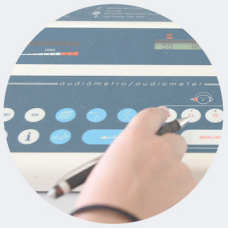
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA BIOMÉDICA





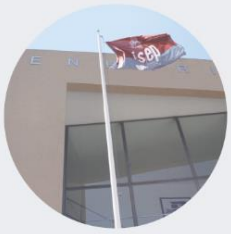

Estimação das propriedades óticas de tecidos a partir de dados de reflectância difusa usando modelos de aprendizagem automática

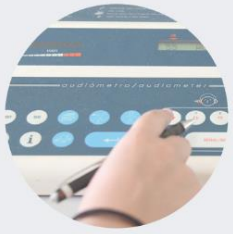
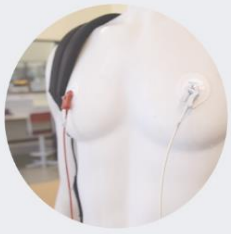
LUÍS EMANUEL PEREIRA PINTO FERNANDES
julho de 2021

P.PORTO

INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

MESTRADO EM ENGENHARIA BIOMÉDICA



P.PORTO

Estimation of tissue's optical properties from reflectance data using machine learning models

Luís Emanuel Pereira Pinto Fernandes

"Dissertação apresentada no Instituto Superior de Engenharia do Porto para a obtenção de grau de Mestre em Engenharia Biomédica"

Orientador: Professor Luís Oliveira PhD.

Coorientador: Professor Hélder Oliveira PhD.

Julho de 2021

“Do not go gentle into that good night.

Rage, rage against the dying of the light.”

Dylan Thomas

Agradecimentos

Antes de mais, agradeço a paciência generosa dos meus orientadores, o Professor Luís Oliveira e o Professor Hélder Oliveira. Os meus agradecimentos vão igualmente para o CIETI e para o Professor Luís Oliveira, pelo equipamento disponibilizado para realizar as medições necessárias a este trabalho. Agradeço também ao Departamento de Anatomia Patológica do IPO-Porto por fornecer as amostras de tecidos.

Por fim, mas não menos importante, agradeço à minha família e aos meus amigos pelo suporte emocional que me deram.

Resumo

As doenças oncológicas consistem numa das maiores preocupações dos sistemas de saúde a nível mundial, visto que o cancro é anualmente responsável por milhões de mortes. Além disso, se o cancro for detetado num estágio tardio poderá tornar o seu tratamento mais difícil, ou mesmo ineficaz. Tendo em conta estes fatores, novas tecnologias de diagnóstico de cancro são necessárias para melhorar o tratamento médico fornecido aos pacientes.

A ótica médica é uma disciplina que tem produzido muitos progressos nos últimos 40 anos e que pretende a longo prazo substituir as tecnologias de diagnóstico e de tratamento que utilizam radiação ionizante, permitindo procedimentos não invasivos ou minimamente invasivos. Para qualquer procedimento clínico que utilize a luz é necessário conhecer as propriedades óticas dos tecidos e a sua dependência no comprimento de onda. Tais propriedades são características dos tecidos e apresentam diferenças nos casos de patologia oncológica, devido às mutações genéticas que o cancro produz. Estas alterações nos tecidos com cancro podem ser detetadas com medições sensíveis não invasivas como a reflectância difusa. Devido à falta de formalismos matemáticos que relacionem a reflectância difusa com as propriedades óticas dos tecidos, existem na atualidade ainda muito poucas formas de se proceder a um diagnóstico com estas medições. As Simulações de Monte Carlo são normalmente utilizadas para estimar as propriedades óticas desejadas numa determinada aplicação, no entanto, estas tendem a ser computacionalmente dispendiosas e morosas. Atualmente, os modelos de aprendizagem automática estão a assumir um papel muito importante na ótica médica, em especial na estimação das propriedades óticas dos tecidos biológicos. No entanto, os dados para treino destes modelos não provêm de medições experimentais de tecidos, mas sim de simulações de Monte Carlo ou de medições de estruturas criadas em laboratório como modelos de tecidos. Assim, o objetivo do presente trabalho consistiu em investigar mais aprofundadamente o uso de modelos de aprendizagem automática para estimar as propriedades óticas de tecidos biológicos, tais como: o coeficiente de absorção, o coeficiente de espalhamento e o índice de refração. Para além da estimação das propriedades óticas dos tecidos, estudamos também a classificação em normal ou patológico dos espectros de reflectância difusa, como forma de detetar o cancro colorretal.

Os modelos de aprendizagem automática que foram usados para estimar as propriedades óticas de tecidos da mucosa colorectal humana são os seguintes: *Single Layer Perceptron, Random Forest Regressor, K Nearest Neighbor, Decision Tree for Multioutput Regression and Linear Regression for Multioutput*. Estes modelos foram treinados utilizando o método *Leave One Out* e para avaliar a sua performance foi calculada a distância euclidiana entre os espectros estimado e de referência que provem das medições experimentais em tecidos. Nestes estudos, o modelo que teve a melhor performance foi o *RFR*, visto ter sido capaz de estimar com a melhor precisão os espectros do coeficiente de absorção e de produzir a segunda melhor precisão na estimação do coeficiente de espalhamento.

Para classificar os espectros de reflexão difusa em espectros normais ou patológicos, tentamos numa primeira experiência usar o declive dos espectros para discriminar entre os dois tipos de espectros, mas sem sucesso. Numa segunda experiência utilizámos uma *Selector Vector Machine* para classificar os espectros. Depois de selecionar os comprimentos de onda que iriam ser utilizados como *features*, o modelo foi treinado utilizando o método *Leave One Out*, obtendo uma precisão de 90%.

Em suma, neste estudo nós propomos uma nova *framework* para estimar as propriedades óticas do tecido colorretal a partir do respetivo espectro de reflectância difusa e usando modelos de aprendizagem automática. Foi também abordada a classificação automática dos espectros de reflectância difusa para a deteção do cancro colorretal. Os modelos usados foram capazes de estimar os espectros desejados de uma forma precisa e com uma distância euclidiana abaixo de dois. Os espectros de absorção foram estimados com uma precisão suficiente para serem usados para calcular o índice de refração dos tecidos. A classificação dos espectros foi atingida com sucesso e com uma precisão de 90%.

Palavras-chave:

Propriedades óticas espetrais de tecidos biológicos, índice de refração, coeficiente de espalhamento, coeficiente de absorção, aprendizagem máquina, diagnóstico não invasivo, cancro colorretal.

Abstract

Oncologic diseases are one of the main concerns of the health systems around the world due to the fact that cancer is responsible for millions of deaths per year. In general, cancer diseases are detected at a late stage of development, turning its treatment more difficult, or even impossible. Taking these factors into account, new cancer diagnostic technologies are necessary to improve the clinical therapies for the patient.

Biophotonics is a field that has produced many improvements in the past 40 years and whose long-term objective is to replace diagnostic and treatment procedures that use ionizing radiation, allowing the application of noninvasive or minimally invasive procedures. For any clinical procedure that uses light, it is necessary to know the optical properties of tissues and their wavelength dependence. Such properties are characteristic to the biological tissues and present differences in cases of oncological pathologies due to the genetic mutations that cancer produces. Such cancer-induced mutations can be detected with noninvasive sensitive measurements, such as diffuse reflectance. Due to the lack of mathematical formalisms that relate the diffuse reflectance with the optical properties of tissues, there are still very few ways to obtain a reliable diagnosis with such measurements. Monte Carlo simulations are normally used to estimate the optical properties in particular applications, but they are time consuming and computational expensive. Nowadays, the machine learning models are assuming a very important role in biophotonics, especially in the estimation of tissue's optical properties. Nevertheless, the experimental data used to train such models are not obtained from biological tissues, but from tissue phantoms or from Monte Carlo simulations. This way, the objective of the present work was to deep investigate the use of machine learning models to estimate the optical properties of biological tissues, such as the absorption coefficient, the scattering coefficient and the refractive index. Additionally, the automated classification of the experimental diffuse reflectance spectra into normal or diseased, as a way to detect colorectal cancer, was also studied.

The machine learning models that were used to estimate the optical properties of human colorectal mucosa tissues were the following: Single Layer Perceptron, Random Forest Regressor, K Nearest Neighbor, Decision Tree for Multioutput Regression and Linear Regression for Multioutput. These models were trained using the Leave One Out method and the euclidean distance between the reference (experimental) and estimated spectra was used to evaluate their performance. In these studies, the Random Forest Regressor model was the one with better performance, since it was able to estimate the spectra of the absorption coefficient with the highest precision and presented the second-best performance in the estimation of the spectra for the scattering coefficient.

To classify the diffuse reflectance spectra into normal or pathological, a first trial consisted on evaluating the slope of the spectra, but this turned into a failure. A second attempt was made by using a Selector Vector Machine to classify the spectra. After selecting the wavelengths that would be considered as features, the model was trained using the Leave One Out method. Such procedure resulted in a precision of 90%.

As a conclusion, in this work we propose a new framework to estimate the optical properties of colorectal tissues from the diffuse reflectance spectrum through machine learning models. The automatic classification of the diffuse reflectance spectra for cancer detection was also performed. The models used in this task were capable to estimating the desired spectra with an euclidean distance below 2. The absorption spectra were estimated with a sufficient precision to be used for the calculation of the refractive index of the tissues. The spectral classification was successfully reached with a precision of 90%.

Key words:

Spectral optical properties of biotissues, refractive index, scattering coefficient, absorption coefficient, machine learning, noninvasive diagnostics, colorectal cancer.

Index

Figure List

List of Tabels

Abbreviation List

AI	Artificial Intelligence
CIETI	Center for Innovation in Engineering and Industrial Technology
T_c	Collimated Transmittance
DTFMR	Decision Tree for Multioutput Regression
R_d	Diffuse Reflectance
ED	Euclidian Distances
IAD	Inverse Adding-Doubling
KNN	K Nearest Neighbors
K-K	Kramers-Kroning
LOO	Leave One Out
LRFMO	Linear Regression for Multioutput
LUT	Look Up Table
ML	Machine Learning
MES	Mean Estimated Spectrum
MRS	Mean Reference Spectrum
MC	Monte Carlo
NIR	Near Infrared
NN1	Neural Network 1
NN2	Neural Network 2
N	Normal
NM	Normal Mucosa
P	Pathological
PM	Pathological Mucosa
PMT	Photomultiplier Tube

RFR	Random Forest Regression
RI	Refractive Index
SVM	Selector Vector Machine
SLP	Single Layer Perceptron
SD	Standard Deviation
TCSPC	Time Correlated Single Photon Counting
TRS	Time-resolved spectroscopy
R_t	Total Reflectance
T_t	Total Transmittance
TS	Trained Separately
TT	Trained Together
UV	Ultraviloet
WHO	World Health Organization

CHAPTER 1 – INTRODUCTION

1. Introduction

According to the world Health Organization (WHO), cancer related diseases were responsible for 9.6 million deaths in 2018 and the burden of cancer is still growing globally affecting individuals financially and emotionally [1] . One of the most common types of cancer is colorectal cancer, being solely responsible for 10% of the new cancer cases in 2020, as represented in Figure 1 [2].

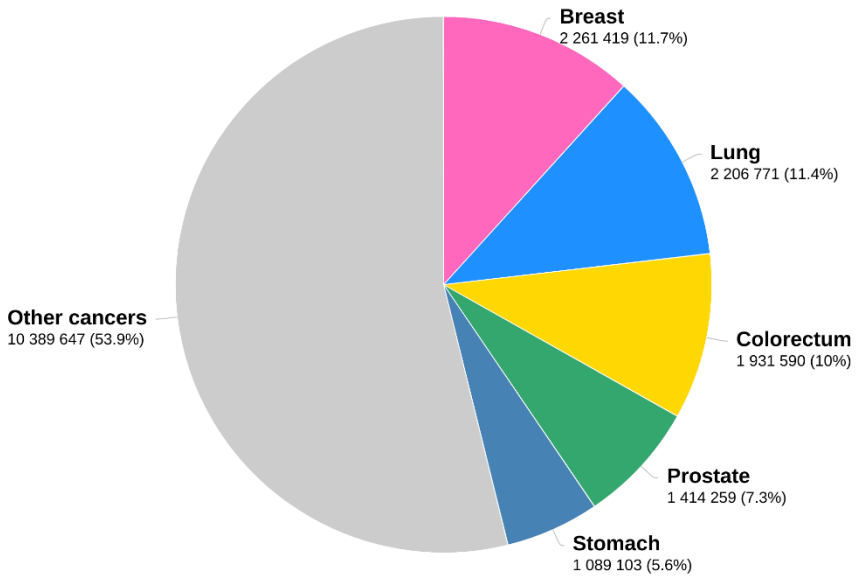


Figure 1.Estimated number of new cancer cases in 2020, worldwide, both sexes, all ages [1].

The systematic screening for cancer detection in the population could increase the chances to detect the malignant tumors in their early stage while they are in a treatable phase. Such procedure would increase the probability of a successful treatment for the patient and extend life expectancy. The study of the optical properties of biological tissues presents an opportunity to develop new diagnostic tools that could help in the screening of cancer in patients. Due to physiological changes that occur in tissues as cancer develops, the spectral optical properties of normal and pathological tissues will differ, and the identification of such differences can be used for diagnostic purposes [3][4].

The interaction of light with mater inside biological tissues occurs in two ways: absorption and scattering of photons. Biological tissues present two optical properties that

quantify the number of photons that are absorbed (or scattered) per unit length: the absorption coefficient, μ_a, and the scattering coefficient, μ_s, both measured in cm^{-1} [5]. Furthermore, when light scatters, it can take any direction. The scattering function $p(\vec{S}, \vec{S}')$ describes the probability of a photon that travels in the $\vec{S}$ direction to be deviated into the $\vec{S}'$ direction. If the scattering is symmetric, which is the most common case for biological tissues, using $\vec{S}$ direction as reference, the scattering function will only depend on θ, the angle between the $\vec{S}$ and $\vec{S}'$ directions. An additional optical property, which characterizes such scattering directionality, is designated as the anisotropy-factor (g). Such property represents the mean of the cosine of θ ($g = <\cos\theta>$) [3][5] . The reduced scattering coefficient μ'_s is another widely studied optical property and it is mathematically related with g and μ_s [3][6]. μ'_s contains both scattering and scattering directionality properties and is defined as: $\mu'_s = \mu_s (1-g)$ [3].

The evaluation of tissue's optical properties in a wide spectral range is necessary for the optimization of current and development of new optical methods in clinical practice [7]. Current biophotonics techniques range from the deep-ultraviolet, where induced transparency windows have been recently discovered [8], to the THz frequencies [7], where new methods are emerging. Furthermore, the calculation (or estimation) of those properties can be used for direct diagnostic purposes, considering that they are different between normal and pathological tissues [9].

Traditional methods to evaluate the optical properties of tissues rely on inverse simulations that are based on the Monte Carlo (MC) [3], or the Adding-Doubling algorithms [10]. These simulation methods use optical measurements, such as the total transmittance, T_t, the total reflectance, R_t, and the collimated transmittance, T_c, to estimate the optical properties that would result in those experimental measurements when *ex vivo* tissue samples are stimulated by light at certain wavelengths [6]. Apart from the inconvenient that these measurements must be made from excised *ex vivo* tissues, the above referred simulations only estimate a set of optical properties for a single wavelength at a time. To obtain complete spectra for the optical properties of a tissue within a wavelength range would be a time consuming task [9]. An alternative method was recently proposed to evaluate almost all optical properties of a tissue for a wide spectral range at once [7][10]. In this method, using only simple calculations, which use as input T_t, R_t and T_c spectra that were collected from tissue samples, it is possible to calculate the wavelength dependence for the optical properties of the tissue in a wide spectral range at

once. The method requires only the need of a few inverse adding-doubling (IAD) simulations to estimate μ'_s at discrete wavelengths within the desired spectral range[11][12].

While the above referred method is fast, it still needs measurements from excised tissues. A possible approach to develop new noninvasive methods for estimation of the optical properties of biological tissues consists on combining reflectance measurements with automated learning techniques. Although there are mathematical equations that describe the relations between the optical properties of tissues, to our knowledge, there is no mathematical equation that can be used to estimate those properties from reflectance data. All the optical properties of tissues, with the exception of μ_a, present a smooth decreasing behavior with increasing wavelength from the ultraviolet to the near infrared (UV-NIR), a behavior that is well described mathematically [7]. In the case of μ_a, no equation has been defined due to the presence of several absorption bands that correspond to the biological components in the tissues [6]. Biological tissues contain proteins, blood, cells, DNA, lipids and other absorbers that present absorption bands located in the UV-NIR range. Those bands will appear in the μ_a spectrum, turning impossible the description of such spectrum by a mathematical equation.

The T_t and R_t spectral measurements are not very sensitive to the detection of all absorption bands that correspond to tissue components, meaning that more sensitive spectral measurements are required. For the *ex vivo* situation, the measurement of T_c spectra can be used to calculate the μ_a spectrum of a tissue [7], but once again, tissue excision needs to be done. The measurement of diffuse reflectance (R_d) spectra is also sensitive to the absorption bands of tissue components [6], but no mathematical relation exists between this measurement and μ_a or with the other optical properties. R_d measurements can be made both from *ex vivo* or from *in vivo* tissues, meaning that they can be acquired with noninvasive procedures, which allow a higher comfort for the patient and clinician and avoid surgical procedures. In this case, and if the spectral optical properties of a tissue are known from other measurements, machine learning (ML) algorithms can be developed to recalculate them from the R_d spectra.

Although some research has already been made using ML models to estimate optical properties from pre-acquired data, there is still a need to extensively investigate and compare various ML models to access which one is the best at estimating the desired spectral shape and also to establish a reference of performance for future ML works.

Therefore, this research has the following objectives: the study of various ML algorithms to recalculate the optical properties of biological tissues based only on noninvasive R_d spectra, the automated labelling of the R_d spectra in normal or pathological for colorectal cancer detection and the improvement of the healthcare service given to the patients with colorectal cancer. An additional objective of this work is to improve the spectral estimation process, by replacing the traditional use of inverse simulations by ML models. By using ML algorithms as a replacement of the traditional simulation procedure, we intend to accelerate the optical properties estimation process. The present work is organized in five chapters: the current chapter – the introduction, which is followed by the state of the art, the materials and methods, the results and discussion and the conclusion. The conventional strategies that are used to estimate the optical properties of tissues are described in the following chapter, while the different experimental approaches used in the present work are described in the chapter of materials and methods.

CHAPTER 2 – STATE OF THE ART

2. State of the Art

Several methods to estimate optical properties of tissues have been developed though the last 40 years and their description can be found in some reference works such as Ref. [5]. We can categorize these methods in three main categories: mathematical modelling of the light interactions with tissues, inverse simulation methods and the use of ML algorithms. A brief description of these methods is made in the following sub-sections.

2.1. Mathematical models

Considering an example of mathematical models to estimate the optical properties, the authors of Ref. [13] developed a relation between R_d and the absorption and reduced scattering coefficients - μ_a, μ'_s of the tissues, as presented in Eq. (1):

$$R_d = \frac{1}{k_1 \times \frac{1}{\mu'_s} + k_2 \times \frac{\mu_a}{\mu'_s}}, \quad (1)$$

where k_1 and k_2 are parameters related to the geometry of the probe and the refractive indices of the tissues. Considering that both μ_a and μ'_s are represented in cm^{-1}, k_1 is represented in cm^{-1} and k_2 is a dimensionless parameter. After calculating k_1 and k_2 from R_d data that was measured from tissue phantoms, the authors of Ref. [13] used Eq. (1) to fit the R_d spectra acquired from human tissues, as presented in Fig. 2.

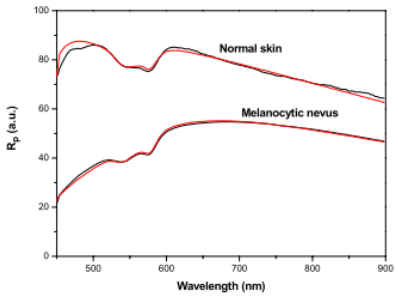


Figure 2. Diffuse reflectance spectrum of human tissues *in vivo*: experimental data (black) and model fitting (red) [13].

With such model fitting the authors obtained a precision generally better than 6%. By analysing Fig. 2, it was expected a better precision. However, this is an average

precision of the two combinations of μ_a and μ'_s for a given value of R_d. Although the precision was not relatively high, the authors were able to avoid using Monte Carlo Simulations.

Guyon et al. [14] proposed a new way to evaluate the model goodness by fitting it to empirical data to evaluate the χ^2 criterium. To retrieve the necessary experimental data, the apparatus in Figure 4 was used.

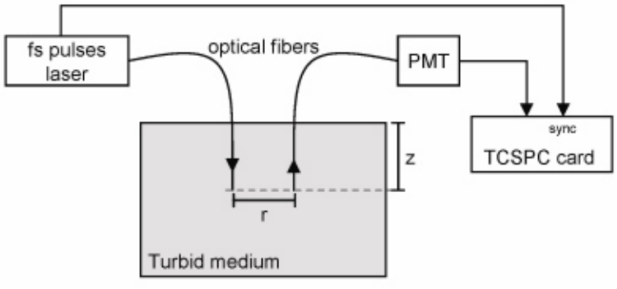


Figure 3. Experimental setup used in the study of Ref. [14].

According to Fig. 3, the optical fiber cables are separated by a distance *r* and both are at depth *z*. After the photons are emitted from the laser, they travel throw the medium, to be captured by an optical fiber cable connected to a fast photomultiplier tube (PMT). Then, a time correlated single photon counting card (TCSPC) does the electronic acquisition chain of the photons.

The scattering properties of a turbid medium (Acronal stock solution) were measured at a fixed value of μ_a of 5.35 cm^{-1} using time-resolved spectroscopy (TRS). In addition, Monte Carlo simulations were executed to investigate the domains where there is strong absorption. Such simulations were adopted because the inverse algorithm that is used to estimate the optical properties uses the diffusion equation as a mathematical model and therefore, the estimations will have an increase in their error where there is a strong absorption and a weak scattering. Following the diffusion assumptions, the authors formulated the following mathematical model that is proportional to the optical properties of a medium:

$$m \propto \frac{c}{4\pi D c^{\frac{3}{2}} t^{\frac{3}{2}}} \times e^{\left(-\frac{r^2}{4Dct} - \mu_a ct\right)}, \quad (2)$$

where c is the speed of light, $D = \frac{1}{3\mu's}$ is the diffusion coefficient, r is the distance between an isotropic source and the detector fiber.

Furthermore, to evaluate the fitting goodness, the reduced error function χ_R^2 was used in the following form [10]:

$$\chi_R^2 = \frac{1}{T-p}\sum(W_{Ri}) = \frac{1}{T-p}\sum\left(\frac{m_i - s_i}{\sqrt{s_i}}\right), \quad (3)$$

where T represents the number of temporal channels i used during the fitting process, p is the number of estimated parameters, m_i and s_i are the model and the measured temporal profile per channel t_i and W_{Ri} are the weighted residuals.

After making the necessary measurements, the authors concluded that when there was an increase in the acquired photons, the χ_R^2 also increased, indicating a bad fit and that the model was bias. Furthermore, a low distance between the fibers could result in an increase of the measurement error of μ_a and μ'_s.

2.2. Inverse simulation methods

Considering the estimation of tissue's optical properties with inverse simulations, many studies have been performed, both for normal and for pathological tissues. Some studies have been reported for human stomach mucosa[14], human cranial bone [15], skin, subcutaneous and muscle tissues [16] or normal and pathological colorectal tissues [17][18].

Apart from the basic inverse simulation procedure, some hybrid models that use Monte Carlo simulations have been presented in recent years. The main objective of such methods is to avoid using invasive measurements such as T_t and T_c. *Hsieh et al.* [19], developed an hybrid model to extract the optical properties of two layered tissues from R_d spectroscopy. The first step of this method was to establish a tissue model to simulate the reflectance spectra using Monte Carlo simulations. The first layer of the tissue model represents the epithelium and the second layer represents the stroma. Both layers had homogenous absorption and scattering coefficients. The absorption coefficient of the first layer of the model is described by Eq. (4):

$$\mu_{a,epi} = C_{epi} \times E(\lambda), \quad (4)$$

where C_{epi} is the scaling factor to quantify variability and $E(\lambda)$ is the μ_a spectrum of the epithelial cells.

The μ_a spectrum of the second layer of the model is described by equation (5):

$$\mu_{a,str} = \mu_a^{collagen} + \mu_a^{Hb}, \quad (5)$$

with $\mu_a^{collagen}$ representing the absorption coefficient of collagen fibers and μ_a^{Hb} representing the absorption coefficient of the hemoglobin in the blood vessels. The $\mu_a^{collagen}$ can be calculated using equation (6):

$$\mu_a^{collagen} = C_{Col} \times C(\lambda), \quad (6)$$

where C_{Col} is the volume of collagen in the stroma and $C(\lambda)$ is the μ_a spectrum of collagen.

Furthermore, the μ_a^{Hb} can be calculated through equation (7):

$$\mu_a^{Hb} = \ln(10) \times C_{Hb} \times [StO_2 \times \varepsilon^{oxy}(\lambda) + (1 - StO_2) \times \varepsilon^{deoxy}(\lambda)], \quad (7)$$

with the concentration of hemoglobin represented by C_{Hb}, the oxygen saturation represented by StO_2 and the extinction coefficients of the oxygenated and deoxygenated hemoglobin represented by $\varepsilon^{oxy}(\lambda)$ and $\varepsilon^{deoxy}(\lambda)$, respectively.

The reduced scattering coefficient for both layers can be calculated by the following equation [13]:

$$\mu_s'(\lambda) = A \times \lambda^{-k}. \quad (8)$$

The *A* and *k* coefficients in Eq. (8) are fitting parameters that should be adjusted to the data obtained from the epithelial (A_{epi} , k_{epi}) and stroma (A_{str} , k_{str}) layers.

Using the hybrid method proposed by the authors, it is possible to estimate the following parameters: k_{epi}, k_{str}, C_{Hb} , StO_2 and the thickness of the epithelium (TH_{epi}) of the tissue model. The C_{epi}, C_{Col}, A_{epi} and A_{str} parameters were fixed with values previously described in the literature. After making the necessary parameter estimations, it is possible to estimate the optical properties of the model, using the R_d data.

The estimation process starts with an initial rough estimation of the parameters. This step is important to reduce the total computational cost of the estimation process. To do the initial estimations, MC simulations were previously performed to create a Look Up Table (LUT). The LUT can be consulted quickly and establishes a relation between

the k_{epi}, k_{str} and TH_{epi} parameters with an R_d spectrum. Since k_{epi}, k_{str} and StO_2 are the parameters with higher influence in the reflectance sensitivity and in the spectral shape, they were the first to be estimated.

