photonics

Article

Classifying Raman Spectra of Colon Cells Based on Machine Learning Algorithms

Maria Lasalvia, Crescenzio Gallo, Vito Capozzi and Giuseppe Perna *

Department of Clinical and Experimental Medicine, University of Foggia, 71122 Foggia, Italy; maria.lasalvia@unifg.it (M.L.); crescenzio.gallo@unifg.it (C.G.); vito.capozzi@unifg.it (V.C.)
* Correspondence: giuseppe.perna@unifg.it

Abstract: Colorectal cancer is very widespread in developed countries. Its diagnosis partly depends on pathologists' experience and their laboratories' instrumentation, producing uncertainty in diagnosis. The use of spectroscopic techniques sensitive to the cellular biochemical environment could aid in achieving a reliable diagnosis. So, we used Raman micro-spectroscopy, combined with a spectral analysis by means of machine learning methods, to build classification models, which allow colon cancer to be diagnosed in cell samples, in order to support such methods as complementary tools for achieving a reliable identification of colon cancer. The Raman spectra were analyzed in the 980–1800 cm^{-1} range by focusing the laser beam onto the nuclei and the cytoplasm regions of single FHC and CaCo-2 cells (modelling healthy and cancerous samples, respectively) grown onto glass coverslips. The comparison of the Raman intensity of several spectral peaks and the Principal Component Analysis highlighted small biochemical differences between healthy and cancerous cells mainly due to the larger relative lipid content in the former cells with respect to the latter ones and to the larger relative amount of nucleic acid components in cancerous cells compared with healthy ones. We considered four classification algorithms (logistic regression, support vector machine, k nearest neighbors, and a neural network) to associate unknown Raman spectra with the cell type to which they belong. The built machine learning methods achieved median values of classification accuracy ranging from 95.5% to 97.1%, sensitivity values ranging from 95.5% to 100%, and specificity values ranging from 93.9% to 97.1%. The same median values of the classification parameters, which were estimated for a testing set including unknown spectra, ranged between 93.1% and 100% for accuracy and between 92.9% and 100% for sensitivity and specificity. A comparison of the four methods pointed out that k nearest neighbors and neural networks better perform the classification of nucleus and cytoplasm spectra, respectively. These findings are a further step towards the perspective of clinical translation of the Raman technique assisted by multivariate analysis as a support method to the standard cytological and immunohistochemical methods for diagnostic purposes.

Keywords: colon cells; Raman spectroscopy; machine learning algorithms

Citation: Lasalvia, M.; Gallo, C.; Capozzi, V.; Perna, G. Classifying Raman Spectra of Colon Cells Based on Machine Learning Algorithms. *Photonics* **2024**, *11*, 275. https://doi.org/10.3390/photonics11030275

Received: 15 February 2024
Revised: 14 March 2024
Accepted: 15 March 2024
Published: 21 March 2024

1. Introduction

Nearly 2 million new cases of colorectal cancer were diagnosed in 2020; they contributed to 10.7% of total cancer cases according to the data from the World Cancer Research Fund International [1]. The gold standard procedure for colon cancer screening and diagnosis is mainly based on colonoscopy, followed by histopathological examinations of biopsy samples taken from patients. Indeed, performing colon cancer diagnosis by means of quantifying tumor markers in the blood, such as the carcinoembryonic antigen (CEA), sometimes provides unreliable results because some healthy people have high blood levels of CEA, especially if they are smokers [2]. A histopathologist makes a diagnosis according to the morphological characteristics of the cells and lesions present in the biopsy specimen after specific staining of the same specimen has been carried out. Therefore, the expertise and ability of a physician can influence the sensitivity and specificity of a cancer diagnosis.

Because the evolution of a healthy cell into a cancerous one requires changes at the biochemical level, it is of interest to confirm the diagnostic results of a histological analysis with the results obtained by chemical analytical techniques. Raman spectroscopy is a vibrational technique that is suitable to detect biochemical modification occurring as a consequence of cancer onset at the cellular level [3]. In recent years, many authors reviewed Raman spectroscopy applications for cancer detection in human breast [4], brain [5], lung [6,7], and gastric [8,9] tissues and cells. Although promising results have been obtained in cancer detection, Raman spectroscopy remains, at present, only a technique to assist standard histopathological techniques. In fact, much work remains to be conducted to implement this spectroscopic technique in diagnosis and screening activities.

Raman spectra provide information about the biochemical differences between cancer cells and healthy cells. However, such spectral differences in many cases are only slightly larger than the signal-to-noise ratio, so the spectra measured for cells of unlike types almost overlap. In this case, it is necessary to process the spectral data by means of proper analysis techniques in order to allow useful information to be extracted. In particular, multivariate statistical methods have been recently used to classify Raman spectra [10–12].

The classification problem consists of the assignment to the right class of a given spectrum, which is measured on a sample whose class is unknown. Recently, W. Wang et al. reported an interesting method, based on Raman spectra measured on tissue biopsies, that is able to identify the prognoses of patients suffering from gastric cancer [13]. This method is based on the estimation of the Euclidean distance of various Raman spectra with respect to a poor prognosis reference spectrum. The good results obtained (sensitivity of 75% and specificity of 96.8%) refer to a situation in which the average Raman spectra of the favorable and poor prognosis groups are quite different from each other.

In the case of spectra that are very similar to each other but belong to different classes, the classification problem can be solved using machine learning methods. They are mathematical methods that first analyze spectra whose classes they belong to are known in order to obtain the main spectral characteristics that distinguish the different classes, and then apply these characteristics to classify spectra whose classes are not known. Some of these methods classify the spectra according to a limited number of wavenumber values and the corresponding spectral intensities; therefore, they require a careful feature selection step. Nonetheless, several of them are freely available online. As an example, the "Orange" software product (https://orangedatamining.com/, accessed on 31 January 2024) comprises many machine learning methods [14], such as neural networks (NNs), support vector machines (SVMs), k nearest neighbors (kN N), and logistic regression (LR).

