Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy

View Article Online
DOI: 10.1039/C8AN01384E

Holly J. Butler[a,b], Benjamin R. Smith[c], Robby Fritzsch[d], Pretheepan Radhakrishnan[e], David S. Palmer[b,c]* and Matthew J. Baker[a,b]*

[a] WestCHEM, Department of Pure and Applied Chemistry, University of Strathclyde, Technology and Innovation Centre, 99 George Street, Glasgow, G1 1RD, UK;
[b] ClinSpec Dx, Technology and Innovation Centre, 99 George St, Glasgow, G1 1RD, UK;
[c] WestCHEM, Department of Pure and Applied Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow, G1 1XL, UK;
[d] Department of Physics, University of Strathclyde, 107 Rottenrow East, Glasgow, G4 0NG, UK;
[e] Department of Biomedical Engineering, University of Strathclyde, 50 George Street, Glasgow, G1 1QE, UK;

*Corresponding authors:
Dr Matthew J Baker, matthew.baker@strath.ac.uk, +44(0)141 548 4700, @ChemistryBaker
Dr David S. Palmer, david.palmer@strath.ac.uk, +44(0)141 548 4178

Abbreviations: ATR, attenuated total reflectance; FTIR, Fourier-transform infrared, IR, infrared; biofluid

Abstract

Pre-processing is an essential step in the analysis of spectral data. Mid-IR spectroscopy of biological samples is often subject to instrumental and sample specific variances which may often conceal valuable biological information. Whilst pre-processing can effectively reduce this unwanted variance, the plethora of possible processing steps has resulted in a lack of consensus in the field, often meaning that analysis outputs are not comparable. As pre-processing is specific to the sample under investigation, here we present a systematic approach for defining the optimum pre-processing protocol for biofluid ATR-FTIR spectroscopy. Using a trial-and-error based approach and a clinically relevant dataset describing control and brain cancer patients, the effects of pre-processing permutations on subsequent classification algorithms were observed, by assessing key diagnostic performance parameters, including sensitivity and specificity. It was found that optimum diagnostic performance correlated with the use of minimal binning and baseline correction, with derivative functions improving diagnostic performance most significantly. If smoothing is required, a Sovitzky-Golay approach was the preferred option in this investigation. Heavy binning appeared to reduce classification most significantly, alongside wavelet noise reduction (filter length ≥ 6), resulting in the lowest diagnostic performances of all pre-processing permutations tested.

Introduction

A single IR spectrum obtained from a biological specimen contains not only the information of interest, but also underlying contributions from unwanted signals. Optimised sampling methods are essential to reducing this variability; however, these contributions are often still apparent in the subsequent dataset. Pre-processing can be defined as the reduction of these uncontrolled variables and can improve the experimental outcomes of spectral investigations.

Extracting biological variance arising from the sample itself is often the key aim of spectroscopic studies of biological materials. Whether this is exploratory or diagnostic, differences in biological content, molecular structure and distribution can allow differences to be observed within the dataset. However, spectra can also contain variance as a result of environmental, experimental and technical conditions. Respectively, factors such as humidity, sample morphology, and instrumental drift can all have negative impacts on spectral quality, repeatability and reproducibility (1).

The purpose of pre-processing is to reduce this unwanted variance, thus exposing the important underlying information from the spectral dataset. Consequently, pre-processing can improve exploratory analysis, classification and calibrations models, and interpretability whilst also removing outliers and trends, and reducing dimensionality (2) . It is important to acknowledge that pre-processing is not a solution to poor spectral data that arises from inherent issues with sample preparation and spectral acquisition. Whilst pre-processing may improve poor spectra, it is first imperative to obtain the highest quality spectra possible, within the constraints of sample and instrument (3) .

Fourier-transform Infrared (FTIR) spectroscopy has been widely applied to biological applications, due to its ability to identify chemical bonds characteristic of biological samples. More specifically, FTIR spectroscopy has been increasingly used as a tool to identify and differentiate disease status, in combination with machine learning and classification algorithms (4). For such approaches to perform optimally – that is with the highest sensitivity, specificity, accuracy and precision, in combination with low false positive and negative rates – the data must be pre-processed to ensure the important biological information is not concealed or diluted by systematic variance. Different combinations of pre-processing techniques have been shown to have a drastic impact on the diagnostic performance of machine learning algorithms, and thus an optimised approach to data handling must be employed prior to this form of analysis (Trevisan *et al.*, 2012; Gajjar *et al.*, 2013)

Sources of variance in FTIR spectroscopy

One of the primary sources of unwanted variance in an infrared (IR) spectrum derives from the phenomena of light scattering. Biological molecules absorb light in the mid-IR (MIR) region due matched frequencies between the incoming light and specific chemical bond vibrations (8). FTIR spectroscopy is able to produce an information rich spectrum that is indicative of the sample's discrete biochemical fingerprint. Due to this, the technology is widely implemented in the field of biological sciences, with applications spanning clinical, microbiological, pharmaceutical and food fields (9,10). However, despite the suitability of MIR light for analysing molecular vibrations of interest, the wavelength of this light (2.5 – 25 μm) is also highly correlated with the size of many biological samples, including cells and their subcellular components. These are ideal conditions for light scattering which can cause aberrations to the spectral baseline, and thus presents one of the most common issues in FTIR

investigations (11). This particular form of scattering is defined as Mie scattering and results in spectra that do not obey the principles of Beer-Lambert's Law, often altering the intensity and position of the amide I band (12,13). Scattering is also apparent in the analysis of powders, or other solids with uneven surfaces. As scattering is wavelength dependent (shorter wavelengths are more prone to scattering), a subsequent spectrum will often have higher absorption in the high wavenumber region (14) .

Furthermore, noise is an inherent issue with FTIR and other photonic techniques, that is apparent as high frequency signals within a spectrum. This noise can arise from electrical signals, mechanical vibrations, and environmental parameters, which are often unavoidable. A cooled detector, such as a deuterated triglycine sulfate detector (DGTS) can reduce thermal, or dark, noise in an IR system although not entirely (3) . Increased spectral noise can often overshadow subtle spectral features, and thus spectral quality is often assessed as a value of signal-to-noise, or the signal-to-noise ratio (SNR).

The optical pathlength of a system is directly implicated in Beer-Lambert's Law, and as such, FTIR spectra can also contain evidence of pathlength heterogeneity. This can commonly arise as a consequence of disparity in sample thickness, but can also occur due to intensity changes in the IR source (2) . The evidence of this is what may initially appear are gross spectral differences in absorbance, but are in actual fact indicative of inconsistencies in the sampling. Although not exhaustive, these factors can conceal interesting biological information by reducing the quality, accuracy and precision of IR spectra. By pre-processing data with evidence of such spectral features, the repeatability and reproducibility of the approach can be improved dramatically, leading to more insight in the data.

Pre-processing for FTIR spectroscopy

There is a wealth of pre-processing options available for spectral datasets, often providing more than one solution. Although this flexibility allows for optimised signal processing, the number of processing options available can often prove too large to systematically determine the best approach (15) . In this instance, we focus upon relevant pre-processing steps for FTIR spectroscopy, although this may draw from information from other techniques. Some processes are technique-specific, such as cosmic ray removal in Raman spectroscopy, but many are applicable to a wide range of spectroscopies including NMR and near-infrared (NIR) spectroscopy (16) .The development of spectral pre-processing methods has been largely developed in the field of NIR spectroscopy, and also has much overlap with Raman spectroscopy, due to similar susceptibilities to scattering and noise (14).

For an in depth overview of pre-processing applications in IR (and Raman) spectroscopy, the authors direct the reader to the following comprehensive review, which covers an array of pre-processing steps, namely: exclusion, normalisation, filtering, de-trending, transformations, feature selection, folding and other methods (2). In short, FTIR datasets often first undergo a form of quality control; an exclusion step where spectra with poor SNR or high water contributions for example, can excluded from the subsequent analysis. It is often important to undergo this step first, so that highly variable spectra do not influence subsequent analysis (such as processes that use the dataset mean, such as mean centering)(17) . Normalisation steps are required to negate differences in optical pathlength, allowing spectra to scaled relative to each other (18) . Baseline correction procedures are also commonly acquired to remove scattering as well as additive or multiplicative baselines (19). Additionally, filtering, or smoothing, can reduce the appearance of noise regions, thus potentially improving the clarity of spectral features and SNR. Spectral derivation is a useful

filtering tool that can be remove baseline effects and deconvolute complex spectra, whilst also improving diagnostic performances of classification algorithms (20,21)

As spectral datasets are highly dimensional, with a single spectrum alone often containing nearly around 3600 absorbance values, the computational burden of data processing can be high. A feature selection step that selects only the variables that are important to the post-analysis, can often make a large dataset more manageable, whilst also improving overall analysis accuracy (22). This can often be as simple as reducing the spectral range under investigation, or more sophisticated multivariate approaches such as principal component analysis (PCA) and partial least squares (PLS) which can also describe spectral differences between given experimental classes (23).

