Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Raman spectroscopy for classification of neoplastic and non-neoplastic CAM colon tumors

B. Esteves [a], S. Pimenta [a,b,*], M.J. Maciel [a,b], M. Costa [c,d], F. Baltazar [c,d], M.F. Cerqueira [e,f], P. Alpuim [e,f], C.A. Silva [a,b], J.H. Correia [a,b]

[a] *CMEMS-UMinho, Department of Industrial Electronics, University of Minho, Guimarães, Portugal*
[b] *LABBELS – Associate Laboratory, Braga, Guimarães, Portugal*
[c] *Life and Health Sciences Research Institute (ICVS), University of Minho, Campus of Gualtar, Braga, Portugal*
[d] *ICVS/3B's - PT Government Associate Laboratory, Braga, Guimarães, Portugal*
[e] *International Iberian Nanotechnology Laboratory (INL), Braga, Portugal*
[f] *Centre of Physics of Minho and Porto Universities (CF-UM-UP), University of Minho, Braga, Portugal*

ARTICLE INFO

*Keywords:*
Colon cancer
Chicken chorioallantoic membrane assay
Raman peaks
Principal component analysis
Support vector machine

ABSTRACT

This paper demonstrates the potential of Raman spectroscopy for differentiating neoplastic from non-neoplastic colon tumors, obtained with the CAM (chicken chorioallantoic membrane) model. For the CAM model two human cell lines were used to generate two types of tumors, the RKO cell line for neoplastic colon tumors and the NCM460 cell line for non-neoplastic colon tumors. The Raman spectra were acquired with a 785 nm excitation laser. The measured Raman spectra from the CAM samples ($n = 14$) were processed with several methods for baseline correction and to remove artifacts. The corrected spectra were analyzed with PCA (principal component analysis). Additionally, machine learning based algorithms were used to create a model capable of classifying neoplastic and non-neoplastic tumors. The principal component scores showed a clear differentiation between neoplastic and non-neoplastic colon tumors. The classification model had an accuracy of 93 %. Thus, a complete methodology to process and analyze Raman spectra was validated, using a rapid, accessible, and well-established tumor model that mimics the human tumor pathology with minor ethical concerns.

## 1. Introduction

According to WHO (World Health Organization), cancer is a leading cause of death worldwide, with approximately 10 million deaths in 2020. Colon cancer is one of the most common types of cancer. Early detection and effective treatment can increase the probability of patient cure [1] Colonoscopy is the main examination procedure used for colon cancer screening and it is a technique completely based on morphological changes in the tissues. It can be used with biopsy of suspicious tissue areas to increase the probability of cancer detection. However, biopsies are invasive and time-consuming, which can delay patient treatment. Thus, researchers are constantly trying to develop new tools to predict colon cancer, specifically, techniques capable of detecting very small

* Corresponding author. CMEMS-UMinho, Department of Industrial Electronics, University of Minho, Guimarães, Portugal.
*E-mail addresses:* bruno.esteves@dei.uminho.pt (B. Esteves), sara.pimenta@dei.uminho.pt (S. Pimenta), mmaciel@dei.uminho.pt (M.J. Maciel), martafcosta@med.uminho.pt (M. Costa), fbaltazar@med.uminho.pt (F. Baltazar), fcerqueira@fisica.uminho.pt (M.F. Cerqueira), pedro.alpuim.us@inl.int (P. Alpuim), csilva@dei.uminho.pt (C.A. Silva), higino.correia@dei.uminho.pt (J.H. Correia).

https://doi.org/10.1016/j.heliyon.2024.e36981
Received 26 June 2023; Received in revised form 24 August 2024; Accepted 26 August 2024
Available online 28 August 2024

molecular changes in the tissues associated with colon cancer at an early stage [2,3].

Raman spectroscopy is based on the inelastic scattering of light by matter. Raman spectrum (also known as the molecular fingerprint of a sample) measures the energy shift of the scattered photons, which depends on the chemical composition of the molecules from the sample that cause scattering. It is a non-destructive technique, does not require sample pretreatment with chemical dyes or genetic manipulation, and is suitable for non-invasive *in-vivo* applications [4–10].

Due to Raman ability to detect small biochemical changes in the tissue associated with cancer, this technique has been previously reported in several *ex-vivo* and *in-vivo* studies as a technique able to differentiate between healthy and diseased gastrointestinal tissue, with the potential to eliminate or reduce biopsies [5,7]. In 2003, Molckovsky et al. [11] reported a PCA (principal component analysis) and LDA (linear discriminant analysis) based diagnostic models for differentiating adenomas from hyperplastic colon polyps, using *ex-vivo* and *in-vivo* Raman spectroscopy, with an accuracy of 93 % and 95 %, respectively. In 2007, Chowdary et al. [12] acquired the Raman spectra of normal and malignant colon tissues (*ex-vivo*) and used PCA, Mahalanobis distance, spectral residuals, and limit tests to obtain biochemical differences and classify the samples into two groups, obtaining a sensitivity and specificity of 99.5 %. In 2015, Bergholt et al. [13] used an image-guided Raman endoscopy system (based on a fiber probe) to characterize molecular differences in Raman spectra from different sites of colorectal tissues. The authors applied PCA-DA analysis and showed that was possible to distinguish normal and cancerous tissue with 88 % accuracy, referring the huge potential for *in-vivo* early detection of colorectal cancer. More recently, Petersen et al. [14] performed the analysis of a large number of biopsy samples ($n = 343$) with Raman spectroscopy from different sites of the colon. The authors applied SVM (support vector machine) analysis and successfully differentiated cancer from normal and high-risk from low-risk lesions, with an accuracy of 81 % and 77 %, respectively. Besides gastrointestinal tissues, Raman spectroscopy and machine learning methods are widely used to characterize and differentiate cancerous and normal skin, brain and breast tissues [15–17].