NNs are algorithms that acquire classification ability by means of many mathematical functions (artificial neurons) arranged in layers; each neuron of a layer receives input data and provides output data to the neurons of the next layer until it reaches the output layer, which provides the response about the classification of the input data [15]. Recently, D. Kalatzis et al. applied a type of NN method, known as a convolutional NN (CNN), for the classification of the Raman spectra of colon tissues; they found that the CNN algorithm achieved an accuracy of 83.4% and a sensitivity of 85.9% in the classification of a spectral dataset including 248 spectra measured in the 800–1800 cm^{-1} range [16]. Also, H. Yan et al. obtained excellent classification accuracy (97.2%), sensitivity (99.1%), and specificity (95.4%) in the discrimination of tongue squamous cell carcinoma from adjacent non-tumorous tissues with Raman spectroscopy and a CNN [17]. As for cell samples, W. Shuyun et al. reported an accuracy mean of 99.2 $\pm$ 5.1%, a sensitivity mean of 99.2 $\pm$ 5.1%, and a specificity mean of 99.8 $\pm$ 1.0% for the classification of different kinds of liver cancer cell lines by means of laser tweezer Raman spectroscopy combined with a deep neural network [18].

The SVM algorithm uses the input data, belonging to known classes, to identify a hyperplane in the space of the selected spectral features, which optimize the separation of data belonging to different classes; next, the projection of unknown data onto the hyperplane allows them to be classified correctly [10,15]. The SVM algorithm was able

to detect colorectal cancer by analyzing the Raman spectra of blood serum samples from 75 patients; in particular, the analysis of 43 properly selected spectral features in the 800–3000 cm^{-1} range allowed the investigated samples to be classified with accuracy, sensitivity, and specificity values of 96%, 93%, and 98%, respectively [19]. X. Fang et al. implemented the surface enhanced Raman technique with SVM method to distinguish lung cancer cells from normal cells including blood cells and immortalized lung cells; they achieved a classification accuracy between 98.8% and 100% for differentiation of cancer cells from normal ones [20].

The kNN algorithm classifies unknown data according to the classes of their nearest neighbors. Specifically, first, an appropriate value of the number k is chosen, and then an unknown spectrum is assigned to the class to which the k closest spectra in the calibration set belong [15]. X. Li et al. developed an algorithm for colon cancer diagnosis based on the kNN method applied to Raman spectra measured for serum taken from 75 healthy volunteers, 65 colon cancer patients, and 60 postoperative colon cancer patients; an accuracy of 91.0% and a specificity of 92.6% were achieved [21]. X. Wang et al. investigated the feasibility of using Raman spectroscopy combined with kNN to discriminate between healthy volunteers, breast cancer, and ductal carcinoma in situ (DCIS); the kNN method applied to Raman spectra collected from the serum of 241 healthy volunteers, including 463 patients with breast cancer and 100 DCIS patients, achieved an accuracy of 78.93%, while larger accuracy values were obtained for binary classifications [22].

LR is an algorithm, mainly used for binary classification, based on a logistic function (also known as sigmoid function) whose parameters are optimized during the calibration phase; then, unknown data are classified according to the value of sigmoid function and calculated with the optimized parameters [23]. In a work several years ago, S.K. Teh et al. achieved accuracies of 92% and 94% for tissue classification of gastric adenocarcinomas of intestinal and diffuse type, respectively, by the LR method applied to Raman spectra of such tissues [24]. More recently, L.A. Arevalo et al. successfully discriminated vibrational spectra of cerebral–spinal fluid from healthy and Alzheimer's patients by the LR method [25].

Recently, we differentiated healthy colon cells from cancerous cells by analyzing FTIR spectra and applying machine learning methods. In particular, the NN algorithm was very effective in discriminating the two cell types, with excellent accuracy, sensitivity, and specificity [26]. The aim of this work is to investigate which of the above machine learning algorithms (NN, SVM, kNN, and LR) allow reliable classification of Raman spectra measured in the nucleus and cytoplasm region of healthy and cancerous colon cells. We found that all algorithms can discriminate the Raman spectra from the two classes of spectra with accuracy, sensitivity, and specificity values larger than 92%. In particular, excellent accuracies were obtained for the classification of nucleus spectra by the kNN method and cytoplasm spectra by the NN method.

2. Materials and Methods

2.1. Cell Growth

Fetal human colon (FHC) cell line was used as a model of healthy colon cells. DMEM F12 was used as the growth medium, to which 10 mM Hepes, 10 ng/mL cholera toxin, 5 µg/mL insulin, 5 µg/mL transferrin, 100 ng/mL hydrocortisone, 20 ng/mL EGF, and fetal bovine serum with a 10% final concentration were added. The human colorectal adenocarcinoma (CaCo-2) cell line was used as a colon cancer cell model. CaCo-2 cells were grown in Dulbecco's Modified Eagle's medium (DMEM), supplemented with 4 $mmol/dm^3$ L-glutamine, 1% penicillin/streptomycin, 10% fetal bovine serum (FBS), and 1% non-essential amino acids (NEAA) at 37 °C and 5% CO_2. Both cell lines were purchased from ATCC (Manassas, VA, USA).

The cells were allowed to adhere to a glass coverslip, and after that, a proper poly-lysine coating was deposited on the glass surface. Both healthy and cancer cells were fixed in 3.7% paraformaldehyde and stored in Petri dishes with phosphate-buffered saline (PBS)

solution until Raman spectra acquisition. Each cell sample was rinsed in deionized water to remove residual PBS before Raman measurements.

2.2. Raman Spectra

Raman spectra were measured with a Raman confocal micro-spectrometer (Labram from Jobin-Yvon Horiba, Montpellier, France), using an Olympus 100x oil-immersion objective, in the range 600–1800 cm^{-1}. The 514.5 nm line of an Ar ion laser was used to excite two different positions of single FHC and CaCo-2 cells, i.e., a cell volume that includes the nucleus and a cell volume that excludes it (and includes only the cytoplasm region). The diffraction-limited spot focused on the sample had a diameter of less than 1 μm. In particular, before measuring each Raman spectrum, an optical image of the single cell to be measured was obtained using a charge-coupled device camera, in order to select the cell compartment on which to focus the laser beam and from which to collect the signal. Each measured cell was excited with a laser power of 6 mW. The spectrum obtained from each single cell consisted of the average signal of three consecutive acquisitions of 10 s each. About 50 randomly chosen cells were measured, both for healthy and cancer samples. The backscattered Raman light was analyzed by a diffraction grating with 600 grooves/mm and it was detected by a charge-coupled device. The spectral resolution was about 5 cm^{-1}/pixel. The background signal was measured within volumes where no cell was located.

