Chaos

Diffuse reflectance and machine learning techniques to differentiate colorectal cancer *ex vivo*

Cite as: Chaos **31**, 053118 (2021); https://doi.org/10.1063/5.0052088
Submitted: 29 March 2021 . Accepted: 20 April 2021 . Published Online: 17 May 2021

Luís Fernandes, Sónia Carvalho, Isa Carneiro, Rui Henrique, Valery V. Tuchin, Hélder P. Oliveira, and Luís M. Oliveira

AIP Publishing

Diffuse reflectance and machine learning techniques to differentiate colorectal cancer *ex vivo*

Cite as: Chaos **31**, 053118 (2021); doi: 10.1063/5.0052088
Submitted: 29 March 2021 · Accepted: 20 April 2021 ·
Published Online: 17 May 2021

View Online Export Citation CrossMark

Luís Fernandes,[1,2] Sónia Carvalho,[3,4] Isa Carneiro,[3] Rui Henrique,[3,5] Valery V. Tuchin,[6,7,8] Hélder P. Oliveira,[9,10] and Luís M. Oliveira[1,2,a)]

AFFILIATIONS

[1]Center for Innovation in Engineering and Industrial Technology, Polytechnic of Porto—School of Engineering, 4249-015 Porto, Portugal
[2]Physics Department, Polytechnic of Porto—School of Engineering, 4249-015 Porto, Portugal
[3]Department of Pathology and Cancer Biology and Epigenetics Group-Research Center, Portuguese Oncology Institute of Porto, 4200-072 Porto, Portugal
[4]Department of Pathology, Santa Luzia Hospital, ULSAM, 4904-858 Viana do Castelo, Portugal
[5]Department of Pathology and Molecular Immunology, Institute of Biomedical Sciences Abel Salazar-University of Porto (ICBAS-UP), 4050-313 Porto, Portugal
[6]Science Medical Center, Saratov State University, Saratov 410012, Russia
[7]Interdisciplinary Laboratory of Biophotonics, National Research Tomsk State University, Tomsk 634050, Russia
[8]Laboratory of Laser Diagnostics of Technical and Living Systems, Institute of Precision Mechanics and Control of the Russian Academy of Sciences, Saratov 410028, Russia
[9]Institute for Systems and Computer Engineering, Technology and Science, INESC TEC, 4200-465 Porto, Portugal
[10]Faculty of Science, University of Porto, FCUP, 4169-007 Porto, Portugal

Note: This paper is part of the Focus Issue, In Memory of Vadim S. Anishchenko: Statistical Physics and Nonlinear Dynamics of Complex Systems.

[a)]**Author to whom correspondence should be addressed:** lmo@isep.ipp.pt

ABSTRACT

In this study, we used machine learning techniques to reconstruct the wavelength dependence of the absorption coefficient of human normal and pathological colorectal mucosa tissues. Using only diffuse reflectance spectra from the *ex vivo* mucosa tissues as input to algorithms, several approaches were tried before obtaining good matching between the generated absorption coefficients and the ones previously calculated for the mucosa tissues from invasive experimental spectral measurements. Considering the optimized match for the results generated with the multilayer perceptron regression method, we were able to identify differentiated accumulation of lipofuscin in the absorption coefficient spectra of both mucosa tissues as we have done before with the corresponding results calculated directly from invasive measurements. Considering the random forest regressor algorithm, the estimated absorption coefficient spectra almost matched the ones previously calculated. By subtracting the absorption of lipofuscin from these spectra, we obtained similar hemoglobin ratios at 410/550 nm: 18.9-fold/9.3-fold for the healthy mucosa and 46.6-fold/24.2-fold for the pathological mucosa, while from direct calculations, those ratios were 19.7-fold/10.1-fold for the healthy mucosa and 33.1-fold/17.3-fold for the pathological mucosa. The higher values obtained in this study indicate a higher blood content in the pathological samples used to measure the diffuse reflectance spectra. In light of such accuracy and sensibility to the presence of hidden absorbers, with a different accumulation between healthy and pathological tissues, good perspectives become available to develop minimally invasive spectroscopy methods for *in vivo* early detection and monitoring of colorectal cancer.

The application of machine learning methods to noninvasive-like diffuse reflectance spectra allowed us to reconstruct the absorption coefficient spectra of human healthy and pathological mucosa tissues from the colorectal wall. Consequently, we were able to obtain differentiated blood and pigment content in both tissues, which can be used for the development of new noninvasive diagnostic methods for colorectal cancer.

I. INTRODUCTION

The optical properties of biological tissues condition how light beams propagate inside those tissues and interact with their biological components. There are some optical properties that can be estimated/calculated for a biological material, but the most commonly used are the refractive index (RI), the absorption coefficient (μ_a), the scattering coefficient (μ_s), and the anisotropy-factor (g).[1–3] These fundamental properties quantify the speed of light inside the material and the number of photons that are absorbed and/or scattered per unit length inside the medium and characterize the mean directionality of such scattering.[2] The wavelength dependence of the optical properties of a tissue provides useful information for the optimization of current optical methods in clinical practice or for the development of new methods that operate at individual wavelengths within the electromagnetic spectrum.[4] There are some traditional tissue windows, located at certain wavelength ranges, where current optical diagnostic and therapeutic methods work: I (625–975 nm), II (1100–1350 nm), III (1600–1870 nm), and IV (2100–2300 nm).[5,6] In addition to these natural tissue windows, where local maxima for the light penetration depth are observed,[3,7] other optical diagnostic and treatment windows can be induced through the application of optical clearing treatments, as recently demonstrated for the ultraviolet (UV) range with transmittance efficiency peaks at 230, 275, and 300 nm.[8,9]

Considering such recent discovery of UV-induced-windows to diagnose and treat pathologies, the current optical methods that work at visible and near infrared (NIR) wavelengths, and the emerging THz techniques for clinical application, the need to map the optical properties of both normal and pathological tissues in a wide spectral range becomes urgent.[10] The estimation or calculation of a set of optical properties for any tissue provides individual information, such as an identity card for that tissue, and consequently their evaluation might help in the discrimination of pathologies. By knowing the optical properties of tissues, it is also possible to create individual light propagation models that can be used in the development of novel noninvasive optical diagnostic and treatment procedures.[2,11,12]

The traditional methods to estimate the optical properties of biological tissues rely on performing inverse simulations that use codes,[2,3] which were constructed based on the radiation transfer theory, such as the Monte Carlo,[13] or the adding-doubling[14] methods. Being part of a wider range of estimation methods, these two are the most precise in the estimation of the optical properties.[2] To perform an estimation of the optical properties of a tissue with any of these methods, a set of optical measurements that were experimentally acquired from a tissue sample are used as input in the inverse simulation. The simulation code uses arbitrary optical properties to generate the total transmittance (T_t), the total reflectance (R_t), and possibly the collimated transmittance (T_c) for the tissue under study. These generated values are compared with the corresponding experimental values that were introduced as input to the simulations to check the difference. While the difference between the generated and the experimental values is above a certain limiting value, the optical properties in the simulation are corrected and the simulation runs again.[3,15] When the difference between the generated and the experimental values is minimal, the simulation ends and the last set of optical properties used is presented to the user. The major problem with these estimation codes is that each simulation generates results for a single wavelength,[3,14,15] turning these methods time-consuming if we want to estimate the optical properties for a tissue in a wide spectral range.