The development and validation of a new methodology to process Raman spectra and extract biochemical features involves the use of normal and abnormal tissues, using *ex-vivo* or *in-vivo* samples from patients or animals, which requires ethical procedures (e. g. permissions or approval from ethics committees). Thus, the use of a tumor model can be an interesting alternative, easily evaluating the methodology efficacy to differentiate between normal from abnormal tissues. The CAM (chicken chorioallantoic membrane) model has been previously used to study human tumor growth. In the CAM model, human cells are transplanted onto the membrane that surrounds the embryo of a fertilizer chicken egg. After a few days, a tumor is formed, which has human tumor features, including a highly vascularized structure, multiple cells, and an extracellular matrix. The main advantage of using the well-established CAM models, besides the formation of tumors being very similar to human tumor pathology, is the rapid tumor formation and accessibility [18–20]. Moreover, CAM models have minor ethical concerns and they are in agreement with the “Three Rs” initiative, which aims to find alternatives to reduce animal testing or animal use in research [21].

In this paper, we present the quantitative analysis of Raman spectra measured from neoplastic and non-neoplastic colon tumors obtained with the CAM model. The Raman spectra were processed with several stated methods to achieve baseline correction and to remove artifacts. Then, the corrected Raman spectra were analyzed with PCA, and machine learning algorithms were implemented to create a classification model to differentiate between neoplastic and non-neoplastic samples. To our knowledge, this is the first report of a methodology to process and analyze Raman spectra of samples generated from the CAM model, a rapid, accessible, and a well-established tumor model with minor ethical concerns. This study was a pre-validation for a future *in-vivo* application of this new methodology for quantitative analysis of Raman spectra and classification of suspicious colon tissues during conventional colonoscopy.

## 2. Materials and methods

### 2.1. Cell lines and culture conditions

This study used RKO (human colon cancer) and NCM460 (normal derived colon mucosa) cell lines from IPATIMUP, Porto, Portugal and from INCELL Corporation upon MTA approval (LLC, San Antonio, USA), respectively. These cell lines were grown in standard conditions [22]. To cultivate cells for the assays, they were allowed to grow to approximately 90 % of confluence in the culture flasks. Then, they were reaped by washing the flasks with phosphate-buffer saline (PBS 1×) and detached by trypsin (TrypLETM Express, Gibco) at 37 °C. The respective medium was added to the flasks to inactivate trypsin and then cells were collected and centrifuged for 5

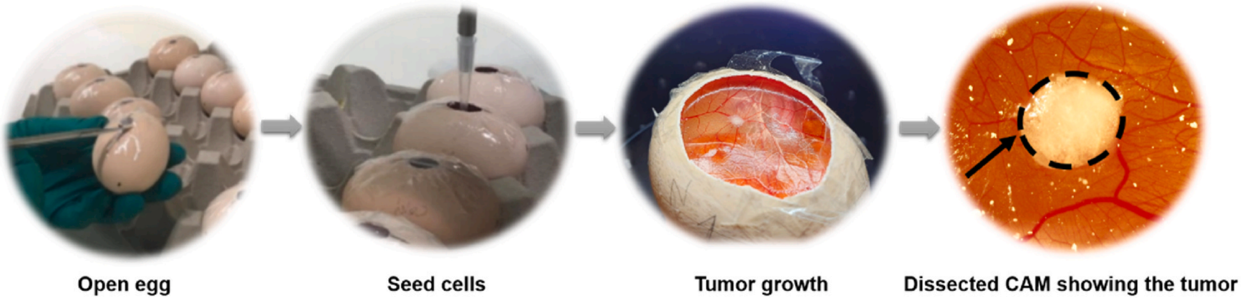


**Fig. 1.** Overview of the CAM assay procedure.

min, at 900 rpm and 4 °C. The culture medium was discarded, cells were re-suspended in a fresh culture medium, and density calculations were performed using trypan blue solution (Trypan Blue Solution, 0.4 %, Gibco).

### 2.2. *Chicken chorioallantoic membrane (CAM) assay*

The CAM assay was used to obtain neoplastic (from RKO cell line) and non-neoplastic (from NCM460 cell line) tumors for analysis. Fertilized chicken eggs were incubated at 37 °C and 70 % humidity and, after three days of incubation, a window was opened into the egg shell and the eggs returned to the incubator. On day 9 of development, RKO or NCM460 cells ($2 \times 10^6$ cells/egg) were injected with Matrigel (10 μL) on the respective CAM and allowed to grow into 3D structures (microtumours). On day 16 of development, the chicken embryos were sacrificed through $CO_2$ exposure for 20 min, followed by exposure to 80 °C for 20 min. Next, the CAM was dissected, properly washed with PBS, and analyzed using Raman spectroscopy. Fig. 1 shows an overwide of the CAM assay procedure.

### 2.3. *Data set and Raman measurements*

This study used 14 CAM samples, 7 of neoplastic colon tumors and 7 of non-neoplastic colon tumors. The Raman spectra were obtained with a WiTec Alpha300 R confocal Raman microscope, from Witec Ulm, Germany. This equipment has a lateral resolution of 200–300 nm, a depth resolution of 500 nm, a spectral resolution of 1 $cm^{-1}$, and a spectrometer with several focal lengths. Several spectra were acquired for each sample with a 785 nm excitation laser at approximately 50 mW and with 3 s of integration time. The Raman spectra were obtained between 800 $cm^{-1}$ and 1800 $cm^{-1}$, since previous studies indicated this range as the most relevant for colon tissue analysis and classification [7].

After the acquisition, the Raman spectra were processed and analyzed, according to the following diagram – Fig. 2. Sections 2.4 and 2.5 also detail data processing and data analysis and classification, respectively. After spectra processing, only one Raman spectrum for each sample was considered (selection explained in section 2.4).

