

# Developing **Infrared** Spectroscopic Detection for Stratifying Brain Tumour Patients: Glioblastoma Multiforme vs. Lymphoma

James M. Cameron<sup>1</sup>, Holly J. Butler<sup>2</sup>, Benjamin R. Smith<sup>2</sup>, Mark G. Hegarty<sup>2</sup>, Michael D. Jenkinson<sup>3</sup>,  
Khaja Syed<sup>4</sup>, Paul M. Brennan<sup>5</sup>, Katherine Ashton<sup>6</sup>, Timothy Dawson<sup>6</sup>, David S. Palmer<sup>2,7</sup>, Matthew  
J. Baker<sup>1,2</sup>

<sup>1</sup>WestCHEM, Department of Pure and Applied Chemistry, Technology and Innovation Centre, University of Strathclyde, 99 George St, Glasgow, G1 1RD, UK

<sup>2</sup>ClinSpec Diagnostics, University of Strathclyde, Technology and Innovation Centre, 99 George Street, Glasgow, G1 1RD, UK

<sup>3</sup>Institute of Translational Medicine, University of Liverpool & The Walton Centre NHS Foundation Trust, Lower Lane, Fazakerley, Liverpool, L9 7LJ, UK

<sup>4</sup>Walton Research Tissue Bank, Neurosciences Labs, The Walton Centre NHS Foundation Trust, Lower lane, Fazakerley, Liverpool, L9 7LJ, UK

<sup>5</sup>Translational Neurosurgery, Department of Clinical Neurosciences, Western General Hospital, Edinburgh, EH4 2XU, UK

<sup>6</sup>Neuropathology, Lancashire Teaching Hospitals NHS Trust, Royal Preston Hospital, Sharoe Green Lane North, Preston, Lancashire, PR2 9HT, UK

<sup>7</sup>WestCHEM, Department of Pure and Applied Chemistry, Thomas Graham Building, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, UK

\*Corresponding Author:

Email: [matthew.baker@strath.ac.uk](mailto:matthew.baker@strath.ac.uk); [matthew.baker@clinspecdx.com](mailto:matthew.baker@clinspecdx.com)

Twitter: @ChemistryBaker

## **Abstract**

Over a third of brain tumour patients visit their general practitioner more than five times prior to diagnosis in the UK, leading to 62% of patients being diagnosed as emergency presentations.

Unfortunately, symptoms are non-specific to brain tumours, and the majority of these patients complain of headaches on multiple occasions before being referred to a neurologist. As there are

currently no methods in place for the early detection of brain cancer, the affected patients' average life expectancy is reduced by 20 years. These statistics indicate that the current pathway is ineffective, and there is a vast need for a rapid diagnostic test.

Attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy is sensitive to the hallmarks of cancer, as it analyses the full range of macromolecular classes. The combination of serum spectroscopy and advanced data analysis has previously been shown to rapidly and objectively distinguish brain tumour severity. Recently, a novel high-throughput ATR accessory has been developed, which could be cost-effective to the National Health Service **in the UK**, and valuable for clinical translation.

**In this study, 765 blood serum samples have been collected from healthy controls and patients diagnosed with various types of brain cancer**, contributing to one of the largest spectroscopic studies to date. Three robust machine learning techniques - **random forest, partial least squares-discriminant analysis and support vector machine - have all provided promising results**. The novel high-throughput technology has been validated by separating brain cancer and non-cancer with balanced accuracies of 90% **which is comparable to the traditional fixed diamond crystal methodology**.

Furthermore, the differentiation of brain tumour type could be useful for neurologists, as some are difficult to distinguish through medical imaging alone. For example, the highly aggressive glioblastoma multiforme and primary cerebral lymphoma can appear similar on magnetic resonance imaging (MRI) scans, thus are often misdiagnosed. Here, we report the ability of infrared spectroscopy to distinguish between glioblastoma and lymphoma patients, at a sensitivity and specificity of 90.1% and 86.3%, respectively. A reliable serum diagnostic test could avoid the need for surgery and speed up time to definitive chemotherapy and radiotherapy.

## **Introduction**

Brain tumour incidence rates have been increasing since the early 1990s, rising by 34% in the UK alone [1]. Despite an improvement in patient survival, only 14% of patients survive 10 years or more after diagnosis, and the average reduction in life expectancy of 20 years is the highest of all cancers [2]. Rapid and timely diagnosis and determination of tumour type is crucial to expediting management and improving patient outcomes [3].

The symptoms most frequently associated with brain tumour are non-specific, such as headache, presenting a challenge for doctors in identifying which patients with these common symptoms are most likely to have a brain tumour, and should have expedited brain imaging [4]. Consequently patients often visit their general practitioner (GP) multiple times before diagnosis and for nearly two thirds of patients diagnosis is in the emergency department once they deteriorate [5,6]. Existing referral guidelines lack sensitivity and specificity, with as few as 1.6% of patients referred for urgent brain imaging from primary care having a brain tumour, suggesting many brain scans are unnecessary [7].

Brain tumours are diagnosed on magnetic resonance imaging (MRI) or computed tomography (CT) brain imaging. There are many different types of tumours, depending on the underlying cell of origin, each with its own optimal treatment regimens. Crucially, it is not possible to identify the tumour type with certainty from imaging alone. For example, primary cerebral lymphoma and glioblastoma (GBM) can have similar appearance on MRI [8], but very different therapy options. Patients therefore require a gold standard histological tissue diagnosis. This subjects them to surgical tumour biopsy, with attendant risks, including stroke or death, and with consequent delay to commencement of the chemotherapy and/or radiotherapy that will best impact on their disease. A rapid blood test that can identify patients with tumours amongst those with similar symptoms, and that can stratify tumour type would have a profound impact.

Analytical techniques based on vibrational spectroscopy, **such as Raman and infrared (IR) spectroscopy**, have emerged in the field of disease diagnostics [9–12]. Fourier-transform infrared

spectroscopy (FTIR) in particular has become increasingly popular in medical research, because of its rapid, non-invasive analysis [13]. In FTIR spectroscopy, biological samples are irradiated with infrared light. The absorbance of this light causes molecular **excitation and enables transitions between vibrational states**, resulting in an IR spectrum. **A typical spectrum of a biological sample** represents a biochemical fingerprint, and can characterise and quantify the levels of proteins, lipids, carbohydrates and nucleic acids that are present. The imbalances in these biomolecular components can give an indication of disease states [14]. Machine learning algorithms learn the differences in IR biosignatures that are exclusive to disease and can provide a diagnostic output with a prediction on the patient's state [15].

Many spectroscopic disease diagnostic pilot studies to date have analysed human tissue, indicating the possibility to differentiate healthy and cancerous tissue, as well as benign and malignant tumours [16]. Breast, lung, colorectal and prostate lesions have been studied, providing a platform of promising results [17–21]. Blood serum contains over 20,000 different proteins and is one of the most complex biofluids [22]. Serum perfuses all body organs, gaining proteomes from surrounding tissues and cells, making it rich in information, hence the spectroscopic biosignature of serum ideal for indicating disease states [23,24].

Attenuated total reflection (ATR)-FTIR spectroscopy is well suited to the analysis of biofluids, as only tiny volumes are required and sample preparation is minimal [25,26]. Backhaus *et al.* successfully differentiated breast cancer and healthy controls in a study using blood serum [27]. Ollesch *et al.* introduced automated sampling for the first time, robotically spotting serum for high throughput FTIR measurements, in their quest to identify and validate spectroscopic biomarker candidates for urinary bladder cancer [28]. Ovarian cancer can be detected from both serum and plasma samples, with accuracies of ~95% and ~97% respectively [29]. The serum biosignature for cirrhotic patients, with and without hepatocellular carcinoma (HCC), could also be differentiated effectively [30]. In brain tumours, Hands *et al.* were the first to use serum for ATR-FTIR spectroscopic analysis. Comparisons have been made between glioma and non-cancer patients [31], as

well as different brain tumour types, effectively predicting tumour grade, through spectroscopic separation of low grade gliomas from glioblastoma (high grade) [32]. In a further study of blood serum from 433 patients [33], a range of primary (glioma, meningioma) and secondary (metastatic) lesions were analysed, with sensitivity and specificity values for discrimination of cancer versus non-cancer of 92.8% and 91.5% respectively [15].

Traditional ATR-FTIR instrumentation was used in these studies, which has barred the clinical translation of the technique for a number of reasons. Conventionally, an ATR-FTIR spectrometer has a fixed point of analysis, known as the internal reflection element (IRE). IREs for ATR analysis are made from materials with high refractive indices, the most common being diamond, zinc selenide or germanium [34], to contrast with the sample that has a lower refractive index. Biofluid samples are deposited directly onto the surface of the IRE before being air dried, in order to combat the spectral interference of water [35]. IR light is directed into the IRE and internally reflected, forming an evanescent wave at the IRE-sample interface. This evanescent wave interrogates the sample at a defined penetration depth, which is dependent upon the refractive indices of the IRE and sample, the angle of incidence and the wavelength of IR beam [36].

The traditional approach is limited in both cost and time. The IRE materials tend to be high cost, therefore would be expensive to replace. The fixed IRE needs to be cleaned between each sample, which is extremely time consuming. It takes approximately 8 minutes to adequately dry 1  $\mu$ L of human serum on to a diamond crystal, and with the necessary cleaning steps, as well as the technical and biological repeats, it would take over an hour to process one patient [32]. Also, scratches on the surface of fixed IREs are known to affect the sample-IRE contact, which is essential for ATR-FTIR measurements [37]. These limitations have inhibited the progression of the technique thus far, however high-throughput ATR-FTIR could overcome these barriers for successful clinical translation. A recently published health economic study has suggested a high-throughput alternative to the traditional IRE would be cost-effective to the [UK's National Health Service](#) (NHS). Gray *et al.* highlight the clinical and economic benefits of implementing a quick diagnostic test for brain cancers

into the current pathway [38]. They reported that a serum blood test at the primary care level could prioritise patients for neuroimaging, improving patient survival and quality of life, whilst also saving on the cost of unnecessary brain scans. Furthermore, since blood tests at the primary care level are already in place, an additional test at this stage would not significantly disrupt current practices.

Silicon (Si) has a high refractive index, and its relatively low cost - cf. diamond - and transparency to infrared light makes it an ideal material for Si IREs (SIREs) [39,40]. High-throughput disposable SIREs are now commercially available, that allow single-bounce ATR-FTIR (ClinSpec Diagnostics Ltd, UK) [41]. The SIRE replaces the expensive fixed crystal, and allow multiple sampling points. The design enables the slides to be batch processed, as well as having the option of repeating analysis if required, a feature that would not be possible with conventional fixed SIREs.

In this study, we further explore the largest retrospective dataset curated to date of serum samples from patients with brain tumours, with a specific focus on the spectroscopic interpretation of the variances within the brain tumour cohort. Specifically, we elucidate the ability of **SIRE-based ATR-FTIR spectroscopy to successfully identify the cancerous biosignature in serum**, and to differentiate between glioblastoma and primary cerebral lymphoma.