Consensus in the community

Although, there have been several attempts to unify the field of biological FTIR spectroscopy (3,24), there still remains a distinct lack of consensus with regards to pre-processing (16). This issue has been directly highlighted as one of the key objectives of The International Society of Clinical Spectroscopy, stressing the importance of establishing consensus between researchers (25).

Due to the variability between biological samples, spectral artefacts will be specific for each sample type, and even each individual experimental set-up. This therefore requires *a priori* knowledge of the sample, and the spectral response, in order to apply appropriate pre-processing steps. Through visual inspection of the dataset, indicators of unwanted spectral variance may be noticeable and thus pre-processing steps can be applied when deemed necessary by the analyst (16)

fig. This highly subjective approach may be the efficient with regards to analysis time, but will be variable between individuals. It has been shown that this may be improved using a trial-and-error based approach which systematically implements a range of pre-processing options, with the highest performing choice determined as the optimum protocol (26). A search algorithm, such as a genetic algorithm (GA), can optimise this process using machine learning to predict the optimal pre-processing steps (27) . However, despite the obvious benefits of this method, it can still be considered computationally heavy and is often not easily implemented in each spectroscopic experiment.

The order in which pre-processing steps are implemented is also another aspect of pre-processing to be optimised. It could be suggested that the largest source of spectral variance is minimised in the first instance, so that this is not influential in the next stage of analysis. For instance it is suggested that baseline effects should be removed prior to a normalisation step (15,28) . It has been suggested that the most effective approach for pre-processing is often the simplest, and as such the number of processes in a pre-processing protocol should be kept to the minimum (15).

The optimum sample pre-processing procedure is likely entirely sample specific, with suggestions that this may even be specific to the classification question being asked of the dataset (29). For instance, samples prone to contamination, such as paraffin embedded tissue, may undergo specific quality tests to automatically exclude spectra containing evidence of the contaminant (in this case, paraffin) (30). Whereas in contrast, a cell based investigation may be more prone to scattering and thus require a specific baseline correction (31).

Biofluid FTIR spectroscopy

The analysis of biofluids, such as blood serum, using IR spectroscopy is a rapidly progressing field that is nearing ever closer to clinical translation (17,32,33). Due to its simplicity and robust methodology, the analysis of easily obtained bodily fluids using ATR-FTIR spectroscopy lends itself well to a rapid and cost-effective technology for clinical diagnostics (34,35).

The diagnostic capabilities of this approach have been explored in a range of cancers and disease (6,22,36–41). The application of ATR-FTIR serum analysis for the early detection of brain tumour provides an example of where a spectroscopic technique is distinctly addressing an unmet clinical need. Due to a combination of non-specific symptoms, pressure in the health service diagnostic pathway, expensive neuroimaging and highly invasive biopsies – the diagnosis of brain tumours is often made in the case of an emergency, when the patient will likely have a well-developed tumour. A method of early detection would greatly benefit this patient pathway, allowing screening or triage into secondary healthcare (33). Recently, we have shown that glioblastoma (GBM) patients can be correctly identified at sensitivities and specificities of 91.5% and 83% respectively, using a feature-fed support vector machine (SVM) analysis (42). This same dataset was reanalysed using a random forest (RF) approach, which resulted in an improved classification performance (92.8% and 91.5%, sensitivity and specificity respectively) (43). The classification process was iterated up to 96 times to generate a robust result, and thus small differences in sensitivity and specificity can be expected due to effectively altering the population of patients in the training and test sets.

The range of pre-processing methods for biofluid spectroscopy described in the literature are variable, with a baseline correction, normalisation step the most commonly implemented. A specific review of pre- and post-processing in ATR-FTIR has been recently published, highlighting technique specific approaches to data analysis (44) . It is evident therefore, that even in this highly specific application there is no defined pre-processing approach that has been accepted.

This study aims to optimise the spectral pre-processing approach for biofluid ATR-FTIR spectroscopy, for the purpose of improving a subsequent classification model. Although largely specific to this sample-technique scenario, the optimum pre-processing approach as defined by this thorough investigation may also be applicable to other sample types and techniques, as the approaches highlighted address many sources of variance non-discriminately. The spectral investigation of samples, such as bodily fluids, using techniques that are sensitive to differences in sample thickness and inherent heterogeneity, would be considered the best suited application of this approach.

Experimental

Data processing was conducted using the PRRFECT toolbox written the R programming language (https://github.com/Palmer-Lab/PRFFECT (45)).The aim of this programme is to provide a comprehensive, robust, and interpretable system for pre- and post-processing of spectral datasets. In its current format, this programme follows pre-processing steps commonly implemented in the field of biospectroscopy, with scope for altering the order and parameters inputted into the processing options available. An overview of the pre-processing dataflow investigated in this study can be seen in **Figure 1.**

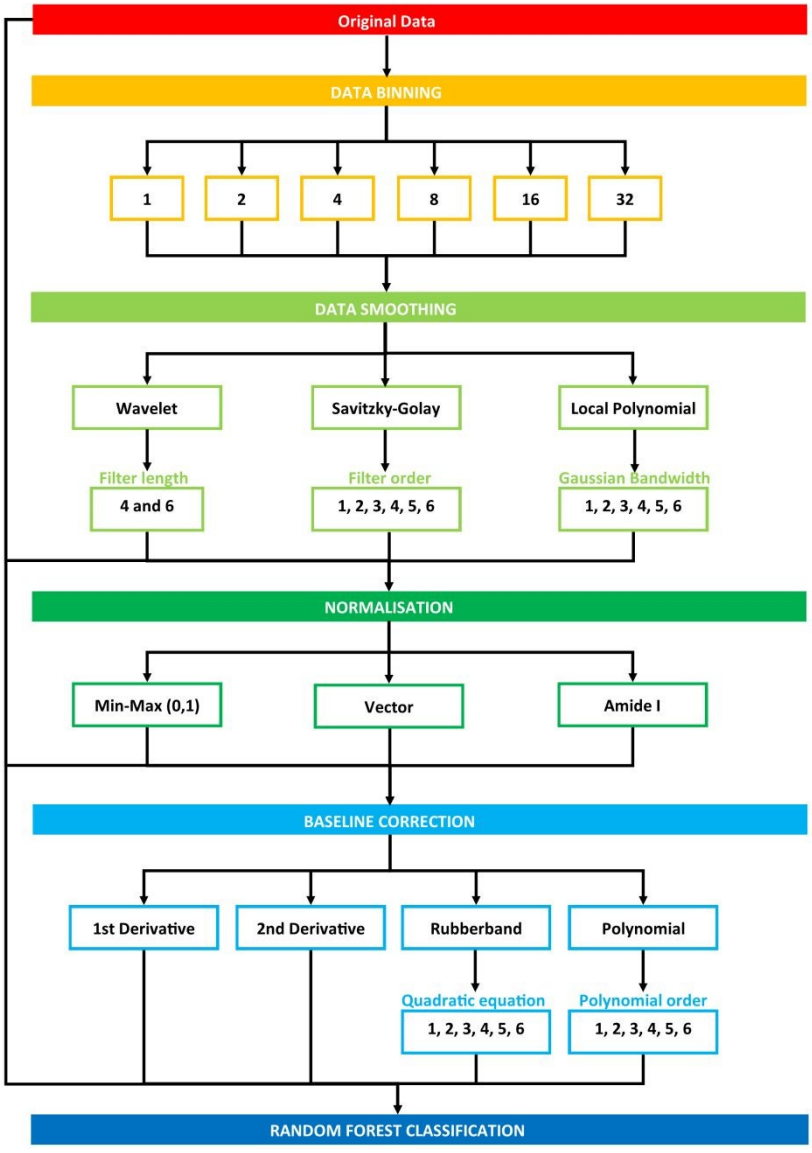

Figure 1. Schematic overview of pre-processing steps explored in this study. Numbers describe the cumulative total of pre-processing combinations.

Each pre-processing permutation from this point onwards will be described by a 6 (or 7, in the case of binning with a factor of 16 or 32) identifier, which is described by **Table 1.** The calculations were run in serial on Dual Intel Xeon X5650 2.66 GHz processors at the ARCHIE-WeSt supercomputing center located at the University of Strathclyde in Glasgow, Scotland and each run performed took approximately 2 – 3 minutes.

