


**FULL ARTICLE**

# Rapid analysis of disease state in liquid human serum combining infrared spectroscopy and “digital drying”

Alexandra Sala<sup>1</sup>  | Katie E. Spalding<sup>1</sup> | Katherine M. Ashton<sup>2</sup> | Ruth Board<sup>3</sup> | Holly J. Butler<sup>1</sup> | Timothy P. Dawson<sup>2</sup> | Dean A. Harris<sup>4</sup> | Caryn S. Hughes<sup>1</sup> | Cerys A. Jenkins<sup>5</sup> | Michael D. Jenkinson<sup>6</sup> | David S. Palmer<sup>7</sup> | Benjamin R. Smith<sup>1,7</sup> | Catherine A. Thornton<sup>8</sup> | Matthew J. Baker<sup>1\*</sup>

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

<sup>2</sup>Neuropathology, Lancashire Teaching Hospitals NHS Trust, Royal Preston Hospital, Preston, UK

<sup>3</sup>Rosemere Cancer Centre, Lancashire Teaching Hospitals NHS Trust, Royal Preston Hospital, Preston, UK

<sup>4</sup>Swansea Bay University Local Health Board, Singleton Hospital, Swansea, UK

<sup>5</sup>Department of Physics, College of Science, Swansea University, Swansea, UK

<sup>6</sup>University of Liverpool & The Walton Centre NHS Foundation Trust, Liverpool, UK

<sup>7</sup>WestCHEM, Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, UK

<sup>8</sup>Institute of Life Science, Swansea University Medical School, Swansea University, Swansea, UK

**\*Correspondence**

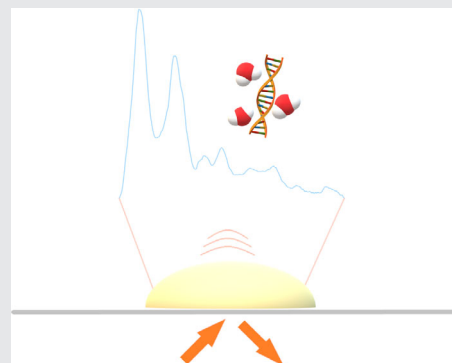
Matthew J. Baker, WestCHEM,  
Department of Pure and Applied  
Chemistry, Technology and Innovation  
Centre, University of Strathclyde, 99  
George Street, G1 1RD, Glasgow, UK.  
Email: matthew.baker@strath.ac.uk

**Funding information**

Engineering and Physical Sciences  
Research Council, Grant/Award Number:  
EP/L505080/1

**Abstract**

In recent years, the diagnosis of brain tumors has been investigated with attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectroscopy on dried human serum samples to eliminate spectral interferences of the water component, with promising results. This research evaluates ATR-FTIR on both liquid and air-dried samples to investigate “digital drying” as an alternative approach for the analysis of spectra obtained from liquid samples. Digital drying approaches, consisting of water subtraction and least-squares method, have demonstrated a greater random forest (RF) classification performance than the air-dried spectra approach when discriminating cancer vs control samples,



**Abbreviations:** ATR, attenuated total reflection; ATR-FTIR, attenuated total reflection-Fourier transform infrared; DFIR, discrete frequency infrared; EMSC, extended multiplicative signal correction; FTIR, Fourier transform infrared; GBM, glioblastoma multiforme; IRE, internal reflection element; LS, least-squares method; QCL, quantum cascade laser; QCL-IR, quantum cascade laser-infrared; RF, random forest; SD, standard deviation; SMOTE, synthetic minority over-sampling technique; WS, water subtraction.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Journal of Biophotonics* published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

reaching sensitivity values higher than 93.0% and specificity values higher than 83.0%. Moreover, quantum cascade laser infrared (QCL-IR) based spectroscopic imaging is utilized on liquid samples to assess the implications of a deep-penetration light source on disease classification. The RF classification of QCL-IR data has provided sensitivity and specificity amounting to 85.1% and 75.3% respectively.

#### KEYWORDS

ATR-FTIR, cancer, digital drying, infrared spectroscopy, serum

## 1 | INTRODUCTION

In 2018, the global cancer burden was reported to amount to over 18 million new cases with almost 10 million deaths; worldwide, during their lifetime, one man in five and one woman in six develop cancer, which is lethal for one man in eight and one woman in eleven [1]. Health services are struggling to cope with the large number of suspected cancer patients referred for further examination, often leading to late diagnosis and treatment. Despite the development of novel therapies, patients are not always able to benefit from them in time [2]. Accurate and early diagnostic tools are urgently required for use in the clinical setting.

Brain cancer is a primary example of this diagnostic problem. Only 3.7% of patients referred for medical imaging have been found to have major abnormal structural lesions indicative of cancer [3]. An early and accurate method of disease detection would aid health professionals in patient prioritization during the referral process, leading to a reduced demand on medical imaging resources and subsequent economic benefits to the health care system [4]. For patients, a diagnosis time improvement would potentially lead to higher life quality and expectancy [5].

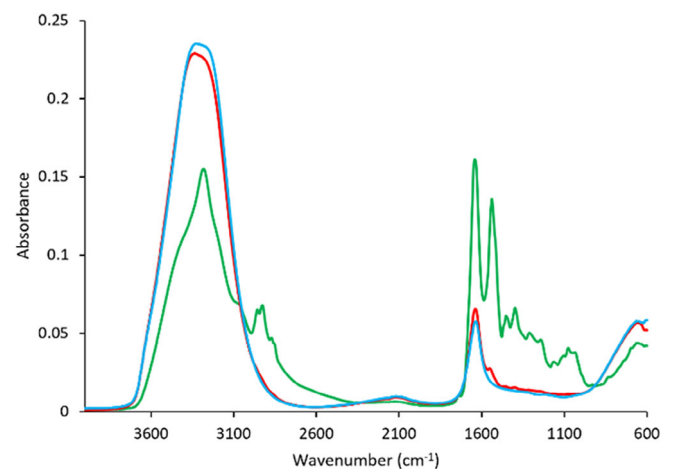
Human blood serum and other biofluids are easily obtained following a minimally invasive procedure, providing the possibility of repeated sampling. They contain several biomolecular components, which are useful to determine the disease state of the patient [6, 7]. Extensive research has been carried out to explore the potential of biological fluids, including early detection of cancer using IR spectroscopy [8–10].

Attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectroscopy represents a well-established technique to analyze biofluids, and a potential solution to counteract the issues related to a late diagnosis [11]. The acquisition of a unique spectrum, serving as a spectral fingerprint, allows the discrimination between various biological matrices and has been shown to detect disease signals within biofluids [12]. This instrumental analysis can be complementary to other techniques currently employed in

the disease classification process, for instance histopathology [13, 14]. Hands *et al.* showed that ATR-FTIR spectroscopy can be used to discriminate between healthy and diseased patients, as well as decipher disease severity, during the investigation of different types of brain tumors [15].

Although the great efficiency proved by ATR-FTIR spectroscopy, the impact of water on the biological spectra represents a major issue in the development of many studies [16–18]. The 1700 to 1500  $\text{cm}^{-1}$  range of a biological serum spectrum has been shown to be a key region for the discrimination of cancer vs non-cancer state; it is given by the proteins associated with amide I and amide II bands, which lie in this region. In liquid spectra, their vibrations are overlapped by the O-H bending vibrational mode (Figure 1) [19]. A solution could be to air-dry the sample on top of the internal reflection element (IRE) in order to acquire spectra with defined biological features.

The air-drying process requires a minimum of 8 minutes per sample, depending upon sample volume and ambient conditions [15]. Moreover, the cleaning of the IRE between each sample, as well as the background spectrum



**FIGURE 1** Overlay of an air dried, a liquid and a water spectrum, highlighting the water contribution in the liquid spectrum. (Blue: water; red: liquid serum; green: air-dried serum)

collection, need to be considered and limit the technique. In the study by Hands *et al.* in 2016, a dataset of 433 patients was analyzed using the air-drying process; each patient's serum sample took over 45 minutes to be analyzed in triplicate, leading to more than 8 weeks of analysis of the complete dataset [20]. A patient serum sample analyzed in liquid form requires less than half of the time, decreasing the time of analysis to less than 4 weeks. Spectroscopic bio-fluid analysis has the grounds to become a powerful technique for translation; however, the methodology needs to be improved to achieve a high throughput, while also maintaining comparable sensitivity and specificity demonstrated in previous studies [8, 21, 22].

To optimize the discrimination, the discrete frequency infrared (DFIR) spectroscopic analysis can also be used, with tunable narrow-bandwidth quantum cascade lasers (QCLs) as a source. DFIR has already shown to be successful in rapid diagnostics through the analysis of dried serum spots [23]. The QCLs enable an enhanced classification via a more brilliant source than the thermal Global sources used in FTIR imaging [24]. Furthermore, once the significant frequency range for discrimination has been detected, this technique allows the narrowing of the data collection to those discrete frequencies [25].

In this paper, we pursue the possibility of using serum samples in liquid form for disease discrimination of non-cancerous patients from brain and metastatic cancer patients. We present two different approaches to overcome the water interference on biological spectra and hence achieve a high-throughput liquid analysis. The first approach consists of removing the drying step by performing a later “digital drying” process on ATR-FTIR collected spectra, representing a strategy to clear the water spectral component via computer software. Digital dewaxing has already employed similar techniques to remove the paraffin contribution from tissue samples, such as extended multiplicative signal correction (EMSC) and a combination of independent component analysis and nonnegatively constrained least squares [26]. The second approach includes the use of DFIR spectroscopic imaging with a QCL microscopy, by selecting discrete frequencies to achieve a rapid and efficient analysis. These approaches could provide a disease indication within a few minutes and transform the diagnostic and patient care environment, leading to increased survival rates and quality of life, alongside health and economic benefits [4].

## 2 | MATERIALS AND METHODS

### 2.1 | Samples collection and preparation

Blood samples were acquired from 150 patients; (a) Grade IV Glioblastoma multiforme (GBM) ( $n = 50$ ), (b)

metastatic brain cancer (secondary brain cancer from multiple sites within the body) ( $n = 50$ ) and (c) non-cancer controls ( $n = 50$ ). 43% females and 57% males, across an age range from 16 to 85 years old. In detail, cancer samples were found in the age range from 19 to 85 years old and included a total of 55% males and 45% females; control samples were found in the age range from 16 to 71 years old, with a total of 60% males and 40% females. Samples were age and gender matched to the best of the cohort possibilities. The research was granted full ethical approval (Walton Research Bank BTNW/WRTB 13\_01/BTNW Application #1108).

The whole blood samples were collected in BD Vacutainer<sup>®</sup> SST tubes and left to stand upright for 30 minutes to allow clotting. They were then centrifuged at 2200 g for 15 minutes at room temperature to allow serum separation. Subsequently, the serum samples were aliquoted and stored in separate vials at  $-80^{\circ}\text{C}$  until required.

All serum samples were fully thawed at room temperature before spectral collection and the sample set was randomized prior to analysis. For every patient, 1  $\mu\text{L}$  serum aliquot was pipetted onto the IRE using a calibrated pipette (Gilson, United Kingdom). Following spectral collection, Virkon disinfectant (FisherScientific, United Kingdom) followed by 99.5% ethanol (Thermo Scientific, United Kingdom) were used to clean the crystal prior to the next analysis.

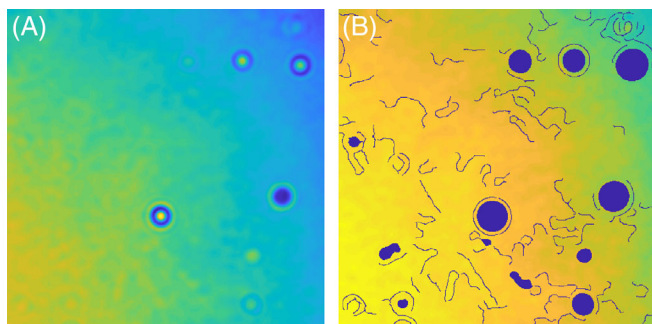
### 2.2 | Data collection with ATR-FTIR spectrometer

ATR-FTIR spectra were collected using a Cary-600 series FTIR spectrometer (Agilent Technologies, Santa Clara, California) with a MIRacle<sup>™</sup> single reflection ATR configured with a diamond (Di) IRE plate (PIKE Technologies, Fitchburg, Wisconsin). 32 co-added scans, covering a wavenumber range of 4000 to 600  $\text{cm}^{-1}$ , were combined to produce the spectrum, using a spectral resolution of 4  $\text{cm}^{-1}$ . A background spectrum of the ambient conditions was automatically subtracted by the Resolution Pro software (Agilent Technologies) to produce each sample spectrum.