### 2.4. *Data processing*

After acquisition, the Raman spectra were corrected to minimize background fluctuations derived from sample fluorescence and environment variation effects, using the asymmetric least square smoothing (ALS) method, described in section 2.4.1. Secondly, the artifacts originating from high-energy particles (cosmic spikes) were removed using a variation of the modified Z score outlier detection method, described in section 2.4.2. Finally, the spectra were normalized, in order to remove the amplitude variation due to sample geometry and position during signal acquisition. The algorithms for data processing were implemented in python 3.9 (NumPy and SciPy libraries).

After spectra processing, only one Raman spectrum for each sample was considered. Raman spectra whose cosmic spikes coincided with Raman relevant peaks were automatically excluded, and the selected Raman spectrum for each sample was the one with the best signal-to-noise ratio.

#### 2.4.1. *Asymmetric least square smoothing*

The ALS method, used for baseline correction, was developed by Eilers and Boelens [23], and consists of a correction made to the original regularized least squares smoothing, shown as follows in equation (1):

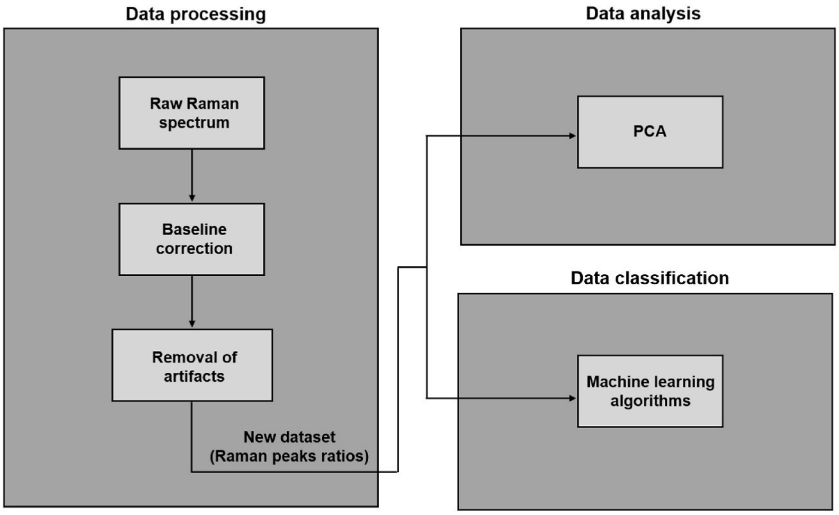


**Fig. 2.** The workflow of data processing, data analysis and data classification.

$$z = \left(W + \lambda D^T D\right)^{-1} W y \tag{1}$$

where $y$ is the original vector, $z$ is the smoothed signal to be found, $W$ is the diagonal matrix for the weight vector $w$, $D$ is the difference matrix, and $\lambda$ is a parameter introduced to regulate the balance between curve fitness and smoothness. On ASL, a special parameter ($p$) is used to modify $w$, by asymmetrically set weights for positive and negative residuals ($y - z$). $w$ is given by equation (2):

$$w_i = \begin{cases} p, y_i \leq z_i \\ 1 - p, y_i \leq z_i \end{cases} \tag{2}$$

The parameters were optimized to the dataset, resulting in a $p$ of 0.02 and a $\lambda$ of 70,000.

### 2.4.2. Z score outlier detection method

A variation of the modified Z score outlier detection method developed by Whiteker and Hayes was used for the removal of artifacts generated by random high-energy spikes [24]. For this method, modified Z scores ($Z_t$) are defined by equation (3):

$$Z_t = \frac{0,6745 \times (\nabla Y_t - M)}{MAD} \tag{3}$$

where $Y_t$ represents the input values, $\nabla Y_t$ represents the detrended differenced series, $M$ represents the median of $\nabla Y_t$ and $MAD$ represents the median of $|\nabla Y_t - M|$. After $Z_t$ calculation, a threshold ($\tau$) was used to determine which points indicate spikes. This value was optimized in order to ensure optimal peak detection and removal. Then, final interpolated values ($\widetilde{Y}_t$) were obtained for each detected point, as the mean of their correspondent neighbors, defined by equation (4):

$$\widetilde{Y}_t = \frac{1}{W} \sum_{t-m}^{t+m} Y_t \times p \tag{4}$$

Where $W$ is defined by equation (5):

$$W = \sum_{t-m}^{t+m} |Z_t| \times p \tag{5}$$

$p$ is defined by equation (6):

$$p = \begin{cases} 1, |Z_t| < \tau \\ 0, |Z_t| \geq \tau \end{cases}, \tag{6}$$

and $m$ defines the width of the moving average. These steps ensure that the values of $Y_t$ and neighbors $Y$ values flagged as contributing to spike formation are removed from $\widetilde{Y}_t$ determination. The optimization of the parameters resulted in a $\tau$ of 4 and a $m$ of 5.

## 2.5. Data analysis and classification

A new dataset was created where each sample is composed by all possible ratios between the Raman peaks located at 855, 875, 936, 1004, 1070, 1218, 1265, 1302, 1335, 1445, 1576, 1618, 1655, and 1745 cm$^{-1}$ of each processed spectrum. These Raman peaks are referred at the literature as crucial Raman spectrum features to the differentiation of benign and cancerous gastrointestinal tissues [25–27]. The use of ratios instead of Raman peaks allows the dataset to incorporate the interdependent relationships between the original variables. Table 1 shows the chemical significance of each peak.