The k_{epi} and k_{str} parameters were estimated using the LUT in the 650-800 nm range. This estimation was made three times for the set value of TH_{epi} in the 200 µm, 400 µm and 600 µm. After that, 100 MC simulations with randomized parameters were executed and a correlation between the minimum in the wavelength of 414-432 nm and the StO_2 was established for future estimations of this same parameter. From the MC simulations, presented in Fig. 4, it is possible to see that the StO_2 tends to decrease with the increase of the minimum reflectance wavelength.

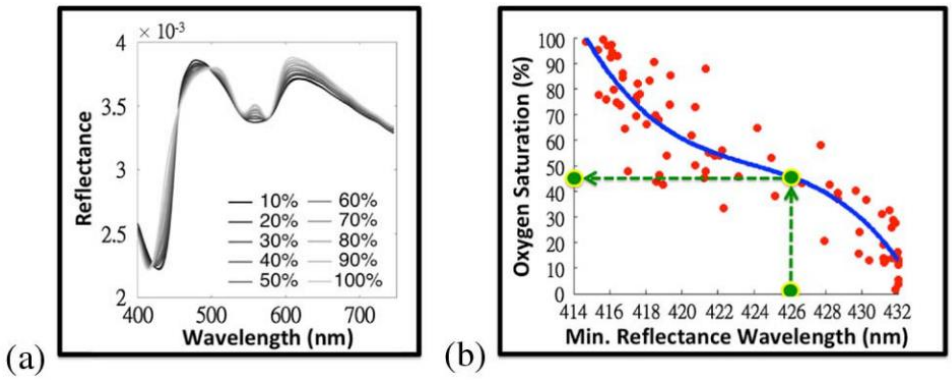


Figure 4. Monte Carlo simulations (a), and relation between the Minimum reflectance wavelength and the StO_2 (b) [19].

Afterwards, the C_{Hb} was estimated for the three TH_{epi} values using MC simulations. The authors varied the C_{Hb} value for each MC simulation and then compared the simulated R_d spectra with the original spectra to evaluate the goodness of the estimation. After estimating the k_{epi} , k_{str} and the C_{Hb} for each of the TH_{epi} values, it was necessary to select the better suited value within the three TH_{epi}. To do this assessment, MC simulations were executed with one of the TH_{epi} values, resulting in a simulated R_d spectrum, which was then compared to the real R_d spectra. In this second run of MC simulations, the k_{epi} , k_{str} and the C_{Hb} values, for each TH_{epi} value are already known from the previous MC simulations.

After estimating these initial parameters, they were used to run MC simulations again to reach the final solution faster. Between the MC simulations, a pseudo-inverse operation was performed to determine the amount of adjustment that the parameters needed. This algorithm was executed a maximum of 30 times or until the simulated spectrum presented an error bellow 6%. If 30 iterations were reached, the selected combination of parameters was the one with the least spectral error. The final estimated R_d spectra are presented in Fig. 5.

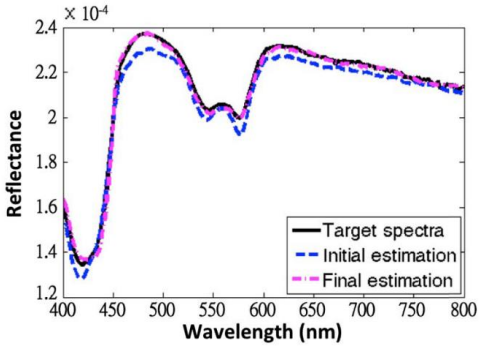


Figure 5. Comparison between the initial estimation (blue), the final estimation (pink) and the target spectra (black) [19].

The presented hybrid method also showed that it was capable of estimating the parameters with a lower associated error, when compared to the conventional fitting process, as its possible to see in Fig. 6.

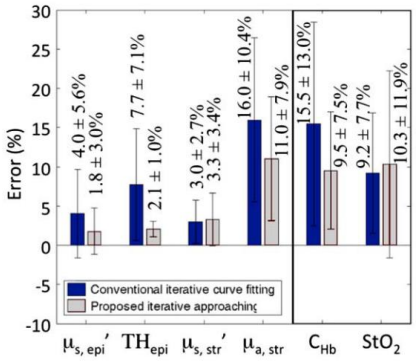


Figure 6. Comparison between the error associated with the present approach and the one associated with the conventional fitting process [19].

Finally, the authors were able to reduce the computational cost in the fitting process by using pre-simulated spectra and creating LUT. Furthermore, the estimated spectrum was close to the targeted spectrum and the estimated parameters had a low error. However, due to the fact that it was necessary to do an initial estimation of the parameters to reduce the computational cost of the MC simulations, this new hybrid method tends to be more complex than other approaches reported in the literature.

2.3. Machine Learning Models

ML is an evolving field that has the capability to revolutionize different services in healthcare such as the diagnostic of diseases like cancer, pharmacogenomics and the administration of hospitals [20]. ML models are also increasingly used in the field of biophotonics to perform classification and estimation tasks. The objective of such research is to estimate the optical properties of tissues or phantoms using the non-invasive R_d measurements.

The work in Ref. [21] uses neural networks to estimate optical properties of tissues. To our knowledge, this paper was one of the first that used neural networks with the objective of estimating tissue's optical properties. First, the authors created a database from spatially resolved diffuse reflectance at eight different distances between the source and the detector. Then, the architecture of the neural network was defined with an input layer and a hidden layer with eight nodes and an output layer with two nodes as seen in Fig. 7.

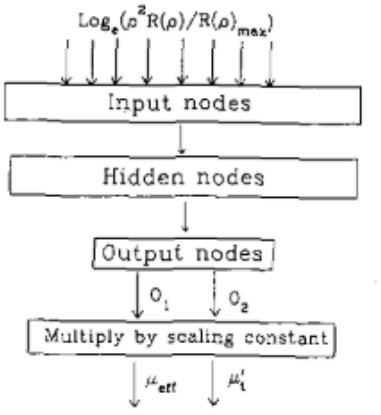


Figure 7. Architecture of the neural network used in Ref. [21].

For the training process, 100 sets were generated using the diffusion theory [22] , and the same number of datasets were generated in the testing process. Figure 8 presents a correlation graph between the parameters returned by the neural network and the data used as input in the estimations.

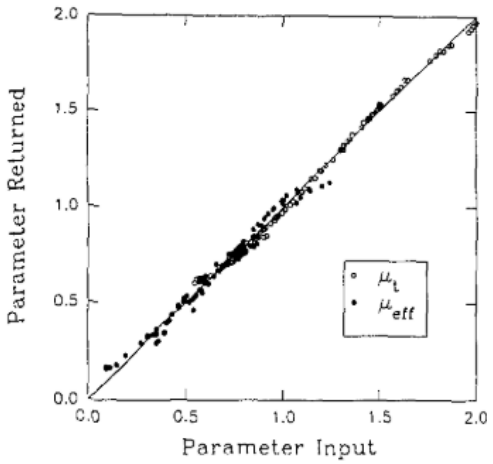


Figure 8. Correlation between the optical properties estimated by the neural network and the data used as input in that estimation [21].

The neural network was able to perform the necessary estimations with an error smaller than 7%. Such error shows a better performance for the neural network estimations when compared with traditional nonlinear least-square techniques. However, the performance of the neural network could be further improved if the simulated data that was used to train the neural network, was replaced with experimental data collected from real tissues. *Kienle et al.* [23], developed neural networks that were capable of estimating optical properties from reflectance data. Such procedure is of great interest, since it mimics the realistic noninvasive method to estimate the optical properties of *in vivo* tissues. One of the neural networks, nominated neural network 1 (NN1), consisted of 11 input nodes, 11 hidden nodes and 2 outputs and a smaller neural network, nominated neural network 2 (NN2), consisted of 9 input and hidden nodes and two output nodes. The authors selected to use two different neural networks to accommodate the range of attenuation that light suffers when interacting with real tissues. To create a dataset to train the neural networks, 120 MC simulations were executed and to evaluate the performance of the neural networks, 13 phantoms, with their optical properties measured, were used.

The comparison between the measured optical properties and the results from the neural networks are presented in table 1.

Table 1. Measured optical properties of tissue phantoms and the estimated results from the neural networks [23].

True Optical Properties		Neural Network Results	
μ_a (mm^{-1})	μ_s' (mm^{-1})	μ_a (mm^{-1})	μ_s' (mm^{-1})
0.0022	1.99	0.0023	1.99
0.0057	1.98	0.0047	1.95
0.0143	1.97	0.0150	1.97
0.0033	0.98	0.0034	1.00
0.0088	0.98	0.0083	1.03
0.025	0.97	0.022	0.99
0.070	0.94	0.075*	0.95*
0.100	0.93	0.107*	0.96*
0.0022	0.50	0.0017	0.52
0.0065	0.50	0.0053	0.52
0.020	0.49	0.020	0.51
0.043	0.49	0.048*	0.49*
0.073	0.49	0.083*	0.48*

The values marked with an asterisk in Table 1 are the results from NN2 and the non-marked results are from NN1 [23]. The resulting root mean square error for the μ_s' was 2.6 % and 14 % for the μ_a which shows that ML models are capable of estimating optical properties with some degree of trust.

Zhang et al. [24] developed a neural network, which was trained with MC simulations and R_d spectra that were measured using the setup seen in Fig. 9.

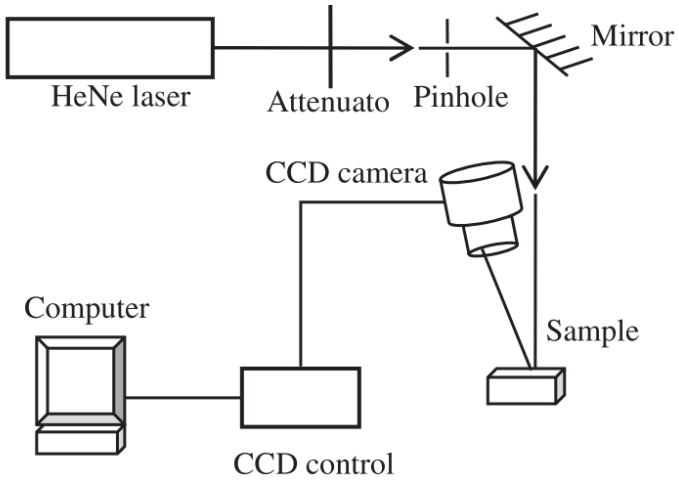


Figure 9. Experimental setup to measure spatially resolved R_d spectra [24].

The architecture of the neural network consisted on three layers in which the output layer had 2 nodes. Each output node of the neural network represented μ_a or μ'_s. The training dataset was composed by 32 MC simulations of R_d spectra and tested with 10 simulated spectra. The maximal relative errors observed from the neural network estimations was 6% for μ_a and 4.5% for μ'_s. After training and testing the neural network with MC simulations, experimental R_d spectral data from 11 phantoms were used as input in the neural network to evaluate its performance with experimental data. The resulting maximal relative errors were 25 % and 50% for μ'_s and μ_a, respectively. This increase in the error of the neural network estimations is due to the fact that the training process was made using data from MC simulations and the validation process was made using experimental data. However, when the neural network was trained and tested with experimental data, the maximal relative errors dropped to 8% and 16% for μ_a and μ'_s respectively. As a conclusion to this study, one should consider that ML models that are trained and tested with data from MC simulations may not be suitable to be used with experimental measurements.

In Ref. [25], the authors created a metamodel to efficiently predict optical properties of liquid phantoms from spatially resolved spectroscopy measurements. In total 36 liquid phantoms were prepared using intra-lipids and Indian ink. They measured the R_d spectra of these fluids using the experimental setup presented in Fig. 10. A total of 102 R_d spectra were measured from each liquid phantom.

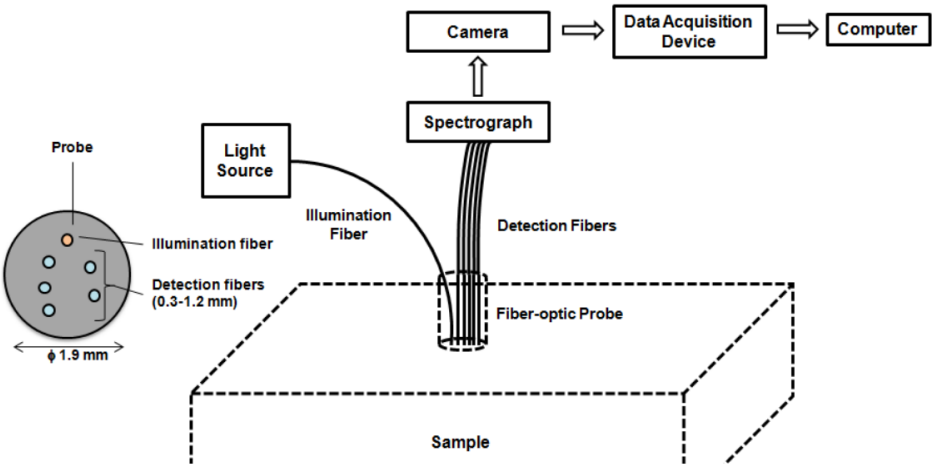


Figure 10. Experimental setup to measure the R_d spectrum of the liquid phantoms [25].

The experimental data was divided in two sets: 16 samples for the test set and 20 samples for the training set. The performance of the neural network was evaluated using the cross-validation process. This resulted in 1632 (16 × 102) estimated optical properties. The comparison between the predicted and the measured optical properties is presented in Fig. 11.

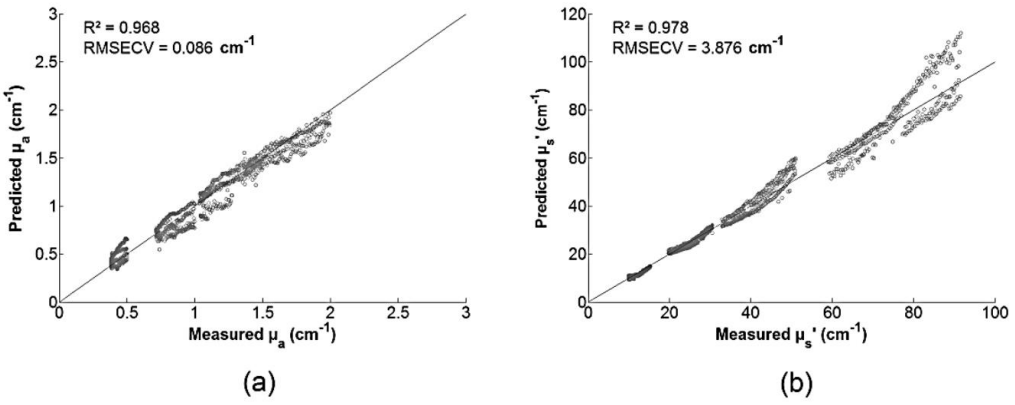


Figure 11. Scatter plot of the predicted and measured optical properties μ_a (a) and μ'_s (b) of the phantoms used in the test set [25].

The absolute error associated with the estimation of μ_a was 0.0086 cm^{-1} and the error for the μ'_s was 3.876 cm^{-1}. These results show that it is possible to apply an inverse metamodeling approach for the estimation of the optical properties. However, it is still necessary to evaluate the impact of noisy data in the performance of the metamodel.

Panigrahi el al. [26] developed a Random Forest Regression (RFR) model to estimate the optical properties from reflectance measurements. Training and validation datasets were generated using data from MC simulations. From the validation process, the RFR model achieved an error of 0.556 % for μ_a and 0.126% for μ'_s. Additionally, the performance of the ML model was compared with the performance of other estimation algorithms such as the LUT, as it is seen in Fig. 12.

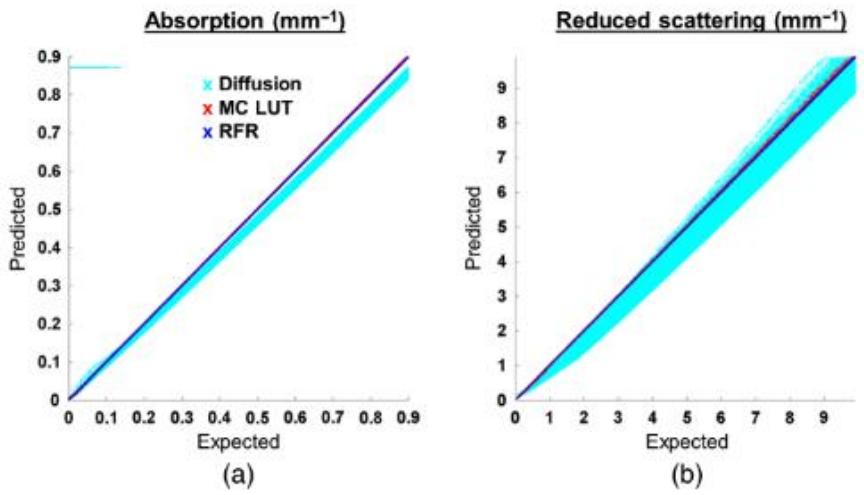


Figure 12. Comparison between the estimated and real values of μ_a (a) and μ'_s (b) from the study in Ref. [27].

The estimations from the different algorithms that are represented in Fig. 12 are as follow: diffusion theory is represented in cyan, MC LUT is represented in red and RFR is represented in blue [26]. Afterwards, the RFR model was tested with reflectance data acquired from tissue phantoms and its performance was compared to the one of the LUT. The RFR model estimated on average with an associated error of 0.16% in μ_a and 0.01% in μ'_s. As a final test, the optical properties of the surface of a human hand were estimated using LUT and the RFR model. The mean error obtained using the LUT was 0.8% and 0.14% for μ_a and μ'_s respectively, while for the RFR model was 0.16% for μ_a and 0.01% for μ'_s. From this result, it is possible to see that ML algorithms are capable of estimating optical properties with a similar error to the one obtained in the standard LUT method. Furthermore, ML models have the advantage of being able to processing more complex data environment.

2.4.Summary

After analyzing all these studies, it is possible to see that there is a tendency to use data from MC simulations in the process of estimating the desired optical properties. However, these simulations tend to be computationally expensive and therefore not very practical to be used in a clinical scenario. Furthermore, by using the diffusion theory or mathematical models, we have to make assumptions about the tissues that eventually could influence our estimations in a negative way. Even though the strategies here discussed are capable of estimating the optical properties with low associated error, there is still room for improvement. Also, the ML models that are described in the literature are

normally trained with simulated data or phantom data and not with real data originated from measurements of tissues and therefore, to access the real performance of these ML models, more studies are needed. However, because there is still no mathematical relation between the optical properties of a tissue and its R_d data, ML models could be used to estimate the desired optical properties of a tissue and, by using noninvasive experimental data to train the ML models, there would be no need to use MC simulations.

CHAPTER 3 – MATERIALS AND METHODS

3. Materials and Methods

In this work, the following ML algorithms were used to estimate the μ_a, the μ_s and the refractive index (RI) spectra of human colorectal mucosa tissues: Single Layer Perceptron (SLP), K Nearest Neighbors (KNN), RFR, Decision Tree for Multioutput Regression (DTFMR) and the Linear Regression for Multioutput (LRFMO). The SLP algorithm was developed using the TensorFlow framework for Python and the other algorithms can be found in the scikit-learn library for Phyton as well. The μ_a, the μ_s and the RI spectra that were used to train the models were previously obtained in the study of Ref. [18]. In total 20 spectra were used, 10 of which correspond to normal and 10 correspond to pathological colorectal mucosa tissues from humans. All these spectra were obtained with the following procedure.

3.1. Tissue sample collection and preparation

The experimental setups presented in Fig. 13 were used to measure the T_t and the R_t spectra from 10 normal and 10 pathological tissue samples. The pathology was confirmed by the Portuguese Oncology Institute of Porto to be adenocarcinoma, commonly designated as colorectal cancer. All spectra were acquired between 200 and 1000 nm.

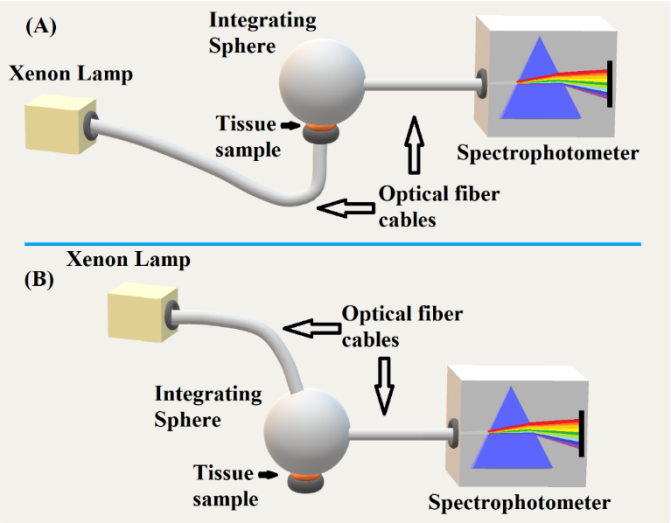


Figure 13. T_t (A) and R_t (B) experimental setups.

All the equipment used in these measurements belongs to the Center for Innovation in Engineering and Industrial Technology (CIETI) and was acquired from Avantes™, in the Netherlands.

The calculation of the μ_a spectra was made according to the following equation [9] [18]:

$$\mu_a(\lambda) = \frac{1-\left(\frac{T_t(\lambda)+R_t(\lambda)}{100}\right)}{d} \quad (9)$$

where $T_t(\lambda)$ and $R_t(\lambda)$ are the spectra measured with the setups in Fig. 13 and d is the sample thickness, which was set to 0.5 mm. This calculation was made for all the 20 samples, 10 normal and 10 pathological. An example of a μ_a spectra, that was calculated using Eq. (9) is represented in Fig. 14.

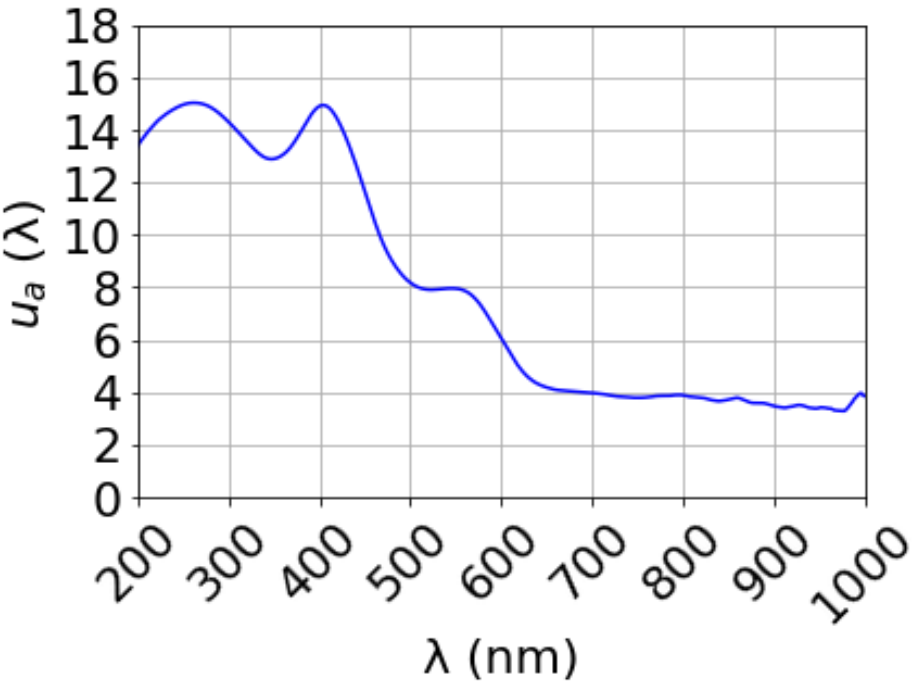


Figure 14. Example of a μ_a spectrum from the training dataset.

To use as input in the development of the ML algorithms, a set of 20 R_d spectra, 10 from normal and 10 from pathological mucosa tissues were measured, particularly for the present study. The setup used to perform those measurements is presented in Fig. 15.

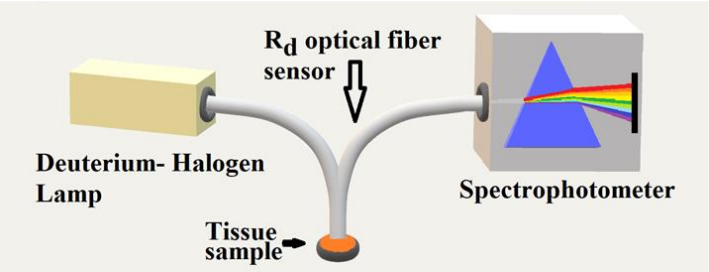


Figure 15. R_d experimental setup.

The spectrophotometer used in these measurements is the same that was used in the T_t and R_t measurements. The Deuterium-Halogen lamp also belongs to CIETI and was also acquired from Avantes™. The spectra acquired with this setup also range from 200 to 1000 nm.

3.2.Machine learning methods

3.2.1. Classification of the Diffuse Reflectance spectra

In a hypothetical clinical scenario, the pathology diagnosis from the R_d spectra is not direct. However, by using a distinguishing feature or a ML model, the diagnosis process could be automated. The parameter that was first studied to be used as a distinguishing feature between the two categories was the slope of the R_d spectra in the 750-850 nm domain. The equation that was used to fit the R_d values to calculate the slopes of the spectra in the 750-850 nm was the following:

$$R_d(\lambda) = m \times \lambda + b \tag{10}$$

where m is the slope of the R_d spectra and b is the y-intercept. After retrieving the slopes, they were analyzed to verify if this was a reliable feature to classify the R_d spectra. To classify the spectra, the Selector Vector Machine (SVM) algorithm was used. This ML algorithm can be found in the scikit-learn library for Phyton.

3.2.2. Estimation of the optical proprieties

The R_d spectra was first retrieved from Ref. [18] and then organized in a matrix where the lines corresponded to the number of the samples and each column corresponded to a single wavelength. Then, this matrix was selected as input data to each ML model.

To estimate the optical properties of the tissue samples from the R_d spectra, different ML models were studied. For each ML algorithm, two experiments were made. In the first experiment, two machine learning models were trained only with normal or pathological samples (Trained Separately-TS). In the second experiment, only one ML model was trained with all the available normal and pathological samples (Trained Together – TT model).

After defining which ML models to use, the training process needed to be adjusted, taking into account the low number of samples available. Therefore, the Leave One Out (LOO) method was adopted to train the models. This method consists in the following steps: select a random sample out of the training dataset, train the model with the remaining samples and validate the model using the sample that was left out by comparing the estimated μ_a spectrum with the measured μ_a spectrum of the tissue sample. This process is repeated equally to the number of used samples. The final result of this process is the average of the estimated spectra.

