structure (i.e. no hierarchy). In this CV strategy spectra from the same subject and sample could be present in both the training and test set. In the other instance CV was performed such that all spectra belonging to the same sample would remain as a single block so that all would either be in the training or test set. As with the baseline experiments, default values for the model hyperparameters were used and only the method of CV splitting changed: splitting by spectra or splitting by subject.

Results and Discussion

As can be seen in table 5.1, splitting by subject significantly reduces the accuracy achieved across both datasets and across all models. This is consistent with findings in the literature review; both from those studies which explicitly investigated this strategy and implicitly in the numerous studies which employed various strategies with those splitting by spectra reporting generally higher accuracies. This accords with the theoretical concern that splitting data at a lower level means that the independence assumption is violated, as spectra from the same sample would be present in both the training and test sets leading to an over-estimation of the generalisation error. We can conclude that splitting by subject leads to more accurate estimates. Also of note is the variance achieved in these results. The accuracies achieved when splitting by spectra have a generally small standard deviation compared to when splitting by subject. This is likely a manifestation of the former strategy incorporating subjects in both the training and test sets, whereas the latter strategy is forced to attempt to

generalise to previously unseen subjects. This is a more realistic replication of what the models would encounter in the clinical setting, and also demonstrates that patient variability is a significant contribution to the uncertainty of the generalisation error with small datasets.

Model/Dataset	Accuracy: Spectra Split	Accuracy: Subject Split	Difference in performance
PCA-LDA			
Lynch	85.32% (+/- 1.56%)	68.24% (+/- 7.10%)	17.08%
Ovarian	78.65% (+/- 1.69%)	60.00% (+/- 11.86%)	18.65%
SVM			
Lynch	85.63% (+/- 1.49%)	65.71% (+/- 7.31%)	19.92%
Ovarian	80.21% (+/- 0.91%)	51.07% (+/- 9.91%)	29.14%
CNN			
Lynch	70.62% (+/- 8.15%)	60.73% (+/- 6.05%)	9.89%
Ovarian	83.37% (+/- 2.45%)	64.29% (+/- 7.01%)	19.08%

Table 5.1: Accuracy according to how the data was split: by spectra or by subject

This firmly establishes the necessity of splitting data during CV at the correct level in order to appropriately measure the generalisation error. In the baseline correction experiment of section 5.3.2 the data were split by subject.

5.5 Accounting for sample heterogeneity

5.5.1 Lessons from the literature: sample heterogeneity

In the literature there were two methods by which sample heterogeneity is currently incorporated. Averaging spectra provides one means, while including multiple spectra from a single sample provides another. Indeed, all studies in the literature employed one of these strategies, regardless of the intended clinical use. There was no direct comparison of these methods in the literature reviewed. *A priori*, there is no reason to suspect that one approach will work better than the other and will likely depend upon the intended application. For instance, a mean spectrum typically has a higher SNR than individual spectra, and a model trained on the former may not generalise to applications requiring individual spectra to be classified, such as

post-surgery cancer edge detection. It is also an open question whether averaging spectra from a sample before analysis would be an effective use of data for deep learning models, which are notoriously data intensive. Overall, it is not clear that averaging spectra will always provide a benefit, and the decision to do so should take into account the nature of the application and the model being used.

For all applications in this thesis, it is the sample label which is pertinent to the pathologist as this will be what guides patient management decisions. This allows for an alternative method of classifying or predicting disease class: using the classifications of individual spectra within a single sample to 'vote' on the overall sample class. I shall call this method simple consensus classification. One way of doing this would be to classify each individual spectrum in a sample and then pick the most common label amongst all the spectra as the sample's predicted class. However, this is not indicative of current histopathology practice, in which even if only a small amount of tissue in a sample unambiguously shows the worst possible disease class, then the whole sample will be labelled as that class. A simple consensus voting would miss such cases. An alternative is to assess whether the worst disease class is present above some pre-defined threshold (20% was used in the subsequent experiment). If not, the next worst disease is assessed against the same threshold until a class is selected. I shall call such a voting system as a proportional consensus classification.

In the following section I shall explore 4 methods of taking sample heterogeneity into account: classifying each individual spectra, by the mean spectrum of each sample, by sample as determined by simple consensus and by sample as determined by proportional consensus.

5.5.2 Sample heterogeneity: Experiments

Methods

The Lynch and Ovarian datasets were used for this experiment, utilising the three models using the 5 x 3 CV strategy as defined in section 5.2. All model hyperparameters were held constant across each dataset. The CV strategy was

applied to each of the four above methods described above to account for sample heterogeneity. The method classifying individual spectra still ensured training/test splits were made at the appropriate level of the hierarchy.

Results and Discussion

Tables 5.2 and 5.3 show the performance by different method with the Lynch and Ovarian dataset respectively. Although the error margins are large and overlap, a general trend can be seen in that labelling by sample with simple consensus produces higher accuracy. However, classifying samples by their average spectra is largely competitive across most models and for both datasets. Given the computational advantage of this method, it is an attractive option. However, the speed up is not sufficient for the applications in this thesis to warrant even the slight drop in performance.

Note that the error margins tend to be narrower for the 'by spectra' method. This is because that method classifies thousands of individual spectra, and as sample size increases the variance of the sample mean decreases. As the other methods are taking a consensus over each entire sample they are classifying per sample rather than per spectrum and so have tens rather than thousands of classification attempts.

Lynch Data	By Spectra	Proportional Consensus	Simple Consensus	Sample Average Spectrum
PCA-LDA 15 PCs	73.1% +/- 6.7	71.3% +/- 10.2	82.6% +/- 12.4	79.6% +/- 9.7
SVM $c = 10$, $\gamma = 0.01$	72.6% +/- 15.7	70.0% +/- 12.1	74.7% +/- 10.9	71.4% +/- 15.7
CNN $LR = 0.001$, $BS = 64$	66.2% +/- 6.3	63.2% +/- 13.8	74.5% +/- 11.5	66.7% +/- 15.7

Table 5.2: Lynch data accuracy by method of sampling

As discussed in section 3.3, focusing on accuracy alone can obscure rather than clarify. In particular, to determine what difference the proportional consensus method makes we can look at the confusion matrices of the best performing model: the PCA-LDA model on the Lynch dataset. Despite the simple consensus model performing better in terms of accuracy we might expect the proportional method to

identify more true disease classes at the expense of giving more false disease classes. Indeed, the proportional consensus model gives far fewer errors in classifying the MSS class, 8% error compared to the 26% error of the simple consensus. However, it misclassifies MSI-H as MSS in 64% of cases, which is only 22% in the simple consensus method. There may still be cases where this is desirable. If the cost (human and financial) of misclassifying one disease class is particularly high then reliably identifying that class at the expense of misclassifying other classes may be preferable.

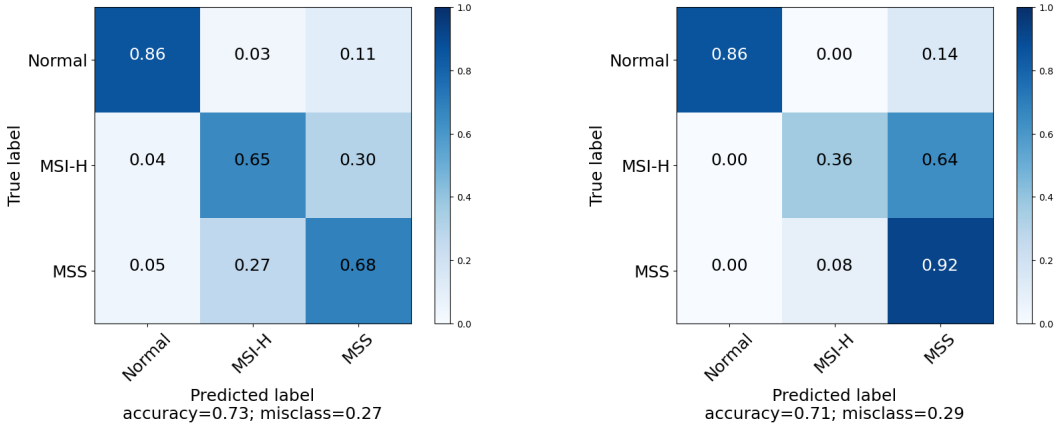
Ovarian Data	By Spectra	Proportional Consensus	Simple Consensus	Sample Average Spectrum
PCA-LDA 3 PCs	53.4% +/- 7.4	52.2% +/- 13.4	55.6% +/- 14.5	53.3% +/- 16.3
SVM $c = 10$, $\gamma = 0.01$	54.4% +/- 7.5	56.7% +/- 24.1	53.3% +/- 12.5	40.0% +/- 20.9
CNN $LR = 0.001$, $BS = 64$	49.5% +/- 15.0	44.4% +/- 19.9	50.0% +/- 23.6	47.7% +/-16.3

Table 5.3: Ovarian data accuracy by method of sampling

The choice of threshold for the proportional consensus in this experiment (20%) was arbitrarily selected. It may be a useful metric but would require expert histopathological input to choose a range of possible values, then further, independent, data to properly calibrate the choice. This is beyond the scope of this thesis, and so to evaluate the models in section 6, I shall use the simple consensus model to determine their performance. This results in improved performance, and also more directly mimics current histopathological practice for all three clinical problems under consideration in that whole slides, rather than individual pixels, are the unit of interest.

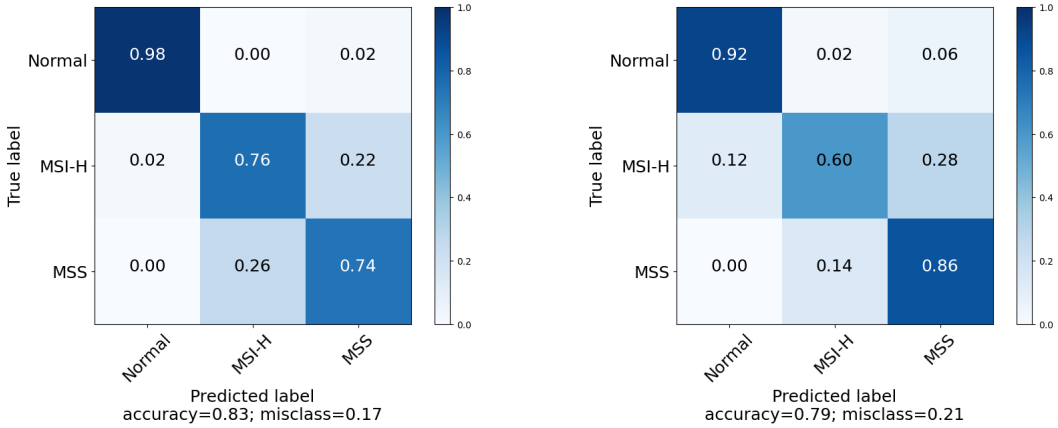
5.6 Imbalanced Data

Class balance refers to the distribution of classes within a dataset. In a balanced dataset the classes are approximately equally distributed; in an imbalanced dataset one or more of the classes has fewer members. It is common for medical datasets



(a) Classify by spectra

(b) Classify by proportional consensus



(c) Classify by sample: simple consensus

(d) Classify by average spectrum

Figure 5.13: Confusion Matrices by method of classification: taken from Lynch data PCA-LDA model

to have some degree of class imbalance as this often reflects the true distribution of disease classes in a population, or the distribution of cases presenting with a suspected disease. This is particularly true of rare diseases.

This can be a problem for ML if the aim is to optimise the training error regardless of class distribution. A toy example of how this becomes problematic is illustrated in a case where there are 100 samples, 99 belonging to a 'healthy' class and 1 to a 'diseased class'. A model need simply classify all samples as healthy in order to achieve 99% accuracy. This is also a lesson in the dangers of using a single metric to assess performance. For instance, the sensitivity of this toy classifier would be 0%, giving a clear indication of a problem. Regardless of the metrics, imbalanced datasets

in the medical literature have been shown to bias towards the majority class [176]. However, as discussed in section 3.2.1, disease class distribution represents important information from the real world, and this knowledge is critical in constructing the best possible classifier. Therefore, there is a tension between these competing concerns.

This tension has not been addressed in the recent oncological Raman literature (section 1.5), as research datasets tend to be reasonably balanced. Hence, we turn to the more general medical literature for insights into the problem.

There is no precise definition of what constitutes class imbalance as many factors interact to determine how much of a problem it is. For instance, Krawczyk showed that for classes with non-overlapping distributions which are linearly separable, then any amount of imbalance is acceptable [177]. However, this is an extremely unlikely scenario for medical datasets. Factors such as the degree of imbalance, the sample size and the degree of separability all interact to determine how much of a problem, if at all, it presents [178].

A common metric of class imbalance is the imbalance ratio (IR): dividing the highest class membership by the lowest.

$$IR = \frac{Instances_{majority}}{Instances_{minority}} \tag{5.2}$$

Studies can suffer a large degree of imbalance with $IR = 100$ or even $IR = 1000$. However, even studies with an $IR = 10$ have been found to hinder the training of models [178]. It has been shown that class imbalance becomes less of a problem as overall sample size increases (even if the IR remains constant) [179]. Unfortunately, medical sample sizes are typically very low compared to the general ML literature. Additionally, the degree of separability between classes determines how much of a problem imbalance is, with findings showing that as the degree of data complexity increases, so too does the susceptibility to any imbalance [180]. There is very little literature regarding multi-class classification with imbalanced datasets, which would only complicate matters [178].

The Lynch and Ovarian datasets were deliberately constructed to suffer from no class imbalance (i.e. $IR = 1$). This sampling strategy has been referred to

as separate sampling [130]. The SMART dataset was perhaps closer to random sampling. However, it is not clear whether this was truly random, or merely a sample of convenience based on samples that were available in the biobank at the time. A common mistake is to assume that a sample of convenience is the same as a random sample, but this is not necessarily true. For instance, an unrelated study may have recently sought samples of a particular disease class, thus depleting these from the biobank, skewing the class distribution for the subsequent study. Regardless, the SMART data has some class imbalance with $IR = 3.5$ counting by spectra and $IR = 3.1$ counting by patient. Given the previous discussion that we should split data at the highest level present in the data, we shall use the latter value. This is not as severe as examples found in the literature, but the sample size is relatively small and, as shown in section 2.4.3, there is a significant degree of class overlap, particularly between the clinically intermediate groups of HGD, LGD and IM.

As it is unknown whether the SMART data was truly sampled at random, and without knowing the true disease class distributions in the clinical setting and given that small samples sizes are known to exacerbate class imbalance issues, it is perhaps most prudent to address the class imbalance in the dataset. There are two main remedies: data-level and algorithm-level.

Data level balancing

These methods alter the original data during the pre-processing stage. This can involve either under-sampling, where members from the majority class are deliberately excluded from analysis, or over-sampling, where members from the minority class are increased by one of several methods. Both methods rebalance the data to more equal levels. The former method may be undesirable as it discards what could be scarce and expensive data. The latter method may seem more prudent, but it can potentially lead to over-fitting and so needs to be implemented with care.

I will not explore any under-sampling methods. Not only does this discard precious medical data, but many such methods have been shown on medical image datasets to under-perform compared to over-sampling methods, while increasing the variance in the generalisation error, which suggests instability in models so trained

[176]. This is exacerbated the more imbalanced the dataset is as more data needs to be discarded from the majority class to achieve equity with the minority class.

The simplest over-sampling method is random over-sampling (ROS) in which samples from the minority class are randomly replicated with replacement and added to the data until all the classes are equal ($IR = 1$). Despite its simplicity, over two datasets and three classifiers, ROS has been found to be the best rebalancing method [176]. Another popular method is the SMOTE (Synthetic Minority Over-Sampling Technique) algorithm, which produces synthetic instances by interpolating between the minority instances using kNN (k-Nearest Neighbours) [181]. An extension of this is borderline SMOTE which performs SMOTE on borderline instances which are often misclassified by their nearest neighbour [182].

Algorithm level balancing

We can also adjust the model in a way to better accommodate imbalanced data. With CNNs the loss function can be adjusted to account for the class imbalance, as was shown in equation 4.5. This weights the loss function such that the weight multiplied by the class distribution frequency of each class is equal. This method is not available for LDA or SVM. Other algorithms, not considered here, have been found to be robust to imbalanced data, such as the AdaBoost model [183].

To determine the extent of any class imbalance problem in the SMART dataset, and to determine the best method of ameliorating it, we first explore class imbalance in a simulated environment constructed from the Lynch dataset before exploring a subset of the SMART dataset itself.

5.6.1 Imbalanced data: Lynch rebalancing experiments

In order to simulate class imbalance in a multi-class classification setting I took the Lynch data and artificially imbalanced the data. This was done by systematically reducing the MSI-H class until only three samples were left. MSI-H was selected as this has the lowest degree of class separability, particularly in relation to the MSS group: this distinction is both the point of the study and a harder problem than

distinguishing healthy from MSS AC. By selecting the class with less separability any class imbalance is exacerbated. This also more closely mimics the class imbalance present in the SMART dataset, in which the minority classes of HGD and LGD also happen to be intermediate, less separable, states of oncogenesis, making any lessons learnt from this simulation more applicable.

Method

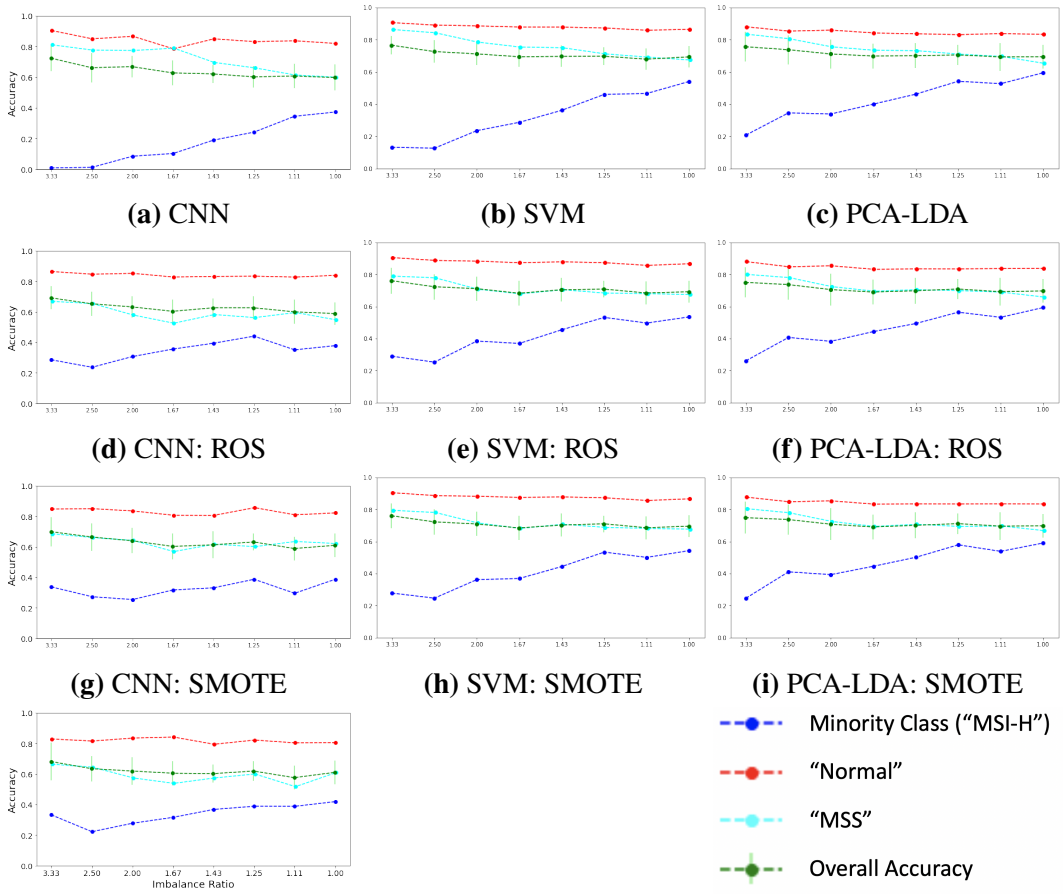
As described in section 2.2, the Lynch dataset contains three classes, each with ten samples. From the MSI-H disease class, one sample was removed at a time until only three samples were left. Thus the maximum IR was 3.33, approximately equal to that of the SMART data, and the minimum IR 1. For each removal the 5 x 3-fold CV training strategy was used for all three models. The CV splits were stratified such to ensure at least one MSI-H sample was always present in the training and test set. This reduced dataset was then subject to four treatments: no treatment (imbalanced data), ROS, SMOTE and weighted loss function. These 'rebalanced' datasets were then trained on a PCA-LDA, SVM and CNN using the same hyperparameters as in previous experiments. Note that the weighted loss function is only available for the CNN. There is a rebalancing available for the hinge loss used by SVM, but it is only possible for linear SVM, which I have not been using.

Confusion matrices provide a reasonable summary of the performance of the various models. However, for this simulation an unwieldy number of comparisons are being made, hence a more simple collection of metrics is reported. Along with the overall accuracy, the class specific accuracy for each of the three classes is also given. This allows us to track the behaviour of several classes as the IR changes.

Results and Discussion

Figures 5.14a, 5.14b and 5.14c show that class imbalance is having an impact upon all three models, especially upon the minority class which has a lower class accuracy as the IR increases. The CNN is particularly sensitive to this imbalance. Interestingly, the overall accuracy slightly decreases as the IR trends towards unity. This is a consequence of the majority classes performing better when the minority class is less represented: these being the majority class they increase the overall

accuracy, even though there is limited performance in the minority group. It has been suggested that SVMs are less prone to class imbalance [178], though here its performance is comparable to PCA-LDA, which is known to be sensitive (figures 5.14b and 5.14c) [146].



(a) CNN **(b)** SVM **(c)** PCA-LDA

(d) CNN: ROS **(e)** SVM: ROS **(f)** PCA-LDA: ROS

(g) CNN: SMOTE **(h)** SVM: SMOTE **(i)** PCA-LDA: SMOTE

(j) CNN: weighted **(k)** Key

Figure 5.14: Overall and per class of accuracy as a function of imbalance ratio (IR) over three models, CNN, SVM and PCA-LDA, and under different imbalance mitigation strategies. Top row is the original data with no attempt to mitigate class imbalance. Green line represent overall accuracy, red line the 'normal' class, cyan line the 'MSS' class and the blue line the minority class, 'MSI-H'. Note that SDs have not been reported for the class accuracies as these were so wide as to render the results meaningless. Instead we will focus on trends in the data.

All the rebalancing techniques similarly improve the CNNs minority class performance, with the most pronounced improvement achieved with the highest IR. The benefit of data level techniques to the SVM and PCA-LDA performance is less clear. These experiments show that even an $IR = 3.33$ can introduce difficulties to

model training with biomedical Raman data.

Another explanation for why rebalancing data may help performance is that the technique *de facto* performs bootstrap sampling. This is similar to other resampling CV strategies but spectra are sampled from a dataset *with replacement*, meaning that the same sample can be taken more than once. Testing is then performed on those spectra which have never been taken, forming an out-of-bootstrap test set. By rebalancing the training data with ROS, this process is being imitated.

I next extend the experiments to the SMART dataset which has true class imbalance.

5.6.2 Imbalanced Data: SMART Experiments

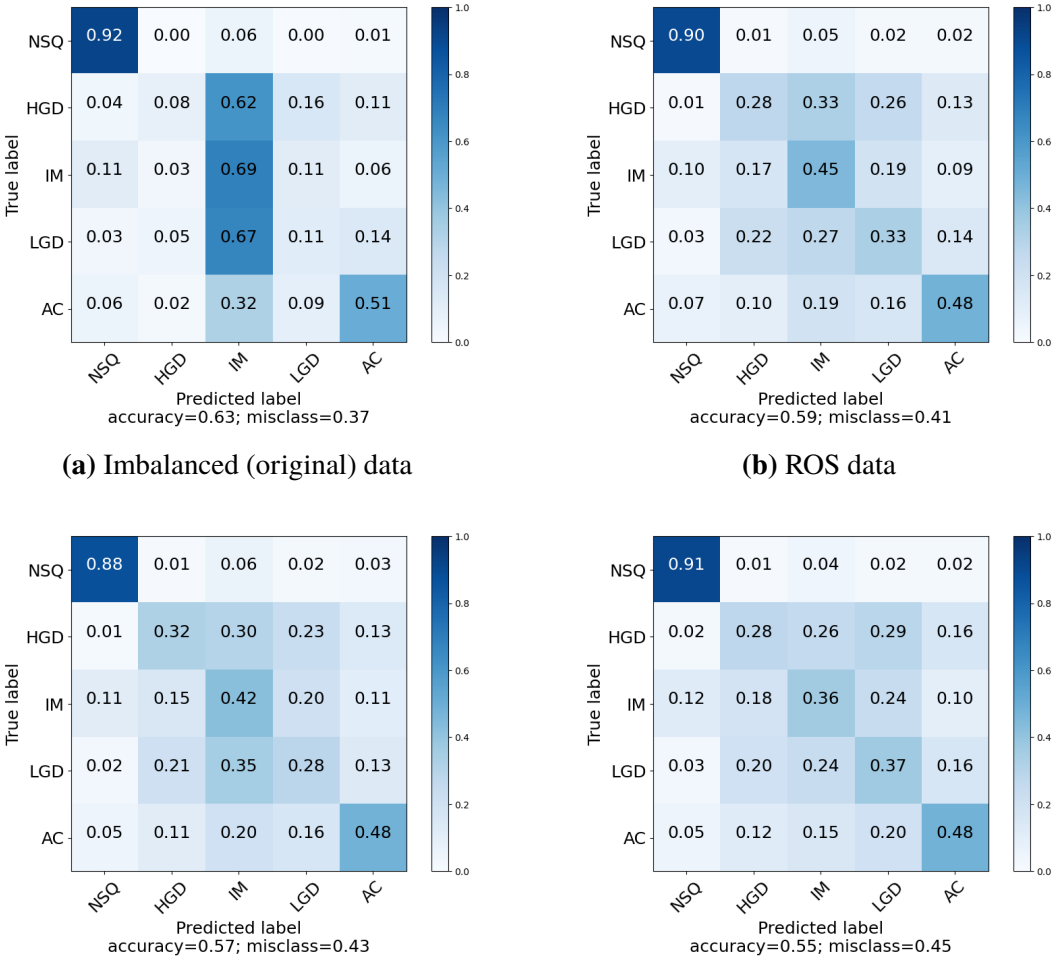
Method

A subset of the SMART data, representing a single study centre, was taken. This subset had an $IR = 3.0$. The data was already imbalanced but otherwise it was subject to the same process as the Lynch rebalancing experiment, whereby different methods of rebalancing were compared to no rebalancing.

Results and Discussion

Figure 5.15 shows that in terms of overall accuracy the imbalanced data performs best. This accords with the Lynch rebalancing experiment. However, this performance comes at the expense of misclassifying the minority classes HGD and LGD. These were only classified correctly 8.2% and 11.4% of the time respectively, both often being misattributed to IM. ROS improves performance in these minority classes to 27.7% and 33.3% at the expense of the majority classes. SMOTE achieves comparable results with 32.3% and 28.4% as does a weighted loss function with 27.6% and 36.8%. In all cases it is the most prevalent class, IM, that suffers the most from balancing the data, from a class accuracy of 68.7% to 45.5%, 42.4% and 35.9% respectively.

The results from this, and the Lynch rebalancing experiment, prompt a clinical decision: whether it is preferable to misclassify the minority groups in favour of



(a) Imbalanced (original) data

(b) ROS data

(c) SMOTE data

(d) Reweighted loss function

Figure 5.15: SMART: CNN Confusion Matrix by different balancing strategies

classifying majority groups. In many medical decisions this problem often occurs in the context of rare disease versus no disease. However, here the decision is whether to detect intermediary disease states that are, perhaps, under-represented in the data. Much depends upon the clinical consequences of such misclassifications.