Several studies using the inverse Monte Carlo (IMC) or the inverse Adding-Doubling (IAD) methods were performed for various tissues to estimate their optical properties at individual wavelengths, which were later used to determine their wavelength dependence for a selected spectral range.[7,11,16–23] Since only the wavelength dependence of μ_a is not well described by an equation, an alternate calculation method has been recently proposed, which obtains almost all spectral optical properties directly from experimental spectra that were acquired from the tissue samples.[4,24] In this method, the T_t, R_t, and T_c spectra are measured from thin tissue samples for a wide spectral range. To obtain the μ_a spectrum of the tissue sample, a simple calculation, as described by Eq. (1) that uses the sample thickness (d) and the T_t and R_t spectra, can be made,[24]

$$\mu_a(\lambda) = \frac{1 - \left(\frac{T_t(\lambda)+R_t(\lambda)}{100}\right)}{d}. \tag{1}$$

Usually, $T_t(\lambda)$ and $R_t(\lambda)$ are measured in percentage, meaning that to perform the calculation with Eq. (1), these spectra need to be normalized (divided by 100) to vary between 0 and 1. If d is represented in cm, $\mu_a(\lambda)$ will be calculated in cm^{-1}.[4,24] To obtain the scattering coefficient spectrum, $\mu_s(\lambda)$, the Bouguer–Beer–Lambert (BBL) equation[25] can be used as represented by Eq. (2), where $T_c(\lambda)$ is also divided by 100 to vary between 0 and 1, as described above for $T_t(\lambda)$ and $R_t(\lambda)$ in Eq. (1),[2,3,24]

$$\mu_s(\lambda) = -\frac{ln\left[\frac{T_c(\lambda)}{100}\right]}{d} - \mu_a(\lambda). \tag{2}$$

In Eq. (2), $\mu_a(\lambda)$ is the one calculated through Eq. (1).[4,24] To obtain the wavelength dependence for the reduced scattering coefficient (μ_s'), a set of IAD simulations, performed at individual wavelengths within the range of interest,[4,24] needs to be performed. The IAD code generates μ_s' with significant precision,[26] and since its wavelength dependence is well described mathematically,[1] simulations at a certain number of wavelengths are sufficient.[24] Once these discrete values of μ_s' are estimated, they can be fitted with a curve described by Eq. (3),[24]

$$\mu_s'(\lambda) = a' \times \left(f_{Ray} \times \left(\frac{\lambda}{500\,\text{nm}}\right)^{-4} + (1 - f_{Ray}) \times \left(\frac{\lambda}{500\,\text{nm}}\right)^{-b_{Mie}} \right), \tag{3}$$

which accounts both for the Rayleigh and the Mie scattering regimes.[1] In Eq. (3), a' is a normalizing factor that represents the

reduced scattering coefficient of the tissue at 500 nm, f_{Ray} represents the Rayleigh scattering fraction, and b_{Mie} is the mean size of the Mie scatterers.[24] Equation (3), which can also be used to fit $\mu_s(\lambda)$,[24] has been successfully used to fit the wavelength dependence for data of many biological soft tissues.[1,27,28] Once $\mu_s(\lambda)$ is obtained through Eq. (2) and $\mu'_s(\lambda)$ is obtained through Eq. (3), they can be combined in Eq. (4) to obtain $g(\lambda)$,[24]

$$g(\lambda) = 1 - \frac{\mu'_s(\lambda)}{\mu_s(\lambda)}. \quad (4)$$

Another useful optical property, the light penetration depth (δ), can also be calculated from μ_a and μ'_s,[3,16,24]

$$\delta(\lambda) = \frac{1}{\sqrt{3\mu_a(\lambda) \times (\mu_a(\lambda) + \mu'_s(\lambda))}}. \quad (5)$$

This calculation method, which only needs IAD simulations to estimate $\mu'_s(\lambda)$, is fast in the determination of the wavelength dependencies of the optical properties of biological tissues. A disadvantage of this method is that to perform such calculations, spectral measurements, which are collected from *ex vivo* tissue samples, are necessary. A noninvasive or minimally invasive method that could estimate the optical properties of *in vivo* tissues and their wavelength dependence without the need for sample excision would be a valuable tool in clinical practice and in the detection of pathologies, even at their early stage of development.

To develop such an innovative method, new approaches are necessary. One particular and interesting approach relies on the combination of noninvasive-like spectral measurements, such as diffuse reflectance (R_d), with *machine learning* (ML) techniques. Some of these ML techniques consist of the application of neural network geometries, without the need for significant programing to develop a model that can predict a desired outcome from specific experimental data.[29] Such techniques have already proven useful in the imaging and spectroscopy fields of biophotonics,[30] and they can be used for a fast estimation of the spectral optical properties of biological tissues from noninvasive optical measurements.

The ML method, which consists of the development of a model that learns to calculate desired outcomes, was first proposed by McCulloch and Pitts.[31] In this paper, the authors presented a mathematical model that was able to reproduce the behavior of the nervous system from experimentally collected data. Since then, different learning strategies have been used, and nowadays ML is used to automate functions such as classification or estimation of features, without any specific programming.[32] During the learning process, the parameters of the ML model are incrementally adjusted so that it can reproduce the desired outcome.

By exploring this ability, it is possible to produce a model that can estimate $\mu_a(\lambda)$ from R_d spectra, without the need to define a specific relation between the two variables. The use of Monte Carlo simulations to create Look-Up tables (LUT), and more recently, the use of ML algorithms to estimate the optical properties of tissues have been reported.[33–36] An example of the use of mathematical models to estimate the optical properties of tissues is described in Ref. 33, where the authors defined a formula and fitted it to empirical data to retrieve the μ_a values to evaluate the goodness of the fit. In Ref. 34, the authors used previously generated LUT to find an initial estimation of the optical properties. Using this initial estimation in a Monte Carlo simulation, they generated the diffuse reflectance spectrum to compare with measured spectra. By adjusting the simulation parameters, the following simulations generated new R_d spectra that were better matched to the measured $R_d(\lambda)$. In another work,[35] a neural network was used to estimate the optical properties, based on R_d values estimated with the radiation transfer equation. In this work, the neural network consisted of three layers: one input layer with eight nodes, a second layer with also eight nodes, and an output layer of two nodes. Reference 36 shows further examples on the use of ML methods to estimate the optical properties of biological tissues. In this study, the authors used a random forest regressor with 15 estimators to obtain the μ'_s and μ_a values. Once again, the data used to train the model were generated with Monte Carlo simulations.

Considering the estimation of the spectral optical properties of biological tissues, the ML method can also be a powerful tool. In opposition to the traditional IMC or IAD simulations, it allows estimating the entire spectral properties at once, and the model can be trained with *in vivo* spectral data as input, which opens the possibility of developing noninvasive diagnostic protocols.

There are several ML algorithms available that can be applied to spectral data, such as the single layer perceptron (SLP), the random forest regressor, the K-nearest neighbor (KNN), the decision tree for multioutput regression (DTFMR), and the linear regression for multioutput (LRFMO).[37] Depending on the available experimental data to use in the learning process to develop the desired model, some of these algorithms may have better performance than others. This means that for a specific task, the various ML algorithms must be tested first to check which is the one that can produce better estimations of the desired outcome.

With the objective of exploring the use of ML techniques in the estimation of tissues' spectral optical properties, we have measured the R_d spectra from human normal and pathological (adenocarcinoma) colorectal mucosa tissues to reconstruct their $\mu_a(\lambda)$. Since the μ_a spectra for these tissues were previously calculated with Eq. (1) and allowed the discrimination of colorectal cancer through the evaluation of differentiated content of a pigment,[24] we tried to reproduce those calculations and obtain the same results. In this study, and having the objective to establish a protocol that can be used in future *in vivo* and noninvasive (or minimally invasive) detection of colorectal cancer, we used the methodology described in Sec. II to obtain the results presented in Sec. III.

II. MATERIALS AND METHODS

A. Tissue sample collection and preparation

All the tissue samples used in the present study were collected from the mucosa layer of the human colorectal wall. Following the guidelines of the Ethics Committee of the Portuguese Oncology Institute of Porto (Portugal), the healthy and pathological (adenocarcinoma) areas were separated from the surgical resections of patients under treatment at that institution. To confirm the diagnosis, a histological analysis of the surgical specimens was the gold standard for tissue examination. All the cases were classified (according to the current World Health Organization classification) as "colorectal adenocarcinoma not otherwise specified" (a malignant

epithelial tumor originating from glandular cells in the superficial colorectal layer, which comprises about 90% of all colorectal cancers).[38]

A cryostat (Leica™, model CM 1850 UV) was used to prepare the mucosa samples for the present study. Ten samples were prepared from the healthy areas and ten samples were prepared from the pathological areas, having an approximated circular form, with a diameter of about 1 cm and uniform thickness (d) of 0.5 mm. All these samples were submitted to spectral diffuse reflectance ($R_d(\lambda)$) measurements, as described next.

B. Spectral measurements

To calculate the reference μ_a spectra to use in the present study, T_t and R_t spectra were acquired from both healthy and pathological mucosa samples between 200 and 1000 nm. Ten healthy and ten pathological samples were submitted to those measurements, using the setups presented in Figs. 1(a) and in 1(b).

Although those measurements and calculations were made in the study of Ref. 24, we will describe them here for better perception. In the present study, to perform the estimation of $\mu_a(\lambda)$ through ML algorithms, R_d spectra were also necessary to be acquired. Considering ten new healthy and ten new pathological colorectal tissue samples, we performed those measurements using the setup presented in Fig. 1(c). All the R_d spectral measurements were also acquired between 200 and 1000 nm.