The estimated μ_a spectra were then used to calculate a single mean spectrum, both for the normal and pathological samples. For the TT models, the estimated spectra were separated in the normal and pathological category, before the calculation of the mean spectrum. Afterwards, the mean reference spectra (mean of the calculated μ_a spectra from the experimental measurements with tissue samples [18]) and the mean estimated spectra were compared to access the model performance. To perform a further model performance evaluation, the Euclidian Distances (ED) were calculated between the measured and estimated spectra, according to equation (11):

$$ED = |a - b| \tag{11}$$

where a represents the estimated μ_a value and b represents the reference μ_a value.

As previously mentioned, in the models that were trained with all the samples, after the LOO method, the estimated spectra were separated depending on their previously known category. For example, if the R_d spectra that was used in the validation process was classified as pathological, then the estimated μ_a spectra was labelled in the same category.

During the training process, that was described above, the hyperparameters of the models were tuned, when necessary, to improve their individual performance. The range in which the hyperparameters were tested can be seen in table 2.

Table 2. Range of the experimented hyperparameter for each ML model.

		ML models	
ML algorithm	**Hyperparameter**	TS models	TT models
SLP	Number of Layers	2 to 5	2 to 5
KNN	Number of Neighbours	1 to 9	1 to 19
RFR	Numbers of tress	1 to 9	1 to 19
DTFMR	Depth of the tree	1 to 4	1 to 4

The SLP model was defined with an input shape of 801, so that each wavelength of the spectrum, which ranged between 200 and 1000 nm with 1 mm of resolution, was interpreted as a feature. The architecture was defined as an input layer with 10 nodes and an output layer with 801 output nodes. This way, it was possible to obtain an estimated spectrum with the same resolution as the measured R_d spectrum. The dimension of the architecture was set only with an input and an output layer to minimize the parameters of the neural network and consequently preventing overfitting. An example of the code used to define and train a SLP model can be seen in Annex 1 in Figure S41.

In the KNN algorithm, the number of neighbours (k value) was set as 5 because more increments resulted in error increasing associated with the estimated μ_a spectrum. During the fine-tuning process, the k value was set between 1 and 9 for the models that were trained with only one type of sample (TS model), and between 1 and 19 for the model that was trained with all the samples (TT model). An example of the code used to define and train a KNN model can be seen in Annex 1 in Figure S42.

Regarding the RFR algorithm, the number of trees was set as 5 because, in the initial computational experiments, further increments did not improve the spectral shape estimation.

The DTFMR and the LRFMO are simpler algorithms with less complex hyperparameters. In the DTFMR algorithm, the depth of the tree was fixed at 4 to prevent overfitting and the LRFMO algorithm automatically finds the best slope for the input data.

Following the procedure presented in Ref. [18], it was necessary to subtract the μ_a spectrum of Lipofuscin from the μ_a spectra of the tissues, for an accurate evaluation of the blood content in both types of tissues. Lipofuscin is a pigment that accumulates in tissues and induces tissue and cell degeneration, a process that is associated with the tissue ageing progression [27]–[29]. The μ_a spectrum of Lipofuscin for colorectal tissues has been describe as the following [9][18]:

$$\mu_{a-\mathrm{lip}}(\lambda) = A \times (5.2 + e^{(3.524-0.01087\times\lambda)}), \quad (12)$$

where A is set to 1 for the normal tissue samples and 1.1 for the pathological tissue samples. This adjustment is necessary because cancerous tissues tend to have more pigment than healthy tissues as reported in Refs. [9][18] .

Following the estimation of $\mu_a(\lambda)$, the experiments regarding the estimation of $\mu_s(\lambda)$ were started. These spectra where calculated using the Bouguer–Beer–Lambert law [5][6][18]:

$$\mu_s(\lambda) = -\frac{\ln[T_c(\lambda)]}{d} - \mu_a(\lambda), \quad (13)$$

with $T_c(\lambda)$ representing the measured T_c spectra, which were available from the study reported in Ref. [18], d representing the sample thickness (0.5 mm) and $\mu_a(\lambda)$ the absorption spectra that were calculated with Eq. (9). The experimental μ_s spectra had a small bump near 400 nm, possibly some remaining evidence of the Soret band from hemoglobin. To eliminate this bump, the μ_s data between 350 and 450 nm were ignored and the remaining data was fitted with the following equation [3]:

$$\mu_s = a \times (f_{Ray} \times \left(\frac{\lambda}{500\ (nm)}\right)^{-4} + (1 - f_{Ray}) \times \left(\frac{\lambda}{500\ (nm)}\right)^{-b_{Mie}}) \quad (14)$$

with a representing the μ_s value of the tissue at 500 nm, f_{Ray} representing the fraction of Rayleigh scattering and b_{Mie} representing the mean size of Mie scatterers. After fitting the μ_s data with curves described by Eq. (14), all μ_s spectra became smooth. Once the reference μ_s spectra were calculated, we initiated the μ_s estimations with the same procedure as in the estimations of the μ_a spectra.

After finishing the μ_s estimations, the next step was to estimate the RI of the normal and pathological tissues. To perform these estimations, the first step was to calculate 10 RI spectra from experimental data. As described in Ref. [18], the reference RI spectra

were calculated from μ_a spectra through Kramers-Kronig (K-K) relations. First, the imaginary part of the RI spectra ($\kappa(\lambda)$) was calculated with [30][31]:

$$\kappa(\lambda) = \frac{\lambda}{4\pi}\mu_a(\lambda), \tag{15}$$

After calculating $\kappa(\lambda)$, the real part of the RI spectra could be calculated with [30]–[32]:

$$n_{tissue}(\lambda) = 1 + \frac{2}{\pi}\int_0^{\infty} \frac{\lambda_1}{\Lambda} \times \frac{\lambda_1}{\Lambda^2 - \lambda_1^2}\kappa(\Lambda)\mathrm{d}\Lambda \tag{16}$$

where Λ is the integrating variable and λ_1 is a fixed wavelength that can be adjusted for optimal calculation of the RI spectra of the tissue [30]. Once the reference RI spectra were calculated through Eqs. (15) and (16), the estimations of the RI spectra were made using the μ_a spectrum estimations from the RFR TS model. The reference and estimated RI spectra were then compared.

Since we have used the same experimental data as in Ref. [18] to calculate the reference spectral optical properties of human colorectal mucosa tissues, most of the above description to prepare the samples, to perform the spectral measurements and following calculations is the same as described in that research work. After retrieving the necessary spectral data, the ML experiments were initiated. To classify the R_d spectra an SVM model was used and its performance evaluated. Furthermore, the spectral shape estimation was also performed using ML models. Each ML algorithm was carefully tuned to achieve the best performance and to prevent overfitting. The ML models were trained using the LOO method. The ED between the reference and the estimated μ_a spectra were calculated for each ML algorithm, so that it would be possible to access their performance.

CHAPTER 4 – RESULTS AND DISCUSSION

4. Results and Discussion

In this chapter, we present the spectral estimation of the optical porperties and also the R_d spectra classification as normal or pathological. The μ_a and the μ_s spectra were estimated using the mentioned above ML models and the RI was calculated using the estimated μ_a spectra from a ML model in Eqs. (15) and (16). The dataset used in the ML experiments can be found in Ref. [18]. R_d spectral measurements were acquired from each tissue sample and the reference spectral optical properties, to be used in the ML estimations, were also calculated for each sample. In total, 10 normal and 10 diseased tissue samples were used to acquire the spectral measurements that were used to calculate the reference optical properties. To perform the R_d measurements, a similar number of tissue samples was used. The pathological tissue samples were histologically analysed at the Portuguese Oncology Institute of Porto and they were diagnosed with colorectal adenocarcinoma, following the WHO classification [9].

4.1. Diffuse Reflectance classification

During the training process of the ML TT models, the estimations were automatically labelled in the same category of the R_d spectrum that originated them. Furthermore, the labels of the R_d spectra were known because the tissues that originated these spectra were previously labelled by medical professionals as: normal or pathological mucosa diagnosed with colorectal cancer. In a hypothetical clinical scenario, it would be necessary to use the R_d spectrum as a way to diagnose the mucosa of the patient and, after labelling the R_d spectrum, it would be possible to choose the TS model that was trained with the samples of the same category.

Before starting the classification process using ML models, it is necessary to have general view of the available samples. The individual R_d spectra that were measured from each individual sample are presented in Fig. 16.

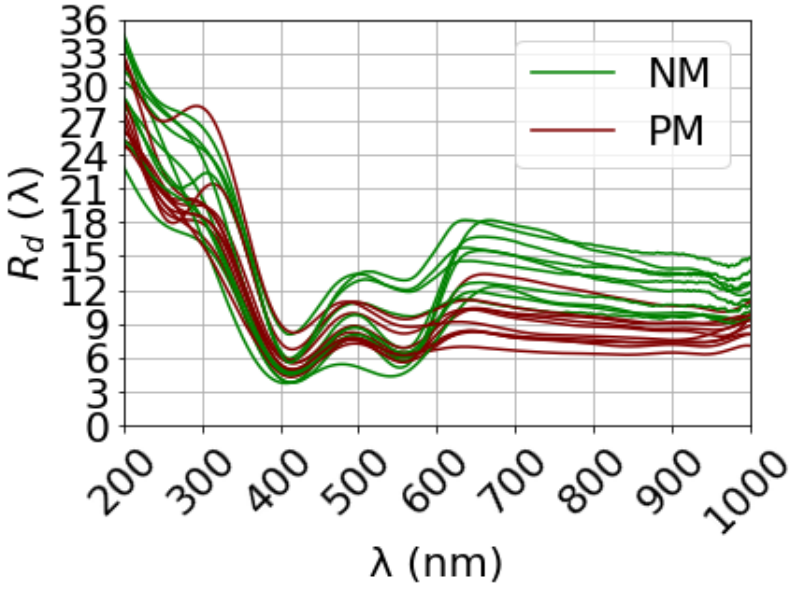


Figure 16. Individual R_d spectra of the Normal Mucosa (NM-green) and Pathological Mucosa (PM-red)

From Fig. 16, it is possible to see that between 650 and 1000 nm, a difference in the slope of R_d occurs between the normal and pathological tissues. Therefore, a restricted spectral range (750-850 nm) where the R_d spectra present a linear behaviour was selected to calculate the slopes that correspond to the normal and pathological tissue samples. Fig. 17 presents those spectra in that restricted range of wavelengths.

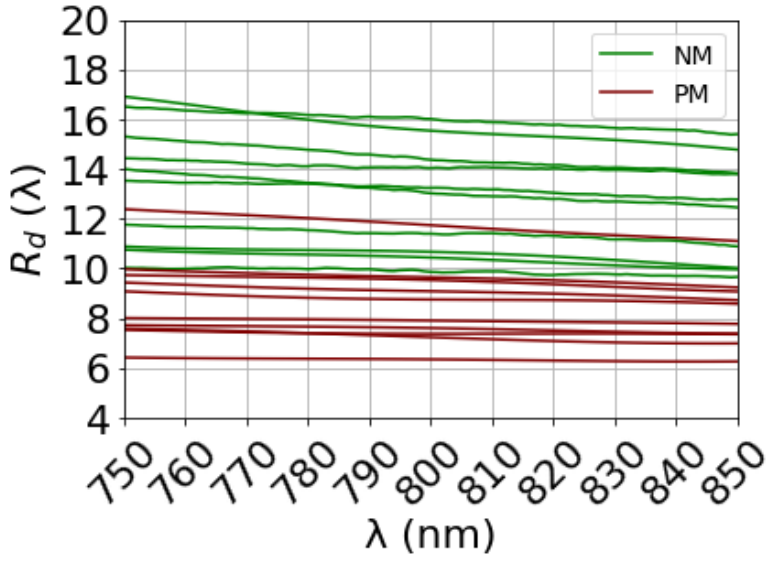


Figure 17. Spectral range selected to calculate the slopes of the individual R_d spectra for further classification.

After analyzing and restricting the spectra to the 750-850 nm range, Eq. (10) was used to fit each spectrum and retrieve the corresponding slope. These slopes were organized in two categories, as presented in Fig. 18.

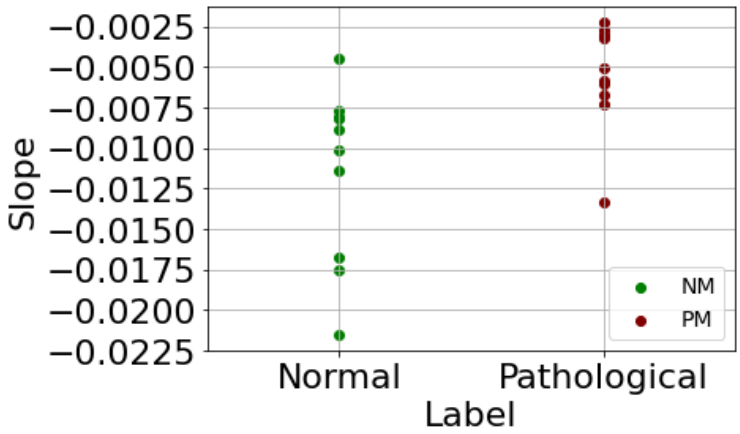


Figure 18. Slopes from the R_d spectra separated in Normal Mucosa (NM-green) and Pathological Mucosa (PM-red)

From Fig. 18, we see that the slopes of the R_d spectra from normal and pathological tissue samples tend to overlap and therefore, this is not the best discriminating feature.

The next step in attempting to classify the R_d spectra was to use a polynomial SVM to do the classification process.

In preliminary testing, all of the R_d spectra were used as input data without restraining the domain, but the accuracy only reached 75%. To improve the performance of the SVM model, it would be necessary to select the wavelengths where the R_d spectra present the most significant differences between normal and pathological samples.

In the next experiment, only one wavelength was used as feature and then the LOO method was used to train the SVM model. For each wavelength, there were 10 R_d values from the normal and other 10 from the pathological samples. This experiment had the goal of analysing the spectra and finding the best wavelengths were the SVM model presents the best performance. The error associated with the SVM model classification is presented in Fig. 19.

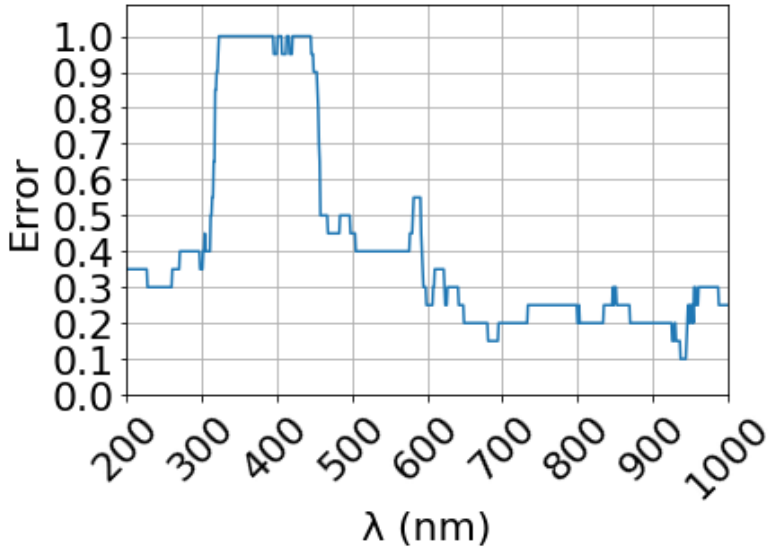


Figure 19. Error rate of the SVM model when the R_d values from each wavelength are used as input data. The SVM model used in this experiment has a polynomial kernel of third degree.

From the experiment above, it is possible to notice that there is a small domain between the 900 and 1000 nm, where the error rate is at its lowest. In further inspection of the graph in Fig. 37, it was noticed that such domain is located for wavelengths between 937 and 945 nm and therefore, this is the domain that would be used as input to the SVM model.

After training the SVM model with the R_d data within the 937-945 nm spectral range, we reached an accuracy of 95%, as it is seen in the Confusion Matrix presented in Fig. 20.

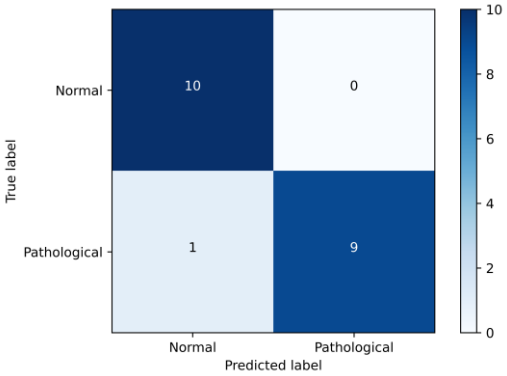


Figure 20. Confusion Matrix of the SVM model trained to classify the R_d spectra in the 937-945 nm spectral range.

Figure 20 shows that only one sample was wrongly labelled. This false negative could mean, in a hypothetical clinical scenario, that the patient that in reality had the disease, would not be diagnosed with colorectal cancer. Nonetheless, besides the discussed outliar, the model performed as planned and with a performance comparable with other SVM models from other studies [33]–[35]. However, using only eight wavelengths can be a small number of features in certain applications and therefore a bigger number of wavelengths was selected to increase the robustness of the SVM model. By analysing Fig. 34, it is possible to see that the spectral range where the overlapping of R_d spectra between normal and pathological tissues is smaller is between 700 – 1000 nm. Therefore, the SVM model was re-trained using the R_d in this domain and also the LOO method. The SVM model reached an accuracy of 90%, as seen in the Confusion Matrix in Fig. 21.

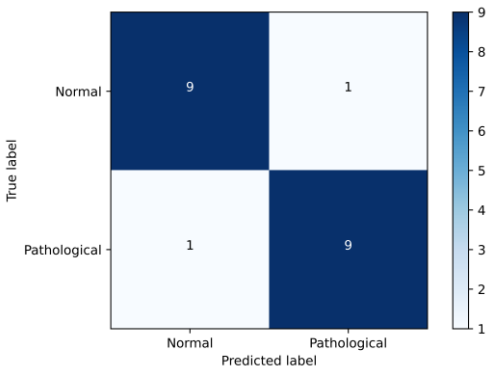


Figure 21.Confusion Matrix of the SVM model trained to classify the R_d spectra in the 700-1000 nm spectral range.

Figure 21 shows that an additional false negative was predicted by using the wavelength between the 700-1000 nm. This could mean that in a hypothetical clinical scenario, the patient would be diagnosed with colorectal cancer and probably undergo unnecessary treatment.

4.2. Absorbance coefficient estimation

As in the study of Ref. [18], we started our research by calculating the 10 μ_a spectra from the T_t and R_t spectral measurements that were made available from that study. For that calculation we used Eq. (9) and the calculated spectra were separated according to their provenience from normal and pathological mucosa. Considering the R_d spectra that were acquired from the colorectal mucosa tissues to be used as input in the ML algorithms, we have catalogued them also as normal or pathological. First and second order statistics were applied to the R_d and μ_a spectra, resulting in the graphs presented in Fig. 22.

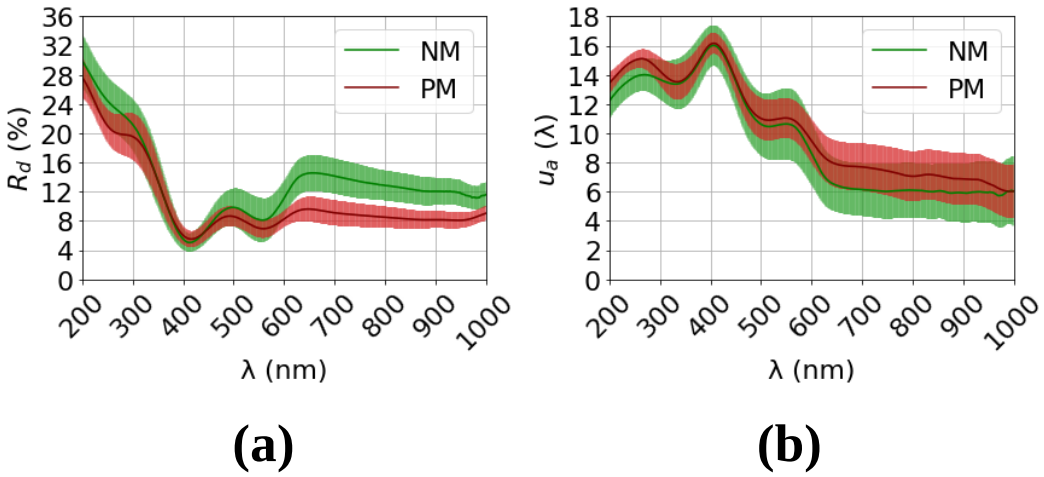


Figure 22. Mean and dispersion of R_d (a) and μ_a (b) spectra of the normal (NM-green) and pathological (PM-red) mucosa.

After collecting all necessary data, the individual samples that originated the graphs in Fig. 16 were used to train the already mentioned ML models. To compensate for the low number of samples, the LOO method was applied during the training process of the ML models.

The mean spectral estimations obtained with the SLP model are presented in Fig. 17, while the individual estimations that were used to calculate the mean spectra can be seen in Annex 1. The mean spectrum that was calculated using the individual estimations from the LOO method is denominated as Mean Estimated Spectrum (MES) while the mean spectrum that results from the experimental measurements is designed as Mean

Reference Spectrum (MRS). In Fig. 23, the panels are divided in two columns, normal (N) and pathological (P) spectra, and two lines – TS model and TT model.

4.2.1. Single Layer Perceptron

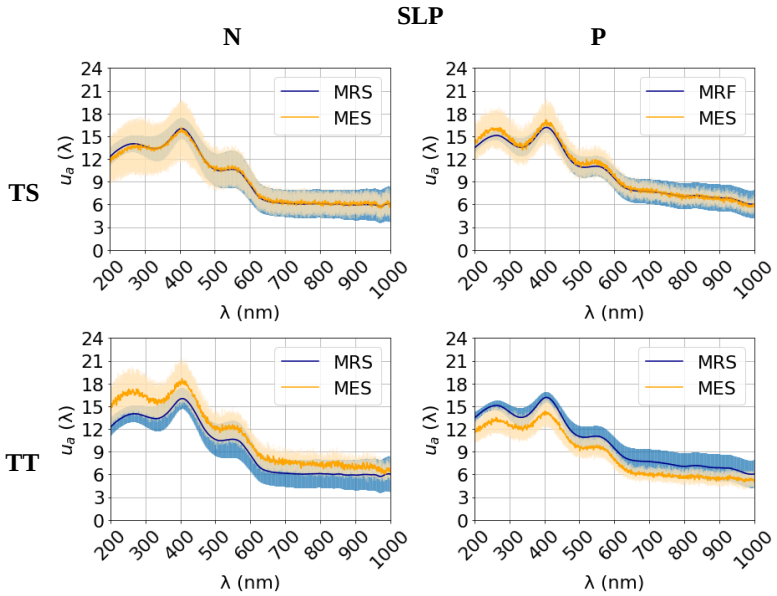


Figure 23. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the SLP algorithm. The individual estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S1-S4).

By comparing between the TT and the TS models, it is possible to see that the later has a better performance because the MRS and the MES tend to overlap more. In addition, the MES tends to have a higher standard deviation (SD) in the nonlinear domain (200-600 wavelength) when compared to the MRF. This fact can be justified by the linearity of the mathematical model that governs the perceptron response to input data. Because the perceptron is the building block for neural networks, the SLP model will have more difficulty in estimating nonlinear data.

4.2.2. K Nearest Neighbour

The estimations obtained with the KNN model are presented in Fig. 18, which is organized in the same manner as Fig. 24.

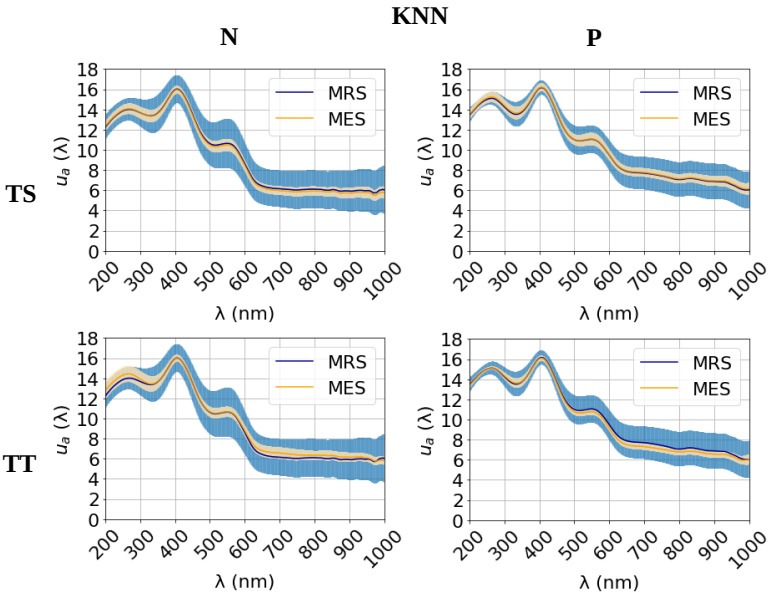


Figure 24. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the KNN algorithm. The individual estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S5-S8).

By analysing the estimations form the KNN models, it is possible to see that overall, this algorithm has a good performance. However, the MES tends to be close to the MRS and with a low SD. This could mean that the KNN models are overfitted. Still, by analysing the individual estimations (Figure S5-S8 in Annex 1) it is possible to see that in each individual sample the model estimates differently, which means that the models are not overfitted. The overfitting process occurs when there is a memory leak from the data to the ML model. In other words, the model memorizes the data and does not reach an optimal solution. To prevent this, the training process of the SLP was stopped before reaching overfitting values and in the rest of the ML models the hyperparameters where fine-tuned to reach better performance.

4.2.3. Random Forest Regression

Similarly to the KNN models, the estimations made by the RFR models, seen in Fig. 25, have an overall good performance except the mean pathological spectra from the TT model. This can be due to the fact that the individual estimations were lower than expected, which consequently leads to a lower MES.