Unfortunately, histopathological definitions vary worldwide, as do clinical pathways [184]. In LGD, molecular architecture is largely preserved with only subtle morphological changes, making it difficult to distinguish from non-dysplastic tissue [185]. HGD displays more marked atypical features and hence different morphological features. However, there is no clear demarcation between the two, leading to high inter and intra-rater variability [186, 187]. The distinction between grades is clinically important as their management options vary. LGD is managed

via endoscopic surveillance with endoscopic resection as needed, or potentially with radiofrequency (RFA) ablation. Patients with HGD undergo far more intensive surveillance with options including RFA +/- endoscopic mucosal resection and surgical resection.

The clinical relevance of IM is less clear. IM is characterised by the presence of goblet cells, normally present in the intestine. It is associated with a risk of progressing to LGD [185]. However, its clinical relevance is dependent upon the precise location in the oesophagus from which it is taken [185]. This is further complicated by the fact that IM can be further subdivided into three types, one of which is not deemed a risk factor for gastric cancer, and the remaining two having an association with developing cancer, but with an unclear causal pathway [188]. There is debate in the community regarding the necessity of the presence of IM for the diagnosis of Barrett's oesophagus, an inflammatory disease associated with gastric reflux [189].

Due to the more clear clinical management pathways for LGD and HGD compared to IM it may be prudent to focus on the former at the expense of the latter, although this should ultimately be a medical decision. To this end, a weighted loss function for the CNN seems optimal: it has comparable performance to the data level balancing techniques with minimal computational cost.

The high proportion of IM cases in the SMART dataset may reflect the presence of subjects with Barrett's oesophagus who would be under increased surveillance compared to the general population. It is therefore not clear whether this representation reflects the true class distribution, or has been skewed by the data collection process itself. By rebalancing the data, we make the assumption that all classes have an equal probability of being present in a dataset, which is unlikely to be true.

5.7 Data Augmentation

A common technique used to both increase the size of a dataset and to help prevent overfitting to training data is data augmentation.

Data augmentation is a technique by which the number of samples in a dataset is inflated by transforming the original data in various ways, thereby creating new data with the same class label which is then added to the original data pool. This increases the training set size, particularly important for data hungry DL models, and also provides a degree of regularisation of the data, ameliorating overfitting by representing the same data with transformations. In the context of deep learning in standard medical imaging, such as radiographs or digital pathology, popular transformations include taking random crops of images, rotating those images and/or adding noise to them [190]. However, care is needed that a transformation does not change the label of the sample. For instance, a model trained to classify hand written digits would be confounded if images of the digit "6" were rotated 180° thus changing it to resemble a "9" (figure 5.16) . However, the same transformation performed on a histology slide has no impact upon the label, as the images orientation does not typically matter to disease morphology. Here, augmentation transforms the data such that a model can potentially learn that the orientation of an image is not important to its class, thus making it invariant to orientation.

5.7.1 Lessons from the literature: data augmentation

RS data is similarly amenable to augmentation, and several methods have been explored in the reviewed literature of section 1.5. Shu *et al.*, in classifying spectra derived from nasopharyngeal tissue, adopted an augmentation technique mimicking the rotations and reflections used in medical images, flipping spectra both vertically and horizontally [87]. Chen *et al.* added white Gaussian noise of varying levels to spectra, increasing the training data by a factor of five [191]. Lee *et al.* similarly added Gaussian noise to spectra, increasing the entire dataset by a factor of four [78]. Ma *et al.* also added random Gaussian noise in addition to shifting the wavenumber axis up to $2cm^{-1}$ and adding a random scale coefficient, thus increasing the sample size from 600 to 5000 spectra [192]. Wu *et al.* also performed wavenumber shifting,

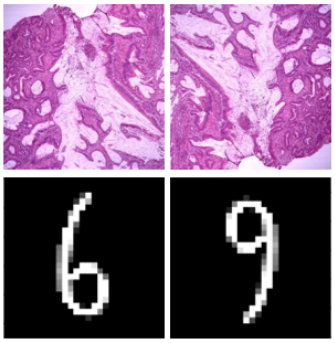


Figure 5.16: Rotating images by 180°. *Top left:* H&E stained image of diseased colonic tissue. *Top right:* same image rotated by 180°. This rotation does not change the image label. *Bottom left:* a digit "6" from the MNIST dataset. *Bottom right:* the same image rotated by 180°. This rotation means the first label no longer represents the data.

up to $4cm^{-1}$, as well as adding linear combinations of 2-5 random spectra from the same class to create a new spectrum, thus increasing the sample size from 233 to 2420 spectra [88]. Fang *et al.* also linearly combined several spectra to create a new spectrum and additionally performed 'wavenumber shifting' and added 'random noise' (neither were precisely defined), creating 6600 spectra from 510 spectra [73]. Xia *et al.* augmented their training set up to an unspecified number, shifting the wavenumber axis and adding noise to the magnitude at each wavenumber, a process which more closely resembles the Poisson noise inherent in Raman spectra, compared to adding Gaussian noise [89].

Only two studies assessed the impact that their data augmentation had upon classification performance. Chen *et al.* found it consistently increased performance across multiple data subset analyses [191], and Ma *et al.* found it increased the overall accuracy from 75% to 92% [192].

Bjerrum *et al.* offer additional augmentation techniques in the context of pharmaceutical applications of near-infrared (NIR) spectroscopy which could translate to RS. This includes adding a random offset and slope to spectra [170]. A technique used in image classification that was not utilised in the RS studies is random erasing. In the context of images this means 'blacking out' a random segment of an image

[193]. This would easily translate to spectra by randomly flattening a wavenumber region to zero.

With so many augmentation techniques available, it is not obvious which techniques, or combinations thereof, will yield the best results. Even in the field of general image analysis there are not many comparative studies [193], and none particular to spectroscopy. Generative Adversarial Networks (GANs), a type of deep learning architecture, potentially mitigate this problem. This architecture trains two competing networks. The first network, the discriminator, tries to classify samples correctly as usual. The second, generative, network tries to create new samples of sufficient quality to convince the discriminator network of their fidelity. Thus the task of picking the 'best' augmentations becomes part of the training regime. Wu *et al.* explored this in the context of RS of skin cancer tissues [194]. They employed two strategies using GANs to produce augmented training data: one to create augmentations such that the dataset becomes balanced, so that every class has the same number of spectra, and a stratified approach which maintained the prior class distribution. They also explored augmenting the training set by different amounts and a suite of ML methods including traditional and deep models. With three performance metrics, and minimal difference between them, it was not clear whether the balanced or stratified strategy performed best, both giving good results. Increasing the number of augmented spectra increased performance up to a point, beyond which performance started to suffer, suggesting that for any task there is an optimal number of spectra to augment; an additional hyper-parameter to optimise. GANs have also be used to augment RS in the context of paint analysis for cultural heritage characterisation to good effect [195].

Another pertinent factor seems to be the size of the original data set. Perez *et al.*, while investigating data augmentation for images of skin lesions, found that augmenting training images with less than 500 original images only worsened performance and that performing training and test augmentation on more than 500 images significantly improved performance [196]. If and how this translates to Raman spectra is not obvious and worth investigating if deep learning techniques

become established in the field.

Augmentation is traditionally only performed upon the training data, as data inflation and its regularisation effect is only pertinent during training and to inject noise in the test data risks introducing systematic error. However, test-time augmentation is becoming more common, particularly with small datasets. This technique augments test samples to produce several transformations of a single observation and then takes an average of performance of these as a single prediction. It has been shown to increase model performance [193]. This could allow for multiple predictions on the same spectrum, augmented several times, of which an average can be taken - essentially creating an ensemble approach. Four of the above studies applied data augmentation to the entire dataset, conducting *de facto* test-time augmentation [192, 78, 88, 73]. However, it is not clear from their methods that they exploited this merging of predictions.

It is clear from the literature that there are a wide range of available techniques and considerations to make during data augmentation. The following sections will explore two of these methods: adding noise to the Raman intensity together with wavenumber axis shifting and taking linear combinations of random spectra within the same class. Both methods will be compared to simple resampling on the Ovarian and Lynch datasets.

5.7.2 Wavenumber axis shifting experiments

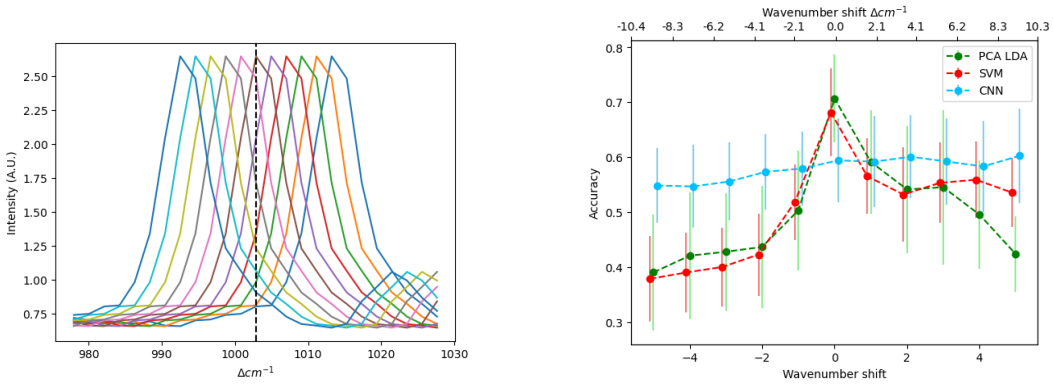
Wavenumber axis shifting can occur in Raman systems, where the wavenumber axis of two or more Raman spectra becomes misaligned. It is defined in terms of the difference between the wavenumber of an observed Raman peak to the wavenumber of its well established (theoretically and empirically) position. These can drift due to temperature fluctuations or tiny mechanical deviations in the high precision engineering of a spectrometer. Certain substances have particularly prominent and well studied peaks to allow for such comparisons, like polystyrene and cyclohexane. Such substances are therefore often used in Raman systems as an internal standard by

which to measure, and correct for, any internal drift in a spectrometers wavenumber axis alignment. In biomedical applications, the phenylalanine peak has been consistently found at about $1003cm^{-1}$, and can be used as a guide for the presence of wavenumber axis shifting. However, even with such monitoring it remains a common problem in many Raman datasets and is particularly prominent when comparing spectra taken across different systems [121]. There are wavenumber calibration algorithms which can be added to the pre-processing pipeline to help ameliorate this problem. However, these cannot completely remove system induced wavenumber axis shifting [197], and can even exacerbate the problem if not applied correctly [121]. This shifting is one of the exacerbating factors which makes system transferability difficult and so is pertinent to the SMART dataset, which has been taken on three Raman systems. Data augmentation may provide a means to mitigate the problem, as adding wavenumber axis shifting could make the model invariant to this source of noise.

5.7.2.1 Impact of wavenumber axis shift upon model performance

First I establish the extent of the problem by taking the Lynch dataset which, due to very precise experimental conditions, has no wavenumber axis shifting. To assess the difference that wavenumber shifting would have upon the performance of various models I again trained a PCA-LDA, SVM and CNN model via the stratified 3-fold CV repeated 5 times established in section 5.2. Wavenumber shifting was then induced to increasing degrees within the test data. This was done by shifting the entire wavenumber axis, represented by a vector. This involves shifting every element in the vector up to a specified amount, which I vary from -5 to 5. Each element in the vector corresponds to approximately $2cm^{-1}$, thus the induced shifting varies from -10 to $10cm^{-1}$ (figure 5.17a). This also necessarily truncates the resulting spectrum, which was conducted to the maximum induced wavenumber shift so that all of the vectors were the same size. Thus, for every fold, eleven tests sets were created and tested, each varying up to a wavenumber shift, selected at random from a uniform

distribution.



(a) **(b)**

Figure 5.17: **a**.) Shifting around the phenylalanine peak: black dashed line represents the original peak, all other spectra have been shifted. **b**.) Effect of wavenumber shift upon classifier accuracy

Figure 5.17b shows that PCA-LDA and SVM are sensitive to wavenumber shifting, with performances falling to barely better than guessing (33% for a 3 class problem). Even though no data augmentation was conducted and there was no wavenumber shifting present in the training data, the CNN is invariant to shifting present in the test data, though its performance when there is no shifting is worse than the traditional models.

It should be noted that this artificially induced wavenumber shifting does not represent how such a phenomenon would occur in practice, as such shifting would also be present in the training data via the use of different instrument, less controlled experimental conditions and less experienced operators. However, the results establish that wavenumber shifting can impact upon some traditional models and CNNs may provide a way of mitigating the problem. If the CNN can be improved to perform at least as well as the traditional models, this robustness would make it the more favourable model.

5.7.2.2 Wavenumber axis shift for augmentation

Whereas the above experiment trained on wavenumber-unshifted data and tested on shifted data, the following experiment induces shifting in the training data during augmentation and none in the test data, as is typical in the data augmentation process. The Lynch and Ovarian datasets were used for this experiment. The Lynch data had no intrinsic wavenumber shifting, but the Ovarian dataset had a small degree.

A choice needs to made regarding how much wavenumber shifting to induce. The Raman spectrometer used for these experiments performs an internal calibration to check for wavenumber misalignment, tolerating a maximum shift of $10cm^{-1}$ before the automatic calibration system reports an error. Therefore I induce varying amounts of wavenumber axis shifting (WNS) up to this number, but also explored a range of values beyond this in order assess the effect of larger WNS. The reported value of WNS represents the maximum possible shift every augmented spectrum can undergo, the actual value being randomly selected from a uniform distribution (so for WNS = 5, each augmented spectrum will be shifted by some integer value between -5 and 5).

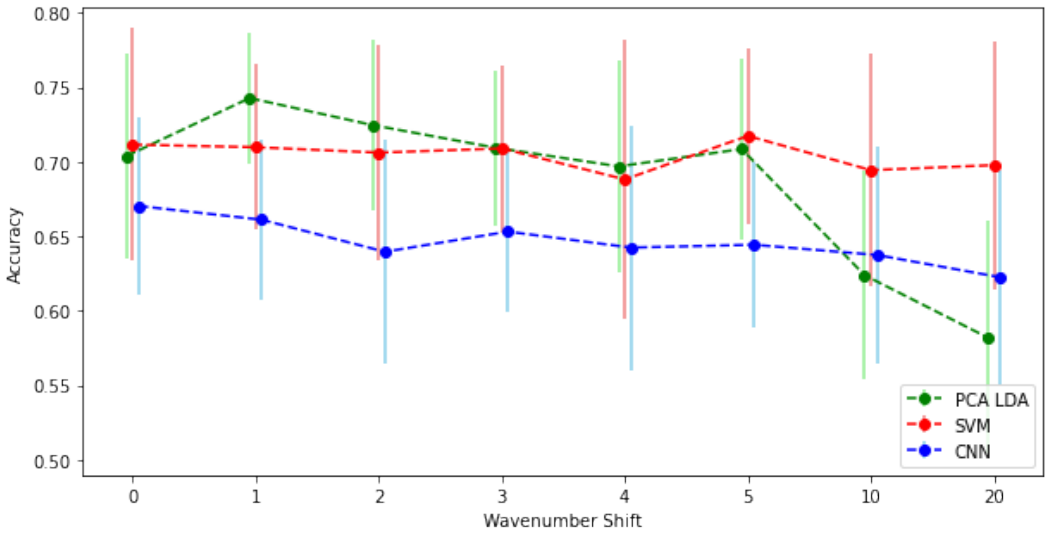
Method

The 5 x 3-fold CV strategy was again used, keeping all other hyperparameters constant. After the data was split the training data was augmented by a factor of 16. This was done 8 times, each time varying the degree of wavenumber shifting from 0-20 (corresponding to a shift of $0-40cm^{-1}$). This number represents the maximum amount of shifting, the precise number being drawn from $-max$ to max from a uniform distribution. The performance of all three models was assessed on the test data, which itself had not been subjected to any transformations other than normalisation.

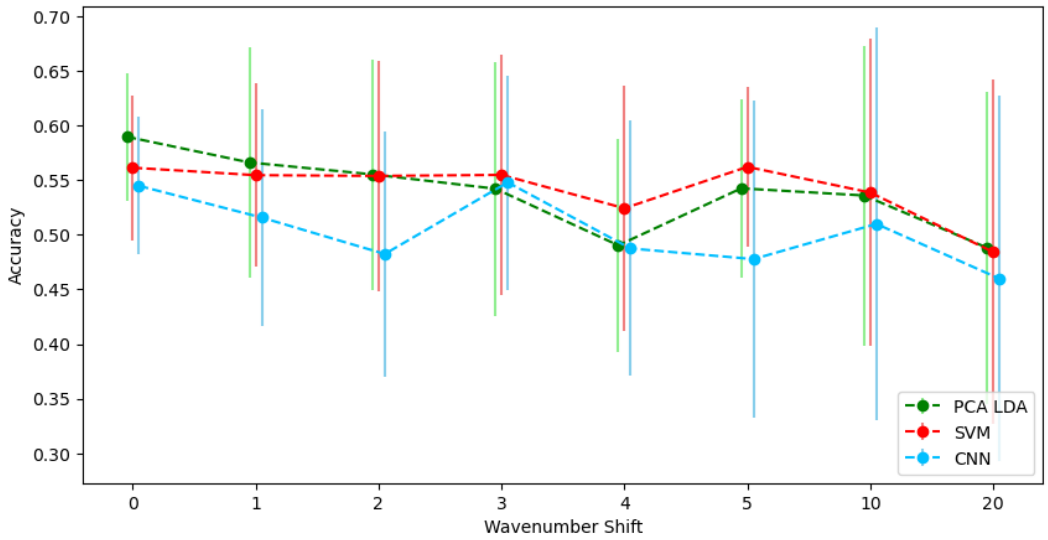
Results

Although there is a general mild downwards trend in figure 5.18, all models seem to tolerate wavenumber shifting in the training set, except for PCA-LDA which starts to suffer when the shifting exceeds that which we would expect from the instrument. However, neither was there any appreciable improvement in performance over the range up to 5 wavenumber shifts, which is far greater than the noise we expect from

the spectrometer.



(a)



(b)

Figure 5.18: Accuracy as a function of the maximum amount of wavenumber shifting induced during augmentation (a) Lynch data (b) Ovarian data. Note that each shift of 1 (which is in terms of the vector representation) corresponds to a shift of $2cm^{-1}$

5.7.3 Poisson noise on Raman peaks for augmentation

Several sources of noise in RS were discussed in section 1.3.5. These sources are generally Poissonian in nature; the amount of noise is dependent upon the Raman intensity at any given wavenumber. Therefore, inducing such noise during augmentation could help prevent models from overfitting. The question then arises of how much Poisson noise should be added. In the following experiment I induce varying amounts of Poisson noise during the augmentation process, using the Lynch and Ovarian datasets, to assess how this impacts upon test performance.

Method

To simulate this process I will assume that shot noise dominates, as this source tends to dominate other sources at high signal levels. Poisson noise was added in a particular manner in order to more closely resemble the noise generated by the spectrometer. The intensity axis of the spectrometer is calibrated in electrons/cm^{-1} where the detected electrons are the physical entity that follow Poisson statistics. Hence we need to correct the reported intensity for the spectrometer dispersion d:

$$d_i = \frac{x_{i-1} - x_{i+1}}{2} \tag{5.3}$$

for the i^{th} wavenumber of a given spectrum x (x_i being the intensity at wavenumber i). The spectrum will then be multiplied by this dispersion factor to given the electron count per wavenumber pixel in the CCD. The spectrum is then adjusted so that its maximum intensity was equal to a given constant, $h = 10^2, 10^3, ..., 10^6$, and Poisson noise added according to the (scaled) signal intensity at each wavenumber. This means that the noise added in the case of e.g. $h = 100$ is characteristic of the spectrum having a maximum intensity of $100e/cm^{-1}$, as illustrated in figure 5.19.

Again, the 5 x 3-fold CV strategy was used to assess the influence this has on model performance. After the data was split the training data was augmented by a factor of 16. This was done 5 times, each time varying the degree of induced Poisson noise. The performance of all three models was assessed on the test data, which itself had not been subjected to any transformations other than normalisation.

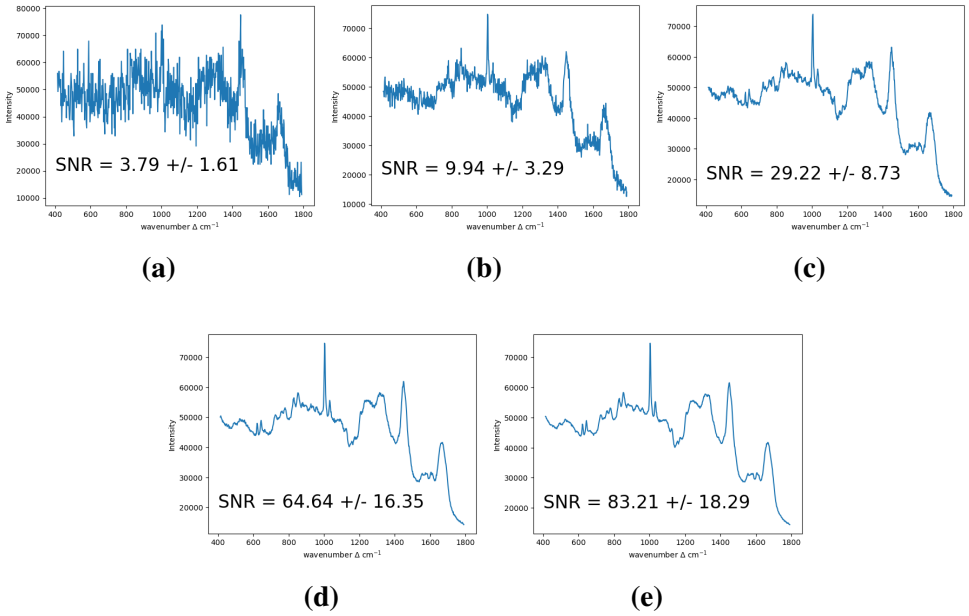


Figure 5.19: The same spectrum with different amounts of Poisson noise added during augmentation. Maximum scaled intensity: a = 10^2, b = 10^3, c = 10^4, d = 10^5, e = 10^6. SNR calculated per spectrum as described in section 1.3.6. SNR indicates the mean for the Lynch dataset +/- 1 standard deviation.

Results and Discussion

Figure 5.20 shows how augmentation with increasing levels of Poisson noise influence the accuracy of models over the Lynch and Ovarian data. It presents a mixed picture. The Lynch data shows a trend of improving results across all three models, plateauing around an induced noise level of 10^4. The Ovarian data shows no such improvement. This is perhaps a consequence of the Ovarian data SNR being lower than the Lynch to begin with, and so has a generally lower performance which does not improve with augmentation.

These two sources of noise, Poisson noise and wavenumber shifting, will be combined into a single augmentation technique, hereafter referred to as noise augmentation. Poisson noise will be induced at a level of 10^4 and wavenumber shifting to a maximum of 3 wavenumbers. Together these represent an attempt to augment data in a physically meaningful way, mimicking the noise generating processes known to be present in RS. Another method was also explored, which allows the data to dictate the nature of the perturbance of augmented spectra.

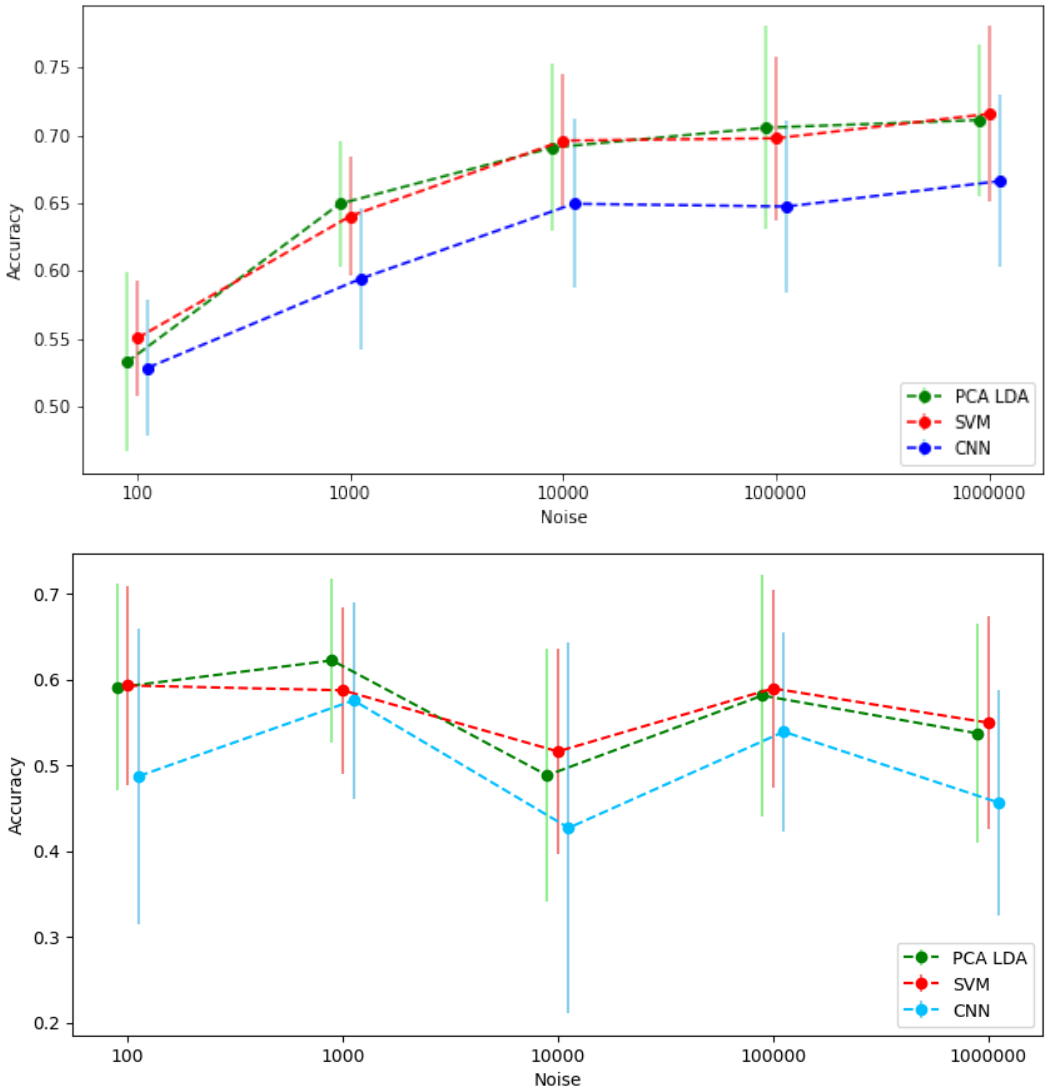


Figure 5.20: Accuracy as a function of the amount of Poisson noise induced during augmentation. Note that lower values of noise represent more induced noise. (a) Lynch data (b) Ovarian data

5.7.4 Linear combination augmentation

Another common form of augmenting Raman spectra identified in the literature was taking some form of combination of existing spectra with the same class label to create new spectra. I will explore this form of augmentation by randomly taking five spectra, $\mathbf{S}$, from the same class, then linearly combining them:

$$\alpha_1\mathbf{S}_1 + \alpha_2\mathbf{S}_2 + \alpha_3\mathbf{S}_3 + \alpha_4\mathbf{S}_4 + \alpha_5\mathbf{S}_5 \tag{5.4}$$

where each α_i is a random coefficient and $\sum_i \alpha = 1$. Figure 5.21 visually demonstrates an augmented spectrum which is a linear combination of five spectra.