For every patient, three measurements of the same 1  $\mu\text{L}$  serum aliquot were taken immediately following deposition. After the optimal drying time of 8 minutes, determined from previous drying experiments [16], the same 1  $\mu\text{L}$  serum aliquot was analyzed in triplicate again.

### 2.3 | Data collection with QCL spectroscopic microscopy

Liquid transmission measurements were performed using a Specac<sup>™</sup> FTIR micro-compression cell (Specac Ltd,



**FIGURE 2** A, QCL raw image of a sample. B, QCL pre-processed image of a sample, showing removed sample artifacts. QCL, quantum cascade laser

Kent, United Kingdom). The cell comprised of an o-ring component, followed by a 1 mm thick circular  $\text{CaF}_2$  substrate with a diameter of 14 mm, on which 10  $\mu\text{L}$  of serum were placed and compressed with a second  $\text{CaF}_2$  substrate and another o-ring. For background measurements, a single 2 mm thick  $\text{CaF}_2$  substrate was used. Once assembled, the cell was placed in the spectroscopic microscope for data acquisition.

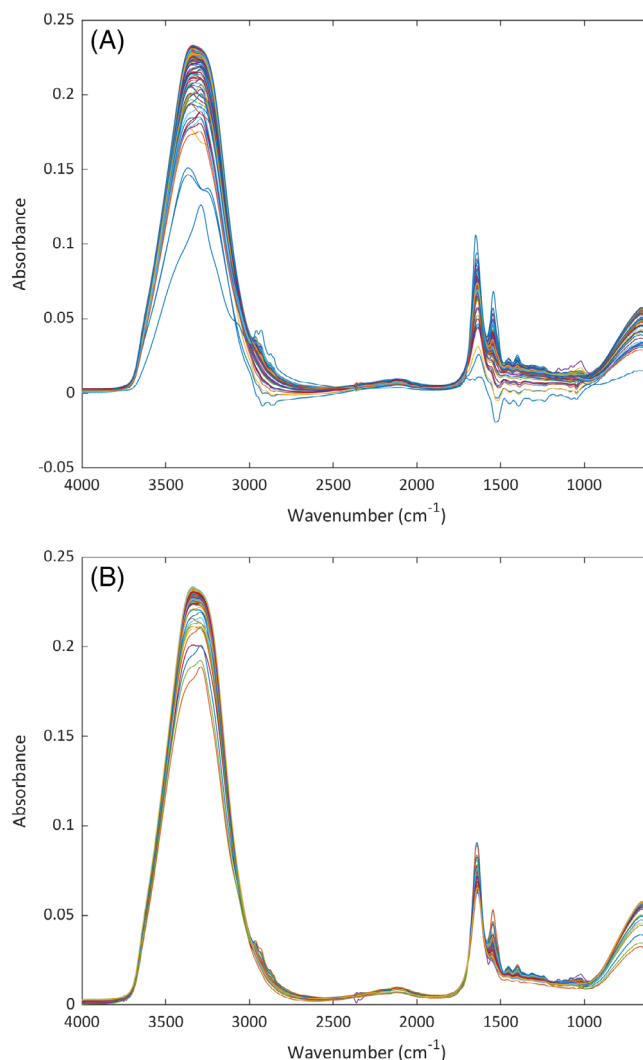
Data, in the form of an image (Figure 2A) were acquired using a QCL Spero<sup>TM</sup> microscope (Daylight Solutions Inc., San Diego, CA, USA) in the 1800 to 948  $\text{cm}^{-1}$  range with a data spacing of 4  $\text{cm}^{-1}$ . A single frame was collected with a  $\times 4$  magnification objective, a field of view of 2 mm  $\times$  2 mm, a numerical aperture of 0.15 and a pixel size of 4.25  $\times$  4.25. 230 400 spectra were obtained from a single image.

## 2.4 | Pre-processing of ATR-FTIR spectra of air-dried samples

The spectral range was reduced to the fingerprint region (1800-1000  $\text{cm}^{-1}$ ), and baseline correction (Savitzky-Golay filter: second derivative with a window size of five points) and vector normalization were applied to all the 450 spectra using PRFFECT [27] (v2); an in-house developed software package running in the R environment (R Studio software).

## 2.5 | Pre-processing of ATR-FTIR spectra of liquid samples

Using PRFFECT v2, the spectral range was reduced to the 1800 to 1000  $\text{cm}^{-1}$  region, and baseline correction (second derivative with a window size of five points) and vector normalization were applied to all the spectra, as in the air-dried dataset. The drying process of the serum is



**FIGURE 3** A, ATR-FTIR 450 raw liquid spectra; first, second and third collection for each sample. B, ATR-FTIR 150 raw liquid spectra; first collection only for each sample. ATR-FTIR, attenuated total reflection-Fourier transform infrared

affected by temperature and humidity [15, 28], and several spectra (second and third replicates of the same sample) presented visible signs of the drying process (Figure 3A); therefore, the dataset was reduced to 150 spectra (first replicate only) to ensure the dataset included liquid samples only (Figure 3B).

## 2.6 | Pre-processing of QCL spectral images of liquid samples

A different preprocessing was applied to this dataset to guarantee the best classification outcomes, and therefore obtain an equitable comparison between the techniques.

An image quality test was performed with Matlab<sup>®</sup> software (The Mathworks) prior to extracting the mean



spectra of QCL data collected. First, the spectral range of the QCL image (Figure 2A) was reduced to 1800 to 1000  $\text{cm}^{-1}$ , then the image was contrast enhanced for additional structural resolution of image artefacts boundaries (eg, air bubbles and fibers). Subsequently, using the “edge” and “imfill” functions, a binary mask was created and overlaid onto the original hyperspectral data to remove the undesirable spectra (Figure 2B). Spectra were then averaged and smoothed by a three-points moving average filter. Before performing classification, a first derivative with a window size of nine points was also applied to all the 150 spectra using PRFFECT v2.