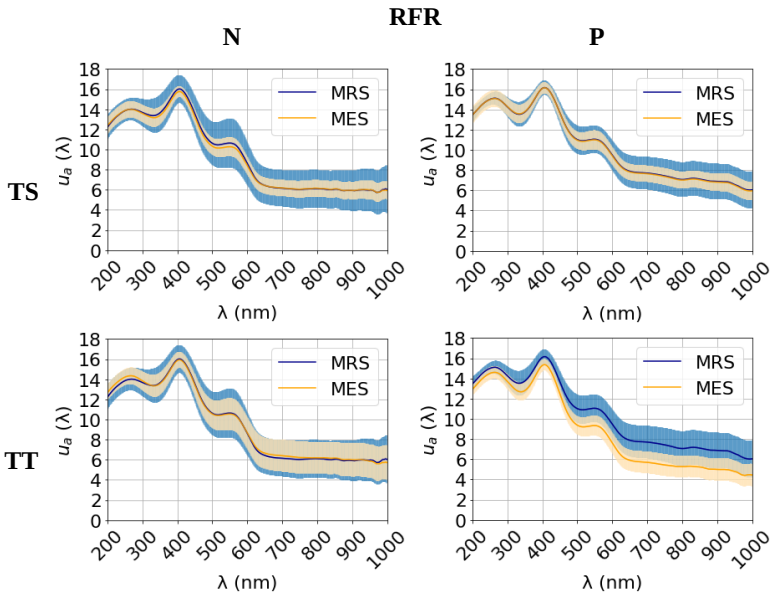


Figure 25. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the RFR algorithm. The individual estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S9-S12).

4.2.4. Decision Tree for Multioutput Regression

The DTFMR algorithm did not have a good performance as the other ML algorithms presented above. Such results are presented in Fig. 26.

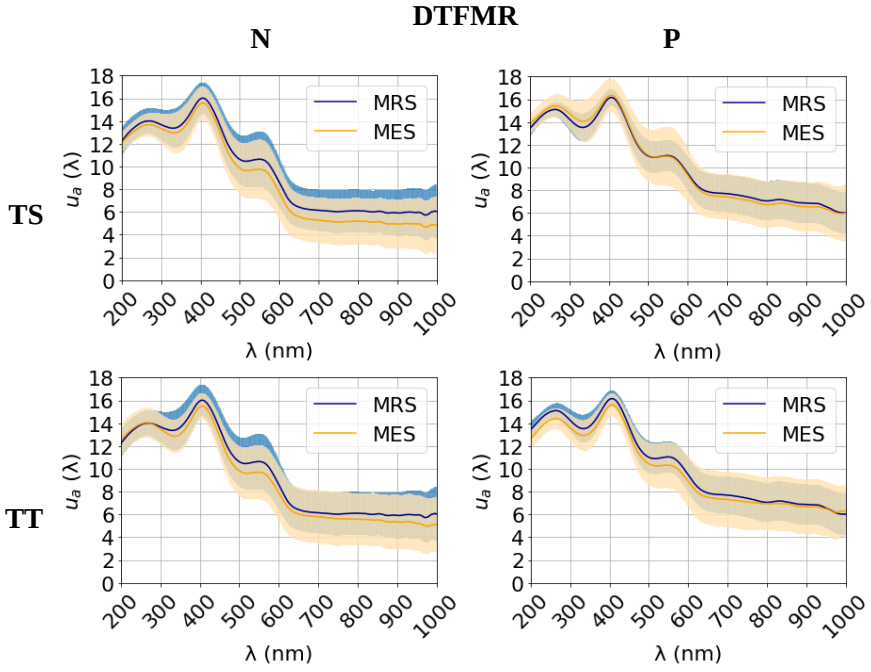


Figure 26. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the DTFMR algorithm. The individual

estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S13-S16).

This bad performance can occur because when compared with the RFR algorithm, the DTFMR algorithm uses only one decision tree to perform the required spectral estimations and so the estimations given by the models are going to be less accurate.

4.2.5. Linear Regression for Multioutput

The final ML algorithm that was studied was the LRFMO and was the one with the worst performance among the presented ML algorithms (see Fig. 27). As it is possible to see in Annex 1 (Fig. S13-S16), the individual estimations tended to be significantly above or below the reference spectra, which consequently lead to a MES with a higher SD than expected.

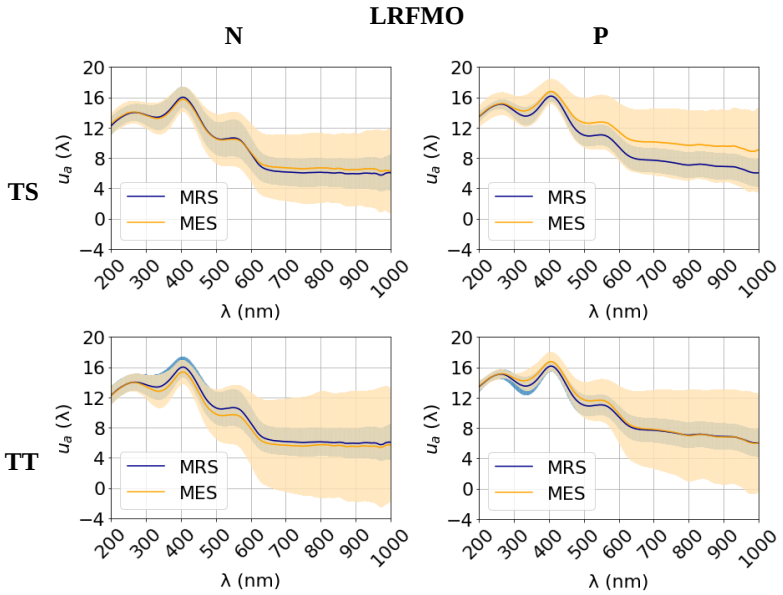


Figure 27. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the RFR algorithm. The individual estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S14-S20).

To quantify the performance of each ML algorithm, the ED between the MRS and the MES for each ML model was calculated using Eq. (11). Fig. 28 presents the ED for each ML model, allowing to make a performance comparison between models.

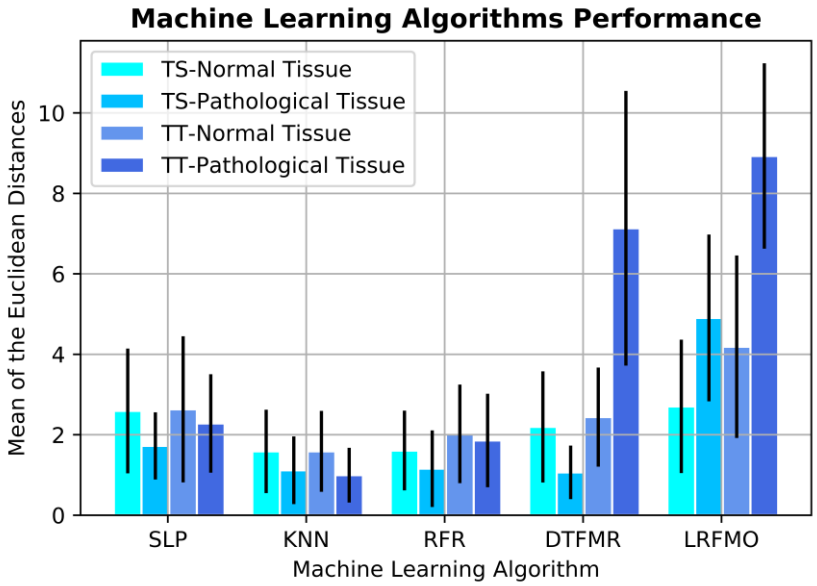


Figure 28. Average of the Euclidean distances for the different models when they are trained with data from separated samples (TS) or with data from all samples (TT).

From Fig. 28, we see that within all ML algorithms, the TT approach seems to reproduce the worst results. This difference in performance seems to have the highest values in the LRFMO and the DTFMR algorithms. Besides these two ML algorithms, the SLP, KNN and the RFR models seem to present an overall better accuracy when estimating μ_a.

Considering these factors, the MES from the three best ML algorithms were used to replicate the calculations seen in Ref. [18] to subtract the absorption of lipofuscin ($\mu_{a\text{-}lip}$) from the absorption of the mucosa tissues. The goal of such calculations is to allow the evaluation of the true blood content in the normal and pathological tissues. Figure 29 shows the results of these calculations, presenting the original $\mu_a(\lambda)$ of the tissue as obtained in the estimations, $\mu_{a\text{-}lip}(\lambda)$ and $\mu_a(\lambda)–\mu_{a\text{-}lip}(\lambda)$, where the hemoglobin ratios were evaluated.

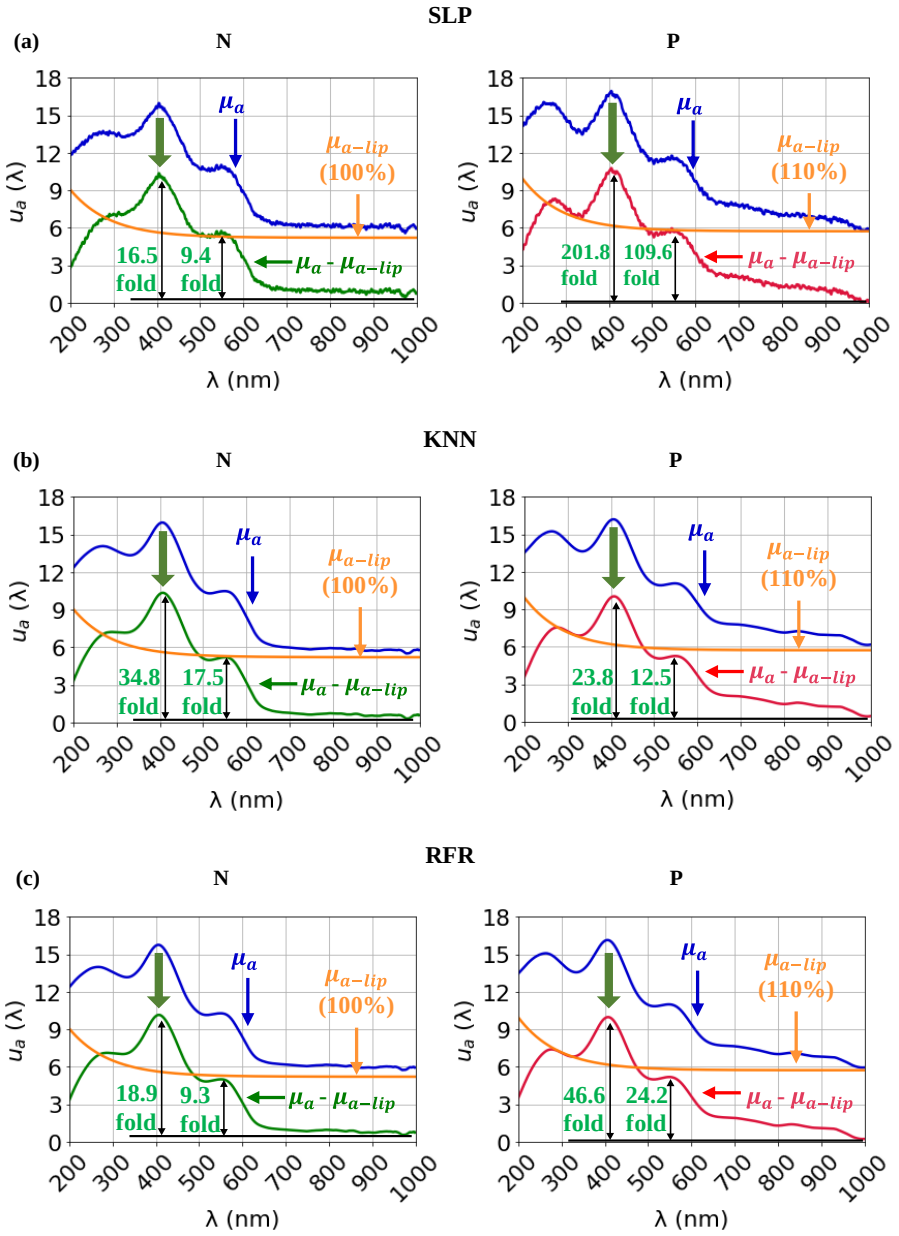


Figure 29. Wavelength dependencies of μ_a for lipofuscin (orange), for healthy (N) and pathological (P) mucosa, before (blue) and after (green or red) subtracting the absorption of lipofuscin. Results obtained with the SLP (a), KNN (b), and RFR (c) algorithms.

The curve for $\mu_{a\text{-}lip}(\lambda)$ that is presented in graphs of Fig. 29 was calculated with Eq. (12), where *A* was set to 1 (or to 1.1) for the normal (or pathological) mucosa. Comparing the ratios in the graphs of Fig. 29 with the ones previously reported in Ref. [18], we see that the SLP model originated lower values in the normal than in the pathological mucosa. In this case, although the ratios for the normal mucosa are in good agreement to the ones presented in Ref. [18], the ones obtained for the pathological mucosa are too high. We remind that the ratios that were reported in Ref. [18] were: 19.7-

fold (at 410 nm) and 10.1 (at 550 nm) for the normal mucosa and 33.1-fold (at 410 nm) and 17.3-fold (at 550 nm) for the pathological mucosa. The excessive higher ratios obtained in our study for the pathological mucosa were caused by a minimum value in the μ_a spectrum that is lower than expected, which consequently increases the ratios produced by our estimations in the pathological mucosa.

For the case of the KNN algorithm, something not expected has occurred, since the ratios in the normal mucosa are higher than in the pathological mucosa. Such erroneous results have occurred due to the fact that the minimum value in the μ_a spectrum that was estimated for the normal mucosa is lower than expected.

The RFR algorithm was the algorithm that produced the best results for the hemoglobin ratios. It has maintained the lower ratios in the normal mucosa and the generated values are very approximated to the ones reported in Ref. [18]. The exception to this fact is that the ratios obtained with our estimations for the pathological mucosa are somehow higher than the ones presented in Ref.[18], but such discrepancy can be justified. The spectral measurements that originated the reference μ_a spectra were obtained from a set of tissue samples, while the spectral R_d measurements were acquired from a different set of samples. The later samples could have bigger blood content than the previous, leading to higher magnitude ratios in the estimations with the RFR algorithm.

4.3.Scattering coefficient estimation

In a similar way to the estimation of the μ_a spectrum, we tried using ML models to estimate the μ_s spectra from the R_d measurements. The first step in these estimations was to calculate the reference μ_s spectrum from the T_s and μ_a measurements using Eq. (13). After calculating 10 normal and 10 pathological μ_s spectra, we noticed a small increase in the 400 nm wavelength. To improve these measurements, Eq. (14) was used to fit the μ_s spectra and obtain the smooth μ_s spectra. During the fitting process, the values between the 350-450 nm wavelength, where the abnormal increase was located, were ignored.

These new spectra were averaged to produce the mean normal and mean pathological reference spectra that are represented in Fig. 30.

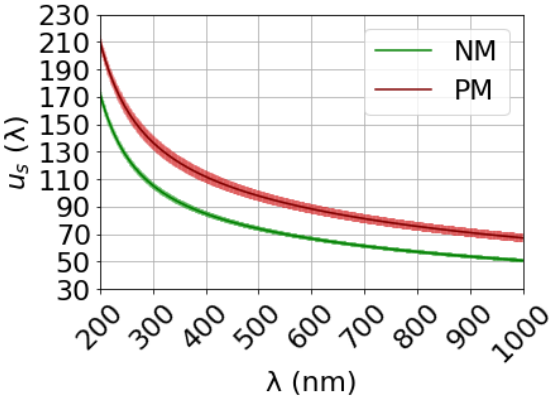


Figure 30. Mean μ_s spectra from the Normal Mucosa (NM) and from the Pathological Mucosa (PM).

4.3.1. Single Layer Perceptron

As in the estimation of the μ_a spectrum, the first experiment with ML models was started using the SLP. The estimations with this ML algorithm are presented in Fig. 31.

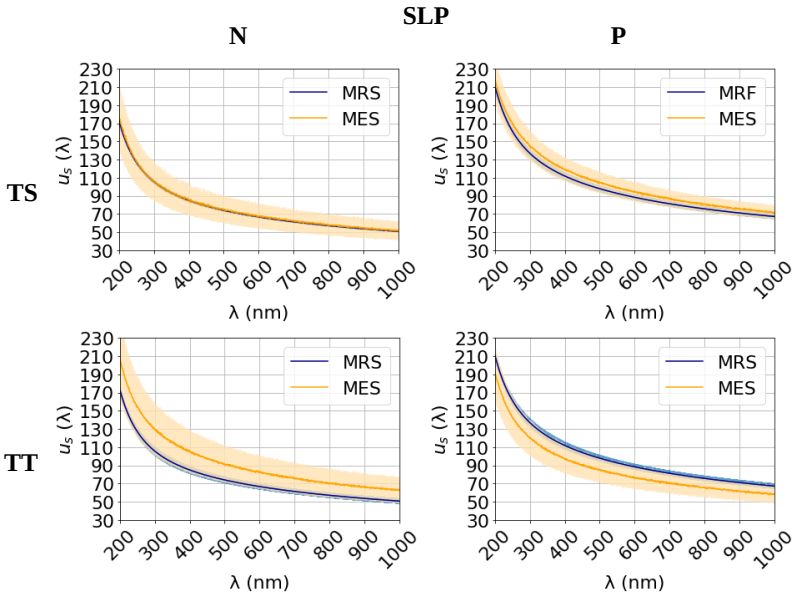


Figure 31. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the SLP algorithm. The individual estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S21-S24).

In a first analysis, the SLP models seem to have a good performance. However, during the training process, the training error tended to be higher than the validation error. This is a common sign of overfitting, but in the individual estimations that can be seen in the Annex 1, it is possible to see that the estimations made by the models change for

different input R_d spectra, which indicates that the model did not suffer from memory leakage.

4.3.1. K Nearest Neighbor

The KNN models had a good performance especially the TS models, as demonstrated by the data in Fig. 32. However, the SD of the estimated spectra from the TS models is lower than the one in the MRS. This could be because the KNN estimates the μ_s spectrum by interpolation using the "neighbors values", which consequently tends to produce estimations with higher bias.

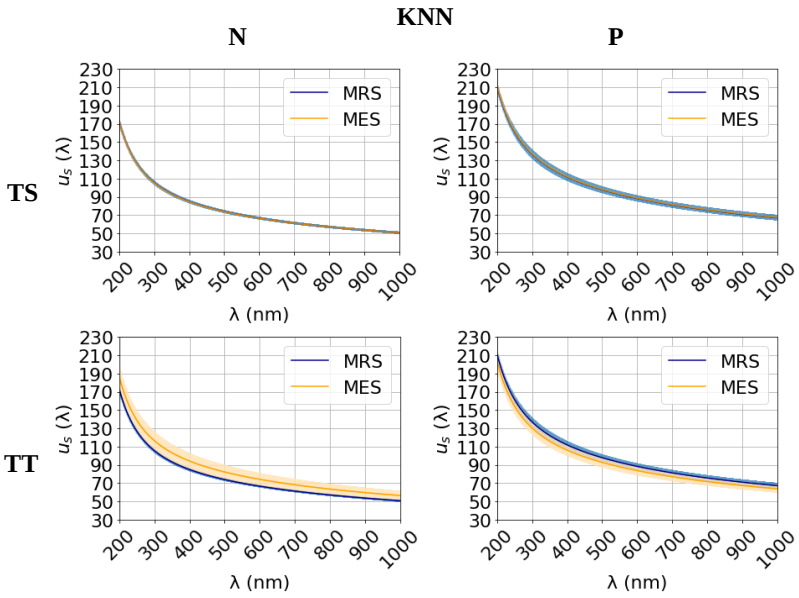


Figure 32. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the KNN algorithm. The individual estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S25-S28).

4.3.2. Random Forest Regression

In the estimations with the RFR models, seen in Fig. 33, the MES also tended to have a lower SD than the MRS. This could also be a consequence of the algorithm that the RFR uses to estimate each individual spectrum because the estimation given by the RFR algorithm is the mean of the estimations given by the trees that make up the random forest. Therefore, the estimation given by the RFR models, will be similar to the MRS because they are also an average of a random subset of samples that were used to train the trees.

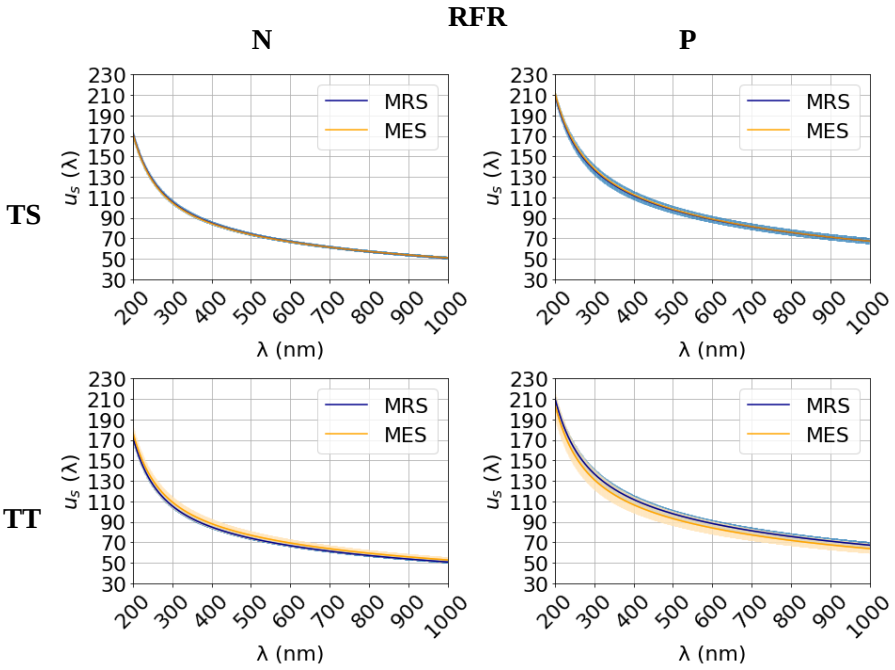


Figure 33. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the RFR algorithm. The individual estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S29-S32).

4.3.3. Decision Tree for Multioutput Regression

Similarly, to the ML models mentioned above, the DTFMO algorithm presented a good performance, except in the estimations from the TT models (see Fig. 34).

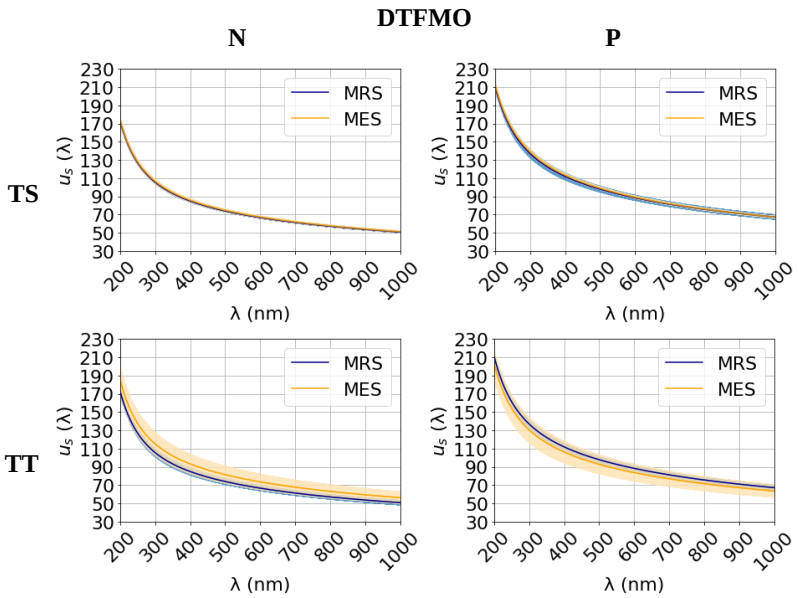


Figure 34. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the DTFMO algorithm. The individual

estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S33-S36).

This similar behaviour is expected because likewise to the RFR, the DTFMO uses a decision tree to make its predictions. However, a big difference between the models is that the DTFMO uses only one tree, while the RFR uses more than one.

4.3.4. Linear Regression for Multioutput

The LRFMO was the last algorithm that we tested and it presented the worst performance from all the studied ML algorithms. Figure 35 presents such results.

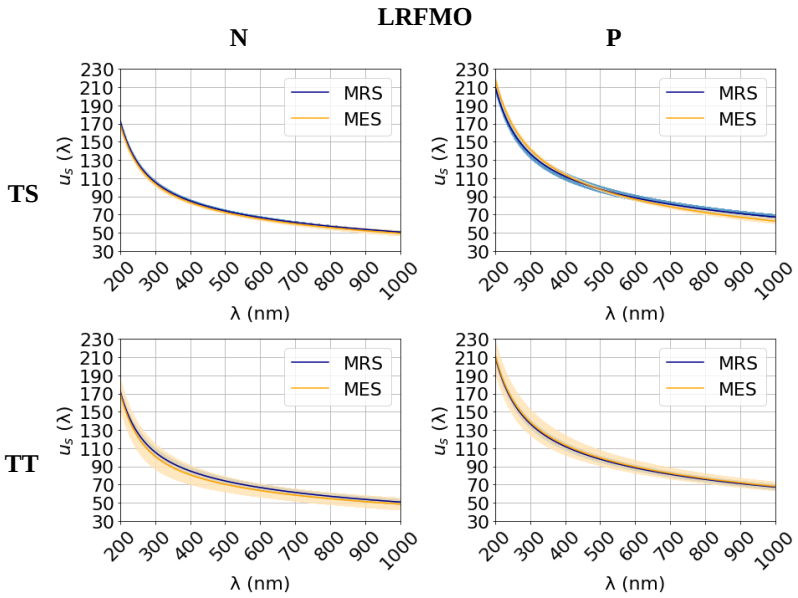


Figure 35. Comparison between the mean reference spectra (MRS) and the mean estimated spectra (MES) that result from the LRFMO algorithm. The individual estimations that were used to calculate the mean spectra can be seen in Annex 1 (Figure S37-S40).

To evaluate the performance of the ML models in a quantitative way, the ED between the estimated and the reference μ_s spectra were calculated for each algorithm. The mean data for the ED is presented in Fig. 36, where the TT models show a worse performance than the TS models in all ML algorithms. Considering all ML algorithms, the KNN and RFR seem to have the best performance.

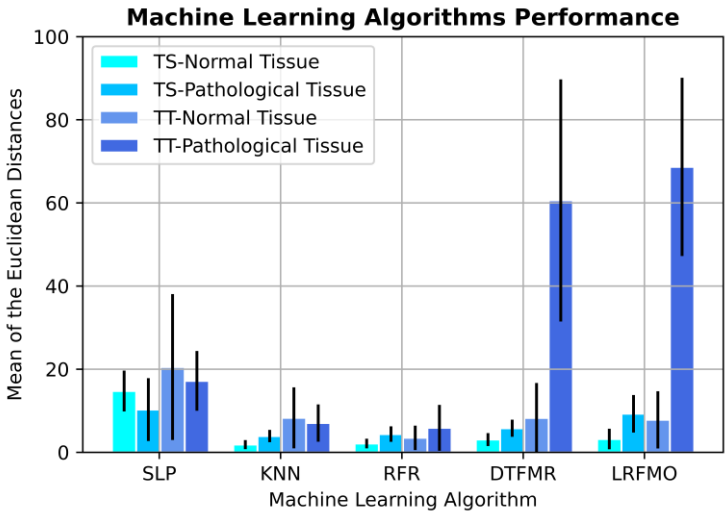


Figure 36. Average of the Euclidean distances for the different models when they are trained with data from separated samples of μ_s (TS) or with data from all samples of μ_s (TT).