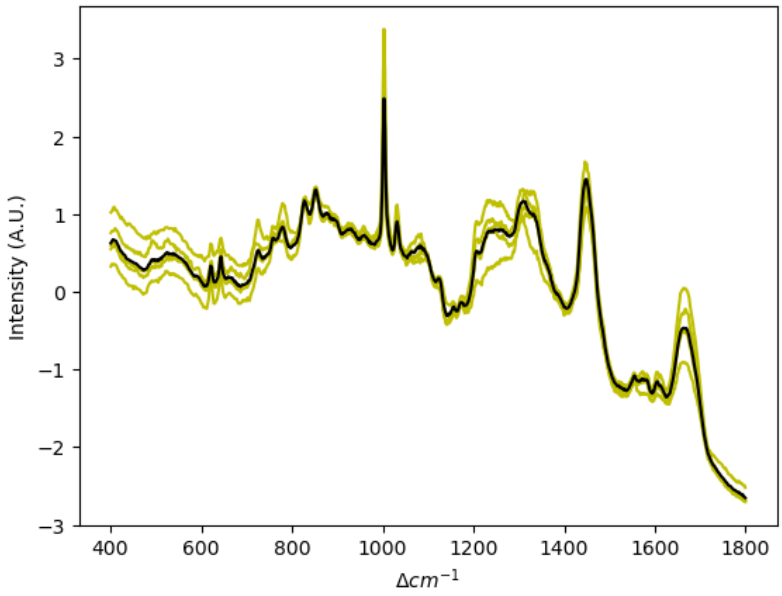


Figure 5.21: Yellow spectra: 5 normalised spectra randomly selected from the same class from the Lynch data. Black spectrum: a linear combination of the 5 yellow spectra weighted by random coefficients, α_i

In the next section this linear recombination augmentation technique will be explored alongside the noise augmentation technique and simple resampling to determine which would be most suitable for the datasets in this thesis.

5.7.5 Augmentation inflation factor experiment

In the above explorations, the number of augmented spectra created has been held constant. The question of how much data to augment is open, with the review of the literature in section 5.7.1 showing that several choices have been made, sometimes varying by many orders of magnitude. The effect of the chosen quantity of additional spectra has not been systematically explored in the context of RS, but has been found to be an important consideration in adjacent domains [196]. Hence, in this section, I explore this effect with the above developed augmentation techniques and decide upon a final augmentation strategy.

Method

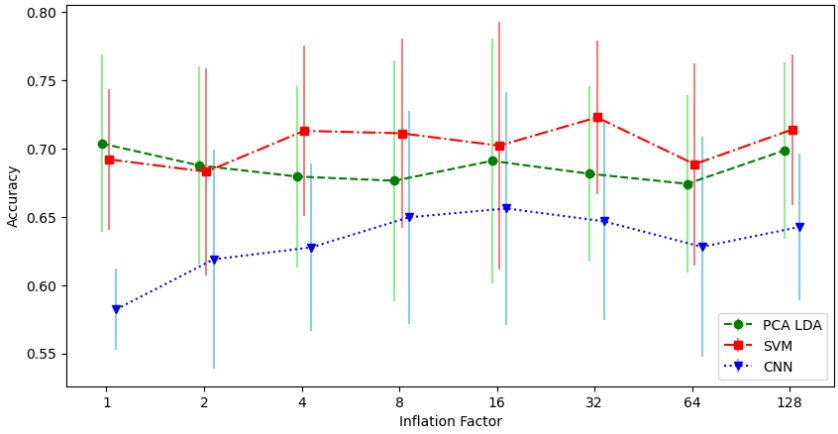
The Lynch and Ovarian datasets were explored using PCA-LDA, SVM and a CNN, to assess the role that the augmentation inflation factor has upon training. Three augmentation techniques are used: simple resampling, random linear combinations (as detailed in section 5.7.4) and adding noise (as detailed in section 5.7.3). In the former, no actual augmentation is performed but the training data is simply resampled. This provides a baseline performance to which to compare the augmentation techniques. When adding noise, the amount of poisson noise and wavenumber shifting added during augmentation has been fixed at 10000 and 3 respectively.

The Lynch and Ovarian training data were augmented by a factor of 1 (i.e. no augmentation), 2, 4, 8, 16, 32 and 64. The inflation factor refers to the number of times the training set is increased - so an inflation factor of 8 means that the training data is augmented until it is 8 times larger. The Lynch data was additionally augmented by a factor of 128 as it is a smaller dataset to begin with. All models were trained using the 5 x 3-fold CV strategy as described in section 5.2.

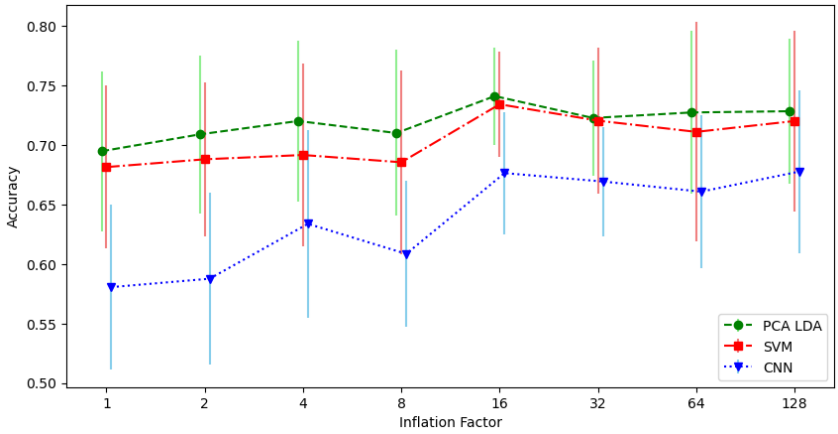
Results and Discussion

Figure 5.22 shows how the augmentation techniques scale as a function of inflation factor for the Lynch data. PCA-LDA and SVM do not gain, or suffer, from any of the augmentation techniques. This is consistent with the nature of these models: LDA classifies based on data centroids and SVMs on a subset of the data around the margin. The augmentation techniques are not expected to influence either of these. The CNN does show a moderate improving trend up to an inflation factor of approximately four before seeming to plateau. However, the error bars overlap such that it is difficult to draw firm conclusions. Additionally, any improvement has been achieved by simply resampling the data, with neither augmentation technique improving this above resampling.

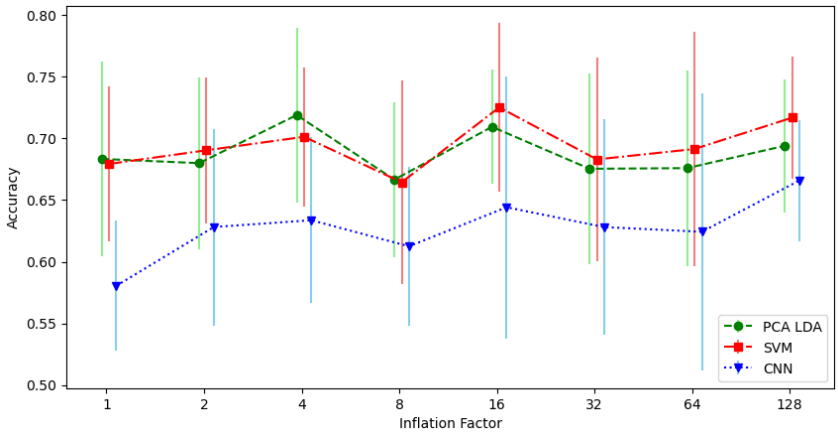
The results from the Ovarian dataset (figure 5.23) show no such improvement from any augmentation method. Even the CNN does not seem to benefit from the techniques. It was noted in section 5.7.1 that augmentation was only of benefit when the original dataset was of sufficient size. Though the Lynch dataset is smaller than



(a)



(b)

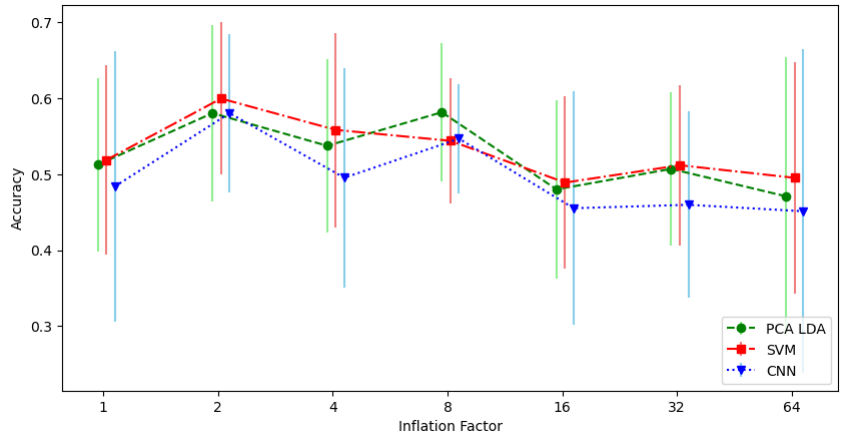


(c)

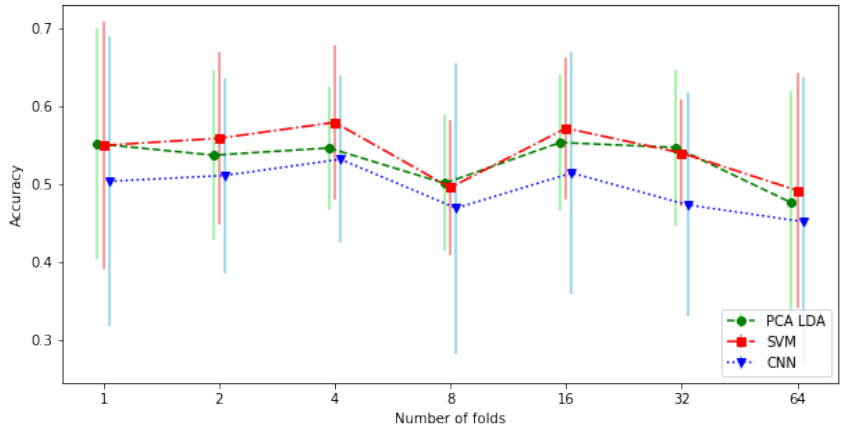
Figure 5.22: Lynch data: accuracy across the three models as a function of data inflation by three methods (a) Resampling (b) Noise added augmentation (c) linear recombination augmentation

the Ovarian dataset, the latter has a far lower SNR. It may be that the SNR of the original data also has a significant impact upon the benefits of augmentation, whence if the noise component is too great, adding more noise does not have a regularising effect.

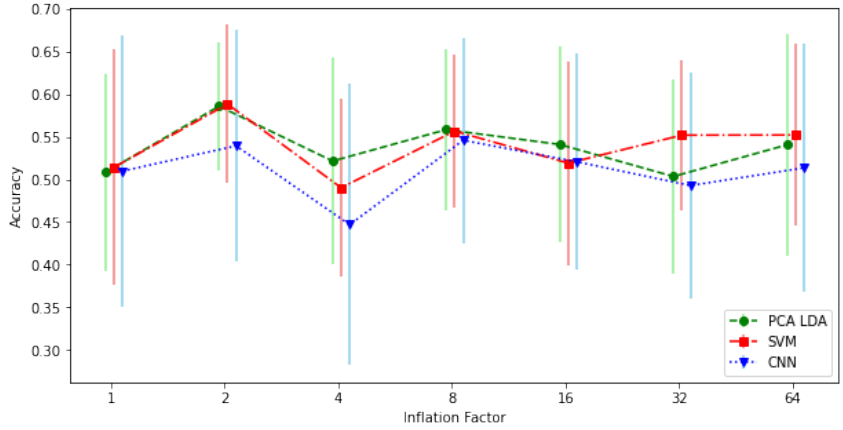
As a consequence of these augmentation experiments the following strategies for each dataset will be adopted. For the Lynch dataset the noise added augmentation technique will be used for the CNN, using a factor of 16 upon the training data. Although this technique only improves performance marginally, there is a trend that suggests a real effect. This is supported by knowledge of the noise generating processes in RS, which the noise added technique directly seeks to mimic. However, this effect does not seem to apply to the Ovarian data, possibly due to its low SNR. However, no deleterious effects were seen from augmentation, and so the noise added augmentation technique will be applied with an inflation factor of 4. The SMART dataset is larger than either of these datasets, but has far lower SNR than the Lynch or Ovarian datasets. The size of the SMART data means that augmentation increases computational time significantly enough to restrict the amount of subsequent experiments that could be performed. While there is no convincing evidence that this would be beneficial, the computational cost means that I will not augment this dataset.



(a)



(b)



(c)

Figure 5.23: Ovarian data: accuracy across the three models as a function of data inflation by three methods (a) Resampling (b) Noise added augmentation (c) linear recombination augmentation

5.8 Nested Cross Validation

We have thus far focussed on many of the hyperparameter choices available before the final model is applied to the data. The aim was to establish general trends in order to find which of these data hyperparameters may improve performance. However, there are two outstanding issues that will be here addressed: hyperparameter choices pertaining to the models themselves and the possibility of overfitting to hyperparameters. Thus we draw a distinction between data-dependent hyperparameters (i.e. the order of a baseline to fit to during baseline removal or the factor by which to inflate a dataset during augmentation) and model-dependent hyperparameters (i.e. number of PCs to retain during PCA).

As detailed in chapter 4 the models used all have a number of hyperparameters which need to be chosen before the model is trained. These can have profound differences upon the performance of a model. It is common to explore these either with a process of trial and error or a more systematic search. However, it has been shown that using standard CV to both select the best hyperparameters and to estimate the generalisation performance of a model can result in significant bias [198]. This may not be true of models trained with sufficiently large datasets, but becomes more of a problem as sample size decreases [199]. This is due to overfitting at the second level of inference: in the same way it is possible to overfit the training of a model to a particular dataset, it is possible to overfit the hyperparameter choices such that they may perform well on the training data, but fail to generalise well to new data [155]. This can also be a problem when there are many hyperparameter choices from which to select.

One rationale for exploring some potential data hyperparameters, such as the choice of the number of augmented spectra, was to restrict our choices. However, model specific hyperparameters remain. To mitigate against this source of bias I have adopted a nested CV strategy. This embeds one CV process inside another, resulting in an inner CV loop and an outer CV loop (figure 5.24). The outer loop first detaches a portion of the data: in this case I selected 3-fold CV for the outer loop, resulting in one third of the overall data being held out, while the remaining

two thirds is passed to the inner loop. The inner loop then splits the data again, in this case a 2-fold CV process, resulting in one third of the data remaining for training and the other third, called the validation set, for assessing the performance of the hyperparameters. Every combination of hyperparameters is trained in this manner. Just as with normal CV this process can be repeated to decrease the variability of the estimated performance due to small sample size. I repeat the inner loop 3 times. The best performing set of hyperparameters are then trained upon the combined training and validation data and passed back to the outer loop where the model is assessed against the held-out test data. This outer CV loop can also be repeated, which I do five times. Thus the inner loop selects the best performing hyperparameters and the outer loop gives a less biased estimate of the generalisation performance of the chosen model. Note, that for each outer loop, different data is passed into the inner loop and so it is possible that the inner loop selects a different set of hyperparameters each loop.

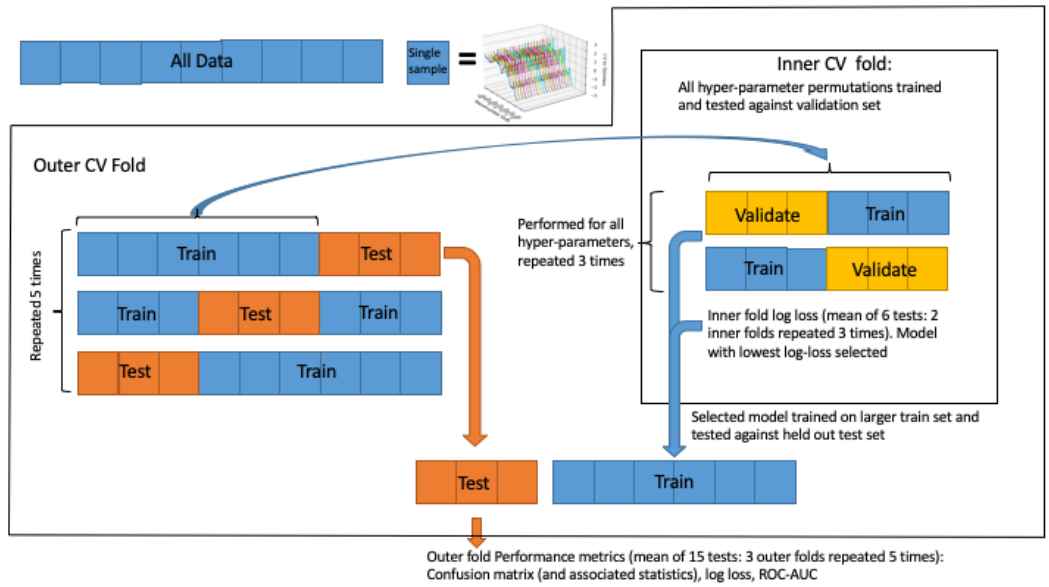
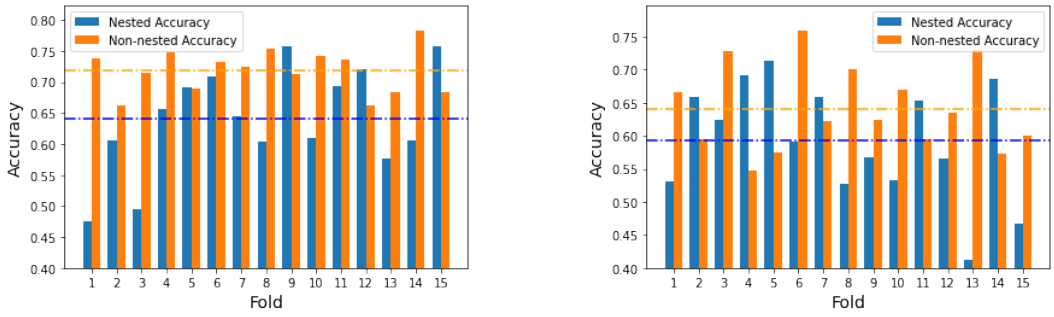


Figure 5.24: Schematic of Nested Cross Validation

Figure 5.25 shows the difference that nested CV makes to estimating the generalisability of a model. The non-nested scores represent the inner fold scores. These are what would be reported in a flat (i.e. non-nested) CV strategy. This contrasts to the lower nested scores, which have been further tested upon the test set,

which was not available to the models during hyperparameter selection. This clearly shows that flat CV provides an overly optimistic assessment of performance. This trend persists with other performance metrics, such as log loss.



(a) Lynch Data **(b)** Ovarian Data

Figure 5.25: Nested vs Non-nested Cross Validation. PCA-LDA model on the Lynch and Ovarian data, the number of principle components to retain being the hyperparameter searched, from 2-35 PCs retained. Dashed lines represent the means across all 15 folds.

Thus, for the selection of model hyperparameters, a nested CV strategy is used for each of the datasets. Unfortunately nested CV comes with a considerable computational cost, limiting its applicability. This comes from the number of models that need to be trained and assessed: for instance, the inner fold of figure 5.25 trained a number of models equal to: the number of hyperparameters x number of folds x number of repetitions, 34 x 2 x 3 = 204. As this is repeated 15 times across the outer fold a total of 3060 models were thus trained. This is possible with PCA-LDA as it is not computationally expensive but SVMs and especially CNNs are expensive. This limits how many hyperparameters can be assessed using nested CV for these latter models. This was another reason for exploring the data hyperparameters in such detail, as it leaves only the model hyperparameters to be explored.

The details of the final CV strategy are explained for each dataset in the following chapter.

5.9 Limitations

Searching particular hyperparameters while holding all others constant allows us to explore the effect a given hyperparameter has upon performance estimates. However,

it potentially miss interactions between those hyperparameters, which may combine in complex and unpredictable ways if allowed to vary together. Thus even though our exploration may provide some insights, it also ignores these interactions and so none of the above conclusions can be considered definitive.

Chapter 6

Results

> *"Errors using inadequate data are much less than those using no data"*
>
> Charles Babbage

6.1 Lynch and Ovarian Data nested CV strategy

For the Lynch and Ovarian datasets, and for all three models, a nested CV strategy was employed to explore the model hyperparameters. This involved 3-fold CV repeated 5 times for the outer fold with 2-fold CV repeated 3 times for the inner fold, as illustrated in figure 5.24. The inner folds require a single performance metric to select the best model: as discussed in section 3.3 the log-loss captures the distributions of outcomes to the true labels rather than a single point measure for performance. Hence, the model with the lowest average log-loss was selected for retraining with the larger data pool and tested against the held-out test set from the outer fold. The best performing hyperparameter combination was then selected and subjected to a flat 3-fold CV repeated 5 times for a final assessment of model performance (otherwise the performance estimate will include estimates from models with different hyperparameters, which does not reflect how the model would be deployed in the clinical setting). This final estimate extracted a prediction per sample, as opposed to per spectrum, using the simple consensus method described in section 5.5. The same hyperparameter space was searched over the Lynch and Ovarian

datasets using a grid-search strategy, in which all hyperparameter combinations in a grid are systematically explored (tables 6.1 and tables 6.2 for the SVM and CNN respectively). For PCA-LDA only one hyperparameter was searched, the number of PCs to retain over the range 2-35.

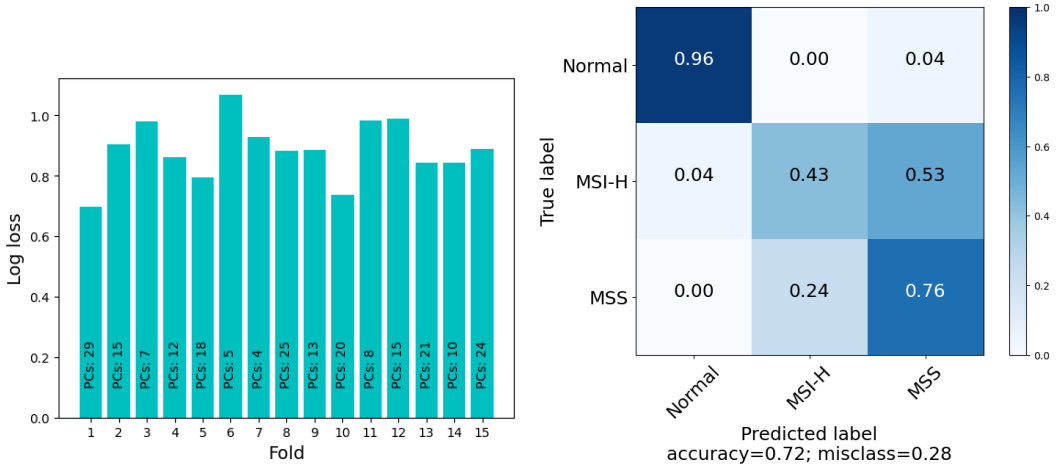
Analyses were conducted using Python version 3.10. The Scikit-learn library was used for the PCA-LDA models [133]. The ThunderSVM library was used for the SVM models [200]. This allows SVM models to be computed on a graphics processing unit (GPU), which was necessary as kernel SVM computations scale quadratically with the data. The CNN was developed using Pytorch [152]. An NVIDIA A10 GPU with 32 GB RAM was used. On the largest dataset, the SMART data, the approximate time to run through 5 x 3 CV for the PCA-LDA, SVM and CNN models were one hour, 12 hours and 14 hours respectively.

6.2 Lynch Data

6.2.1 Lynch Data: Results

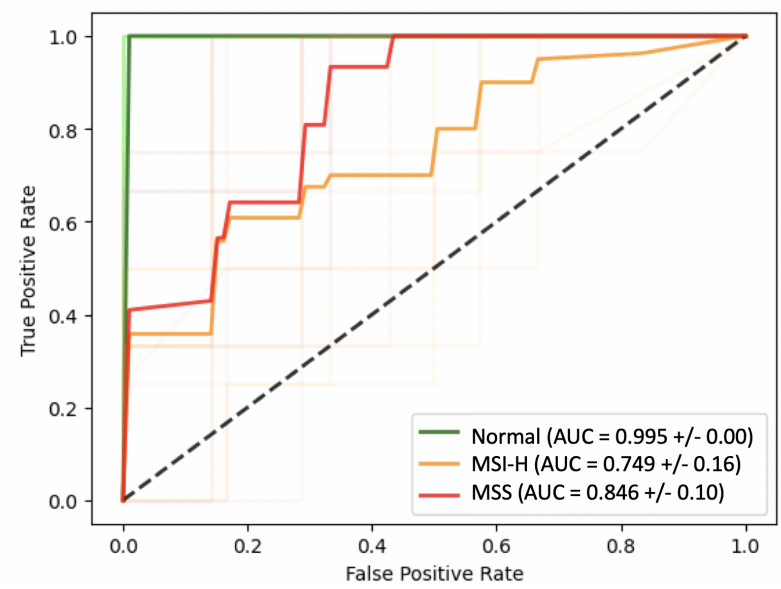
PCA-LDA

Figure 6.1a shows the outer-fold performance of the PCA-LDA with the corresponding best hyperparameters selected from the inner-fold. This shows that there is considerable variance in the number of PCs the inner-fold selects. This is indicative of an unstable model which is sensitive to the training data, which is common with small datasets. Selecting the mode (and employing the principle of parsimony to break any ties by selecting the simplest model) the 15 PC model was selected. The model was retrained using this hyperparameter on a flat 5 x 3 CV strategy to produce a confusion matrix (figure 6.1b) and a ROC curve (6.2).

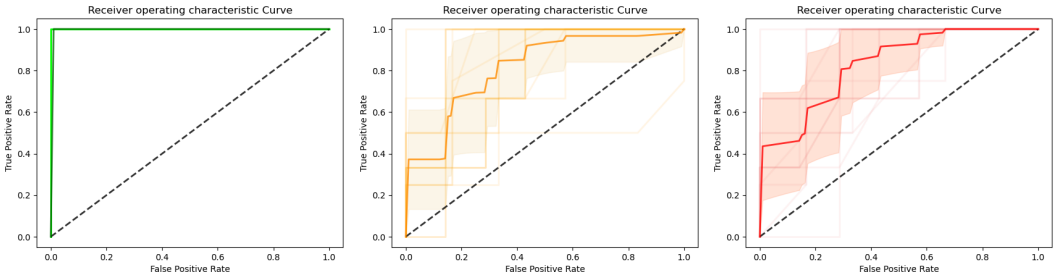


(a) **(b)**

Figure 6.1: (**a**) Lynch Data: PCA-LDA Log-Loss. Each outer fold was subject to an inner fold hyperparameter search. The bars represent the log-loss of the best performing hyperparameter in that fold. The ordering of the folds is arbitrary. (**b**) Lynch Data: PCA-LDA Outer Fold Confusion Matrix.



(a) All ROC curves



(b) Normal ROC curves **(c)** MSI-H ROC curves **(d)** MSS ROC curves

Figure 6.2: PCA-LDA ROC. Top, all disease classes plotted on same axis. Bottom row, individual disease classes. Thick lines represent the mean performance over 15 folds. Pale lines represents the performance of each individual fold. Shaded regions indicate 1 standard deviation.

It is clear from the ROC curves in figure 6.2 that distinguishing normal from diseased tissue (either MSI-H or MSS) is a relatively easy task, but the task of distinguishing MSI-H from MSS is more difficult. This is consistent with expert opinion. The confusion matrix (figure 6.1b) shows that the PCA-LDA confuses MSI-H for MSS over half the time (54.0%), but only confuses MSS for MSI-H 24.0% of the time.