## Materials and Methods

### Sample collection and preparation

**Following a specified standard operating procedure**, a total of 765 serum samples were obtained from three sources; the Walton Centre NHS Trust (Liverpool, UK), Royal Preston Hospital (Preston, UK), and the commercial source Tissue Solutions Ltd (Glasgow, UK). Ethical approval for this study was obtained (Walton Research Bank and Brain Tumour North West/WRTB 13\_01/BTNW Application #1108). **All experiments were performed in accordance with the University of Strathclyde and NHS Lothian ethical guidelines and approved by ethics committees at both institutions, Informed consents were obtained from human participants of this study. The primary care triage study contains 724 cases**

- 487 brain tumour samples and 237 healthy controls. A respectable balance of male and female patients has been included, with a widespread age range (Table S1). A large variety of tumour types are involved in the brain cancer cohort (Table S2). An additional 41 serum samples were collected from patients with primary cerebral lymphoma for comparison against GBM samples (Table S3).

In order to be included in this study, the cancer patients must have had a pathologically confirmed primary or secondary brain tumour, and must not have been undergoing chemo- or radio-therapy at the time of collection. For control patients, obtained as above, inclusion criteria stated that they should not be undergoing any medical treatments, nor have any history of cancer. Blood samples were collected in serum collection tubes and allowed to clot for up to one hour. The tubes were centrifuged at 2200 g for 15 minutes at room temperature, then the separated serum component was subsequently aliquoted stored in an -80°C freezer.

Prior to spectral analysis, the frozen serum samples were removed from storage and thawed at room temperature (18-25°C) for an average time of 15-20 minutes. Using a micropipette, 3µL of serum from one individual patient was deposited onto each of the three sample wells of the optical sample slide (wells 1, 2 and 3), whilst ensuring well '0' remained clean for background collection. The serum drops were spread across the well using the pipette tip, in order to create a thin serum film and cover the whole IRE for more uniform deposition. Prepared slides were stacked in 3D printed polylactic acid (PLA) slide holders, which were designed to enable batch drying. The stacked slides were then stored in a drying unit incubator (Thermo Fisher™ Heratherm™, GE) at 35°C for 1 hour. Pre-analytical work prior to beginning this study showed this step to be vital, as it provides even heat and airflow for controlled drying dynamics of the serum droplet, to obtain a smooth, flat homogenous sampling surface [42–44].

Spectral collection

For this study, a Perkin Elmer Spectrum 2 FTIR spectrometer (Perkin Elmer, UK) was used for the spectral collection. A Specac Quest ATR accessory unit was fitted with a specular reflectance puck (Specac Ltd, UK), allowing the SIRE to sit on top of the aperture and replace the traditional fixed diamond IRE. The Slide Indexing Unit (ClinSpec Diagnostics Ltd, UK) enabled accurate and reproducible movement across the specular reflectance puck, indexing the optical slide between sample wells. With the first well acting as a background, the three sample wells provide the biological repeats. Each well was analysed in triplicate - resulting in nine spectra per patient. The spectra were acquired in the range  $4000\text{-}450\text{cm}^{-1}$ , at a resolution of  $4\text{cm}^{-1}$ , with  $1\text{cm}^{-1}$  data spacing and 16 co-added scans. In total 6885 spectra have been collected from all serum samples.

### Spectral pre-processing

Here we have used the PRFFECT toolbox within R Statistical Computing Environment software for the spectroscopic analysis [45,46], which can be divided into two parts; spectral pre-processing and spectral classification. The pre-processing step is commonly applied in spectroscopic studies, as it reduces unwanted variance in the dataset. A combination of baseline correction, normalisation and data reduction enables the significant biological information be emphasised and improves the classification performance [47]. The optimum pre-processing protocol was determined using a trial-and-error iterative approach. The PRFFECT toolbox offers various pre-processing methods, such as binning, smoothing, normalisation and numerical derivatives - we direct the reader towards Smith *et al.* [45] for more information on the use of this open-source program. Figure 1 gives an example of the data pre-processing; (a) **raw spectra as the mean plot per patient, for the whole 724-patient dataset**, and (b) shows the spectra cut to the fingerprint region, with baseline correction and a vector normalisation applied - greatly reducing the spectral variation.

Extended multiplicative signal correction (EMSC) has recently been shown to be a reliable pre-processing tool, that allows more selective correction for various types of scattering [48]. The EMSC process scales the IR spectra according to a given reference spectrum. In this case, the reference was

an average spectrum of 10 background measurements of the SIRE, which was chosen to minimise the spectral variance caused by silicon lattice vibrations. The optimal pre-processing parameters were found to be (in order); EMSC, spectral cut to the fingerprint region (1800-1000  $\text{cm}^{-1}$ ), a minmax normalisation and a binning factor of 8.

## Spectral analysis

Spectral analysis was carried out to identify the cancerous biosignature from a known patient cohort, to develop a trained classification model, and then to use this information to predict the presence of cancer in an unknown population. **Prior to running the classifications, bootstrapping analysis on the training set was carried out to search for an acceptable number of iterations. This technique resamples a dataset with replacement, to determine an optimal resample value which will maximise classification accuracy [49].** To develop the models, patients were randomly split into training and test sets, with a 70:30 split. Models were tuned on the training set (70%) and then used to make predictions for the spectra in the test set (30%). Model tuning was carried out by a grid-search of model hyper-parameters to maximise Cohen's Kappa statistic computed on a per-spectra basis for 5-fold cross-validation on the training set. In order to ensure that the models were trained and validated correctly, spectra from a single patient's sample could only appear in one cross-validation fold, and in either the training or test set. The consensus vote amongst the nine spectra that were analysed for each patient was reported as the diagnostic outcome (cancer or non-cancer). Model performance is reported in terms of sensitivity, specificity, kappa, and balanced accuracy. For those classifications for which the class prevalence in the clinical population were known, positive and negative predictive values were also computed.

Sensitivities and specificities (Eq. 1 and 2), are based on the number of correct and incorrect predictions in the external test set. The sensitivity refers to the ability of the test to correctly identify the patients with the disease (brain cancer), and specificity is the ability to correctly pick out those without the disease (controls) [50]. True positives (TP) result from a patient with the target disease

with five or more spectra out of their nine spectra correctly identified, whereas true negatives (TN) refer to the patients without the target disease who have at least five out of their nine spectra correctly identified. False positives (FP) are where a control patient has five or more spectra incorrectly identified as cancer, and a false negative (FN) is from a patient with the target disease who has five or more spectra incorrectly classified as non-cancer.

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{TP}{P} \quad (1)$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{TN}{N} \quad (2)$$

where P is the number of real positives and N is the number of real negatives.

When employing binary classifications on imbalanced datasets, the overall model performance is commonly measured using balanced accuracy (Eq. 3), which can be defined as the average accuracy obtained on either class [51].

$$\text{Balanced Accuracy} = \left( \frac{TP}{P} + \frac{TN}{N} \right) / 2 \quad (3)$$

In order to understand the reliability of the diagnostic model the Kappa value,  $\kappa$ , can give a quantitative measure of the magnitude of agreement between observers (Eq. 4).

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

where  $p_o$  is the relative observed agreement and  $p_e$  is hypothetical probability of the chance agreement, which can be calculated from consideration of the number of times that cancer and non-cancers occur in the real and predicted data. Values of  $\kappa$  range from below zero to one and equate to the level of agreement. Where in general,  $\kappa \leq 0$  indicates no agreement, 0.01–0.20 accounts for slight, 0.21–0.40 fair, moderate agreement is 0.41–0.60, 0.61–0.80 is substantial and lastly 0.8–1.00 is almost perfect agreement [52,53].

Predictive values are useful to clinicians as they indicate the true likelihood of the test results. The positive predictive value (PPV) is the proportion of patients with positive test results who are correctly diagnosed, and the negative predictive value (NPV) relates to those with correctly assigned negative results [54,55]. The predictive values are dependent upon the sensitivity, specificity and the prevalence of the disease (Eq. 5 and 6) - in this case the prevalence of brain tumours [56]. In this study the PPV and NPV have been calculated using the mean values of sensitivity and specificity from each of the resampled classification models and the prevalence is equal to that of the positive class (i.e. cancer) in the clinical environment.

$$PPV = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1-specificity) \times (1-prevalence)} \quad (5)$$

$$NPV = \frac{specificity \times (1-prevalence)}{(1-sensitivity) \times prevalence + specificity \times (1-prevalence)} \quad (6)$$

To determine the optimum values for the tuning parameters, a 5-fold cross-validation was performed - on a patient basis - on the training data. Due to the class imbalance present when examining the difference between cancer (487 patients) vs. non-cancer (237 patients), various sampling methods were used throughout this study to ensure no bias was present within the models; up-sampling, down-sampling and synthetic minority over-sampling technique (SMOTE). The up-sampling method consists of repeatedly sampling the minority class with replacement to increase the number of samples, whereas down-sampling selects a subset of the majority class at random, removing the extra samples to make it the same size as the minority class [57]. SMOTE is unique in that it artificially mixes the data to, creating 'new' samples to achieve a more balanced dataset [58].

## Random forest

Random forest (RF) is a robust machine learning technique that for classification problems builds an ensemble of decision trees from the training data using the Classification and Regression Trees (CART) algorithm [59]. There are three main tuning parameters employed in this technique; *n<sub>tree</sub>* is the number of trees, *m<sub>try</sub>* is the number of variables available for splitting at each tree node and

*nodesize* refers to the depth of the trees. To encourage diversity amongst the forest, each of *n*tree trees is built on a bootstrap sample of the training data, and at each node in each tree the optimum feature is selected from a random subset of *m*try available features. Each tree is grown until the terminal nodes contain no fewer than *nodesize* observations. Classification predictions are reported as the majority vote of all of the decision trees in the forest, which allows the method to benefit from the “wisdom of the crowds”. Random Forest is well-known to be insensitive to the values of *n*tree, *nodesize*, and *m*try [60]. Here, default values were adopted for *n*tree = 500 and *nodesize* = 1 and *m*try = 30. Additionally, spectral importance results can be graphically viewed in the form of Gini plots. The Gini impurity metric accounts for how often a randomly selected component from a training set would be incorrectly labelled if it was randomly labelled according to the class distribution in a subset [61]. The mean decrease in the Gini, also known as Gini importance, is the total decrease in node impurities from splitting on the variable, averaged over all trees [62]. This is essentially a measure of how important a variable is for estimating the value of the target variable across all trees in the forest. Hence, by using this metric, RF can rank the spectral features in order of significance - for example, which wavenumbers are the most discriminating between the two classes [15].

#### Partial least squares-discriminant analysis

Partial Least Squares – Discriminant Analysis (PLS-DA) is supervised machine learning method that combines PLS regression (PLSR) and Linear Discriminant Analysis (LDA). This technique can extract important information from complex datasets, by reducing the dimensionality to reveal hidden patterns within the data. For binary classification problems, the technique separates classes by looking for a straight line that divides the data space into two distinct regions [63]. The data points are projected perpendicularly to the line, which is known as the discriminator [64]. The distances from the discriminator are referred to as the discriminant scores [65]. This information is provided in the form of new variables called PLS components, where the first PLS component (PLS1) accounts for the greatest variation in the dataset, PLS2 represents the next greater variation, and so on. PLS scores plots give an overview of the general inconsistencies within large datasets, and loadings plots further

explain the variance, by suggesting where the most variable regions exist e.g. which spectral regions display the highest disparity. The optimal number of components, *ncomp*, is determined when tuning the classification models. The best value for *ncomp* provides the most reliable results, so that the cross-validation error is minimized.