Then, 5 component PCA was used on the new dataset to perform data analysis, i. e., to see if it was possible to differentiate between

**Table 1**
Raman peaks and respective biochemical assignments (δ: bending mode; ν: stretching mode; $\nu_S$: symmetric stretching mode), adapted from Ref. [27].

| Raman peak (cm$^{-1}$) | Assignment |
|---|---|
| 855 | δ(CCH) ring breathing of collagen |
| 875 | ν(C-C) of hydroxyproline |
| 936 | ν(C-C) in α conformation of proteins |
| 1004 | $\nu_S$(C-C) symmetric ring breathing of phenylalanine |
| 1070 | O-P-O stretching of nucleic acids |
| 1218 | ν(C-$C_6H_5$) of proteins |
| 1265 | ν(C-N), δ (N-H) of amide III (α-helix) of proteins |
| 1302 | $CH_3CH_2$ wagging of proteins |
| 1335 | $CH_3CH_2$ twisting f proteins and nucleic acids |
| 1445 | ν(C-N) in-plane vibration of proteins and lipids |
| 1576 | δ(C=C) pyrimidine ring of nucleic acids |
| 1618 | δ(C=C) (in-plane) of proteins |
| 1655 | ν(C=O) of amide I of proteins |
| 1745 | ν(C=O) stretching of phospholipids |

neoplastic and non-neoplastic samples. PCA is a dimension reduction approach widely used, which allows the extraction of mathematical information from a set of variables, by concentrating the information on a smaller sequence of components organized in decreasing order, from the highest to the lowest variance [28,29]. PCA with several components were tested, concluding that 5 component PCA produced the best results in terms of the differentiation between tumor types. Above 5 component PCA no significant improvements were verified in the differentiation of neoplastic and non-neoplastic colon tumors.

Finally, machine learning based algorithms were used to create a model capable of classifying neoplastic and non-neoplastic samples. In the medical field, machine learning techniques are widely used for biomedical data analysis, making it possible to analyze a large amount of data in a shorter time, speeding up cancer detection [30].

The performance of a model can be largely dependent on the dataset training/testing division, especially for datasets with a reduced number of samples. As such, to attenuate that effect, all the 49 training/testing division possibilities were evaluated using 12 samples for training (6 neoplastic and 6 non-neoplastic) and the remaining 2 for testing (1 neoplastic and 1 non-neoplastic). Based on the SciKit library, SVM classification was tested using the LinearSVC method and the SVC (support vector classifier) method with linear, polynomial, RBF (radial basis function), and sigmoid kernels. The effect of transforming the dataset was also tested, using a scaler (remove mean and scale to unit variance) and PCA (5 components), on the performance of the classifier.

For each one of the 49 training/testing division possibility described above, the best parameters were searched individually and independently. For each, a grid search algorithm using cross-validation (leave-one-out) was used to determine the best parameters for each method and kernel, and using or not using the scaler and PCA transformations. The described transformations (scaler and PCA) are always first fitted to the training data and then applied to the testing data. For each fold of the leave-one-out cross-validation, the transformations, if in use, are first fitted to the n-1 training samples, and then the classifier is trained. Finally, the test sample is transformed by the previously fitted models and used for testing. Fig. 3 shows a diagram explaining the methodology for the dataset division, grid search, and model training and testing. All algorithms were implemented in python 3.9 (NumPy and Scikit Learn libraries).

## 3. Results and discussion

### 3.1. Raman spectra of neoplastic and non-neoplastic membranes

Fig. 4 (a) and 4 (b) show the normalized mean Raman spectra of the neoplastic and non-neoplastic colon tumors, respectively. Both mean spectra present the signature of lipids and proteins, in the range of 800 $cm^{-1}$ to 1800 $cm^{-1}$, and are very similar to the literature Raman spectra of colorectal tissues [7,26]. Fig. 4 (c) shows the difference between the mean Raman spectra of neoplastic colon tumors and non-neoplastic colon tumors. As it can be seen, the neoplastic colon tumors have a stronger protein signature (at ~ 1450 $cm^{-1}$ and ~1650 $cm^{-1}$), compared to the non-neoplastic colon tumors. A stronger protein signature is usually indicative of malignancy in colon mucosa [26]. On the other hand, the non-neoplastic colon tumors have stronger lipid signatures (especially at the region of ~1750 $cm^{-1}$ and below ~ 1400 $cm^{-1}$), when comparing with the neoplastic colon tumors, which is also expected for colon mucosa [26].

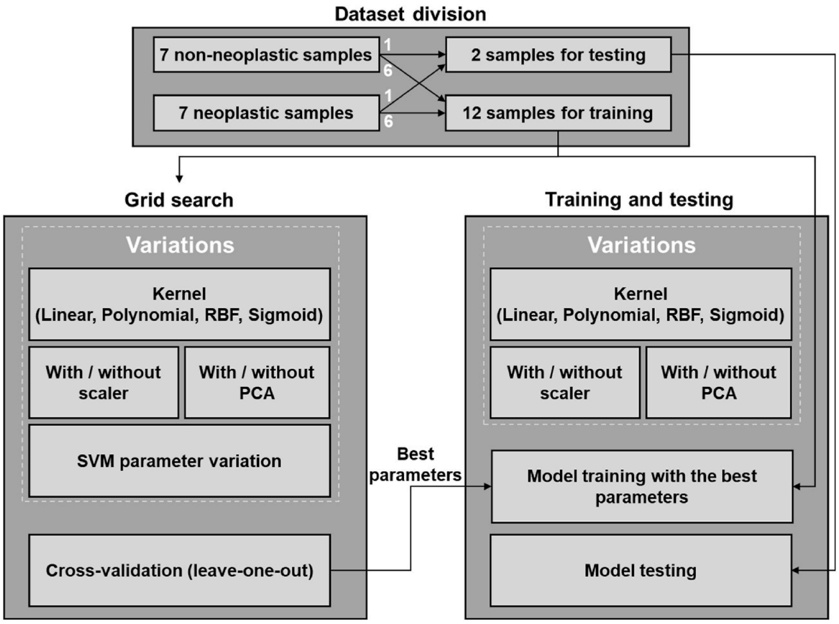


**Fig. 3.** Workflow of the dataset division, grid search and model training and testing, used for each one of the 49 training/testing dataset division possibilities.