2.3. Spectral Processing and Data Analysis

Each Raman spectrum was preprocessed by first subtracting the corresponding background signal and then performing a subtraction of the cell fluorescence and stray light signal by means of the adaptive algorithm of the Spectragryph software (version 1.2.16.2023) [27], with a coarseness value of 30. That algorithm creates a baseline that fits to the lower part of the spectra to remove the underlying broad and featureless signals while keeping actual peaks. Next, area normalization was performed with the goal of normalizing each spectrum to the total amount of biological material in the sampling volume. In particular, the background spectrum, measured in a region of cell-free coverslip, is due to the Raman signal of the glass coverslip and water, and it is mainly characterized by a strong band centered at about 940 cm^{-1}, related to the glass coverslip [28]; this band cannot be fully removed by means of the background subtraction. Therefore, because the presence of this spurious and cell-to-cell variable signal could affect the results of area normalization, each single spectrum was normalized only in the 980–1800 cm^{-1} spectral range. Finally, each spectrum was smoothed with a Savitzky–Golay filter with interval size 5 and polynomial order 3 using Spectragryph software [27]. These values were optimized so that the intensity ratios of the peaks in each spectrum did not vary drastically.

Exploratory data analysis was performed by Principal Component Analysis (PCA) with Unscrambler X (CAMO software, version 10.4), in order to visualize the discrimination of the two types of samples (in the score plots) and the spectral variables to which this separation is related (in the loading plots). A full cross-validation was used to validate the PCA results.

After comparison of the mean spectra and PCA analysis, each of the sets of the healthy and cancerous spectra was divided to obtain two subsets: a calibration set (70% of the whole set from each cell type) for training the algorithms and a test set (the remaining 30% of the whole set from each cell type) for testing them. The calibration and test sets were obtained for both the sampled regions of nucleus and cytoplasm. A random number generator was used to select the spectra of the calibration set. Four classification models included in the Orange software 3.35.0 were used for training: LR, SVM, kNN, and NN. During the training step, the values of some parameters that the algorithms use to perform the learning process were optimized in order to obtain maximum accuracy values and full cross-validation was used to validate the obtained results. The random selection of the calibration set spectra from the whole spectral dataset measured in both the nucleus

and cytoplasm regions was repeated 25 times. The classification parameters were reported according to median values.

3. Results and Discussion

The area-normalized Raman spectra of healthy FHC and cancerous CaCo-2 cells are shown in Figure 1, where the mean (continuous lines) and standard deviation (dashed lines) spectra are plotted for Raman signals detected within a cell region comprising the nucleus (a) and a cell region including only the cytoplasm (b). In particular, when the laser beam is focused on the nucleus region, Raman signals are mainly due to nucleus constituents (nucleic acids, DNA/RNA, and proteins) and less to components of the plasmatic membrane (mainly proteins and lipids); when the laser beam is focused within the cytoplasm, the components of the nucleus are not sampled. The spectra of CaCo-2 cells have been shifted vertically in Figure 1 for the sake of clarity. We measured the Raman spectra of these two cell regions for each cell to test to determine whether one of these two regions is more suitable than the other for effective discrimination of cancerous cells from healthy cells. The standard deviation values in Figure 1 indicate that the Raman signals of healthy cells are more variable than those of cancerous cells; this points out that healthy cells are characterized by larger differences in the relative content of cellular constituents than cancerous ones.

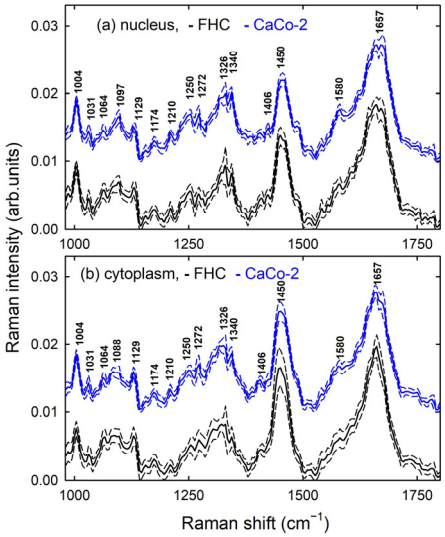

Figure 1. Mean Raman spectra of healthy FHC (continuous black line) and cancerous CaCo-2 (continuous blue line) cells after area normalization. The spectra were measured within a cell volume including the nucleus (**a**) and including only the cytoplasm (**b**). Standard deviation spectra are shown as dashed lines. The spectra have been vertically shifted for clarity. The labels indicate the spectral position of the main Raman features.

Similar spectra have been reported by other authors about colon cells [29,30] and tissues [31]. In particular, all spectra in Figure 1 show many peaks, due to Raman scattering of distinct functional groups located within the cellular components. The assignment of spectral features, in agreement with the results of the previous literature [32], is shown in Table 1. In particular, the most intense peaks in the spectra in Figure 1 are due to the contribution of the amide I (~1657 cm^{-1}) and CH_2 deformation (~1450 cm^{-1}) peaks. Other well-resolved protein-related peaks include amide III (~1260 cm^{-1}), phenylalanine ring breathing vibrations (1004 cm^{-1}), C-N stretching (1088 and 1129 cm^{-1}), and aromatic ring vibrations associated with phenylalanine, tryptophan, and tyrosine (e.g., 1031, 1174, and 1210 cm^{-1}). The contribution of DNA and RNA components is more evident in the spectra measured in the nucleus region than in the cytoplasm one; it is mostly related to the peaks at 1097 cm^{-1} (PO_2^- phosphodioxy bond of the phosphate group), 1326 and 1340 cm^{-1}

(CH_3CH_2 wagging mode in purine bases of DNA and ring breathing modes of DNA/RNA bases, respectively), and 1580 cm^{-1} (ring breathing vibrational modes characteristic of adenine and guanine). Lipid components weakly contribute to the Raman spectra in Figure 1. In particular, a well-resolved lipid-related peak is located at 1064 cm^{-1}, where other lipid peaks are mainly overlapped with the protein ones, as occurs at 1088, 1129, and 1450 cm^{-1}.

Table 1. Attribution of Raman bands, according to previous literature results [32] and in the present investigation. Abbreviations: p: proteins; l: lipids; n.a.: nucleic acids.

Spectral Position (cm^{-1})	Assignment
1004	C-C symmetric ring breathing of Phenylalanine (p.)
1031	C-H in plane bending of Phenylalanine (p.)
1064	C-C stretching (l.)
1088	C-N stretching (p.) and C-C stretching (l.)
1097	Symmetric PO_2^- stretching of DNA (n.a.)
1129	C-N stretching (p.), C-O stretching (c.), C-C stretching (l.)
1174	C-H bending aminoacids (p.)
1210	C-C_6H_5 stretching aminoacids (p.)
1250	Amide III (p.)
1272	Amide III (p.)
1326	CH_3CH_2 wagging mode in purine bases of DNA (n.a.)
1340	Ring breathing modes of DNA bases (n.a.)
1406	$(C=O)O^-$ stretching of aminoacids (p.)
1450	CH_2 bending modes (p., l.)
1580	Ring breathing modes in DNA bases (n.a.)
1657	Amide I (p.)