### Support vector machine

A support vector machine (SVM) is a supervised algorithm, commonly employed for classification purposes [66]. From known data, SVM outputs an optimal dimension for the separation of the data, known as the hyperplane. Support vectors are the co-ordinates of the individual observation and the hyperplane can be used to categorise new samples [67]. Linear, radial basis function, and polynomial kernels have all been used in SVM models. Here we use the linear kernel that has previously been shown to perform well in spectral classification studies [68]. The optimization of SVM tuning parameters can change the classification efficiency dramatically. The penalty parameter *cost* is responsible for the trade-off between smooth boundaries and the ability to classify the data [69].

## Results

### Brain Cancer vs Control

Each classification model was executed multiple times to ensure the variance within the dataset was fully encompassed. Bootstrapping analysis on the training set showed 51 resamples to be a reliable number of iterations, as shown in Figure 2, the standard error adequately converges at around 51 resamples for both (a) sensitivity and (b) specificity. A higher number of iterations reduces the variance, but also increases the time required to run the classification models. At 51 iterations, the standard error for the test set was 0.13% for sensitivity and 0.19% for specificity, which was deemed to be an acceptable level of error, with reasonable analysis time.

### *Random forest results*

Following optimal pre-processing, a single RF classification model was trained on the training data and used to predict the test set. The value of *mtry* was chosen to be 30, which gave sensitivity and specificity of 93.8% and 80.1% for the 5-fold cross-validation on the training data. Initial analysis of the single RF model emphasised the ability to successfully pick out the brain cancer patients from the training set, with the test set reporting a sensitivity of 92.5%. However, the specificity was much lower at 76.1%, meaning this particular model incorrectly predicted many of the control patients as having the disease. That said, the lower specificity could be attributed to the class imbalance within the dataset. As the sensitivity of a model relates to the ability to detect patients with disease, and 487 out of the 724 patient samples were cancerous, there was a bias towards the prediction of a brain cancer. The addition of statistical sampling techniques can reduce the class imbalance and improve the accuracy of the model. Table 1 compares classification output of the initial RF model with the three different resampling techniques. The up-sampling method was not effective in improving this model, losing 0.7% on sensitivity and only increasing specificity by 1.4%. On the other hand, down-sampling greatly improved the specificity, rising from 76.1% to 85.9%. Some studies have been critical of down-sampling, as the technique ‘ignores’ data that could provide important differences and/or similarities between the two classes [70] That being said, multiple iterations could potentially overcome this data loss, as there would be different samples down-sampled during each iteration. With SMOTE sampling technique, the minority class (controls in this case) is synthetically up-sampled in order to be more comparable with the majority class [58]. This was found to provide the best output, with a sensitivity and specificity of 94.5% and 88.7% respectively.

**Table 1 - Sampling comparison for single random forest classifications**

	<b>RF only</b>	<b>Up-sampling</b>	<b>Down-Sampling</b>	<b>SMOTE</b>
<b>Sensitivity (%)</b>	92.5	91.8	90.4	94.5
<b>Specificity (%)</b>	76.1	77.5	85.9	88.7

Figure S1 shows the confusion matrices of the (a) initial and (b) SMOTE models which describe the predictions that were made in the random forest test sets. As outlined above in Table 1, the specificity

increased to 88.7%, predicting 63 out of 70 non-cancer patients correctly. The sensitivity remained high, only falsely predicting 8 out of 146 cancer patients as non-cancer, resulting in a sensitivity of 94.5%. As the SMOTE sampling was found to be optimal in this case, it was used for the resampled classification.

The single model results were promising, but to ensure they were reliable, the RF model was resampled 51 times. The 51 independent RF models were combined to provide mean sensitivity and specificity values, as well as the standard deviations to account for the statistical variance. Table 2 lists the mean and standard deviation (SD) values for the sensitivity, specificity,  $\kappa$  and balanced accuracy relating to the 51 RF iterations. The PPV and NPV were also calculated from the mean sensitivity and specificity, as well as the prevalence of brain tumours - reported as approximately 1.6% [7]. The ability to successfully predict the brain cancer patients was high, with an average sensitivity of 93.1%. However, despite SMOTE being more beneficial for the singular RF model, the average specificity over the 51 iterations dropped to 81.1%, ranging from 73.2% to 92.9%. This particular RF classification performed well in the detection of brain cancer within this dataset, but it was incorrectly assigning more of the non-cancer patients as ‘cancer’, resulting in a higher number of false positives. Clearly this would not be very efficient for the clinic, as the excess brain scans would be costly to the health services, meanwhile putting healthy patients through needless stress and anxiety. However, these findings are still relatively promising, with a health economic study reporting statistics >80% would be cost-effective to the NHS [38].

**Table 2 - Statistical results for the test set in the RF model with 51 iterations**

	<b>Mean</b>	<b>SD</b>
<b>Kappa Value</b>	0.75	0.05
<b>Sensitivity (%)</b>	93.1	1.97
<b>Specificity (%)</b>	81.1	3.90
<b>Balanced Accuracy (%)</b>	87.1	2.35

<b>PPV (%)</b>	7.4	-
<b>NPV (%)</b>	99.9	-

The Gini impurity metric was examined to identify the most important features within the dataset. The accuracy and reliability of the model can be determined from the RF statistical value outputs, with the Gini plot highlighting wavenumbers responsible for the results for the optimal model (Figure 3).

Table 3 gives an overview of the top 15 identified wavenumbers in order of importance, with their corresponding wavenumber assignments and vibrational modes. The column " $\sum$ Gini" in the table is a summation of the mean decrease in Gini for each wavenumber, over all nodes in all trees in the ensemble.

**Table 3 - Top 15 wavenumbers from RF classification of brain cancer vs non-cancer with tentative biochemical assignments [14,16]**

Wavenumbers (cm <sup>-1</sup> )	$\sum$ Gini	Tentative Assignments	Vibrational Modes
1524.5	619.1	Amide II of proteins	$\delta$ (N-H), $\nu$ (C-N), $\delta$ (C-O), $\nu$ (C-C)
1516.5	430.0		
1532.5	425.7		
1508.5	193.2		
1028.5	177.8	Glycogen	$\nu$ (C-O), $\nu$ (C-C), def(C-OH)
1036.5	150.0		
1500.5	120.7	Amide II of proteins	$\delta$ (N-H), $\nu$ (C-N), $\delta$ (C-O), $\nu$ (C-C)
1540.5	95.2		
1020.5	90.9	DNA/Glycogen	$\nu$ (PO <sup>2-</sup> )/ $\nu$ (C-O)
1788.5	83.2	Lipids	$\nu$ (C=O)
1044.5	79.3	Nucleic Acids	$\nu$ (PO <sup>2-</sup> )
1796.5	79.1	Lipids	$\nu$ (C=O)
1668.5	76.1	Amide I of proteins	$\nu$ (C=O), $\nu$ (C-N), $\delta$ (N-H)
1012.5	67.2	Carbohydrate	$\nu$ (C-O)
1492.5	67.1	Amide II of proteins	$\delta$ (N-H), $\nu$ (C-N), $\delta$ (C-O), $\nu$ (C-C)

$\nu$  = stretching;  $\delta$  = bending; def = deformation

The most discriminatory region is the Amide II band, making up the top 4 wavenumbers with extremely high  $\sum$ Gini values. The out-of-phase combination of the NH bending and the CN stretching

vibrations, as well as minor contributions from the CO in-plane bend and the CC/NC stretching vibrations, give rise to the Amide II band [71]. Certain wavenumbers in the lower wavenumber region - relating to carbohydrates, glycogen and nucleic acids – were also shown to be highly discriminating. These areas of importance are closely followed by lipid and other protein (Amide I) contributions.

#### *PLS-DA results*

The optimal value of *ncomp* was found to be 14, which was selected from a tuning grid with a range 1:20. This gave sensitivity of 89.4% and specificity of 88.7% for the 5-fold cross-validation on the training data. An initial PLS-DA model reported a sensitivity of 95.9% and specificity of 81.7% for the external test set. Similar to the RF analysis, the sampling techniques were used to balance the classes. All three methods greatly enhanced the specificity, each improving by 10% or greater (Table 4), but this was costly for the sensitivity values, each falling below 90%. It is likely that the fall in sensitivity is caused by the class imbalance, as the initial model can be biased and overpredict the majority of the patients as ‘cancer’, simply because there is many more within the dataset. The sampling techniques balance the classes, giving an impartial representation and hence more reliable predictions. Again, as with the RF, the best results were obtained using SMOTE sampling (Figure S2), hence 51 iterations of the RF + SMOTE classification was employed.

**Table 4 - Sampling comparison for single PLS-DA classifications**

	<b>PLS-DA only</b>	<b>Up-sampling</b>	<b>Down-Sampling</b>	<b>SMOTE</b>
<b>Sensitivity (%)</b>	95.9	86.3	86.3	89.7
<b>Specificity (%)</b>	81.7	92.9	94.3	91.6

Table 5 lists the results from the 51 resamples of the PLS-DA/SMOTE model. PLS-DA was not as effective at predicting the brain cancer patients correctly, reporting an average sensitivity of 90.5%, in comparison to 93.1% for the RF model. However, the average specificity was 91.1%, meaning PLS-DA was far superior in correctly assigning the control samples as ‘non-cancer’. Out of the three

classification models, PLS-DA reported the best PPV, at 14.2%, almost double that of the RF model (7.4%).

**Table 5 - Statistical results for the test set in the PLS-DA + SMOTE model with 51 iterations**

	<b>Mean</b>	<b>SD</b>
<b>Kappa Value</b>	0.71	0.10
<b>Sensitivity (%)</b>	90.5	2.09
<b>Specificity (%)</b>	91.1	3.28
<b>Balanced Accuracy (%)</b>	90.8	1.83
<b>PPV (%)</b>	14.2	-
<b>NPV (%)</b>	99.8	-

Figure 4 shows the scores plot between the first and second PLS components. There is a substantial amount of overlap between the two classes, with some separation across the 2<sup>nd</sup> PLS component (PLS2). The loadings plot for PLS2 is described in Figure 5, which suggests the biggest variance within the brain tumour dataset exists in the Amide II region (1500-1600cm<sup>-1</sup>), and in the lower wavenumber region (1000-1100cm<sup>-1</sup>). This agrees with the RF Gini importance values, in that the Amide II of proteins, and the bands from glycogen/carbohydrate/phosphate vibrations are most discriminatory.

#### *SVM results*

Similar analysis was carried out using an SVM-based classification (Table 6). The SVM model was tuned using the optimal value for *cost*, which was determined to be 0.019 by running a sequence between 0.001 and 0.03 at intervals of 0.018. Again, the use of the sample balancing techniques greatly improved the accuracy of the model, with SMOTE being the preferred method. The linear-SVM with SMOTE single model performed slightly better than RF and PLS-DA, with both sensitivity

and specificity above 90%, as described in the test set confusion matrices (Figure S3). Table 7 reports the statistics for the 51 SVM iterations using SMOTE.