Dataset and Spectral Acquisition

The dataset explored in this study is that from Hands *et al.* 2016, where more detail can be found. In short, ATR-FTIR spectra obtained from 433 patients; 122 control patients and 311 brain cancer patients. Each patient was analysed in triplicate, with three spectra obtained per sample, providing a total of 3,897 spectra. Spectra were obtained at a spectral resolution of 4 cm^{-1}, and a data spacing of 1.9 cm^{-1} (42). Spectra were collected using a triangular apodization, no zero filling factor, a 5kHz collection speed, 1.28 kHz electronic low pass filter, an interferogram sample interval of 2, a sensitivity factor of 1 and an asymmetric single sided interferogram symmetry. Primarily, the binary classification between cancer and control is explored as an indicator of pre-processing optimisation. However, as a further post-comparative analysis, the effect of pre-processing on determining more difficult clinical questions, specifically differentiating between primary and metastatic brain cancer, is also explored. The research described in this paper was performed with full ethical approval (Walton Research Bank BTNW/WRTB 13_01/BTNW Application #1108).

Table 1. - Explanation of identifiers for each pre-processing combinations, using numerical values. Each column represents each respective digit of the code, with some parameters dependent upon the option (or digit) before it. For example, 212133 would refer to a binning factor of 2, a SG smoothing with a filter order of 2, a min-max normalisation, and a rubberband baseline correction with a quadration equation of 3.

Binning Factor*	Smoothing	Smoothing Parameters	Normalisation	Baseline Correction	Baseline Correction Parameters
1	0 – None	0 – None	0 – None	0 – None	0 – None
2	1 – SG Filter	1,2,3,4,5,6 Filter Order	1 – Min/Max	1 – 1st Derivative	
4	2 – Wavelet Denoise	4 or 6 Length of Filter	2 – Vector	2 – 2nd Derivative	
8	3 – Local Polynomial	1,2,3,4,5,6 Bandwidth of Gaussian	3 – Amide I	3 – Rubberband	1,2,3,4,5,6 Factor of Quadratic Equation
16				4 – Polynomial	1,2,3,4,5,6 Polynomial Degree
32					

Pre-processing Options

A concise overview of the pre-processing options explored in this study are presented; full details are described here can be found in the article by Smith *et al.,* 2018. The option that have been selected for this study encompass a range of approaches that are commonly implemented in IR studies of biological materials. A total of 3,528 possible pre-processing permutations were considered initially in this study, each generating a new dataset that is subsequently fed into the same classification algorithm (**Figure 1**). Some permutations were excluded from the study due to an insufficient number of data points that resulted in unviable spectral outputs; this was particularly significant with increased binning and larger processing parameters.

Binning

The spectral resolution of a spectrometer system will determine how well a spectrometer can distinguish between features in an IR spectrum, with a resolution of 4 and 8 cm^{-1} considered common in biological applications of IR spectroscopy. Alongside zero filling and interpolation of data points, the spectral resolution influences the data spacing of the resultant spectrum and its relative smoothness. Spectral resolution can be lost by increasing the data spacing within a spectrum, whilst also often increasing SNR. Binning is a method that finds the average of adjacent data points, thus physically reducing the number of data points in the dataset, that can also help reduce the dimensionality of a dataset. This can reduce computational burden when conducting multivariate and classification algorithms. In this instance, the binning factor describes the how many data points are averaged and replaced; for example, a bin factor of 8, represents that every 8 data points are averaged, and replaced by the mean value of absorbance intensity. A bin factor of 1 can be considered no binning.

Smoothing

Smoothing is a process that will reduce the appearance of inherent noise in the dataset; specifically retaining low frequency components of the spectrum, whilst removing high frequency noise. The significant risk of smoothing is the potential loss of information from smoothing unresolved peaks, or spectral features that may be mistaken for noise. **Savitzky-Golay (SG) filtering** is one of the most widely implemented smoothing techniques in FTIR spectroscopy due to its ability to minimise high frequency noise, whilst maintaining peak morphology. This is a local least-squares approximation of a given window size (the number of data point considered; always odd) that is fitted with a polynomial of a fixed degree. Here we describe alterations to the polynomial order rather than window size. SG filtering is often conducted in conjunction with derivative filtering, to overcome the reduction in SNR. There is a risk that incorrect tuning of these parameters will lead to peak distortions. **Wavelet denoising** could be considered an alternative approach that is suited to IR spectra (5). The discrete wavelet transformation that is used can visible improve spectral quality, particularly when the input dataset is of a high SNR (46). Also investigated was **local polynomial fitting with Gaussian weighting**, where Gaussian curves are fitted to the spectrum with varying Gaussian bandwidths (45).

Normalisation

Due to intrinsic differences between samples or within instruments – for instance varying thicknesses of a dried blood serum film – it is possible that pathlength variations can have an impact on IR spectra. This unwanted variance can be addressed by a normalisation step, which reduces intra-dataset discrepancies that can inhibit comparison of spectra (47). **Min-max scaling (0,1)** as a first option, allows the user to scale all areas of the spectrum so that are shifted in relation to each other. The minimum and maximum absorbance values are always assigned to 0 and 1 respectively, with all other data points scaled accordingly. Alternatively, **vector normalisation** works by calculating the average intensity across the spectrum, subtracting this average value from the spectrum, then dividing by the square root of the sum of the squares of all intensity values. As a result of this process, the subsequent vector norm of the spectrum is 1, effectively mean centering and scaling the spectra (48). Feature-led normalisation, such as **normalisation to Amide I band,** scales all data point in the spectrum by the maximum intensity of the given feature. The amide I peak can be found between 1600-1700 cm^{-1} and is often the most intense peak in the biological IR spectrum, thus a commonly used feature for scaling (24). By normalising to this region, one can

introduce exaggerated spectral alterations in lower wavenumber regions, and minimise differences in protein related bands (3).

Baseline correction

As a technique that fundamentally observes the interaction of light with matter, IR spectroscopy is also influenced by scattering, as well as absorbance, of radiation. Scattering is undesirable in IR spectroscopy as it can reduce reproducibility and repeatability of studies. The wavelength of MIR (2.5 – 25 μm) used in IR spectroscopy almost matches the dimensions of a biological cell, meaning that there are ideal conditions for scattering (13). Furthermore, unless a sample can be described as truly flat, without the presence of surface features such as cracks, there is the possibility of scattering (49). Baseline correction algorithms can reduce the impact of scattering artefacts in IR datasets, as well as generally reducing unwanted slopes and offsets. **Derivative filters** are arguably the most powerful processing options used for pre-processing of IR spectroscopy, as they not only reduce baseline differences, but can also resolve overlapping spectral bands based on differentiation of the spectra. By deconvolution of the broad peaks of an IR spectrum further information can be resolved visually, as well as the benefit of often improved classification performance in diagnostic studies (22). Whilst the use of first and second derivatives can be beneficial, it is recommended only on spectra with a high SNR due to the introduction of noise during this process (50). **Rubberband baseline correction** fits a convex polygonal to troughs of the spectrum, typically beneath spectral peaks and effectively pulls the baseline down at these points (5). Whilst this is more common for processing of IR spectra, **polynomial baseline correction** is perhaps more common in Raman spectroscopy analysis, where baselines are often found to be less consistent than IR spectroscopy (50). In this instance, a localised polynomial is used to estimate the baseline, requiring user input defined the polynomial order of choice (51).

Assessment of Performance

Initially, each pre-processing combination was analysed using a random forest (RF) machine learning algorithm. This approach uses a defined number of decision trees (in this case, 500 trees), which subdivide during training at each or fork, or node, using randomly chosen descriptors (wavenumbers). The number of descriptors used was determined as the square root of the total number of wavenumbers in the dataset, and a minimum of 5 nodes for each tree was chosen (45). Each of the pre-processed datasets were split into training and test sets at a ratio of 2:1 based upon patient identity, with no spectra from a single patient appearing in both the training and test sets. The process is iterated a total of 96 times, in order to produce the average results of the classification for both the training and test sets. It is possible to identify which descriptors contribute to the split at each node using the relative Gini importance. This useful aspect of RF is considered in 'Further classification'.

The output of this process is a binary classification between cancer versus non-cancer (and afterwards metastatic versus GBM) with the following metrics; prediction accuracy (PAC), sensitivity, specificity, Matthew's correlation coefficient (MCC), positive predictive value (PPV) and negative predictive value (NPV). A description of each of these metrics can be found in the following articles (43,45). There are also corresponding standard error values for each metric. The model is iterated 96 times in order to ensure the population of the training and test set is changed at each iteration, providing results more representative of the total patient population. As such, there is less opportunity for bias in the test set. To encompass all

measures of performance, a representative metric was created (Equation 1). This found the cumulative total of the standard error (se) for all test measures, and subtracted this value from the cumulative total performance of each measure over 96 iterations. As such, a simple method of observing overall stability of the pre-processing method, as well as overall performance can be conducted.

$$(PAC + MCC + Spec + Sens + PPV + NPV) - (PAC_{se} + MCC_{se} + Spec_{se} + Sens_{se} + PPV_{se} + NPV_{se})$$

Equation 1. Equation representing the Overall Metric

In order to visualise the overall results for each of these metrics, the performance of each combination was ranked in terms of test performance, and displayed as a line chart of decreasing efficiency. The corresponding validation dataset performance was shown for comparative purposes. Standard error bars are shown to display the variance across the 96 iterations.

