Received: 7 August 2020 | Revised: 21 September 2020 | Accepted: 24 September 2020

DOI: 10.1002/tbio.202000019

FULL ARTICLE

TRANSLATIONAL BIOPHOTONICS

Data consistency and classification model transferability across biomedical Raman spectroscopy systems

Fabien Picot[1,2] | **François Daoust**[1,2] | **Guillaume Sheehy**[1,2] | **Frédérick Dallaire**[2] | **Layal Chaikho**[1] | **Théophile Bégin**[1] | **Samuel Kadoury**[1,2] | **Frédéric Leblond**[1,2,3]

[1]Department of Engineering Physics, Polytechnique Montréal, 2500 chemin de Polytechnique, Montreal, Quebec, Canada

[2]Centre de recherche du Centre Hospitalier de l'Université de Montréal, Montreal, Quebec, Canada

[3]Institut du Cancer de Montréal, Montreal, Quebec, Canada

Correspondence
Frédéric Leblond, Department of Engineering Physics, Polytechnique Montreal, 2500 chemin de Polytechnique, Montreal QC H3T 1J4, Canada.
Email: frederic.leblond@polymtl.ca

Funding information
Canadian Institutes of Health Research; Canadian Network for Research and Innovation in Machining Technology, Natural Sciences and Engineering Research Council of Canada; TransMedTech Institute

Abstract

Surgical guidance applications using Raman spectroscopy are being developed at a rapid pace in oncology to ensure safe and complete tumor resection during surgery. Clinical translation of these approaches relies on the acquisition of large spectral and histopathological data sets to train classification models. Data calibration must ensure compatibility across Raman systems and predictive model transferability to allow multi-centric studies to be conducted. This paper addresses issues relating to Raman measurement standardization by first comparing Raman spectral measurements made on an optical phantom and acquired with nine distinct point probe systems and one wide-field imaging instrument. Data standardization method led to normalized root-mean-square deviations between instruments of 2%. A classification model discriminating between white and gray matter was trained with one point probe system. When used to classify independent data sets acquired with the other systems, model predictions led to >95% accuracy, preliminarily demonstrating model transferability across different biomedical Raman spectroscopy instruments.

KEYWORDS

cancer, classification models, medical imaging, Raman spectroscopy

1 | INTRODUCTION

Raman spectroscopy (RS) is an optical technique which can assess a sample's molecular content by probing its vibrational states. Over the last decades, it has been used in the medical field, in particular toward the development of imaging systems to detect pathologies [1–3]. The technique was used to guide tumor resection where the

Abbreviations: (N)RMSD, (normalized) root-mean-square deviation; CCD, charged-coupled device; FOV, field of view; FP, fingerprint; HWN, high wavenumber; LOPOCV, leave one patient out cross-validation; PBS, phosphate-buffered saline; PCA-LDA, principal component analysis linear discriminant analysis; PLS-DA, partial least-square discriminant analysis; QF, quality factor; RS, Raman spectroscopy; SNV, standard normalized value; SORS, spatially offset Raman spectroscopy; SVM, support vector machine; WF, wide-field.

objective was to use the vibrational spectroscopy information in combination with machine learning technology to maximize the volume of resected malignant tissue while preserving healthy tissue. Raman spectroscopy also showed promises in other clinical applications such as guided biopsy procedures in prostate surgery and neurosurgery [4, 5]. In 2011, Vargis et al presented a Raman probe used in an ex vivo human study [6] to discriminate between normal, benign and malignant areas of the cervix, leading the way towards in vivo diagnosis. These results were followed in 2014 by another group, Shaikh et al, which was able to diagnose in vivo cervix cancer tissue with an accuracy superior to 95% [7]. Additional in vivo human diagnostic applications using RS can be found for multiple pathologies and especially multiple cancer types, such as gastrointestinal cancers [8–14] and malignant ulcers [15], lung cancers [16, 17], skin cancers [18–21], oral cancers [22, 23] and brain cancers [4, 24–28]. The number of patients in these studies varies greatly, ranging from 2 [17] to 848 [20], but in most cases does not exceed 100. A remarkable feat is that most RS studies consistently reported classification accuracies superior to 80%, while the lowest sensitivity and specificity reported were 75% [9] and 65% [16]. Most of these results were obtained by training a classification algorithm, usually combined with a leave-one-patient-out cross-validation (LOPOCV). Principal component analysis combined with linear discriminant analysis (PCA-LDA) and partial least-square analysis combined with discriminant analysis (PLS-DA) stands as the most common algorithms used.

An emergent method to improve Raman imaging systems is to take advantage of a complementary signature to the fingerprint (FP) domain, that is, the high wavenumber (HWN) range between 2500 and 3400 cm^{-1}. This was achieved in 2016 for oral [29] and colon [30] cancers and in a swine brain model in 2018 demonstrating its feasibility for human brain surgery [31]. Innovative clinical applications of RS, with in vivo human surgery potential, are still in development. It is partially done by conducting new animal experiments [32] and by designing new hardware systems, including Raman wide-field imaging systems [33, 34] as well as fused navigation platforms with Raman probe [5]. Overall, RS has demonstrated its great potential as an assisting diagnosis tool in the clinical environment. However, there are still relatively few large-scale trials in the literature which not only limit the performance of the statistical models built but also mitigate their transferability to new trials.

To ensure high accuracy surgical guidance for a real-time human in situ application, the clinical translation relies on the quality and quantity of spectral and histopathological data sets from numerous patients. Due to the rapid pace of the RS development in oncology, multicentric studies stand out as the best way to build these large-scale data sets. As a result, data calibration must ensure compatibility across Raman imaging systems. To answer this problem, we present in this paper a quantitative evaluation of the Raman data reproducibility when measurements are acquired and processed using multiple systems with different optical designs. This study's scope is limited to the compatibility of Raman biomedical spectroscopy systems regardless of the particularities of their clinical application. Hence, the assessment is focusing on the inter-system deviation when comparing Raman spectra on a reference nylon phantom and the diagnosis error of a classification model trained with one system and then exported and tested on multiple other systems.

2 | METHODS

2.1 | Imaging systems and acquisition protocol

2.1.1 | Point probe systems

Single-point interrogation in this study was performed using the fiber optics probe system developed by our group for neurosurgical applications [28]. Briefly, the hand-held probe (Emvision, LLC) was connected to a 785 nm laser source (Innovative Photonics Solution) and a spectrometer (Emvision, LLC) with a resolution of ≈ 2 cm^{-1}. The probe had an outer diameter of 2.1 mm and was designed for tissue interrogation in reflectance. It integrated a central 272 μm (diameter) core excitation fiber surrounded by seven 300 μm core detection fibers. To minimize the Raman signature of the silica in the fibers, a short-pass filter and a notch filter were disposed of in front of the illumination fiber and the detection fibers, respectively. Furthermore, a two-component lens was placed at the probe tip to ensure overlap between the excitation and detection areas, and a 3-m long fiber bundle connected the probe to the laser and the spectrometer. Three probes, three spectrometers and one laser source were combined to assemble a total of nine distinct systems, labeled #1 to #9 (Figure 1). The optical design and fabrication process was the same for all probe systems but the spectrometers differed in their slit width: 100 μm for systems #1 to #6 and 75 μm for systems #7 to #9.