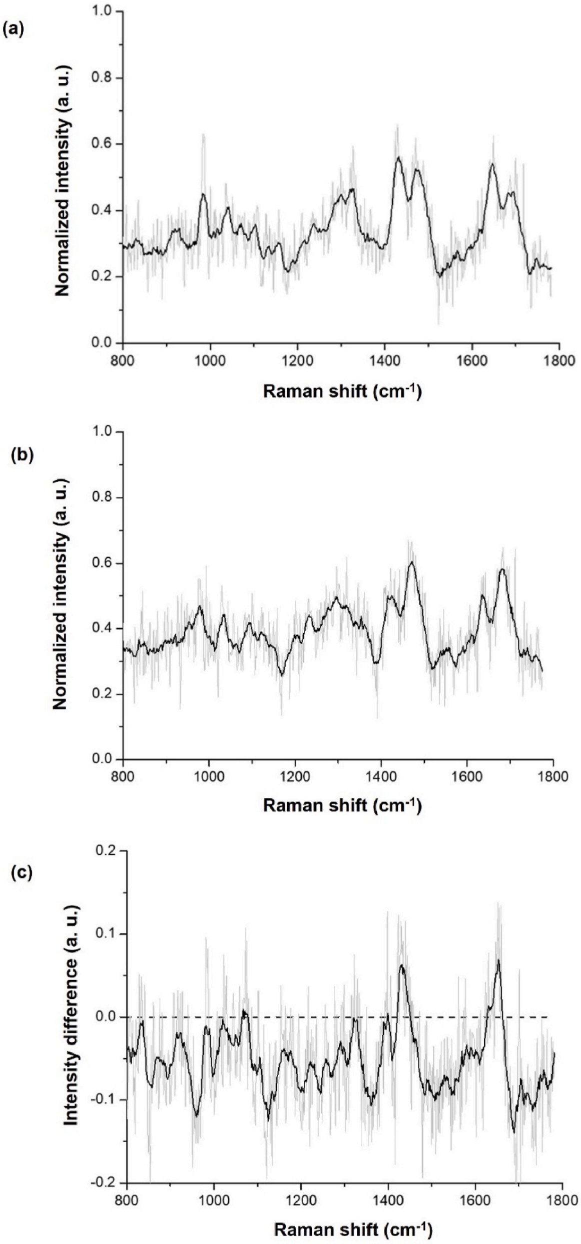


**Fig. 4.** (a) Mean of all the normalized Raman spectra corresponding to the neoplastic colon tumor membranes (using RKO cell line), n = 7. (b) Mean of all the normalized Raman spectra corresponding to the non-neoplastic colon tumor membranes (using NCM460 cell line), n = 7. (c) Difference between the mean Raman spectra of neoplastic tumor and non-neoplastic tumor membranes. Curves in black were obtained using a Savitzky-Golay filter applied to the data points (only for data visualization).

### *3.2. Principal component analysis*

Fig. 5 shows the principal component scores obtained for all the analyzed neoplastic and non-neoplastic membranes, when applying a 5 component PCA. As it can be seen, when plotting the third and fourth components, there is a clear separation between the neoplastic and non-neoplastic samples, which indicates the existence of distinguishable features between the two types of membranes on the acquired Raman spectrum.

The eigenvalues of the covariance matrix show different contributions from each principal component: first component 32.7 %, second component 19.7 %, third component 14.0 %, fourth component 10.3 %, and fifth component 8.2 %. The importance of each component is reflected on the corresponding value in the eigenvectors, and the importance of different peaks can be deduced from the largest absolute values of the ratios in the eigenvectors. For the third component, the most important peaks are at 936, 1004, 1070,