Order of Pre-processing

The initial processing ordered as described, with first binning (B), followed by smoothing (S), normalisation (N) and baseline correction (C). However, as this order is somewhat debated in the literature, the impact of order was assessed. The top 12 performing pre-processed datasets were re-analysed using RF classification in a variety of orders, denoted by their 'B', 'S', 'N', 'C' label. Re-analysis of the data results in alterations to performance values previously derived. The overall performance metric, sensitivity and specificity were compared against the default 'BSNC' results ascertained through this reanalysis. To reduce the number of order options and to take advantage of data reduction, binning was kept at the first position throughout.Net percentage change was used to visualise these differences.

Further Classification

The 12 permutations that had the best diagnostic performance occording to the overall metric were recorded, as were the worst performing permutations. These best performing pre-processing combinations were than re-analysed using; (i) a RF-fed support vector machine (SVM), and (ii) a genetic algorithm (GA) fed SVM.

This was conducted to compare alternative classification approaches and to observe any relationships between specific pre-processing protocols and classifiers. A SVM was employed as a non-linear model that is known to minimise empirical error and maximise inter-class geometric margin (52). The top 30 Gini descriptors that were extracted from the orginal RF analysis were thus fed into the SVM, producing a feature-fed classification system which should focus on wavenumbers that best describe the variance in the dataset. 30 Gini descriptors were chosen due to preliminary investigation that suggested this provided the optimum performance in comparison to higher and lower values (data not shown). A GA was used as a comparison to the trial-and-error based approach described here, in order to optimise the pre-processing combination. The output of this was then also fed into an SVM. Net percentage change in the overall performance metric was used to describe the effect of alternative classifiers on overall classification.

Results and Discussion

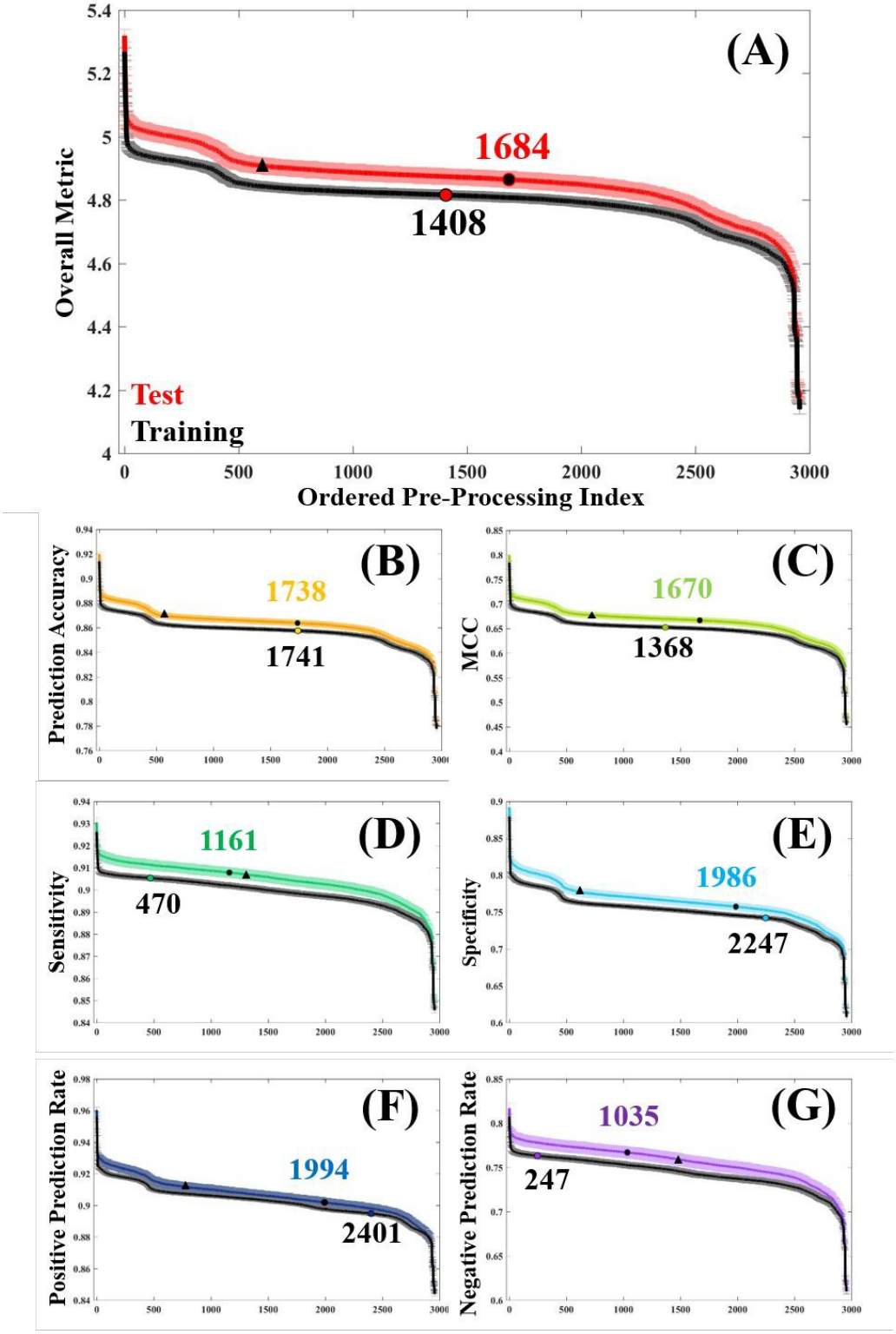

Figure 2. Performance of each pre-processing permutation within the training (validation) and test stage with regards to: (A) overall performance metric, (B) prediction accuracy, (C) Matthew's correlation co-efficient, (D) specificity, (E) sensitivity, (F) positive predictive value and (G) negative predictive value. The original, non-processed dataset is marked upon each test and validation set as a vertical marker, and a standard pre-processing option combination '112236' is marked with a triangle marker.

View Article Online
DOI: 10.1039/C8AN01384E

Table 2. Top twelve pre-processing combinations following random forest binary classification between cancer and non-cancer and the relative diagnostic performances of the test dataset.

Rank	Overall Metric		Prediction Accuracy		Matthew's CC		Sensitivity		Specificity		PPV		NPV	
1	100220	5.319 ± 0.021	100220	0.920 ± 0.002	100220	0.799 ± 0.005	100220	0.930 ± 0.002	100220	0.892 ± 0.004	100220	0.960 ± 0.002	100220	0.817 ± 0.006
2	100020	5.262 ± 0.021	100020	0.913 ± 0.002	100020	0.783 ± 0.005	100020	0.925 ± 0.002	100020	0.884 ± 0.005	100020	0.957 ± 0.002	100020	0.805 ± 0.006
3	416220	5.215 ± 0.023	416220	0.907 ± 0.002	416220	0.769 ± 0.006	416210	0.924 ± 0.002	416210	0.868 ± 0.006	100210	0.951 ± 0.002	3215310	0.802 ± 0.006
4	100210	5.192 ± 0.021	416320	0.905 ± 0.002	416320	0.762 ± 0.005	100210	0.924 ± 0.002	100210	0.857 ± 0.004	416210	0.946 ± 0.002	416320	0.801 ± 0.006
5	416210	5.172 ± 0.021	100210	0.902 ± 0.002	100210	0.757 ± 0.005	416320	0.923 ± 0.002	416320	0.847 ± 0.005	416320	0.941 ± 0.002	100210	0.799 ± 0.006
6	416310	5.122 ± 0.023	416210	0.895 ± 0.002	416310	0.742 ± 0.005	416220	0.923 ± 0.002	3215310	0.830 ± 0.005	416220	0.935 ± 0.002	416220	0.798 ± 0.005
7	100010	5.101 ± 0.023	416310	0.894 ± 0.002	416210	0.735 ± 0.005	3215310	0.919 ± 0.002	416310	0.825 ± 0.005	3215310	0.934 ± 0.002	415032	0.797 ± 0.005
8	416133	5.084 ± 0.024	100010	0.891 ± 0.002	100010	0.731 ± 0.006	416310	0.918 ± 0.002	134136	0.824 ± 0.005	416310	0.934 ± 0.002	416310	0.790 ± 0.005
9	414144	5.060 ± 0.023	416133	0.889 ± 0.002	424210	0.724 ± 0.005	134136	0.918 ± 0.002	416220	0.824 ± 0.005	213035	0.933 ± 0.002	416210	0.790 ± 0.006
10	413132	5.060 ± 0.022	414144	0.888 ± 0.002	414010	0.724 ± 0.005	132143	0.918 ± 0.002	815042	0.823 ± 0.006	412136	0.933 ± 0.003	416335	0.790 ± 0.006
11	413136	5.058 ± 0.024	413132	0.888 ± 0.002	413132	0.723 ± 0.005	100010	0.918 ± 0.002	435042	0.821 ± 0.005	235043	0.932 ± 0.002	135135	0.789 ± 0.006
12	414132	5.053 ± 0.021	413136	0.888 ± 0.002	413136	0.722 ± 0.005	234144	0.917 ± 0.002	435034	0.820 ± 0.005	114134	0.932 ± 0.002	132035	0.789 ± 0.006

Table 3. Twelve worst pre-processing combinations following random forest binary classification between cancer and non-cancer and the relative diagnostic performances of the test dataset.