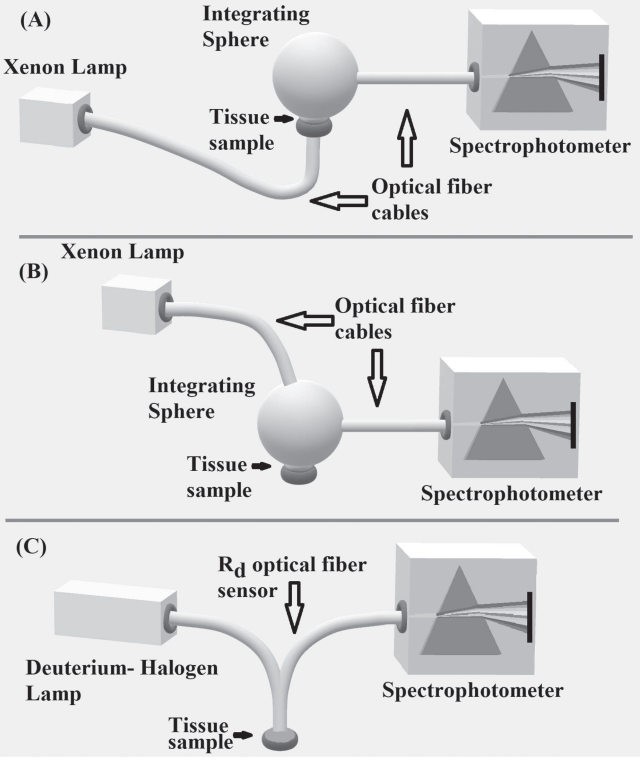

FIG. 1. Experimental setups to measure T_t (a), R_t (b), and R_d (c) spectra.

Considering the T_t setup [Fig. 1(a)], a broadband pulsed xenon lamp was used to illuminate the tissue sample. The beam was delivered to the sample by using an optical fiber cable and a collimating lens (below the tissue sample), which limited the beam diameter to 6 mm before reaching the sample. The transmitted light entered an integrating sphere, where it was integrated before being delivered to the spectrometer through another optical fiber cable. The R_t setup is similar to the one used to acquire the T_t measurements and it uses the same optical components. The only difference is that sample illumination is made through the integrating sphere, to integrate the reflected beam, instead of integrating the transmitted beam—see Fig. 1(b). For the R_d measurements, to which all healthy and pathological samples were also submitted, the sample was illuminated with a broadband deuterium-halogen lamp. An R_d optical fiber sensor was used both to illuminate the sample and to collect its diffuse reflected light—see Fig. 1(c). In these measurements, the tip of the R_d sensor was kept at a constant distance of 2 mm above the tissue sample's surface. All of this equipment was acquired from Avantes™ (Apeldoorn, The Netherlands), with the exception of the optical fiber sensor to measure R_d, which was kindly supplied to our research by ArtPhotonics™ (Berlin, Germany). This sensor contained a detection fiber at the center, which was surrounded by seven irradiation fibers. All fibers were made of silica with a core diameter of 400 μm and a cladding thickness of 20 μm. Their numerical aperture was 0.22. The detection fiber at the center had an aluminum cover, having a total diameter of 560 μm. The irradiation fibers had a polyamide cover, and their global diameter was 465 μm. These fibers were packed around the detection fiber, without any spacing. With this geometry and dimensions, the source-detector separation was 512.5 μm.

C. Calculation of $\mu_a(\lambda)$

After all spectra were collected, calculations were made to obtain the wavelength dependence for all optical properties of the healthy and pathological colorectal mucosa. The description of those calculations and the corresponding results for the mean optical properties of the mucosa tissues were already published in Ref. 24. Considering, in particular, the wavelength dependence of μ_a, sample thickness, T_t and R_t spectra were used in Eq. (1) to calculate $\mu_a(\lambda)$ for each mucosa condition. After calculating 10 μ_a spectra for the healthy and 10 μ_a spectra for the pathological mucosa, mean and standard deviation (SD) were calculated for each case. These mean spectra were considered as reference in the present study in the estimations with ML models. Those estimations were made from R_d spectra measured from similar tissue samples, with the objective of reconstructing the mean $\mu_a(\lambda)$ of both mucosa tissues. Our objective was to evaluate if the estimated data can also be used to discriminate cancer. The ML estimation procedure is described in Sec. II D, and the final calculations to check if the estimated data can be used for cancer detection are described in Sec. II E.

D. Machine learning to estimate $\mu_a(\lambda)$

Different ML algorithms were tested, with empirical parameter tuning, to access the best model to estimate $\mu_a(\lambda)$ from $R_d(\lambda)$ data. Furthermore, during these computational experiments, the

models were trained in either of the two ways: (a) using only normal or pathological samples (Trained Separately—TS model) or (b) using all the samples (Trained Together—TT model) with further separation of the estimated spectra in normal and pathological categories.

The Single Layer Perceptron (SLP) model was created using the TensorFlow framework for Python. Since the acquired experimental spectra ranged from 200 to 1000 nm with a 1 nm resolution, the input dimension was set to 801 in the model so that data at each wavelength from an experimental R_d spectrum are interpreted by using the SLP model as a feature. The SLP model architecture was set into two layers: one with ten output nodes (to generate ten spectra) and a second with 801 output nodes (to obtain same spectral resolution and bandwidth as the measured R_d spectra). Since we had a small amount of spectra (ten for healthy and ten for pathological mucosa tissues), we decided to minimize the architecture dimension in order to prevent overfitting.

In the K-Nearest Neighbor (KNN) algorithm, the number of neighbors was set as 5, since further increments tended to increase the error on the spectral shape estimation. When fine-tuning the parameters of the KNN model, the k value was set between 1 and 9 for the models that were trained with only normal or pathological data and between 1 and 19 for the model that was trained with all the data. Taking into account the initial computational experiments, and since higher increments did not improve the spectral shape estimation, the number of trees in the Random Forest Regression (RFR) was set to 5. For the Decision Tree for Multioutput Regression (DTFMR) algorithm, the depth was fixed at 4, in order to prevent overfitting. With respect to the Linear Regression for Multioutput (LRFMO) algorithm, the simplest models used in this study automatically found the best slope for the data fitting.

The neural network was implemented using the TensorFlow framework for Python, using all the ML algorithms available in the *scikit learn library*, also available for the Python language. Due to the low amount of experimental spectra, the Leave One Out (LOO) method was adopted.[39] This method consists of the following:

a. place a random μ_a spectrum out,
b. train the model using the other μ_a spectra,
c. evaluate the model performance with the μ_a spectrum that was left out in the training,
d. compare between the estimated μ_a spectra and the reference ones that were calculated from invasive measurements, and
e. repeat the entire process, leaving a different μ_a out at each time (this process is repeated the number of times equal to the number of samples used).

After having the individual estimated μ_a spectra, the mean $\mu_a(\lambda)$ was calculated and compared with the mean $\mu_a(\lambda)$ that was calculated from invasive measurements (considered as reference spectra). To evaluate the performance of the different ML algorithms, the Euclidean Distance (ED) was calculated for each wavelength, between the estimated and the corresponding reference spectrum, which was calculated from invasive measurements. The formula used to calculate the ED in one dimension is the following:

$$ED = |a - b|, \tag{6}$$

where a represents μ_a for the estimated spectrum and b represents μ_a for the reference spectrum, as retrieved from the calculations based on invasive measurements. As previously referred, for each ML algorithm, the models were trained with μ_a data from normal or pathological samples or with μ_a data from all samples available with later classification of the estimated spectra as normal or pathological, depending on the previous known sample category. In future works, we plan to perform an automatic classification of the estimated μ_a spectra.

E. Pigment accumulation estimation

After estimating the mean μ_a spectra both for the healthy and the pathological mucosa tissues, we selected the estimations that presented better performance to implement a final calculation. In our previous study,[24] we suspected that a hidden absorber in the mucosa tissues was camouflaging the true blood content. In that study, we concluded that such an absorber was a pigment called lipofuscin with an absorption coefficient spectrum as described by Eq. (7),[40]

$$\mu_{a\text{-lip}}(\lambda) = A \times 5.2 + A \times e^{(3.524 - 0.01087 \times \lambda)}. \tag{7}$$

In Eq. (7), $\mu_{a-lip}(\lambda)$ represents the wavelength dependence for the absorption coefficient of lipofuscin, represented in cm^{-1}; λ is the wavelength (in nm) for the range between 200 and 1000 nm; and A represents the lipofuscin content in the tissue, which should be 1 for the normal mucosa and 1.1 for the pathological mucosa (10% more content), as determined in the study of Ref. 24. Here, we performed the same calculation to compare results with the ones obtained only from invasive measurements and so quantify the accuracy of the different ML algorithms. Assuming here also a different lipofuscin content in the healthy and pathological tissues ($A = 1$ for the healthy mucosa and $A = 1.1$ for the pathological mucosa), we subtracted $\mu_{a-lip}(\lambda)$ from the μ_a spectra of both tissues to obtain the accurate blood ratios at the hemoglobin bands (410 and 550 nm). Such calculation was made with the estimations produced by the SLP, KNN, and RFR algorithms, which were the ones that presented better performance. A comparison between the hemoglobin ratios obtained in the present study and the ones obtained in the study of Ref. 24 was made.