All systems were controlled by a computer and the acquisition parameters were pre-set using a custom Matlab (Mathworks) software. The laser power, P and the number of repeat measurements at each point, n, were set at 50 mW and 20, respectively. The exposure

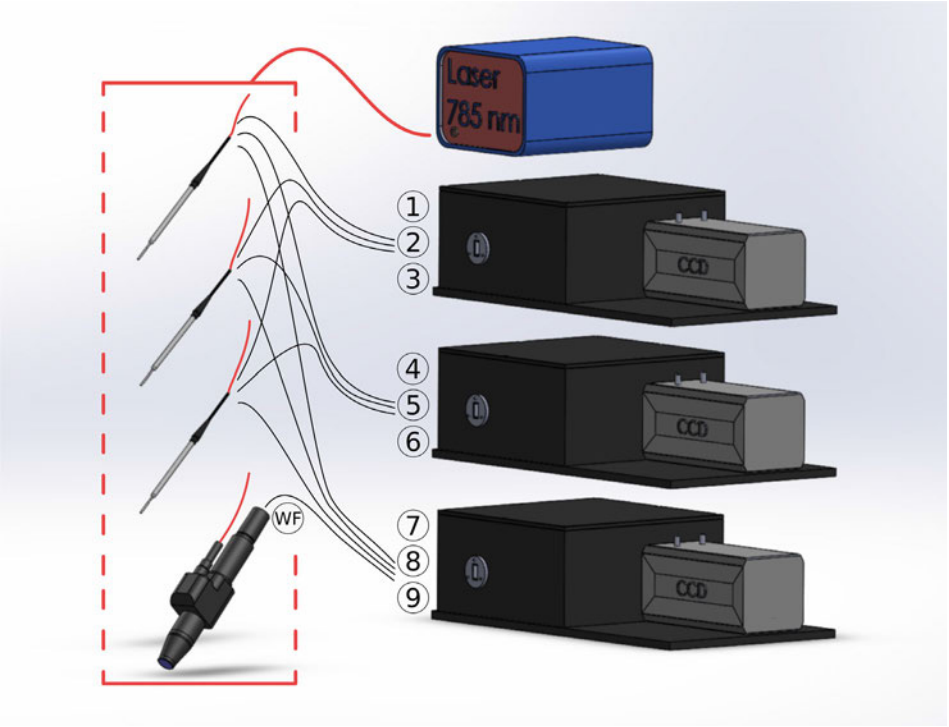

FIGURE 1 Three hand-held probes, one portable wide-field imaging instrument (labeled WF), three spectrometers and one 785 nm laser source were combined to build a total of 10 different Raman spectroscopy systems

time per spectrum, T, was computed and adjusted automatically to ensure raw detected light intensity (photon count) >85% of the charged-coupled device (CCD) camera dynamic range, while avoiding saturation. The following measurements were made for normalization purposes: a dark count measurement with the laser off, an acetaminophen (DiN 00789801, Trianon inc, Canada) measurement for x-axis calibration, and a measurement on a relative intensity RS standard for excitation at 785 nm (NIST2241, NIST).

2.1.2 | Wide-field Raman imaging system

The wide-field RS system used in this study was a modified version of a custom line-scanning instrument developed by our group [34]. The improvements made in the new system included the addition of a bright-field reflectance channel and a series of optical components in the light path to the spectrometer. The new optics effectively decreased image size at the spectrometer slit and increased the signal-to-noise ratio (SNR). Briefly, the system consisted of a portable imaging instrument with a working distance of 40 mm, a FOV of 4×4 mm^2 and a dichroic beam-splitter (LP02-785RE-25, Semrock) combining the excitation and collection optical paths. Raman excitation was achieved using a 785 nm laser (Innovative Photonics Solution) with an average laser intensity of 6.0 W/cm^2 at the sample along a single line scanned over the sample. Light detection was done through a series of collection optics, including a notch filter to eliminate elastic scattering from the excitation source and a 91 cm long coherent imaging bundle with a 4×4 mm^2 light-sensitive area (IG-154, Schott). The imaging bundle conveyed images through a flexible conduit of fiber optics to the collection branch of the system. The collection branch employed a series of optics and a scanning dichroic mirror to separate Raman scattering (810 to 922 nm) from the bright field signal (400 to 700 nm). Spectroscopic images were acquired through line scanning across the 75 μm slit of the spectrometer (HT model, EmVision LLC) and the spectral content of each line was collected using a -60°C cooled CCD camera (Newton 920, Oxford Instruments). A bright-field reflectance image was collected with a high-sensitivity CMOS camera (DCC1240C, Thorlabs) upon sample excitation with an exterior white light source (LED lamp).

Each component of the system was controlled via a custom software developed using LabVIEW (LabVIEW 2017, National Instruments). Control and synchronization of the main system components (laser, galvanometer, cameras) were achieved using a digital acquisition device (150099A-04L, Texas Instruments). Exposure time was kept at 4 seconds per line and laser intensity was set at its maximum value for each measurement. A total of 3 hyperspectral images were acquired and averaged for measurements on tissue phantoms while a total of 5 images were acquired and averaged for measurements on ex vivo tissue samples. Raman images were

reconstructed by scanning 20 lines across a 16 mm^2 FOV. Pixels of the CCD camera were binned over 6 pixels along the spatial direction and 3 pixels along the spectral axis using the camera software (Solis S, Oxford Instruments). This led to an instrument spatial resolution of 250 μm and a spectral resolution of 6 cm^{-1}. Room lights and the system's white light source were turned off during all spectroscopic measurements. White light images were acquired before every Raman image acquisition and the spectral data were processed in the same manner as the point probe systems. Table 1 summarizes the main imaging specifications distinguishing the imaging system from the point probe systems. The sampling depth for the Raman systems is estimated for both types of systems to ~500 μm based on [35].

2.2 | Tissue phantom experiment

Nylon is associated with a distinctive Raman signature in the 400-1800 cm^{-1} region [36]. A pure nylon disk (diameter = 5 cm, thickness = 1 cm) was cut and used as an optical phantom to evaluate signal variability between all 10 Raman systems. For the point probe systems, three measurements were made at three different locations on the phantom with the instrument placed in contact with the surface. The wide-field system was used to acquire a Raman image of the phantom and a nylon spectrum computed by averaging over three randomly selected pixels. The intra-system and inter-system deviations were computed as outlined in section 2.5.