Rank	Overall Metric		Prediction Accuracy		Matthew's CC		Sensitivity		Specificity		PPV		NPV	
1	3226141	4.179 ± 0.023	3226141	0.782 ± 0.002	3226144	0.465 ± 0.005	3226141	0.849 ± 0.002	3226141	0.618 ± 0.004	3226136	0.847 ± 0.003	3226134	0.615 ± 0.005
2	3232131	4.187 ± 0.020	3232131	0.783 ± 0.002	3226132	0.468 ± 0.004	3226132	0.849 ± 0.002	3226132	0.618 ± 0.004	3226135	0.847 ± 0.003	3226145	0.616 ± 0.006
3	3226134	4.195 ± 0.020	3226134	0.784 ± 0.002	3226131	0.470 ± 0.005	3226146	0.850 ± 0.002	3226146	0.619 ± 0.004	3226146	0.847 ± 0.003	3226144	0.617 ± 0.006
4	3226136	4.195 ± 0.021	3226145	0.784 ± 0.002	3226131	0.470 ± 0.005	3226136	0.850 ± 0.002	3226131	0.620 ± 0.004	3226133	0.848 ± 0.003	3226135	0.620 ± 0.005
5	3226145	4.195 ± 0.023	3226136	0.784 ± 0.002	3226145	0.470 ± 0.004	3226131	0.850 ± 0.002	3226145	0.621 ± 0.004	3226132	0.849 ± 0.003	3226131	0.621 ± 0.006
6	3226143	4.197 ± 0.020	3226133	0.785 ± 0.002	3226143	0.471 ± 0.004	3226143	0.851 ± 0.002	3226144	0.621 ± 0.004	3226141	0.849 ± 0.003	3226141	0.621 ± 0.006
7	3226133	4.198 ± 0.020	3226143	0.785 ± 0.002	3226146	0.471 ± 0.005	3226145	0.851 ± 0.002	3226136	0.621 ± 0.004	3226143	0.849 ± 0.003	3226132	0.621 ± 0.005
8	3226132	4.201 ± 0.019	3226142	0.786 ± 0.002	3226134	0.471 ± 0.004	3226133	0.851 ± 0.002	3226143	0.621 ± 0.004	3226131	0.849 ± 0.003	3226146	0.622 ± 0.005
9	3226146	4.202 ± 0.022	3226146	0.786 ± 0.002	3226135	0.472 ± 0.005	3226135	0.851 ± 0.002	3226134	0.624 ± 0.004	3226145	0.851 ± 0.003	3226136	0.622 ± 0.006
10	3226135	4.208 ± 0.021	3226135	0.786 ± 0.002	3226141	0.474 ± 0.005	3226134	0.851 ± 0.002	3226135	0.625 ± 0.004	3226142	0.853 ± 0.003	3226142	0.624 ± 0.006
11	3226142	4.214 ± 0.020	3226132	0.786 ± 0.002	3226133	0.476 ± 0.005	3226144	0.853 ± 0.002	3226133	0.626 ± 0.004	3226144	0.853 ± 0.003	3226143	0.626 ± 0.006
12	3226144	4.235 ± 0.023	3226144	0.789 ± 0.002	3226136	0.482 ± 0.005	3226142	0.854 ± 0.002	3226142	0.631 ± 0.005	3225134	0.854 ± 0.003	3226133	0.630 ± 0.006

In order to assess general trends in classification performance of the pre-processed datasets, test and training sets were ranked according to each metric (**Figure 2**). By plotting the distribution of each pre-processing permutation it is possible to identify how pre-processing effects the overall classification between brain cancer and control, in comparison to the raw data (shown as a circle on each plot). Both training and test datasets are displayed in order to identify any discrepancies and stability within the model. The raw, unprocessed dataset is highlighted as a marker for classification performance, as is a standard processing step commonly used in the literature (3).

Initially it is clear that the trend in overall performance is similar across the board, with a number of permutations yielding higher results than the vast majority, and similarly a number of permutations that have detrimental effects on the overall classification. Generally, it appears that around 2000 or so options in the central area do not drastically alter the classification. What is also apparent in **Figure 2**, is a dip in efficiency at around the 500th ranked combination. From investigating the data further (data not shown), this corresponds with the use of a min-max normalisation and subsequent rubberband or polynomial baseline

correction. The combination of these approaches may be well suited to diagnostic studies using IR spectra.

The overall metric (**Figure 2A**) encompasses this trend, and is also evident in each of the other performance measures. It is noticeable that the unprocessed dataset appears at a slightly higher rank in both the training dataset, coinciding with smaller standard error in this dataset too. This is as expected given the cross validation of the training dataset will not be as variable as the predicting test data. The sharp incline represents the best performing combinations, of which the top 12 are given in **Table 2**. Consistently, the top performing processing combination was a simple vector normalised and second derivative filtered dataset. Similarly, the second best classification result also came from the dataset only corrected using a second derivative, indicating the suitability of this processing step for the analysis of FTIR data. As this method removes baseline effects, has an in-built smoothing SG step, and also has the ability to resolve spectral features, it is a simple yet powerful approach for diagnostic applications. The minimal number of steps in these approaches could also be considered preferable (15).

Below the two highest ranked pre-processing options, there is less uniformity across the different classification metrics (**Table 2)**. Whilst simple procedures such as first order derivation with and without a normalisation step appear spordically in this table, the majority of pre-processing permutations have multiple steps. Using PAC as an example, options ranked 3 to 12 vary quite significantly, with binning, smoothing, normalisation, and baseline corrections having a postive effect on the overall accuracy. This metric is indicative of the correct prediction of true positives and negatives, in this case predicting the presence or absence of brain cancer, and ranges from 92.0 – 88.8% in the top pre-processing approaches. Interestingly, a binning factor of 4 appears more regulary than any other binning option, representing a four-fold reduction in the number of data points within the dataset. Binning is known to improve the SNR across the spectrum, by averaging out the signal of a given number of wavenumbers. With a data spacing of every four wavenumbers, closer matched to the original spectral resolution of 4 cm^{-1}, this binning option may increase SNR without smoothing out spectral features important for classification.

In this clinical dataset, a binning step is usually associated with a smoothing procedure, with SG filtering being the most commonly chosen option. Looking at the top 12 permutations with regards to optimum MCC, SG filtering with a filter order of 6 generates the best classification. It is worth noting that the value of MCC is lower than the other metrics, with values ranging between 0.799 – 0.722 (**Table 2**). Rather than being expressed as a percentage, MCC is representative of a scale between -1 and +1; with positive values indicating a strong correlation between the observed and predicted classifications and negative values indicating a worse performance than random choice. As expected, in the test dataset the classification error is higher than in cross validation, and unprocessed spectra as a comparion differ between these two datasets (**Figure 2C**).

For the remaining classification metrics, a number of processing combinations already mentioned also perform well. For sensitivity, our ability to detect brain cancer patients in this case, ranges from 93.0 – 91.7%. Local polynomial smoothing appear to have a postiive impact on sensitivity on this dataset, as well as on the NPR. However, it appears as though pre-processing generally has a greater impact on sensitivity of the classifier, shown by a steady increase in performance from the unprocessed dataset (**Figure 2E**). In cross validation of the algorithm, this raw dataset is ranked 470th in specificity, compared to a 2247th in

sensitivity; indicating that our ability to identify true negatives, or control patients, without pre-processing is higher than our ability to detect disease patients. This may be an inherent characteristic of this classifier, also influenced by the patient population. An unbalance in patient numbers in each class may be further investigated with up- or down-sampling methods (53).