**Table 6 - Sampling comparison for single SVM classifications**

	<b>SVM only</b>	<b>Up-sampling</b>	<b>Down-Sampling</b>	<b>SMOTE</b>
<b>Sensitivity (%)</b>	93.2	89.7	87.7	91.7
<b>Specificity (%)</b>	81.7	94.4	94.4	90.1

**Table 7 - Statistical results for the test set in the SVM model with 51 iterations**

	<b>Mean</b>	<b>SD</b>
<b>Kappa</b>	0.80	0.03
<b>Sensitivity (%)</b>	92.1	2.1
<b>Specificity (%)</b>	88.7	3.3
<b>Balanced Accuracy (%)</b>	90.4	1.5
<b>PPV (%)</b>	13.5	-
<b>NPV (%)</b>	99.9	-

Here, the average sensitivity was 92.1% but the specificity was slightly lower at 88.7%. Again, we achieve a balanced accuracy over 90%, and the PPV of 13.5% is relatively high. The SVM model produced a mean  $\kappa$  value of 0.8, indicating almost perfect agreement which suggests this particular model was robust and reliable.

### *ROC curves*

In addition, receiver operating characteristic (ROC) curves can illustrate the diagnostic ability of machine learning classifiers, and aid with tuning the classification model for clinical applications. The area under the curve (AUC) represents the measure of separability, with the higher AUC the better the model is at distinguishing between classes [72]. The ROC graph in Figure 6 suggests the diagnostic

performance of all three models is extremely promising. The ROC curves are all relatively symmetrical across sensitivity and specificity - the PLS-DA curve is slightly better with an AUC value of 0.948, which is regarded as excellent.

In general, the results from this study confirm the ability of serum spectroscopy - coupled with computational analysis - to be effective in differentiating brain lesions from healthy controls. Using basic machine learning techniques, we have successfully separated brain cancer and non-cancer with accuracies greater than 90%. Both PLS-DA and SVM performed extremely well, with the PLS-DA model reporting an average sensitivity and specificity of 90.5% and 91.1% respectively, meanwhile the linear-SVM produced 92.1% sensitivity and 88.7% specificity. These results are just slightly inferior to the 92.8% and 91.5% reported by Hands *et al.* [33]. However, there are various differences between these studies, meaning they are not entirely comparable. The patient cohorts were comprised of different patients, and this dataset contains almost 300 more serum samples. Hands *et al.* employed a radial basis function (RBF) based SVM, whereas we compare the capability of linear-SVM to basic RF and PLS-DA models, all of which provided promising results. Despite **these differences**, accuracies above 90% suggests that the new SIRE technology is comparable with the traditional fixed diamond IRE, indicating the high-throughput design could now be implemented into the clinical environment, to enable a quick blood test for the early detection of brain cancer.

#### Glioblastoma Multiforme vs Lymphoma

Neuro-oncologists are particularly interested in the challenge of differentiating of primary cerebral lymphoma from GBM. It can often be difficult to distinguish between these diagnoses on brain imaging alone, such as magnetic resonance imaging (MRI). This therefore necessitates patients to have surgical biopsy in order to identify the tumour pathology, and to determine the most appropriate regimen of surgery, chemotherapy and radiotherapy. A reliable serum diagnostic test could avoid the need for surgery and speed up time to definitive chemotherapy and radiotherapy.

Additional serum samples were collected from the Walton Centre and the Royal Preston hospital, to provide a total of 41 lymphoma samples. A random subset of the GBM samples from the 724 patient dataset were used for the comparisons. **The patient information is summarised in supplementary information Table S3.**

Similar to the 724 dataset, bootstrapping analysis was done on the lymphoma vs GBM training set to search for an acceptable number of iterations; 51 resamples were also found to be sufficient, with the standard error converging at this point (**Figure S4**). An initial RF model provides us with the biochemical differences between the lymphoma and GBM patients. The Gini plot (**Figure 7**) suggests the Amide II region is of particular importance, closely followed by the Amide I band. Between 1150-1000 $\text{cm}^{-1}$  there are various significant bands, relating to vibrations within nucleic material, glycogen and carbohydrates (**Table 8**).

**Table 8 - Top 15 wavenumbers from RF classification of lymphoma vs GBM with tentative biochemical assignments [14,16]**

Wavenumbers ( $\text{cm}^{-1}$ )	$\Sigma$ Gini	Tentative Assignments	Vibrational Modes
1556.5	95.9	Amide II of proteins	$\delta(\text{N-H})$ , $\nu(\text{C-N})$ , $\delta(\text{C-O})$ , $\nu(\text{C-C})$
1564.5	91.4		
1676.5	57.9	Amide I of proteins	$\nu(\text{C=O})$ , $\nu(\text{C-N})$ , $\delta(\text{N-H})$
1684.5	50.1		
1572.5	42.9	Amide II of proteins	$\delta(\text{N-H})$ , $\nu(\text{C-N})$ , $\delta(\text{C-O})$ , $\nu(\text{C-C})$
1548.5	32.6		
1668.5	32.2	Amide I of proteins	$\nu(\text{C=O})$ , $\nu(\text{C-N})$ , $\delta(\text{N-H})$
1660.5	30.5		
1020.5	19.7	DNA/Glycogen	$\nu(\text{PO}_2^-)$ / $\nu(\text{C-O})$ , def(C-OH)
1100.5	19.0	Nucleic Acids	$\nu(\text{PO}_2^-)$
1036.5	17.4	Glycogen	$\nu(\text{C-O})$ , $\nu(\text{C-C})$
1692.5	15.3	Amide I of proteins	$\nu(\text{C=O})$ , $\nu(\text{C-N})$ , $\delta(\text{N-H})$
1108.5	14.6	Carbohydrate	$\nu(\text{C-O})$ , $\nu(\text{C-C})$
1628.5	14.5	Amide I of proteins	$\nu(\text{C=O})$ , $\nu(\text{C-N})$ , $\delta(\text{N-H})$
1620.5	13.2		

$\nu$  = stretching;  $\delta$  = bending; def = deformation

For the PLS-DA model, in this case the optimal value for *ncomp* was found to be 4. As with the C/NC set, the scores plot separates the lymphoma and GBM patients across the 2<sup>nd</sup> PLS component (Figure 8). Again, we see the highest discrimination arises from the Amide bands and the lower wavenumber region on the loadings plot (Figure 9). For lymphoma vs GBM, the Amide I region is also highly discriminatory, substantiating the RF Gini findings outlined previously in Table 8. **The confusion matrix for a single PLS-DA model is described in Figure S5.** SMOTE showed to be the best sampling technique for RF and PLS-DA, but up-sampling was found to be optimal for the SVM-based model (Table 9). The PPV and NPV are not included here, as the prevalence value for this particular classification is difficult to estimate.

**Table 9 - Statistical results for the lymphoma vs GBM test sets from the three different classification models with 51 iterations**

	RF + SMOTE		PLS-DA + SMOTE		SVM + UP	
	Mean	SD	Mean	SD	Mean	SD
<b>Kappa</b>	0.63	0.13	0.76	0.09	0.72	0.11
<b>Sensitivity (%)</b>	90.9	5.8	90.1	5.7	86.6	8.5
<b>Specificity (%)</b>	70.8	14.9	86.3	9.4	86.3	9.5
<b>Balanced Accuracy (%)</b>	80.8	7.2	88.2	5.0	86.4	5.4

For this particular dataset, the sensitivities refer to the ability to detect GBM, and the specificity relates to lymphoma. As shown in Table 9, the inferior model for this dataset was found to be RF – despite having a high sensitivity, the specificity was rather low at 70.8%. Although the balance between sensitivity and specificity could be tuned by optimising the probability threshold of the classifier, the fact that Kappa and balanced accuracy have significantly lower values than for the other models suggests that it would not change the rankings of the models. SVM combined with up-sampling performed well, reporting a balanced accuracy of 86.4%. The PLS-DA + SMOTE method seemed to be the optimal model, with a sensitivity of 90.1%, a specificity of 86.3%, and the highest  $\kappa$  value of all three models – mean  $\kappa = 0.76$ . The sensitivities were relatively stable, but the predictions

for lymphoma were more variable, for example, one of the RF resamples reported a sensitivity of 42%, which ultimately lowered the mean value. That said, the ROC curves for both the SVM-based and PLS-DA models still indicate promising diagnostic capability, with AUC values of  $\sim 0.9$  (Figure 10). The RF ROC curve is substantially lower across sensitivity and specificity compared to the other two techniques, reporting a much lesser AUC value of 0.829, which coincides with the lower balanced accuracy from the RF classification. For this diagnostic test to be validated, it could be said that more patients would have to be introduced. However, these results indicate the potential for a serum diagnostic tool at the secondary care stage, that could aid clinicians when brain scans are inconclusive. Furthermore, it would be beneficial if this type of blood test could prevent patients undergoing unnecessary surgical biopsy.

## **Conclusion**

The implementation of a quick blood serum test for the early detection of brain tumours at a GP setting could have a huge impact on the quality of life and prognosis for patients. The traditional fixed diamond crystal limited the translation of the technique, as the methodology was laborious, requiring long drying times and cleaning of the crystal between measurements [33]. An earlier health economic assessment in advance of prospective clinical data, stated the development of a high-throughput ATR accessory would be cost-effective to the UK's NHS if sensitivities and specificities  $>80\%$  were achieved [38]. This study validates the capability of the novel ClinSpec Dx SIRE optical sample slides [41].

We report a sensitivity and specificity of 90.5% and 91.1% respectively when separating brain cancer patients from healthy controls, through PLS-DA. Despite the prevalence of brain tumours being extremely low (1.6%), the PLS-DA model reported a PPV of 14.2%. Three different machine learning techniques have been compared, all of which report balanced accuracies of  $\sim 90\%$ , which would be deemed sufficient to be cost-effective to the NHS. Analysis of blood serum using this novel technique would fit ideally in the clinical pathway as a primary care triage tool for brain cancer. For the

effective treatment of this disease, it is vital to identify the tumours early. A test at this stage in the diagnostic pathway would provide GPs with further information to inform their referral decision. **If test a positive result for cancer was reported from a spectroscopic triage test, then the ‘at risk’ patients would progress into secondary care quicker, whilst a negative result would provide reassurance.** The time taken to diagnose brain cancer patients could be reduced significantly, whilst also saving on NHS funds.

By training a classification algorithm on this known population, new patient samples can be predicted based upon the derivation of cancer signals from the retrospective dataset, providing an appropriate measure of the true diagnostic accuracy. The first prospective clinical study employing ATR-FTIR spectroscopy is presented in Butler *et al.* [41] which is the first initial analysis of an ongoing clinical study at the Western General Hospital, in Edinburgh.

Additionally, we present the ability of this technique to differentiate between brain tumour types. Notably, the separation of lymphoma and GBM through ATR-FTIR spectroscopy would be particularly attractive for neurologists in a secondary care setting, when imaging results are not clear. This proof-of-principle study involved 112 patients, providing a sensitivity of 90.1% and a specificity of 86.3%. These statistics are hugely promising, and a  $\kappa$  value of 0.76 indicates the technique is reliable. Further analysis with a larger cohort of patients would be valuable, in order to make the diagnostic model more robust.