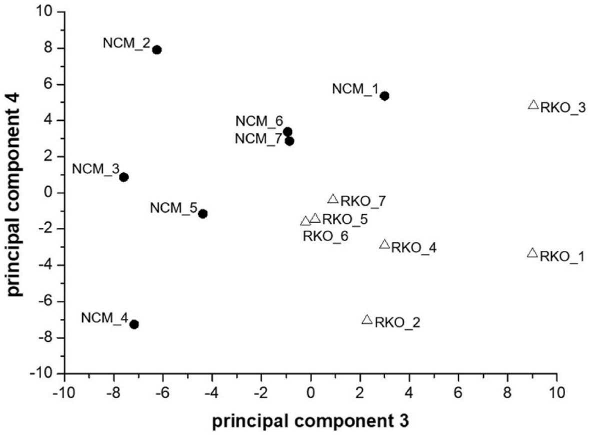


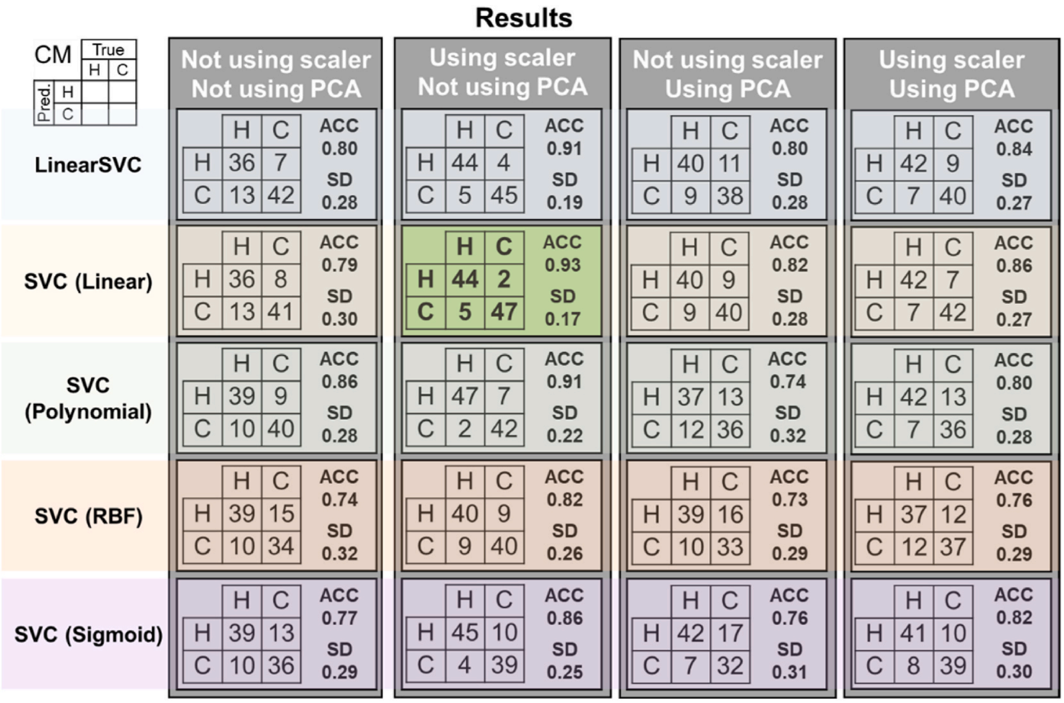
**Fig. 5.** Principal component scores of neoplastic colon tumors (using RKO cell line) and non-neoplastic colon tumors (using NCM460 cell line).

1335, 1618, and 1745 cm$^{-1}$. For the fourth component the most important peaks are at 1265, 1302, 1335, 1618, and 1745 cm$^{-1}$. The corresponding loading vectors are available alongside the source code (see data availability statement).

### 3.3. Support vector machine classification

Fig. 6 shows the results of the SVM classification analysis, where each confusion matrix corresponds to the sum of the 49 confusion matrices (from the training/testing division possibilities) for a specific kernel and transformations. As a result, the accuracy is the mean accuracy of those 49 training/testing division possibilities. The approach was used to avoid a non-representative high-performance result obtained from random train/test data division, and to maximize the number of training data. This was necessary due to the low sample size, which is the main limitation of this study.

The best accuracy (93 %) was obtained with a linear kernel, using the SVC method, and using the scaler transformation previously described. The obtained accuracy is the range of those obtained in literature, as described in the Introduction section. The most similar study reported in the literature that have used SVM to differentiated cancer from normal and high-risk from low-risk lesions, reported


Results

CM: True H C; Pred. H, C

| | Not using scaler<br>Not using PCA | Using scaler<br>Not using PCA | Not using scaler<br>Using PCA | Using scaler<br>Using PCA |
|---|---|---|---|---|
| LinearSVC | H C<br>H 36 7<br>C 13 42<br>ACC 0.80<br>SD 0.28 | H C<br>H 44 4<br>C 5 45<br>ACC 0.91<br>SD 0.19 | H C<br>H 40 11<br>C 9 38<br>ACC 0.80<br>SD 0.28 | H C<br>H 42 9<br>C 7 40<br>ACC 0.84<br>SD 0.27 |
| SVC (Linear) | H C<br>H 36 8<br>C 13 41<br>ACC 0.79<br>SD 0.30 | **H C**<br>**H 44 2**<br>**C 5 47**<br>ACC 0.93<br>SD 0.17 | H C<br>H 40 9<br>C 9 40<br>ACC 0.82<br>SD 0.28 | H C<br>H 42 7<br>C 7 42<br>ACC 0.86<br>SD 0.27 |
| SVC (Polynomial) | H C<br>H 39 9<br>C 10 40<br>ACC 0.86<br>SD 0.28 | H C<br>H 47 7<br>C 2 42<br>ACC 0.91<br>SD 0.22 | H C<br>H 37 13<br>C 12 36<br>ACC 0.74<br>SD 0.32 | H C<br>H 42 13<br>C 7 36<br>ACC 0.80<br>SD 0.28 |
| SVC (RBF) | H C<br>H 39 15<br>C 10 34<br>ACC 0.74<br>SD 0.32 | H C<br>H 40 9<br>C 9 40<br>ACC 0.82<br>SD 0.26 | H C<br>H 39 16<br>C 10 33<br>ACC 0.73<br>SD 0.29 | H C<br>H 37 12<br>C 12 37<br>ACC 0.76<br>SD 0.29 |
| SVC (Sigmoid) | H C<br>H 39 13<br>C 10 36<br>ACC 0.77<br>SD 0.29 | H C<br>H 45 10<br>C 4 39<br>ACC 0.86<br>SD 0.25 | H C<br>H 42 17<br>C 7 32<br>ACC 0.76<br>SD 0.31 | H C<br>H 41 10<br>C 8 39<br>ACC 0.82<br>SD 0.30 |

**Fig. 6.** Results of the classification attempts using SVM models (C stands for neoplastic samples, H for non-neoplastic samples and Pred. for Predicted).

an accuracy of 81 % and 77 %, respectively [14]. Other studies using PCA and LDA and PCA-DA analysis do differentiate colon lesions reporting accuracies between 88 % and 95 % [11,13].

The result contains 2 false negatives and 5 false positives, considering that a negative result is non-neoplastic and a positive result is neoplastic. Thus, the lower false negative rate (0.04) is a good indicator for diagnostic applications, as false negatives can delay patient treatment and cure and, ultimately, have grievous consequences. It is also important to refer that all the obtained false negatives were from the RKO_2 sample, and all false positives were from the NCM_4 sample. Analyzing the raw Raman spectra of these samples, it was observed a weak signal-to-noise ratio in both samples. Additionally, for the RKO_2 sample only one Raman spectrum acquisition was performed, since the Raman signal was truly weak.