## 2.7 | Digital drying of ATR-FTIR spectra of liquid samples

First, a water spectral reference was collected onto the same ATR-FTIR spectrometer used for the liquid serum analysis. One microliter of MilliQ water—collected from the MilliQ water unit in the University of Strathclyde, Glasgow (United Kingdom)—was deposited onto the center of the Diamond IRE after air-background collection, and three spectra were then recorded and averaged to a single spectrum. Subsequently, three approaches were investigated in order to remove the water spectral component in liquid spectra via computer software; (a) a direct water subtraction (WS) using Matlab<sup>®</sup>; the water spectral reference was subtracted from each spectrum of the liquid dataset. (b) The application of the least-squares method (LS) using Matlab<sup>®</sup>. This consisted of the “lscov” function, which returned the coefficients of the least-squares solution to the linear systems of each liquid spectra and the water spectral reference. These coefficients were then multiplied by the water spectral reference to produce the water spectral contribution, which was subsequently subtracted from each liquid spectrum. (c) The EMSC algorithm [29] application, using PRFFECT v2. This algorithm allows the correction of spectra for specific features using an input reference; the water spectral reference was used to correct each liquid spectrum.

The three datasets were then pre-processed using PRFFECT v2. The spectral range was reduced to 1800 to 1000  $\text{cm}^{-1}$ , and vector normalization and a second derivative with a window size of five points were applied.

## 2.8 | Data analysis: Random forest classification

The machine learning package randomForest by Liaw and Wiener, was used as method of classification [30,

31]. The algorithm allows the detection of features associated with input classes, presenting easily interpretable results. The performance of the model can be determined from the random forest (RF) statistical outputs, while the Gini plot highlights features responsible for the results and their importance in the discrimination [32].

Using PRFFECT v2, each dataset was split into 70% training set and 30% test set based upon patient population, to ensure distinct sets; a failure in separating the sets may result in a biased model, which would not allow a proper classification of yet unseen data. The training set was used to train the model by pairing the input with the known, expected output; also, a 5-fold cross-validation was used to validate the training set before each classification. This was then used as a measure of the model training performance. *n*tree (number of trees) and *nodesize* (minimum size of the node) default settings were maintained to 500 and 1 respectively, while *m*try (number of variables selected for splitting at each node) was set to 30; nevertheless, it has previously been seen that the RF algorithm is relatively insensitive to its tuning parameters [33]. The sampling, which consists in training each RF model on different samples, was reiterated 100 times to ensure that the reported results were not biased to a certain patient population and the variance within the sample dataset was fully encompassed. Due to class imbalance when distinguishing between cancer (100 patients) vs non-cancer (50 patients), synthetic minority over-sampling technique (SMOTE) [34] was used throughout the classifications to eliminate any eventual bias on statistical metrics within the model [35].

## 3 | RESULTS AND DISCUSSION

### 3.1 | ATR-FTIR of air-dried and liquid samples

Optimal results were achieved with the RF classification of air-dried spectra. The test set (TS) reached a sensitivity of 91.8% and a specificity of 83.2%. The classification model of liquid spectra showed lower results; sensitivity and specificity amounted to 89.9% and 81.2% respectively. Although SMOTE was used to overcome class imbalance, both precision and F1 score were reported along with accuracy of the test sets (Table 1). This was done as it has been previously reported that accuracy is not always a reliable indicator of the classifier performance in case of an uneven class distribution [35]. Notwithstanding the high standard deviation (SD) values of both TS specificity outputs, the models produced high precision and F1 score values above 90%, suggesting stability and reliability of both models. Moreover, the cross-validation sets

**TABLE 1** Statistical outputs of RF classification of air-dried and liquid samples

		Air-dried (%)		Liquid (%)	
		Value	SD	Value	SD
CV	Sensitivity	91.7	2.2	88.9	2.7
	Specificity	81.8	3.4	79.4	4.5
	Accuracy	88.4	2.3	85.7	2.5
	Precision	91.0	1.6	89.6	2.1
	F1 Score	91.3	1.7	89.2	1.9
TS	Sensitivity	<b>91.8</b>	4.7	<b>89.9</b>	5.3
	Specificity	<b>83.2</b>	7.7	<b>81.2</b>	10.9
	Accuracy	88.9	3.4	87.0	4.4
	Precision	<b>91.8</b>	3.4	<b>90.8</b>	4.7
	F1 Score	<b>91.7</b>	2.6	<b>90.2</b>	3.3

Abbreviations: CV, cross-validation set; RF, random forest; TS, test set.

(CV) results were also reported (Table 1) to provide an estimate of the training sets performance [32].

The RF (mean decrease Gini) importance plots (Figures S1 and S2) identified the main spectral features involved in the classification. Proteins, lipids, nucleic acids and carbohydrates are the biomacromolecules responsible for class discrimination in both datasets. 20 higher RF importance spectral features to the corresponding IR molecular vibrations of both air-dried and liquid spectra were assigned in Table 2 [4, 14, 35–38]. Only the air-dried dataset classification used the amide II features (1600–1500  $\text{cm}^{-1}$ ) for discrimination.

The peaks assigned to the carbohydrates in the region 1200 to 1000  $\text{cm}^{-1}$  are typical of the furanose structures, which can also be found in nucleic acids [36, 38].

### 3.2 | Digital drying of liquid samples spectra

Results of three different approaches were investigated to evaluate their potential in removing the water spectral component and, subsequently, their performance in RF classification. Figure 4 shows the spectral fingerprint regions of liquid spectra after using the three different approaches.

Both WS (Figure 4A) and LS (Figure 4B) approaches highlighted enhanced peak shapes in the 1500 to 1000  $\text{cm}^{-1}$  region and drastically reduced the absorbance intensity of all peaks, when compared to the liquid spectra (Figure 5B). However, WS significantly affected the amide I and amide II peak absorbance ratio, notably