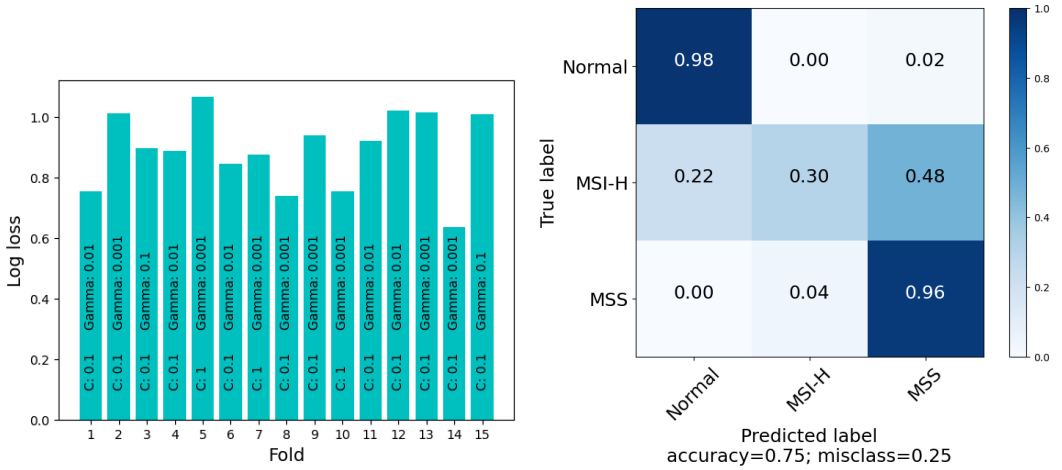
SVM

As discussed in section 4.4, SVM-RBF has 2 model hyperparameters, a range of which was searched as shown in table 6.1. As can be seen from figure 6.3a, SVM

C	0.01	0.1	1	10	100
γ	0.01	0.1	1	10	100

Table 6.1: SVM hyperparameters used during grid search for the Lynch dataset

returns more stable model hyperparameters during nested CV with $C = 0.1$ being heavily favoured and $\gamma = 0.001$ preferred over half the time. This may be due to the ability of the model to draw non-linear decision boundaries. SVM performs better than PCA-LDA across all metrics. From figure 6.1b, this can be attributed to its much improved ability to distinguish MSS, misclassifying it only 4%, though it only correctly identifies MSI-H 30.0% of the time, misclassifying it as normal tissue 22% of the time, which the PCA-LDA did only 4% of the time.

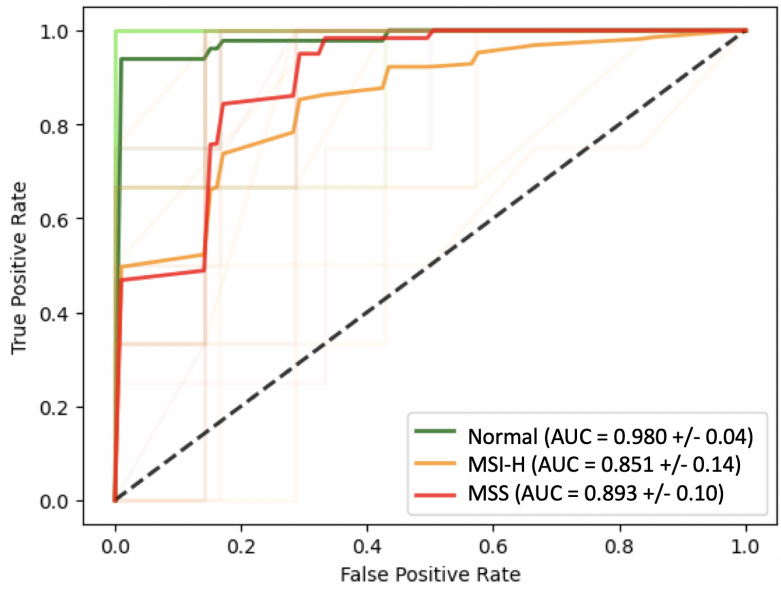


(a) **(b)**

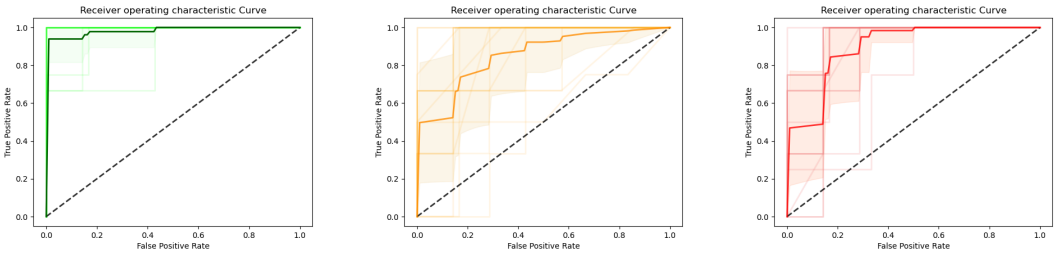
Figure 6.3: (**a**) Lynch Data: SVM Log-Loss. Each outer fold was subject to an inner fold hyperparameter search. The bars represent the log-loss of the best performing hyperparameter in that fold. The ordering of the folds is arbitrary. (**b**) Lynch Data: SVM Outer Fold Confusion Matrix.

CNN

The CNN hyperparameters, discussed in section 4.5, which were searched are shown in table 6.2. The learning rate is stable, with 0.0001 being selected for 8/15 of the nested models. The batch size is relatively unstable, 256 was the most common selection at 5/15 times. Its accuracy is comparable to the previous models, but achieves this in similar way to PCA-LDA being more able to classify MSI-H at the expense of MSS.



(a) All ROC curves



(b) Normal ROC curves **(c)** MSI-H ROC curves **(d)** MSS ROC curves

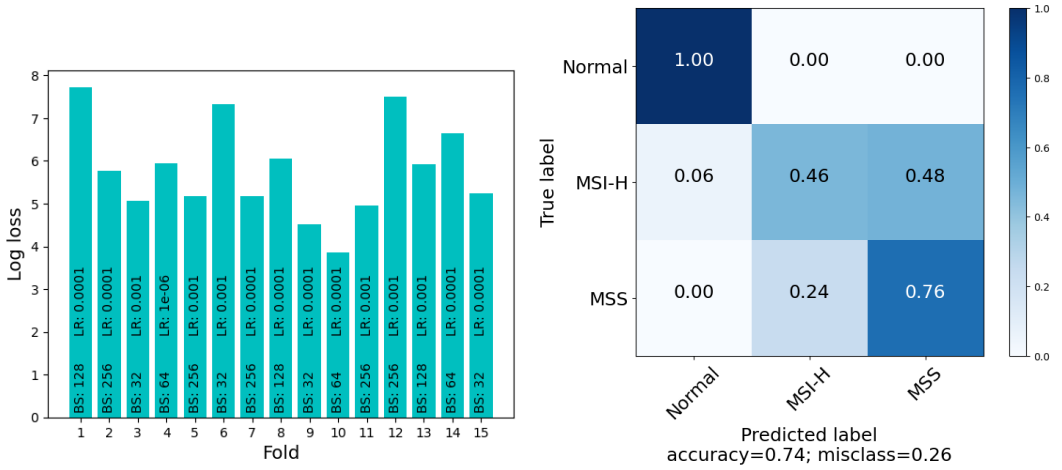
Figure 6.4: SVM ROC. Top, all disease classes plotted on same axis. Bottom row, individual disease classes.Thick lines represent the mean performance over 15 folds. Pale lines represents the performance of each individual fold. Shaded regions indicate 1 standard deviation.

Batch Size	32	64	128	256	516	1032
Learning Rate	10^{-06}	10^{-05}	10^{-04}	10^{-03}	10^{-02}	10^{-01}

Table 6.2: CNN hyperparameters used during grid search for the Lynch dataset

6.2.2 Lynch Data: Discussion

Table 6.3 directly compares the log-loss and overall accuracy across the three models from the Lynch data. Of note is that although PCA-LDA accuracy is worse than the SVM, it was the better in terms of log-loss. This indicates that the PCA-LDA model is more circumspect in its predictions. For instance, an output of a model for a single spectrum might be represented as $[0.33, 0.35, 0.32]$ which would be



(a) **(b)**

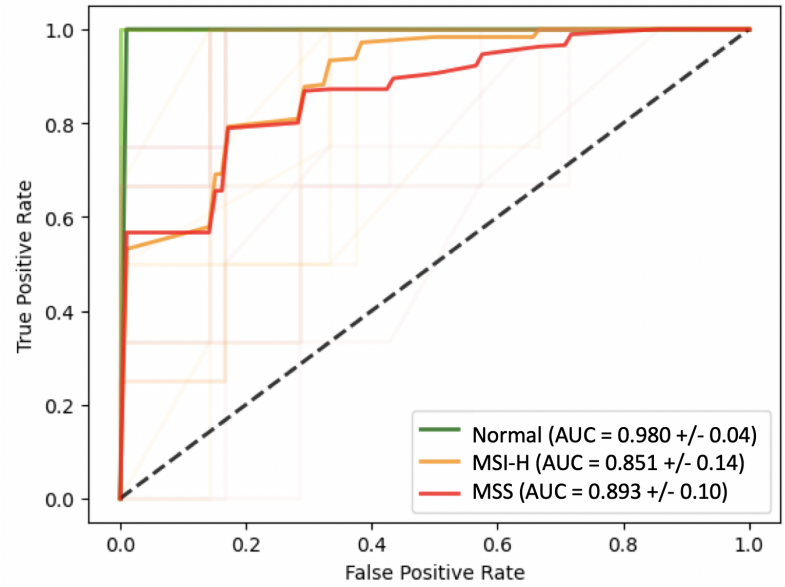
Figure 6.5: (**a**) Lynch Data: CNN Log Loss. Each outer fold was subject to an inner fold hyperparameter search. The ordering of the folds is arbitrary. (**b**) Lynch Data: CNN Outer Fold Confusion Matrix.

	Log Loss	Accuracy
PCA-LDA	0.66 +/- 0.17	71.3% +/- 8.8
SVM	0.80 +/- 0.73	74.7% +/- 7.2
CNN	0.54 +/- 0.17	74.0% +/- 12.0

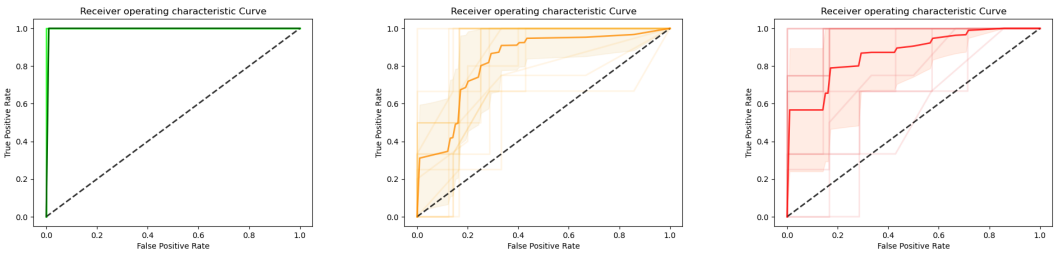
Table 6.3: Lynch data: Final models comparison.

classified as MSI-H, where the true output might be [1,0,0], indicating a normal spectrum. Although the classification is wrong, its log-loss will not be too high as the predictions between the three classes are close. Repeated over many spectra, we would expect a low classification score even though the log-loss is not high. As discussed in chapter 3, this is the advantage of using proper scoring metrics. Across most metrics the CNN is the best performing model, though with such small sample sizes and large standard deviations it is not possible to declare the model definitively better for this application.

The existing criteria used to screen for LS use sensitivity and specificity to measure the generalisation error. The Amsterdam II criteria achieve a sensitivity and specificity of 72% and 78% respectively; the Bethesda protocol 94% and 25%. These pathways are binary classifiers, only taking confirmed CRC cases and classifying them as either MSI-H or MSS. Therefore the sensitivity and specificity are well defined. I retrained the LS model as a binary classifier to allow a like for like comparison.



(a) All ROC curves



(b) Normal ROC curves **(c)** MSI-H ROC curves **(d)** MSS ROC curves

Figure 6.6: CNN ROC. Top, all disease classes plotted on same axis. Bottom row, individual disease classes.Thick lines represent the mean performance over 15 folds. Pale lines represents the performance of each individual fold. Shaded regions indicate 1 standard deviation.

This used the same nested CV strategy as above but excluding the normal group. The best sensitivity and specificity given by a PCA-LDA model was 82% and 51% respectively. For the CNN this was 83% and 45% and the SVM 51% and 57%. These are competitive with the existing protocols, indicating a viable addition to the LS screening pathway, and would likely improve with a larger training cohort. Whether a two or a three class model would be preferable for clinical deployment is ultimately a medical decision. The 3 class model allows the possibility of a different clinical management pathway, where all samples suspected CRC samples could be automatically screened for LS.

Figure 6.7 shows the difference spectrum between the MSI-H and MSS mean spectra. From this it may be possible to infer some biochemical differences between the two classes. In particular, peaks at 714, 1081, 1302, and 1445 cm^{-1} have tentatively been assigned to lipids. Other peak assignments include 1672 cm^{-1} (cholesterol), 494 cm^{-1} (glycogen, nucleic acids), 529 cm^{-1} (amino acids), 732 cm^{-1} (phosphatidylserine, adenine), 787 cm^{-1} (nucleic acids), 852 cm^{-1} (ring-breathing mode of proline, hydroxyproline, tyrosine), 1003 and 1034 cm^{-1}, (phenylalanine, polysaccharides), 1110 cm^{-1} (lipids, proteins), 1366 cm^{-1} (tryptophan, lipids, guanine) and 1583 cm^{-1} (C=C bending mode of phenylalanine). Overall, there seems to be differences in nucleic acids, proteins and lipids. Band assignments were made using findings contained within Movasaghi *et al.* [52].

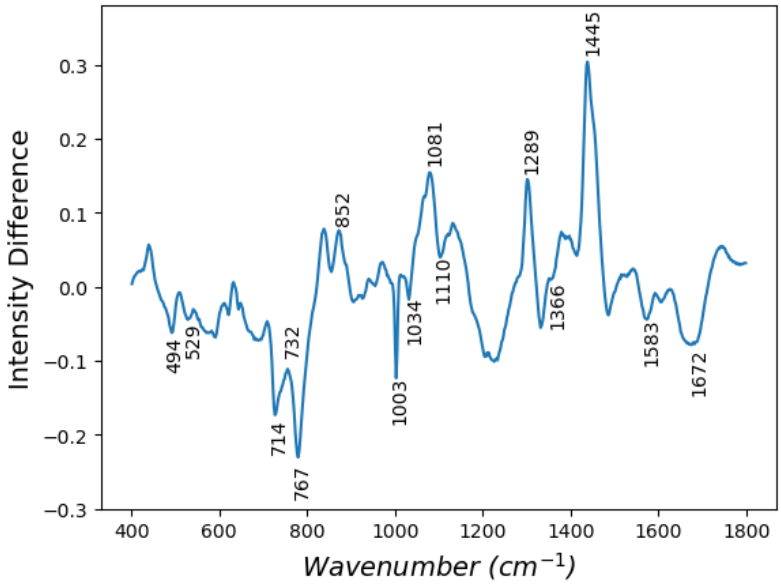


Figure 6.7: Lynch data: MSI-H minus MSS difference spectrum. Numbers indicate peaks mentioned in the text.

However, the small sample size means it is not possible to distinguish which peaks are statistically meaningful - we may be observing natural, healthy, variation between patients. Additionally, there are peaks present with unknown underlying biochemistry, which may be contributing to the predictive capacity. That some of these peaks correlate with known and/or hypothesised biochemical differences between these groups lends credence to these differences being clinically significant.

For instance, higher levels of lipids and lower levels of glucose have been observed in CRCs [201], the latter being consistent with a switch to aerobic glycolysis (the Warburg effect) [202].

Regardless, it is difficult to determine which, if any, of these biochemical differences the models are using to make their classifications. In PCA-LDA we may inspect the PCs which were used for the LDA. PCs 1 - 10 (first two shown in figure 2.2) show multiple potentially biologically relevent peaks characteristic of nucleic acids (727, 780, 1097, 1573 cm^{-1}), lipids (1079, 1301, 1330 cm^{-1}) and proteins (1002, 1030, 1236, 1668 cm^{-1}). However, given that the PCs contain mixed biochemical signatures, there is still no guarantee that any given biochemical difference is informing the models output. Notice also, that the peaks identified here do not correspond to those of the difference spectrum. SVMs are notoriously opaque and no attempt at interpretation has been made for this model. CNNs have been described as a black box technique, but in section 7 I explore some means of extracting biochemical information.

6.3 Ovarian Data

6.3.1 Ovarian Data: Results

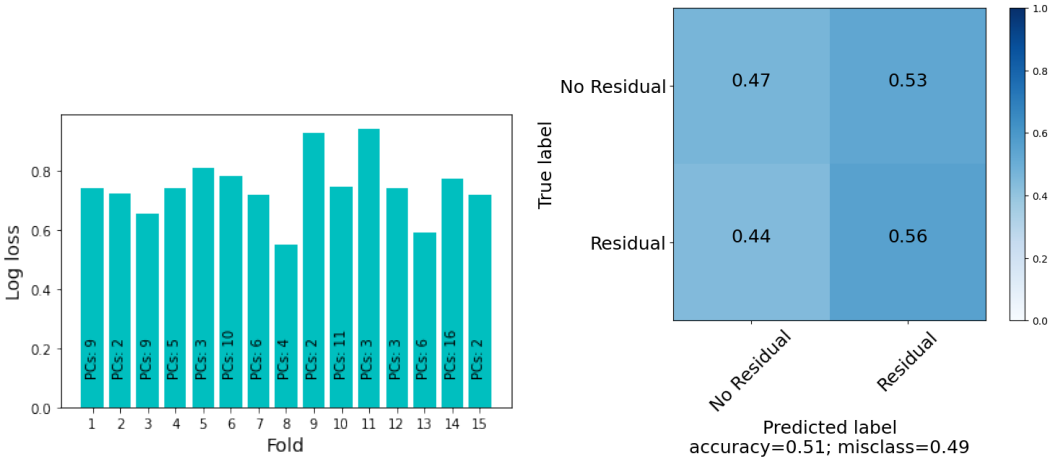
PCA-LDA

Figure 6.8a shows the performance of the inner PCA-LDA fold with the best performing hyperparameters. Similar to the Lynch dataset, there is some variation in choosing the optimal number of PCs. In this case the mode value was 3. The model was retrained using this hyperparameter on a flat 5 x 3 CV strategy, assessing performance per sample using simple consensus, to produce a confusion matrix (figure 6.8b) and ROC curve (6.9).

As a binary classification, there is only one curve to plot. These metrics suggest the model is performing barely above random classification.

SVM

The SVM hyperparameters chosen during nested CV vary significantly, suggesting an unstable model that will not generalise well (figure 6.10a). Selecting the



(a) **(b)**

Figure 6.8: (**a**) Ovarian Data: PCA-LDA Log Loss. Each outer fold was subject to an inner fold hyperparameter search. The ordering of the folds is arbitrary. (**b**) Ovarian Data: PCA-LDA Outer Fold Confusion Matrix.

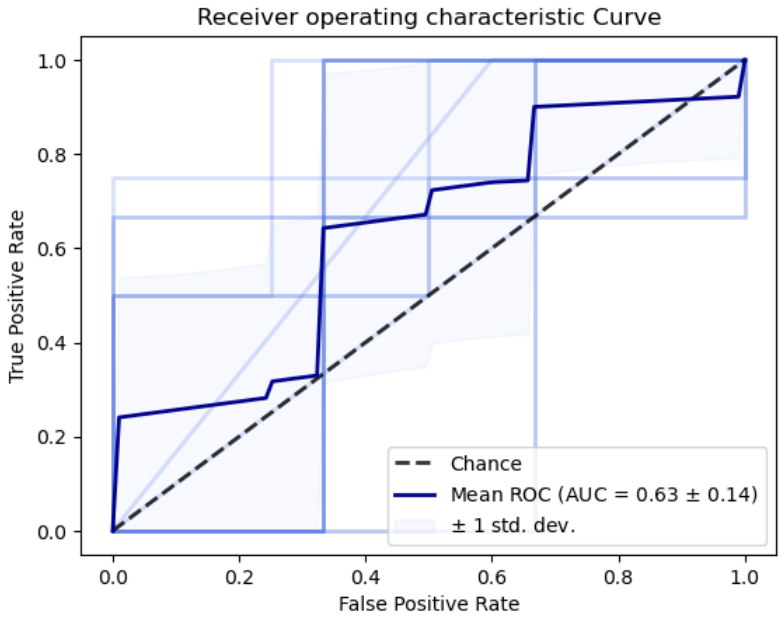
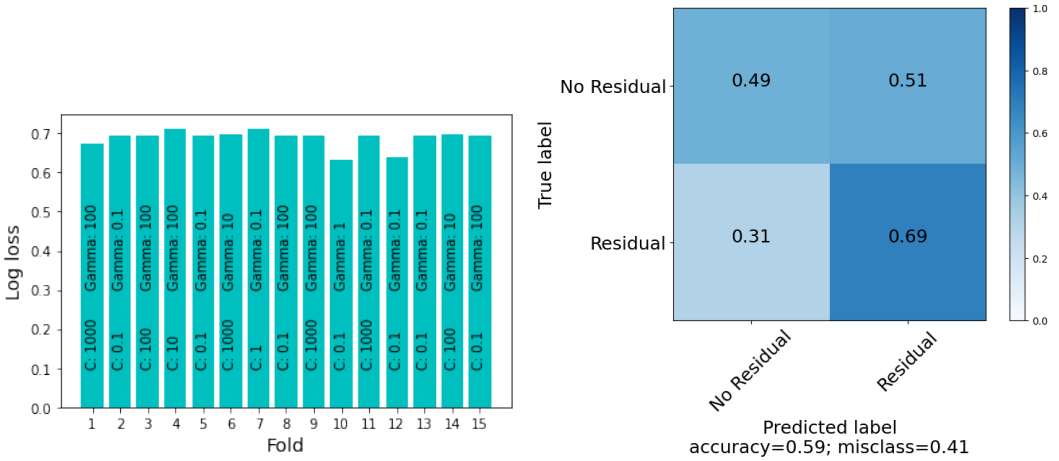


Figure 6.9: PCA-LDA ROC. Thick line represents the mean performance over 15 folds. Pale lines represents the performance of each individual fold. Shaded region indicates 1 standard deviation.

mode of these parameters, $C = 0.1$ and $\gamma = 0.1$, produces the confusion matrix of figure 6.10b. This is a slight improvement upon the PCA-LDA model, but is still only performing slightly better than chance.

CNN



(a) **(b)**

Figure 6.10: (**a**) Ovarian Data: SVM Log Loss. Each outer fold was subject to an inner fold hyperparameter search. The ordering of the folds is arbitrary. (**b**) Ovarian Data: SVM Outer Fold Confusion Matrix.

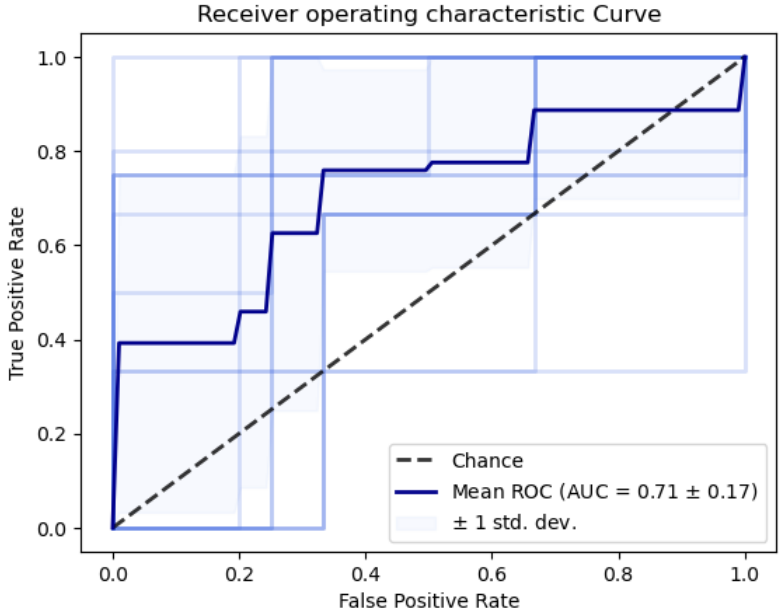
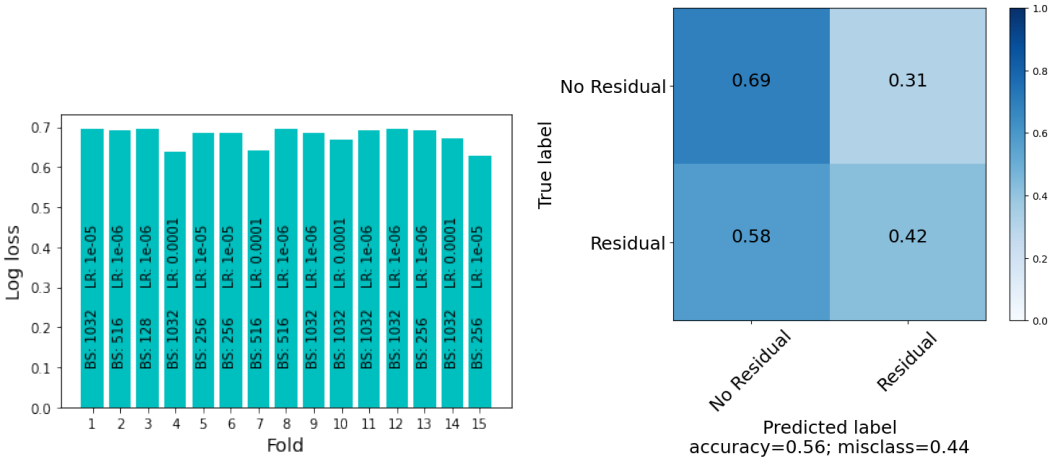


Figure 6.11: SVM ROC. Thick line represents the mean performance over 15 folds. Pale lines represents the performance of each individual fold. Shaded region indicates 1 standard deviation.

Figure 6.12a suggests a reasonably stable model, with a learning rate of 10^{-6} and batch size of 1032 being the most common values. The corresponding confusion matrix (figure 6.12b) again shows a slight improvement over both PCA-LDA and SVM, though the performance remains only slightly better than chance.



(a) **(b)**

Figure 6.12: (**a**) Ovarian Data: CNN Log-Loss. Each outer fold was subject to an inner fold hyperparameter search. The ordering of the folds is arbitrary. (**b**) Ovarian Data: CNN Outer Fold Confusion Matrix.

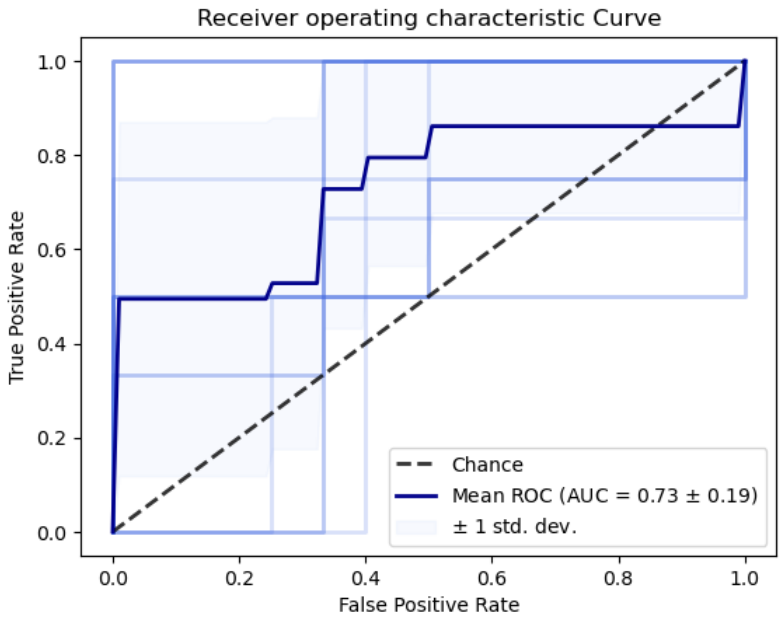


Figure 6.13: CNN ROC. Thick line represents the mean performance over 15 folds. Pale lines represents the performance of each individual fold. Shaded region indicates 1 standard deviation.