Analysis of the ROC curves and consideration of the prevalence of the diseases to be diagnosed suggests that in some cases the models presented here could be optimised for clinical applications by modifying the probability threshold that each classifier uses to discriminate between positive and negative classes. For example, to identify brain tumours, which have low prevalence of  $\sim 1.6\%$  in the clinical population, the balance between sensitivity and specificity is important, since the probability that a positive test result is truly indicative of a brain tumour (i.e. PPV) increases rapidly with specificity. Nonetheless, we believe this kind of optimisation can be done more accurately once more

is understood about the clinical population, which for the lymphoma vs GBM test in particular is not yet fully known. Therefore, the classifiers presented here all correspond to a default probability threshold of 0.5 used to discriminate between positive and negative classes, which is sufficient to demonstrate their efficacy.

The potential for high-throughput spectroscopy in the clinical environment goes further than purely brain tumour detection. With more research, this platform technology could address the clinical need for various malignancies and other diseases. There are also gaps for this technology in secondary care scenarios, for patient disease progression, treatment monitoring, and potentially aiding oncologists when MRI scans are inconclusive, as discussed in this study. The ultimate goal is to make the diagnostic pathways more efficient, cost effective and allow patients to obtain early treatment for optimal outcome.

#### ACKNOWLEDGMENTS

The authors would like to thank the EPSRC (EP/L505080/1) for funding

## References

- [1] Cancer Research UK. Brain, other CNS and intracranial tumours statistics 2019. <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/brain-other-cns-and-intracranial-tumours/survival>.
- [2] Burnet NG, Jefferies SJ, Benson RJ, Hunt DP, Treasure FP. Years of life lost (YLL) from cancer is an important measure of population burden — and should be considered when allocating research funds. *British Journal of Cancer* 2005;92:241–5. doi:10.1038/sj.bjc.6602321.
- [3] Spalding K, Board R, Dawson T, Jenkinson MD, Baker MJ. A review of novel analytical diagnostics for liquid biopsies: spectroscopic and spectrometric serum profiling of primary and secondary brain tumors. *Brain and Behavior* 2016;6:e00502. doi:10.1002/brb3.502.
- [4] Latinovic R. Headache and migraine in primary care: consultation, prescription, and referral rates in a large population. *Journal of Neurology, Neurosurgery & Psychiatry* 2005;77:385–7. doi:10.1136/jnnp.2005.073221.
- [5] The Brain Tumour Charity. Defeating Brain Tumours 2017. [https://www.thebraintumourcharity.org/media/filer\\_public/49/b5/49b5e1d3-6cff-4399-9d7e-9fabbb5c76a3/the-strategy-publication-rgb-v3.pdf](https://www.thebraintumourcharity.org/media/filer_public/49/b5/49b5e1d3-6cff-4399-9d7e-9fabbb5c76a3/the-strategy-publication-rgb-v3.pdf).

- [6] The Brain Tumour Charity. Finding a Better way? 2017. [http://cdn.basw.co.uk/upload/basw\\_21512-10.pdf](http://cdn.basw.co.uk/upload/basw_21512-10.pdf).
- [7] Zienius K, Grant R, Brennan P. Impact of Open access CT (OACT) for headache suspected of brain cancer in Lothian. *Neuro-Oncology* 2018;20:i8–i8. doi:10.1093/neuonc/nox237.035.
- [8] Toh C-H, Castillo M, Wong AM-C, Wei K-C, Wong H-F, Ng S-H, et al. Primary Cerebral Lymphoma and Glioblastoma Multiforme: Differences in Diffusion Characteristics Evaluated with Diffusion Tensor Imaging. *AJNR Am J Neuroradiol* 2008;29:471–5. doi:10.3174/ajnr.A0872.
- [9] Baker MJ, Byrne HJ, Chalmers J, Gardner P, Goodacre R, Henderson A, et al. Clinical applications of infrared and Raman spectroscopy: state of play and future challenges. *The Analyst* 2018;143:1735–57. doi:10.1039/C7AN01871A.
- [10] Byrne HJ, Baranska M, Puppels GJ, Stone N, Wood B, Gough KM, et al. Spectroscopy for the next generation: Quo vadis? *The Analyst* 2015;140:2066–73. doi:10.1039/C4AN02036G.
- [11] Kong K, Kendall C, Stone N, Notingher I. Raman spectroscopy for medical diagnostics — From in-vitro biofluid assays to in-vivo cancer detection. *Advanced Drug Delivery Reviews* 2015;89:121–34. doi:10.1016/j.addr.2015.03.009.
- [12] Petrich W. MID-INFRARED AND RAMAN SPECTROSCOPY FOR MEDICAL DIAGNOSTICS. *Applied Spectroscopy Reviews* 2001;36:181–237. doi:10.1081/ASR-100106156.
- [13] Perkin Elmer. FTIR Spectroscopy: Attenuated Total Reflectance (ATR) 2018. [https://shop.perkinelmer.com/content/TechnicalInfo/TCH\\_FTIRATR.pdf](https://shop.perkinelmer.com/content/TechnicalInfo/TCH_FTIRATR.pdf).
- [14] Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, et al. Using Fourier transform IR spectroscopy to analyze biological materials. *Nature Protocols* 2014;9:1771–91. doi:10.1038/nprot.2014.110.
- [15] Smith BR, Ashton KM, Brodbelt A, Dawson T, Jenkinson MD, Hunt NT, et al. Combining random forest and 2D correlation analysis to identify serum spectral signatures for neuro-oncology. *Analyst* 2016;141:3668–3678.
- [16] Movasaghi Z, Rehman S, ur Rehman DrI. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. *Applied Spectroscopy Reviews* 2008;43:134–79. doi:10.1080/05704920701829043.
- [17] Walsh MJ, Holton SE, Kajdacsy-Balla A, Bhargava R. Attenuated total reflectance Fourier-transform infrared spectroscopic imaging for breast histopathology. *Vibrational Spectroscopy* 2012;60:23–8. doi:10.1016/j.vibspec.2012.01.010.
- [18] Bird B, Remiszewski S, Akalin A, Kon M, Diem M, others. Infrared spectral histopathology (SHP): a novel diagnostic tool for the accurate classification of lung cancer. *Laboratory Investigation* 2012;92:1358.
- [19] Lasch P, Haensch W, Naumann D, Diem M. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 2004;1688:176–86. doi:10.1016/j.bbadis.2003.12.006.
- [20] Baker MJ, Gazi E, Brown MD, Shanks JH, Gardner P, Clarke NW. FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer. *British Journal of Cancer* 2008;99:1859–66. doi:10.1038/sj.bjc.6604753.
- [21] Gazi E, Baker M, Dwyer J, Lockyer NP, Gardner P, Shanks JH, et al. A Correlation of FTIR Spectra Derived from Prostate Cancer Biopsies with Gleason Grade and Tumour Stage. *European Urology* 2006;50:750–61. doi:10.1016/j.eururo.2006.03.031.

- [22] Bonnier F, Brachet G, Duong R, Sojinrin T, Respaud R, Aubrey N, et al. Screening the low molecular weight fraction of human serum using ATR-IR spectroscopy. *Journal of Biophotonics* 2016;9:1085–97. doi:10.1002/jbio.201600015.
- [23] Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD. Characterization of the Low Molecular Weight Human Serum Proteome. *Molecular & Cellular Proteomics* 2003;2:1096–103. doi:10.1074/mcp.M300031-MCP200.
- [24] Petricoin EF, Belluco C, Araujo RP, Liotta LA. The blood peptidome: a higher dimension of information content for cancer biomarker discovery. *Nat Rev Cancer* 2006;6:961–7. doi:10.1038/nrc2011.
- [25] Baker MJ, Hussain SR, Lovergne L, Untereiner V, Hughes C, Lukaszewski RA, et al. Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chemical Society Reviews* 2016;45:1803–1818.
- [26] Dorling K, Baker MJ. Highlighting attenuated total reflection Fourier transform infrared spectroscopy for rapid serum analysis. *Trends in Biotechnology* 2013;31:325–7. doi:10.1016/j.tibtech.2013.03.009.
- [27] Backhaus J, Mueller R, Formanski N, Szlama N, Meerpohl H-G, Eidt M, et al. Diagnosis of breast cancer with infrared spectroscopy from serum samples. *Vibrational Spectroscopy* 2010;52:173–7. doi:10.1016/j.vibspec.2010.01.013.
- [28] Ollesch J, Heinze M, Heise HM, Behrens T, Brüning T, Gerwert K. It's in your blood: spectral biomarker candidates for urinary bladder cancer from automated FTIR spectroscopy: Spectral cancer biomarkers from high-throughput FTIR spectroscopy. *Journal of Biophotonics* 2014;7:210–21. doi:10.1002/jbio.201300163.
- [29] Gajjar K, Trevisan J, Owens G, Keating PJ, Wood NJ, Stringfellow HF, et al. Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. *Analyst* 2013;138:3917–26. doi:10.1039/C3AN36654E.
- [30] Zhang X, Thiéfin G, Gobinet C, Untereiner V, Taleb I, Bernard-Chabert B, et al. Profiling serologic biomarkers in cirrhotic patients via high-throughput Fourier transform infrared spectroscopy: toward a new diagnostic tool of hepatocellular carcinoma. *Translational Research* 2013;162:279–86. doi:10.1016/j.trsl.2013.07.007.
- [31] Hands JR, Abel P, Ashton K, Dawson T, Davis C, Lea RW, et al. Investigating the rapid diagnosis of gliomas from serum samples using infrared spectroscopy and cytokine and angiogenesis factors. *Analytical and Bioanalytical Chemistry* 2013;405:7347–55. doi:10.1007/s00216-013-7163-z.
- [32] Hands JR, Dorling KM, Abel P, Ashton KM, Brodbelt A, Davis C, et al. Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples: Serum spectroscopy gliomas. *Journal of Biophotonics* 2014;7:189–99. doi:10.1002/jbio.201300149.
- [33] Hands JR, Clemens G, Stables R, Ashton K, Brodbelt A, Davis C, et al. Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy. *Journal of Neuro-Oncology* 2016;127:463–72. doi:10.1007/s11060-016-2060-x.
- [34] Ferrer N. Forensic Science, Applications of IR Spectroscopy. *Encyclopedia of Spectroscopy and Spectrometry*, Elsevier; 1999, p. 603–15. doi:10.1006/rwsp.2000.0100.
- [35] Bonnier F, Petitjean F, Baker MJ, Byrne HJ. Improved protocols for vibrational spectroscopic analysis of body fluids: Improved protocols for vibrational spectroscopic analysis of body fluids. *Journal of Biophotonics* 2014;7:167–79. doi:10.1002/jbio.201300130.