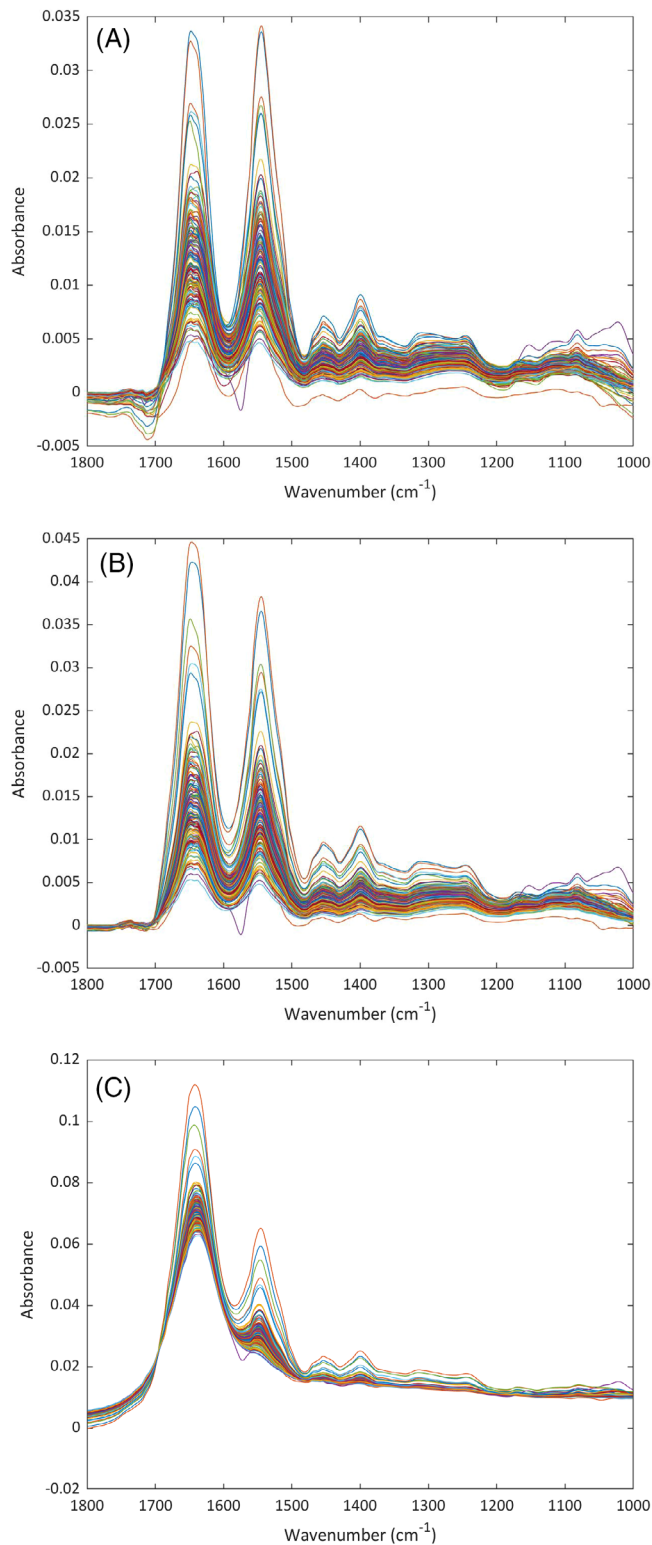
**TABLE 2** Assignment of the main spectral features in the RF importance plots for air-dried and liquid datasets

Wavenumber ( $\text{cm}^{-1}$ )		
Air-dried	Liquid	Tentative assignment
1645	1664	“Amide I” of proteins $\nu(\text{CO})/\nu(\text{CN})/\delta(\text{NH})$
1547-1539	/	“Amide II” of proteins $\delta(\text{NH})/\nu(\text{CN})$
1498	/	$\nu(\text{CC})/\delta(\text{CH})$ of proteins
1463-1461	1464	$\delta_s(\text{CH}_2)$ of lipids
1415-1413	1415-1412	$\nu(\text{CN})$ of proteins
1390-1386	1390-1386	$\nu_s(\text{COO}^-)$ of proteins
1377	1379	$\delta_s(\text{CH}_3)$ of proteins
/	1298	“Amide III” of proteins $\delta(\text{NH})/\nu(\text{CN})$
/	1153-1151	$\nu_{\text{as}}(\text{CO-O-C})$ of carbohydrates
1078-1076	1097-1079	$\nu_s(\text{PO}_2^-)$ of nucleic acids
1060	/	$\nu_s(\text{CO})$ of carbohydrates
1041-1033	1041-1033	$\nu_s(\text{CO-O-C})$ of carbohydrates

Abbreviation: RF, random forest.

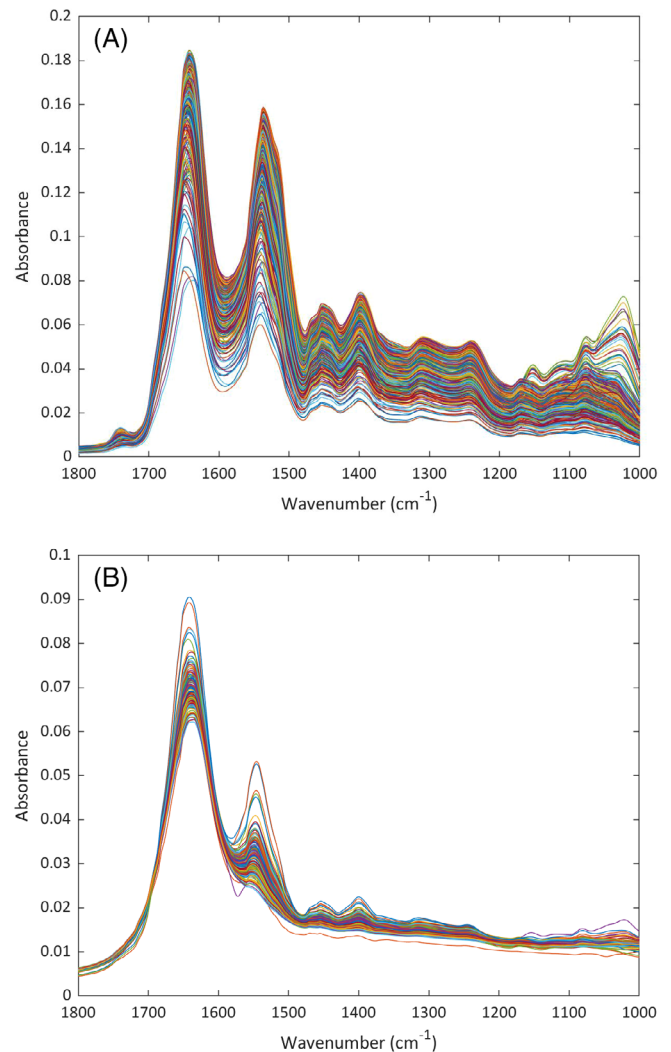
deviating from the common peak ratio, as it can be seen in both air-dried and liquid spectra (Figure 5). Although LS also affected the amide I and amide II peak absorbance ratio, the peak of amide II remained lower than the amide I peak (Figure 4B). This ratio is clearly different from the one seen in liquid spectra (Figure 5B); however, it appears to be more similar to the one commonly seen in air-dried spectra (Figure 5A), as well as the rest of the fingerprint spectral region, showing a great performance of LS as digital drying method. On the other hand, EMSC approach (Figure 4C) did not show an enhancement of the 1500 to 1000  $\text{cm}^{-1}$  region, nor significant visible variations from the liquid spectra (Figure 5B); only a slight increase of absorbance values can be seen in the amide I and amide II region of the spectra (1700–1500  $\text{cm}^{-1}$ ).