Somewhat surprisingly, data processed with a binning factor of 32 appears to perform favourably with regards to sensitivity (7th), specificity (6th), PPR (7th), and NPR (3rd). Whilst 'heavy' binning has the benefit of improved SNR in the dataset, there is also the likelihood of removing spectral information, with some spectral features broader than the 32 wavenumber spacing. The evidence for this can be seen when exploring the pre-processing permutations that contribute to the worst classification values, visualised as the steep drop in performance across **Figure 2**. Of the 12 least efficient pre-processing models, a binning factor of 32 appears in every combination (**Table 3**), as well as wavelet denoising (with a filter length of 6), min-max normalisation and a baseline correction of either rubberband or polynomial corrections (with varying parameters). It is likely that the binning aspects of these permutations is reducing spectral resolution to a point where few features are visible, and thus classification is reduced. However, in the instance where a binning factor of 32 performs well, it is coupled with a standard SG filter (filter order of 5), but also with Amide I normalisation and a first derivative filter. The latter of these processes can resolve spectral features and may account for an improvement in classification, whilst Amide I may be amplifying subtle differences between cancer and control patients.

To further explore the impact of pre-processing on classification of IR spectra, the un-processed dataset was used to split the ranked pre-processing permutations into two portions; a list of pre-processing protocols that improves classification performance compared to the raw data, and a list that reduced classification performance. The frequency that each processing option occurred to both increase or decrease the performance was recorded. **Figure 3A** displays the how frequently each binning choice occurred, and how this impacted the overall classification with regards to the overall metric. It is clear to see that when an increase in diagnostic performance was seen overall, a binning factor of 2 or 4 was more common, whilst no binning made up a total of 22%. Increasing the binning factor was more influential in decreasing the overall classification in comparison to raw spectra, with a clear shift towards 16 and 32 seen.

Normalisation looks to have less influence on the overall metric, as the frequency of each of the options appears relatively equally. Min-max and amide I normalisation contribute to improved classification more commonly than no or vector normalisation, yet both only make up 57% of the overall selections (**Figure 3B**). The parameters are all standard choices for use in pre-processing and have been used extensively in the literature. This could indicate that normalisation, in any capacity, is beneficial to diagnostic performance, regardless of the approach chosen. It is also of considerable interest that no normalisation performs well. Comparisons of smoothing and baseline correction, as well as their respective parameters are shown in **Supplementary Information**. As some steps, such as rubberband baseline correction, have multiple parameters compared to others, these graphs are not shown to avoid confusion. For smoothing, the parameters have little effect on overall performance particularly in SG filtering, which appears equally across all the ranked permutations (Supplementary Information: **Figure S1**). Local polynomial smoothing has a more positive impact on classification, although again the relative parameters have little effect. The same is

seen with baseline corrections that have tuneable parameters, namely rubberband and polynomial corrections. (Supplementary Information: **Figure S2**).

View Article Online
DOI: 10.1039/C8AN01384E

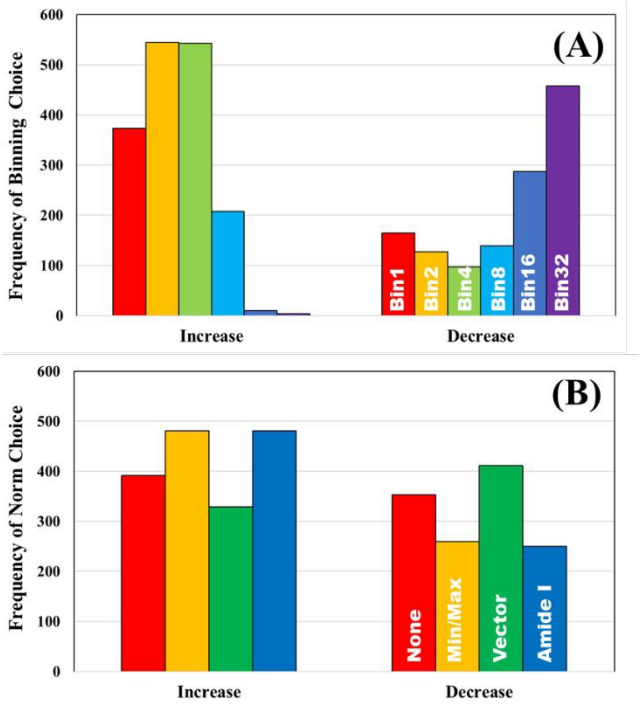

Figure 3. The frequency of pre-processing options that increase and decrease the classification performance of the unprocessed clinical dataset; (A) binning and (B) normalisation choices

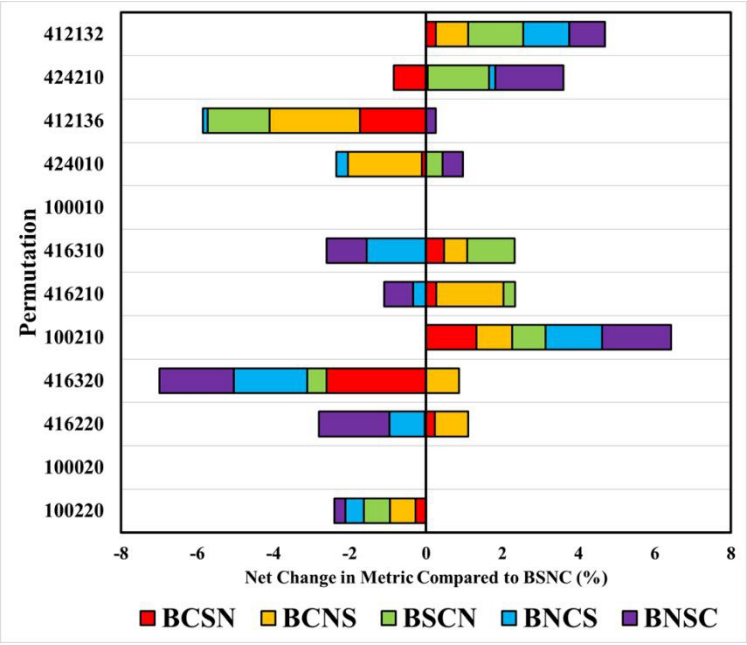

Figure 4. The comparative change (%) in diagnostic performance measured by the overall metric by altering the order of the top twelve pre-processing permutations. Each order

combination is compared to an order of Binning (B), Smoothing (S), Normalisation (N) and Baseline Correction (C).

The order in which processing steps are implemented is explored in the top twelve processing combinations. By comparing each new arrangement of these processing steps against the default order described previously (Binning (B), Smoothing (S), Normalisation (N) and Baseline correction (C)), the impact of order can be seen. It is important to note that these comparisons are made from BSCN values generated separately from the previously described analyses. This can result in slight variations in performance metrics and can suggest unexpected variance in some combinations. A full breakdown of these comparisons can be found in **Supplementary Information; Table S1-2**and **Figures S3-5.**

As expected, when only a single processing step is conducted, such as a first or second derivative (100020 and 10010), order has no impact on the overall performance. Some permutations are equivalent yet not identical; for example, BSNC, BNSC and BNSC for '100220', and other combinations where only two variables are altered. The result of this is only small changes to overall performance values.

Beginning with the highest ranked combination (100220: No binning or smoothing, vector normalised and second derivative correction), it is clear that any alteration to the default order has a negative impact on the overall classification by an average of 5% (**Figure 4**). Most significantly affected was the permutation '416320', that appears sensitive to order of implementation. Other pre-processing protocols with smoothing steps also appear to be sensitive to order, suggesting that smoothing may be better implemented earlier in the processing order.

Altering the order can also have positive impacts, shown particularly in '100210' representative of a vector normalisation and a first derivative filter. Each different arrangement improved the overall classification, illustrating that each processing protocol may require bespoke tuning with regards to order. With regards to the analysis of clinical data of biofluids, it remains clear that the top permutation of 100220 is well suited for this application, however, may lose diagnostic accuracy if re-ordered.

Throughout this study, a RF model has been used to classify patients as either cancer or non-cancer; the computational burden of such approach is low and allows rapid analysis of multiple datasets and was thus ideal for this application. However, there are a wide variety of machine learning algorithms available, which may be more appropriate for this study and yield better diagnostic results. To investigate this, two additional algorithms were explored as alternatives to a standalone RF classifier (**Table 4**). Comparing the overall metric, sensitivity and specificity of all three approaches shows that feature fed classification can improve overall performance. This is more clearly visualised in **Figure 5**, where the percentage change in diagnostic performance (compared to RF) is illustrated. The pattern described by the overall metric indicates that for each of the permutations, RF-SVM improves classification to some degree, whereas GA-SVM has a more variable response (**Figure 6A)**. It is also clear that the top performing pre-processing combinations do not vary much between the three classifiers. This could indicate stability in the dataset due to pre-processing steps revealing an optimum level of diagnostic information.