## 4. Conclusions

This study used the CAM model and human colon cell lines to produce neoplastic and non-neoplastic tumor samples for Raman measurements. The measured Raman spectra were baseline corrected, filtered, normalized and analyzed with PCA. The principal component scores indicated a clear distinction between the Raman spectra obtained from neoplastic and non-neoplastic samples. SVM models were used for data classification. Several SVM models were tested, and the best result (accuracy of 93 %) was achieved with a linear kernel, using the SVC method, and a scaler transformation. The high accuracy value obtained indicates that the data contains enough information for training a classifier with a high-performance level, even when using a low number of samples ($n = 14$). All the obtained results gave promising signs for further use of this quantitative methodology to analyze the Raman spectra and classify a sample. This study shows that the CAM model can be a valid and very interesting alternative for the study of human colon tumors and that Raman spectroscopy can be used to distinguish between neoplastic and non-neoplastic tissue. Future directions include the application of this new and optimized methodology in a larger sample and, ultimately, in *in-vivo* studies.

## Data availability statement

The data and code required to reproduce the above findings are available to download from:
data link – https://zenodo.org/records/11060720
code link – https://github.com/BBrunoEsteves/Raman_spectra_processing_and_classification.

## CRediT authorship contribution statement

**B. Esteves:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **S. Pimenta:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **M.J. Maciel:** Writing – review & editing, Validation, Project administration, Investigation, Funding acquisition, Conceptualization. **M. Costa:** Writing – review & editing, Resources, Investigation, Data curation. **F. Baltazar:** Writing – review & editing, Resources, Investigation, Data curation. **M.F. Cerqueira:** Writing – review & editing, Resources, Investigation, Data curation. **P. Alpuim:** Writing – review & editing, Resources, Investigation, Data curation. **C.A. Silva:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **J.H. Correia:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by: project OCT-Raman, reference PTDC/FIS OTI/28296/2017, operation code NORTE 01-0145-FEDER-028296, through the COMPETE 2020; project MPhotonBiopsy, reference PTDC/FIS-OTI/1259/2020, https://doi.org/10.54499/PTDC/FIS-OTI/1259/2020; CMEMS-UMinho Strategic Project, UIDB/04436/2020 and UIDP/04436/2020; and Infrastructures Micro&NanoFabs@PT, operation code NORTE 01-0145-FEDER-022090, POR Norte, Portugal 2020; project UIDB/50026/2020, UIDP/50026/2020 and by the project NORTE-01-0145-FEDER-000055, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). B. Esteves thanks FCT (Fundação para a Ciência e a Tecnologia) for the Ph.D. grant, 2021.07239. BD. S. Pimenta thanks FCT for the grant 2022.00101. CEECIND/CP1718/CT0008, https://doi.org/10.54499/2022.00101.CEECIND/CP1718/CT0008. The authors also thank Dr. Raquel Seruca from IPATIMUP, Porto, Portugal for kindly provided the RKO cell line.

## References

[1] World Health Organization, Cancer - key facts [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. (Accessed 1 April 2022).

[2] B. Brozek-Pluska, A. Dziki, H. Abramczyk, Virtual spectral histopathology of colon cancer - biomedical applications of Raman spectroscopy and imaging, J. Mol. Liq. 303 (Apr. 2020) 112676, https://doi.org/10.1016/j.molliq.2020.112676.

[3] H. Noothalapati, K. Iwasaki, T. Yamamoto, Non-invasive diagnosis of colorectal cancer by Raman spectroscopy: recent developments in liquid biopsy and endoscopy approaches, Spectrochim. Acta Mol. Biomol. Spectrosc. 258 (Sep. 2021) 119818, https://doi.org/10.1016/j.saa.2021.119818.

[4] P. Rostron, S. Gaber, D. Gaber, Raman spectroscopy, review, International Journal of Engineering and Technical Research (IJETR) 6 (1) (2016).

[5] K.J.I. Ember, et al., Raman spectroscopy and regenerative medicine: a review, NPJ Regen Med 2 (1) (Dec. 2017) 12, https://doi.org/10.1038/s41536-017-0014-3.

[6] C.A. Jenkins, et al., A high-throughput serum Raman spectroscopy platform and methodology for colorectal cancer diagnostics, Analyst 143 (24) (2018) 6014–6024, https://doi.org/10.1039/C8AN01323C.

[7] H. Noothalapati, K. Iwasaki, T. Yamamoto, Non-invasive diagnosis of colorectal cancer by Raman spectroscopy: recent developments in liquid biopsy and endoscopy approaches, Spectrochim. Acta Mol. Biomol. Spectrosc. 258 (Sep. 2021) 119818, https://doi.org/10.1016/j.saa.2021.119818.

[8] J. Fleureau, et al., Characterization of renal tumours based on Raman spectra classification, Expert Syst. Appl. (Jun. 2011), https://doi.org/10.1016/j.eswa.2011.05.092.

[9] N.A. Correia, et al., Detection of prostate cancer by Raman spectroscopy: a multivariate study on patients with normal and altered PSA values, J. Photochem. Photobiol., B 204 (Mar. 2020) 111801, https://doi.org/10.1016/j.jphotobiol.2020.111801.

[10] E.E. Rossi, A.L.B. Pinheiro, O.C. Baltatu, M.T.T. Pacheco, L. Silveira, Differential diagnosis between experimental endophthalmitis and uveitis in vitreous with Raman spectroscopy and principal components analysis, J. Photochem. Photobiol., B 107 (Feb. 2012) 73–78, https://doi.org/10.1016/j.jphotobiol.2011.12.001.