For a better visualization of the performance between the KNN and the RFR algorithm, Fig. 37 presents a new graph with only these two models. The ED tells us that the TS models of both algorithms have a similar performance and that the KNN algorithm is slightly better than the RFR algorithm.

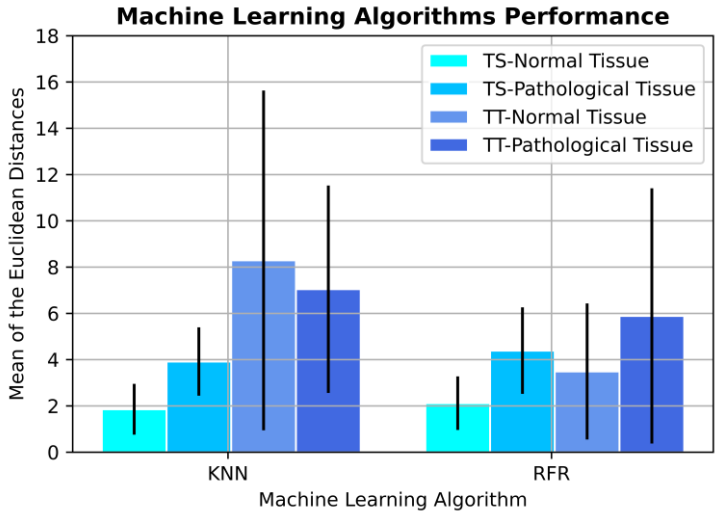


Figure 37. Average of the Euclidean distances for the different models when they are trained with data from separated samples of μs (TS) or with data from all samples of μs (TT).

4.4.Refractive index estimation

The reference RI of the samples was calculated using the μ_a spectra through the K-K relations (Eqs. (15) and (16)). After calculating the individual RI spectrum for each normal and pathological sample, the mean spectra that are presented in Fig. 38 were calculated.

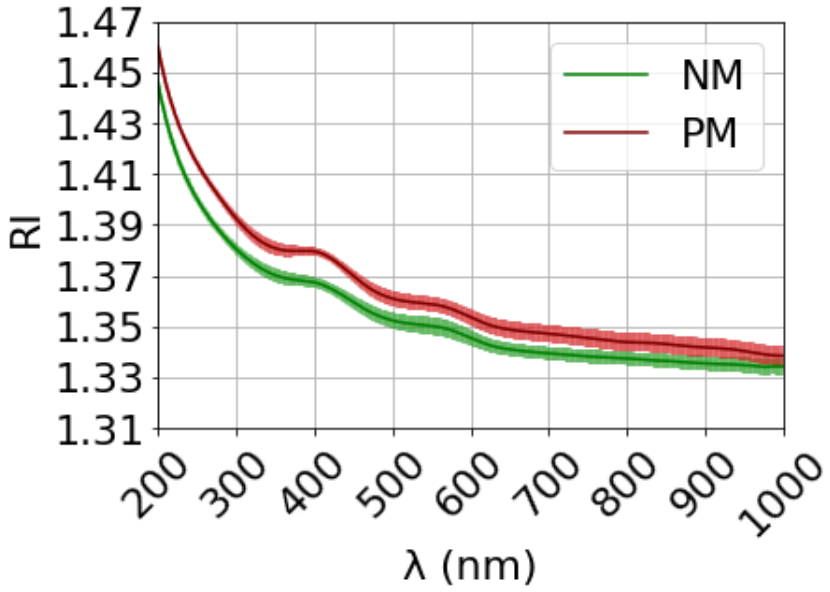


Figure 38. Mean RI spectra of the normal (NM-green) and pathological (PM-red) mucosa.

Since the K-K relations can be used to calculate RI from the μ_a spectra, there is no need to use ML models. However, we can calculate the RI spectra in the same manner, but using the μ_a spectra that were estimated with the ML algorithms. This means that hypothetically, instead of using μ_a spectra that were calculated from invasive techniques, it is possible to use μ_a spectra that were estimated with ML models, and therefore, avoiding the invasive measurement procedure to obtain the μ_a. The estimations of μ_a that were used to calculate the new RI spectra came from the RFR TS model, due to the fact that this was the model that presented the best performance in the μ_a estimations. Fig. 39 presents the mean estimated spectra for the normal and pathological samples.

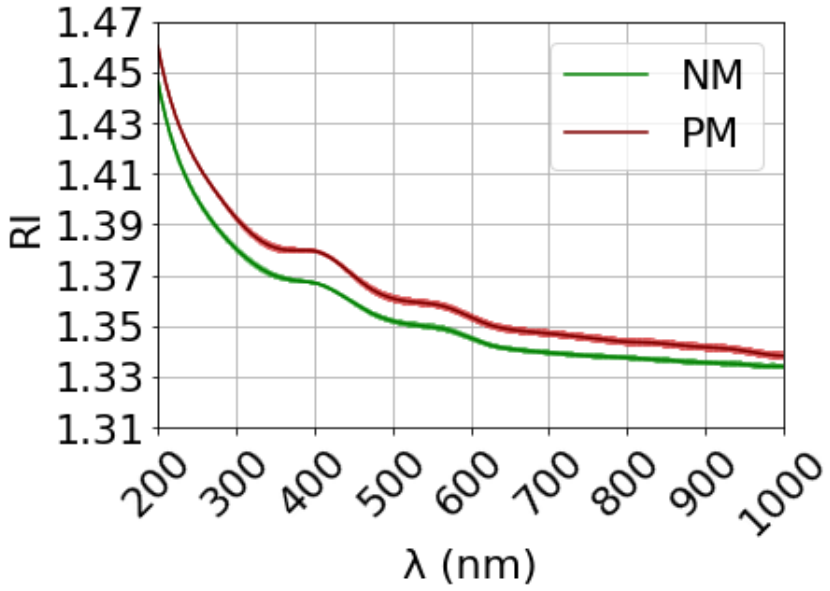


Figure 39. Mean RI estimated spectra of the normal (NM-green) and pathological (PM-red) mucosa.

The results in Fig. 39 show that by using the estimations from the RFR model it is possible to replicate the RI spectrum in a trustworthy way. Also, it shows how accurate are the μ_a spectrum estimations.

In summary, in this chapter different ML models were studied to estimate the optical properties of colorectal tissue and also the classification of R_d data in normal or pathological. During the estimation of the μ_a and μ_s, the TS approach yielded better results than the TT approach. In the estimation of μ_a, the RFR TS Model achieved the best performance with an ED bellow 2 and with spectral estimations that were trustworthy enough to replicate the calculations in Ref. [18] (see Fig. 29). Furthermore, the time consumed for each MES using the RFR TS model was below 1.5 seconds, which means this method is relatively fast when compared to other approaches such as the use of MC simulations. During the estimation of μ_s, the KNN and the RFR were the ML models that performed the best estimations with a similar ED of 4, with the KNN models being slightly better. The RI of the tissues were calculated using the K-K relations and the μ_a estimations from the RFR TS model. Instead of using the experimental μ_a measurements, we were able to replicate the calculations of the RI spectra with non-invasive ML estimations, as represented in Fig. 33.

In the classification task, the R_d data was analysed and different spectra domains were selected to achieve the best performance of the SVM model. When the wavelength domain was restrained to the 937-945 nm, the SVM model achieved an accuracy of 95 %. However, eight wavelengths represent a low number of features and therefore the domain of wavelengths used as input in the SVM model was increase to 700-1000 nm. The retrained SVM model achieved an accuracy of 90 %. Even though some accuracy was lost, the increase in features makes the SVM model more robust for future applications.

CHAPTER 5 – CONCLUSION

5. Conclusions and Future works

5.1. Conclusion

Biophotonics methods can provide a big help in the fight against cancer, since it allows the possibility of non-invasive approaches to retrieve diagnostic data. Considering human colorectal mucosa tissues, both in healthy and pathological versions, several spectral measurements were made from surgical samples to develop the present work with the objective of testing ML methods to diagnose cancer. Using invasive measurements from excised tissue samples, the reference optical properties of both tissues were calculated. Three of these properties, namely the absorption coefficient, the scattering coefficient and the refractive index are of major interest, since they are known to contain information that can be used to discriminate between normal and pathological tissues. Due to ethical considerations, diffuse reflectance measurements were acquired also from a new set of freshly excised tissue samples, but with appropriate hydration conditions to mimic the in vivo situation. Such data were submitted as input to ML studies to evaluate the capability of these methods to reconstruct the reference spectra of the optical properties of the tissues and retrieve diagnostic information. An additional objective of the present study was to investigate the possibility of creating a fast and non-invasive procedure for colorectal cancer diagnosis, which will be less aggressive for the patient and can provide data with higher accuracy for the clinician to establish a reliable diagnosis.

By testing different ML models, it was possible to select the RFR TS model as the one that could perform an optimized recreation of the reference absorption coefficient of the colorectal mucosa tissues with an ED bellow 2. This same model also presented the second highest performance in the recreation of the reference scattering coefficient spectrum for both tissues. The DTFMO and the LRFMO models were not reliable to perform the recreation of the reference spectra of the absorption and scattering coefficients due to the fact that their mean estimated spectra tended to be most distant from the mean reference spectra, especially when the TT approach is considered. The SLP algorithm presented a good performance in the reconstruction of the absorption coefficient spectrum, but when recreating the scattering coefficient spectrum, its performance was not so good. Considering the reconstruction of the absorption coefficient spectra, the results obtained with the models that had higher performance

allowed to discriminate cancer through the detection of discriminated lipofuscin content. The calculations to evaluate such pigment content were the same that have previously allowed such discrimination from invasive measurements. Finally, the work developed in this study resulted in an article that was published in the CHAOS magazine [9].

5.2. Future works

In a further attempt to obtain diagnostic information directly from the diffuse reflectance measurements, such spectra were used to train an SVM model. An accuracy of 90% was obtained in this procedure, a value which is similar to the one obtained in other studies that use different diagnostic approaches. The results obtained in the present study show that ML procedure can be applied to non-invasive spectroscopy methods to develop new non-invasive diagnostic procedures to be used in cancer screening exams, even in their early stage of development.

As a future perspective in this line of research, other studies to obtain diagnostic information from tissues can be performed, provided that the spectral reflectance of tissues becomes available. One case of interest is to evaluate the pigment content in colorectal cancer in different stages of development, a study that can also be made for other types of cancer. Exploratory studies to perform the diagnosis of other pathologies can also be made with the combination of reflectance spectra and ML techniques. One example of interest is to evaluate degeneration of brain cells due to the accumulation of melanin and lipofuscin in the various stages of development of the Alzheimer's or Parkinson's diseases. Another case that can be studied is the evaluation of glycaemia in tissue blood, as a marker of diabetes disease.

References:

[1] "Cancer Today." https://gco.iarc.fr/today/online-analysis-pie?v=2020&mode=cancer&mode_population=continents&population=900&populations=900&key=total&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=5&group_cancer=1&include_nmsc=1&include_nmsc_other=1&half_pie=0&donut=0 (accessed Mar. 13, 2021).

[2] "Cancer." https://www.who.int/health-topics/cancer#tab=tab_1 (accessed Mar. 13, 2021).

[3] S. L. Jacques, "Optical properties of biological tissues: A review," *Phys. Med. Biol.*, vol. 58, no. 11, 2013, doi: 10.1088/0031-9155/58/11/R37.

[4] E. Salomatina, B. Jiang, J. Novak, and A. N. Yaroslavsky, "Optical properties of normal and cancerous human skin in the visible and near-infrared spectral range," *J. Biomed. Opt.*, vol. 11, no. 6, p. 064026, 2006, doi: 10.1117/1.2398928.

[5] V. Tuchin, "Tissue Optics Light Scattering Methods and Instruments for Medial Diagnosis," *SPIE*, vol. 13, 2000, doi: 10.1117/3.684093.

[6] L. M. Couto Oliveira and V. V. Tuchin, *The Optical Clearing Method. A New Tool for Clinical Practice and Biomedical Engineering*. 2019.

[7] 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 *Biophotonics—Riga 2020*, 2020, vol. 11585, pp. 12–26, doi: 10.1117/12.2584999.

[8] I. Carneiro, S. Carvalho, R. Henrique, L. Oliveira, and V. Tuchin, "Moving tissue spectral window to the deep-ultraviolet via optical clearing.," *Journal of biophotonics*, vol. 12, no. 12. Germany, p. e201900181, Dec-2019, doi: 10.1002/jbio.201900181.

[9] L. Fernandes, S. Carvalho, I. Carneiro, R. Henrique, V. V Tuchin, and P. Hélder, "Diffuse reflectance and machine learning techniques to differentiate colorectal cancer," vol. 053118, no. March, 2021, doi: 10.1063/5.0052088.

[10] 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.*, vol.

32, no. 4, pp. 559–568, 1993, doi: 10.1364/AO.32.000559.

[11] 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. Quantum Electron.*, vol. 27, no. 4, pp. 1–8, 2021, doi: 10.1109/JSTQE.2020.3012350.

[12] N. M. Gomes, V. V Tuchin, and L. M. Oliveira, "Refractive Index Matching Efficiency in Colorectal Mucosa Treated With Glycerol," *IEEE J. Sel. Top. Quantum Electron.*, vol. 27, no. 4, pp. 1–8, 2021, doi: 10.1109/JSTQE.2021.3050208.

[13] G. Zonios and A. Dimou, "Modeling diffuse reflectance from semi-infinite turbid media: application to the study of skin optical properties," *Opt. Express*, vol. 14, no. 19, p. 8661, 2006, doi: 10.1364/oe.14.008661.

[14] L. Guyon, A. da Silva, A. Planat-Chrétien, P. Rizo, and J.-M. Dinten, "X^2 Analysis for Estimating the Accuracy of Optical Properties Derived From Time Resolved Diffuse-Reflectance," *Opt. Express*, vol. 17, no. 22, p. 20521, 2009, doi: 10.1364/oe.17.020521.

[15] A. N. Bashkatov *et al.*, "Optical properties of human stomach mucosa in the spectral range from 400 to 2000nm: Prognosis for gastroenterology," *Med. Laser Appl.*, vol. 22, no. 2, pp. 95–104, 2007, doi: https://doi.org/10.1016/j.mla.2007.07.003.

[16] A. N. Bashkatov, E. A. Genina, and V. V. Tuchin, "Optical properties of skin, subcutaneous, and muscle tissues: A review," *J. Innov. Opt. Health Sci.*, vol. 4, no. 1, pp. 9–38, 2011, doi: 10.1142/S1793545811001319.

[17] I. Carneiro, S. Carvalho, R. Henrique, and V. V Tuchin, "Optical properties of colorectal muscle in visible/NIR range," *Biophotonics Photonic Solut. Better Heal. Care VI*, no. May, 2018, doi: 10.1117/12.2306586.

[18] S. Carvalho, I. Carneiro, R. Henrique, V. Tuchin, and L. Oliveira, "Lipofuscin-type pigment as a marker of colorectal cancer," *Electron.*, vol. 9, no. 11, pp. 1–14, 2020, doi: 10.3390/electronics9111805.

[19] 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.*, vol. 57, no. 12, p. 3038, 2018, doi: 10.1364/ao.57.003038.

[20] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nat. Med.*, vol. 25, no. 1, pp. 30–36, 2019, doi: 10.1038/s41591-018-0307-0.

[21] T. J. Farrell, 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.*, vol. 37, no. 12, pp. 2281–2286, 1992, doi: 10.1088/0031-9155/37/12/009.

[22] T. J. Farrell, M. S. Patterson, and B. Wilson, "A diffusion theory model of spatially resolved, steady-state diffuse reflectance for the noninvasive determination of tissue optical properties in vivo," *Med. Phys.*, vol. 19, no. 4, pp. 879–888, 1992, doi: https://doi.org/10.1118/1.596777.

[23] A. Kienle, L. Lilge, M. S. Patterson, R. Hibst, R. Steiner, and B. C. Wilson, "Spatially resolved absolute diffuse reflectance measurements for noninvasive determination of the optical scattering and absorption coefficients of biological tissue," *Appl. Opt.*, vol. 35, no. 13, p. 2304, 1996, doi: 10.1364/ao.35.002304.

[24] L. Zhang, Z. Wang, and M. Zhou, "Determination of the optical coefficients of biological tissue by neural network," *J. Mod. Opt.*, vol. 57, no. 13, pp. 1163–1170, 2010, doi: 10.1080/09500340.2010.500106.

[25] R. Watté *et al.*, "Metamodeling approach for efficient estimation of optical properties of turbid media from spatially resolved diffuse reflectance measurements," *Opt. Express*, vol. 21, no. 26, p. 32630, 2013, doi: 10.1364/oe.21.032630.

[26] S. Panigrahi and S. Gioux, "Machine learning approach for rapid and accurate estimation of optical properties using spatial frequency domain imaging," *J. Biomed. Opt.*, vol. 24, no. 07, p. 1, 2018, doi: 10.1117/1.jbo.24.7.071606.

[27] T. JUNG, N. BADER, and T. GRUNE, "Lipofuscin," *Ann. N. Y. Acad. Sci.*, vol. 1119, no. 1, pp. 97–111, 2007, doi: https://doi.org/10.1196/annals.1404.008.

[28] J. J. Hunter, J. I. W. Morgan, W. H. Merigan, D. H. Sliney, J. R. Sparrow, and D. R. Williams, "The susceptibility of the retina to photochemical damage from

visible light," *Prog. Retin. Eye Res.*, vol. 31, no. 1, pp. 28–42, 2012, doi: https://doi.org/10.1016/j.preteyeres.2011.11.001.

[29] M. E. Gosnell, A. G. Anwer, J. C. Cassano, C. M. Sue, and E. M. Goldys, "Functional hyperspectral imaging captures subtle details of cell metabolism in olfactory neurosphere cells, disease-specific models of neurodegenerative disorders," *Biochim. Biophys. Acta - Mol. Cell Res.*, vol. 1863, no. 1, pp. 56–63, 2016, doi: https://doi.org/10.1016/j.bbamcr.2015.09.030.

[30] I. Martins, H. Silva, V. V Tuchin, and L. Oliveira, "Estimation of Rabbit Pancreas Dispersion Between 400 and 1000 nm," vol. 7, no. May, pp. 1–10, 2021, doi: 10.18287/JBPE21.07.020303.

[31] O. Sydoruk, O. Zhernovaya, V. Tuchin, and A. Douplik, "Refractive index of solutions of human hemoglobin from the near-infrared to the ultraviolet range: Kramers-Kronig analysis," *J. Biomed. Opt.*, vol. 17, no. 11, p. 115002, 2012, doi: 10.1117/1.jbo.17.11.115002.

[32] J. Gienger, H. G. J. Neukammer, and M. Bär, "Determining the refractive index of human hemoglobin solutions by Kramers--Kronig relations with an improved absorption model," *Appl. Opt.*, vol. 55, no. 31, pp. 8951–8961, Nov. 2016, doi: 10.1364/AO.55.008951.

[33] E. J. M. Baltussen *et al.*, "Diffuse reflectance spectroscopy as a tool for real-time tissue assessment during colorectal cancer surgery," vol. 22, no. 10, 2021, doi: 10.1117/1.JBO.22.10.106014.

[34] S. Terenborg, T. H. E. O. J. M. R. Uers, and B. Ehdad, "Optimizing algorithm development for tissue classification in colorectal cancer based on diffuse reflectance spectra," vol. 10, no. 12, pp. 6096–6113, 2019.

[35] M. S. Nogueira *et al.*, "Evaluation of wavelength ranges and tissue depth probed by diffuse reflectance spectroscopy for colorectal cancer detection," *Sci. Rep.*, pp. 1–17, 2021, doi: 10.1038/s41598-020-79517-2.

6. Annex 1

The following figures present the individual estimated spectra that originated the mean and standard deviation (SD), as obtained from each model and training method used. Figure S1 presents the individual estimated spectra that were obtained with the SLP model trained with data from healthy samples only.

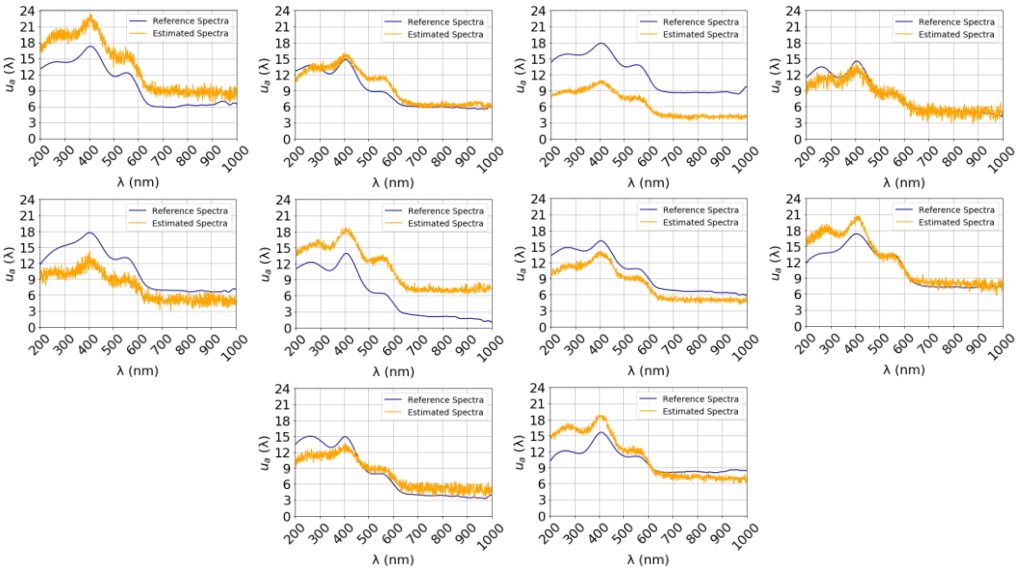


Figure S1. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the SLP model.

Figure S2 shows the estimated spectra from the SLP model trained with data from pathological samples only.

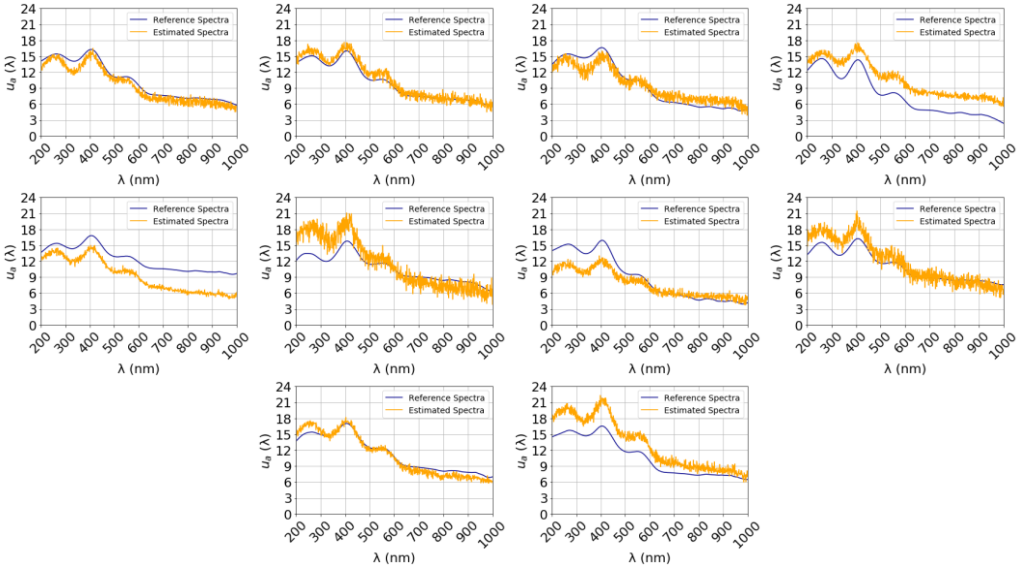


Figure S2. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the SLP model.

Figure S3 presents the estimated healthy spectra from the SLP model that was trained with data from normal and pathological samples.

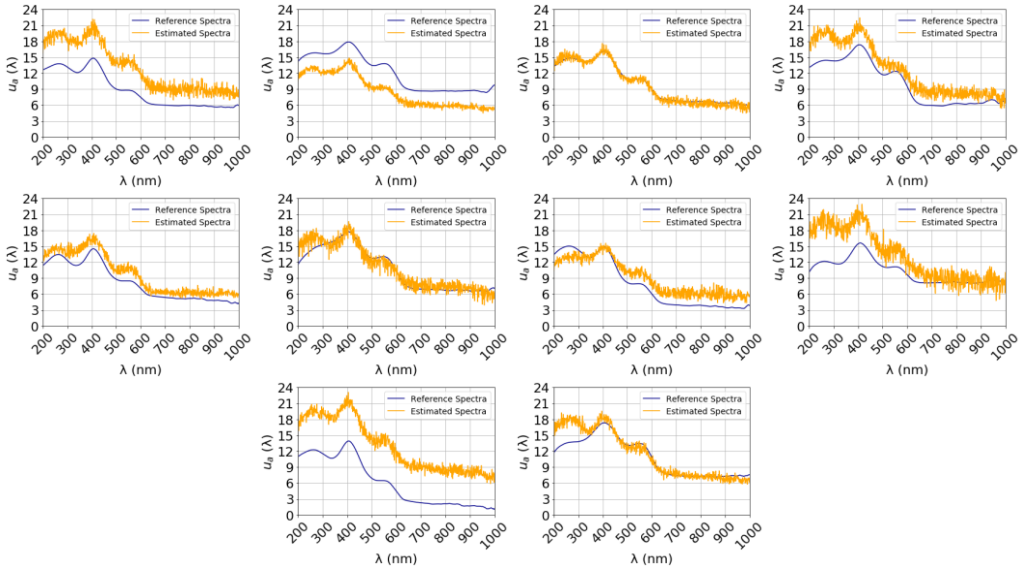


Figure S3. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the SLP model.

Figure S4 shows the estimated pathological spectra from the SLP model that was trained with data from the healthy and pathological samples.