In contrast, RF-SVM and GA-SVM both dramatically increase the sensitivity of these pre-processed datasets, with only small decreases apparent (**Figure 5B**). Sensitivity was found to be high in this clinical dataset using RF classification, tentatively associated with the 3:1

imbalance of cancer to control patients. This may contribute to heightened sensitivity with feature fed classifiers, which should contain specific information for distinguishing cancer. Specificity on the other hand is more likely to be decreased when using these classifiers (**Figure 5C**). Apart from a couple of improvements in performance, on the whole RF-SVM and GA-SVM reduce the specificity of the model. Again, this could be attributed to the fact that these approaches extract disease specific information from the dataset, and thus the capabilities of identifying true negatives, or control patients, is inhibited.

Table 4. A comparison of random forest (RF), RF fed support vector machine (SVM), and genetic algorithm fed SVM classifiers with regards to overall metric, sensitivity and specificity.

	Random Forest			Random Forest - SVM			Genetic Algorithm - SVM		
Permutation	**Overall**	**Sens**	**Spec**	**Overall**	**Sens**	**Spec**	**Overall**	**Sens**	**Spec**
100220	5.319	0.930	0.892	5.315	0.936	0.869	5.300	0.934	0.868
100020	5.262	0.925	0.884	5.289	0.938	0.855	5.307	0.929	0.881
416220	5.215	0.923	0.868	5.273	0.929	0.868	5.215	0.946	0.807
416320	5.192	0.923	0.857	5.219	0.937	0.829	5.008	0.919	0.787
100210	5.172	0.924	0.847	5.209	0.947	0.803	5.147	0.935	0.806
416310	5.101	0.918	0.830	5.146	0.948	0.776	5.072	0.945	0.754
416210	5.122	0.924	0.824	5.279	0.927	0.874	4.915	0.930	0.727
100010	5.084	0.918	0.825	5.229	0.936	0.835	5.294	0.942	0.848
424210	5.023	0.913	0.811	5.083	0.937	0.775	4.770	0.906	0.723
424010	5.044	0.914	0.817	5.071	0.927	0.794	4.964	0.925	0.756
412132	5.009	0.913	0.805	5.320	0.937	0.870	5.380	0.950	0.864
412136	5.038	0.917	0.807	5.159	0.923	0.835	5.403	0.951	0.872

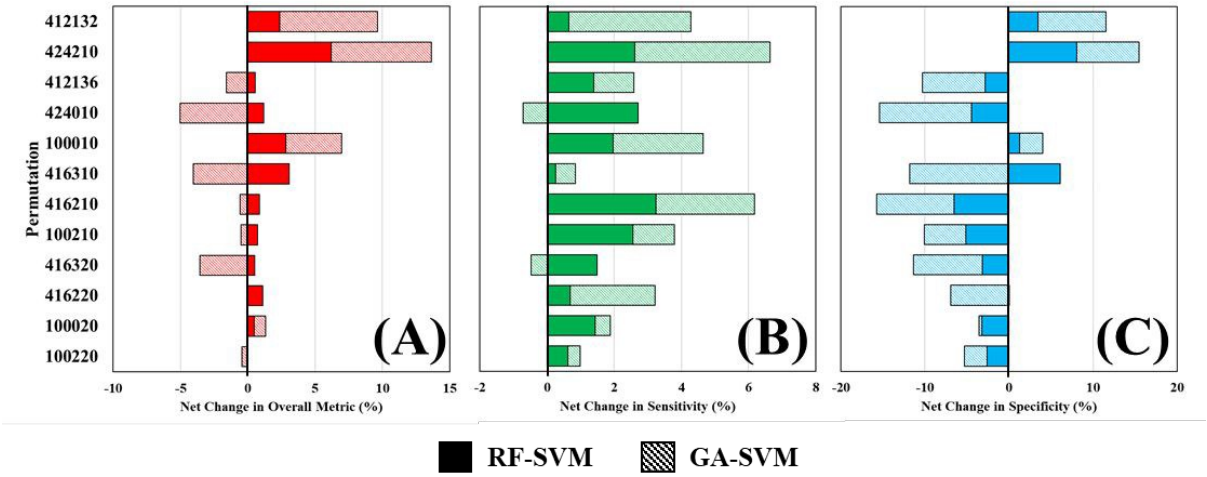

Figure 5. Percentage change in overall performance metric (x-axis), sensitivity and specificity of random forest fed support vector machine (RF-SVM) and genetic algorithm

fed SVM (GA-SVM) classifiers. The top twelve performing permutations are uses a sub-selection of the total pre-processing options (y-axis).

View Article Online
DOI: 10.1039/C8AN01384E

Conclusions

For all metrics, it is evident that there are a number of highly favourable pre-processing permutations, a larger group of that have incremental improvements in classification, and a group of unfavourable pre-processing combinations. The overall metric is a good method of viewing all the statistical patterns in the data, and mimics the morphology of all performance curves. Variability between the validation and test sets is evident throughout, however this is expected due to the nature of predictions on unknown populations.

With regards to the best pre-processing combination for discriminatory biofluid analysis, two clear permutations came out on top; a simple second order derivative filter, or a first derivative filter with a vector normalisation. Differentiation, of first or second order, has the benefit of removing baseline effects as well as revealing further spectral information by peak deconvolution.

Although no binning features highly in the top performing combinations, a binning factor of 4, also appears to be a beneficial step in pre-processing. This moderate binning factor has the benefit of dimension reduction, thus improving analysis times, as well as enhancing SNR across the spectral; all of which can have a positive influence on the diagnostic performance of the RF classifier. On the other hand, binning factors above four tend to have a detrimental effect on diagnostic performance. This approach may reduce the information contained in the spectrum and is consistent with the worst diagnostic performers. A normalisation step appears to be preferable, although of the approaches discussed in this study, none are clear frontrunners. The same can also be said for smoothing and baseline correction approaches, despite derivative and SG filters featuring prominently in the top twelve pre-processing permutations. The order in which pre-processing steps are implemented can have significant impact on overall classification. This could be dependent on specific combinations of processes, such as normalisation alongside derivative filters.

Whilst it is important to explore the range of classification algorithms available, it is important to first note the desired output of the study. In this given example, the diagnosis of brain cancer would require a high level of sensitivity, in order to ensure the false negative rate is low and no tumours are missed. The use of feature fed algorithms that have been trained on datasets with a higher proportion of positives, may provide this higher sensitivity. However, if sensitivity, or another metric is desirable, the choice of machine learning approach should be carefully considered.

References

1. Naumann D, Fabian H, Lasch P. FTIR spectroscopy of cells, tissues and body fluids. Biol Biomed Infrared Spectrosc. IOS Press Amsterdam; 2009;2:312.

2. Lasch P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. Chemom Intell Lab Syst [Internet]. 2012;117(Supplement C):100–14. Available from: http://www.sciencedirect.com/science/article/pii/S0169743912000561

3. Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 Jul 3;9:1771. Available from: http://dx.doi.org/10.1038/nprot.2014.110

4. Wang L, Mizaikoff B. Application of multivariate data-analysis techniques to biomedical diagnostics based on mid-infrared spectroscopy. Anal Bioanal Chem. Springer; 2008;391(5):1641–54.

5. Trevisan J, Angelov PP, Carmichael PL, Scott AD, Martin FL. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. Analyst [Internet]. 2012;137(14):3202–15. Available from: http://dx.doi.org/10.1039/C2AN16300D

6. 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(14):3917–26.

7. Theophilou G, Lima KMG, Martin-Hirsch PL, Stringfellow HF, Martin FL. ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue: classifying subtypes of human cancer. Analyst. Royal Society of Chemistry; 2016;141(2):585–94.

8. Stuart B. Infrared Spectroscopy. In: Kirk-Othmer Encyclopedia of Chemical Technology [Internet]. John Wiley & Sons, Inc.; 2000. Available from: http://dx.doi.org/10.1002/0471238961.0914061810151405.a01.pub2

9. Movasaghi Z, Rehman S, ur Rehman DI. Fourier transform infrared (FTIR) spectroscopy of biological tissues. Appl Spectrosc Rev. Taylor & Francis; 2008;43(2):134–79.

10. Singh B, Gautam R, Kumar S, Kumar BNV, Nongthomba U, Nandi D, et al. Application of vibrational microspectroscopy to biology and medicine. Curr Sci. 2012;102(2):232–44.

11. Afseth NK, Kohler A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemom Intell Lab Syst. Elsevier; 2012;117:92–9.