6.3.2 Ovarian Data: Discussion

Table 6.4 compares the results between models. In this instance the AUROC is also given as the problem is binary. The difference across metrics is marginal, with large variance, but the log-loss for PCA-LDA is much worse compared to the other models.

	Log Loss	Accuracy	AUROC
PCA-LDA	4.08 +/- 2.08	51.1% +/- 19.7	0.63+/- 0.14
SVM	1.18 +/- 1.14	58.9% +/- 17.1	0.71 +/- 0.17
CNN	1.09 +/- 0.64	55.5% +/- 23.3	0.73 +/- 0.19

Table 6.4: Ovarian data results table

This indicates that this model is returning very confident but incorrect predictions.

These results are actually a slight improvement upon previous findings that used molecular signatures of residual disease using gene expressions. Using several models and thresholds the genetic method could not return an AUROC > 0.65 [114]. The same study hypothesised that the maximum AUROC achievable with an oracle would be 0.83 or 0.71 depending upon how residual disease is defined (the latter value being indicative of a stricter criterion, not used here, which may lead to better patient outcomes when acted upon) and so the ceiling of optimal performance may be particularly low for this clinical problem.

Despite the CNN giving an underwhelming accuracy, its low log-loss and relatively high AUROC indicate that the model has potential. The AUROC, in particular, suggests a better threshold other than the default 0.5 may be preferable for this problem. A thorough investigation of this value may yield sufficiently satisfactory sensitivities and specificities to be clinically useful.

Figure 6.14 shows the difference spectrum between the mean spectra of the two classes. It is difficult to discern any obvious biochemical features that distinguish the classes. Instead, the obvious difference is in the general intensity, with the no residual disease class having a higher overall Raman intensity. This could be considered an artefact of the lack of pre-processing which omitted a baseline correction step, which explicitly seeks to remove such a component. However, the baseline correction experiments of section 5.3.2 showed that baseline correction did not improve model performance (figure 5.8). This may be a further hint that, at least for certain applications, the baseline contains diagnostically useful information, as suggested by Gaifulina *et al.* [169].

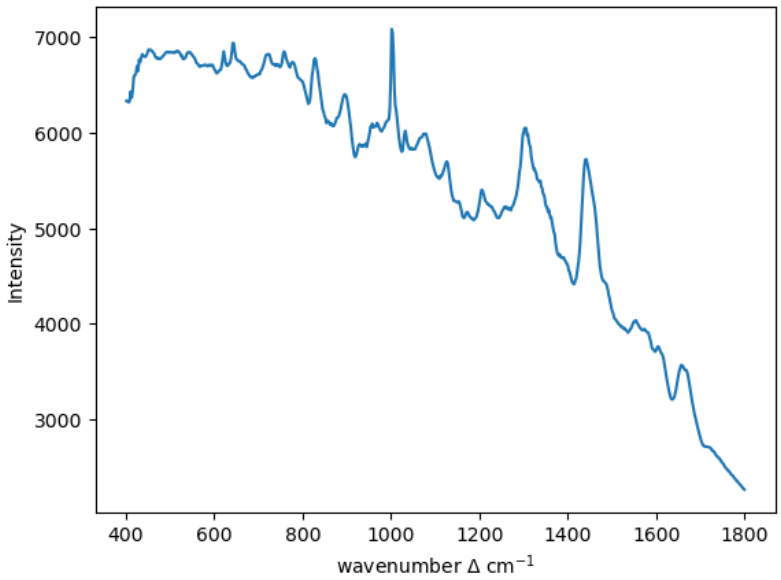


Figure 6.14: Lynch data: difference spectrum between average residual and no residual classes

6.4 SMART Data

6.4.1 Hyperparameters for SMART

Based on the Ovarian and Lynch results, I selected hyperparameters for the SMART data. In particular, it was found that taking the mean (or the geometric mean for the log-distributed hyperparameters searched in SVMs and the CNNs) resulted in the best performing models. For the PCA-LDA model, 11 PCs were retained. For the SVM model $C = 0.9$ and $\gamma = 0.1$ were selected and for the CNN $LR = 5 \times 10^{-5}$ and $BS = 184$ were used.

As the SMART hyperparameters were preselected I did not employ nested CV. This circumvents the possibility of overfitting the data at the second level of inference, making the generalisability of the results more robust, at the expense of perhaps not finding the optimal model for the SMART data.

6.4.2 CV strategy for SMART

As described in section 2.4.2 the principal purpose of this five class dataset was to assess the transferability of data taken at one centre to the same data taken at different centres with the same make of spectrometer. To this end, the data was subject to an

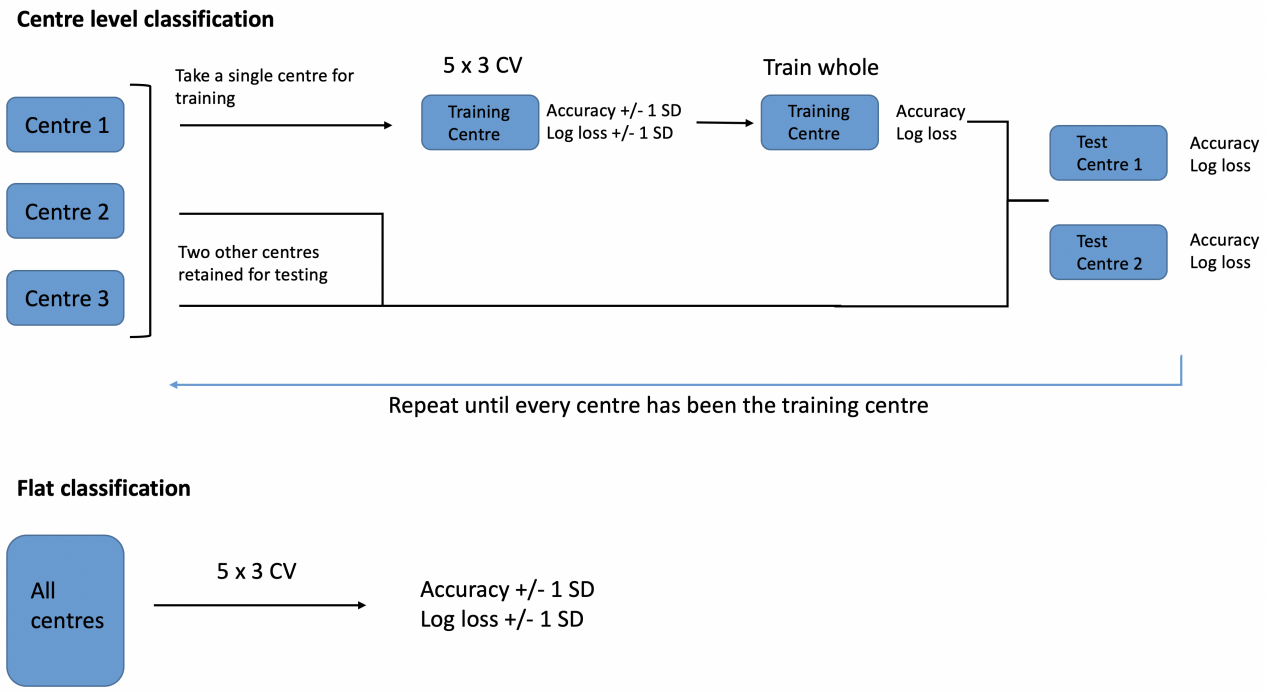


Figure 6.15: SMART data: Schematic of CV strategy. Upper schematic shows an iterated centre level CV strategy that uses 5 x 3 CV to produce performance estimates +/- 1SD, then the training estimates on the full single centre data, followed by the test estimates for each other centre. Lower schematic ignores the centre level hierarchy, pooling all data together and subject to the 5 x 3 CV strategy giving performance estimates +/- 1SD

instrument correction, giving a 'corrected' and an 'uncorrected' dataset. Compared with the previous datasets, the SMART data has an additional level to its hierarchical structure: the centre from which it was taken. To tease out information about the differences in performance across centres, three CV strategies were employed. First, a single centre was taken and subjected to the 5 x 3 CV strategy described in section 5.2. This produced an estimate of performance with error bars, and is the estimated generalisability of the model trained on a single centre. The models were then retrained on the entirety of that training centres data, as would be done if deploying a model in the real world. This model was then tested against the two held-out test centres. If the instrumentation and protocol were sufficient to mitigate against system transferral issues then we would expect little difference between the single centre 5 x 3 performance and that of the held-out centres. This process was repeated 3 times such that each centre was used to train the models (figure 6.15).

Additionally, a 5 x 3 CV strategy was employed on the entire SMART datasets, ignoring the centre level splitting. This provides an estimate of performance if the centre level structure of the hierarchy is ignored. The corrected and uncorrected datasets were each independently subjected to this training regime. The CNN used class rebalancing through a weighted loss function, as described in section 5.6.2. This process produced a plethora of results and so only the accuracy and log-loss will be shown in this initial comparison.

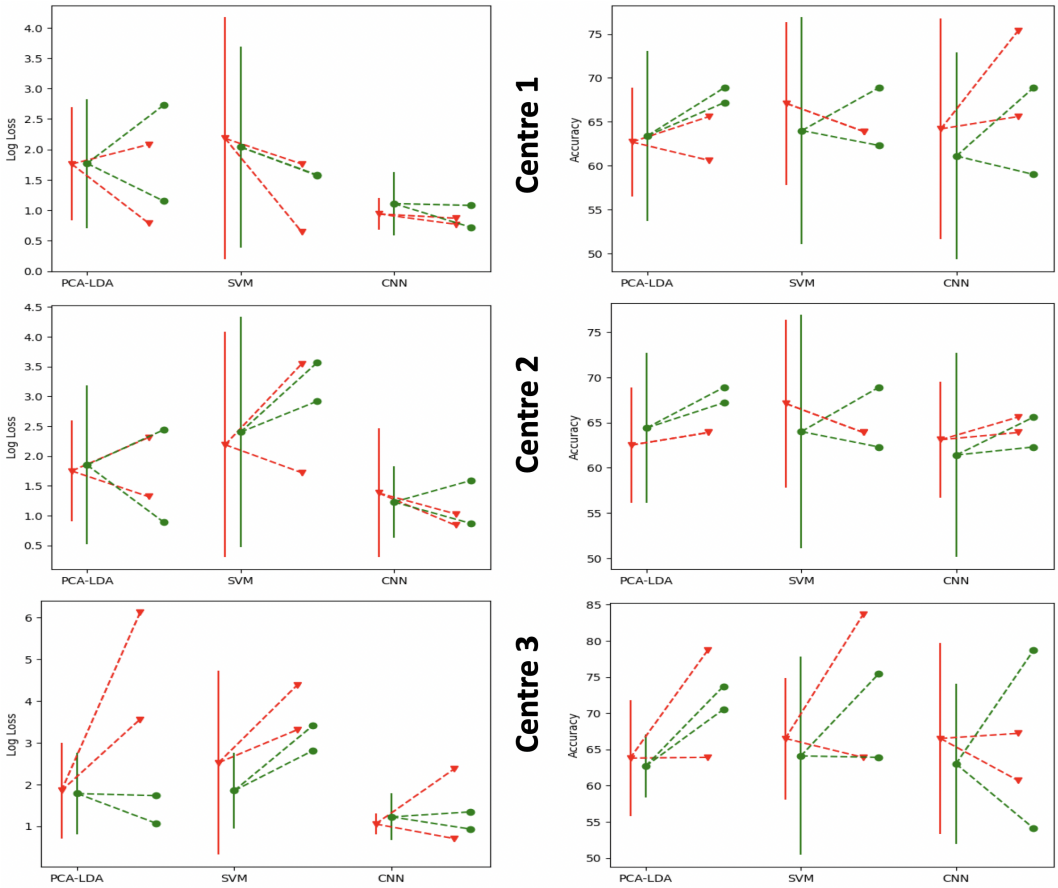


Figure 6.16: Log-loss and accuracy visualised across centres. Left panels indicate change in log-loss, right panels change in accuracy. Points with error bar indicate 5x3 CV performance, with dotted lines indicating performance at the held out centres. Red indicates uncorrected data, green instrument corrected.

6.4.3 SMART data: corrected vs uncorrected results

Tables 6.5 and 6.6 compare the uncorrected and instrument corrected SMART dataset performances. In order to more easily compare these tables some information has

been extracted to visualise the change in performance from the 5 x 3 CV score per centre to the two held out test centres. Figure 6.16 shows there is no appreciable drop in accuracy from the training centres to the test centres as would be expected if transferability between centres was problematic. Additionally, there is little to distinguish the performance of the three models on either dataset. Although there is the merest suggestion that CNNs perform better, most evident in the all centre 5 x 3 CV results, the degree of improvement and variability in results are insufficient to confirm superior performance.

		5 x 3 CV one centre	Centre 1	Centre 2	Centre 3	5 x 3 CV All centres
PCA-LDA	Centre 1	63.8% +/- 8.0 1.76 +/- 0.93	*72.1%* *1.03*	63.9% 2.08	78.7% 0.79	
	Centre 2	62.5% +/- 6.4 1.85 +/- 1.15	63.9% 3.56	*70.5%* *1.13*	63.9% 6.12	64.2% +/- 6.9 1.83 +/- 1.01
	Centre 3	62.7% +/- 6.2 1.75 +/- 0.85	65.6% 1.32	60.6% 2.31	*68.9%* *1.04*	
SVM	Centre 1	66.5% +/- 8.4 2.18 +/- 1.99	*75.4%* *0.51*	63.9% 1.76	83.6% 0.64	
	Centre 2	67.1% +/- 9.3 2.52 +/- 2.21	63.9% 3.31	*77.0%* *1.22*	63.9% 4.38	63.9% +/- 11.2 1.63 +/- 1.41
	Centre 3	66.5% +/- 10.7 2.19 +/- 1.89	67.2% 1.72	63.9% 3.55	*75.4%* *0.74*	
CNN	Centre 1	66.5% +/- 13.2 0.94 +/- 0.26	*100%* *0.29*	67.2% 0.87	60.7% 0.77	
	Centre 2	63.1% +/- 6.4 1.05 +/- 0.25	65.6% 0.70	*90.2%* *0.30*	63.9% 2.37	68.4% +/- 5.6 1.02 +/- 0.20
	Centre 3	64.2% +/- 12.6 1.38 +/- 1.08	75.4% 0.84	65.6% 1.03	*91.8%* *0.30*	

Table 6.5: SMART Uncorrected results. Italicised entries indicate the training results and so are not indicative of performance.

However, as discussed in section 3.3, it may be that accuracy is too blunt a measure. Figure 6.16 additionally presents the same information for the log loss. This similarly shows no difference in performance when transferring models across systems, and that instrument correction does not improve performance. Unlike the accuracy results though, the superiority of the CNN is more evident, giving marginally lower log-loss scores, but with much narrower variability (both within a

single centre and transferring across to other centres), indicating a more confident estimate of performance.

		5 x 3 CV one centre	Centre 1	Centre 2	Centre 3	5 x 3 CV All centres
PCA-LDA	Centre 1	62.7% +/- 4.3	*70.4%*	70.5%	73.7%	
		1.77 +/- 1.06	*1.41*	2.73	1.15	
	Centre 2	64.4% +/- 8.3	67.2%	*75.4%*	68.9%	62.2% +/- 11.2
		1.78 +/- 0.97	1.73	*0.61*	1.07	1.90 +/- 1.37
	Centre 3	63.4% +/- 9.7	68.9%	67.2%	*73.8%*	
		1.85 +/- 1.33	0.89	2.44	*0.78*	
SVM	Centre 1	64.1% +/- 13.7	*80.3%*	63.9%	75.4%	
		2.04 +/- 1.65	*0.39*	1.58	1.57	
	Centre 2	64.0% +/- 12.9	62.3%	*80.3%*	68.9%	66.6% +/- 8.8
		1.85 +/- 0.91	3.41	*0.32*	2.81	1.82 +/- 1.31
	Centre 3	65.1% +/- 12.1	62.2%	65.6%	*78.7%*	
		2.40 +/- 1.93	3.57	2.92	*0.68*	
CNN	Centre 1	63.0% +/- 11.1	*100%*	54.1%	78.7%	
		1.11 +/- 0.52	*0.19*	1.08	0.72	
	Centre 2	61.4% +/- 11.3	62.3%	*100%*	65.6%	68.2% +/- 9.6
		1.22 +/- 0.56	0.93	*0.15*	1.34	1.01 +/- 0.23
	Centre 3	61.1% +/- 11.8	68.9%	59.0%	*93.4*	
		1.23 +/- 0.60	0.87	1.59	*0.19*	

Table 6.6: SMART Corrected results. Italicised entries indicate the training results and so are not indicative of performance.

6.4.4 SMART data: corrected vs uncorrected discussion

These results show that instrument correction does not improve performance for any model, and that using the same make of spectrometer across centres with a common protocol is sufficient to create transferable models. However, as described in section 2.4.2, some samples were discarded due to suspected breaches to that protocol, and that when such samples were present the classification performance suffered. Therefore, strict adherence to a protocol would seem paramount to system transferability, though this necessity is not unusual in healthcare applications and should be well tolerated by pathology technicians should the technology proceed to the clinical setting.

As the engineering and clinical protocol development have been sufficient to allow for model transfer, I have been unable to explore whether the input

invariance property of CNNs would allow them to outperform traditional models when transferring models across spectrometers.

6.4.5 SMART data: model comparison discussion

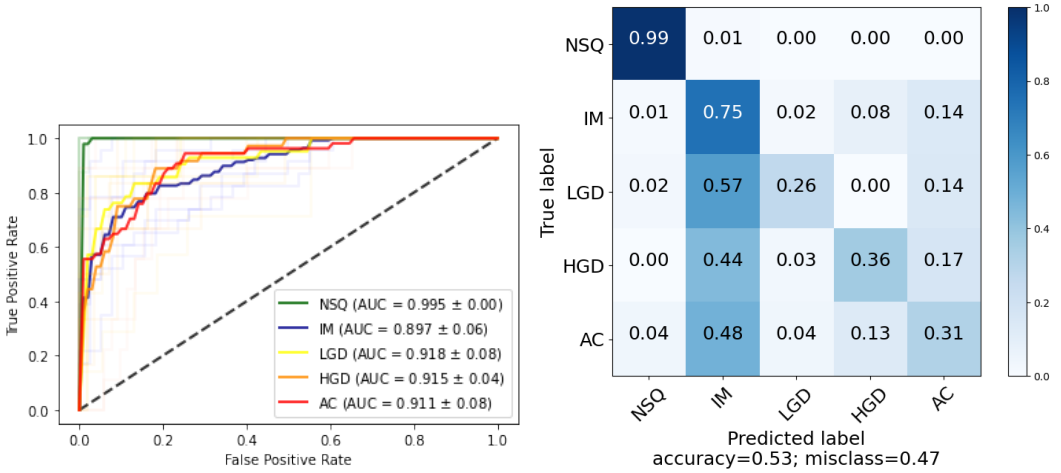
Although the log-loss is a proper scoring metric, and so captures the nuances of the model predictions, it lacks an intuitive interpretation. Table 6.7 shows the AUROC of each class over the three models using the centre level training on the uncorrected data (and so correspond with the red plots given in figure 6.16). This accords with the log-loss findings that the CNN is the best performing model, showing much more discriminative ability over the intermediate classes of IM, LGD and HGD.

	NSQ	IM	LGD	HGD	AC
PCA-LDA	0.99 +/- 0.01	0.78 +/- 0.08	0.69 +/- 0.11	0.51 +/- 0.10	0.74 +/- 0.08
SVM	0.87 +/- 0.06	0.87 +/- 0.09	0.89 +/- 0.08	0.89 +/- 0.08	0.97 +/- 0.02
CNN	0.99 +/- 0.00	0.90 +/- 0.06	0.92 +/- 0.08	0.92 +/- 0.04	0.91 +/- 0.08

Table 6.7: SMART uncorrected data: AUROC by class +/- 1 SD

A recent literature review and meta-analysis of RS applied to *ex vivo* oesophageal cancers found a pooled AUROC of 0.99 when distinguishing between malignant and benign tissues [119]. This is similar to the models developed in this thesis when considered as a binary model of normal (NSQ) vs all other groups. It is not clear in the literature review whether the included studies took into account factors such as the hierarchical structure of the data, which I have found important to consider, but together with the results here does indicate that the technology has the potential for clinical use in distinguishing malignant oesophageal pathologies. The review did highlight that few studies had attempted to grapple with tumour sub-types, as did the SMART study. This is both a more difficult and clinically relevant task.

I have not attempted to extract potential pathogenic biochemical hallmarks from the SMART data, as was done with the Lynch data. There have been a number of studies investigating the use of RS to distinguish potential oesophageal cancers which have tentatively attempted to determine such biochemical antecedents. They have



(a) ROC **(b)** Confusion Matrix

Figure 6.17: SMART uncorrected data: (a) ROC (b) Confusion Matrix

variously found that cancerous oesophageal tissue is associated with an increase of tryptophan (1268 cm^{-1}) and collagen (1454 cm^{-1}) [203] and a decrease in tryptophan (1366, 1627 cm^{-1}) and collagen (849, 1037 cm^{-1}) [204]. Other studies have found an increase in DNA (780 cm^{-1}) and lipids (1440 cm^{-1}) [205], which has been corroborated in another study which assigned a peak at 1334 cm^{-1} to DNA [206]. This difference between a lipid peak and a DNA peak is only 6 cm^{-1}, uncomfortably close to the wavenumber shifting tolerated by some spectrometers. Numerous other biochemicals have been identified, but they have not been corroborated by other studies. These inconsistent and sometimes contradictory findings suggest that at least some of these associations are spurious. The work in this thesis suggests a few reasons why this may occur. For instance, all of the above studies employed baseline correction. As seen in section 5.3.2, this can change the interpretation of mean and difference spectra, which most studies used to detect biochemical differences [203, 206, 205]. One study additionally performed t-tests on a selection of Raman bands [204]. At least some of the statistically significant differences found will be spurious as there was no correction for type one error inflation associated with multiple hypothesis testing. This will only be exacerbated by the small samples sizes, especially in terms of statistical power (i.e. type two errors). Finally, the role of wavenumber shifting in identifying Raman peaks needs to be considered, especially

when using instruments of different makes and according to disparate protocols.

Between the oesophageal literature review [119] and the literature review conducted in section 1.5, I can be reasonably confident that this is a first application of deep learning to RS for classifying potentially malignant oesophageal tissues. Deep learning has been applied to standard H&E slides of oesophageal tissues [207]. This study was also multi-class, considering 4 tissue classes, and similarly found intermediate classes difficult to classify. In section 8.2.2 I consider how these two modalities, biochemical signals through RS and standard morphology, might be combined in a multi-modal deep learning architecture.

The decision not to use baseline correction for the SMART dataset was motivated by the findings from section 5.3 that baseline correction did not improve subsequent ML performance with the smaller datasets. Without a direct comparison of performance of baseline corrected and uncorrected SMART data it is not possible to draw firm conclusions justifying this choice. However, by not so correcting the data, the subsequent models are more parsimonious and have removed a potential hyperparameter choice that could allow for over fitting and exacerbate the generalisation gap.

The results from the SMART data demonstrate that system transferability can be achieved through engineering and clinical protocol alone, bypassing the need for instrument corrections. Additionally, the advantage of CNNs over PCA-LDA and SVMs is becoming apparent and would likely only increase with larger sample sizes. Although the performance metrics are competitive with existing modalities for oesophageal cancer diagnostics, this would require confirmation with a much larger sample size.

6.4.6 Data availability statement

The code used to construct the three models and to perform CV on the SMART data will be released on GitHub upon publication (under institutional review at the time of writing). Release of the SMART data itself will be discussed amongst all the

interested institutions and released if agreed upon and appropriate.

Chapter 7

Post Processing

> *Unweave a rainbow, as it erewhile made*
>
> John Keats

In oncology studies using RS it is common to compare the mean spectra and analyse the difference spectra of disease classes in an attempt to discern biochemical antecedents. However, that spectral differences are present between classes does not necessarily mean that they are clinically meaningful or that a model learned to use those features. Especially with small sample sizes, at least some of those differences will be simple random fluctuations. Some of the studies from the literature review of section 1.5 attempted to distinguish statistically significant Raman bands by comparing a selection between classes [70, 81, 84, 82, 66, 93]. However, none of those studies employed correction for type one error inflation inherent to multiple hypothesis testing (i.e. if multiple Raman bands are tested between classes, 1 in 20 will be 'statistically significant' by chance alone at the traditional p-value > 0.05). Even if properly treated (for instance with Bonferroni correction), and assuming that the sample was truly randomly selected - an oft forgotten but fundamental assumption of hypothesis testing - it does not follow that they are diagnostically significant. The ML model may learn to ignore features that have been determined statistically significant, if those features do not aid the model to discern between classes.

Deep learning architectures are generally considered black box techniques. This

is a key impediment to clinical adoption as being able to relate decisions to biomedical antecedents is a cornerstone of modern medicine. However, CNNs do offer a number of unique methods by which to illuminate the workings of the model and relate classifications to spectral and biochemical precursors.

7.1 Occlusion Studies

Occlusion studies have been developed in the context of image classification whereby a part of an image is occluded and the change in prediction of the occluded image is compared to that of the original, whole image. By moving the occluded patch over the image, a heat map can be created which shows those parts of the image which are diagnostically important to the model (figure 7.1). Thus it is possible to visually map what features the model learns to pay attention to. As an example, this has been used in the context of MRI based diagnosis of Alzheimers disease to find brain areas implicated in the pathology [208].

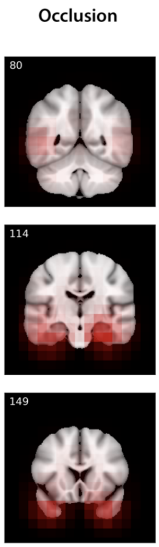


Figure 7.1: CNN occlusion study of the brain under MRI: patches of the image were systematically occluded. Areas that affected the models ability to predict disease are shown in red. Image taken from [208]

I demonstrate how this technique can be used in RS, using the Lynch dataset.

Specifically I have taken the binary CNN model that was used to discern between MSI-H and MSS, described in section 6.2.2. The mean spectra of the MSI-H and MSS class were iteratively 'occluded' by setting their value to zero in 30 cm^{-1} sliding wavenumber segments. Remembering that a model output can be interpreted as a probability (discussed in chapter 3), the change in probability in class membership between the whole spectrum and the occluded spectra was obtained. Figure 7.2 demonstrates how the occluded segments were fed into the model and the attendant change in prediction due to the occlusion tracked.

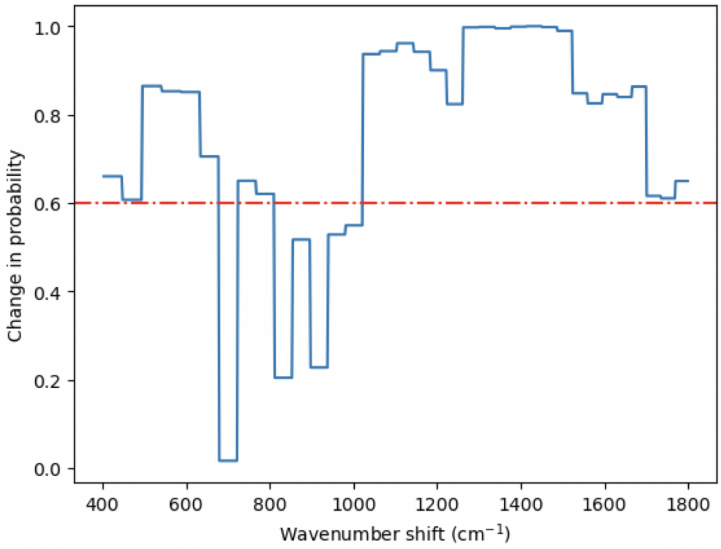


Figure 7.2: Spectral Occlusion Schematic: the 'steps' show 30cm^{-1} spectral segments that were occluded, with their attendant change in prediction for the positive class. Red dotted line shows the prediction probability for the whole spectrum.