- [36] Stuart B. Infrared Spectroscopy. In: John Wiley & Sons, Inc., editor. Kirk-Othmer Encyclopedia of Chemical Technology, Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2005. doi:10.1002/0471238961.0914061810151405.a01.pub2.
- [37] Smith BC. Fundamentals of Fourier transform infrared spectroscopy. Boca Raton, Fla.: CRC Press; 2011.
- [38] Gray E, Butler HJ, Board R, Brennan PM, Chalmers AJ, Dawson T, et al. Health economic evaluation of a serum-based blood test for brain tumour diagnosis: exploration of two clinical scenarios. *BMJ Open* 2018;8:e017593. doi:10.1136/bmjopen-2017-017593.
- [39] Karabudak E, Kas R, Ogieglo W, Rafieian D, Schlautmann S, Lammertink RGH, et al. Disposable Attenuated Total Reflection-Infrared Crystals from Silicon Wafer: A Versatile Approach to Surface Infrared Spectroscopy. *Analytical Chemistry* 2013;85:33–8. doi:10.1021/ac302299g.
- [40] Koç M, Karabudak E. History of spectroscopy and modern micromachined disposable Si ATR-IR spectroscopy. *Applied Spectroscopy Reviews* 2018;53:420–38. doi:10.1080/05704928.2017.1366341.
- [41] Butler HJ, Brennan PM, Cameron JM, Finlayson D, Hegarty MG, Jenkinson MD, et al. A triage blood test for brain cancer: Development of high-throughput ATR-FTIR technology for rapid spectroscopic serum diagnostics. Accepted in *Nature Communications* 2019. doi:10.1038/s41467-019-12527-5.
- [42] Lovergne L, Clemens G, Untereiner V, Lukaszewski RA, Sockalingum GD, Baker MJ. Investigating optimum sample preparation for infrared spectroscopic serum diagnostics. *Anal Methods* 2015;7:7140–9. doi:10.1039/C5AY00502G.
- [43] Lovergne L, Bouzy P, Untereiner V, Garnotel R, Baker MJ, Thiéfin G, et al. Biofluid infrared spectro-diagnosics: pre-analytical considerations for clinical applications. *Faraday Discuss* 2016;187:521–37. doi:10.1039/C5FD00184F.
- [44] Cameron JM, Butler HJ, Palmer DS, Baker MJ. Biofluid spectroscopic disease diagnostics: A review on the processes and spectral impact of drying. *Journal of Biophotonics* 2018;11:e201700299.
- [45] Smith BR, Baker MJ, Palmer DS. PRFFECT: A versatile tool for spectroscopists. *Chemometrics and Intelligent Laboratory Systems* 2018;172:33–42. doi:10.1016/j.chemolab.2017.10.024.
- [46] RStudio Team. RStudio: Integrated Development for R. Boston, MA: RStudio Inc.; 2015.
- [47] Butler HJ, Smith BR, Fritzsche R, Radhakrishnan P, Palmer DS, Baker MJ. Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy. *Analyst* 2018;143:6121–34. doi:10.1039/C8AN01384E.
- [48] Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems* 2012;117:92–9. doi:10.1016/j.chemolab.2012.03.004.
- [49] Efron Bradley, Tibshirani RJ. An introduction to the bootstrap. New York, N.Y.; London: Chapman & Hall; 1993.
- [50] Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain* 2008;8:221–3. doi:10.1093/bjaceaccp/mkn041.
- [51] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey: IEEE; 2010, p. 3121–4. doi:10.1109/ICPR.2010.764.
- [52] Viera AJ, Garrett JM. Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine* n.d.:4.

- [53] McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;22:276–82.
- [54] Altman DG, Bland JM. Diagnostic tests 2: Predictive values. *BMJ* 1994;309:102.
- [55] Molinaro AM. Diagnostic tests: how to estimate the positive predictive value. *Neuro-Oncology Practice* 2015;2:162–6. doi:10.1093/nop/npv030.
- [56] Talluri SK. Positive predictive value. *BMJ* 2009;339:b3835–b3835. doi:10.1136/bmj.b3835.
- [57] SIMAFORE. Managing unbalanced data for building machine learning models 2019. <http://www.simafore.com/blog/handling-unbalanced-data-machine-learning-models>.
- [58] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 2002;16:321–57. doi:10.1613/jair.953.
- [59] Breiman L. Random Forests. *Machine Learning* 2001;45:5–32. doi:10.1023/A:1010933404324.
- [60] Palmer DS, O’Boyle NM, Glen RC, Mitchell JBO. Random Forest Models To Predict Aqueous Solubility. *J Chem Inf Model* 2007;47:150–8. doi:10.1021/ci060164k.
- [61] Louppe G. Understanding Random Forests; From Theory to Practice. University of Liege, 2014.
- [62] Hong Han, Xiaoling Guo, Hua Yu. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China: IEEE; 2016, p. 219–24. doi:10.1109/ICSESS.2016.7883053.
- [63] Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods* 2013;5:3790. doi:10.1039/c3ay40582f.
- [64] Lee LC, Liong C-Y, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst* 2018;143:3526–39. doi:10.1039/C8AN00599K.
- [65] Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics* 2014;28:213–25. doi:10.1002/cem.2609.
- [66] Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning* 1995;20:273–97. doi:10.1023/A:1022627411411.
- [67] de Boves Harrington P. Support Vector Machine Classification Trees. *Anal Chem* 2015;87:11065–71. doi:10.1021/acs.analchem.5b03113.
- [68] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics* 2018;15:41–51. doi:10.21873/cgp.20063.
- [69] Ben-Hur A, Weston J. A User’s Guide to Support Vector Machines. In: Carugo O, Eisenhaber F, editors. *Data Mining Techniques for the Life Sciences*, vol. 609, Totowa, NJ: Humana Press; 2010, p. 223–39. doi:10.1007/978-1-60327-241-4\_13.
- [70] Towards Data Science. Dealing with Imbalanced Classes in Machine Learning 2019. <https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2>.
- [71] Barth A, Zscherp C. What vibrations tell us about proteins. *Quarterly Reviews of Biophysics* 2002;35:369–430. doi:10.1017/S0033583502003815.
- [72] Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med* 2013;4:627–35.