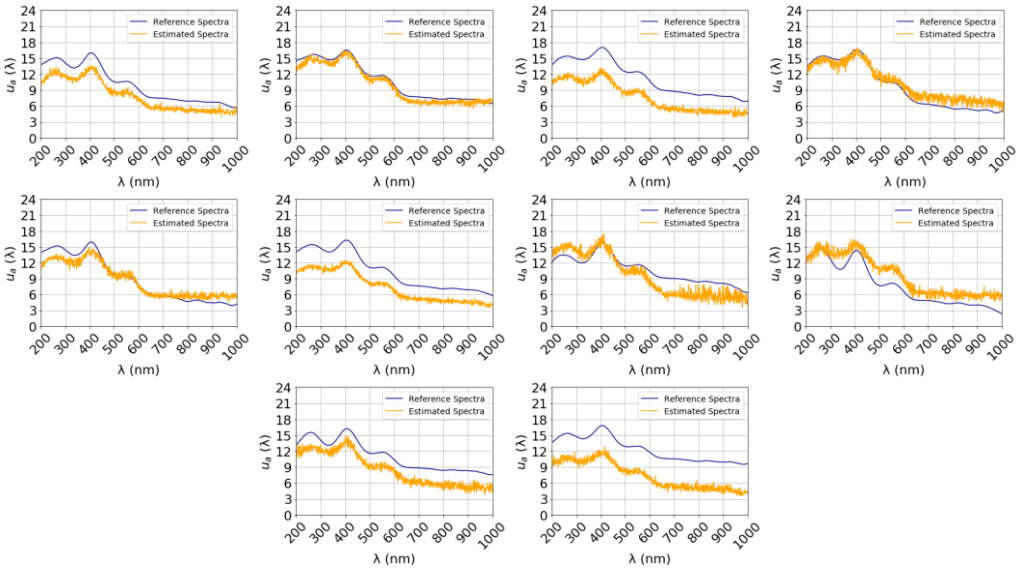


Figure S4. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the SLP model.

In Figure S5, it is possible to see the estimated spectrograms from the KNN model trained with only normal samples.

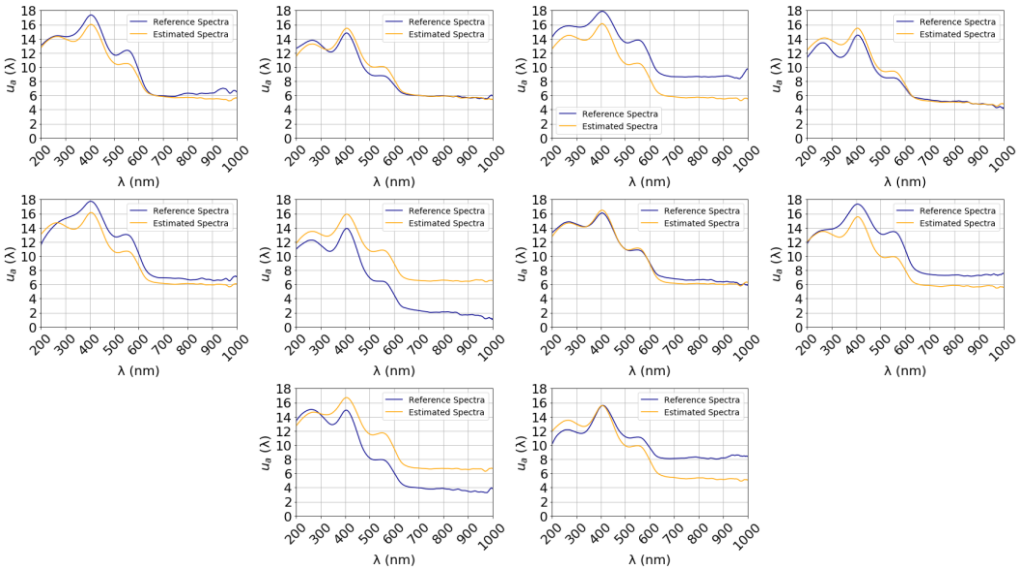


Figure S5. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the KNN model.

In Figure S6, it is possible to see the estimated spectrograms from the KNN model trained with only pathological samples.

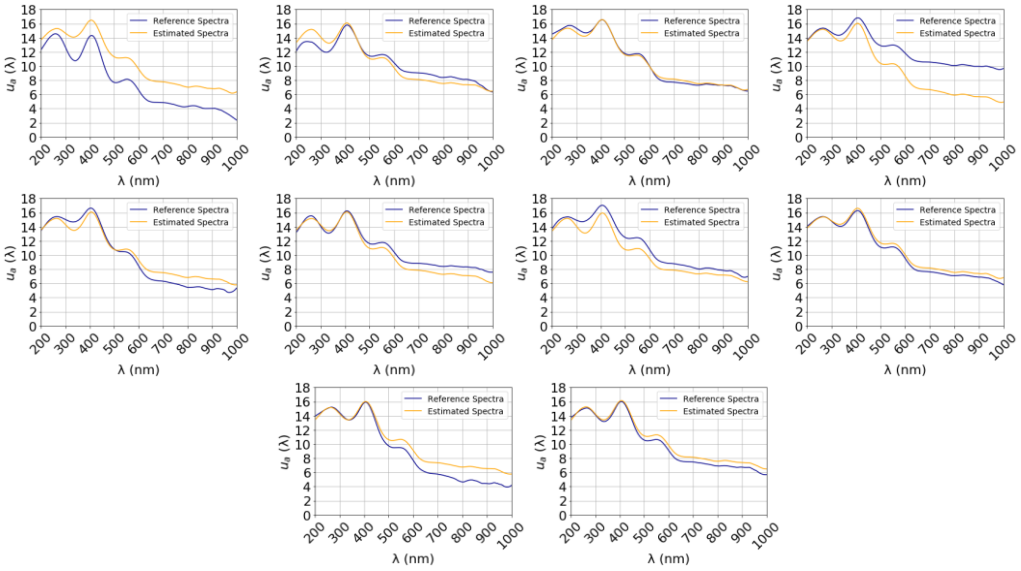


Figure S6. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the KNN model.

In Figure S7, it is possible to see the estimated normal spectrograms from the KNN model that was trained with the normal and pathological samples.

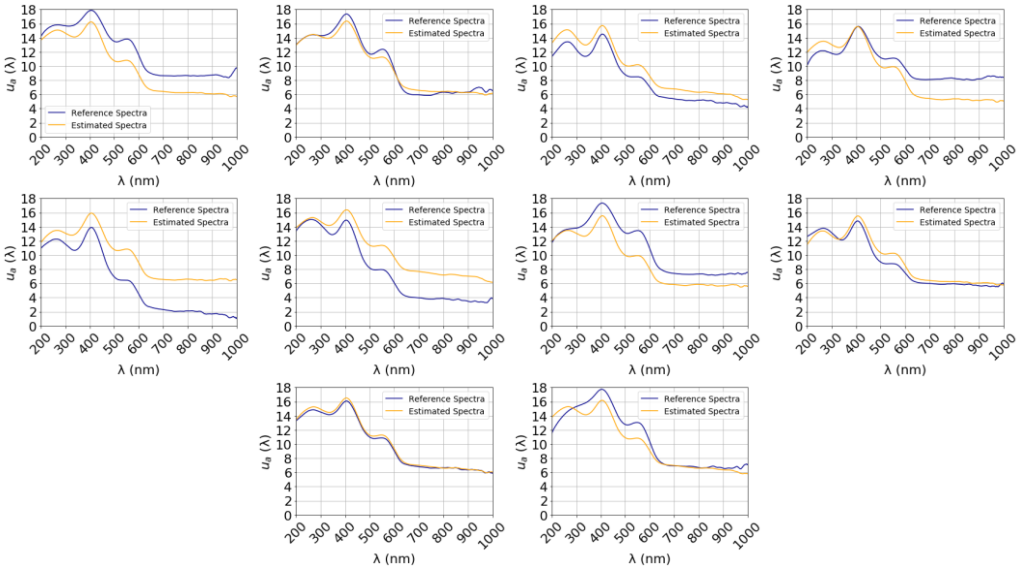


Figure S7. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the KNN model.

In Figure S8, it is possible to see the estimated pathological spectrograms from the KNN model that was trained with the normal and pathological samples.

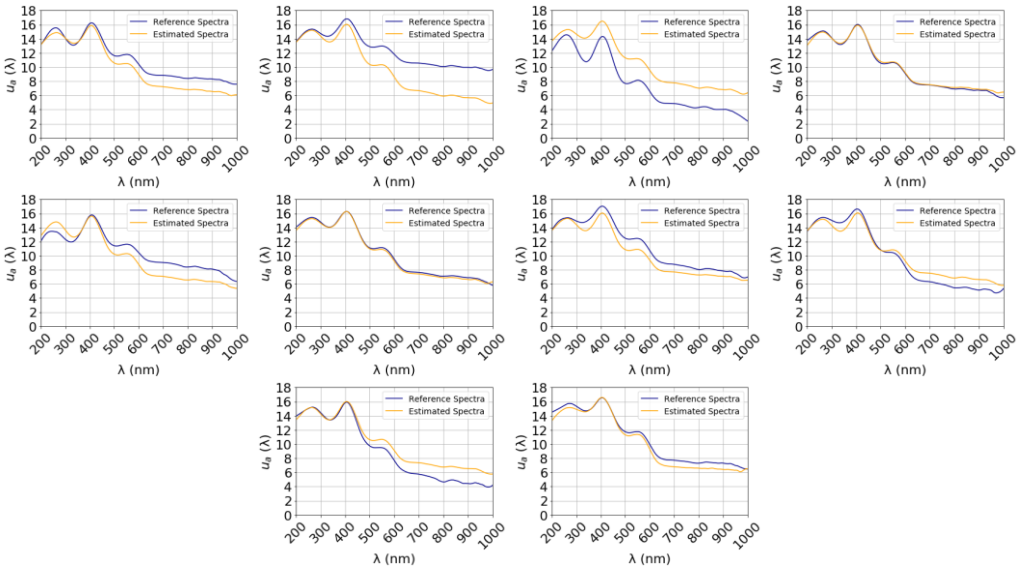


Figure S8. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the KNN model.

In Figure S9, it is possible to see the estimated spectrograms from the RFR model trained with only normal samples.

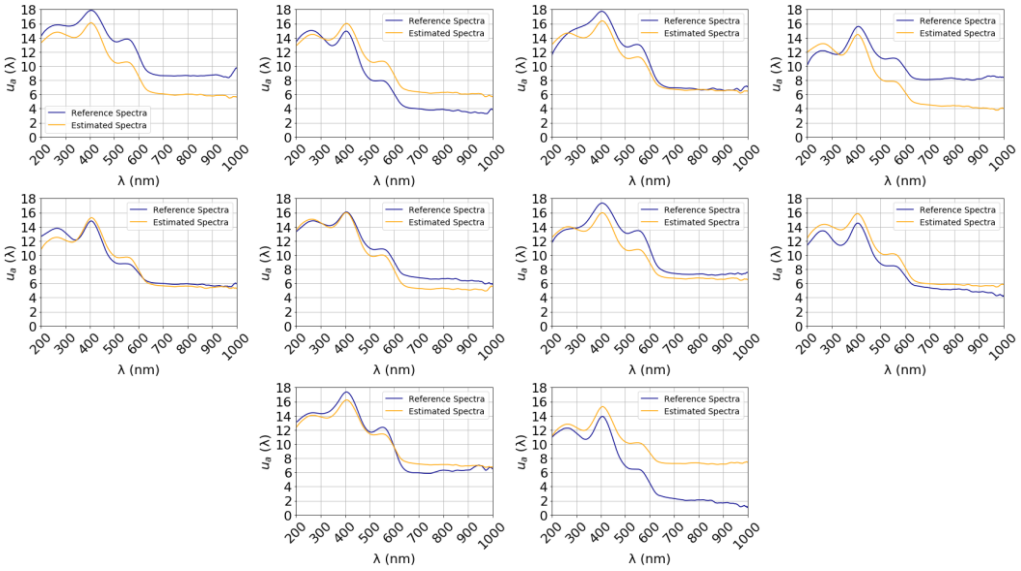


Figure S9. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the RFR model.

Figure S10 presents the estimated spectra from the RFR model that was trained with data from pathological samples only.

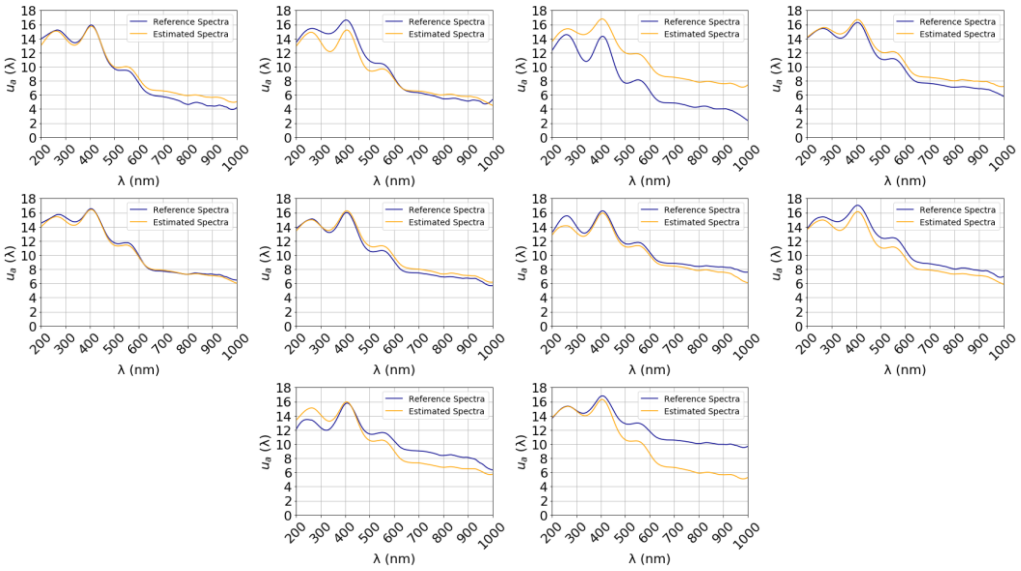


Figure S10. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the RFR model.

Figure S11 shows the estimated healthy spectra from the RFR model that was trained with data from the healthy and pathological samples.

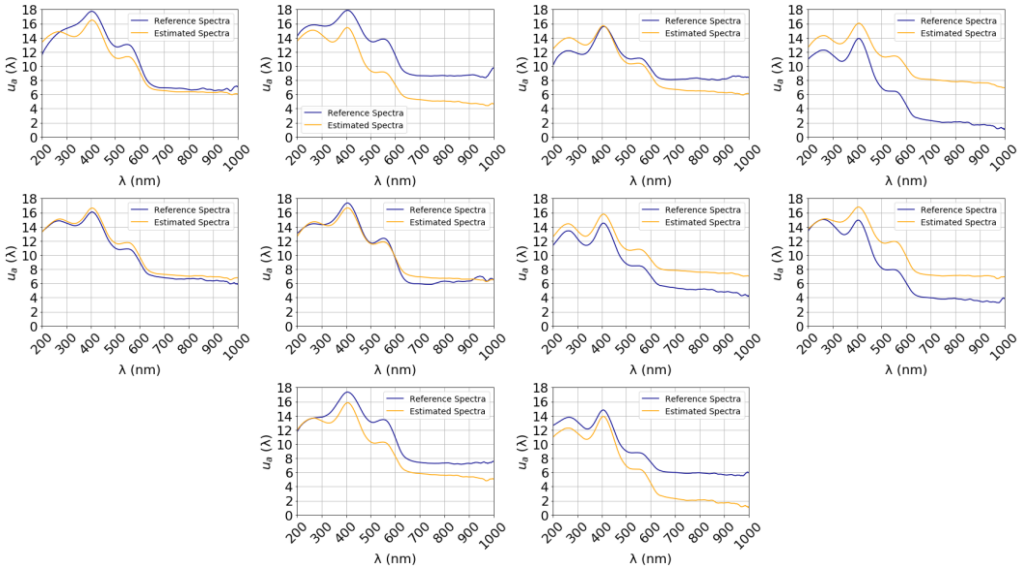


Figure S11. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the RFR model.

Figure S12 presents the estimated pathological spectra from the RFR model that was trained with data from the healthy and pathological samples.

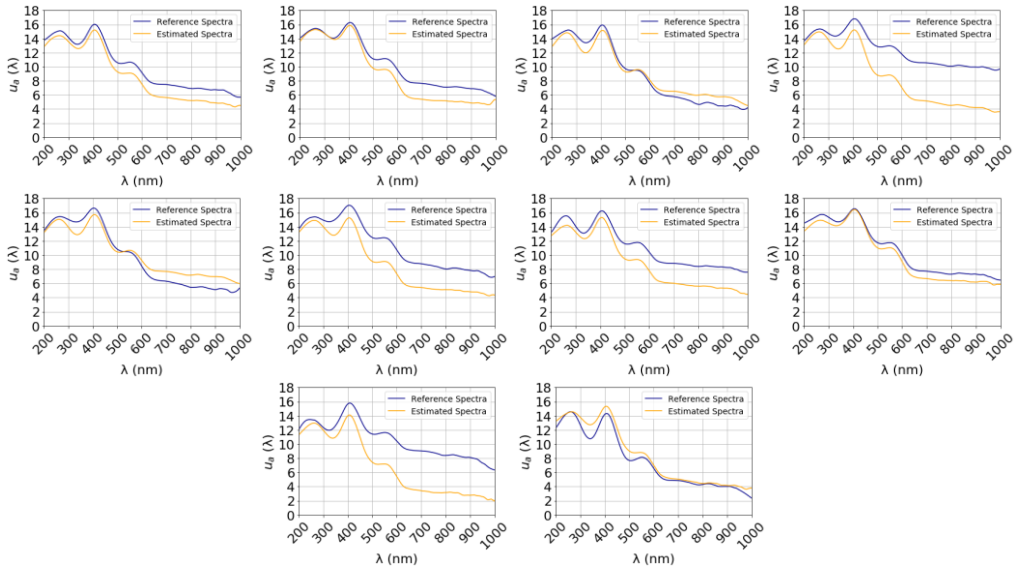


Figure S12. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the RFR model.

Figure S13 shows the estimated spectra from the DTFMR model that was trained with data from the healthy samples only.

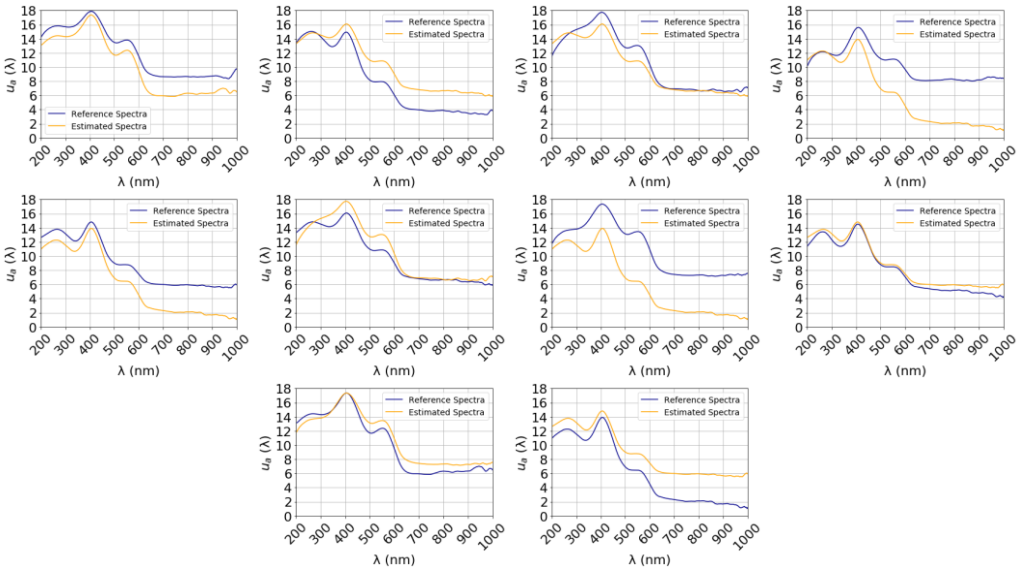


Figure S13. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the DTFMR model.

Figure S14 presents the estimated spectra from the DTFMR model that was trained with data from the pathological samples only.

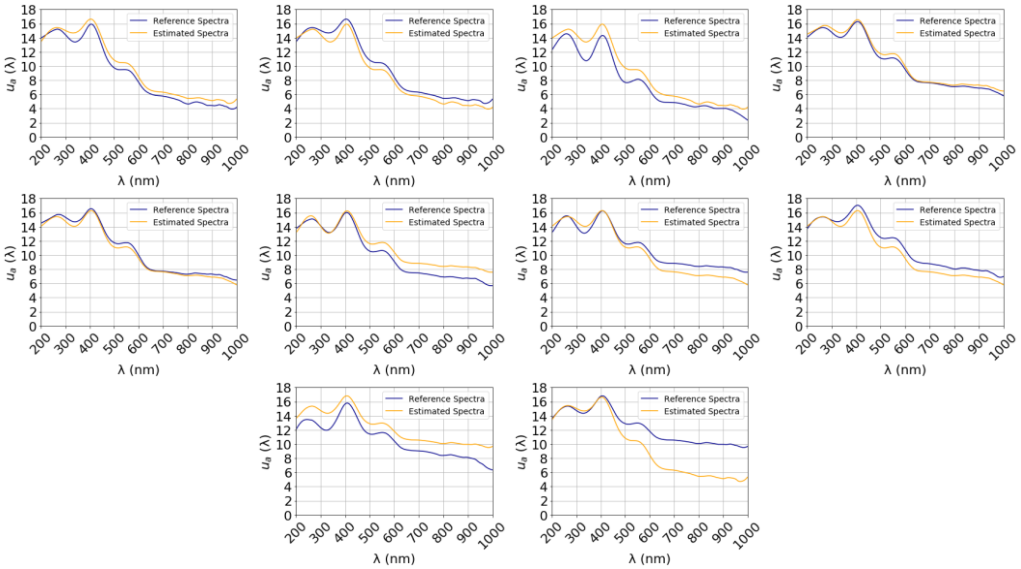


Figure S14. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the DTFMR model.

Figure S15 shows the estimated healthy spectra from the DTFMR model that was trained with data from the healthy and pathological samples.

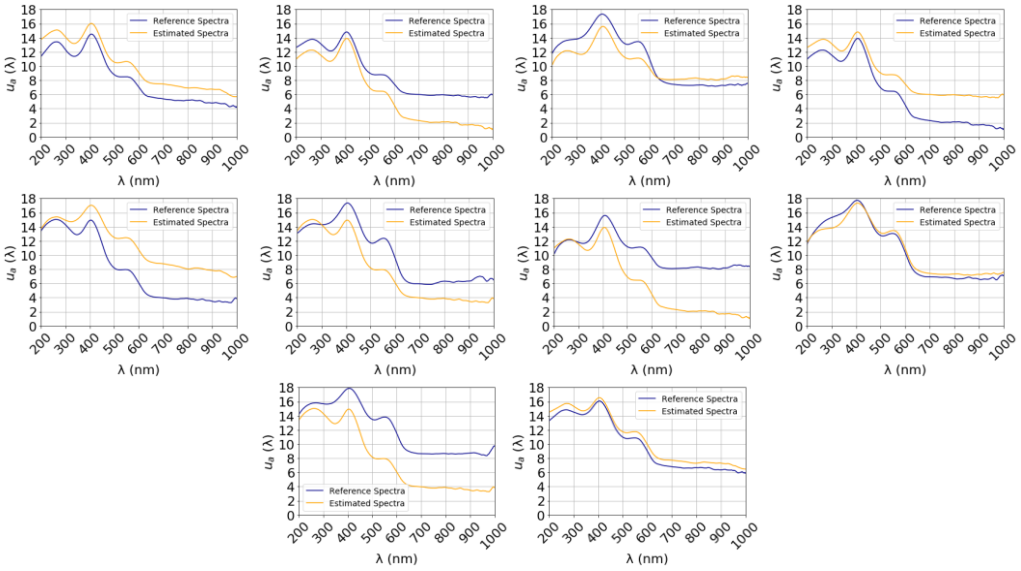


Figure S15. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the DTFMR model.

Figure S16 shows the estimated pathological spectra from the DTFMR model that was trained with data from the healthy and pathological samples.

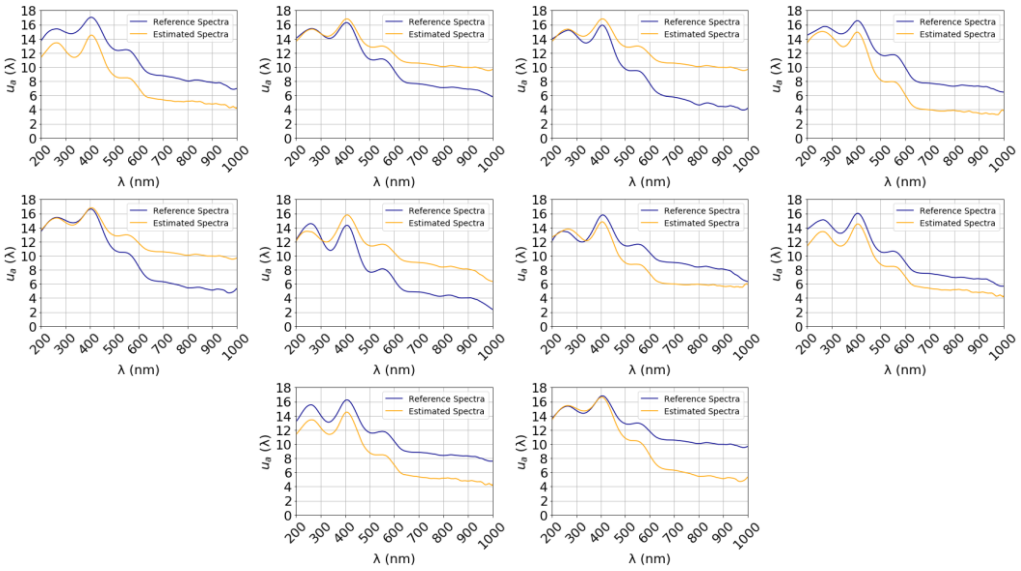


Figure S16. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the DTFMR model.

Figure S17 presents the estimated spectra from the LRFMO model that was trained with data from the healthy samples only.