From the spectral heat maps in figure 7.3 we can see that important features for the CNNs ability to detect MSI-H can be found at $680-710cm^{-1}$, $800-830cm^{-1}$ and $870-900cm^{-1}$. These areas are associated with nucleic acids and proteins. These regions do not correspond to Raman bands thought important by inspecting the difference spectrum between MSI-H and MSS mean spectra (section 6.2.2). Although tentative, it is consistent with the theory that MSI-H is caused by DNA MMR protein disregulation [209]. If we are able accept the results of the occlusion study, then we might treat them as hypothesis generating, guiding clinicians towards potential biomarkers hitherto not considered.

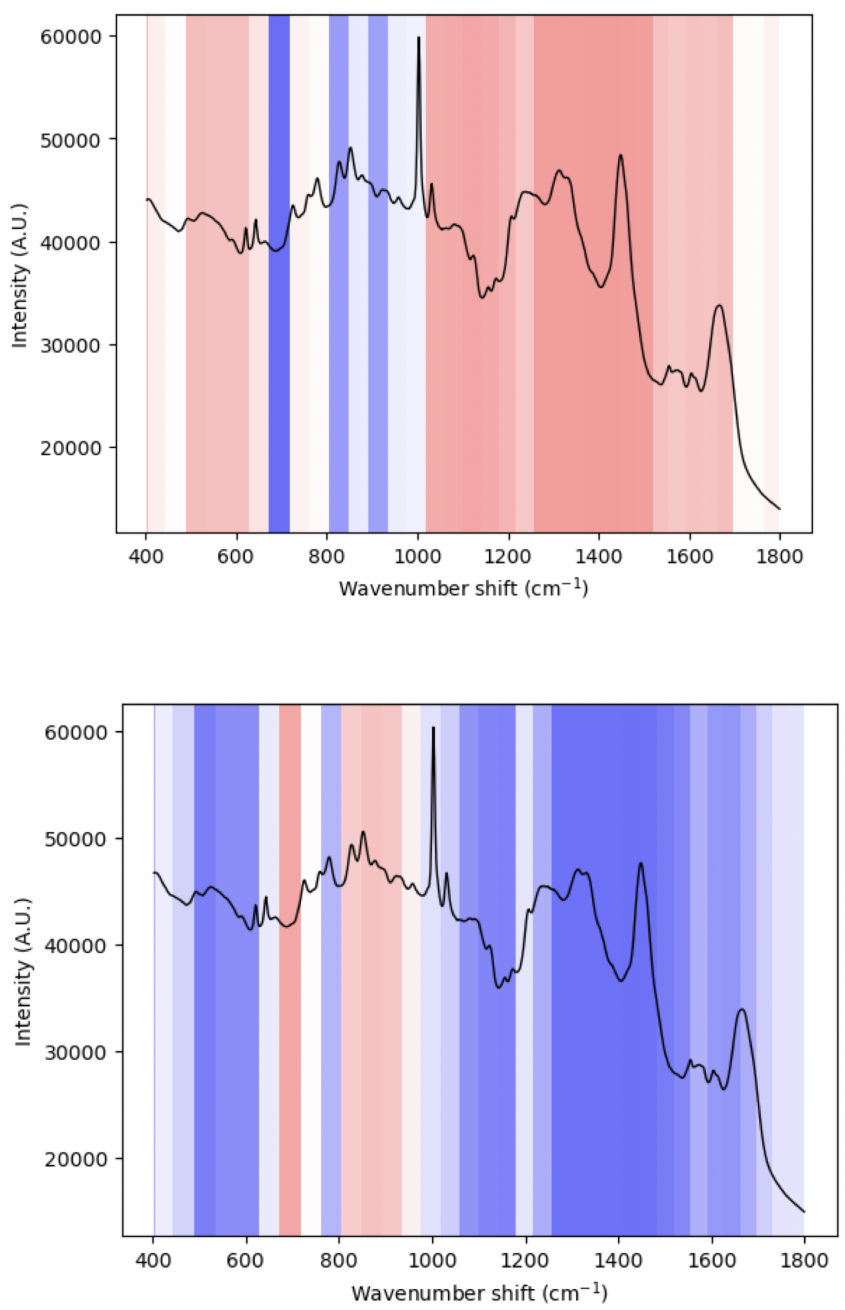


Figure 7.3: Top: MSI-H occlusion map of mean spectrum. Bottom: MSS Occlusion map of mean spectrum. Blue areas indicate decreases in predicting the positive class, red ares to increases. The intensity of colour corresponds to the degree to change in prediction

It is possible to perform spectral occlusion on any single spectrum to discern biochemical information. Here I have used the class average spectra under the assumption that a mean spectrum contains diagnostically pertinent information. This

is likely true to an extent, but it is the same assumption underlying difference spectra and manual comparisons of mean spectra. Additionally, the mean spectra shown above only had probabilities of 0.6 for the positive class, casting some doubt on their utility. A model and spectrum with a more confident prediction would provide a more convincing basis for such analysis. With the next technique we seek to explore a visualisation that uses all spectra the model has seen.

7.2 Sample SMART maps

In this section we explore whether the predictions given by the CNN correspond to any discernible morphological features, which, if present, would add confidence to the predictions. The SMART maps were of sufficient size and spatial quality to attempt such visualisation. There is nothing unique about CNNs for producing such maps and could be done with any model, though I restrict attention to the CNN for brevity.

Figure 7.4 shows a visualisation process from H&E through the Raman mapping process ending with a pseudo-coloured map showing the class labels. The whole H&E sample and the identified region of interest were labelled as AC by the study histopathologist. The CNN labels 69.6% of the spectra as AC, with NSQ the next most common at 16.6%. Unfortunately, the classification map does not correspond to any morphological features, even though some structure is evident in the raw Raman map.

It is evident that no morphological information is being captured by the model and the (correct) prediction is given entirely by the biochemical information extracted from the spectra. The opposite is true of the original classification made by the histopathologists, who used only morphological information. Indeed, most of the attention given to deep learning in the context of histopathology is to networks that utilise morphological information [210]. But it is clear that both biochemical and morphological features can contribute to disease diagnosis and that they contain non-overlapping information that may improve prediction further than any one modality alone. In the next chapter I conclude by briefly considering how multi-model deep

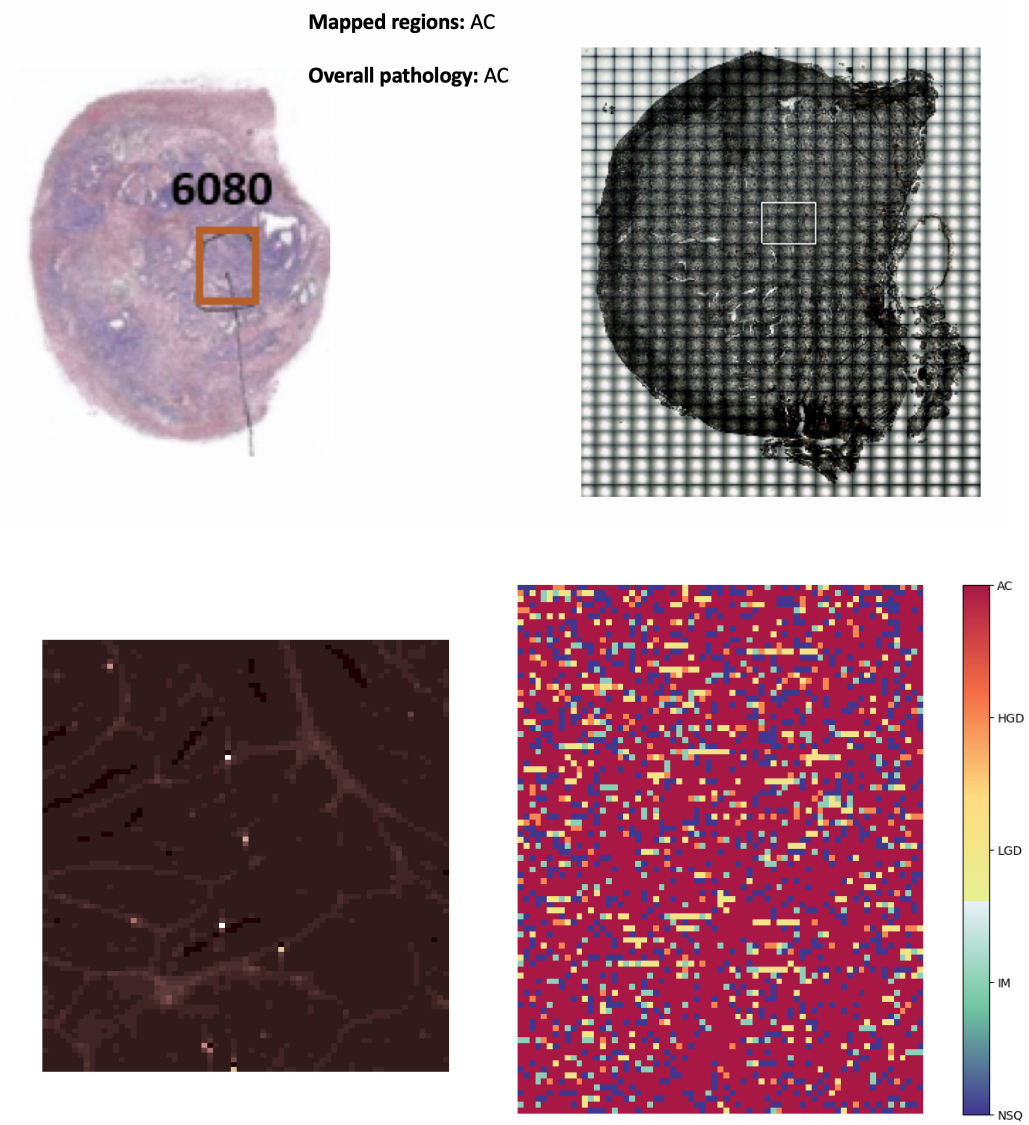


Figure 7.4: SMART Raman Map Visualised. Top left: H&E stained section classified as AC, region of interest outlined in red. Top right: montage view under the microspectrometer of an adjacent section of the same sample, with the area to be mapped outlined in white. Bottom left: raw Raman map corresponding to the white outlined area, visualised at 828 cm^{-1}. Right panel, same Raman map visualised as disease classifications for each spectra.

learning architectures may simultaneously exploit spectral and spatial data as a direction for future research.

Chapter 8

Conclusion

> *The outcome of any serious research can only be to make two questions grow where only one grew before*
>
> Thorstein Veblen

8.1 Conclusions

This thesis has shown that there are a number of subtle methodological considerations that need to be taken into account in order to return realistic assessments of how models applied to RS may perform in real oncology settings. Section 5.3 provided evidence that a staple of RS pre-processing, baseline correction, may not improve model performance, and could even obfuscate biochemical interpretations of mean and difference spectra. This is likely to be a controversial finding given the ubiquity of this family of techniques in the literature. A valid criticism of my approach is that a more systematic search of the baseline correction algorithm hyperparameter space would find those corrections that enhance the performance of a model. For instance, a genetic algorithm may better find optimal hyperparameters compared to the grid search strategy I used. However, as explored in section 5.8 this would need to be combined with a CV strategy that accounts for the search process so that overfitting does not occur at this second level of inference. Based on the three datasets used in this thesis, I have found that the cost of not properly selecting hyperparameters

is an over-estimate in model accuracy of 5-10%. In this thesis I have additionally shown how nested CV can be used to ameliorate this over-estimation. Though not a new finding, the fact that it is far from common, as shown in literature review in section 1.5, shows that this consideration has yet to become established in the oncological Raman community. Regardless, it is clear that the arbitrary selection of hyperparameters based, for instance, on what worked well on unrelated datasets, and ad hoc trial and error approaches, is not tenable. Additionally, the SMART results indicate that algorithmic instrument correction is not necessary to establish system transferability when the same make of instrument is used in conjunction with a common protocol.

In section 5.4 we saw that any study classifying individual spectra but using multiple spectra from the same sample and/or patient must divide spectra into training and test sets such that the required independence assumption is met. The consequences for not doing this are an approximate inflation of 20% in the models estimated accuracy, leading to gross over-estimates of performance. This has been noted elsewhere in the literature, but as evident from the systematic literature review in section 1.5 it is often not applied. This hierarchical structure can be flattened by taking the average spectrum per sample. In section 5.5 I suggested a simple alternative for using spectra to classify whole samples, using simple consensus voting to choose a class. I have shown this improves accuracy by a few percent, and could potentially be further improved with the development of more sophisticated voting regimes.

The role of data augmentation for RS remains unclear, both in terms of the existing literature and the results in this thesis. I explored two methods which provided no benefit on two datasets and a minimal difference to the Lynch dataset. It is not clear why this is so, but a possible reason is that a datasets SNR must be of a certain quality for these augmentation techniques to work. GANs have been found to provide superior data augmentation in RS, and are the most likely method to improve this particular aspect of training.

Taking into account all the above methodological issues I have explored 3

datasets. The results do not achieve the 80-100% accuracy common in the reviewed literature in section 1.5, but likely give a more realistic assessment of how the models would perform when deployed to real clinical settings, given the stringent methodological constraints outlines throughout this thesis.

The Lynch data performed best with the CNN, despite having the lowest sample size in terms of spectra. However, as discussed in section 5.5, it is the number of samples, not spectra, that is more important in this determination. This shows that small datasets with sufficient SNR can still benefit from deep learning architectures, although the extent of this benefit is questionable based on these results. Traditional wisdom would posit that the CNN would benefit most when more data is collected compared to the traditional models, as has been demonstrated with NIR spectroscopy [211]. These results are competitive with existing diagnostic modalities and are being prepared for publication (in peer review at the time of writing).

The Ovarian dataset gave far more ambiguous results. No model was obviously better and the standard deviations of the results make any conclusions tentative. Although the Ovarian dataset was larger than the Lynch in terms of spectra, it was smaller in terms of samples. It also suffered from a lower SNR. It may be that the clinical problem is inherently more difficult, with genetic methods giving similar results. Although the results seem disappointing, the consulting histopathologist (Dr Florian Heintz) is keen to explore the technique further and the data acquired in this thesis is being used in a proposal for research project funding.

The SMART dataset was the largest in terms of spectra and samples collected, though the worst in terms of SNR. However, this dataset gave the least ambiguous results, demonstrating that system transferability can be achieved through engineering and clinical protocol alone, bypassing the need for instrument corrections. Additionally, the advantage of CNNs over PCA-LDA and SVMs is evident and would likely only increase with larger sample sizes. Although the performance metrics are competitive with existing modalities for oesophageal cancer diagnostics, such as ultrasound [212], this would require confirmation with a much larger sample size. The SMART data is being prepared for a publication focussing on clinical RS model

transferability.

Overall, I have shown that deep learning architectures can be competitive with traditional models for RS applications in oncology. Although any benefit is only minimal with low sample sizes, as sample sizes increase, so too will DL performance. Even with small sample sizes, there are a number of additional techniques available which can further enhance DL performance under certain conditions. Some of these have been explored in this thesis, such as data augmentation, while others are the topic of ongoing and future work.

8.2 Future work

8.2.1 Transfer learning

Transfer learning is a mechanism by which a previously trained deep model can be used to classify a different dataset. This works even with models that have been trained on datasets quite different to the orginal application. For instance a model called VGG-16, which was initially trained on everyday images such as pets and vehicles, was transferred to the oncology setting to classify whole slide images [213]. This works because the deeper layers of a CNN are learning to identify simple features such as vertical or horizontal edges, and these are common to all images. It involves taking the model weights learned during training, and transferring them across to a new model with a similar architecture. There are a number of options once this transfer is complete. The weights could be frozen so that they do not adjust in light of the new dataset. This may work for large models trained on huge datasets and transferred to a model performing a similar task. However, this might not work with models trained on smaller datasets or transferring across quite different applications. Fine-tuning may then be preferable, whereby the pre-trained model weights are used during initialisation instead of random weights (such as Kaiming He initialisation described in section 4.5). This allows the model to be adjusted by the existing dataset. Transferring models can be flexible and some combination of freezing and fine-tuning is possible; for instance, the weights of the convolutional layer could be frozen and the weights of the fully connected layers fine-tuned.

A potential candidate for transfer learning in RS is provided by Ho *et al*, who used a ResNet 18 model to classify bacteria type and determine antibiotic resistant using approximately 74000 Raman spectra during training [214]. This model is an order of magnitude larger than the CNN I developed with approximately 1.3 million parameters compared to 0.2 million. As the model has been trained on Raman spectra, it may benefit from freezing the convolutional layers and fine-tuning the fully connected layers to prevent overfitting.

Transfer learning has been shown to improve accuracy by about 5% on a RS dataset of organic compounds [215]. The potential of model transfer provides an additional reason why model and data sharing are beneficial to model development within the medical RS community.

8.2.2 Combined biochemical and morphological models

Traditional histopathology uses H&E samples to distinguish between cancerous tissues. This is done using morphological information based on what is discernible to the microscope-aided eye. In the -omics age of genetic and molecular data these traditional techniques are said to be becoming less relevant [216]. However, there is information in the morphology of a sample, even if it eventually proves less discriminating than molecular information. RS has an advantage over many molecular modalities in that it can easily be combined with existing work flows to produce Raman maps. These can be understood as hyperspectral cubes which contain morphological information in the x-y plane (as seen in a typical H&E section), and biochemical information in the z plane in the form of Raman spectra (figure 8.1). Thus RS can provide combined biochemical and morphological datasets.

A collection of such images could be trained using a typical 3D CNN usually used for image classification. However, this would mean a single sample would constitute a deluge of information requiring copious amounts of training data. Such large datasets are hard, though not impossible, to obtain for medical images, but the need for each to undergo Raman mapping could mount a formidable obstacle to any practical project. An alternative could be to adopt transfer learning as outlined above in order to combine a 1D CNN trained to classify Raman spectra and a 3D CNN

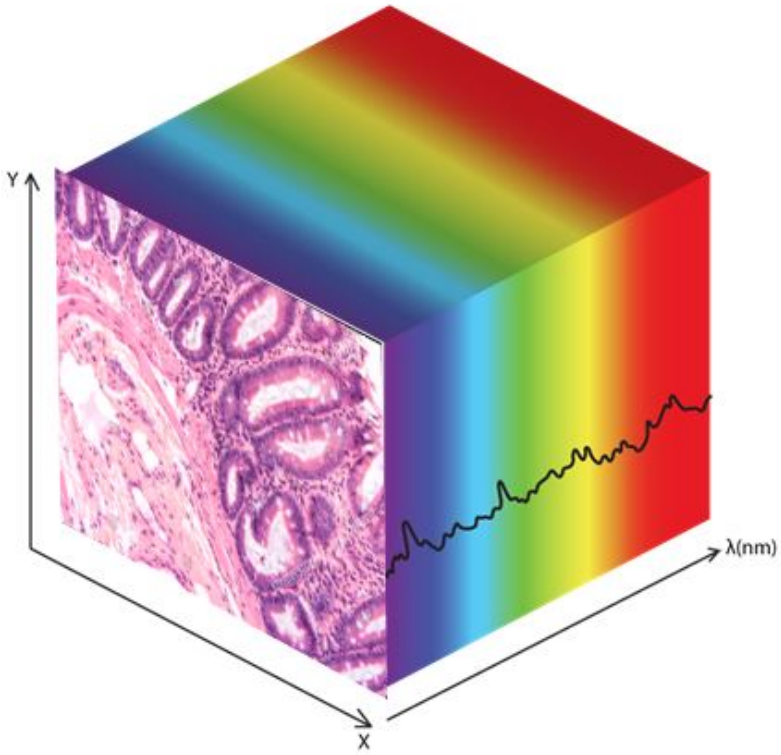


Figure 8.1: Hyperspectral Image: the x, y-plane represents morphological information. The z-plane represents biochemical information contained in Raman spectra

trained to classify images.

Figure 8.2 shows a schematic for how this might proceed. A Raman map would be acquired per sample, as was done with the SMART dataset. Each spectrum in the map could then be fed into a pre-trained 1D CNN. The ResNet model provided by Ho *et al* could be used for this component, adjusted to give three outcomes which would represent biochemical maps. These maps would then be the size of a standard 3D CNN which has been trained on RGB images. A candidate for this is a custom made attention-based 3D CNN which has been trained on data very similar to the SMART dataset, using 123 oesophageal histology images to distinguish between four disease classes [207]. The advantage of these two models is that they have both been trained in Pytorch, making their technical concatenation far simpler. Thus combined, the model may be able to take into account biochemical information via the first component and morphological information in the second component followed by a classification layer. The intermediary stage linking the two models also allows for

the visualisation of biochemical Raman maps which may aid interpretation, opening the black box. This could be compared to a technique such as PCA , which can also reduce the spectral data to three biochemical maps, providing a benchmark to measure if the mixed-modality model does indeed provide a benefit. This is but one possible integration of the two models, and other architectures may prove more fruitful.

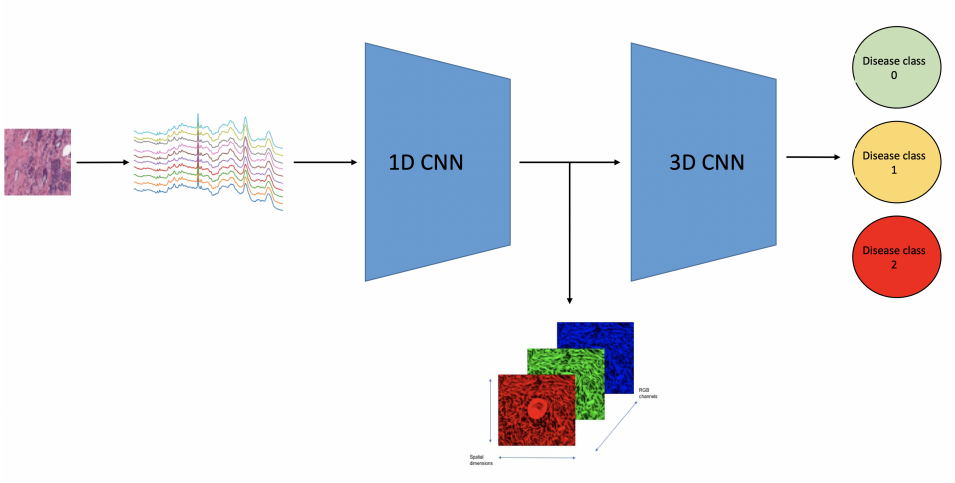


Figure 8.2: A double transferred CNN to combine biochemical and morphological information into a mixed-modality DL model

Deep learning promises to unlock the potential of RS in oncology applications; but great potential calls for restraint. With so many model choices, it is all too easy to train models with extremely impressive performance within a narrow research domain, but that we are unable to apply to the clinical setting where the technology is needed. Deep learning represents a paradigm shift in traditional statistical thinking, allowing data to generate the model. Therefore great care needs to be taken in how data is curated and processed, and providing just one more reason why data and model sharing is so important to the development of models in the biomedical domain.

Appendix A

Table of Inter-rater Reliability Literature

The table on the following pages summarises the literature considered for the meta-analysis in section 1.2.

Study/ Year/ Country	**Study design/ outcomes**	**Sample type**	**Protocol followed**	**Number of samples**	**Number/ type of pathologist**	**Intra-rater reliability** Kappa (95% CI) unless otherwise stated	**Inter-rater reliability** Kappa (95% CI) unless otherwise stated	**Notes**
Thomas et al. 1983 **UK**	Grade of pre and post-operative samples (3 categories)	Rectal carcinomas pre and post-operative samples Paraffin embedded H&E stained	No criteria defined, pathologists asked to grade as per a 'routine specimen'	50 biopsies 50 resections	1 GI specialist 4 non-specialists	Biopsy samples GI specialist 0.613 (0.461 – 0.765) Non-specialist 0.539 (0.387 – 0.691) Resected samples GI specialist 0.630 (0.485 - 0775) Non-specialist 0.546 (0.383 - 0709)	Biopsy samples 0.303 (0.229 – 0.378) Resected samples 0.356 (0.279 – 0.433)	Poor quality samples were excluded Kappa not weighted
Brown et al. 1985 **UK**	Grade (3 categories)	Colorectal adenomas H&E stained	'Published guidelines'	100 for comparison between specialists 50 for comparison including non-specialists	2 GI specialists 2 non-specialists	GI specialists 0.398 (0.273 – 0.521)	GI specialists 0.288 (0.178 – 0.400) Non-Specialists 0.278 (0.043 – 0.513) Specialists vs non-specialists 0.108 (0.000 – 0.216)	Kappa not weighted

Study/ Year/ Country	**Study design/ outcomes**	**Sample type**	**Protocol followed**	**Number of samples**	**Number/ type of pathologist**	**Intra-rater reliability** Kappa (95% CI) unless otherwise stated	**Inter-rater reliability** Kappa (95% CI) unless otherwise stated	**Notes**
Dundas et al. 1988 **UK**	Expanding vs infiltrating tumour edges Peritumoural lymphocytic infiltration	Rectal or sigmoid colon carcinomas H&E stained	Based on a previous publication	60 samples taken from 30 blocks	3 GI specialists 3 other specialists	Tumour Edges Specialists 0.720 (0.608 – 0.832) Non-specialists 0.406 (0.217 - 0.596) Infiltration Specialists 0.328 (0.168 – 0.488) Non-specialists 0.237 (0.085 - 0.388)	Insufficient data given in the results	
Fenger et al. 1990 **Denmark**	Compares 2 dysplasia grading systems (both have 3 categories)	Colorectal adenomas Paraffin embedded H&E stained	Konishi-Morson system (KMS) vs Extended Kozuka system (EKS)	112	2 GI specialists	Weighted Kappa (95% CI) KMS 0.795 (0.608 – 0.982) EKS 0.689 (0.539 – 0.840)	Weighted Kappa (95% CI) KMS - 0.48 (0.35 - 0.61) EKS - 0.42 (0.31 - 0.52)	The 112 samples are from 2 assessments of 56 samples a minimum of 1 week apart
Jensen et al. 1995 **Denmark**	Grading (3 categories) Type (tubular, villous, and tubulovillous)	Archive H&Es of adenomas randomly selected to have different grades	Morson and Dawson's criteria for grade Morson and Sobin's criteria for type	194	3 GI specialists	Weighted kappa (group average) Type 0.8003 Dysplasia 0.6683	Weighted kappa (group average) Type 0.5812 0.5900 Dysplasia 0.3832 0.4540	Also includes unweighted kappa values. Gives no metric to assess variance of these statistics

Study/ Year/ Country	**Study design/ outcomes**	**Sample type**	**Protocol followed**	**Number of samples**	**Number/ type of pathologist**	**Intra-rater reliability** Kappa (95% CI) unless otherwise stated	**Inter-rater reliability** Kappa (95% CI) unless otherwise stated	**Notes**
Schlemper et al. 2000 **Japan**	Compare Western and Japanese pathologists on two grading systems (both 5 category)	Colorectal lesions Archived glass slides	None specified originally then Vienna classification	20	31 GI specialists	Not assessed	Japanese vs Western No system 0.27 (0.04 – 0.49) Japanese vs Western Vienna system 0.47 (0.18 – 0.76)	Kappa not weighted
Cross et al. 2000 **UK**	Hyperplastic vs adenomatous	Colorectal polyps H&E stained	None specified	100	8 experienced 1 in-experienced	Not assessed	Experienced 0.846 (0.838 – 0.854) Inexperienced 0.510 (0.452 – 0.570)	