III. RESULTS AND DISCUSSION

We initiated this experimental study by measuring the R_d spectra from ten healthy and ten pathological mucosa samples. Considering the T_t and R_t spectra from the study reported in Ref. 24 and sample thickness of $d = 0.05$ cm in Eq. (1), we calculated $\mu_a(\lambda)$ for each particular sample. Figure 2 presents the mean R_d spectra and the mean calculated μ_a spectra for both colorectal mucosa tissues, with the data from the normal mucosa identified as NM and the data from the pathological mucosa identified as PM.

Using the individual R_d and μ_a spectra that originated from the mean results presented in Fig. 2, we started to develop the ML models with different approaches. The individual estimations for all models can be seen in Figs. S1–S20 in the supplementary material. The first approach consisted of using the SLP model. As previously indicated, the models were trained in two distinct ways: TS or TT. The ten spectral estimations obtained with the LOO method were

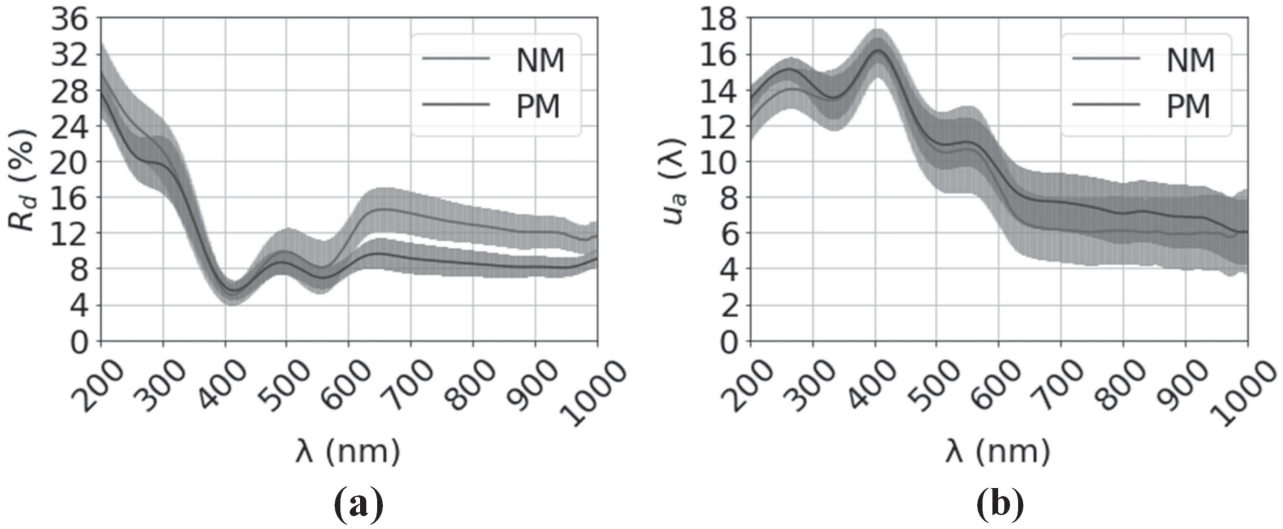

FIG. 2. Mean R_d (a) and μ_a (b) spectra of the normal (NM-green) and pathological (PM-red) mucosa.

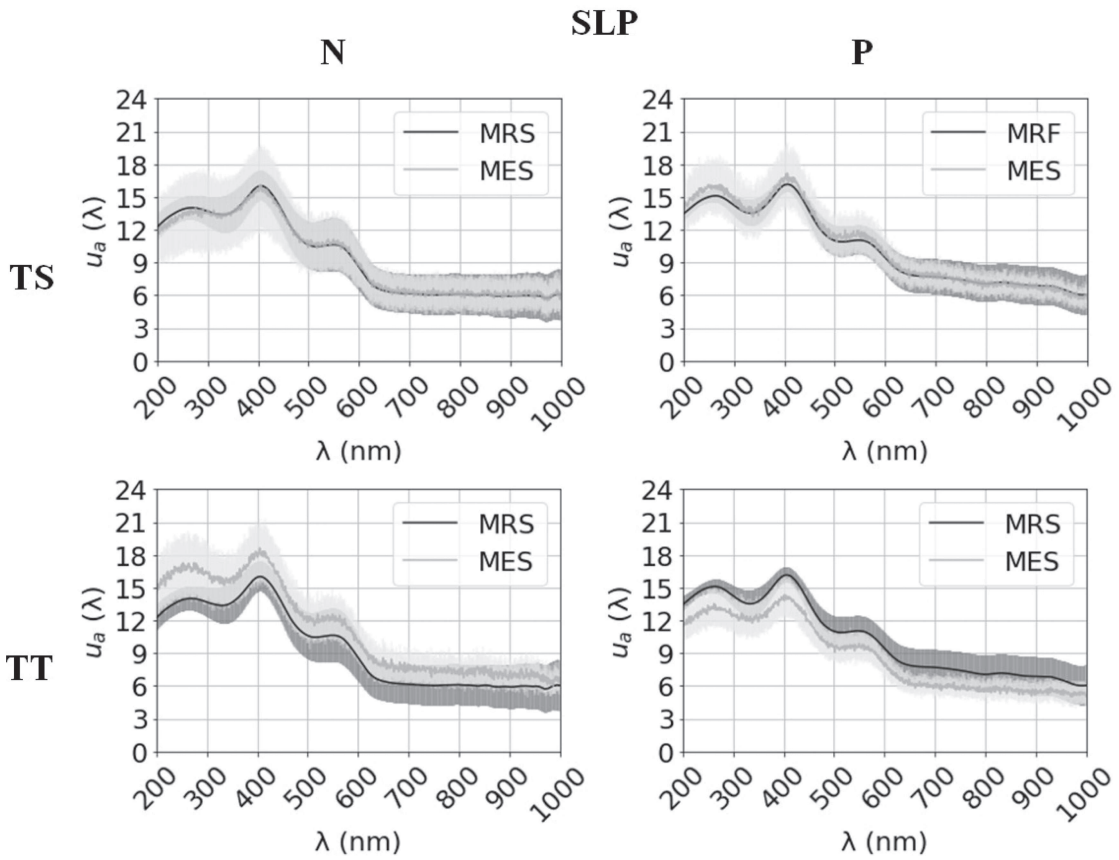

FIG. 3. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the SLP algorithm.