Both WS and LS approaches gave optimal results with RF classification. Sensitivity and specificity values of both datasets exceeded not only the ones of the liquid samples spectral dataset, but also the ones of the air-dried one (Table 1); the outputs accounted for more than 93% of sensitivity and more than 83% of specificity (Table 3), which account for 2.5% to 3.8% increase in overall sensitivity and specificity performance over the liquid samples spectral dataset and a 0.5% to 1.9% increase over the air-dried one. On the contrary, RF classification for the EMSC dataset performed worse than the liquid samples spectral dataset; sensitivity was 89.5% and specificity



**FIGURE 4** Spectral effects of the different digital drying approaches on the fingerprint region ( $1800\text{--}1000\text{ cm}^{-1}$ ). A, water subtraction; B, least-squares method; C, EMSC. EMSC, extended multiplicative signal correction

reached only 80.0%. Moreover, only WS and LS obtained a F1 score above 90%, which is indicative of stable and reliable classification methods.



**FIGURE 5** ATR-FTIR raw spectra reduced to the fingerprint region ( $1800\text{--}100\text{ cm}^{-1}$ ). A, 450 air-dried spectra; B, 150 liquid spectra. ATR-FTIR, attenuated total reflection-Fourier transform infrared

The RF (mean decrease Gini) importance plots (Figures S3, S4, and S5) identified proteins, lipids, nucleic acids and carbohydrates as the biomacromolecules affecting the classification models of the datasets. As performed for the air-dried and liquid spectral datasets, the assignment of 20 higher RF importance spectral features to the corresponding IR molecular vibrations in each dataset was attempted in Table 4 [4, 14, 36–39]. The asymmetric phosphate stretching of the nucleic acids was used as one of the main discriminating features only when LS was applied. It is interesting to notice that the feature was not employed in the classification of the air-dried and the other liquid datasets and may represent a starting point for further investigation on spectral biomarkers involved in cancer vs non-cancer discrimination.

**TABLE 3** Statistical outputs of RF classification of digital drying approaches

		WS (%)		LS (%)		EMSC (%)	
		Value	SD	Value	SD	Value	SD
CV	Sensitivity	92.7	2.3	92.2	2.3	89.1	2.5
	Specificity	82.2	4.7	81.1	4.6	79.2	4.6
	Accuracy	89.1	2.3	88.5	2.3	85.8	2.3
	Precision	91.3	2.1	90.7	2.1	89.6	2.1
	F1 Score	91.9	1.7	91.5	1.8	89.3	1.8
TS	Sensitivity	<b>93.7</b>	4.1	<b>93.4</b>	4.2	<b>89.5</b>	5.6
	Specificity	<b>84.0</b>	9.1	<b>83.5</b>	8.9	<b>80.0</b>	10.6
	Accuracy	90.4	3.7	90.1	3.8	86.3	4.8
	Precision	<b>92.3</b>	4.1	<b>92.1</b>	4.0	<b>90.2</b>	4.7
	F1 Score	<b>92.9</b>	2.7	<b>92.6</b>	2.8	<b>89.7</b>	3.7

Abbreviations: CV, cross-validation set; EMSC, extended multiplicative signal correction; LS, least-squares method; RF, random forest; TS, test set; WS, water subtraction.

Wavenumber (cm <sup>-1</sup> )			
WS	LS	EMSC	Tentative assignment
1545	/	/	“Amide II” of proteins $\delta(\text{NH})/\nu(\text{CN})$
1466-1464	1466-1464	1466-1464	$\delta_s(\text{CH}_2)$ of lipids
1414	1417-1414	1421-1414	$\nu(\text{CN})$ of proteins
1390-1387	1392-1387	1392-1387	$\nu_s(\text{COO}^-)$ of proteins
1379-1377	1379	1379	$\delta_s(\text{CH}_3)$ of proteins
/	1298	/	“Amide III” of proteins $\delta(\text{NH})/\nu(\text{CN})$
/	1225	/	$\nu_{\text{as}}(\text{PO}_2^-)$ of nucleic acids
1161	1184-1155	1171-1155	$\nu_{\text{as}}(\text{CO-O-C})$ of carbohydrates
1097-1078	1097-1078	1091-1082	$\nu_s(\text{PO}_2^-)$ of nucleic acids
1039-1037	1039-1036	1043-1036	$\nu_s(\text{CO-O-C})$ of carbohydrates
1012-1011	1010	/	$\nu(\text{CC})$ of carbohydrates

Abbreviations: EMSC, extended multiplicative signal correction; LS, least-squares method; RF, random forest; WS, water subtraction.

**TABLE 4** Assignment of the main spectral features in the RF importance plots for the digital drying datasets

### 3.3 | QCL spectroscopic imaging of liquid samples

As shown in Table 5, both values of sensitivity (85.1%) and specificity (75.3%) resulted around 5% to 6% lower than the classification of the liquid samples spectral dataset. Both precision and F1 score did not reach 90%, indicating a lower model performance than the previous ones analyzed. However, the trend of the QCL-IR classification results followed the ATR-FTIR ones, showing higher sensitivity than specificity. The CV outputs were also lower but close to the TS outputs, indicating consistency with the training set and the overall classification.

The major reason why the RF classification did not perform as well as the previous classifications can be assessed through the analysis of the RF importance plot (Figure S6) where the major discriminating feature was the region of the amide II only, not detected in the classification of the liquid samples spectral dataset. The other biomacromolecules involved in cancer and non-cancer classification were partially detected, with a RF importance significantly lower than the proteins; however, no carbohydrates specific features were used in this classification.