2.3 | Ex vivo experiment

Six cynomolgus monkey (*Macaca fascicularis*) brains were used for ex vivo tissue experiments. The brains were cut into 52 slices (thickness = 1 mm), with each slice presenting visually distinguishable white matter and gray matter structural features. Sample preparation and storage were done at the CERVO Brain Research Centre. Immediately after animal death, the brains were collected and fixed by immersion, at 4°C, in 4% paraformaldehyde for 24 hours. The organs were then sliced using a vibratome and preserved in phosphate-buffered saline (PBS). This procedure was approved by the animal protection committee of Université Laval, in accordance with the Canadian Council on Animal Cares Guide and Use of Experimental Animals. All animals included in this study served in other experiments prior to our study.

The brain samples were used in an experiment evaluating classifier transferability between different RS systems based on a binary classification model (white matter vs gray matter). To minimize background contributions from nontissue specific Raman signals, slices were imaged on low Raman activity aluminum slides. System #1 was used to build the classification model using a 400 point-measurement data set. Measurements were distributed equally between white matter and gray matter across the 52 slices. Point probe systems #2 to #9 and the wide-field imaging system were then used to acquire a testing data set to evaluate the classification performance of the model. In total, 100 point measurements were acquired with the remaining point probe systems (#2 to #9), for a total of 800 points ensuring equal distribution between white matter and gray matter. All point measurements were acquired using a 1D linear stage to hold the probe and manually modify its position. Every point measurement required ~30s (~1.5 seconds per spectrum for a total of 20 spectra per point measurement). Overall, the duration of the experiment was approximately 20 hours.

Finally, one brain slice was randomly selected (among the 52 that were available) and imaged using the wide-field system. The resulting Raman spectroscopy image was spatially registered with a white light reflectance image to ensure each pixel could be assigned a tissue class (white matter or gray matter) in preparation for classification model testing.

TABLE 1 Parameter differences between Raman wide-field system and point probe systems

Parameters	Point probe systems	Imaging system
Sampling area	Ø500 μm	104 × 250 $μm^2$ per pixel
Laser power	25.5 W/cm^2	6.0 W/cm^2
Spectral resolution	2 cm^{-1}	6 cm^{-1}
Integration time T per spectrum	~1 second per point	4 seconds per line

2.4 | Spectral data processing and statistical analysis

The raw spectroscopic data were first averaged over the number of spectra (n) acquired for each point. A cosmic ray removal algorithm was applied and the dark count signal was subtracted [37]. The signal was subsequently normalized with the NIST standard measurement and an x-axis calibration performed using the acetaminophen measurement. The background was finally removed using a rolling ball algorithm [38]. In addition, a quality factor QF(n), was computed which consists in summing

TABLE 2 Raman band with molecular assignment, based on the literature, found in the current study on the monkey brain classification model

Raman band (cm^{-1})	Associated bond	Assigned molecules	Molecular family	References
700	Vibrational mode of sterol ring	Cholesterol	Lipids	[45]
1001	Symmetric ring breathing|C—C stretching	Phenylalanine|collagen|heme|carotenoid	Proteins	[46]
1064	C—O stretch|C—O—C symmetric stretch|C-C stretch	Proline|phospholipid side chains|cholesterol	Proteins|lipids	[46]
1086	C—C stretch|PO_2-symmetric stretch|C═O vibration	Phospholipids|nucleic acids	Lipids|DNA	[46]
1129	C—C stretching	Skeletal of acyl backbone in lipid	Lipids	[47]
1262	CH_2 in plane deformation	Glycerophospholipid	Lipids	[47, 48]
1298	CH_2 twist and wag|amide III	Phospholipids|palmitic acid|cholesterol|collagen	Lipids|proteins	[46]
1441	CH_2/CH_3 deformation	Lipid side chains|amino acids|cholesterol|collagen	Lipids|proteins	[46]
1580	C═C bending mode of phenylalanine	Phenylalanine	Proteins	[47]
1659	Amide I|C═C stretching	Nucleic acids|collagen|unsaturated fatty acids	DNA|proteins|lipids	[46]

Note: Bold bands were used as features to train the classification model.

the Raman signal-to-noise ratio SNR_R over N spectral bands:

$$SNR_R \approx \sqrt{n\,T\,P} \sum_{j=1}^{N} \frac{r_j}{\sqrt{r_j + a_j}} \quad (1)$$

where r_j and a_j are the Raman signal and the background signal (mostly instrument response and fluorescence from the sample) within spectral band j, respectively [37]. Due to its ubiquitous presence in biological tissues [28, 39–41], the Raman peak at 1441 cm^{-1} was used as the unique band for the quality factor measurement. The final Raman spectrum R was then obtained by applying the standard normal variate (SNV) method. The Raman spectrum for each measurement point was then reduced to a list of K interpretable molecular features $\{\bar{R}\}_{l=1..K}$, around known tissue Raman peaks. Each feature was computed by integrating over many bands extending over 10 cm^{-1} around each peak.

The features used to produce the classification model were limited to $K = 4$ bands and were associated with common tissue molecular bonds: 1298, 1441, 1580 and 1659 cm^{-1} (Table 2). The bands around 1298, 1441 and 1659 cm^{-1} are associated with both lipids and proteins, but the 1659 cm^{-1} band can also be associated with nucleic acid content. The band around 1580 cm^{-1} is typically associated only with protein content, typically phenylalanine.

The next step consisted in producing classification models based on supervised training and testing on independent data. In this study, a two-class model was produced to discriminate between white matter and gray matter using the monkey brain samples by applying the following machine learning workflow. The classification algorithm was based on a support vector machine (SVM) [42] from $K = 4$ features resulting in a three-dimensional decision boundary; a similar modeling technique was used in ref. [43]. Classification between white matter and gray matter was achieved through an optimization process using the geometrical distance of all measurement points from the decision boundary. Specifically, training was performed by minimizing a loss function which depends on SVM hyperparameters, namely: a regularization parameter, the kernel function and the kernel coefficient. The data set acquired with system #1 was divided into a training set (75% of all data points) and a testing set (25% of all data points). A 10-fold cross-validation procedure was then conducted using the training set to find the optimal hyperparameters based on a grid search method. The resulting hyperparameters obtained were 0.22 for the regularization parameter, 0.25 for the kernel coefficient and the kernel function was set to "linear." The resulting model was then applied on an independent testing set composed of data from probes #2 to #9 and the wide-field system. The model performance was reported in terms of classification accuracy, sensitivity and specificity based on a receiver-operating-

characteristic (ROC) curve analysis. All data processing was performed using the Matlab (Mathworks) machine learning library.