## Figures

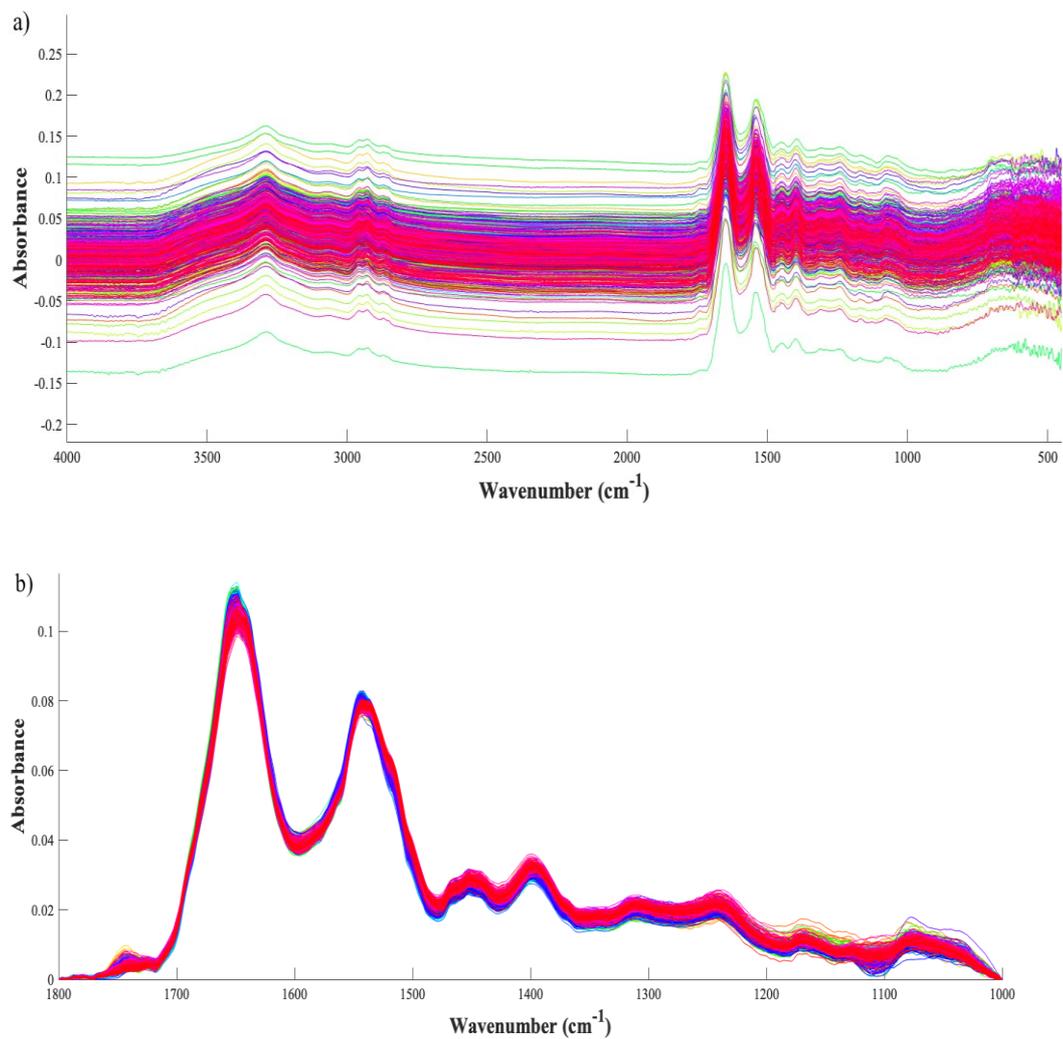


Figure 1 - Pre-processing example; (a) raw infrared data, and (b) pre-processed

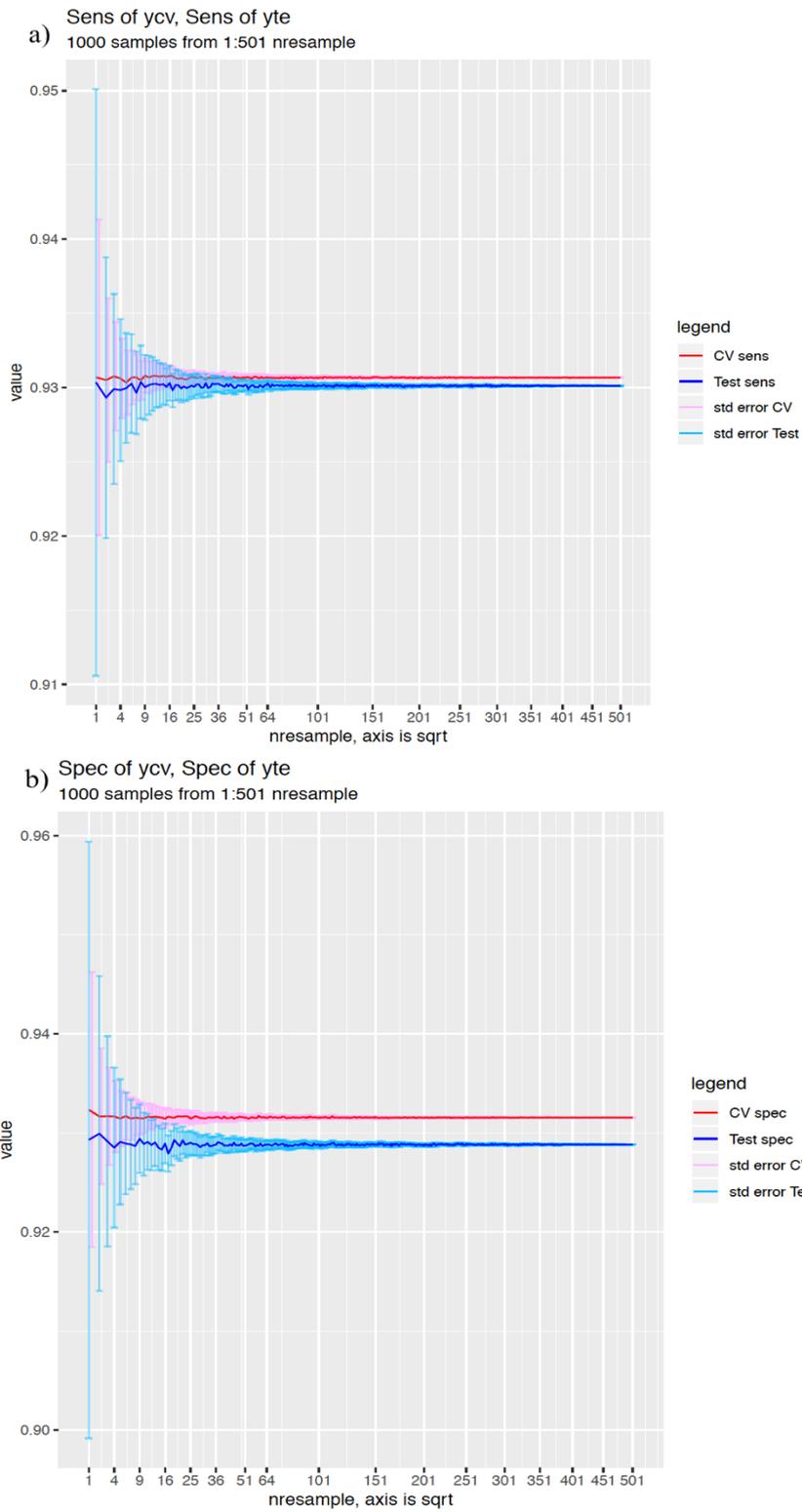


Figure 2 - Bootstrapping analysis to determine sufficient number of resamples required for the 724 patient dataset: (a) the sensitivity and (b) specificity

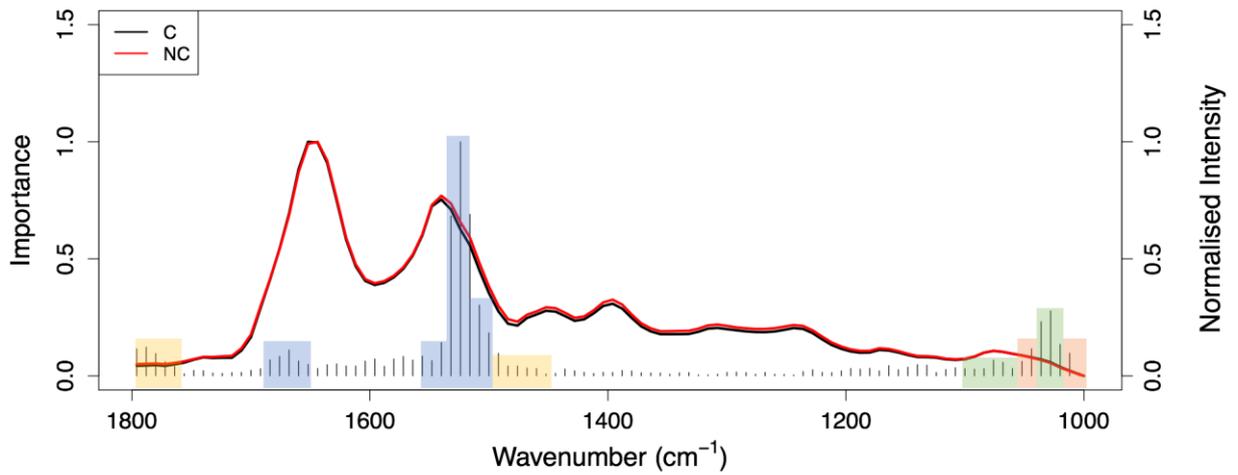


Figure 3 - Gini importance plot from random forest analysis showing the mean spectra from brain cancer (black) and control (red). Blue: Protein; Yellow: Lipid; Green: Nucleic acid and Orange: Carbohydrate

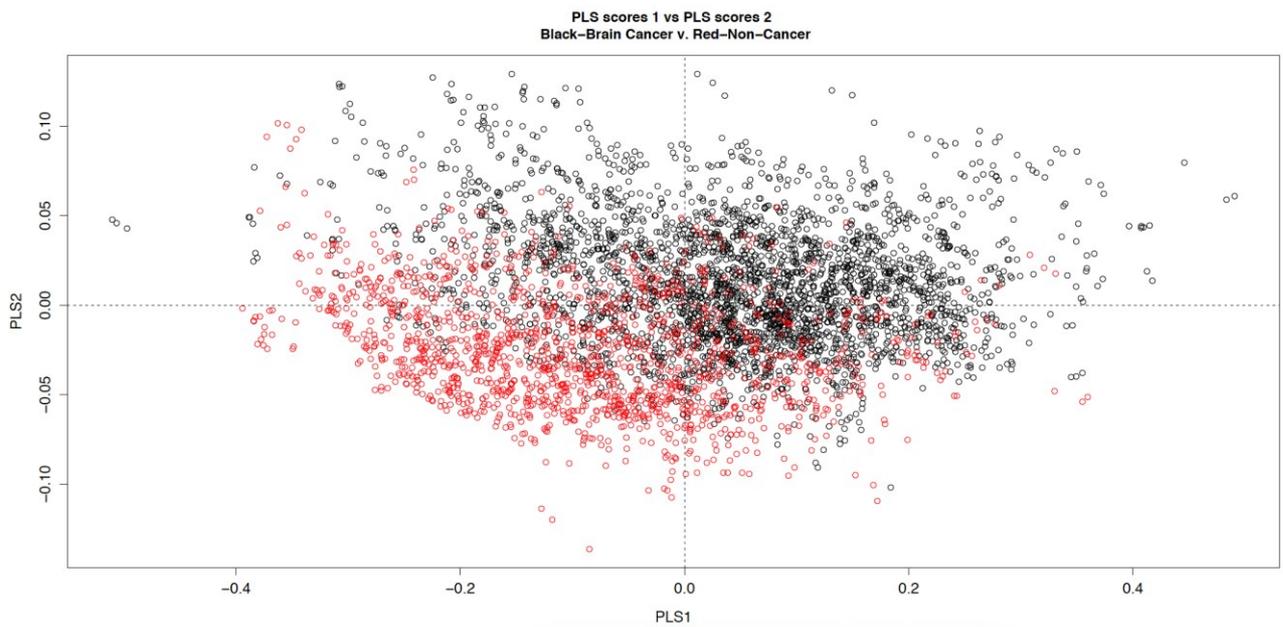


Figure 4 - Partial least squares-discriminant analysis; scores plot for brain cancer (black) vs control (red)

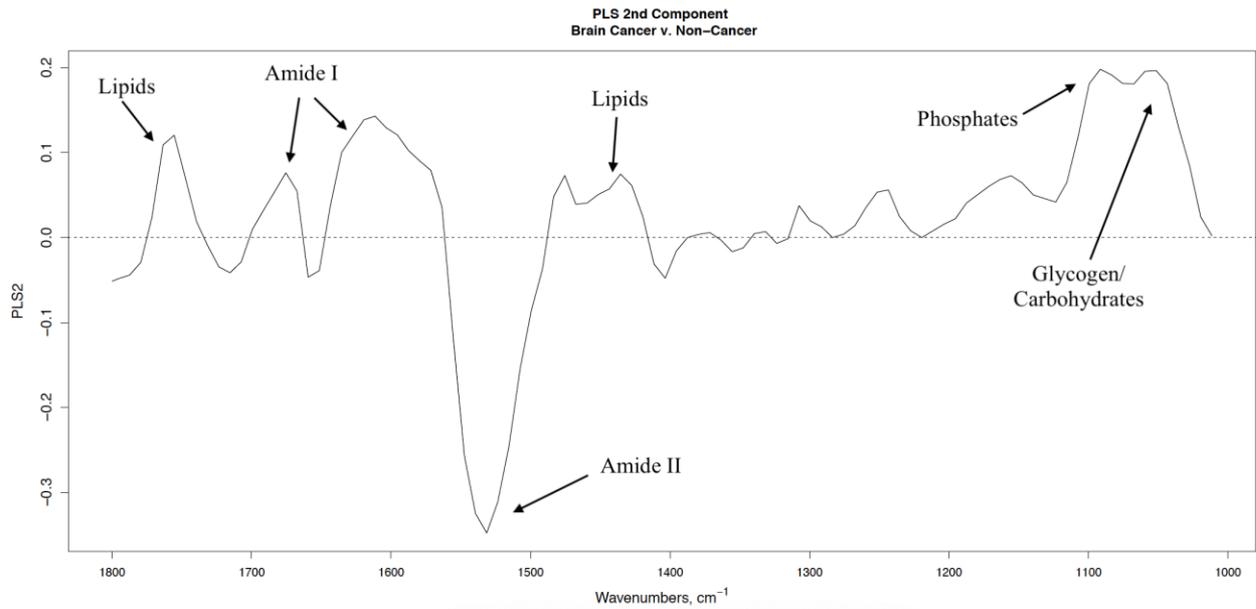


Figure 5 - Loadings plot for the 2<sup>nd</sup> partial least squares component with tentative biological assignments

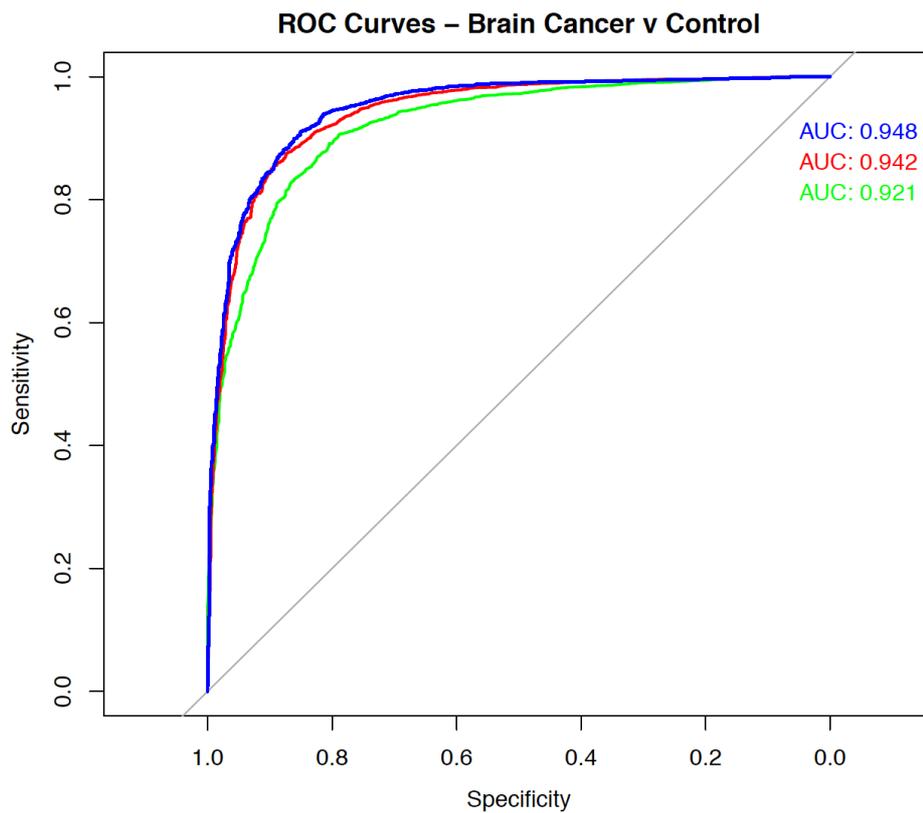


Figure 6 – ROC curves displaying trade-off between sensitivity and specificity of the three classification techniques for the cancer vs non-cancer patients: random forest; Green, partial least squares-discriminant analysis; Blue, support vector machine; Red