The differences of mean spectra, shown in Figure 2a,b for the nucleus and cytoplasm regions, respectively, provide preliminary information on the relative content of the cell components in the two cell types. In fact, the positive peaks in Figure 2 represent a larger relative content of the peak-related components in the healthy cells than in the cancer cells, whereas the opposite occurs for negative peaks. Thus, the positive peaks at about 1060 and 1450 cm^{-1} suggest a larger relative amount of lipids in healthy cells than in cancerous ones. Negative signals at about 1338 and 1580 cm^{-1} also indicate a larger relative content of nucleic acid components in the cancerous cells with respect to the healthy cells. Figure 2 shows larger uncertainty about relative protein content; in fact, positive peaks at about 1660 cm^{-1} suggest larger protein content in healthy cells than in cancerous cells, whereas negative signals in the 1200–1280 cm^{-1} range indicate the opposite.

To test whether the intensity values of individual spectral peaks in Figure 1 were sufficient to discriminate healthy cells from cancerous ones, we compared the intensity values of specific peaks for the two cell types. Figure 3 shows some of these comparisons. In particular, the Raman intensities of the DNA-related peak at about 1340 cm^{-1} in the spectra measured including the nucleus region are quite larger for cancerous cells than for healthy ones, as visible in Figure 3a. Such behavior suggests that cancerous cells contain a larger relative amount of DNA than healthy cells. Such an observation is in agreement with the results reported a few years ago by M.V.P. Chowdary et al., who found that mean Raman spectra of malignant tissue exhibit relatively stronger DNA bands (at 1340 and 1470 cm^{-1}) than mean Raman spectra of normal tissue [33]. Recently, M. Fouskova et al. also reported Raman spectra of normal colorectal mucosa, benign epithelial polyps, and colorectal adenocarcinoma [34]. In particular, they found a larger relative intensity of

the Raman band at about 1333 cm^{-1} (partly due to vibrational modes of DNA bases) in the spectra of adenocarcinoma than in the other tissues. Such a result is confirmed by B. Brozek-Pluska et al. regarding the DNA-related peaks located at 1342 and 1584 cm^{-1} in Raman spectra of colon tissue and colon cells [31]. However, although the statistical difference between the distributions of intensity values in the two types of cells obtained in Figure 3a is significant (as indicated by the box plot on the right side), the separation was not clear-cut, and several Raman intensity values were close to each other for the two groups.

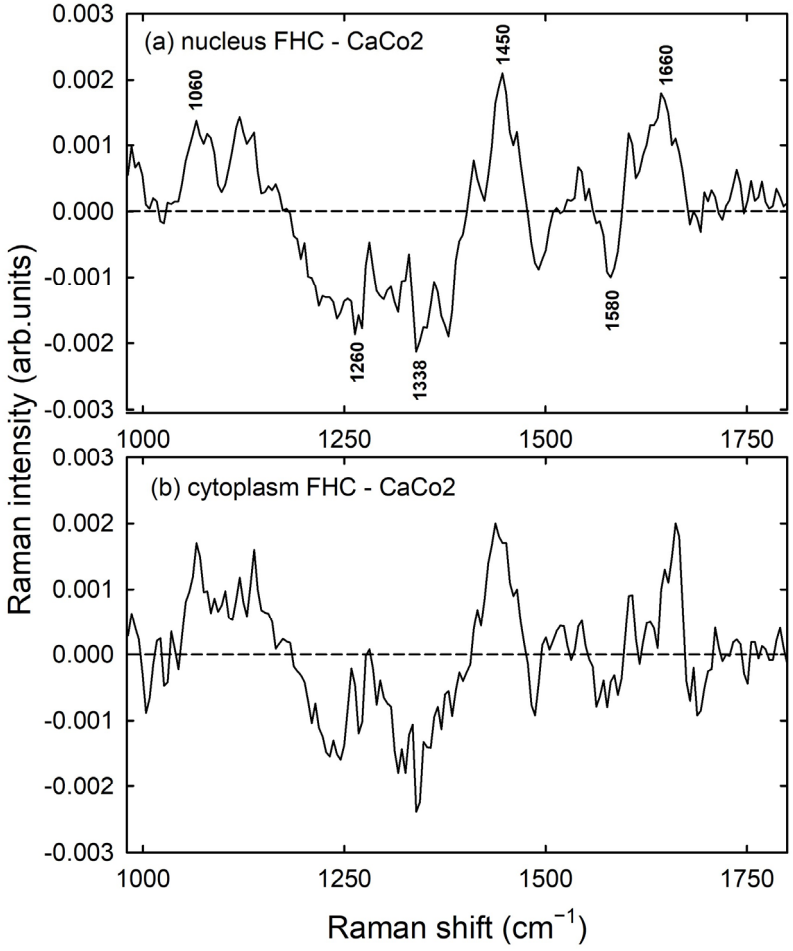

Figure 2. Difference of mean values (healthy–cancerous signal) of the spectra measured within a cell volume including the nucleus (**a**) and including only the cytoplasm (**b**). The labels refer to the spectral position of the main features of the difference Raman spectrum.

Regarding the relative protein content in the two types of cells, contradictory results were obtained from the comparison of peak intensities, as discussed above and as visible in Figure 3b,c. Indeed, Figure 3b suggests a large relative content of proteins in the spectra of cancerous cells compared with those of healthy cells, as the peak at 1450 cm^{-1} is mainly due to proteins for spectra measured in the nucleus region. In contrast, Figure 3c, which shows the comparison of amide I peak intensity values, indicates a larger intensity of such a peak in the healthy cells than in the cancerous ones. These results are in disagreement with similar spectra measured on the nucleus region of normal and cancerous colon cells by other authors [35]. In particular, they found larger intensity values for the Raman peak centered at 1444 cm^{-1} in the spectra of normal cells than in cancerous ones, whereas they measured similar intensity values for the amide I peak in both cell types. Although a statistically significant difference was obtained, as visible in the plots on the right side of Figure 3, in this case, the intensity values in Figure 3b,c are also largely overlapping.