2.5 | Quantification of the intra-system and inter-system variability

Inter-system variability was evaluated by comparing measurements made on a nylon optical phantom with the point probe systems and the imaging system. Three co-located measurements were averaged for the point probe systems, while for the wide-field three randomly selected pixels across the image were averaged. These average values were used to compare system #1 to all the others point probe systems, and to compare system #1 to the wide-field system. The Raman spectrum for each measurement was reduced to a list of $L = 14$ features $\{\bar{R}\}_{l=1..L}$, associated with characteristic nylon Raman bands (491.2, 616.7, 708.0, 788.7, 861.3, 953.8, 1062, 1131, 1234, 1298, 1384, 1443, 1477, 1634 cm^{-1}).

To compare system #1 to all the others point probe systems we calculated the residuals between the features $\{\bar{R}\}_{l=1..L}$ of system #1 and of systems #2 to #9. We then calculated the RMSD by taking the square-root of the residuals averaged over the systems. Finally, the RMSD was normalized by the statistical extent, $(\max_l\{\bar{R}\}_{l=1..L} - \min_l\{\bar{R}\}_{l=1..L})$ of system #1, as shown in the following equation:

$$NRMSD = \frac{RMSD}{\max_l\{\bar{R}\}_{l=1..L} - \min_l\{\bar{R}\}_{l=1..L}} \quad (2)$$

where $\max_l\{\bar{R}\}_{l=1..L}$ and $\min_l\{\bar{R}\}_{l=1..L}$ are the maximum and minimum of $\{\bar{R}\}_{l=1..L}$ over $\{l\}_{l=1...L}$ respectively. Similarly, we used the three point-measurements taken on the nylon phantom to evaluate the intra-system variability with every system. We used the same equation to calculate the NRMSD but the residuals calculated were between the three co-located measurements and their average. This complementary evaluation gives a reference for the inter-system variability to be compared with.

This procedure was repeated identically to compare system #1 to the wide-field system.

2.6 | Quantitative evaluation of model transferability in monkey brain experiment

The classification model was first applied to the testing set from system #1 and the resulting sensitivity and specificity reported. These values were used as a comparison basis for the classifier's performance when applied on data acquired with the testing probe systems (#2 to #9). For each testing system, the entirety of the acquired data was considered as the testing data and the resulting sensitivity and specificity were averaged over 100 iterations of the machine learning workflow described in section 2.4. These values were then directly compared with the base reference classifier's performance. Similarly, the classification model was applied to all pixels within the Raman image acquired with the wide-field system. The performance of the model was estimated based on class labels (white matter or gray matter) assigned from inspection of the white light reflectance image.

3 | RESULTS

3.1 | Quantitative comparison between systems

The inter-system variability was evaluated for all prominent Raman nylon peaks [36]: (491.2, 616.7, 708.0, 788.7, 861.3, 953.8, 1062, 1131, 1234, 1298, 1384, 1443, 1477, 1634) cm^{-1}. Figure 2 illustrates the superposition of the nylon Raman measurements. Figure 2A-C show this superposition when the 9 point probe systems are compared with one another, while Figure 2D-F show the comparison between the point probe system #1 and the imaging system. The NMRSD between the 9 point probe Raman spectra is consistently inferior to 2% (Figure 2C). Similarly, Figure 2E shows how the Raman spectrum acquired using a point probe system compares to one acquired with the imaging system. The NMRSD between the two systems is overall inferior to 6% with only the peaks at 953.8, 1131 and 1443 cm^{-1} being over 4%. This deviation is to be compared with the point probes intra-system deviation, reaching a maximum of 5%.

3.2 | Quantitative evaluation of model transferability in ex vivo experiment

The *ex vivo* experiment illustrate classification model performances on multiple systems is presented in Figure 3. Figure 3A and B show the sensitivity and specificity of the classifier when applied on the testing sets acquired with both point probe system #1 (training system) and point probe systems #2 to #9 (testing systems). Figure 3C shows that the average quality factor, which is a function of n, varies from ~5 to a maximum comprised between 45 and 65 for all systems #1 to #9. As a result, which system is used only has a weak impact on the data quality