The assignment of 20 higher RF importance spectral features to the corresponding IR molecular vibrations in



**TABLE 5** Statistical outputs of RF classification of liquid samples analyzed with QCL-IR vs ATR-FTIR

		QCL-IR (%)		ATR-FTIR (%)	
		Value	SD	Value	SD
CV	Sensitivity	83.6	2.3	88.9	2.7
	Specificity	75.4	4.7	79.4	4.5
	Accuracy	80.9	2.3	85.7	2.5
	Precision	87.2	2.1	89.6	2.1
	F1 Score	85.3	1.7	89.2	1.9
TS	Sensitivity	<b>85.1</b>	4.1	<b>89.9</b>	5.3
	Specificity	<b>75.3</b>	9.1	<b>81.2</b>	10.9
	Accuracy	81.8	3.7	87.0	4.4
	Precision	<b>87.6</b>	4.1	<b>90.8</b>	4.7
	F1 Score	<b>86.1</b>	2.7	<b>90.2</b>	3.3

Abbreviations: ATR-FTIR, attenuated total reflection-Fourier transform infrared; CV, cross-validation set; QCL-IR, quantum cascade laser infrared; RF, random forest; TS, test set.

each dataset was performed [4, 14, 36–39] and correlated to the tentative assignments of the liquid samples spectral dataset in Table 6. The symmetric carbonyl stretching of lipids was highlighted for the first time in these classifications.

## 4 | CONCLUSIONS

The initial investigation of the cancer vs non-cancer samples classification showed comparative outputs with precision and F1 score of the model above 90%. The air-dried dataset produced optimal results with sensitivity and specificity accounting for 91.8% and 83.2%. Furthermore, both classifications were based on a wide range of spectral features, but only the classification of the air-dried dataset used the spectral features of both amide I and amide II ( $1700\text{--}1500\text{ cm}^{-1}$ ) to discriminate the disease state.

The digital drying preliminary study was applied on the liquid samples spectral dataset to bring increased performances in the RF classification models, in terms of statistical outputs and spectral features. Even though the RF classifications were based on less spectral features than the air-dried and liquid datasets classifications, the WS and LS approaches produced outputs even greater than the air-dried dataset; sensitivity and specificity accounted for 93.7% and 84.0% in the WS dataset classification and 93.4% and 83.5% in the LS dataset. On the other side, the EMSC approach operated defectively in RF classification with sensitivity and specificity of 89.5% and 80.0% respectively.

**TABLE 6** Assignment of the main spectral features in the RF importance plots of the liquid dataset analyzed with QCL-IR vs ATR-FTIR

Wavenumber ( $\text{cm}^{-1}$ )		
QCL-IR	ATR-FTIR	Tentative assignment
1732-1716	/	$\nu_s(\text{CO})$ of lipids
/	1664	“Amide I” of proteins $\nu(\text{CO})/\nu(\text{CN})/\delta(\text{NH})$
1592-1524	/	“Amide II” of proteins $\delta(\text{NH})/\nu(\text{CN})$
/	1464	$\delta_s(\text{CH}_2)$ of lipids
/	1415-1412	$\nu(\text{CN})$ of proteins
/	1390-1386	$\nu_s(\text{COO}^-)$ of proteins
/	1379	$\delta_s(\text{CH}_3)$ of proteins
/	1298	“Amide III” of proteins $\delta(\text{NH})/\nu(\text{CN})$
1268	/	$\nu_{\text{as}}(\text{PO}_2^-)$ of nucleic acids
/	1153-1151	$\nu_{\text{as}}(\text{CO-O-C})$ of carbohydrates
1092-1088	1097-1079	$\nu_s(\text{PO}_2^-)$ of nucleic acids
/	1041-1033	$\nu_s(\text{CO-O-C})$ of carbohydrates

Abbreviations: ATR-FTIR, attenuated total reflection-Fourier transform infrared; CV, cross-validation set; QCL-IR, quantum cascade laser infrared; RF, random forest; TS, test set.

The QCL spectroscopic imaging approach was expected to overcome water interference across a defined wavenumber range by using its higher spectral power. Notwithstanding precision and F1 score values of the RF classification model above 90%, preliminary results showed poor sensitivity and specificity values, of 85.1% and 75.3% respectively; these were around 5% lower than the outputs of the liquid samples spectral dataset classification obtained by ATR-FTIR spectroscopy. The analysis of the RF importance plot suggested that the lower classification values were due to the weak ability of the model in detecting common spectral features for classification. However, the model detected spectral features unseen in the other classification, which might represent the grounds for a further investigation of the technique.

In conclusion, it can be considered that these preliminary results are a noteworthy investigation of the potential of ATR-FTIR spectroscopy on liquid serum samples, in terms of needing a rapid, sensitive and specific test. A test with high sensitivity and specificity would be able to accurately identify patients with brain cancer, ensuring their time-to-diagnosis is reduced, and preventing healthy patients from unnecessary investigation. Digital drying could be an optimal tool to overcome the water spectral interferences, increasing sensitivity and specificity during the classification of liquid samples spectra and

therefore, representing a precious aid for the clinical environment.

## ACKNOWLEDGMENTS

The authors would like to thank Rosemere Cancer Foundation, Scottish Enterprise, EPSRC (EP/L505080/1) and ClinSpec Diagnostics Ltd for the funding provided.