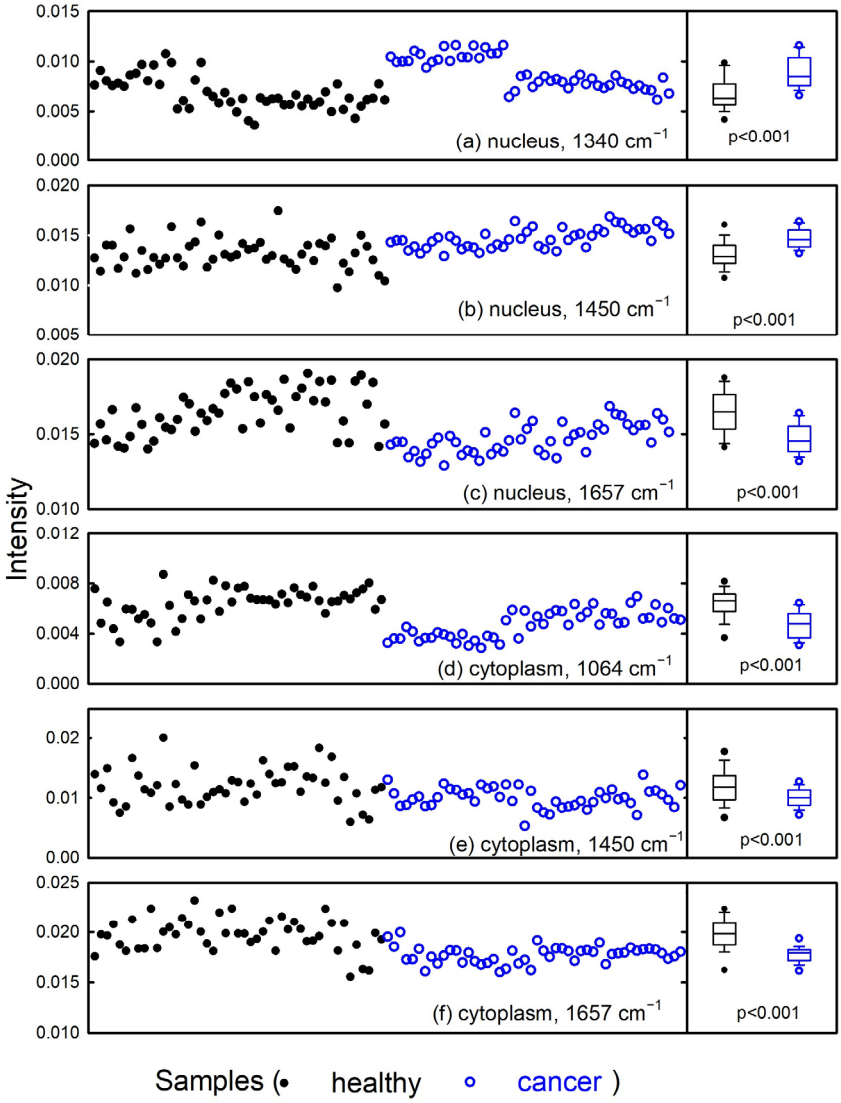

Figure 3. Distribution of intensity values of some spectral peaks due to the nucleic acid (**a**), protein (**b**,**c**,**f**), and lipid (**d**,**e**) components of healthy (black filled circles) and cancerous (blue hollow circles) colon cells. The corresponding box plots of each distribution are shown on the right-hand side.

As for the Raman spectra measured within the cytoplasm region, a larger amount of lipid content in the healthy cells is suggested in Figure 3d, which shows the comparison of peak intensity values at about 1064 and 1450 cm^{-1}, related to C-C stretching and CH_2 bending modes of lipids, respectively. The larger amount of lipid components in healthy compared with cancerous cells has been previously reported by us for similar cells using FTIR measurements [26], as also measured by Dong et al. [36] and E. Kaznowska et al. [37] for colon tissues. However, Brokek-Pluska et al. reported discordant results from those in Figure 3e regarding the 1444 cm^{-1} peak in tissue colon samples [35]. In addition, these authors reported larger intensity values of the Raman signal at 1655 cm^{-1}, due to protein components, in the spectra measured within the cytoplasm region of cancer colon cells than those measured within the cytoplasm of normal colon cells [35]. This protein-related result is also different from that obtained and shown in Figure 3f.

Overall, the intensity values of single spectral peaks cannot be considered as markers of colorectal cancer, because the differences between such values for all the cell constituents in the Raman spectra of normal and cancerous cells are too subtle and overlapping each other; thus, a reliable cancer diagnosis cannot be achieved. Therefore, it is worth considering multivariate analysis techniques (in which the intensity values of many Raman peaks are

simultaneously analyzed, as machine learning methods) to test the possibility of classifying unknown cells as healthy or cancerous samples, according to their Raman spectrum.

Before using machine learning algorithms to evaluate their performance for the classification of unknown cells, we performed the PCA technique as a preliminary step to determine whether spectra belonging to two different groups can be discriminated based on the simultaneous contribution of different spectral intensity values. Figure 4a,b show the PCA score plots for Raman spectra measured within the nucleus and cytoplasm, respectively. The spectra of cancerous cells are well-discriminated from those of healthy cells according to the PC1 and PC3 score values of the nucleus and cytoplasm spectra, respectively. In fact, the spectra of cancerous cells are mainly characterized by negative values of PC1 and PC3 scores, whereas the spectra of healthy cells have mainly positive values of these PC scores. The loading 1 and 3 plots for the nucleus and cytoplasm regions are shown in Figure 3c and 3d, respectively. They should be compared with the difference signals of the average spectra in Figure 2, in order to understand which Raman peaks are related to the discrimination between the two types of spectra. The loading plots in Figure 4 and the difference plots in Figure 2 are quite similar to each other. Therefore, the loading plots confirm that the discrimination of healthy from cancerous cells occurs mainly for the different relative content of lipids and nucleic acids; the content of the former is larger in healthy cells, whereas the content of nucleic acids is larger in cancerous cells.