[11] A. Molckovsky, L.-M.W.K. Song, M.G. Shim, N.E. Marcon, B.C. Wilson, Diagnostic potential of near-infrared Raman spectroscopy in the colon: differentiating adenomatous from hyperplastic polyps, Gastrointest. Endosc. 57 (3) (Mar. 2003) 396–402, https://doi.org/10.1067/mge.2003.105.

[12] M.V.P. Chowdary, et al., Discrimination of normal and malignant mucosal tissues of the colon by Raman spectroscopy, Photomed Laser Surg 25 (4) (Aug. 2007) 269–274, https://doi.org/10.1089/pho.2006.2066.

[13] M.S. Bergholt, et al., Characterizing variability of in vivo Raman spectroscopic properties of different anatomical sites of normal colorectal tissue towards cancer diagnosis at colonoscopy, Anal. Chem. 87 (2) (Jan. 2015) 960–966, https://doi.org/10.1021/ac503287u.

[14] D. Petersen, et al., Raman fiber-optical method for colon cancer detection: cross-validation and outlier identification approach, Spectrochim. Acta Mol. Biomol. Spectrosc. 181 (Jun. 2017) 270–275, https://doi.org/10.1016/j.saa.2017.03.054.

[15] C. Liu, et al., Characterization and discrimination of basal cell carcinoma and normal human skin tissues using resonance Raman spectroscopy, in: Frontiers in Optics 2017, OSA, Washington, D.C., 2017, p. JTu2A.72, https://doi.org/10.1364/FIO.2017.JTu2A.72.

[16] Y. Zhou, et al., Combined spatial frequency spectroscopy analysis with visible resonance Raman for optical biopsy of human brain metastases of lung cancers, J Innov Opt Health Sci 12 (2) (Mar. 2019), https://doi.org/10.1142/S179354581950010X.

[17] E. Bendau, et al., Distinguishing metastatic triple-negative breast cancer from nonmetastatic breast cancer using second harmonic generation imaging and resonance Raman spectroscopy, J. Biophot. 13 (7) (Jul. 2020), https://doi.org/10.1002/jbio.202000005.

[18] B.T. Vu, et al., Chick chorioallantoic membrane assay as an in vivo model to study the effect of nanoparticle-based anticancer drugs in ovarian cancer, Sci. Rep. 8 (1) (Dec. 2018) 8524, https://doi.org/10.1038/s41598-018-25573-8.

[19] J. Eckrich, et al., Monitoring of tumor growth and vascularization with repetitive ultrasonography in the chicken chorioallantoic-membrane-assay, Sci. Rep. 10 (1) (Dec. 2020) 18585, https://doi.org/10.1038/s41598-020-75660-y.

[20] P. Nowak-Sliwinska, T. Segura, M.L. Iruela-Arispe, The chicken chorioallantoic membrane model in biology, medicine and bioengineering, Angiogenesis 17 (4) (Oct. 2014) 779–804, https://doi.org/10.1007/s10456-014-9440-7.

[21] A. Guerra, J. Belinha, N. Mangir, S. MacNeil, R. Natal Jorge, Simulation of the process of angiogenesis: quantification and assessment of vascular patterning in the chicken chorioallantoic membrane, Comput. Biol. Med. 136 (Sep. 2021) 104647, https://doi.org/10.1016/j.compbiomed.2021.104647.

[22] R. Amorim, et al., Monocarboxylate transport inhibition potentiates the cytotoxic effect of 5-fluorouracil in colorectal cancer cells, Cancer Lett. 365 (1) (Aug. 2015) 68–78, https://doi.org/10.1016/j.canlet.2015.05.015.

[23] P. Eilers, H. Boelens, Baseline correction with asymmetric least squares smoothing, Leiden Univ, Med. Cent. Rep. 1 (1) (2005) 5.

[24] D.A. Whitaker, K. Hayes, A simple algorithm for despiking Raman spectra, Chemometr. Intell. Lab. Syst. 179 (Aug. 2018) 82–84, https://doi.org/10.1016/j.chemolab.2018.06.009.

[25] Brozek-Pluska, Musial, Kordek, Abramczyk, Analysis of human colon by Raman spectroscopy and imaging-elucidation of biochemical changes in carcinogenesis, Int. J. Mol. Sci. 20 (14) (Jul. 2019) 3398, https://doi.org/10.3390/ijms20143398.

[26] B. Brozek-Pluska, A. Dziki, H. Abramczyk, Virtual spectral histopathology of colon cancer - biomedical applications of Raman spectroscopy and imaging, J. Mol. Liq. 303 (Apr. 2020) 112676, https://doi.org/10.1016/j.molliq.2020.112676.

[27] Z. Huang, et al., In vivo detection of epithelial neoplasia in the stomach using image-guided Raman endoscopy, Biosens. Bioelectron. 26 (2) (Oct. 2010) 383–389, https://doi.org/10.1016/j.bios.2010.07.125.

[28] S. Xie, Feature extraction of auto insurance size of loss data using functional principal component analysis, Expert Syst. Appl. 198 (Jul. 2022) 116780, https://doi.org/10.1016/j.eswa.2022.116780.

[29] N. Jafarzadeh, A. Mani-Varnosfaderani, K. Gilany, S. Eynali, H. Ghaznavi, A. Shakeri-Zadeh, The molecular cues for the biological effects of ionizing radiation dose and post-irradiation time on human breast cancer SKBR3 cell line: a Raman spectroscopy study, J. Photochem. Photobiol., B 180 (Mar. 2018) 1–8, https://doi.org/10.1016/j.jphotobiol.2018.01.014.

[30] Md A. Talukder, Md M. Islam, M.A. Uddin, A. Akhter, K.F. Hasan, M.A. Moni, Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning, Expert Syst. Appl. 205 (Nov. 2022) 117695, https://doi.org/10.1016/j.eswa.2022.117695.