Study/ Year/ Country	Study design/ outcomes	Sample type	Protocol followed	Number of samples	Number/ type of pathologist	Intra-rater reliability Kappa (95% CI) unless otherwise stated	Inter-rater reliability Kappa (95% CI) unless otherwise stated	Notes
Terry et al. 2002 **USA**	Type (tubular, villous, and tubulovillous) Grade (2 category and 4 category	Advanced and non-advanced adenomas	Grade by WHO classification	99	2 GI specialists	Type GI specialist 0.48 (0.33 – 0.62) weighted 0.53 (0.40 – 0.66)	Type 0.48 (0.33 - 0.62) weighted 0.59 (0.40 - 0.66) Grade (4 category) GI specialists 0.42 (0.29 - 0.55) weighted 0.59 (0.47 - 0.70) Grade (2 category) GI specialists 0.69 (0.55 - 0.83)	A cohort of 'community' pathologists excluded from synthesis due to ambiguous data.
Constantini et al. 2003 **Italy**	Hyperplastic vs adenomatous Type (tubular, villous, and tubulovillous) Grade (2 category)	Colorectal polyps	WHO classification	100	4 GI specialists	Not assessed	Hyperplastic vs adenoma 0.90 (0.82 – 0.98) Type 0.34 (0.28 - 0.41) Grade 0.54 (0.48 – 0.61)	Kappa not weighted

Study/ Year/ Country	**Study design/ outcomes**	**Sample type**	**Protocol followed**	**Number of samples**	**Number/ type of pathologist**	**Intra-rater reliability** Kappa (95% CI) unless otherwise stated	**Inter-rater reliability** Kappa (95% CI) unless otherwise stated	**Notes**
Komuta et al. 2004 **Japan**	Grade (2 categories) Invasion (5 categories)	Malignant colorectal polyps	Haggit's classification for Invasion	88	3 GI specialists	Not assessed	Pooled weighted Kappa Grade 0.163 Invasion 0.682	Request for additional info sent - pending
Denis et al. 2008 **France**	Type (tubular, villous, and tubulovillous) Grade (4 categories)	Screen detected colorectal polyps	WHO classification Vienna classification for grade	297	2 GI specialists	Not assessed	Type 0.41 Grade 0.67	Request for additional info sent - unable to contact
van Putten et al. 2011 **Netherlands**	Non-adenomatous vs adenomatous Type (tubular or villous component) Grade (2 categories)	Screen detected colorectal polyps	WHO classification	Adeno-matous - 440 Type - 322 Grade - 322	1 GI specialist 1 non-specialist	Not assessed	Adenomatous Specialists 0.86 (0.73 - 0.98) Non-specialist vs specialist 0.89 (0.83 - 0.95) Grade 0.62 (0.46 - 0.79) Type 0.55 (0.44 - 0.66)	

Study/ Year/ Country	**Study design/ outcomes**	**Sample type**	**Protocol followed**	**Number of samples**	**Number/ type of pathologist**	**Intra-rater reliability** Kappa (95% CI) unless otherwise stated	**Inter-rater reliability** Kappa (95% CI) unless otherwise stated	**Notes**
Foss et al. 2012 **UK**	Adenomatous vs non-Adenomatous Type (4 categories) Dysplasia (2 categories) Stromal Invasion (yes/no) Excision status (complete, incomplete)	Screen detected colorectal polyps	NHS BCSP	239	2 GI specialists	Not assessed	Adenoma 0.83 (0.73 - 0.93) Type 0.45 (0.38 - 0.52) Grade 0.60 (0.53 - 0.67) Stromal Invasion 0.84 (0.72 - 0.96) Excision 0.64 (0.55 - 0.73)	Serrated adenoma data excluded from synthesis
Osmond et al. 2014 **Canada**	Type (tubular, villous, and tubulovillous) Grade (2 categories)	Archived H&E slides of colorectal adenomatous polyps	Pre and post distribution of National Colorectal Cancer Screening Network (Canadian) guidelines	40	6 GI specialists 6 non-specialists	Insufficient data for synthesis	Type Non-specialists Pre: 0.38 (0.32-0.45) Post: 0.51 (0.45-0.57) Specialists Pre: 0.50 (0.44-0.56) Post: 0.56 (0.49-0.62) Dysplasia Non-specialists Pre: 0.58 (0.49-0.67) Post: 0.41 (0.32-0.51) All Pre: 0.57 (0.52-0.61) Post: 0.48 (0.44-0.53)	Distribution of samples chosen to maximise power Request for additional info sent - pending Also includes data on reliability by polyp size

Study/ Year/ Country	**Study design/ outcomes**	**Sample type**	**Protocol followed**	**Number of samples**	**Number/ type of pathologist**	**Intra-rater reliability** Kappa (95% CI) unless otherwise stated	**Inter-rater reliability** Kappa (95% CI) unless otherwise stated	**Notes**
Davenport et al. 2016 **UK**	Margin involvement Grade Tumour differentiation Lymphovascular invasion Tumour budding	H&E 'proven polyp cancer'.	WHO classification and European guidelines	56	4 GI specialists	Margin 0.23 Grade 0.07 Lymphovascular 0.35	Kappa stats but no SE/CI - might be sufficient info to reconstruct	Request for additional info sent - pending

Appendix B

Lynch Data Patient Characteristics

Sample ID	Sample Type	TNM Stage	Tumour Grade
MSI-H1	Resection cancer	T2 N0 M0	Mod. Diff.
MSI-H2	Resection cancer	T2 N0 M0	Mod. Diff.
MSI-H3	Resection cancer	T2 N0 M0	Mod. Diff.
MSI-H4	Resection cancer	T3 N0 M0	Poor diff.
MSI-H5	Resection cancer	T3 N0 M0	Mod. Diff.
MSI-H6	Resection cancer	T3 N0 M0	Mod. Diff.
MSI-H7	Resection cancer	T3.N1.Mx	Poor diff.
MSI-H8	Resection cancer	T3 N1 M0	Poor diff.
MSI-H9	Resection cancer	T4 N0 M0	Mod. Diff.
MSI-H10	Resection cancer	T4 N1 M0	Mod. Diff.
AC1	Resection cancer	T2 N2 M0	Mod. Diff.
N1	Normal	-	-
AC2	Resection cancer	T2 N0 M0	Mod. Diff.
N2	Normal	-	-
AC3	Resection cancer	T2 N0 M0	Mod. Diff.
N3	Normal	-	-
AC4	Resection cancer	T3 N1 M0	Mod. Diff.
N4	Normal	-	-
AC5	Resection cancer	T3 N3 M0	Mod. Diff.
N5	Normal	-	-
AC6	Resection cancer	T3 N0 M0	Mod. Diff.
N6	Normal	-	-
AC7	Resection cancer	T3 N0 M0	Mod. Diff.
N7	Normal	-	-
AC8	Resection cancer	T3 N0 M0	Mod. Diff.
N8	Normal	-	-
AC9	Resection cancer	T4 N2 M0	Mod. Diff.
N9	Normal	-	-
AC10	Resection cancer	T4 N0 M1	Poor diff.
N10	Normal	-	-

Table B.1: Breakdown of patient samples used to build the ML models for the classification of normal (N), sporadic adenocarcinoma (MSS AC) and micro-satellite instability high (MSI-H) patients.

Bibliography

[1] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: a cancer journal for clinicians*, 70(1):7–30, 2020.

[2] Nigel Hawkes. Cancer survival data emphasise importance of early diagnosis. British Medical Journal Publishing Group, 2019.

[3] Great Britain. *Cancer survival in England: national estimates for patients followed up to 2017*. Office for National Statistics, 2019.

[4] Kevin M Elias, Jing Guo, and Robert C Bast. Early detection of ovarian cancer. *Hematology/Oncology Clinics*, 32(6):903–914, 2018.

[5] Abel Joseph, Siva Raja, Suneel Kamath, Sunguk Jang, Daniela Allende, Mike McNamara, Gregory Videtic, Sudish Murthy, and Amit Bhatt. Esophageal adenocarcinoma: A dire need for early detection and treatment. *Cleveland Clinic Journal of Medicine*, 89(5):269–279, 2022.

[6] Enrique Rodríguez de Santiago, Nerea Hernanz, Héctor Miguel Marcos-Prieto, Miguel Ángel De-Jorge-Turrión, Eva Barreiro-Alonso, Carlos Rodríguez-Escaja, Andrea Jiménez-Jurado, María Sierra-Morales, Isabel Pérez-Valle, Nadja Machado-Volpato, et al. Rate of missed oesophageal cancer at routine endoscopy and survival outcomes: A multicentric cohort study. *United European gastroenterology journal*, 7(2):189–198, 2019.

[7] Marzieh Araghi, Isabelle Soerjomataram, Mark Jenkins, James Brierley, Eva Morris, Freddie Bray, and Melina Arnold. Global trends in colorectal cancer

mortality: projections to the year 2035. *International journal of cancer*, 144(12):2992–3000, 2019.

[8] Hugh Alderwick and Jennifer Dixon. The nhs long term plan. *British Medical Journal*, 364, 2019.

[9] Scarlet Fiona Brockmoeller and Nicholas Paul West. Predicting systemic spread in early colorectal cancer: Can we do better? *World journal of gastroenterology*, 25(23):2887, 2019.

[10] Francesco Del Giudice, Martina Pecoraro, Hebert Alberto Vargas, Stefano Cipollari, Ettore De Berardinis, Marco Bicchetti, Benjamin I Chung, Carlo Catalano, Yoshifumi Narumi, James WF Catto, et al. Systematic review and meta-analysis of vesical imaging-reporting and data system (vi-rads) inter-observer reliability: An added value for muscle invasive bladder cancer detection. *Cancers*, 12(10):2994, 2020.

[11] Mieke R Van Bockstal, Martine Berlière, Francois P Duhoux, and Christine Galant. Interobserver variability in ductal carcinoma in situ of the breast. *American journal of clinical pathology*, 154(5):596–609, 2020.

[12] Matthew Fleming, Sreelakshmi Ravula, Sergei F Tatishchev, and Hanlin L Wang. Colorectal carcinoma: pathologic aspects. *Journal of gastrointestinal oncology*, 3(3):153, 2012.

[13] Douglas K Rex and John R Goldblum. Should hgd or degree of villous changes in colon polyps be reported? *American Journal of Gastroenterology*, 103(6):1327–1329, 2008.

[14] Sidney J Winawer, Ann G Zauber, Robert H Fletcher, Jonathon S Stillman, Michael J O'Brien, Bernard Levin, Robert A Smith, David A Lieberman, Randall W Burt, Theodore R Levin, et al. Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the us multi-society task force on colorectal cancer and the american cancer society. *CA: a cancer journal for clinicians*, 56(3):143–159, 2006.

[15] C C Compton. Updated protocol for the examination of specimens from patients with carcinomas of the colon and rectum, excluding carcinoid tumors, lymphomas, sarcomas, and tumors of the vermiform appendix: a basis for checklists. cancer committee. *Arch Pathol Lab Med*, 124(7):1016–1025, Jul 2000.

[16] C C Compton, L P Fielding, L J Burgart, B Conley, H S Cooper, S R Hamilton, M E Hammond, D E Henson, R V Hutter, R B Nagle, M L Nielsen, D J Sargent, C R Taylor, M Welton, and C Willett. Prognostic factors in colorectal cancer. college of american pathologists consensus statement 1999. *Arch Pathol Lab Med*, 124(7):979–994, Jul 2000.

[17] Mary Beth Terry, Alfred I Neugut, Roberd M Bostick, John D Potter, Robert W Haile, and Cecilia M Fenoglio-Preiser. Reliability in the classification of advanced colorectal adenomas. *Cancer Epidemiology and Prevention Biomarkers*, 11(7):660–663, 2002.

[18] Jesse A Berlin and Robert M Golub. Meta-analysis as evidence: building a better pyramid. *Jama*, 312(6):603–606, 2014.

[19] R Brian Haynes, Nancy Wilczynski, K Ann McKibbon, Cynthia J Walker, and John C Sinclair. Developing optimal search strategies for detecting clinically sound studies in medline. *Journal of the American Medical Informatics Association*, 1(6):447–458, 1994.

[20] Mariska MG Leeflang, Jonathan J Deeks, Constantine Gatsonis, and Patrick MM Bossuyt. Systematic reviews of diagnostic test accuracy. *Annals of internal medicine*, 149(12):889–897, 2008.

[21] Lucas M Bachmann, Reto Coray, Pius Estermann, and Gerben Ter Riet. Identifying diagnostic studies in medline: reducing the number needed to read. *Journal of the American Medical Informatics Association*, 9(6):653–658, 2002.

[22] Lucas M Bachmann, Pius Estermann, Corinna Kronenberg, and Gerben ter Riet. Identifying diagnostic accuracy studies in embase. *Journal of the Medical Library Association*, 91(3):341, 2003.

[23] Nicholas P Lucas, Petra Macaskill, Les Irwig, and Nikolai Bogduk. The development of a quality appraisal tool for studies of diagnostic reliability (qarel). *Journal of clinical epidemiology*, 63(8):854–861, 2010.

[24] Arzu Ensari, Banu Bilezikçi, Fatima Carneiro, Gülen Bülbül Doğusoy, Ann Driessen, Ayşe Dursun, Jean-François Flejou, Karel Geboes, Gert De Hertogh, Anne Jouret-Mourin, et al. Serrated polyps of the colon: how reproducible is their classification? *Virchows Archiv*, 461(5):495–504, 2012.

[25] Yark Hazewinkel, Thomas R de Wijkerslooth, Esther M Stoop, Patrick M Bossuyt, Katharina Biermann, Marc J van de Vijver, Paul Fockens, Monique E van Leerdam, Ernst J Kuipers, and Evelien Dekker. Prevalence of serrated polyps and association with synchronous advanced neoplasia in screening colonoscopy. *Endoscopy*, 46(03):219–224, 2014.

[26] Charles Muller, Akihiro Yamada, Sachie Ikegami, Haider Haider, Yuga Komaki, Fukiko Komaki, Dejan Micic, and Atsushi Sakuraba. Risk of colorectal cancer in serrated polyposis syndrome: a systematic review and meta-analysis. *Clinical Gastroenterology and Hepatology*, 20(3):622–630, 2022.

[27] C Fenger, Martin Bak, O Kronborg, and H Svanholm. Observer reproducibility in grading dysplasia in colorectal adenomas: comparison between two different grading systems. *Journal of clinical pathology*, 43(4):320–324, 1990.

[28] Massimo Costantini, Stefania Sciallero, Augusto Giannini, Beatrice Gatteschi, Paolo Rinaldi, Giuseppe Lanzanova, Luigina Bonelli, Tino Casetti, Elisabetta Bertinelli, Orietta Giuliani, et al. Interobserver agreement in the histologic diagnosis of colorectal polyps: the experience of the multicenter adenoma

colorectal study (smac). *Journal of clinical epidemiology*, 56(3):209–214, 2003.

[29] Simon S Cross, Samar Betmouni, Julian L Burton, Asha K Dubé, Kenneth M Feeley, Miles R Holbrook, Robert J Landers, Phillip B Lumb, and Timothy J Stephenson. What levels of agreement can be expected between histopathologists assigning cases to discrete nominal categories? a study of the diagnosis of hyperplastic and adenomatous colorectal polyps. *Modern Pathology*, 13(9):941–944, 2000.

[30] LJ Brown, NC Smeeton, and MF Dixon. Assessment of dysplasia in colorectal adenomas: an observer variation and morphometric study. *Journal of clinical pathology*, 38(2):174–179, 1985.

[31] GD Thomas, MF Dixon, NC Smeeton, and NS Williams. Observer variation in the histological grading of rectal carcinoma. *Journal of Clinical pathology*, 36(4):385–391, 1983.

[32] RJ Schlemper, RH Riddell, Y e al Kato, F Borchard, HS Cooper, SM Dawsey, MF Dixon, CM Fenoglio-Preiser, JF Fléjou, Karel Geboes, et al. The vienna classification of gastrointestinal epithelial neoplasia. *Gut*, 47(2):251–255, 2000.

[33] Bernard Denis, Carol Peters, Catherine Chapelain, Isabelle Kleinclaus, Anne Fricker, Richard Wild, Bernard Auge, Isabelle Gendre, Philippe Perrin, Denis Chatelain, et al. Diagnostic accuracy of community pathologists in the interpretation of colorectal polyps. *European journal of gastroenterology & hepatology*, 21(10):1153–1160, 2009.

[34] Fiona A Foss, Steve Milkins, and Angus H McGregor. Inter-observer variability in the histological assessment of colorectal polyps detected through the nhs bowel cancer screening programme. *Histopathology*, 61(1):47–52, 2012.

[35] Paul G Van Putten, Lieke Hol, Herman Van Dekken, J Han van Krieken, Marjolein Van Ballegooijen, Ernst J Kuipers, and Monique E Van Leerdam.

Inter-observer variation in the histological diagnosis of polyps in colorectal cancer screening. *Histopathology*, 58(6):974–981, 2011.

[36] Allison Osmond, Hector Li-Chang, Richard Kirsch, Dimitrios Divaris, Vincent Falck, Dong Feng Liu, Celia Marginean, Ken Newell, Jeremy Parfitt, Brian Rudrick, et al. Interobserver variability in assessing dysplasia and architecture in colorectal adenomas: a multicentre canadian study. *Journal of clinical pathology*, 67(9):781–786, 2014.

[37] SA Dundas, RW Laing, A O'Cathain, I Seddon, DN Slater, TJ Stephenson, and JC Underwood. Feasibility of new prognostic classification for rectal cancer. *Journal of clinical pathology*, 41(12):1273–1276, 1988.

[38] A Davenport, J Morris, SA Pritchard, E Salmo, M Scott, and NY Haboubi. Interobserver variability amongst gastrointestinal pathologists in assessing prognostic parameters of malignant colorectal polyps: a cause for concern. *Techniques in coloproctology*, 20(9):647–652, 2016.

[39] Pia Jensen, Michael R Krogsgaard, John Christiansen, Otto Brændstrup, Aage Johansen, and Jens Olsen. Observer variability in the assessment of type and dysplasia of colorectal adenomas, analyzed using kappa statistics. *Diseases of the colon & rectum*, 38(2):195–198, 1995.

[40] K Komuta, K Batts, J Jessurun, D Snover, J Garcia-Aguilar, D Rothenberger, and R Madoff. Interobserver variability in the pathological assessment of malignant colorectal polyps. *British journal of surgery*, 91(11):1479–1484, 2004.

[41] Stanley R Hamilton, Lauri A Aaltonen, et al. Who classification of tumours. pathology and genetics of tumours of the digestive system. *Geneva: World health organization*, 2000.

[42] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[43] Matthew J Baker, Hugh J Byrne, John Chalmers, Peter Gardner, Royston Goodacre, Alex Henderson, Sergei G Kazarian, Francis L Martin, Julian Moger, Nick Stone, et al. Clinical applications of infrared and raman spectroscopy: state of play and future challenges. *Analyst*, 143(8):1735–1757, 2018.

[44] Angela WS Fung, Vijithan Sugumar, Annie He Ren, and Vathany Kulasingam. Emerging role of clinical mass spectrometry in pathology. *Journal of Clinical Pathology*, 73(2):61–69, 2020.

[45] Ina Aretz and David Meierhofer. Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *International journal of molecular sciences*, 17(5):632, 2016.

[46] Michael S. Feld, Ramasamy Manoharan, Juha Salenius, Jacobo Orenstein-Carndona, Tjeerd J. Roemer, James F. Brennan III, Ramachandra R. Dasari, and Yang Wang. Detection and characterization of human tissue lesions with near-infrared raman spectroscopy, 1995.

[47] Matthew D Keller, Elizabeth Vargis, Nara de Matos Granja, Robert H Wilson, Mary-Ann Mycek, Mark C Kelley, and Anita Mahadevan-Jansen. Development of a spatially offset raman spectroscopy probe for breast tumor surgical margin evaluation. *Journal of biomedical optics*, 16(7):077006–077006, 2011.

[48] Minbiao Ji, Daniel A Orringer, Christian W Freudiger, Shakti Ramkissoon, Xiaohui Liu, Darryl Lau, Alexandra J Golby, Isaiah Norton, Marika Hayashi, Nathalie YR Agar, et al. Rapid, label-free detection of brain tumors with stimulated raman scattering microscopy. *Science translational medicine*, 5(201):201ra119–201ra119, 2013.

[49] Michael Jermyn, Joannie Desroches, Kelly Aubertin, Karl St-Arnaud, Wendy-Julie Madore, Etienne De Montigny, Marie-Christine Guiot, Dominique Trudel, Brian C Wilson, Kevin Petrecca, et al. A review of raman spectroscopy advances with an emphasis on clinical translation challenges in oncology. *Physics in medicine and biology*, 61(23):R370, 2016.

[50] Chandrasekhara Venkata Raman. The molecular scattering of light. In *Proceedings of the Indian Academy of Sciences-Section A*, volume 37, pages 342–349. Springer, 1953.

[51] John R Ferraro. *Introductory raman spectroscopy*. Academic press, 2003.

[52] Abdullah Chandra Sekhar Talari, Zanyar Movasaghi, Shazza Rehman, and Ihtesham Ur Rehman. Raman spectroscopy of biological tissues. *Applied spectroscopy reviews*, 50(1):46–111, 2015.

[53] Richard L McCreery. *Raman spectroscopy for chemical analysis*. John Wiley & Sons, 2005.

[54] Ishan Barman, Chae-Ryon Kong, Gajendra P Singh, and Ramachandra R Dasari. Effect of photobleaching on calibration model development in biological raman spectroscopy. *Journal of biomedical optics*, 16(1):011004, 2011.

[55] Fabien Picot, François Daoust, Guillaume Sheehy, Frédérick Dallaire, Layal Chaikho, Théophile Bégin, Samuel Kadoury, and Frédéric Leblond. Data consistency and classification model transferability across biomedical raman spectroscopy systems. *Translational Biophotonics*, 3(1):e202000019, 2021.

[56] LM Uriarte, LJ Bonales, J Dubessy, A Lobato, VG Baonza, and M Cáceres. The self-absorption phenomenon in quantitative raman spectroscopy and how to correct its effects. *Microchemical Journal*, 139(134–138), 2018.

[57] Dean J. Rajkomar, A. and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[58] A Esteva, A Robicquet, B Ramsundar, V Kuleshov, M DePristo, K Chou, C Cui, G Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[59] Pranita Pradhan, Shuxia Guo, Oleg Ryabchykov, Juergen Popp, and Thomas W. Bocklitz. Deep learning a boon for biophotonics? *Journal of biophotonics*, 13(6), 2020.

[60] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11):703–715, 2019.

[61] I P Santos, E.M Barroso, T.C.B Schut, P.J Caspers, C.G van Lanschot, D.H Choi, M.F Van Der Kamp, R.W Smits, R Van Doorn, R.M. Verdijk, and V.N. Hegt. Raman spectroscopy for cancer detection and cancer surgery guidance: translation to the clinics. *Analyst*, 142(17):3025–3047, 2017.

[62] M Roberts, D Driggs, M Thorpe, Yeung M Gilbey, J., S Ursprung, A.I Aviles-Rivero, C Etmann, C McCague, L. Beer, and J.R. Weir-McCall. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[63] Gibney E. Could machine learning fuel a reproducibility crisis in science?, 2022.

[64] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science. *arXiv preprint arXiv:2207.07048*, 2022.

[65] M.J Page, J.E McKenzie, P.M Bossuyt, I Boutron, T.C Hoffmann, C.D Mulrow, L Shamseer, J.M Tetzlaff, E.A Akl, S.E. Brennan, and R. Chou. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

[66] Kelly Aubertin, Vincent Quoc Trinh, Michael Jermyn, Paul Baksic, Andrée-Anne Grosset, Joannie Desroches, Karl St-Arnaud, Mirela Birlea, Maria Claudia Vladoiu, Mathieu Latour, et al. Mesoscopic characterization of prostate

cancer using raman spectroscopy: potential for diagnostics and therapeutics. *BJU international*, 122(2):326–336, 2018.

[67] Enrico Baria, Riccardo Cicchi, Francesca Malentacchi, Irene Mancini, Pamela Pinzani, Marco Pazzagli, and Francesco S Pavone. Supervised learning methods for the recognition of melanoma cell lines through the analysis of their raman spectra. *Journal of Biophotonics*, 14(3):202000365, 2021.

[68] Danielle Bury, Guy Faust, Maria Paraskevaidi, Katherine M Ashton, Timothy P Dawson, and Francis L Martin. Phenotyping metastatic brain tumors applying spectrochemical analyses: segregation of different cancer types. *Analytical Letters*, 52(4):575–587, 2019.

[69] Fengye Chen, Chen Sun, Zengqi Yue, Yuqing Zhang, Weijie Xu, Sahar Shabbir, Long Zou, Weiguo Lu, Wei Wang, Zhenwei Xie, et al. Screening ovarian cancers with raman spectroscopy of blood plasma coupled with machine learning data processing. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 265:120355, 2022.

[70] Zhang H. Yang X. Shao X. Li T. Chen N. Chen Z. Xue W. Pan J. Chen, S. and 2021. Liu, S. Raman spectroscopy reveals abnormal changes in the urine composition of prostate cancer: An application of an intelligent diagnostic model with a deep learning algorithm. *Advanced Intelligent Systems*, 3(4):2000090, 2021.

[71] Wei Wu Cheng Chen Fangfang Chen Xiaogang Dong Mingrui Ma Ziwei Yan Xiaoyi Lv Yuhua Ma Chen, Chen and Min Zhu. Rapid diagnosis of lung cancer and glioma based on serum raman spectroscopy combined with deep learning. *Journal of Raman Spectroscopy*, 2021.

[72] Amuthachelvi Daniel, Aruna Prakasarao, and Singaravelu Ganesan. Near-infrared raman spectroscopy for estimating biochemical changes associated with different pathological conditions of cervix. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 190:409–416, 2018.

[73] QiuYao Zeng XinLiang Yan Zuyi Zhao Na Chen QianRu Deng MengHan Zhu YanJiao Zhang Fang, XiangLin and ShaoXin Li. Fast discrimination of tumor and blood cells by label-free surface-enhanced raman scattering spectra and deep learning. *Journal of Applied Physics*, 129(12), 2021.

[74] Wu X. Zhou J. Chen Y. He, C. and J. Ye. Raman optical identification of renal cell carcinoma via machine learning. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 252:119520, 2021.

[75] Hiroaki Ito, Naoyuki Uragami, Tomokazu Miyazaki, William Yang, Kenji Issha, Kai Matsuo, Satoshi Kimura, Yuji Arai, Hiromasa Tokunaga, Saiko Okada, et al. Highly accurate colorectal cancer prediction model based on raman spectroscopy using patient serum. *World Journal of Gastrointestinal Oncology*, 12(11):1311, 2020.

[76] Sharma M. Sharma L. Chao T.Y. Huang S.F. Chang L.B. Wu S.L. Jeng, M.J. and L. Chow. Raman spectroscopy analysis for optical diagnosis of oral cancer detection. *Journal of clinical medicine*, 8(9):1313, 2019.

[77] Brusatori M. Yurgelevic S. Huang C. Werner C.W. Kast R.E. Shanley J. Sherman M. Honn K.V. Maddipati K.R. Koya, S.K. and G.W. Auner. Accurate identification of breast cancer margins in microenvironments of ex-vivo basal and luminal breast cancer tissues using raman spectroscopy. *Prostaglandins and Other Lipid Mediators*, 151, 2020.

[78] Lenferink A.T. Otto C. Lee, W. and H.L Offerhaus. Classifying raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *Journal of raman spectroscopy*, 51(2):293–300, 2020.

[79] Linwei Shang Jinlan Tang Yilin Bao Juanjuan Fu Ma, Danying and Jianhua Yin. Classifying breast cancer tissue by raman spectroscopy with one-dimensional convolutional neural network. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 256, 2021.

[80] Kanika Mehta, Apurva Atak, Aditi Sahu, Sanjeeva Srivastava, et al. An early investigative serum raman spectroscopy study of meningioma. *Analyst*, 143(8):1916–1923, 2018.