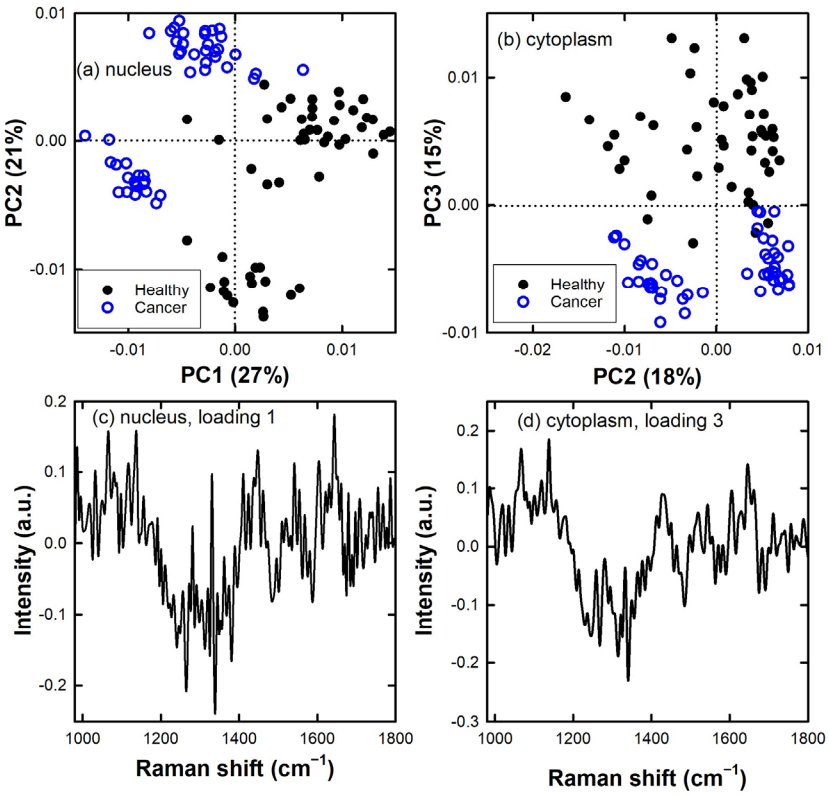

Figure 4. PC2 vs. PC1 (**a**) and PC3 vs. PC2 (**b**) score plot for the healthy FHC cell (black filled circles) and cancerous CaCo-2 cell (blue hollow circles) Raman spectra. Difference of mean spectrum values (healthy–cancer signal) and loading 1 and loading 3 spectra are also reported for the spectra measured within a cell volume including the nucleus (**c**) and including only the cytoplasm (**d**).

Overall, the PCA results demonstrate the possibility of assigning spectra of an unknown nature to the appropriate class by suitable classification techniques, based on the simultaneous use of many Raman peaks as a spectral biomarker for malignancy. Therefore, we estimated the results obtained by training four classification algorithms (kNN, LR, SVM, and NN) from the "Orange" software, for both the nucleus and cytoplasm region. The

spectral features selected for the classification were those corresponding to the spectral positions of the main Raman peaks labelled in Figure 1. The only difference between the features selected for the classification of the spectra acquired in a volume including the nucleus compared to the spectra measured in a volume excluding the nucleus was the choice of the spectral feature located at 1097 cm^{-1} in the former ones and of the feature located at 1088 cm^{-1} in the latter ones. For each algorithm, the parameters controlling the learning process turn oute to be optimized when they assume the following values:

- LR: non-regularization type;
- SVM: radial basis function (RBF) kernel, SVM with cost 1.0 and regression loss epsilon 0.1, tolerance 0.001, and maximum 100 iterations;
- kNN: the number of neighbors equal to two, by using an Euclidean metric and weights by distances;
- NN: 30 neurons in the hidden layer, ReLu activation, Adam solver, and 300 maximum iterations.

The use of different machine learning methods is spreading in medical diagnostics for the classification of cell and tissue samples according to optical [38], electrical [15], and spectroscopic [39] data. The values of the performance parameters achieved by the classification algorithms during the training of the calibration data, randomly chosen 25 times from the original dataset, are reported in Table 2. All the mentioned algorithms accomplished a good classification level, with accuracy values larger than 95% for both cell regions. In particular, the SVM algorithm performs slightly better than the others, with an accuracy value of about 97% for the two cell regions. Note that the classification target was set to detect the cancerous cells; consequently, healthy cells were assessed as negative and cancerous cells as positive.

Table 2. Median values of performance parameters for the investigated classification algorithms applied to the calibration set of Raman spectra of healthy and cancerous colon cells. The parameters are reported for spectra measured with the laser spot focused on both the nucleus region and the cytoplasm region. The 25th and 75th percentile values are reported between brackets for each parameter.

Method	Accuracy Nucleus (%)	Accuracy Cytoplasm (%)	Sensitivity Nucleus (%)	Sensitivity Cytoplasm (%)	Specificity Nucleus (%)	Specificity Cytoplasm (%)
kNN	97.1 (95.7, 97.9)	95.5 (94.0, 96.3)	100.0 (97.1, 100.0)	95.5 (94.1, 97.1)	94.3 (94.3, 97.1)	93.9 (90.9, 97.0)
LR	95.7 (95.7, 97.1)	97.0 (97.0, 97.8)	97.10 (95.7, 97.1)	97.1 (97.1, 98.6)	97.1 (94.3, 97.1)	97.0 (97.0, 97.0)
NN	95.7 (94.2, 97.1)	97.0 (97.0, 98.5)	97.1 (94.1, 100.0)	97.1 (97.1, 100.0)	94.3 (91.4, 95.7)	97.0 (97.0, 97.0)
SVM	97.1 (95.7, 97.1)	97.0 (95.5, 97.8)	97.1 (97.1, 100.0)	97.1 (97.1, 97.1)	94.3 (94.3, 97.1)	97.0 (93.9, 100.0)

Some works have been recently published reporting the comparative analysis of several machine learning techniques applied to Raman spectra measured for tissue samples, with the aim of promoting the adoption of the Raman technique coupled with multivariate methods in colon cancer diagnostics. In particular, M. Fouskova et al. [34], using several methods of machine learning (PCA-Linear Discriminant Analysis, SVM, Decision Tree, and Decision Tree AdaBoost), were able to achieve more than 99% accuracy in distinguishing colorectal lesions from healthy epithelial tissue. However, their results refer to the 10-fold cross-validation of the whole spectral dataset. In addition, J. Depciuch et al. [19] analyzed Raman data collected from 75 blood serum samples of healthy and colon cancer patients using three machine learning methods (Deep Learning, SVM, and eXtreme Gradient Boosting trees) in order to determine the efficiency of discrimination of sick and healthy

people. They obtained accuracy values ranging from 59% to 96%, and both depended on the investigated spectral range and on the number of selected spectral features. The eXtreme Gradient Boosting trees method performed better than the other two methods, although these authors also used the whole dataset to build the classification models, which were tested with a leave-one-out cross-validation approach. As a third example, N. Blake et al. [40] investigated the potential of using Raman spectroscopy to distinguish between normal cells and adenocarcinoma in human colorectal tissue samples. In particular, they obtained discrimination accuracy values between 71% and 75% by means of PCA-LDA, SVM, and CNN. However, even in this case, the validation step was carried out using a cross-validation method rather than using an independent test set.