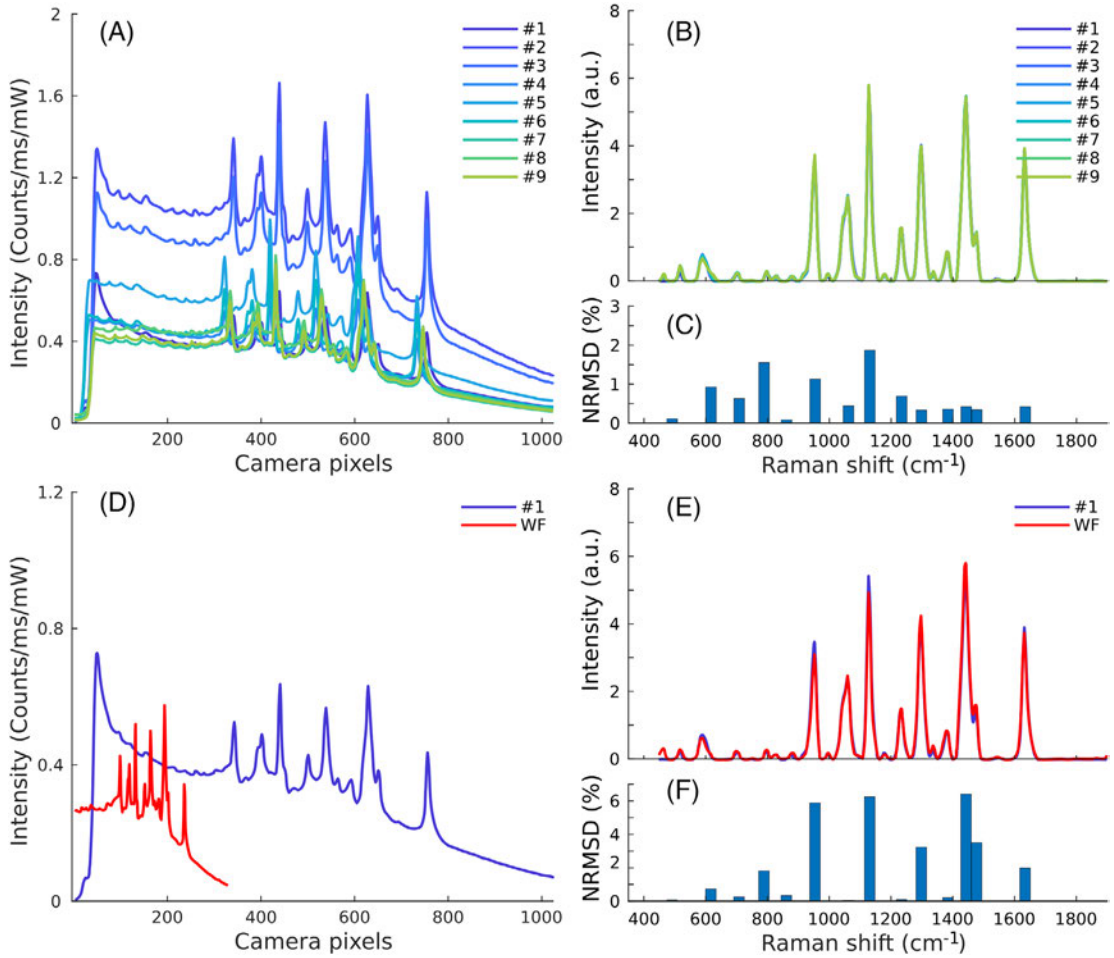

FIGURE 2 Quantitative comparison between point probe and wide field systems. Spectra comparison for point probe systems #1 to #9: A, raw spectra, B, processed Raman spectra, C, inter-system normalized root-mean-square deviation (NRMSD) for nylon features. Spectra comparison between point-probe system #1 and the wide-field system: D, raw spectra, E, processed Raman spectra, F, inter-system NRMSD for extracted features

factor. In contrast, the quality factor is highly dependent on the interrogated tissue type. As shown in Figure 3D, the quality factor averaged on all gray matter sample reaches only a maximum of 15 compared with 100 for white matter. When tested on the testing set acquired with the training system, the classification model has a sensitivity of 99%, regardless of n. Its average sensitivity on data from the testing systems increases from 95% to 98% when n increases from 1 to 5 and is constant otherwise (Figure 3A), meaning the optimum number of measurement per point for the brain classifier is n = 5 even though the quality factor increases for greater n values. Furthermore, the specificity for the training data are constant at 100% and is on average superior to 99% for the testing data with a deviation smaller than 5% (Figure 3B). As a result, when exported from the training system to the testing systems, the model accuracy drops by less than 5% and remains >95%.

The classification model trained using system #1 successfully classified white matter vs gray matter in the Raman image acquired using system #10, as shown in Figure 4A. Figure 4B displays the white light reflectance image used as a gold standard to identify the tissue classes and Figure 4C shows the quality factor for every pixel, with an average of 20 and 60 for gray and white matter pixels, respectively. Nine pixels presented a factor of 0 because of a hardware fault during the acquisition. These pixels were systematically classified as gray matter pixels in the Raman map displayed in Figure 4A, leading to 6 of them, located in the white matter area, to be misclassified. Aside from these misclassified pixels due to a hardware fault, a 500 μm band of gray matter is misclassified as white matter consistently along the border between tissue type. Otherwise, the rest of the image is classified accurately, demonstrating the transferability of the classification model, built with the training system, toward the wide-field system.

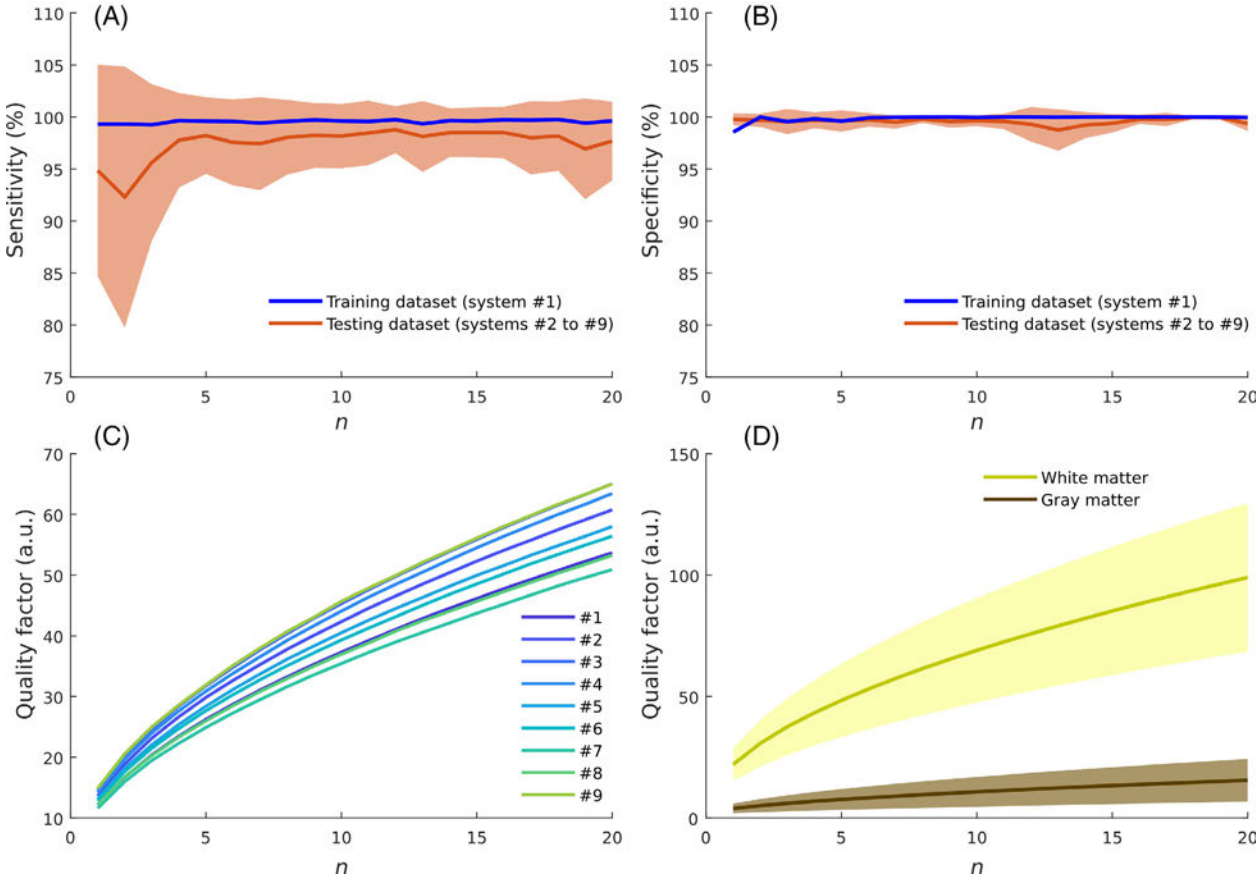

FIGURE 3 Classification model compatibility between the point probe systems and the wide-field imaging system. Average sensitivity, A, and specificity, B, of the classification model built with system #1 and tested on data acquired systems #2 to #9. The average and SD were calculated over 100 iterations of the machine learning workflow. C, Quality factor QF(n) depending on number of spectra per acquisition for systems #1 to #9. D, Quality factor QF(n) depending on number of spectra per acquisition for the two classes white/Gy matter. The average and SD were calculated over data measurements acquired per system, C, and per class, D

Finally, Figure 4D and F show the average spectra measured on white and gray matter using the point probe and the wide-field imaging systems. The main difference in the molecular composition of white and gray matter is a higher myelin content in white matter [44], resulting in a higher lipid concentration. This explains why white matter overall shares the same Raman peaks as gray matter (Table 2). However, the higher signal-to-noise ratio in white matter is partly associated with the strong Raman cross-section of lipids, as evidenced by the higher white matter quality factor in Figure 3.