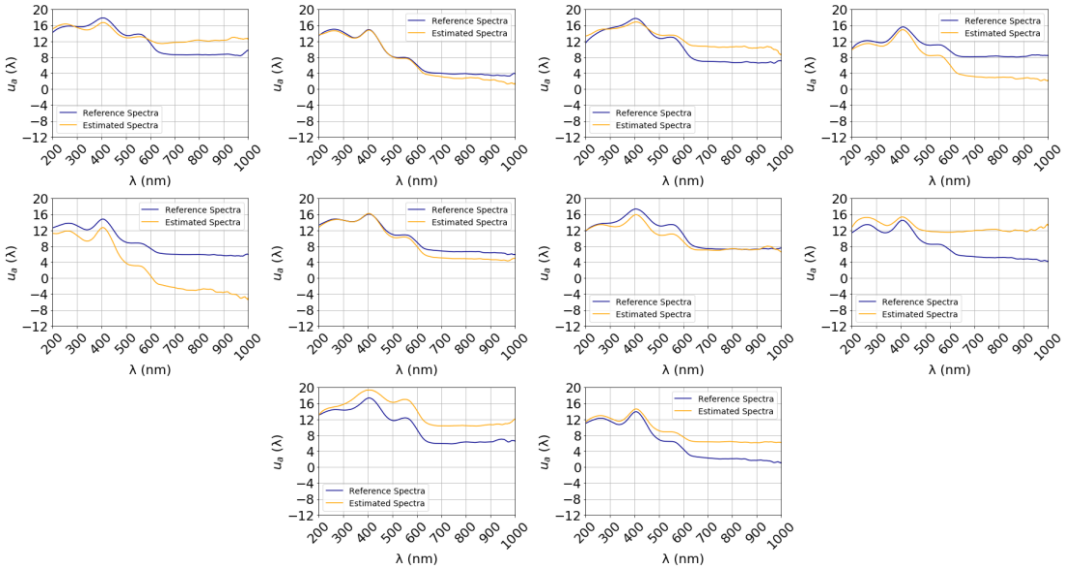


Figure S17. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the LRFMO model.

Figure S18 presents the estimated spectra from the LRFMO model trained with data from the pathological samples only.

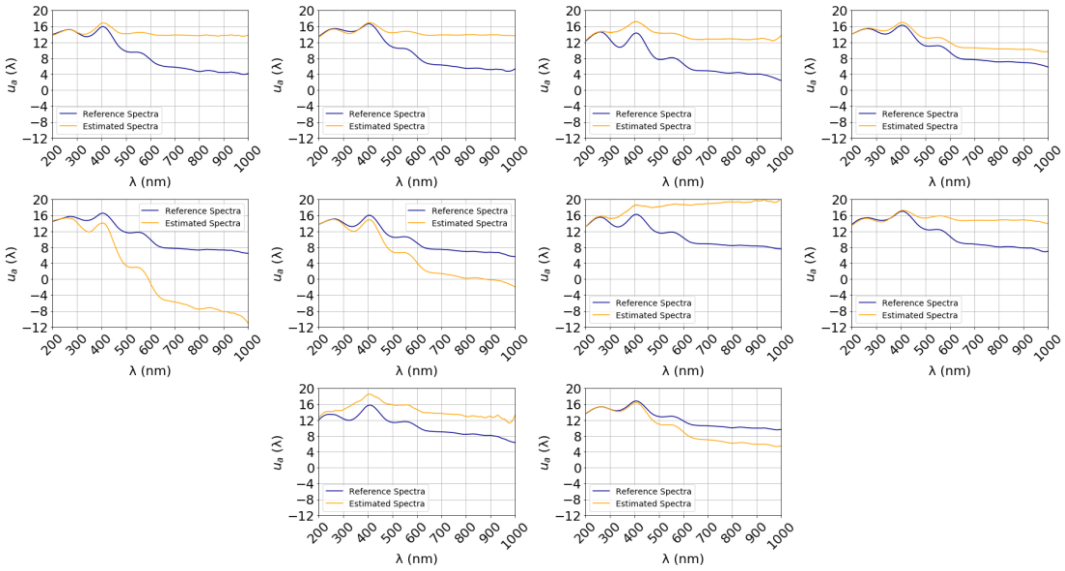


Figure S18. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the LRFMO model.

Figure S19 shows the estimated healthy spectra from the LRFMO model that was trained with data from the healthy and pathological samples.

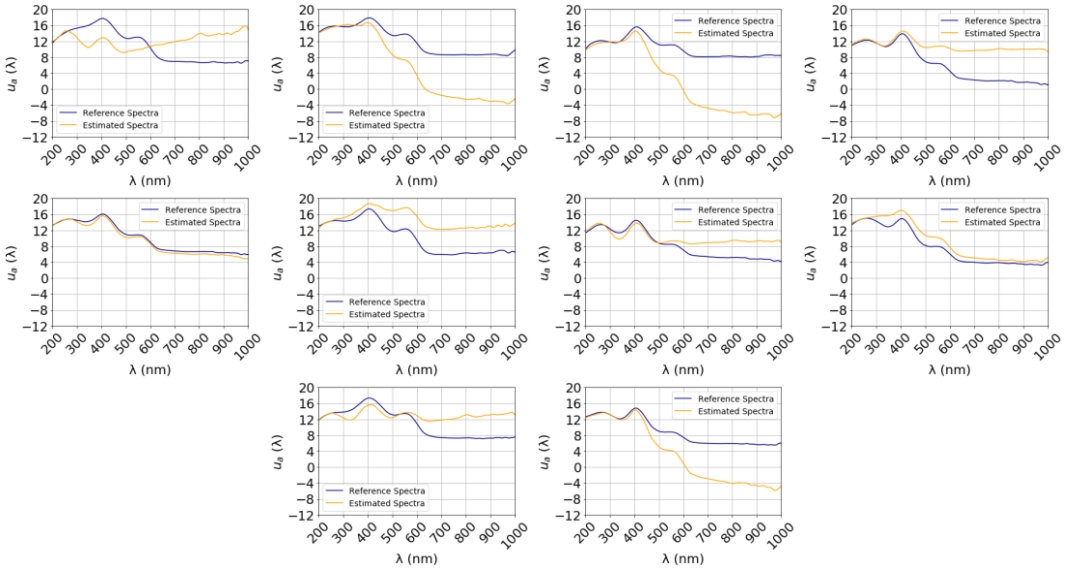


Figure S19. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the LRFMO model.

Figure S20 shows the estimated pathological spectra from the LRFMO model that was trained with data from the healthy and pathological samples.

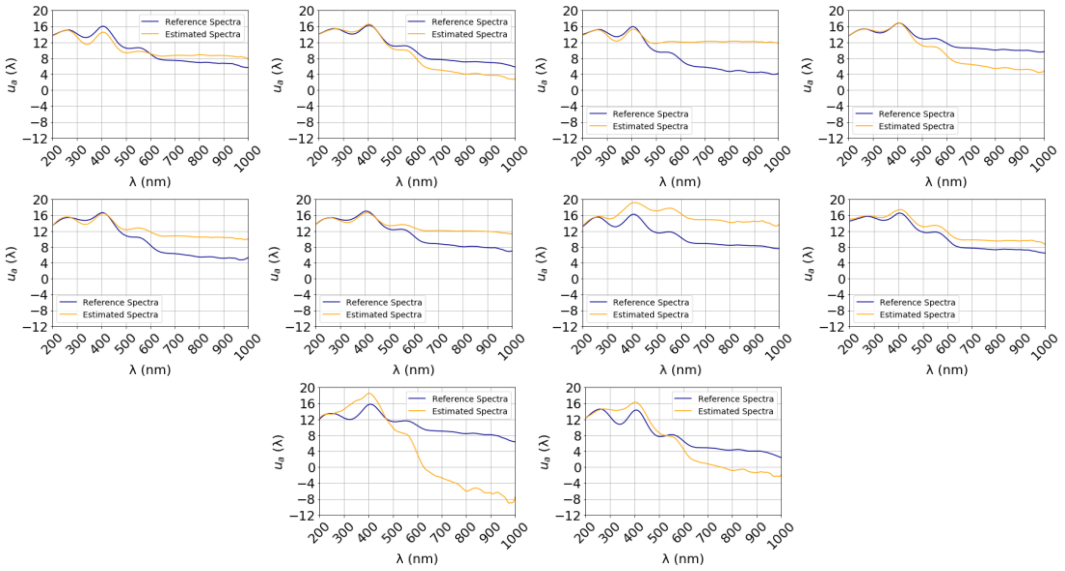


Figure S20. Estimated μ_a spectra for each individual sample (orange), compared to the reference μ_a spectra (blue). The estimated μ_a spectra came from the LRFMO model.

Figure S21 presents the estimated spectra from the SLP model that was trained with data from the healthy samples only.

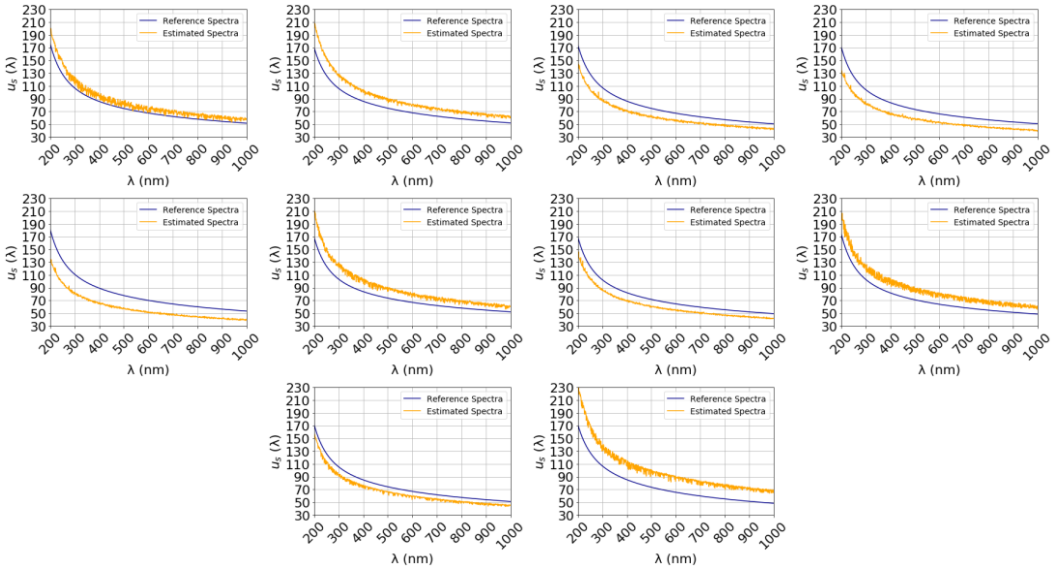


Figure S21. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the SLP model.

Figure S22 presents the estimated spectra from the SLP model trained with data from the pathological samples only.

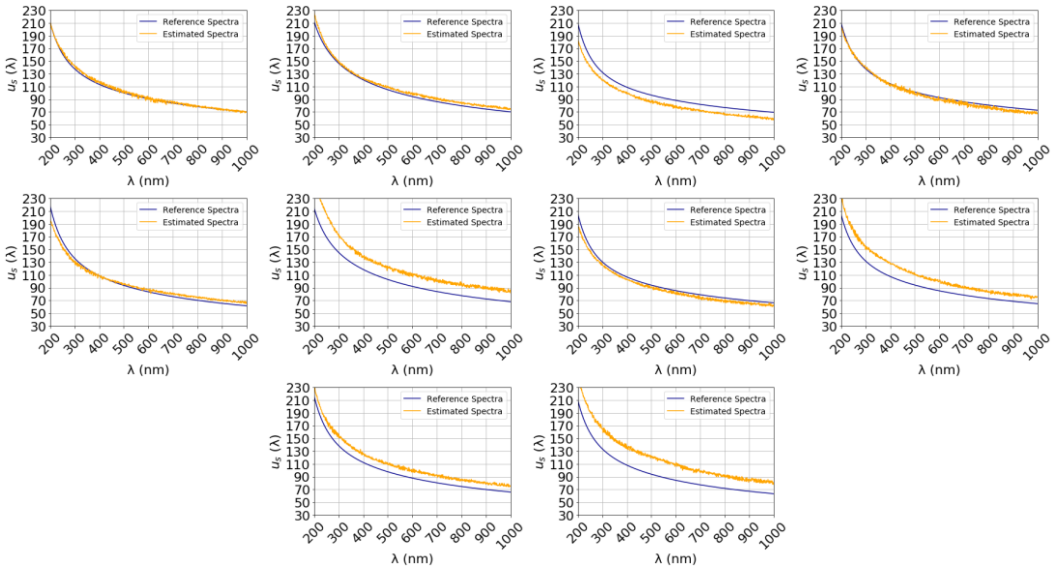


Figure S22. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the SLP model.

Figure S23 shows the estimated healthy spectra from the SLP model that was trained with data from the healthy and pathological samples.

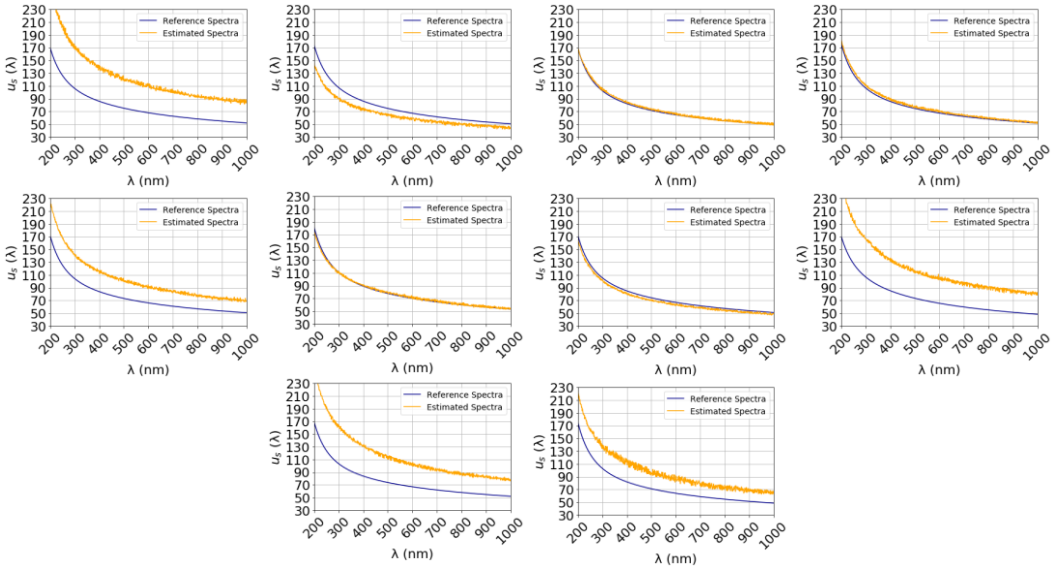


Figure S23. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the SLP model.

Figure S24 shows the estimated pathological spectra from the SLP model that was trained with data from the healthy and pathological samples.

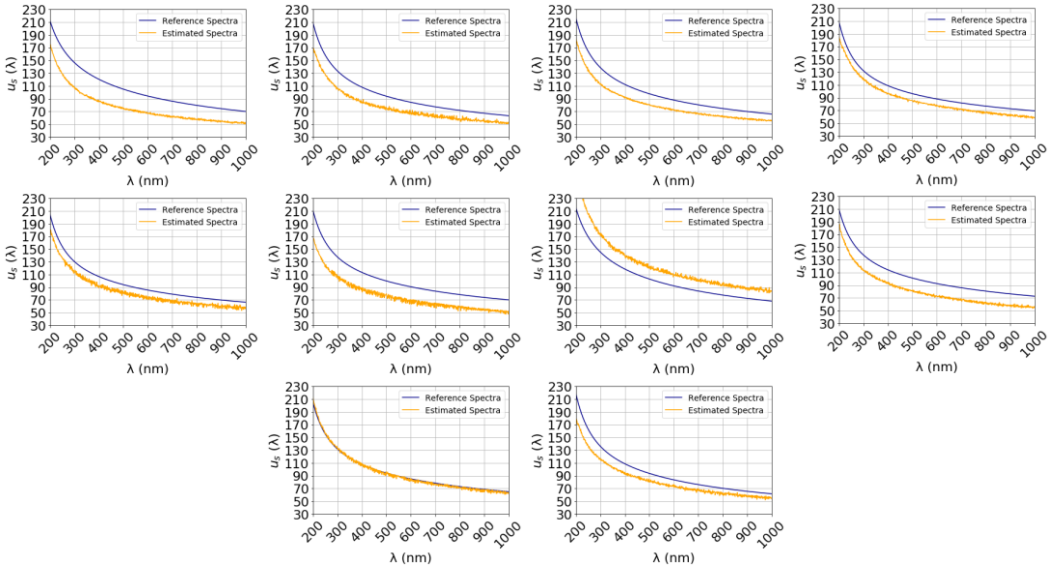


Figure S24. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the SLP model.

Figure S25 presents the estimated spectra from the KNN model that was trained with data from the healthy samples only.

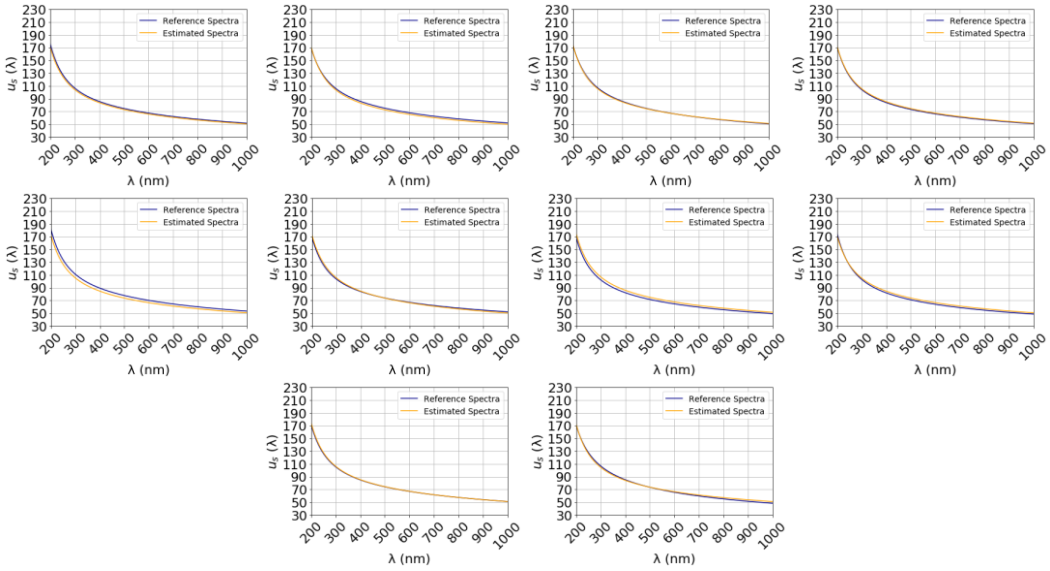


Figure S25. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the KNN model.

Figure S26 presents the estimated spectra from the KNN model trained with data from the pathological samples only.

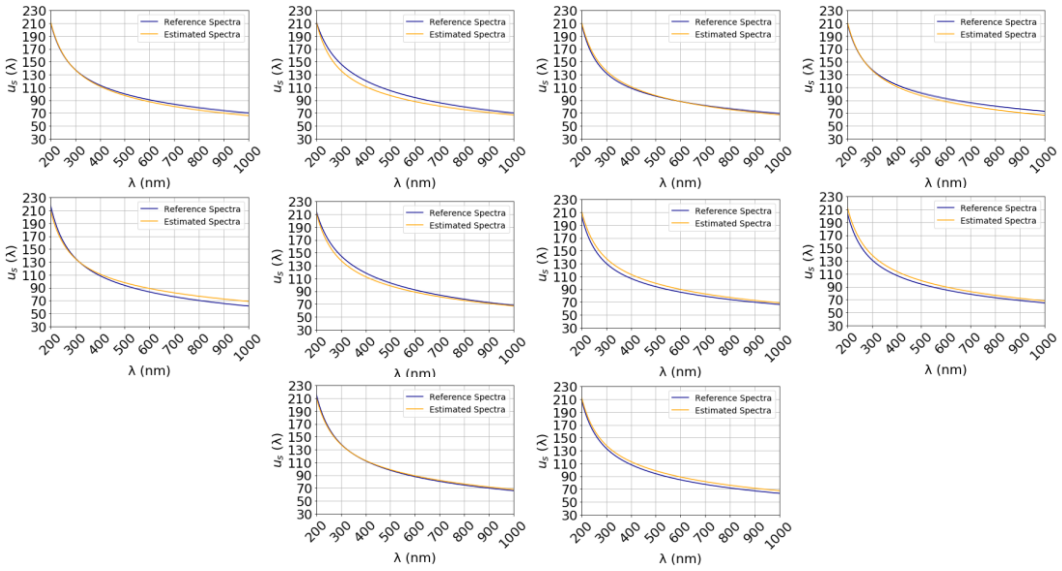


Figure S26. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the KNN model.

Figure S27 shows the estimated healthy spectra from the KNN model that was trained with data from the healthy and pathological samples.

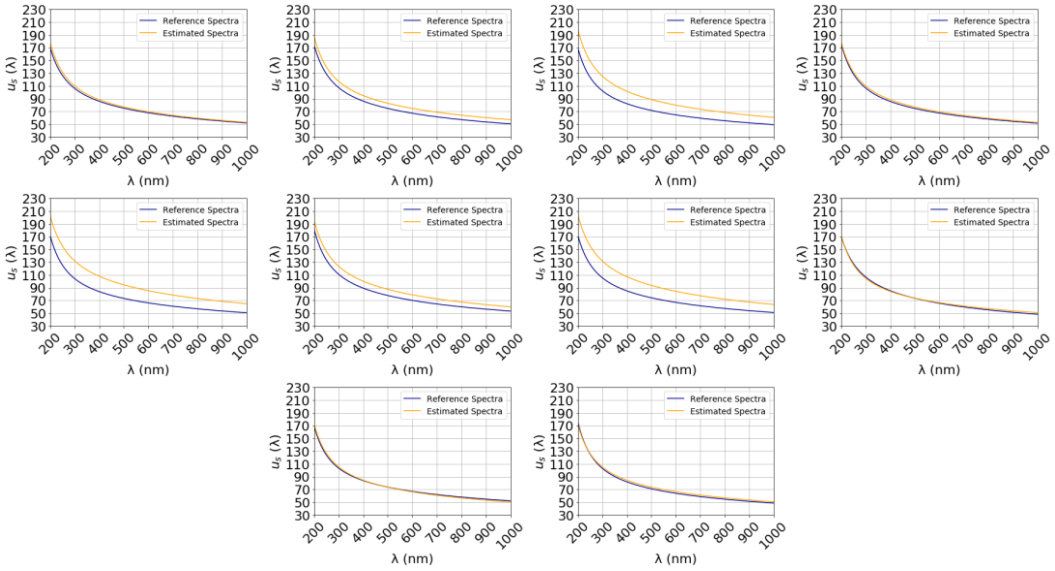


Figure S27. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the KNN model.

Figure S28 shows the estimated pathological spectra from the KNN model that was trained with data from the healthy and pathological samples.

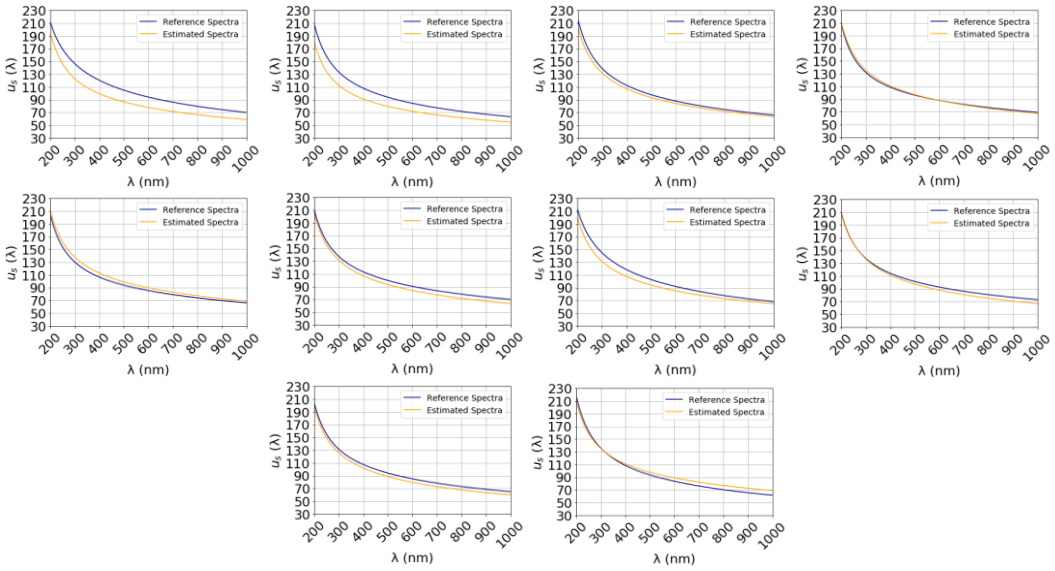


Figure S28. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the KNN model.

Figure S29 presents the estimated spectra from the RFR model that was trained with data from the healthy samples only.

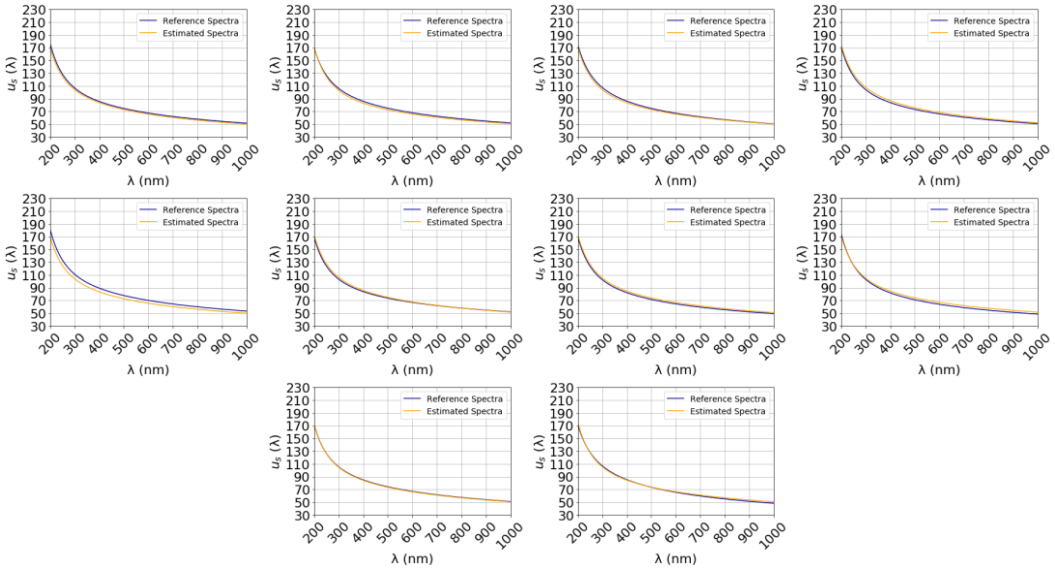


Figure S29. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the RFR model.