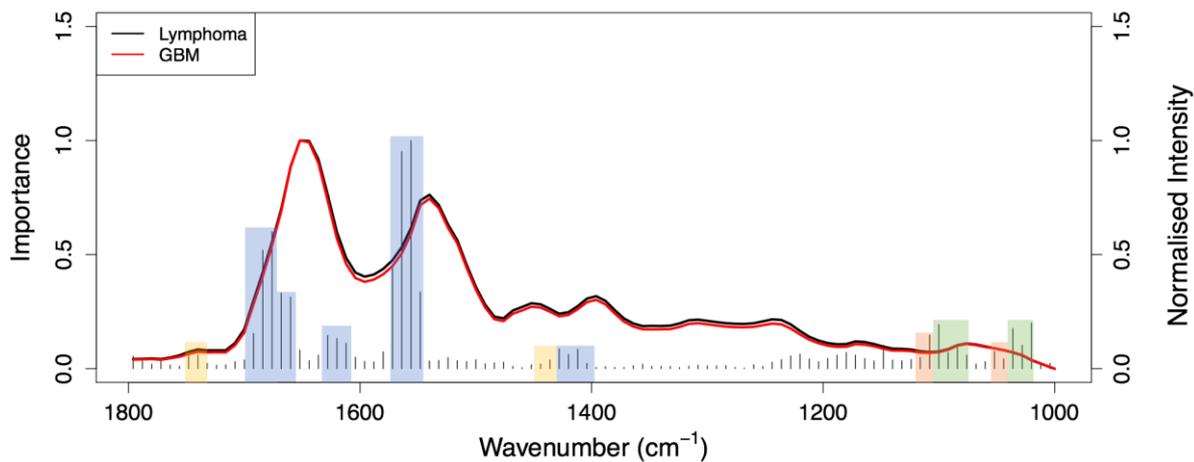


Figure 7 - Gini importance plot from random forest analysis showing the mean spectra from lymphoma (black) and glioblastoma (red). Blue: Protein; Yellow: Lipid; Green: Nucleic acid and Orange: Carbohydrate

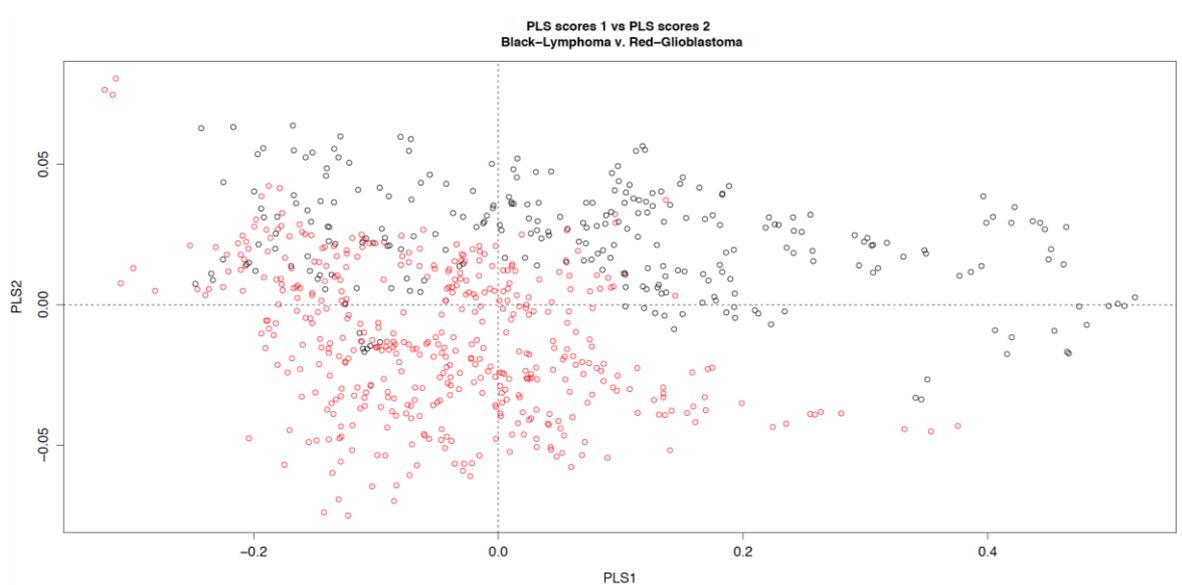
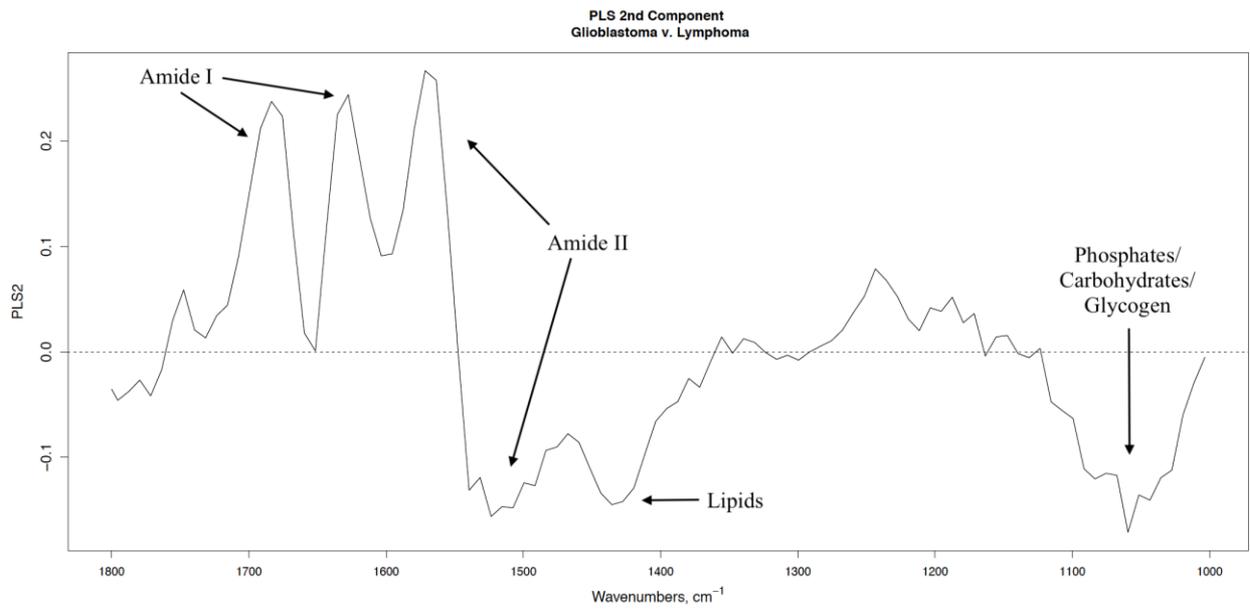
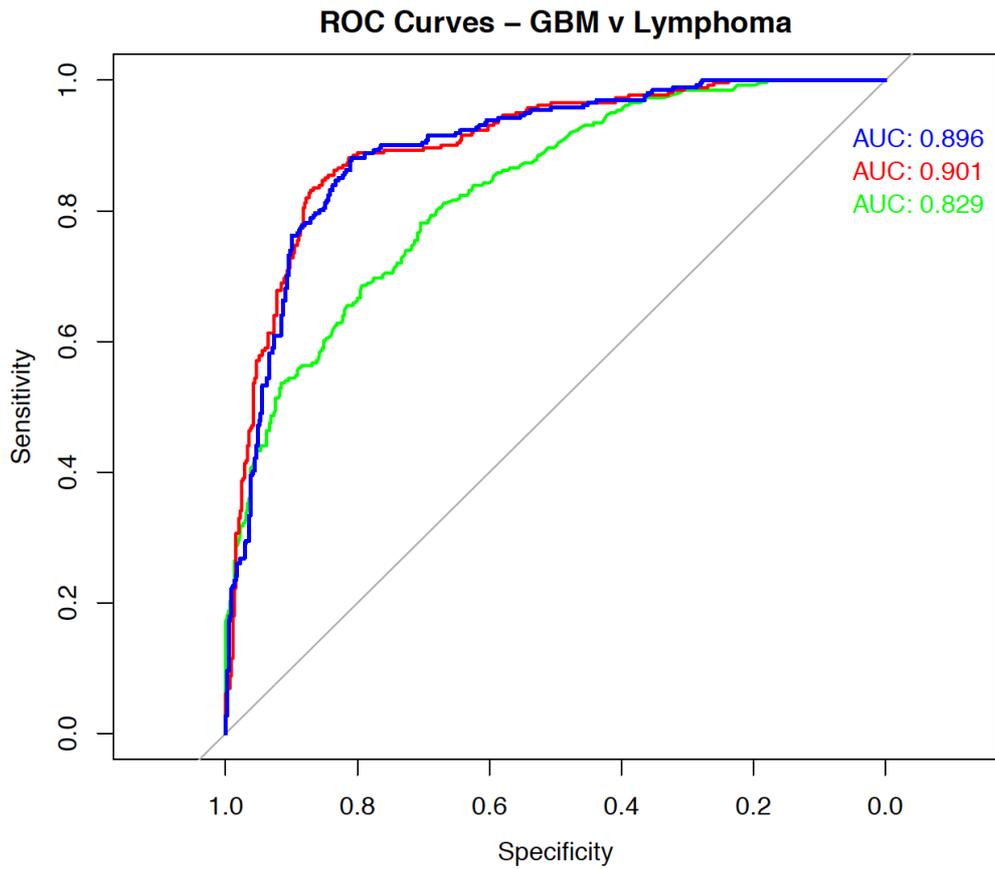


Figure 8 – Partial least squares-discriminant analysis; scores plot for lymphoma (black) vs glioblastoma (red)



*Figure 9 - Loadings plot for the 2<sup>nd</sup> PLS component in the lymphoma vs glioblastoma classification with tentative biological assignments*



*Figure 10 – ROC curve displaying trade-off between sensitivity and specificity of the three classification techniques for the lymphoma vs glioblastoma patients: random forest; Green, partial least squares-discriminant analysis; Blue, support vector machine; Red*