4 | DISCUSSION

Using the nylon phantom experiment, we can compare the intra- and inter-system deviations to assess the various systems measurement reproducibility. The intra-system deviation was typically below 3% for systems #1 to #9 and below 5% for the wide-field system, while these values were similar at 2% and 6% for the inter-system deviations of the multiple point probe systems and point probe vs wide-field systems, respectively. These deviations give, as a result, strong evidence for phantom measurement reproducibility regardless of the system used.

The ex vivo experiment on monkey brain demonstrates that a classification model can be exported from a training to a testing system with a low accuracy loss in the process, while the performance loss is higher in terms of sensitivity than specificity in the ex vivo point probe experiment. In this context, where sensitivity quantify the model ability to detect white matter, we expected that the highest quality factor associated with white matter measurements would be associated with a lower performance loss in sensitivity when transferring the model to other systems compared to the performance loss in specificity. However, due to the tissue identification error being overall negligible, the quantitative correlation between the quality factor and the performance metrics may not be relevant for this association.

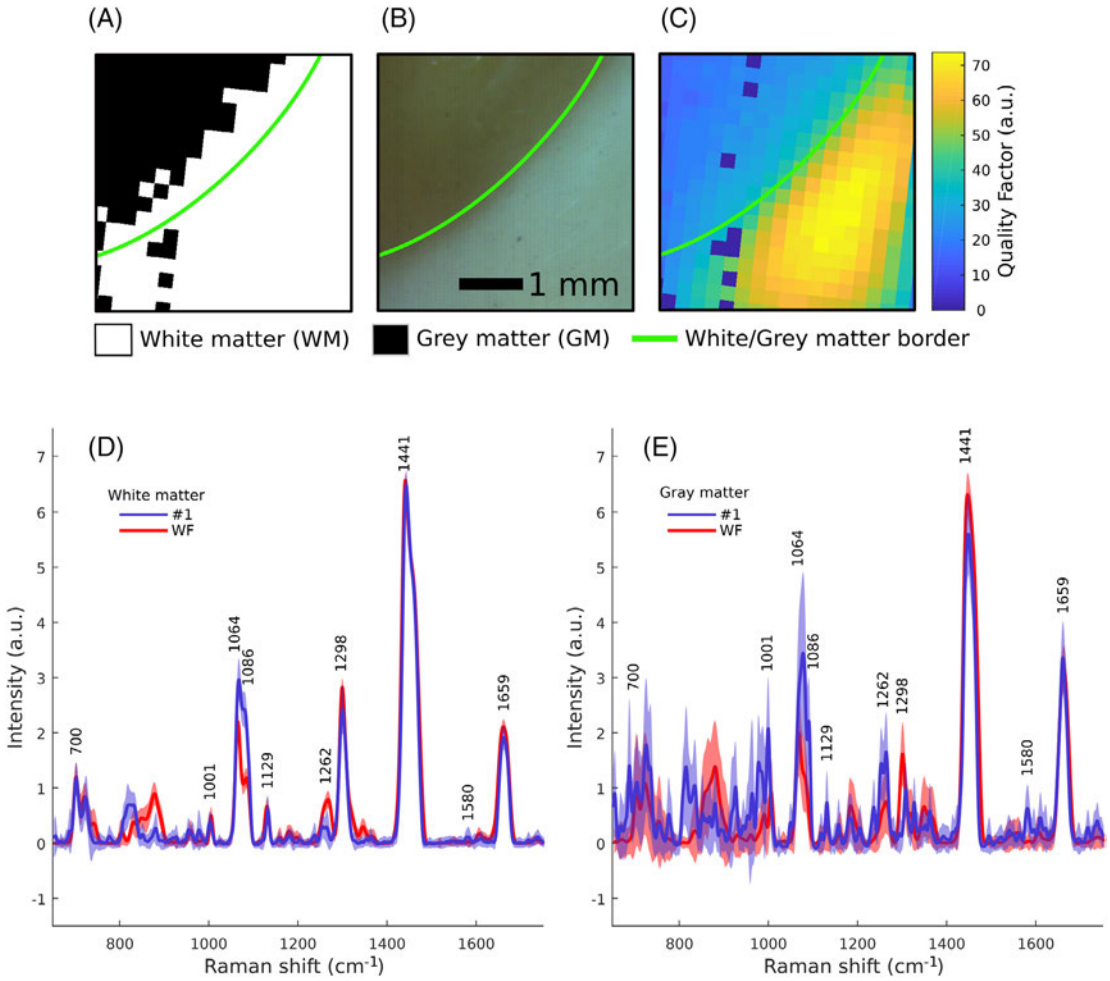

FIGURE 4 Ex vivo classification experimental results. A, Raman map of white matter (white pixels) and gray matter (black pixels) acquired with the wide-field system (#WF) and classified with the model built using the training system #1. B, White light reflectance image of the monkey brain acquired with the wide-field system (#WF). C, Quality factor map. D, Average and SD white matter spectra for system #1 and the wide-field system. E, Average and deviation gray matter spectra for system #1 and the wide-field system. The average and SD were calculated over data measurements acquired for the white matter, D, and for gray matter, E

The Raman map shown in Figure 4A returns a non-symmetrical margin error for diagnosis between white and gray matter, which is similar to results from another study where this system was used [34]. A plausible explanation to this phenomenon is the spatially offset Raman spectroscopy (SORS) effect caused by the wide-field illumination simultaneously to the line-scanning detection. To extract a depth-resolved Raman signature, this Raman imaging technique takes advantage of the spatial distance between the light collection occurring at the tissue surface and the point source Raman scattering event. In 2011, an in vitro study demonstrated Raman photons could be reliably detected over a distance of ≈ 2 mm through tissue [49]. This effect explains why the more prominent white matter signature, with higher lipids content, is detected over gray matter in a 500 μm margin band along the border between white and gray matter. As a result, the Raman map margin error displayed in Figure 4A is probably not due to the classification model or to the data calibration and processing, but instead to the wide-field system detection design. A solution to this limitation was recently proposed and consists in including optical measurements, from the border between tissues, in the training set so that the resulting classification model discriminates the biomolecular features from this region more accurately [50].

The Raman SNR and the corresponding quality factor extracted from the measurements appear to be critical for the development of high accuracy and generalizability of classification models. Optimizing these metrics mitigates Raman spectra deviations due to photonic noise and thus promotes increased ability of the Raman signature to separate between tissue classes. This is exemplified in our study by the difference of quality factor between white

and gray matter. A significant improvement of the acquisition method for these systems would be an automated acquisition parameter control to ensure a minimum quality factor value set by the user, regardless of the interrogated tissue. This would allow for consistent quality Raman signal across the entire data set.