## ORCID

Alexandra Sala  <https://orcid.org/0000-0001-6417-9706>

## REFERENCES

- [1] WHO, IARC, Press Release No. 263, <https://who.int/cancer/PRGlobocanFinal.pdf> (accessed: March 31, 2020).
- [2] R. Cerqua, S. Balestrini, C. Perozzi, V. Cameriere, S. Renzi, G. Lagalla, G. Mancini, M. Montanari, P. Leoni, M. Scerrati, M. Iacoangeli, M. Silvestrini, S. Luzzi, L. Provinciali, *Neurol. Sci.* **2016**, *37*, 23.
- [3] H. Z. Wang, T. M. Simonson, W. R. Greco, W. T. C. Yuh, *Acad. Radiol.* **2001**, *8*, 405.
- [4] E. Gray, H. J. Butler, R. Board, P. M. Brennan, A. J. Chalmers, T. Dawson, J. Goodden, W. Hamilton, M. G. Hegarty, A. James, M. D. Jenkinson, D. Kernick, E. Lekka, L. J. Livermore, S. J. Mills, K. O'Neill, D. S. Palmer, B. Vaqas, M. J. Baker, *BMJ Open* **2018**, *8*, e017593.
- [5] Cancer Research UK, Why is early diagnosis important? <https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important> (accessed: March 31, 2020).
- [6] Z. J. Sahab, S. M. Semaan, Q.-X. A. Sang, *Biomark. Insights* **2007**, *2*, 21.
- [7] J. Liu, Y. Duan, *Oral Oncol.* **2012**, *48*, 569.
- [8] M. J. Baker, H. J. Byrne, J. Chalmers, P. Gardner, R. Goodacre, A. Henderson, S. G. Kazarian, F. L. Martin, J. Moger, N. Stone, J. Sulé-Suso, *Analyst* **2018**, *143*, 1735.
- [9] L. V. Bel'skaya, *J. Appl. Spectrosc.* **2019**, *86*, 187.
- [10] A. Sala, D. J. Anderson, P. M. Brennan, H. J. Butler, J. M. Cameron, M. D. Jenkinson, C. Rinaldi, A. G. Theakstone, M. J. Baker, *Cancer Lett.* **2020**, *477*, 122.
- [11] K. Spalding, R. Board, T. Dawson, M. D. Jenkinson, M. J. Baker, *Brain Behav.* **2016**, *6*, e00502.
- [12] M. J. Baker, C. Hughes, R. A. Lukaszewski, G. Thiéfin, G. D. Sockalingum, V. Untereiner, S. R. Hussain, L. Lovergne, *Chem. Soc. Rev.* **2015**, *45*, 1803.
- [13] F. Bonnier, F. Petitjean, M. J. Baker, H. J. Byrne, *J. Biophotonics* **2014**, *7*, 167.
- [14] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, F. L. Martin, *Nat. Protoc.* **2014**, *9*, 1771.
- [15] J. R. Hands, K. M. Dorling, P. Abel, K. M. Ashton, A. Brodbelt, C. Davis, T. Dawson, M. D. Jenkinson, R. W. Lea, C. Walker, M. J. Baker, *J. Biophotonics* **2014**, *7*, 189.
- [16] G. L. Owens, K. Gajjar, J. Trevisan, S. W. Fogarty, S. E. Taylor, B. Da Gama-Rose, P. L. Martin-Hirsch, F. L. Martin, *J. Biophotonics* **2014**, *7*, 200.
- [17] J. Backhaus, R. Mueller, N. Formanski, N. Szlama, H.-G. Meerpohl, M. Eidt, P. Bugert, *Vib. Spectrosc.* **2010**, *52*, 173.
- [18] H. J. Byrne, M. Baranska, G. J. Puppels, N. Stone, B. Wood, K. M. Gough, P. Lasch, P. Heraud, J. Sulé-Suso, G. D. Sockalingum, *Analyst* **2015**, *140*, 2066.
- [19] R. Adato, H. Altug, *Nat. Commun.* **2013**, *4*, 2154.
- [20] J. R. Hands, G. Clemens, R. Stables, K. Ashton, A. Brodbelt, C. Davis, T. P. Dawson, M. D. Jenkinson, R. W. Lea, C. Walker, M. J. Baker, *J. Neurooncol.* **2016**, *127*, 463.
- [21] K. M. Dorling, M. J. Baker, *Trends Biotechnol.* **2013**, *31*, 327.
- [22] J. Ollesch, S. L. Drees, H. M. Heise, T. Behrens, T. Brüning, K. Gerwert, *Analyst* **2013**, *138*, 4092.
- [23] E. Fotheringham, T. Dawson, A. Brodbelt, G. Clemens, C. Hughes, M. J. Baker, K. M. Ashton, M. Weida, B. Bird, M. D. Jenkinson, J. Rowlette, M. Barre, *Sci. Rep.* **2016**, *6*, 4.
- [24] A. Schwaighofer, M. Brandstetter, B. Lendl, *Chem. Soc. Rev.* **2017**, *46*, 5903.
- [25] G. Clemens, B. Bird, M. Weida, J. Rowlette, M. Baker, *Spectrosc. Eur.* **2014**, *26*, 14.
- [26] D. Sebiskveradze, C. Gobinet, E. Ly, M. Manfait, P. Jeannesson, M. Herbin, O. Piot, V. Vrabie, in: **2008 8th IEEE Int. Conf. Bio-Inform. BioEng.**, Athens, Greece, October **2008**, pp. 1–6.
- [27] B. R. Smith, M. J. Baker, D. S. Palmer, *Chemom. Intel. Lab. Syst.* **2018**, *172*, 33.
- [28] J. M. Cameron, H. J. Butler, D. S. Palmer, M. J. Baker, *J. Biophotonics* **2018**, *11*, 1.
- [29] N. K. Afseth, A. Kohler, *Chemom. Intel. Lab. Syst.* **2012**, *117*, 92.
- [30] A. Liaw, M. Wiener, *R News* **2002**, *2/3*, 18.
- [31] L. E. O. Breiman, *Mach. Learn.* **2001**, *45*, 5.
- [32] B. R. Smith, K. M. Ashton, A. Brodbelt, T. Dawson, M. D. Jenkinson, N. T. Hunt, D. S. Palmer, M. J. Baker, *Analyst* **2016**, *141*, 3668.
- [33] D. S. Palmer, N. M. O'Boyle, R. C. Glen, J. B. O. Mitchell, *J. Chem. Inf. Model.* **2007**, *47*, 150.
- [34] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, *J. Artif. Intell. Res.* **2002**, *16*, 321.
- [35] S. Shekarforoush, R. Green, R. Dyer, in: **2017 Int. Joint Conf. Neur. Netw.**, Anchorage, AK, May **2017**, pp. 1273–1280.
- [36] E. Wiercigroch, E. Szafraniec, K. Czamara, M. Z. Pacia, K. Majzner, K. Kochan, A. Kaczor, M. Baranska, K. Malek, *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2017**, *185*, 317.
- [37] A. Barth, *Biochim. Biophys. Acta - Bioenerg.* **1767**, 2007, 1073.
- [38] M. Banyay, M. Sarkar, A. Gräslund, *Biophys. Chem.* **2003**, *104*, 477.
- [39] F. Bonnier, M. J. Baker, H. J. Byrne, *Anal. Methods* **2014**, *6*, 5155.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Sala A, Spalding KE, Ashton KM, et al. Rapid analysis of disease state in liquid human serum combining infrared spectroscopy and “digital drying”. *J. Biophotonics*. 2020;e202000118. <https://doi.org/10.1002/jbio.202000118>