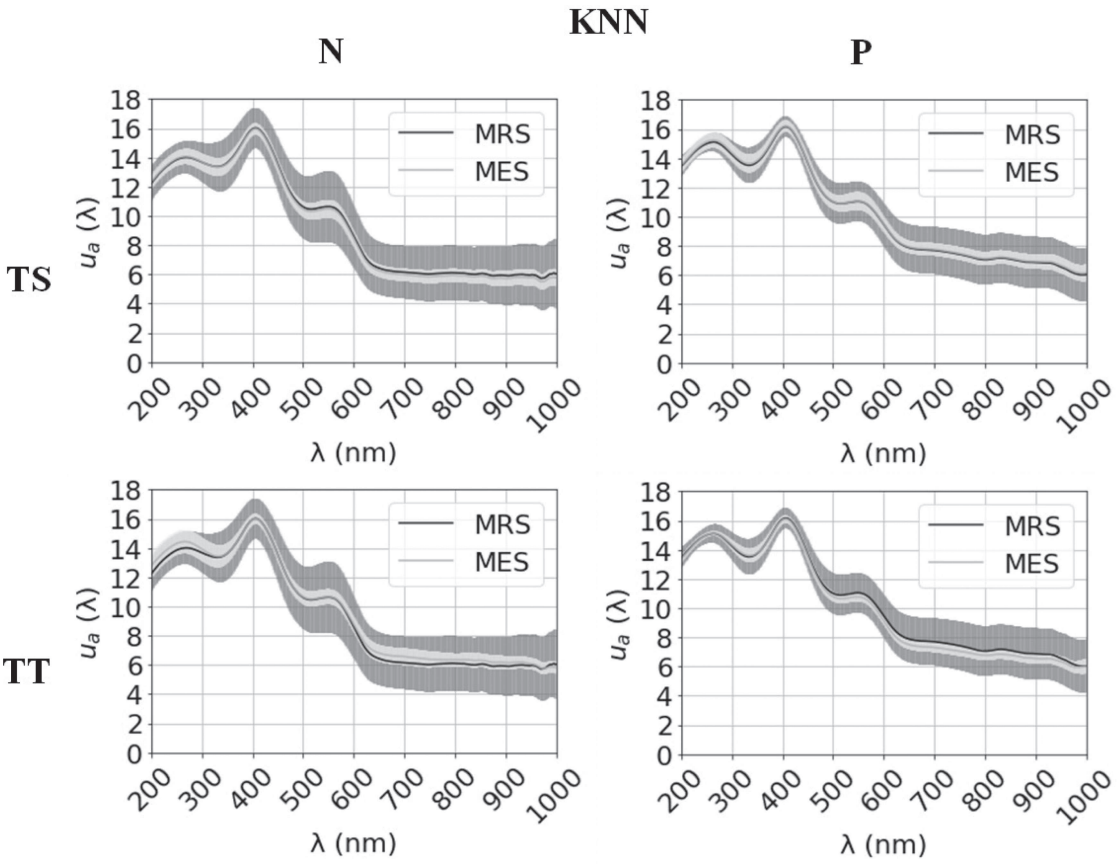

FIG. 4. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the KNN algorithm.

averaged to compare with the mean reference spectra (MRS) for both tissues. The mean results obtained with this approach are presented in Fig. 3, where the left panels correspond to the healthy tissue (N) and the right panels correspond to the pathological (P) tissue.

By making a comparison between the mean estimated spectra that result from training with the TS model and the ones that were trained with the TT model, we see that the TS model has a better performance. Furthermore, all the SLP models tend to output a mean estimated spectra with a higher SD than the reference spectra in the nonlinear domain (200–600 wavelength). Such fact can be related to the linearity of perceptron mathematical model, which is the basis of the SLP building blocks. This may cause the model to have higher difficulty in estimating nonlinear output. The next algorithm that was studied was the KNN. Figure 4, which is organized in the same manner as in Fig. 3, presents the mean estimated spectra that result from this study.

In a first analysis of the panels in Fig. 4, it seems that the KNN model has an overall good performance, but all the estimated spectra tend to be close to the mean with a low SD, which could be a sign of overfitting. However, by analyzing the individual estimations (Figs. S5–S8 in the supplementary material), it is possible to see that the model estimates differently for different samples, and, therefore, there is no overfitting. When the model is over fitted, it means that there is a memory leak from the data to the model. Consequently, the model will have an unwanted better performance in the training set, when compared with the data set. To prevent this from happening not only in the KNN models but also in the other ML algorithms, the learning process was stopped before the validation error increased and the parameters of the models were tuned for the best of their performance.[41] After obtaining the estimations with the KNN algorithm, new estimations were generated using the RFR algorithm. Figure 5 presents the results of those estimations.

Similarly to what was observed with the KNN algorithm, the estimations with the RFR algorithm present an overall good performance of the models except for the mean pathological spectra from the TT model. This could be because most of the individual estimated spectra have a lower value than expected, which results in a lower mean spectrum. Performing other estimation with the DTFMR algorithm, we observe from Fig. 6 a smaller performance when compared with the above algorithms.

The resulting mean estimations presented in the panels of Fig. 6 are more distant from the reference spectra and have a higher

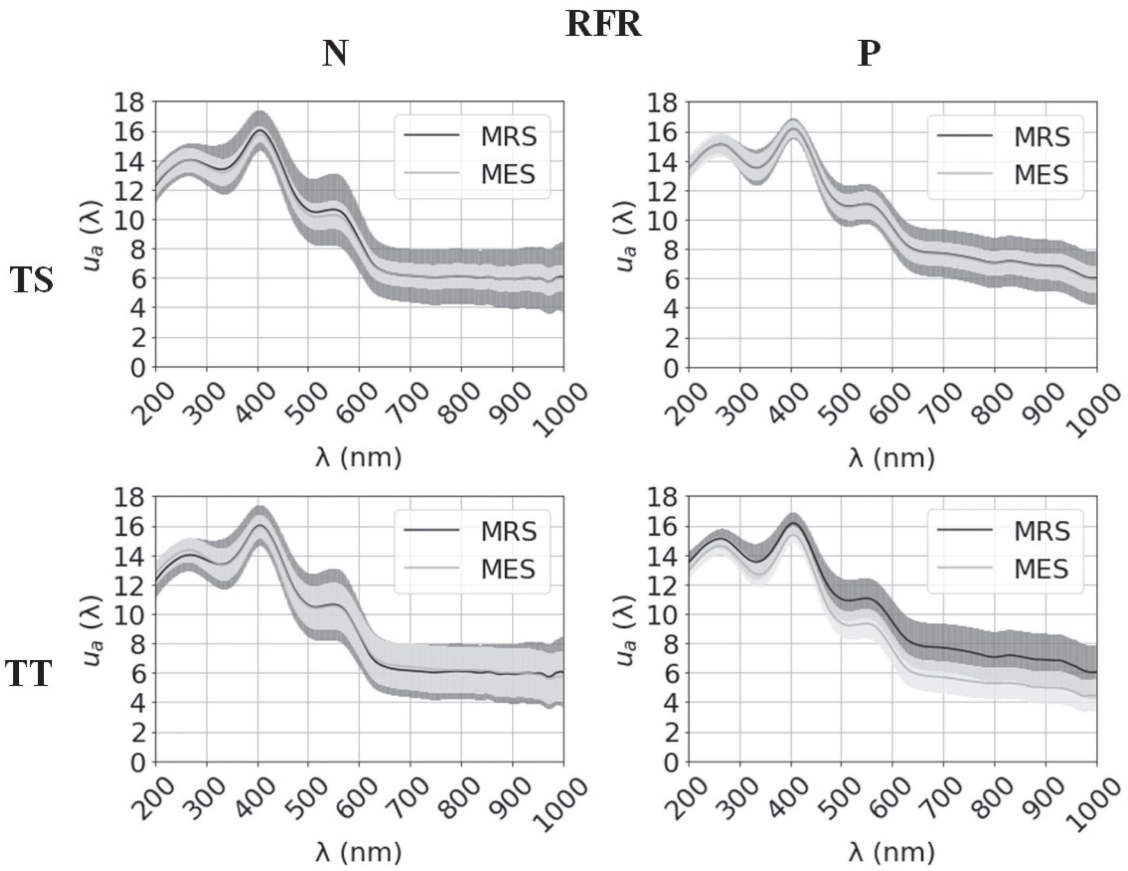

FIG. 5. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the RFR algorithm.

SD. The final estimations were made using the LRFMO algorithm. Figure 7 presents the mean results of these estimations.

As we can see from the panels in Fig. 7, the estimations with the LRFMO algorithm result in mean spectra that indicate the overall worst performance for this algorithm. In Fig. 7, we see that the SD presents a big increase with increasing wavelength. From the individual spectra estimations (see Figs. S13–S16 in the supplementary material), we can see that the models tended to estimate the μ_a spectra significantly above or below the reference spectra, which results in an increase of the standard deviation of the averaged μ_a spectra.

The shortcomings of our study rely on the reduced number of samples used in the development of the machine learning algorithms. If a significant number of samples was available, all the generated models would be more reliable, but unfortunately such samples were not available at the time of this study. Considering the models developed in the present study, with a reduced number of samples, we can say that our approach works as a feasibility test. Nevertheless, and remembering that we decided to use a LOO approach in order to minimize the overfitting, such approach has additionally allowed for pathology discrimination, with reasonable values for the pigment and blood content in both tissue conditions. For future studies for which a reduced number of samples is available, we can open a research line in the machine learning approach, designated as "generative models," in order to artificially generate more samples and to increase the number of samples for training.

After testing the ML algorithms, the average of the ED between the estimated and the reference spectra was calculated in order to compare the performance between the various algorithms. Figure 8 presents the results of this calculation.

The data in Fig. 8 show that the DTFMR and LRFMO algorithms estimate spectra that are significantly distant from the reference spectra. The SLP, KNN, and RFR algorithms, on the other hand, present higher accuracy in the approximation of the estimated to the reference spectra. Within these three algorithms, the results obtained with the TS approach are better than the ones obtained with the TT approach. This means that the R_d spectra contain information about the absorption coefficient and that such information is different for the healthy and pathological tissues.