This study presented preliminary evidence that standardized data calibration and processing enable the training of low complexity classification models, based on few spectral bands and their high transferability towards new data collected with different imaging systems. However, future work should focus on demonstrating the same results for more complex models, including more spectral bands. It should also investigate the classifier's transferability problematic in more challenging clinical problems such as in vivo human brain cancer detection for which spectral signature differences between tissues are more subtle [4, 24, 28].

5 | CONCLUSION

Ten Raman spectroscopy systems, including nine based on hand-held probes and one portable imaging instrument, were used to measure Raman signals on a nylon phantom and on monkey brain samples to evaluate the inter-system signal deviation and its effect on classification problems. Low inter-system deviations were observed and, using our standardized data calibration method, we also demonstrated that a two-class classification model can be trained using a training point probe-based system and transferred to an imaging-based system without significant increase of tissue identification error. These results demonstrate the feasibility of combining measurements from multiple Raman imaging systems in a centralized databank, opening the possibility of future training of statistical models based on a new scale of human clinical data set.

ACKNOWLEDGMENTS

This work is supported by the Discovery Grant program from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Collaborative Health Research Program (CIHR and NSERC) and the TransMedTech Institute. We would also like to thank Damon de Paoli, Daniel Côté and Martin Parent (CERVO brain research center, Université Laval) for providing the animal brains.

CONFLICT OF INTEREST

We do not declare any conflict of interest concerning this study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Fabien Picot https://orcid.org/0000-0001-5679-8638
Frédérick Dallaire https://orcid.org/0000-0002-3333-9014

REFERENCES

[1] M. Jermyn, J. Desroches, K. Aubertin, K. St-Arnaud, W. J. Madore, E. De Montigny, M. C. Guiot, D. Trudel, B. C. Wilson, K. Petrecca, F. Leblond, *Phys. Med. Biol.* **2016**, *61*, R370.
[2] I. Pence, A. Mahadevan-Jansen, *Chem. Soc. Rev.* **2016**, *45*, 1958.
[3] I. P. Santos, E. M. Barroso, T. C. B. Schut, P. J. Caspers, C. G. van Lanschot, D. H. Choi, M. F. van der Kamp, R. W. H. Smits, R. van Doorn, R. M. Verdijk, V. Noordhoek Hegt, J. H. von der Thüsen, C. H. M. van Deurzen, L. B. Koppert, G. J. L. H. van Leenders, P. C. Ewing-Graham, H. C. van Doorn, C. M. F. Dirven, M. B. Busstra, J. Hardillo, A. Sewnaik, I. ten Hove, H. Mast, D. A. Monserez, C. Meeuwis, T. Nijsten, E. B. Wolvius, R. J. Baatenburg de Jong, G. J. Puppels, S. Koljenović, *Analyst* **2017**, *142*, 3025.
[4] J. Desroches, É. Lemoine, M. Pinto, E. Marple, K. Urmey, R. Diaz, M. C. Guiot, B. C. Wilson, K. Petrecca, F. Leblond, *J. Biophotonics* **2019**, *12*, e201800396.
[5] R. Shams1, F. Picot, D. Grajales, G. Sheehy, F. Dallaire, M. Birlea, F. Saad, D. Trudel, C. Menard, F. Leblond, S. Kadoury, *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 1.
[6] E. Vargis, E. M. Kanter, S. K. Majumder, M. D. Keller, R. B. Beaven, G. G. Raod, A. Mahadevan-Jansen, *Analyst* **2011**, *136*, 2981.
[7] R. S. Shaikh, T. K. Dora, S. Chopra, A. Maheshwari, D. K. Kedar, R. Bharat, C. M. Krishna, *J. Biomed. Opt.* **2014**, *19*, 087001.
[8] M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Y. So, Z. Huang, *Abbrev. Biosens. Bioelectron* **2011**, *26*, 4104.
[9] M. S. Bergholt, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Y. So, A. Shabbir, Z. Huang, *J. Biophotonics* **2013**, *6*, 49.
[10] M. S. Bergholt, W. Zheng, K. Lin, Z. Huang, K. Y. Ho, K. G. Yeoh, M. Teh, J. B. Y. So, *J. Biomed. Opt.* **2011**, *16*, 037003.
[11] M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Y. So, Z. Huang, *Int. J. Cancer* **2011**, *128*, 2673.
[12] M. S. Bergholt, W. Zheng, K. Y. Ho, K. G. Yeoh, M. Teh, J. B. Y. So, Z. Huang, *Biomed. Vib. Spectrosc.* **2014**, *8939*, 89390M.
[13] J. Wang, K. Lin, W. Zheng, K. Y. Ho, M. Teh, K. G. Yeoh, Z. Huang, *Anal. Bioanal. Chem.* **2015**, *407*, 8303.
[14] Z. Huang, S. K. Teh, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, *Biosens. Bioelectron.* **2010**, *26*, 383.
[15] M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. Y. So, Z. Huang, *Analyst* **2010**, *135*, 3162.