[81] Yang L. Liu B. Liu L. Liu Y. Zheng Q. Liu D. Qi, Y. and J. Luo. Highly accurate diagnosis of lung adenocarcinoma and squamous cell carcinoma tissues by deep learning. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 265, 2021.

[82] Sciortino T. Secoli R. D'Amico E. Moccia S. Fernandes B. Conti Nibali M. Gay L. Rossi M. De Momi E. Riva, M. and L. Bello. Glioma biopsies classification using raman spectroscopy and machine learning models on fresh tissue samples. *Cancers*, 13(5):1073, 2021.

[83] Inês P Santos, Remco van Doorn, Peter J Caspers, Tom C Bakker Schut, Elisa M Barroso, Tamar EC Nijsten, Vincent Noordhoek Hegt, Senada Koljenović, and Gerwin J Puppels. Improving clinical diagnosis of early-stage cutaneous melanoma based on raman spectroscopy. *British journal of cancer*, 119(11):1339, 2018.

[84] Secoli R. d'Amico E. Moccia S. Conti Nibali M. Gay L. Rossi M. Pecco N. Castellano A. De Momi E. Sciortino, T. and B. Fernandes. Raman spectroscopy and machine learning for idh genotyping of unprocessed glioma biopsies. *Cancers*, 13(16):4196, 2021.

[85] Kirill A Serzhantov, Oleg O Myakinin, Mariya G Lisovskaya, Ivan A Bratchenko, Alexander A Moryatov, Sergey V Kozlov, and Valery P Zakharov. Comparison testing of machine learning algorithms separability on raman spectra of skin cancer. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 11359, page 1135906, 2020.

[86] Oh S. Hong S. Kang M. Kang D. Ji Y.G. Choi B.H. Kang K.W. Jeong H. Park Y. Shin, H. and S. Hong. Early-stage lung cancer diagnosis by deep

learning-based spectroscopic analysis of circulating exosomes. *ACS nano*, 14(5):5435–5444, 2020.

[87] Yan H. Zheng W. Lin K. James A. Selvarajan S. Lim C.M. Shu, C. and Z. Huang. Deep learning-guided fiberoptic raman spectroscopy enables real-time in vivo diagnosis and assessment of nasopharyngeal carcinoma and post-treatment efficacy during endoscopy. *Analytical chemistry*, 93(31):10898–10906, 2021.

[88] Li S. Xu Q. Yan X. Fu Q. Fu X. Fang X. Wu, X. and Y. Zhang. Rapid and accurate identification of colon cancer by raman spectroscopy coupled with convolutional neural networks. *Japanese Journal of Applied Physics*, 60(6):067001, 2021.

[89] Zhu L. Yu M. Zhang T. Zhu Z. Lou X. Sun G. Xia, J. and M. Dong. Analysis and classification of oral tongue squamous cell carcinoma based on raman spectroscopy and convolutional neural networks. *Journal of Modern Optics*, 67(6):481–489, 2020.

[90] Mingxin Yu Jiabin Xia Lianqing Zhu Tao Zhang Zhihui Zhu Yan, Hao and Guangkai Sun. Diverse region-based cnn for tongue squamous cell carcinoma classification with raman spectroscopy. *IEEE Access*, 8, 2020.

[91] Hao Yan Jiabin Xia Lianqing Zhu Tao Zhang Zhihui Zhu Xiaoping Lou Guangkai Sun Yu, Mingxin and Mingli Dong. Deep convolutional neural networks for tongue squamous cell carcinoma classification using raman spectroscopy. *Photodiagnosis and photodynamic therapy*, 26:430–435, 2019.

[92] Li C. Peng D. Yi X. He S. Liu F. Zheng X. Huang W.E. Zhao L. Zhang, L. and X. Huang. Raman spectroscopy and machine learning for the classification of breast cancers. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 264:120300, 2021.

[93] Lin K. Shu C. Zheng W. Lim C.M. Zuvela, P. and Z. Huang. Fiber-optic raman spectroscopy with nature-inspired genetic algorithms enhances real-

time in vivo detection and diagnosis of nasopharyngeal carcinoma. *Analytical chemistry*, 91(13):8101–8108, 2019.

[94] Zhu L. Yu M. Zhang T. Zhu Z. Lou X. Sun G. Xia, J. and M. Dong. Analysis and classification of oral tongue squamous cell carcinoma based on raman spectroscopy and convolutional neural networks. *Journal of Modern Optics*, 67(6):481–489, 2020.

[95] Yan H. Zheng W. Lin K. James A. Selvarajan S. Lim C.M. Shu, C. and Z. Huang. Deep learning-guided fiberoptic raman spectroscopy enables real-time in vivo diagnosis and assessment of nasopharyngeal carcinoma and post-treatment efficacy during endoscopy. *Analytical chemistry*, 93(31):10898–10906, 2021.

[96] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[97] Lenferink A.T. Otto C. Lee, W. and H.L Offerhaus. Classifying raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *Journal of raman spectroscopy*, 51(2):293–300, 2020.

[98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[99] Li S. Xu Q. Yan X. Fu Q. Fu X. Fang X. Wu, X. and Y. Zhang. Rapid and accurate identification of colon cancer by raman spectroscopy coupled with convolutional neural networks. *Japanese Journal of Applied Physics*, 60(6):067001, 2021.

[100] QiuYao Zeng XinLiang Yan Zuyi Zhao Na Chen QianRu Deng MengHan Zhu YanJiao Zhang Fang, XiangLin and ShaoXin Li. Fast discrimination of tumor and blood cells by label-free surface-enhanced raman scattering spectra and deep learning. *Journal of Applied Physics*, 129(12), 2021.

[101] Nathan Blake, Riana Gaifulina, Lewis D. Griffin, Ian M. Bell, and Geraint M. H. Thomas. Machine learning of raman spectroscopy data for classifying cancers: A review of the recent literature. *Diagnostics*, 12(6), 2022.

[102] Aaran T Lewis, Riana Gaifulina, Martin Isabelle, Jennifer Dorney, Mae L Woods, Gavin R Lloyd, Katherine Lau, Manuel Rodriguez-Justo, Catherine Kendall, Nicholas Stone, et al. Mirrored stainless steel substrate provides improved signal for raman spectroscopy of tissue and cells. *Journal of Raman Spectroscopy*, 48(1):119–125, 2017.

[103] Riana Gaifulina, Daren J Caruana, Dahmane Oukrif, Naomi J Guppy, Siân Culley, Robert Brown, Ian Bell, Manuel Rodriguez-Justo, Katherine Lau, and Geraint MH Thomas. Rapid and complete paraffin removal from human tissue sections delivers enhanced raman spectroscopic and histopathological analysis. *Analyst*, 145(4):1499–1510, 2020.

[104] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[105] Rebecca L Siegel, Lindsey A Torre, Isabelle Soerjomataram, Richard B Hayes, Freddie Bray, Thomas K Weber, and Ahmedin Jemal. Global patterns and trends in colorectal cancer incidence in young adults. *Gut*, 68(12):2179–2185, 2019.

[106] Guia Cerretelli, Ann Ager, Mark J Arends, and Ian M Frayling. Molecular pathology of lynch syndrome. *The Journal of Pathology*, 250(5):518–531, 2020.

[107] Mohsin Bilal, Mohammed Nimir, David Snead, Graham S Taylor, and Nasir Rajpoot. Role of ai and digital pathology for colorectal immuno-oncology. *British Journal of Cancer*, pages 1–9, 2022.

[108] Talha Shaikh, Elizabeth A Handorf, Joshua E Meyer, Michael J Hall, and Nestor F Esnaola. Mismatch repair deficiency testing in patients with colorectal cancer and nonadherence to testing guidelines in young adults. *JAMA oncology*, 4(2):e173580–e173580, 2018.

[109] Lindsey A Hildebrand, Colin J Pierce, Michael Dennis, Munizay Paracha, and Asaf Maoz. Artificial intelligence for histology-based detection of microsatellite instability and prediction of response to immunotherapy in colorectal cancer. *Cancers*, 13(3):391, 2021.

[110] Sapna Syngal, Edward A Fox, Charis Eng, Richard D Kolodner, and Judy E Garber. Sensitivity and specificity of clinical criteria for hereditary non-polyposis colorectal cancer associated mutations inmsh2 and mlh1. *Journal of medical genetics*, 37(9):641–645, 2000.

[111] Yanting Zhang, Ganfeng Luo, Mengjie Li, Pi Guo, Yuejiao Xiao, Huanlin Ji, and Yuantao Hao. Global patterns and trends in ovarian cancer incidence: age, period and birth cohort analysis. *BMC cancer*, 19(1):1–14, 2019.

[112] Philipp Harter, Zelal M Muallem, Christine Buhrmann, Dietmar Lorenz, Christine Kaub, Rita Hils, Stefan Kommoss, Florian Heitz, Alexander Traut, and Andreas du Bois. Impact of a structured quality management program on surgical outcome in primary advanced ovarian cancer. *Gynecologic oncology*, 121(3):615–619, 2011.

[113] Giovanni D Aletti, Eric L Eisenhauer, Antonio Santillan, Allison Axtell, Giacomo Aletti, Christine Holschneider, Dennis S Chi, Robert E Bristow, and William A Cliby. Identification of patient groups at highest risk from traditional approach to ovarian cancer treatment. *Gynecologic oncology*, 120(1):23–28, 2011.

[114] Florian Heitz, Stefan Kommoss, Roshan Tourani, Anthony Grandelis, Locke Uppendahl, Constantin Aliferis, Alexander Burges, Chen Wang, Ulrich Canzler, Jinhua Wang, et al. Dilution of molecular–pathologic gene signatures by

medically associated factors might prevent prediction of resection status after debulking surgery in patients with advanced ovarian cancer. *Clinical Cancer Research*, 26(1):213–219, 2020.

[115] Ludwig Geistlinger, Sehyun Oh, Marcel Ramos, Lucas Schiffer, Rebecca S LaRue, Christine M Henzler, Sarah A Munro, Claire Daughters, Andrew C Nelson, Boris J Winterhoff, et al. Multiomic analysis of subtype evolution and heterogeneity in high-grade serous ovarian carcinoma. *Cancer research*, 80(20):4335–4345, 2020.

[116] Antonio Gonzalez Martin, A Oza, Andrew C Embleton, Jacobus Pfisterer, J Ledermann, Eric Pujade-Lauraine, Gunnar Kristensen, M Bertrand, Philip Beale, Andrés Cervantes, et al. Exploratory outcome analyses according to stage and/or residual disease in the icon7 trial of carboplatin and paclitaxel with or without bevacizumab for newly diagnosed ovarian cancer. *Gynecologic Oncology*, 152(1), 2019.

[117] Antonio González Martín, Amit M Oza, Andrew C Embleton, Jacobus Pfisterer, Jonathan A Ledermann, Eric Pujade-Lauraine, Gunnar Kristensen, Monique A Bertrand, Philip Beale, Andrés Cervantes, et al. Exploratory outcome analyses according to stage and/or residual disease in the icon7 trial of carboplatin and paclitaxel with or without bevacizumab for newly diagnosed ovarian cancer. *Gynecologic oncology*, 152(1):53–60, 2019.

[118] Elizabeth C Smyth, Jesper Lagergren, Rebecca C Fitzgerald, Florian Lordick, Manish A Shah, Pernilla Lagergren, and David Cunningham. Oesophageal cancer. *Nature reviews Disease primers*, 3(1):1–21, 2017.

[119] Jianqi Hao, Cong Chen, Hongyu Jin, Nan Chen, Jian Zhou, Yuzhou Zhu, Kayi Chung, and Qiang Pu. The efficacy of raman spectroscopy in the diagnosis of esophageal cancer: a systematic review and meta-analysis. *Translational Cancer Research*, 9(8):4750, 2020.

[120] Durmus Ozdemir, Matt Mosley, and Ron Williams. Effect of wavelength drift on single-and multi-instrument calibration using genetic regression. *Applied spectroscopy*, 52(9):1203–1209, 1998.

[121] Shuxia Guo, Claudia Beleites, Ute Neugebauer, Sara Abalde-Cela, Nils Kristian Afseth, Fatima Alsamad, Suresh Anand, Cuauhtemoc Araujo-Andrade, Sonja Askrabic, Ertug Avci, et al. Comparability of raman spectroscopic configurations: a large scale cross-laboratory study. *Analytical Chemistry*, 92(24):15745–15756, 2020.

[122] M Isabelle, J Dorney, A Lewis, GR Lloyd, O Old, N Shepherd, M Rodriguez-Justo, H Barr, K Lau, I Bell, et al. Multi-centre raman spectral mapping of oesophageal cancer tissues: a study to assess system transferability. *Faraday discussions*, 187(87–103), 2016.

[123] Mariana Bento, Irene Fantini, Justin Park, Leticia Rittner, and Richard Frayne. Deep learning in large and multi-site structural brain mr imaging datasets. *Frontiers in Neuroinformatics*, 15, 2021.

[124] Ling-Li Zeng, Huaning Wang, Panpan Hu, Bo Yang, Weidan Pu, Hui Shen, Xingui Chen, Zhening Liu, Hong Yin, Qingrong Tan, et al. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity mri. *EBioMedicine*, 30(74–85), 2018.

[125] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[126] Shuxia Guo, Jürgen Popp, and Thomas Bocklitz. Chemometric analysis in raman spectroscopy from experimental design to machine learning–based modeling. *Nature protocols*, 16(12):5426–5459, 2021.

[127] Courtland R. Bias detectives: the researchers striving to make algorithms fair. *Nature*, 558(7710):357–360, 2018.

[128] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[129] Roelofs R. Schmidt L. Recht, B. and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[130] Jussi Tohka and Mark Van Gils. Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Computers in Biology and Medicine*, 132:104324, 2021.

[131] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[132] Karel GM Moons and Frank E Harrell. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic radiology*, 10(6):670–672, 2003.

[133] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[134] Moving towards reproducible machine learning. *Nature Computational Science*, 1(10):629–630, 2021.

[135] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586):eabb1655, 2021.

[136] Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pages 381–405. PMLR, 2019.

[137] Elizabeth Vargis, Elizabeth M Kanter, Shovan K Majumder, Matthew D Keller, Richard B Beaven, Gautam G Rao, and Anita Mahadevan-Jansen. Effect of normal variations on disease classification of raman spectra from cervical tissue. *Analyst*, 136(14):2981–2987, 2011.

[138] Claude Nadeau and Yoshu Bengio. Inference for the generalization error. *Advances in neural information processing systems*, 12, 1999.

[139] Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 3–12. Springer, 2004.

[140] Leo Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001.

[141] Moving towards reproducible machine learning. *Nature Computational Science*, 1(10):629–630, 2021.

[142] Benjamin J Heil, Michael M Hoffman, Florian Markowetz, Su-In Lee, Casey S Greene, and Stephanie C Hicks. Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 18(10):1132–1135, 2021.

[143] Kristen M Fedak, Autumn Bernal, Zachary A Capshaw, and Sherilyn Gross. Applying the bradford hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging themes in epidemiology*, 12(1):1–9, 2015.

[144] Félix Lussier, Vincent Thibault, Benjamin Charron, Gregory Q Wallace, and Jean-Francois Masson. Deep learning and artificial intelligence methods for

raman and surface-enhanced raman scattering. *TrAC Trends in Analytical Chemistry*, 124:115796, 2020.

[145] Tao Li, Shenghuo Zhu, and Mitsunori Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and information systems*, 10(4):453–472, 2006.

[146] Jigang Xie and Zhengding Qiu. The effect of imbalanced data sets on lda: A theoretical and empirical analysis. *Pattern recognition*, 40(2):557–562, 2007.

[147] Jacopo Acquarelli, Twan van Laarhoven, Jan Gerretzen, Thanh N Tran, Lutgarde MC Buydens, and Elena Marchiori. Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica chimica acta*, 954:22–31, 2017.

[148] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[149] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[150] David Picard. Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

[151] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. pages 1026–1034, 2015.

[152] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang,

Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[153] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[154] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[155] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[156] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105, 2004.

[157] Claudia Beleites, Richard Baumgartner, Christopher Bowman, Ray Somorjai, Gerald Steiner, Reiner Salzer, and Michael G Sowa. Variance reduction in estimating classification error using sparse datasets. *Chemometrics and intelligent laboratory systems*, 79(1-2):91–100, 2005.

[158] Piotr S Gromski, Elon Correa, Andrew A Vaughan, David C Wedge, Michael L Turner, and Royston Goodacre. A comparison of different chemometrics approaches for the robust classification of electronic nose data. *Analytical and bioanalytical chemistry*, 406:7581–7590, 2014.

[159] Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

[160] George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *Acm Sigkdd Explorations Newsletter*, 12(1):49–57, 2010.

[161] Christopher D Brown and Peter D Wentzell. Hazards of digital smoothing filters as a preprocessing tool in multivariate calibration. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 13(2):133–152, 1999.

[162] Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy. Review of multidimensional data processing approaches for raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2:1–38, 2015.

[163] Dong Wei, Shuo Chen, and Quan Liu. Review of fluorescence suppression techniques in raman spectroscopy. *Applied Spectroscopy Reviews*, 50(5):387–406, 2015.

[164] Thomas Bocklitz, Angela Walter, Katharina Hartmann, Petra Rösch, and Jürgen Popp. How to pre-process raman spectra for reliable and stable models? *Analytica chimica acta*, 704(1-2):47–56, 2011.

[165] Kristian Hovde Liland, Trygve Almøy, and Bjørn-Helge Mevik. Optimal choice of baseline correction for multivariate calibration of spectra. *Applied spectroscopy*, 64(9):1007–1016, 2010.

[166] Philip Heraud, Bayden R Wood, John Beardall, and Don McNaughton. Effects of pre-processing of raman spectra on in vivo classification of nutrient status of microalgal cells. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 20(5):193–197, 2006.

[167] Jane S Greaves, Anita MS Richards, William Bains, Paul B Rimmer, Hideo Sagawa, David L Clements, Sara Seager, Janusz J Petkowski, Clara Sousa-Silva, Sukrit Ranjan, et al. Phosphine gas in the cloud decks of venus. *Nature Astronomy*, 5(7):655–664, 2021.

[168] IAG Snellen, L Guzman-Ramirez, MR Hogerheijde, APS Hygate, and FFS Van der Tak. Re-analysis of the 267 ghz alma observations of venus-no statistically significant detection of phosphine. *Astronomy & Astrophysics*, 644:L2, 2020.

[169] Riana Gaifulina, Abigail DG Nunn, Edward RC Draper, Robin K Strachan, Nathan Blake, Steven Firth, Geraint MH Thomas, Paul F McMillan, and Jayesh Dudhia. Intra-operative raman spectroscopy and ex vivo raman mapping for assessment of cartilage degradation. *Clinical Spectroscopy*, page 100012, 2021.

[170] Esben Jannik Bjerrum, Mads Glahder, and Thomas Skov. Data augmentation of spectral data for convolutional neural network (cnn) based deep chemometrics. *arXiv preprint arXiv:1710.01927*, 2017.

[171] Chad A Lieber and Anita Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Applied spectroscopy*, 57(11):1363–1367, 2003.

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

[173] Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010.

[174] Nils Kristian Afseth and Achim Kohler. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 117(92–99), 2012.

[175] Shuxia Guo, Thomas Bocklitz, Ute Neugebauer, and Jürgen Popp. Common mistakes in cross-validating classification models. *Analytical Methods*, 9(30):4410–4417, 2017.

[176] Matloob Khushi, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A Hameed, Shahadat Uddin, Suhuai Luo, Xiaoyan Yang, and Maranatha Consuelo Reyes. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9:109960–109975, 2021.

[177] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[178] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

[179] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[180] Ronaldo C Prati, Gustavo EAPA Batista, and Maria Carolina Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. In *Mexican international conference on artificial intelligence*, pages 312–321. Springer, 2004.

[181] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[182] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In Springer, editor, *International conference on intelligent computing*, pages 878–887, 2005.

[183] Jiayi Tang, Alex Henderson, and Peter Gardner. Exploring adaboost and random forests machine learning approaches for infrared pathology on unbalanced data sets. *Analyst*, 146(19):5880–5891, 2021.

[184] Kaiyo Takubo, Michael Vieth, Junko Aida, Motoji Sawabe, Yoichi Kumagai, Yoshio Hoshihara, and Tomio Arai. Differences in the definitions used for

esophageal and gastric diseases in different countries. *Digestion*, 80(4):248–257, 2009.

[185] Date Maitra, Ishaan Ravindra Sudhachandra and Francis Luke Martin. Towards screening barrett's oesophagus: current guidelines, imaging modalities and future developments. *Clinical Journal of Gastroenterology*, 13(5):635–649, 2020.

[186] Marjon Kerkhof, Herman Van Dekken, EW Steyerberg, GA Meijer, AH Mulder, Adriaan De Bruïne, Ann Driessen, FJ Ten Kate, JG Kusters, EJ Kuipers, et al. Grading of dysplasia in barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology*, 50(7):920–927, 2007.

[187] Elizabeth Montgomery, Mary P Bronner, John R Goldblum, Joel K Greenson, Marian M Haber, John Hart, Laura W Lamps, Gregory Y Lauwers, Audrey J Lazenby, David N Lewin, et al. Reproducibility of the diagnosis of dysplasia in barrett esophagus: a reaffirmation. *Human pathology*, 32(4):368–378, 2001.

[188] Julian A Abrams and Michael Quante. *Sleisenger and Fordtran's Gastrointestinal and Liver Disease*, volume 11, chapter Adenocarcinoma of the Stomach and Other Gastric Tumors, pages 901–920. Elsevier, 2020.

[189] Lu Zhang, Binyu Sun, Xi Zhou, QiongQiong Wei, Sicheng Liang, Gang Luo, Tao Li, and Muhan Lü. Barrett's esophagus and intestinal metaplasia. *Frontiers in Oncology*, 11:2325, 2021.

[190] Min H. Vandenberg N. Dowling J. Holloway L. Chlap, P. and A. Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 2021.

[191] Wei Wu Cheng Chen Fangfang Chen Xiaogang Dong Mingrui Ma Ziwei Yan Xiaoyi Lv Yuhua Ma Chen, Chen and Min Zhu. Rapid diagnosis of lung cancer and glioma based on serum raman spectroscopy combined with deep learning. *Journal of Raman Spectroscopy*, 2021.

[192] Linwei Shang Jinlan Tang Yilin Bao Juanjuan Fu Ma, Danying and Jianhua Yin. Classifying breast cancer tissue by raman spectroscopy with one-dimensional convolutional neural network. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 256, 2021.

[193] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[194] Man Wu, Shuwen Wang, Shirui Pan, Andrew C Terentis, John Strasswimmer, and Xingquan Zhu. Deep learning data augmentation for raman spectroscopy cancer tissue classification. *Research Square preprint*, 2021.

[195] Stefano Di Frischia, Paolo Giammatteo, Federico Angelini, Valeria Spizzichino, Elena De Santis, and Luigi Pomante. Enhanced data augmentation using gans for raman spectra classification. pages 2891–2898. IEEE, 2020.

[196] Fábio Perez, Cristina Vasconcelos, Sandra Avila, and Eduardo Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018.

[197] Spectrometer calibration protocol for Raman spectra recorded with different excitation wavelengths. Bocklitz, tw and dörfer, t and heinke, r and schmitt, m and popp, j. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 149:544–549, 2015.

[198] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):1–8, 2006.

[199] Jacques Wainer and Gavin Cawley. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 182:115222, 2021.

[200] Zeyi Wen, Jiashuai Shi, Qinbin Li, Bingsheng He, and Jian Chen. ThunderSVM: A fast SVM library on GPUs and CPUs. *Journal of Machine Learning Research*, 19:797–801, 2018.

[201] Eric Chun Yong Chan, Poh Koon Koh, Mainak Mal, Peh Yean Cheah, Kong Weng Eu, Alexandra Backshall, Rachel Cavill, Jeremy K Nicholson, and Hector C Keun. Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (hr-mas nmr) spectroscopy and gas chromatography mass spectrometry (gc/ms). *Journal of proteome research*, 8(1):352–361, 2009.

[202] Matthew G Vander Heiden, Lewis C Cantley, and Craig B Thompson. Understanding the warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930):1029–1033, 2009.

[203] Long Chen, Yue Wang, Nenrong Liu, Duo Lin, Cuncheng Weng, Jixue Zhang, Lihuan Zhu, Weisheng Chen, Rong Chen, and Shangyuan Feng. Near-infrared confocal micro-raman spectroscopy combined with pca–lda multivariate analysis for detection of esophageal cancer. *Laser Physics*, 23(6):065601, 2013.

[204] Mika Ishigaki, Yasuhiro Maeda, Akinori Taketani, Bibin B Andriana, Ryu Ishihara, Kanet Wongravee, Yukihiro Ozaki, and Hidetoshi Sato. Diagnosis of early-stage esophageal cancer by raman spectroscopy and chemometric techniques. *Analyst*, 141(3):1027–1033, 2016.

[205] Wenhua Huang, Qixin Shang, Xin Xiao, Hanlu Zhang, Yimin Gu, Lin Yang, Guidong Shi, Yushang Yang, Yang Hu, Yong Yuan, et al. Raman spectroscopy and machine learning for the classification of esophageal squamous carcinoma. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 281:121654, 2022.

[206] Yamin Lin, Siqi Gao, Qiwen Wang, Mengmeng Zheng, Shuzhen Tang, Yun Yu, and Juqiang Lin. Detection of cancerous esophageal tissue by raman

spectroscopy and multivariate analysis of extracellular fluid. In *Optics in Health Care and Biomedical Optics X*, volume 11553, pages 176–182. SPIE, 2020.

[207] Naofumi Tomita, Behnaz Abdollahi, Jason Wei, Bing Ren, Arief Suriawinata, and Saeed Hassanpour. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA network open*, 2(11):e1914645–e1914645, 2019.

[208] Johannes Rieke, Fabian Eitel, Martin Weygandt, John-Dylan Haynes, and Kerstin Ritter. Visualizing convolutional networks for mri-based diagnosis of alzheimer's disease. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 24–31. Springer, 2018.

[209] C Richard Boland, Dong K Chang, John M Carethers, et al. The biochemical basis of microsatellite instability and abnormal immunohistochemistry and clinical behavior in lynch syndrome: from bench to bedside. *Familial cancer*, 7(1):41–52, 2008.

[210] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[211] Wartini Ng, Budiman Minasny, Wanderson de Sousa Mendes, and José Alexandre Melo Demattê. The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. *Soil*, 6(2):565–578, 2020.

[212] Nirav Thosani, Harvinder Singh, Asha Kapadia, Nobuo Ochi, Jeffrey H Lee, Jaffer Ajani, Stephen G Swisher, Wayne L Hofstetter, Sushovan Guha, and Manoop S Bhutani. Diagnostic accuracy of eus in differentiating mucosal versus submucosal invasion of superficial esophageal cancers: a systematic review and meta-analysis. *Gastrointestinal endoscopy*, 75(2):242–253, 2012.

[213] Shakil Ahmed, Asadullah Shaikh, Hani Alshahrani, Abdullah Alghamdi, Mesfer Alrizq, Junaid Baber, and Maheen Bakhtyar. Transfer learning

approach for classification of histopathology whole slide images. *Sensors*, 21(16):5361, 2021.

[214] Chi-Sing Ho, Neal Jean, Catherine A Hogan, Lena Blackmon, Stefanie S Jeffrey, Mark Holodniy, Niaz Banaei, Amr AE Saleh, Stefano Ermon, and Jennifer Dionne. Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning. *Nature communications*, 10(1):1–8, 2019.

[215] Rui Zhang, Huimin Xie, Shuning Cai, Yong Hu, Guo-kun Liu, Wenjing Hong, and Zhong-qun Tian. Transfer-learning-based raman spectra identification. *Journal of Raman Spectroscopy*, 51(1):176–186, 2020.

[216] Ugljesa Djuric, Gelareh Zadeh, Kenneth Aldape, and Phedias Diamandis. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *NPJ precision oncology*, 1(1):1–5, 2017.