Figure S30 presents the estimated spectra from the RFR model trained with data from the pathological samples only.

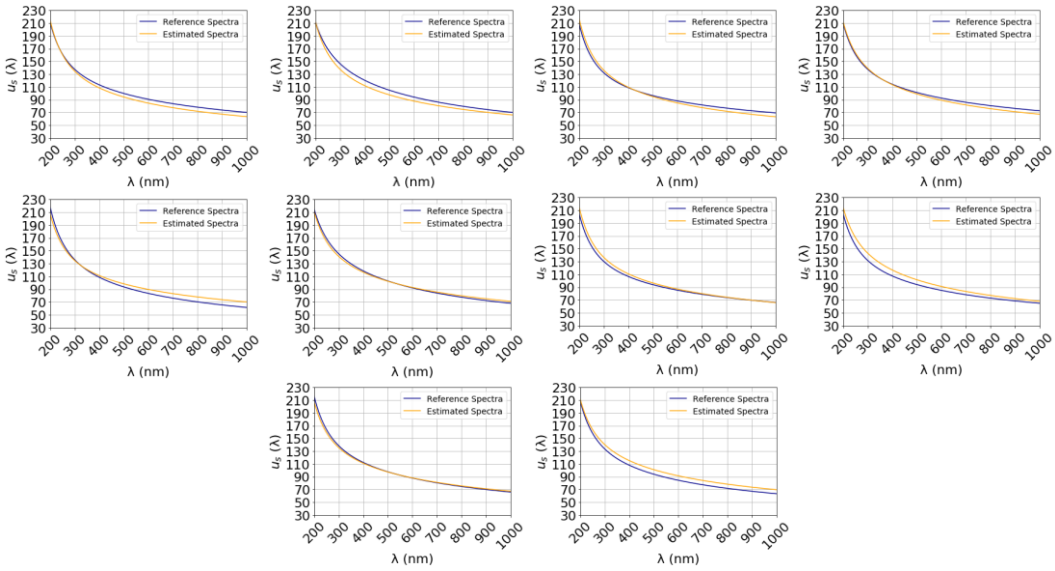


Figure S30. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the RFR model.

Figure S31 presents the estimated healthy spectra from the RFR model that was trained with data from normal and pathological samples.

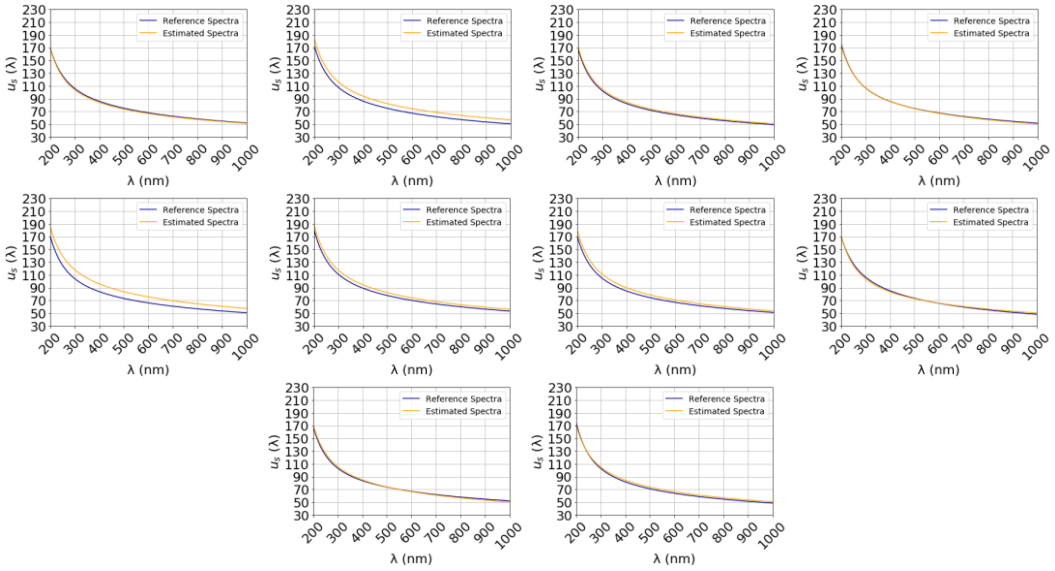


Figure S31. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the RFR model.

Figure S32 shows the estimated pathological spectra from the RFR model that was trained with data from the healthy and pathological samples.

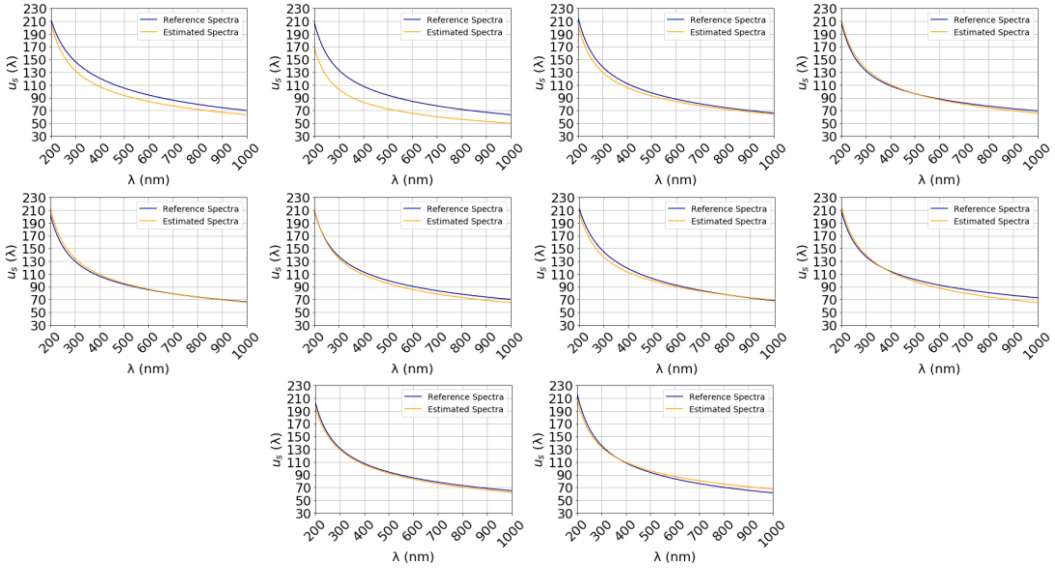


Figure S32. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the RFR model.

Figure S33 presents the individual estimated spectra that were obtained with the DTFMO model trained with data from healthy samples only.

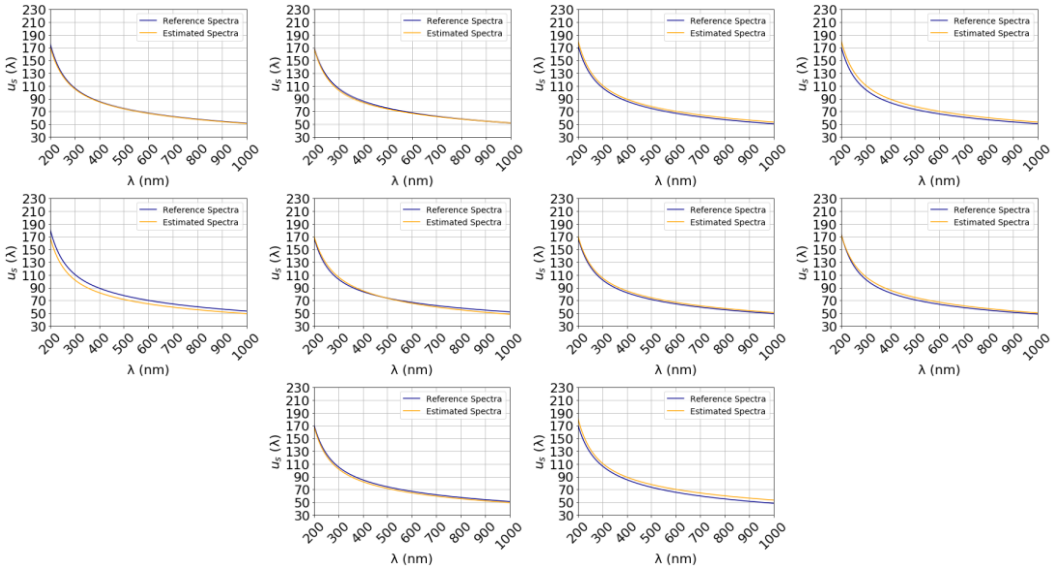


Figure S33. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the DTFMO model.

Figure S34 shows the estimated spectra from the DTFMO model trained with data from pathological samples only.

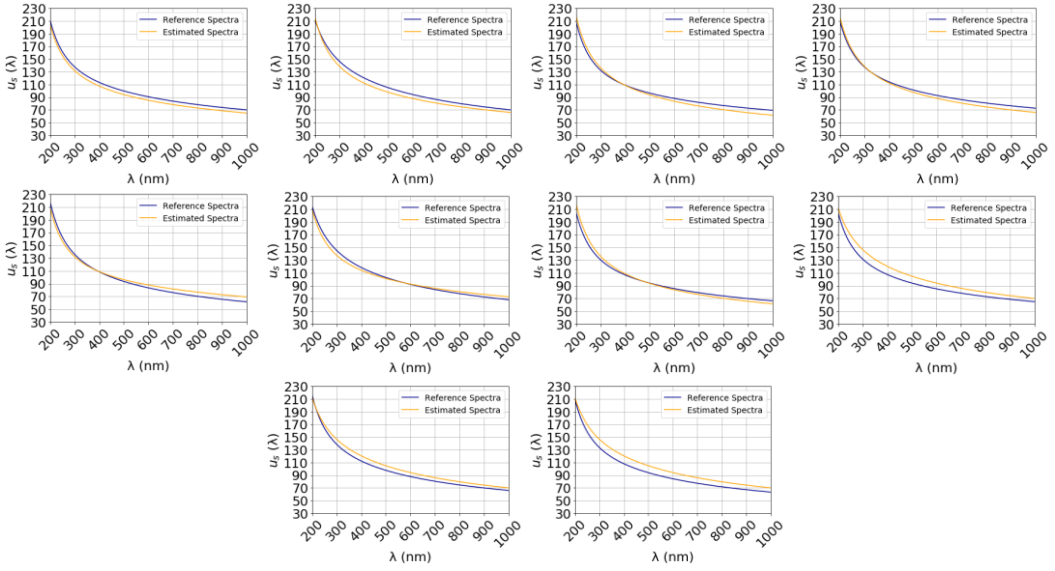


Figure S34. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the DTFMO model.

Figure S35 presents the estimated healthy spectra from the DTFMO model that was trained with data from normal and pathological samples.

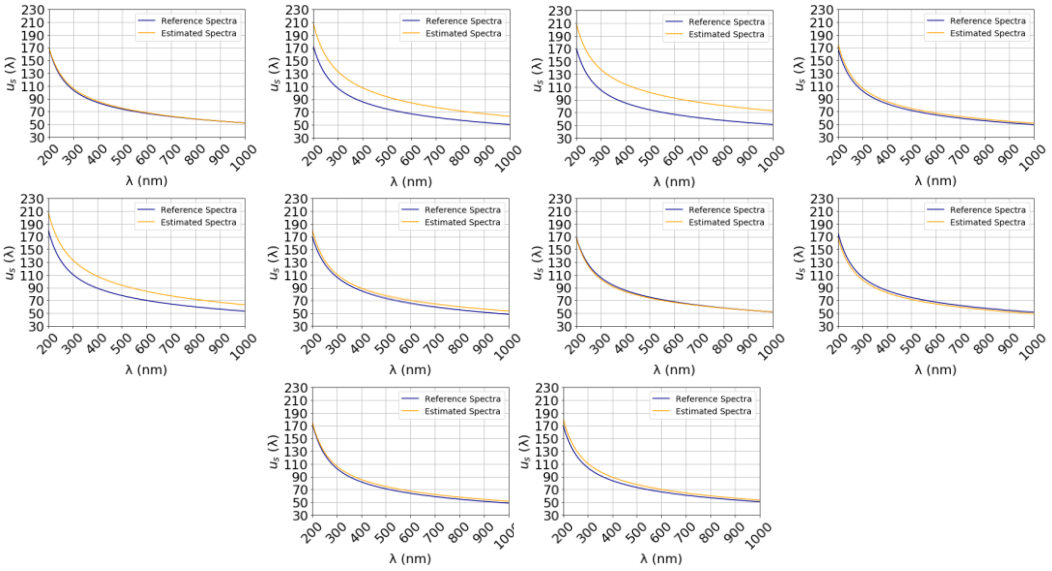


Figure S35. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the DTFMO model.

Figure S36 shows the estimated pathological spectra from the DTFMO model that was trained with data from the healthy and pathological samples.

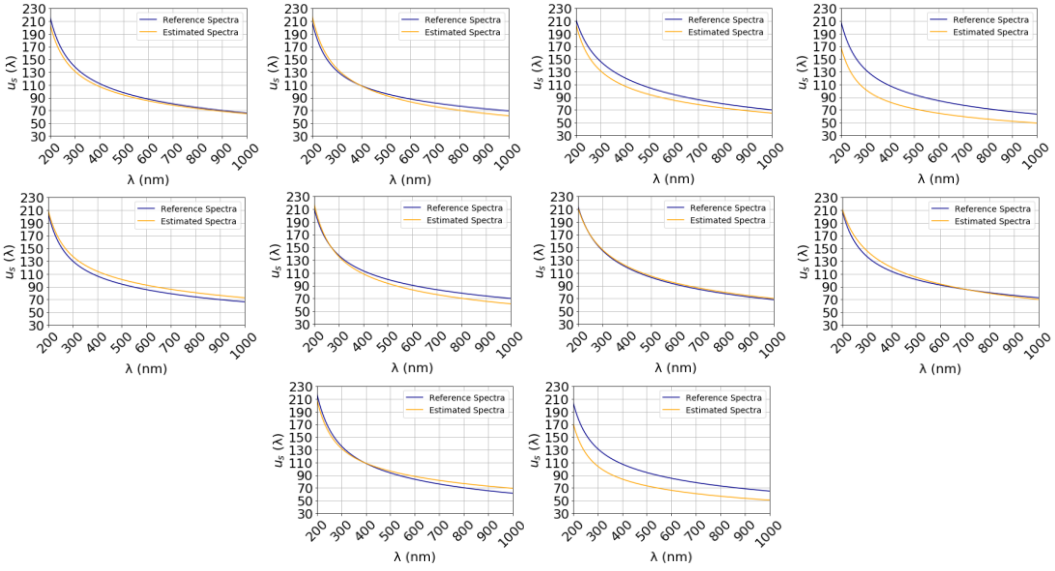


Figure S36. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the DTFMO model.

Figure S37 presents the individual estimated spectra that were obtained with the LRFMO model trained with data from healthy samples only.

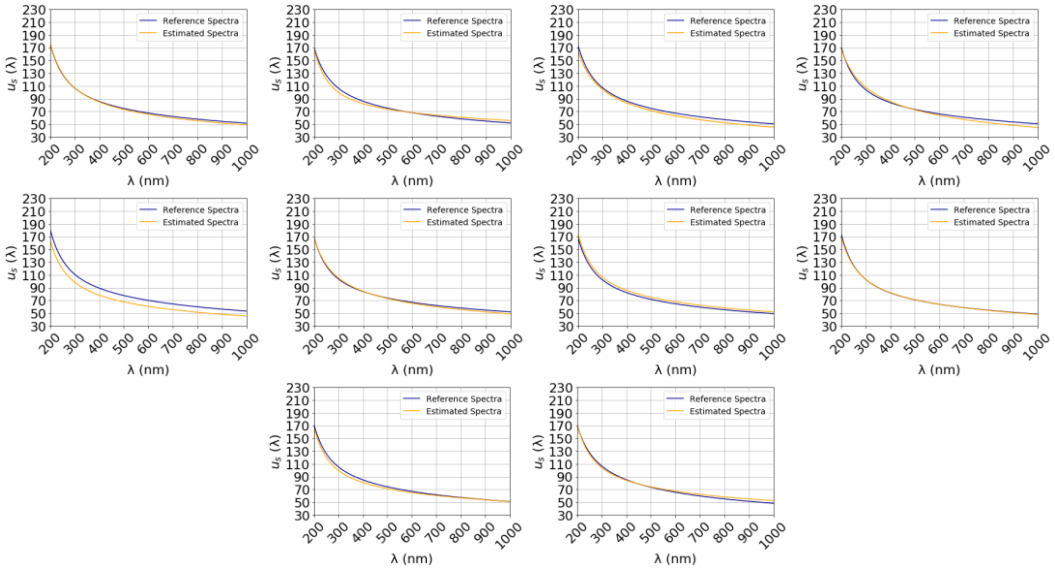


Figure S37. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the LRFMO model.

Figure S38 shows the estimated spectra from the LRFMO model trained with data from pathological samples only.

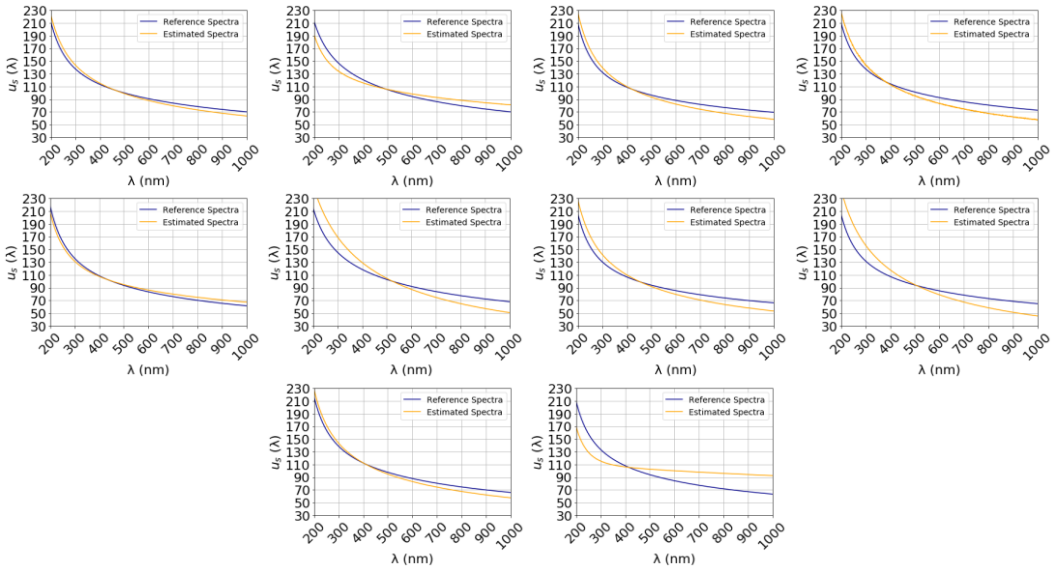


Figure S38. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the LRFMO model.

Figure S39 presents the estimated healthy spectra from the LRFMO model that was trained with data from normal and pathological samples.

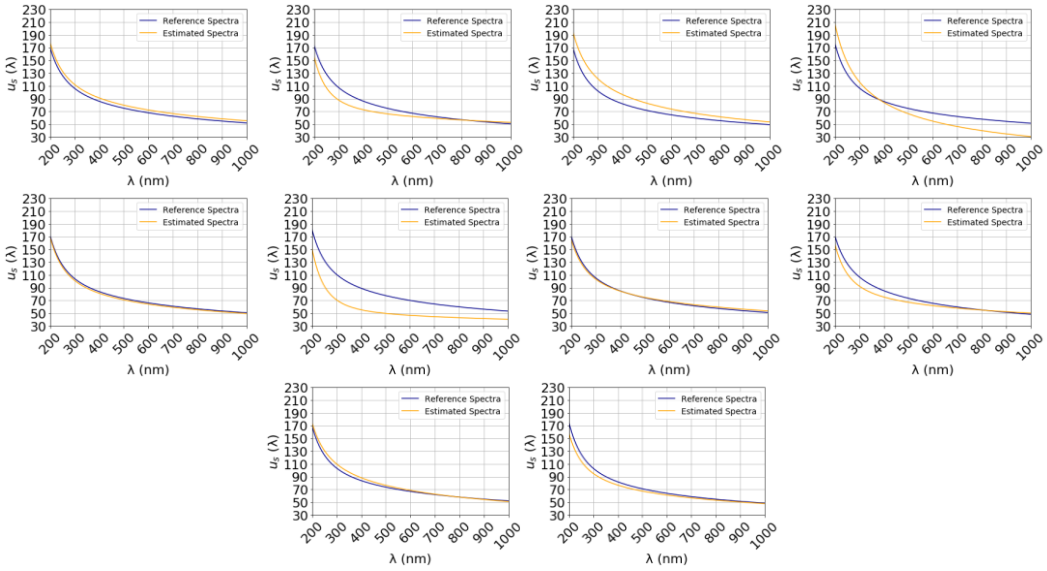


Figure S39. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the LRFMO model.

Figure S40 shows the estimated pathological spectra from the LRFMO model that was trained with data from the healthy and pathological samples.

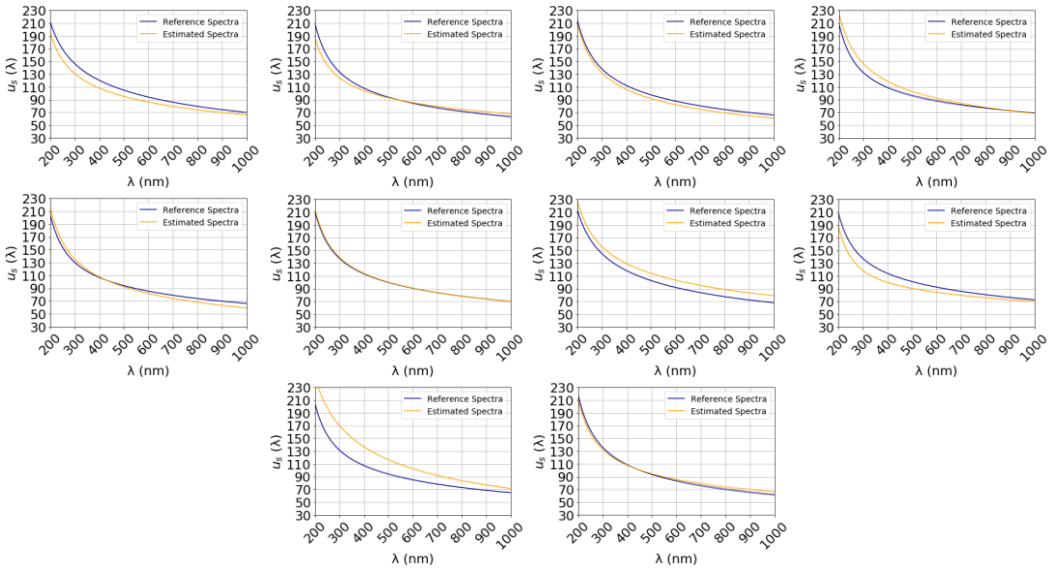


Figure S40. Estimated μ_s spectra for each individual sample (orange), compared to the reference μ_s spectra (blue). The estimated μ_s spectra came from the LRFMO model.

Figure S41 shows the code used to train the SLP TS model that was trained with the normal samples.

```
start=time.time()
for i in range (0,10):
    rd=xns[i].reshape(1,-1)
    x=numpy.delete(xns,i,0)
    y=numpy.delete(yns,i,0)
    model = Sequential()
    model.add(Dense(10, input_shape=input_shape, activation='linear'))
    model.add(Dense(801, activation='linear'))
    model.compile(loss='mean_absolute_error', optimizer='adam', metrics=['mae'])
    history=model.fit(x,y, epochs=50, batch_size=1,verbose=1, validation_split=0.2)
    model.save('modeln'+str(i))
    #model=tensorflow.keras.models.load_model('modeln'+str(i))
    hist = pd.DataFrame(history.history)
    hist['epoch'] = history.epoch
    plt.figure()
    plt.plot(hist['epoch'], hist['mae'],label='Train Error')
    plt.plot(hist['epoch'], hist['val_mae'],label = 'Val Error')
    plt.legend()

    predn[i]=model.predict(rd)
    plt.figure()
    plt.plot(wav,yns[i].transpose(),color='darkblue',label='Reference Spectra')
    #plt.plot(wav,gaussian_filter1d(predn[i],sigma=1))
    plt.plot(wav,predn[i],color='orange',label='Estimated Spectra')
    plt.xlim([200,1000])
    plt.xticks(fontsize=f)
    plt.xticks(rotation=45)
    plt.yticks(numpy.arange(30,250,20))
    plt.ylim([30,230])
    plt.yticks(fontsize=f)
    plt.grid()
    plt.ylabel(r'$u_{s}$ (λ)',fontsize=f)
    plt.xlabel('λ (nm)',fontsize=f)
    plt.legend(prop={'size': 14})
    b=numpy.zeros([1,801],dtype=float)
    for c in range(0,801):
        b[0,c]=abs(yns[i,c].transpose()-predn[i,c])
    euclidean_distancen[0,i]=numpy.mean(b)

end=time.time()
elapsed1 = end - start
```

Figure S 41. Code used to train the SLP TS Model using the LOO method.

Figure S42 shows the code used to train the KNN TS model that was trained with the normal samples.

```
start=time.time()
for i in range (0,10):
    model = KNeighborsRegressor(n_neighbors=5)
    rd=xns[i].reshape(1,-1)
    x=numpy.delete(xns,i,0)
    y=numpy.delete(yns,i,0)
    model.fit(x,y)
    predn[i]=model.predict(rd)
    plt.figure()
    plt.plot(wav,yns[i].transpose(),color='darkblue',label='Reference Spectra')
    plt.plot(wav,predn[i],color='orange',label='Estimated Spectra')
    plt.xlim([200,1000])
    plt.xticks(fontsize=f)
    plt.xticks(rotation=45)
    plt.yticks(numpy.arange(30,250,20))
    plt.ylim([30,230])
    plt.yticks(fontsize=f)
    plt.grid()
    plt.ylabel(r'$u_{s}$ (λ)',fontsize=f)
    plt.xlabel('λ (nm)',fontsize=f)
    plt.legend(prop={'size': 14})
    b=numpy.zeros([1,801],dtype=float)
    for c in range(0,801):
        b[0,c]=abs(yns[i,c].transpose()-predn[i,c])
    euclidean_distancen[0,i]=numpy.mean(b)

end=time.time()
elapsed1 = end - start
```

Figure S 42. Code used to train the KNN TS Model using the LOO method.