Considering such quality factors, we selected the estimations obtained with the TS approach from the SLP, KNN, and RFR algorithms to perform the calculations reported in Ref. 24. These calculations consisted of subtracting μ_{a-lip} to the mean generated μ_a spectra of healthy and pathological mucosa to obtain adequate

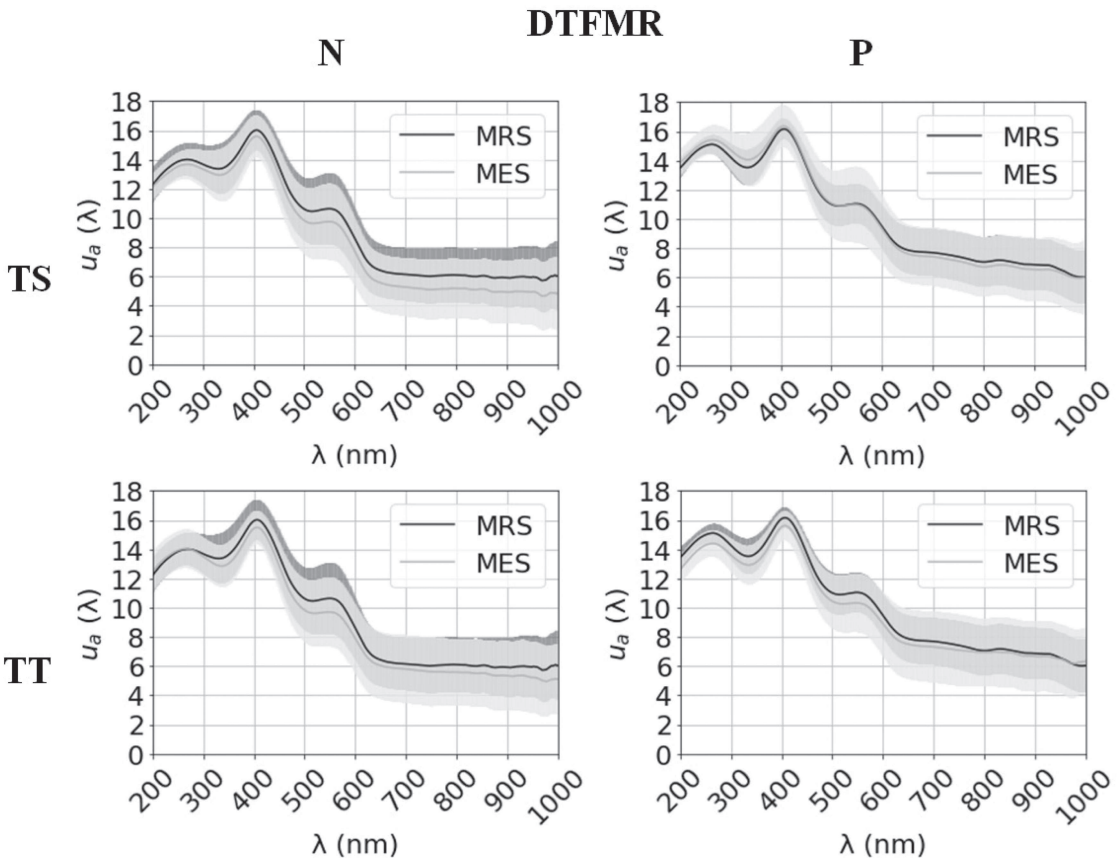

FIG. 6. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the DTFMR algorithm.

blood and pigment content. In the case of healthy mucosa, the $\mu_{a-lip}(\lambda)$ was considered as described by Eq. (7), with A = 1, and in the case of pathological mucosa, the same equation was used, but with A = 1.1 (10% more pigment in the pathological tissue).[24] Figure 9 contains the results of those final calculations for each particular case.

After performing a differentiated subtraction of $\mu_{a-lip}(\lambda)$ to the estimated $\mu_a(\lambda)$ of the tissues, the hemoglobin ratios were calculated at 410 and 550 nm for each case (see Fig. 9). In a first analysis of the data in Fig. 9, we looked to check if the results from all estimations produced higher magnitude ratios for the pathological tissue. For the estimation with the SLP and RFR algorithms, we obtained such results, but for the estimation with the KNN algorithm, the calculated hemoglobin ratios present a lower magnitude for the pathological tissue. Such fact could happen due to a minimum value of μ_a (close to 1000 nm) in the estimated spectrum for the pathological tissue to be lower than expected, which consequently increases the ratios.

Our second concern was to check how close the calculated ratios are to the ones presented in Ref. 24, as obtained from invasive measurements. In that study, the obtained hemoglobin ratios at 410 and 550 nm were: 19.7-fold and 10.1-fold for the healthy mucosa and 33.1-fold and 17.3-fold for the pathological mucosa, respectively.[24] Looking into the data generated with the SLP algorithm in Fig. 9(a), we see that the hemoglobin ratios for the pathological mucosa are excessively high. This means that by subtracting the $\mu_{a-lip}(\lambda)$, with A = 1.1, the minimum value in the red curve becomes too low, which leads to excessively high ratios at both wavelengths.

The ratios obtained from the RFR estimations are more approximated to the ones previously published,[24] although for the case of the pathological mucosa, they are a little higher. Such difference can possibly be related to the colorectal cancer samples that were used to acquire the R_d spectra—they could have a higher blood content than the ones used for the study in Ref. 24. Considering the ratios obtained with the RFR algorithm, we can consider that they are the optimal reconstruction of the ratios previously calculated from invasive measurements. Such results show that the combination of noninvasive R_d spectral measurements with the RFR learning algorithm is a good approach to develop new methods for colorectal cancer detection.

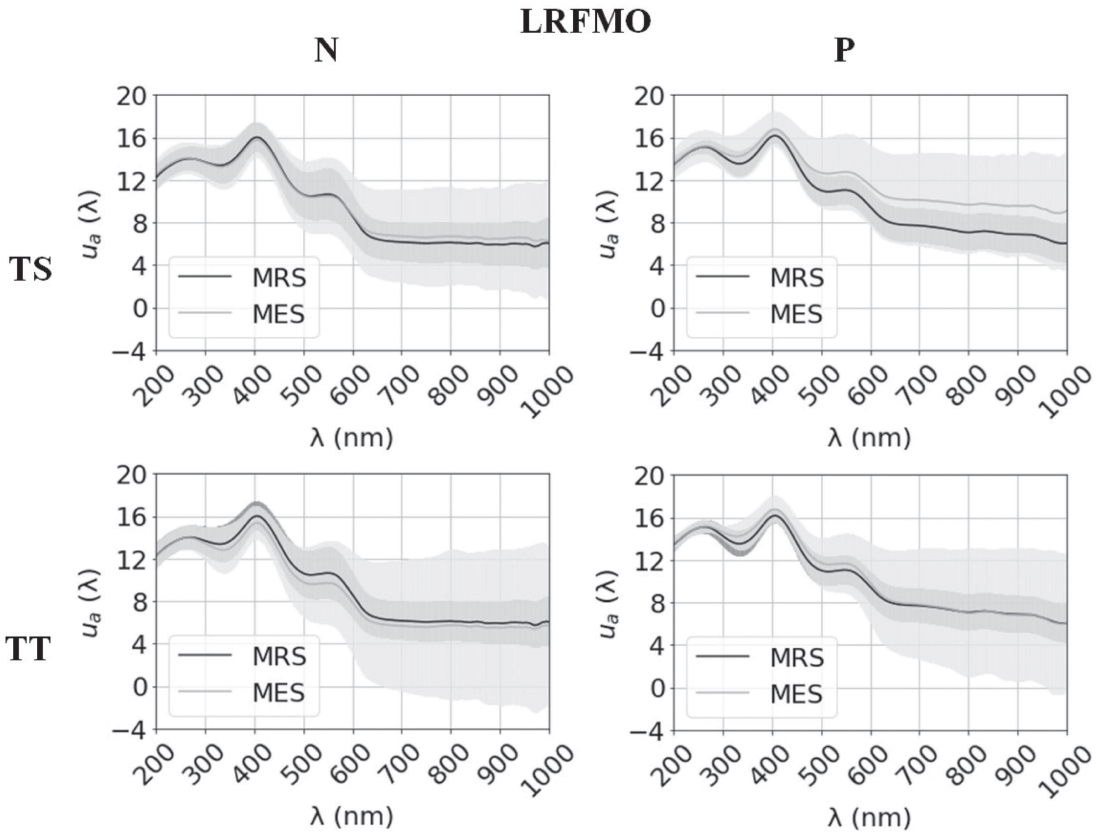

FIG. 7. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the LRFMO algorithm.