12. Mohlenhoff B, Romeo M, Diem M, Wood BR. Mie-Type Scattering and Non-Beer-Lambert Absorption Behavior of Human Cells in Infrared Microspectroscopy. Biophys J [Internet]. 2005;88(5):3635–40. Available from: internal-pdf://157.156.226.121/RSC_ (5).ris

13. Bassan P, Byrne HJ, Bonnier F, Lee J, Dumas P, Gardner P. Resonant Mie scattering in infrared spectroscopy of biological materials–understanding the “dispersion

artefact." Analyst. 2009;134(8):1586–93.

14. Rinnan Å. Pre-processing in vibrational spectroscopy–when, why and how. Anal Methods. 2014;6(18):7124–9.

15. Gerretzen J, Szymańska E, Jansen JJ, Bart J, van Manen H-J, van den Heuvel ER, et al. Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments. Anal Chem [Internet]. 2015;87(24):12096–103. Available from: internal-pdf://107.58.44.223/RSC_ (6).ris

16. Engel J, Gerretzen J, Szymańska E, Jansen JJ, Downey G, Blanchet L, et al. Breaking with trends in pre-processing? TrAC Trends Anal Chem [Internet]. 2013;50:96–106. Available from: internal-pdf://213.82.153.163/RSC_.bib

17. Baker MJ, Hussain SR, Lovergne L, Untereiner V, Hughes C, Lukaszewski RA, et al. Developing and understanding biofluid vibrational spectroscopy: a critical review. Chem Soc Rev. Royal Society of Chemistry; 2016;45(7):1803–18.

18. Aruga R. Closure of analytical chemical data and multivariate classification. Talanta [Internet]. 1998;47(4):1053–61. Available from: http://www.sciencedirect.com/science/article/pii/S003991409800126X

19. Rinnan Å, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal Chem. Elsevier; 2009;28(10):1201–22.

20. Heraud P, Wood BR, Beardall J, McNaughton D. Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalgal cells. J Chemom. 2006;20(5):193–7.

21. Butler HJ, McAinsh MR, Adams S, Martin FL. Application of vibrational spectroscopy techniques to non-destructively monitor plant health and development. Anal Methods. 2015;7(10):4059–70.

22. Ollesch J, Drees SL, Heise HM, Behrens T, Brüning T, Gerwert K. FTIR spectroscopy of biofluids revisited: an automated approach to spectral biomarker identification. Analyst. Royal Society of Chemistry; 2013;138(14):4092–102.

23. Vogt F, Tacke M. Fast principal component analysis of large data sets. Chemom Intell Lab Syst [Internet]. 2001;59(1):1–18. Available from: http://www.sciencedirect.com/science/article/pii/S0169743901001307

24. Martin FL, Kelly JG, Llabjani V, Martin-Hirsch PL, Patel II, Trevisan J, et al. Distinguishing cell types or populations based on the computational analysis of their infrared spectra. Nat Protoc. 2010;5(11):1748.

25. The International Society of Clinical Spectroscopy. Objective 5 | CLIRSPEC Network [Internet]. 2018 [cited 2018 Jul 17]. Available from: https://clirspec.org/uk-network/objectives/objective-5/

26. Bocklitz T, Walter A, Hartmann K, Rösch P, Popp J. How to pre-process Raman spectra for reliable and stable models? Anal Chim Acta. 2011;704(1):47–56.

27. Jarvis RM, Goodacre R. Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. Bioinformatics [Internet]. 2005;21(7):860–8. Available from: http://dx.doi.org/10.1093/bioinformatics/bti102

28. Byrne HJ, Knief P, Keating ME, Bonnier F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. Chem Soc Rev.

2016;45(7):1865–78.

29. Preisner O, Lopes JA, Guiomar R, Machado J, Menezes JC. Fourier transform infrared (FT-IR) spectroscopy in bacteriology: towards a reference method for bacteria discrimination. Anal Bioanal Chem [Internet]. 2007;387(5):1739–48. Available from: internal-pdf://222.249.218.220/RSC_ (4).ris

30. Ly E, Piot O, Wolthuis R, Durlach A, Bernard P, Manfait M. Combination of FTIR spectral imaging and chemometrics for tumour detection from paraffin-embedded biopsies. Analyst. 2008;133(2):197–205.

31. Bassan P, Sachdeva A, Kohler A, Hughes C, Henderson A, Boyle J, et al. FTIR microscopy of biological cells and tissue: data analysis using resonant Mie scattering (RMieS) EMSC algorithm. Analyst. Royal Society of Chemistry; 2012;137(6):1370–7.

32. 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. Analyst [Internet]. The Royal Society of Chemistry; 2018;143(8):1735–57. Available from: http://dx.doi.org/10.1039/C7AN01871A

33. 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 [Internet]. 2018 May 1;8(5). Available from: http://bmjopen.bmj.com/content/8/5/e017593.abstract

34. Mitchell AL, Gajjar KB, Theophilou G, Martin FL, Martin-Hirsch PL. Vibrational spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory to a clinical setting. J Biophotonics. Wiley Online Library; 2014;7(3-4):153–65.

35. Baker MJ. Photonic biofluid diagnostics. J Biophotonics. Wiley Online Library; 2014;7(3-4):151–2.

36. Paraskevaidi M, Morais CLM, Lima KMG, Snowden JS, Saxon JA, Richardson AMT, et al. Differential diagnosis of Alzheimer's disease using spectrochemical analysis of blood. Proc Natl Acad Sci. 2017;201701517.

37. Goodacre R, Baker MJ, Graham D, Schultz ZD, Diem M, Marques MP, et al. Biofluids and other techniques: general discussion. Faraday Discuss. Royal Society of Chemistry; 2016;187:575–601.

38. Menze BH, Petrich W, Hamprecht FA. Multivariate feature selection and hierarchical classification for infrared spectroscopy: serum-based detection of bovine spongiform encephalopathy. Anal Bioanal Chem [Internet]. 2007 Mar;387(5):1801–7. Available from: https://doi.org/10.1007/s00216-006-1070-5

39. Scaglia E, Sockalingum GD, Schmitt J, Gobinet C, Schneider N, Manfait M, et al. Noninvasive assessment of hepatic fibrosis in patients with chronic hepatitis C using serum Fourier transform infrared spectroscopy. Anal Bioanal Chem. Springer; 2011;401(9):2919.

40. 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. J Biophotonics. Wiley Online Library; 2014;7(3-4):210–21.

41. Bonnier F, Blasco H, Wasselet C, Brachet G, Respaud R, Carvalho LFCS, et al. Ultra-filtration of human serum for improved quantitative analysis of low molecular weight biomarkers using ATR-IR spectroscopy. Analyst. Royal Society of Chemistry;

2017;142(8):1285–98.

42. 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. J Neurooncol. Springer; 2016;127(3):463–72.

43. 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. Royal Society of Chemistry; 2016;141(12):3668–78.

44. Lee LC, Liong C-Y, Jemain AA. A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum. Chemom Intell Lab Syst [Internet]. 2017;163(Supplement C):64–75. Available from: http://www.sciencedirect.com/science/article/pii/S0169743916305500

45. Smith BR, Baker MJ, Palmer DS. PRFFECT: A versatile tool for spectroscopists. Chemom Intell Lab Syst. Elsevier; 2018;172:33–42.

46. Alsberg BK, Woodward AM, Winson MK, Rowland J, Kell DB. Wavelet denoising of infrared spectra. Analyst. Royal Society of Chemistry; 1997;122(7):645–52.

47. Randolph TW. Scale-based normalization of spectral data. Cancer Biomarkers. IOS Press; 2006;2(3–4):135–44.

48. Wartewig S. IR and Raman spectroscopy: fundamental processing. John Wiley & Sons; 2006.

49. Hughes C, Brown M, Clemens G, Henderson A, Monjardez G, Clarke NW, et al. Assessing the challenges of Fourier transform infrared spectroscopic analysis of blood serum. J Biophotonics. Wiley Online Library; 2014;7(3-4):180–8.

50. Butler HJ, Ashton L, Bird B, Cinque G, Curtis K, Dorney J, et al. Using Raman spectroscopy to characterize biological materials. Nat Protoc. Nature Publishing Group; 2016;11(4):664–87.

51. Lieber CA, Mahadevan-Jansen A. Automated method for subtraction of fluorescence from biological Raman spectra. Appl Spectrosc. Society for Applied Spectroscopy; 2003;57(11):1363–7.

52. Devos O, Downey G, Duponchel L. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. Food Chem [Internet]. 2014;148:124–30. Available from: http://www.sciencedirect.com/science/article/pii/S0308814613014520

53. Ganganwar V. An overview of classification algorithms for imbalanced datasets. Int J Emerg Technol Adv Eng. Citeseer; 2012;2(4):42–7.