In contrast to these authors' methods, we performed LOOCV only during the building step of the classification models, using the Raman spectra of the calibration set, in order to generalize and optimize such models. Next, we tested the machine learning algorithms on a set of unknown Raman spectra (test set) in order to assess the ability of the investigated machine learning models to classify colon cells. The achieved values of the performance parameters are reported in Table 3; they refer to the median values calculated for 25 instances of randomly chosen spectra of the test set from the original dataset. In particular, the area under the ROC curve (AUC) values, which can vary between 0 and 1, are proportional to the model's ability to distinguish the unknown spectra as belonging to the proper class between the healthy and cancerous cells [41]. The AUC values in Table 3 show that all the investigated models have excellent ability to distinguish between cells with disease and without disease. In addition, all the obtained values of accuracy were almost excellent, particularly for the kNN method for the Raman spectra of the nucleus region and NN method for the Raman spectra of the cytoplasm region. Instead, the obtained sensitivity and specificity values of all models suggest that Raman spectra measured in the cytoplasm region perform better with respect to those measured in the nucleus region in reducing the risk of failing to diagnose the pathology or of misdiagnosing it.

Table 3. Median values of performance parameters for the investigated classification algorithms applied to the test set of Raman spectra of healthy and cancerous colon cells. The parameters are reported for spectra measured with the laser spot focused on both the nucleus region and the cytoplasm region. The 25th and 75th percentile values are reported between brackets for each parameter.

Method	AUC Nucleus	AUC Cytoplasm	Accuracy Nucleus (%)	Accuracy Cytoplasm (%)	Sensitivity Nucleus (%)	Sensitivity Cytoplasm (%)	Specificity Nucleus (%)	Specificity Cytoplasm (%)
kNN	1.00 (0.97, 1.00)	1.00 (0.96, 1.00)	100.0 (96.6, 100.0)	96.4 (92.9, 96.4)	100.0 (100.0, 100.0)	100.0 (92.9, 100.0)	93.3 (93.3, 100.0)	92.9 (92.9, 100.0)
LR	1.00 (0.99, 1.00)	1.00 (0.99, 1.00)	96.6 (93.1, 96.6)	96.4 (94.7, 100.0)	92.9 (89.3, 100.0)	100.0 (92.9, 100.0)	100.0 (93.3, 100.0)9	100.0 (92.9, 100.0)
NN	0.99 (0.99, 1.00)	1.00 (0.99, 1.00)	93.1 (89.7, 96.6)	100.0 (96.4, 100.0)	92.9 (85.7, 100.0)	100.0 (96.5, 100.0)	93.3 (86.7, 96.7)	100.0 (92.9, 100.0)
SVM	1.00 (0.99, 1.00)	1.00 (0.99, 1.00)	96.6 (93.1, 100.0)	96.4 (92.9, 100.0)	100.0 (92.9, 100.0)	100.0 (92.9, 100.0)	100.0 (93.3, 100.0)	100.0 (92.9, 100.0)

4. Conclusions

In summary, we proposed the use of Raman spectroscopy combined with machine learning methods to obtain reliable classification of Raman spectra measured in the nucleus and cytoplasm regions of healthy and cancerous colon cells. First, we randomly selected 70% of the Raman spectra from the whole dataset to form a calibration set to be used to optimize the machine learning parameters. This random selection was repeated 25 times. The remaining 30% of the Raman spectra constituted the test set, to be used to evaluate the performance of the algorithms used. We found a classification accuracy > 93% for the spectra measured in the two cell regions, with the cytoplasm region performing slightly better on average compared to the nucleus region. The sensitivity and specificity values estimated from the Raman spectra measured in the cytoplasm region were also better

on average than those measured in the nucleus region. About the performances of the four machine learning methods, excellent accuracies were obtained for the classification of nucleus spectra with the kNN method and cytoplasm spectra with the NN method.

Although these findings suggest that the subcellular Raman analysis approach combined with machine learning analysis might be a powerful tool to improve cancer diagnosis during clinical examination, some critical issues should be overcome before the Raman measurements and machine learning analysis could be accepted in clinical practice. First, this investigation is only a proof of suitability of the presented diagnostic analysis, because it involves cell lines; hence, the results should be confirmed by measuring and analyzing cytological specimens from patients. In addition, tissue biopsy samples from patients should be investigated too, because histological analysis is as widespread as cytological analysis. In the case of tissue samples, the distributions of the performance parameter values are expected to be broader than those related to cell samples, due to the greater variability of biochemical content present in tissues than in cells. Lastly, the achieved results should be confirmed regarding samples characterized by different degrees of pathology. Nonetheless, our findings are promising in terms of using the vibrational spectroscopy and machine learning algorithms as useful methods in cytology diagnostics.

Despite these promising results, some issues still need to be addressed before Raman spectroscopy and machine learning techniques can be effectively implemented in clinical settings. The first issue concerns the standardization of the measurement protocol, specifically about the preparation of the sample to be measured (for which proper substrates should be evaluated), the choice of the best laser wavelengths (in order to optimize the signal/noise ratio of the spectra), and the planning of the number of measurements to be carried out (in order to obtain reliable results). Then, the preprocessing methods of the spectra should be standardized, by properly choosing the outlier spectra to be removed from the subsequent analysis, the method of subtracting the baseline related to the substrate signal and the fluorescence of the sample, and finally, the normalization method of the spectra. Finally, it is necessary to choose the machine learning techniques that produce the most reliable results (according to the problem to be addressed and what is reported in the literature) for the building of the prediction models to obtain a classification that is as correct as possible. Overcoming these drawbacks would make it possible to correctly diagnose an unknown cellular sample once measurements are made on known samples.