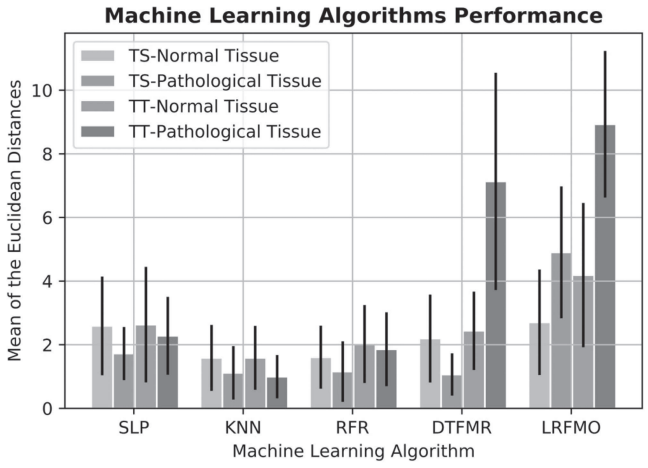

FIG. 8. Average of the Euclidean distances for the different models when they are trained with data from separated samples (TS) or with data from all samples (TT).

IV. CONCLUSION

The present study consisted of using ML methods to reconstruct the μ_a spectra of human colorectal mucosa tissues, both in healthy and in pathological (adenocarcinoma) versions. Such reconstruction was made using various ML algorithms, where the only inputs were noninvasive-like R_d measurements. To train the algorithms, we used μ_a spectra from both tissue conditions that were previously used in another study. The tissues used to perform all experimental measurements were freshly excised from patients via surgical procedure. Since the tissues were kept in saline before measurements to maintain their hydration, it is expected that the obtained optical properties mimic the ones for the *in vivo* situation. The differences should be minimal and within the experimental measurement error.

The efficiency of the various algorithms was analyzed, verifying that the DTFMR and LRFMO algorithms present the worst efficiency in the reconstruction of the μ_a spectra. The SLP, KNN, and RFR algorithms presented the best efficiency in that reconstruction and they were selected to perform additional calculations to obtain the blood and lipofuscin contents in both tissues and compare them with the results obtained from invasive measurements. In this final

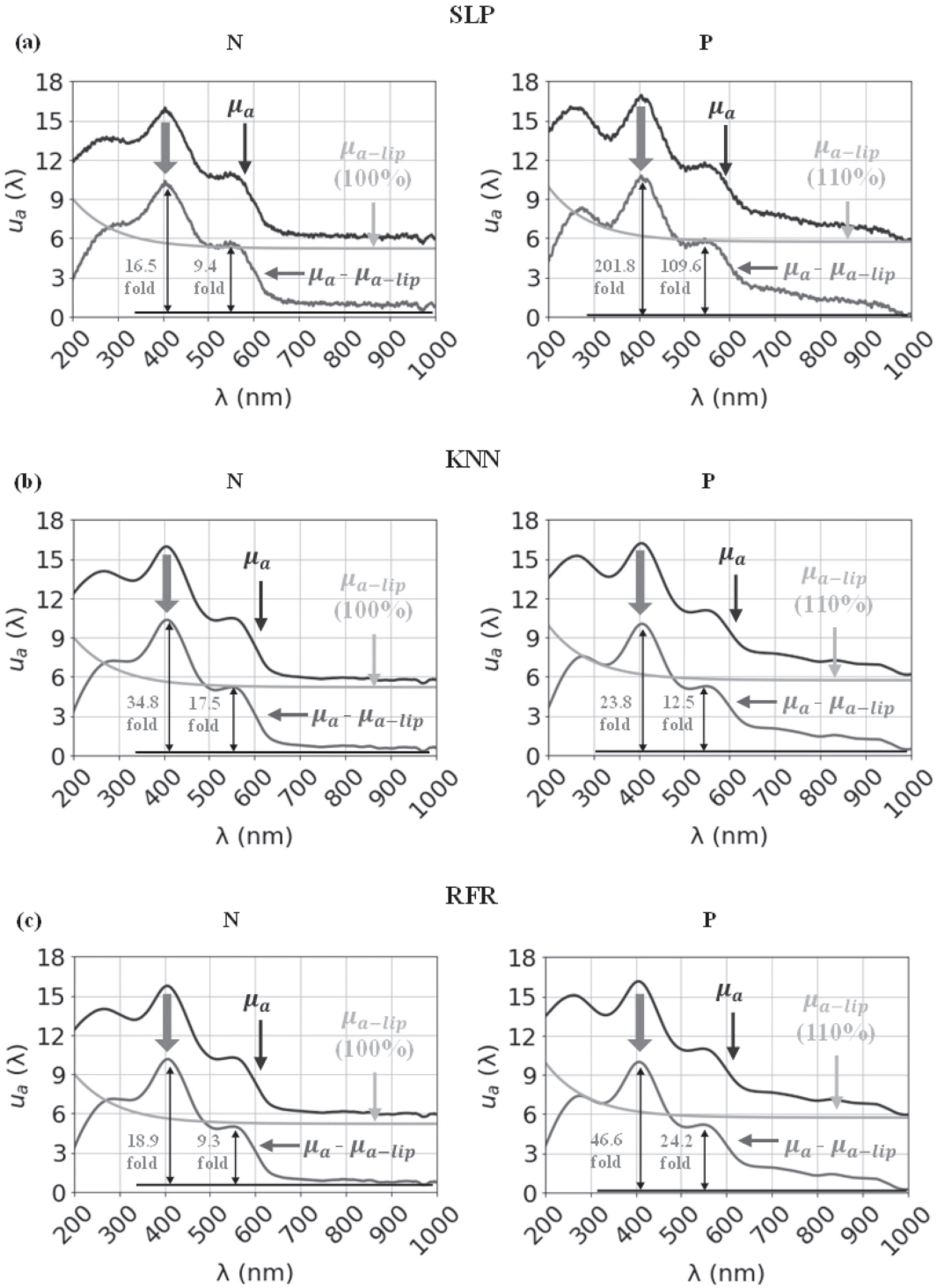

FIG. 9. Wavelength dependencies of μ_a for lipofuscin (orange), for healthy (N) and pathological (P) mucosa, before (blue) and after (green or red) subtracting the absorption of lipofuscin. Results obtained with the SLP (a), KNN (b), and RFR (c) algorithms.

analysis, the estimation made with the RFR algorithm presented the most approximated results to the ones previously obtained. In these calculations, the pathological tissue presented a little higher blood content than in the previous calculation, a difference that can be due to the different samples used in the T_t/R_t and R_d measurements. The results obtained using the TS approach in the RFR algorithm show that such a method presents good sensitivity and specificity for recognizing malignancy in the samples. This study proves the applicability of ML techniques in biophotonics, which, combined with noninvasive spectral measurements, can be used to detect pathologies. Considering the increasing content of blood and lipofuscin in colorectal mucosa tissues, the TS approach in the RFR algorithm can be used to monitor cancer progression. A future research perspective would consist of applying the ML techniques to estimate the spectral RI of tissues from noninvasive or minimally invasive measurements such as R_d spectra. Such research may produce results that can allow pathology differentiation through the spectral refractive index. Another research that can be developed is to use the knowledge gained in this study to perform the estimation of the optical properties for other tissues or other tissue conditions and pathologies. Using the transfer learning procedure, the algorithms developed in this study can be adapted for other tissues, provided that similar spectral measurements are made from those tissues. If different methods are available, the development of new algorithms should be made using machine learning procedures.

SUPPLEMENTARY MATERIAL

See the supplementary material for the individual estimated spectra with all models.

ACKNOWLEDGMENTS

The work of L. M. Oliveira was supported by the Portuguese Science Foundation (Grant No. FCT-UIDB/04730/2020). The work of V. V. Tuchin was supported by a grant of the Government of the Russian Federation (Registration No. 2020-220-08-2389).