[16] H. C. McGregor, M. A. Short, A. McWilliams, T. Shaipanich, D. N. Ionescu, J. Zhao, W. Wang, G. Chen, S. Lam, H. Zeng, *J. Biophotonics* **2017**, *10*, 98.
[17] H. C. McGregor, M. A. Short, S. Lam, T. Shaipanich, E. L. Beaudoin, H. Zeng, *J. Biophotonics* **2018**, *11*, e201800055.
[18] V. P. Zakharov, I. A. Bratchenko, O. O. Myakinin, D. N. Artemyev, Y. A. Khristoforova, S. V. Kozlov, A. A. Moryatov, *Ultrafast Nonlinear Imag. Spectrosc. II* **2014**, *9198*, 919804.
[19] J. Schleusener, P. Gluszczynska, C. Reble, I. Gersonde, J. Helfmann, J. W. Fluhr, J. Lademann, J. Röwert-Huber, A. Patzelt, M. C. Meinke, *Exp. Dermatol.* **2015**, *24*, 767.
[20] H. Lui, J. Zhao, D. McLean, H. Zeng, *Cancer Res.* **2012**, *72*, 2491.
[21] L. Lim, B. S. Nichols, M. R. Migden, N. Rajaram, J. Reichenberg, M. K. Markey, M. I. Ross, J. W. Tunnell, *J. Biomed. Opt.* **2014**, *19*, 117003.
[22] S. P. Singh, A. Sahu, A. Deshmukh, P. Chaturvedi, C. M. Krishna, *Analyst* **2013**, *138*, 4175.
[23] H. Krishna, S. K. Majumder, P. Chaturvedi, M. Sidramesh, P. K. Gupta, *J. Biophotonics* **2014**, *7*, 690.
[24] M. Jermyn, K. Mok, J. Mercier, J. Desroches, J. Pichette, K. Saint-Arnaud, L. Bernstein, M. C. Guiot, K. Petrecca, F. Leblond, *Sci. Transl. Med.* **2015**, *7*, 274ra19.
[25] J. Desroches, M. Jermyn, K. Mok, C. Lemieux-Leduc, J. Mercier, K. St-Arnaud, K. Urmey, M. C. Guiot, E. Marple, K. Petrecca, F. Leblond, *Biomed. Opt. Express* **2015**, *6*, 2380.
[26] M. Jermyn, J. Desroches, J. Mercier, M. A. Tremblay, K. St-Arnaud, M. C. Guiot, K. Petrecca, F. Leblond, *J. Biomed. Opt.* **2016**, *21*, 094002.
[27] M. Jermyn, J. Desroches, J. Mercier, K. St-Arnaud, M. C. Guiot, F. Leblond, K. Petrecca, *Biomed. Opt. Express* **2016**, *7*, 5129.
[28] M. Jermyn, J. Mercier, K. Aubertin, J. Desroches, K. Urmey, J. Karamchandiani, E. Marple, M. C. Guiot, F. Leblond, K. Petrecca, *Cancer Res.* **2017**, *77*, 3942.
[29] K. Lin, W. Zheng, J. Wang, C. M. Lim, Z. Huang, *Photonic Ther. Diagnostics XII* **2016**, *9689*, 96892B.
[30] M. S. Bergholt, K. Lin, J. Wang, W. Zheng, H. Xu, Q. Huang, J. L. Ren, K. Y. Ho, M. Teh, S. Srivastava, B. Wong, K. G. Yeoh, Z. Huang, *J. Biophotonics* **2016**, *9*, 333.
[31] J. Desroches, M. Jermyn, M. Pinto, F. Picot, M. A. Tremblay, S. Obaid, E. Marple, K. Urmey, D. Trudel, G. Soulez, M. C. Guiot, B. C. Wilson, K. Petrecca, F. Leblond, *Sci. Rep.* **2018**, *8*, 1.
[32] T. Bhattacharjee, L. C. Fontana, L. Raniero, J. Ferreira-Strixino, *J. Raman Spectrosc.* **2018**, *49*, 786.
[33] K. St-Arnaud, K. Aubertin, M. Strupler, M. Jermyn, K. Petrecca, D. Trudel, F. Leblond, *Opt. Lett.* **2016**, *41*, 4692.
[34] K. St-Arnaud, K. Aubertin, M. Strupler, W. J. Madore, A. A. Grosset, K. Petrecca, D. Trudel, F. Leblond, *Med. Phys.* **2018**, *45*, 328.
[35] A. Akbarzadeh, E. Edjlali, G. Sheehy, J. Selb, R. Agarwal, J. Weber, F. Leblond, *J. Biomed. Opt.* **2020**.
[36] A. Shoji, T. Ozaki, T. Fujito, K. Deguchi, S. Ando, I. Ando, *Macromolecules* **1989**, *22*, 2863.
[37] F. Dallaire, F. Picot, J. P. Tremblay, G. Sheehy, É. Lemoine, R. Agarwal, S. Kadoury, D. Trudel, F. Lesage, K. Petrecca, F. Leblond, *J. Biomed. Opt.* **2020**, *25*, 040501.
[38] R. Perez-Pueyo, M. J. Soneira, S. Ruiz-Moreno, *Appl. Spectrosc.* **2010**, *64*, 595.
[39] M. Pinto, K. C. Zorn, J. P. Tremblay, J. Desroches, F. Dallaire, K. Aubertin, E. T. Marple, C. Kent, F. Leblond, D. Trudel, F. Lesage, *J. Biomed. Opt.* **2019**, *24*, 025001.
[40] K. Aubertin, V. Q. Trinh, M. Jermyn, P. Baksic, A. A. Grosset, J. Desroches, K. St-Arnaud, M. Birlea, M. C. Vladoiu, M. Latour, R. Albadine, F. Saad, F. Leblond, D. Trudel, *BJU Int.* **2018**, *122*, 326.
[41] K. Aubertin, J. Desroches, M. Jermyn, V. Q. Trinh, F. Saad, D. Trudel, F. Leblond, *Biomed. Opt. Express* **2018**, *9*, 4294.
[42] C. Cortes, *Mach. Learn.* **1995**, *20*, 273.
[43] A. A. Grosset, F. Dallaire, T. Nguyen, M. Birlea, J. Wong, F. Daoust, N. Roy, A. Kougioumoutzakis, F. Azzi, K. Aubertin, S. Kadoury, M. Latour, R. Albadine, S. Prendeville, P. Boutros, M. Fraser, R. G. Bristow, T. van der Kwast, M. Orain, H. Brisson, N. Benzerdjeb, H. Hovington, A. Bergeron, Y. Fradet, B. Têtu, F. Saad, F. Leblond, D. Trudel, *PLoS Med.* **2020**, *17*, e1003281.
[44] J. S. O. Brien, E. L. Sampson, *J. Lipid Res.* **1965**, *6*, 537.
[45] P. Le Cacheux, G. Menard, H. N. Quang, P. Weinmann, M. Jouan, N. Q. Dao, *Appl. Spectrosc. Rev.* **1996**, *50*, 1253.
[46] É. Lemoine, F. Dallaire, R. Yadav, R. Agarwal, S. Kadoury, D. Trudel, M. C. Guiot, K. Petrecca, F. Leblond, *Analyst* **2019**, *144*, 6517.
[47] A. C. S. Talari, Z. Movasaghi, S. Rehman, I. U. Rehman, *Appl. Spectrosc. Rev.* **2015**, *50*, 46.
[48] M. Jové, I. Pradas, M. Dominguez-Gonzalez, I. Ferrer, R. Pamplona, *Redox Biol.* **2019**, *23*, 101082.
[49] M. D. Keller, E. Vargis, A. Mahadevan-Jansen, N. de Matos Granja, R. H. Wilson, M. A. Mycek, M. C. Kelley, *J. Biomed. Opt.* **2011**, *16*, 077006.
[50] F. Daoust, T. Nguyen, P. Orsini, J. Bismuth, M. M. De Denus-Baillargeon, I. Veilleux, A. Wetter, P. Mckoy, I. Dicaire, M. Massabki, K. Petrecca, F. Leblond, *J. Biomed. Opt.* **2020**.

How to cite this article: Picot F, Daoust F, Sheehy G, et al. Data consistency and classification model transferability across biomedical Raman spectroscopy systems. *Translational Biophotonics*. 2021;3:e202000019. https://doi.org/10.1002/tbio.202000019