DATA AVAILABILITY

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

REFERENCES

[1] S. L. Jacques, "Optical properties of biological tissues: A review," Phys. Med. Biol. **58**, R37–R61 (2013).
[2] V. V. Tuchin, *Tissue Optics Light Scattering Methods and Instruments for Medical Diagnosis*, 3rd ed. (SPIE Press, Bellingham, WA, 2015), pp. 245–358.
[3] L. M. Oliveira and V. V. Tuchin, *The Optical Clearing Method—A New Tool for Clinical Practice and Biomedical Engineering* (Springer, Cham, 2019), pp. 1–106.
[4] I. Carneiro, S. Carvalho, R. Henrique, L. Oliveira, and V. V. Tuchin, "Measurement of optical properties of normal and pathological human liver tissue from deep-UV to NIR," Tissue Opt. Photonics **11363**, 113630G (2020).
[5] D. C. Sordillo, L. A. Sordillo, P. P. Sordillo, L. Shi, and R. R. Alfano, "Short wavelength infrared optical windows for evaluation of benign and malignant tissues," J. Biomed. Opt. **22**(4), 45002 (2017).
[6] L. Shi, L. A. Sordillo, A. Rodríguez-Contreras, and R. Alfano, "Transmission in near-infrared optical windows for deep brain imaging," J. Biophot. **9**, 38–43 (2016).
[7] A. N. Bashkatov, E. A. Genina, M. D. Kozintseva, V. I. Kochubei, S. Y. Gorodkov, and V. V. Tuchin, "Optical properties of peritoneal biological tissues in the spectral range of 350-2500," Opt. Spectrosc. **120**, 6–14 (2016).
[8] I. Carneiro, S. Carvalho, R. Henrique, L. Oliveira, and V. V. Tuchin, "Moving tissue spectral window to the deep-ultraviolet via optical clearing," J. Biophot. **12**(12), e201900181 (2019).
[9] I. Carneiro, S. Carvalho, R. Henrique, A. Selifonov, L. Oliveira, and V. V. Tuchin, "Enhanced ultraviolet spectroscopy by optical clearing for biomedical applications," IEEE J. Sel. Top. Quant. Elect. **27**(4), 7200108 (2021).
[10] L. M. Oliveira, K. I. Zaytsev, and V. V. Tuchin, "Improved biomedical imaging over a wide spectral range from UV to THz towards multimodality," In *SPIE PROC of the Third International Conference of Biophotonics Riga 2020*, edited by J. Spigulis (SPIE, Bellingham, WA, 2020), Vol. 11585, p. 11585.
[11] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, V. S. Rubtsov, E. A. Kolesnikova, and V. V. Tuchin, "Optical properties of human colon tissues in the 350–2500 spectral range," Quant. Elect. **44**, 779–784 (2014).
[12] T. Vo-Dinh, *Biomedical Photonics Handbook*, 2nd ed. (CRC Press, Boca Raton, FL, 2014), Vol. 1, pp. 23–168.
[13] L. H. Wang, S. L. Jacques, and L. Q. Zheng, "MCML-Monte Carlo modeling of photon transport in multi-layered tissues," Comp. Methods Progr. Biomed. **47**, 131–146 (1995).
[14] S. A. Prahl, M. J. C. van Gemert, and A. J. Welch, "Determining the optical properties of turbid media by using the adding-doubling method," Appl. Opt. **32**, 559–568 (1993).
[15] S. A. Prahl, see https://omlc.org/software/index.html for "Optics Software;" accessed 27 January 2021.
[16] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, and V. V. Tuchin, "Optical properties of human skin, subcutaneous and mucous tissues in the wavelength range from 400 to 2000 nm," J. Phys. D: Appl. Phys. **38**, 2543 (2005).
[17] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, and V. V. Tuchin, "Optical properties of the subcutaneous adipose tissue in the spectral range 400-2500 nm," Opt. Spectrosc. **99**, 836–842 (2005).
[18] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, and V. V. Tuchin, "Optical properties of human cranial bone in the spectral range from 800 to 2000 nm," in *SPIE Proceedings of the Saratov Fall Meeting 2005: Optical Technologies in Biophysics and Medicine VII, Saratov, Russia, 27–30 September 2005*, edited by V. V. Tuchin (SPIE, Bellingham, WA, 2006), Vol. 6163, p. 6163.
[19] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, A. A. Gavrilova, S. V. Kapralov, V. A. Grishaev, and V. V. Tuchin, "Optical properties of human stomach mucosa in the spectral range from 400 to 2000nm: Prognosis for gastroenterology," Med. Laser Appl. **22**, 95–104 (2007).
[20] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, and V. V. Tuchin, "Optical properties of human sclera in spectral range 370–2500 nm," Opt. Spectrosc. **109**, 197–204 (2010).
[21] S. Carvalho, N. Gueiral, E. Nogueira, R. Henrique, L. Oliveira, and V. V. Tuchin, "Comparative optical properties of colon mucosa and colon precancerous polyps between 400 and 1000 nm," in *SPIE Proceedings of BIOS-Photonics West 2017: Dynamics and Fluctuations in Biomedical Photonics, San Francisco, CA, USA, 28 January–2 February 2017*, edited by V. V. Tuchin, K. V. Larin, M. J. Leahy, and R. K. Wang (SPIE, Bellingham, WA, 2017), Vol. 10063, p. 10063.
[22] I. Carneiro, S. Carvalho, R. Henrique, L. Oliveira, and V. V. Tuchin, "Optical properties of colorectal muscle in visible/NIR range," in *SPIE Proceedings of Photonics Europe Biophotonics—Photonic Solutions for Better Health Care VI, Strasbourg, France, 22–26 April 2018*, edited by J. Popp, V. V. Tuchin, and F. S. Pavone (SPIE, Belingham, WA, 2018), Vol. 10685, p. 10685.
[23] I. Carneiro, S. Carvalho, R. Henrique, L. Oliveira, and V. V. Tuchin, "Measuring optical properties of human liver between 400 and 1000 nm," Quant. Electr. **49**, 13–19 (2019).
[24] S. Carvalho, I. Carneiro, R. Henrique, V. V. Tuchin, and L. Oliveira, "Lipofuscin-type pigment as a marker of colorectal cancer," Electronics **9**, 1805 (2020).

[25]E. A. Genina, A. N. Bashkatov, and V. V. Tuchin, "Optical clearing of human dura mater by glucose solutions," J. Biomed. Photonics Eng. **3**(1), 010309 (2017).
[26]J. W. Pixkering, C. J. M. Moes, H. J. C. M. Sterenborg, S. A. Prahl, and M. J. C. van Gemert, "Two integrating spheres with an intervening scattering sample," J. Opt. Soc. Am. A **9**, 621–631 (1992).
[27]I. Carneiro, S. Carvalho, V. Silva, R. Henrique, L. Oliveira, and V. V. Tuchin, "Kinetics of optical properties of human colorectal tissues during optical clearing: A comparative study between normal and pathological tissues," J. Biomed. Opt. **23**, 121620 (2018).
[28]I. Carneiro, S. Carvalho, R. Henrique, L. Oliveira, and V. V. Tuchin, "Kinetics of optical properties of colorectal muscle during optical clearing," IEEE J. Sel. Top. Quant. Elect. **25**(1), 7200608 (2019).
[29]M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science **349**(6245), 255–260 (2015).
[30]P. Pradhan, S. Guo, O. Ryabchykov, J. Popp, and T. W. Bocklitz, "Deep learning a boom for biophotonics?," J. Biophot. **13**(6), e201960186 (2020).
[31]W. S. McCullock and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull. Mathmat. Biophys. **5**, 115–133 (1943).
[32]A. L. Samuel, "Some studies in machine learning using the game of checkers," IBM J. Res. Develop. **3**(3), 210–229 (1959).
[33]L. Guyon, A. da Silva, A. Planat-Chrétien, P. Riso, and J.-M. Dinten, "X^2 analysis for estimating the accuracy of optical properties derived from time resolved diffuse-reflectance," Opt. Express **17**(22), 20521 (2009).
[34]H.-P. Hsieh, F.-H. Ko, and K.-B. Sung, "Hybrid method to estimate two-layered superficial tissue optical properties from simulated data of diffuse reflectance spectroscopy," Appl. Opt. **57**(12), 3038 (2018).
[35]T. J. Farrel, B. C. Wilson, and M. S. Patterson, "The use of a neural network to determine tissue optical properties from spatially resolved diffuse reflectance measurements," Phys. Med. Biol. **37**(12), 2281–2286 (1992).
[36]S. Panigrahi and S. Gioux, "Machine learning approach for rapid and accurate estimation of optical properties using spatial frequency domain imaging," J. Biomed. Opt. **24**(7), 071606 (2018).
[37]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[38]I. D. Nagtegaal, M. J. Arends, and M. Salto-Tellez, "Colorectal adenocarcinoma," in *WHO Classification of Tumours—Digestive System Tumours*, 5th ed. (The WHO Classification of Tumours Editorial Board, 2019), pp. 177–187.
[39]T. Zhang, "Leave-one-out bounds for kernel methods," Neural Comput. **15**(6), 1397–1437 (2003).
[40]J. D. Johansson and K. Wårdell, "Intracerebral quantitative chromophore estimation from reflectance spectra captured during deep brain stimulation implantation," J. Biophot. **6**(5), 435–445 (2013).
[41]X. Ying, "An overview of overfitting and its solutions," J. Phys. **1168**(2), 022022 (2019).