Author Contributions: Conceptualization, G.P., C.G. and V.C.; methodology, G.P. and M.L.; software, C.G.; validation, G.P. and C.G.; formal analysis, G.P.; investigation, G.P. and M.L.; data curation, G.P. and C.G.; writing—original draft preparation, G.P.; writing—review and editing, G.P., C.G. and V.C.; supervision, V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Available online: https://www.wcrf.org/cancer-trends/worldwide-cancer-data (accessed on 1 February 2024).
2. Hall, C.; Clarke, L.; Pal, A.; Buchwald, P.; Eglinton, T.; Wakeman, C.; Frizelle, F. A Review of the Role of Carcinoembryonic Antigen in Clinical Practice. *Ann. Coloproctol.* **2019**, *35*, 294–305. [CrossRef]
3. Elumalai, S.; Managó, S.; De Luca, A.C. Raman Microscopy: Progress in Research on Cancer Cell Sensing. *Sensors* **2020**, *20*, 5525. [CrossRef]
4. Hanna, K.; Krzoska, E.; Shaaban, A.M.; Muirhead, D.; Abu-Eid, R.; Speirs, V. Raman spectroscopy: Current applications in breast cancer diagnosis, challenges and future prospects. *Br. J. Cancer* **2022**, *126*, 1125–1139. [CrossRef]

5. Murugappan, S.; Tofail, S.A.M.; Thorat, N.D. Raman Spectroscopy: A Tool for Molecular Fingerprinting of Brain Cancer. *ACS Omega* **2023**, *8*, 27845–27861. [CrossRef]
6. Chen, C.; Hao, J.; Hao, X.; Xu, W.; Xiao, C.; Zhang, J.; Pu, Q.; Liu, L. The accuracy of Raman spectroscopy in the diagnosis of lung cancer: A systematic review and meta-analysis. *Transl. Cancer Res.* **2021**, *10*, 3680–3693. [CrossRef] [PubMed]
7. Bourbousson, M.; Soomro, I.; Baldwin, D.; Notingher, I. Ex vivo Raman spectroscopy mapping of lung tissue: Label-free molecular characterization of nontumorous and cancerous tissues. *J. Med. Imaging* **2019**, *6*, 036001. [CrossRef]
8. Liu, K.; Zhao, Q.; Li, B.; Zhao, X. Raman Spectroscopy: A Novel Technology for Gastric Cancer Diagnosis. *Front. Bioeng. Biotechnol.* **2022**, *10*, 856591. [CrossRef] [PubMed]
9. Zhou, X.; Dai, J.; Chen, Y.; Duan, G.; Liu, Y.; Zhang, H.; Wu, H.; Peng, G. Evaluation of the diagnostic potential of ex vivo Raman spectroscopy in gastric cancers: Fingerprint versus high wavenumber. *J. Biomed. Opt.* **2016**, *21*, 105002. [CrossRef]
10. Gautam, R.; Vanga, S.; Ariese, F.; Umapathy, S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Tech. Instrum.* **2015**, *2*, 8. [CrossRef]
11. Morais, C.L.M.; Lima, K.M.G.; Singh, M.; Martin, F.L. Tutorial: Multivariate classification for vibrational spectroscopy in biological samples. *Nat. Protoc.* **2020**, *15*, 2143–2162. [CrossRef]
12. Guo, S.; Popp, J.; Bocklitz, T. Chemometric analysis in Raman spectroscopy from experimental design to machine learning–based modelling. *Nat. Protoc.* **2021**, *16*, 5426–5459. [CrossRef]
13. Wang, W.; Shi, B.; He, C.; Wu, S.; Zhu, L.; Jiang, J.; Wang, L.; Lin, L.; Ye, J.; Zhang, H. Euclidean distance-based Raman spectroscopy (EDRS) for the prognosis analysis of gastric cancer: A solution to tumor heterogeneity. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2023**, *288*, 122163.
14. Demsar, J.; Curk, T.; Erjavec, A.; Gorup, C.; Hocevar, T.; Milutinovic, M.; Mozina, M.; Polajnar, M.; Toplak, M.; Staric, A.; et al. Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349–2353.
15. Kokabi, M.; Tahir, M.N.; Singh, D.; Javanmard, M. Advancing Healthcare: Synergizing Biosensors and Machine Learning for Early Cancer Diagnosis. *Biosensors* **2023**, *13*, 884. [CrossRef] [PubMed]
16. Kalatzis, D.; Spyratou, E.; Karnachoriti, M.; Kouri, M.A.; Orfanoudakis, S.; Koufopoulos, N.; Pouliakis, A.; Danias, N.; Seimenis, I.; Kontos, A.G.; et al. Advanced Raman Spectroscopy Based on Transfer Learning by Using a Convolutional Neural Network for Personalized Colorectal Cancer Diagnosis. *Optics* **2023**, *4*, 310–320. [CrossRef]
17. Yan, H.; Yu, M.; Xia, J.; Zhu, L.; Zhang, T.; Zhu, Z. Tongue squamous cell carcinoma discrimination with Raman spectroscopy and convolutional neural networks. *Vib. Spectrosc.* **2019**, *103*, 102938. [CrossRef]
18. Shuyun, W.; Lin, F.; Pan, C.; Zhang, Q.; Tao, H.; Fan, M.; Xu, L.; Kong, K.V.; Chen, Y.; Lin, D.; et al. Laser tweezer Raman spectroscopy combined with deep neural networks for identification of liver cancer cells. *Talanta* **2023**, *264*, 124753. [CrossRef] [PubMed]
19. Depciuch, J.; Jakubczyk, P.; Paja, W.; Pancerz, K.; Wosiak, A.; Kula-Maximenko, M.; Yaylım, I.; Gültekin, G.I.; Tarhan, N.; Hakan, M.T.; et al. Correlation between human colon cancer specific antigens and Raman spectra. Attempting to use Raman spectroscopy in the determination of tumor markers for colon cancer. *Nanomed. Nanotechnol. Biol. Med.* **2023**, *48*, 102657. [CrossRef] [PubMed]
20. Fang, X.; Li, S.; Fu, Q.; Wang, P.; Wu, X.; Zhang, Y. Label-free identification of lung cancer cells from blood cells based on surface-enhanced Raman scattering and support vector machine. *Optik* **2021**, *248*, 168157. [CrossRef]
21. Li, X.; Yang, T.; Li, S.; Wang, D.; Song, Y.; Zhang, S. Raman spectroscopy combined with principal component analysis and k nearest neighbour analysis for non-invasive detection of colon cancer. *Laser Phys.* **2016**, *26*, 035702. [CrossRef]
22. Wang, X.; Xie, F.; Yang, Y.; Zhao, J.; Wu, G.; Wang, S. Rapid Diagnosis of Ductal Carcinoma In Situ and Breast Cancer Based on Raman Spectroscopy of Serum Combined with Convolutional Neural Network. *Bioengineering* **2023**, *10*, 65. [CrossRef] [PubMed]
23. Varmuza, K.; Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*; CRC Press: Boca Raton, FL, USA, 2009.
24. Teh, S.K.; Zheng, W.; Ho, K.Y.; Teh, M.; Yeoh, K.G.; Huang, Z. Near-infrared Raman spectroscopy for early diagnosis and typing of adenocarcinoma in the stomach. *Br. J. Surg.* **2010**, *97*, 550–557. [CrossRef] [PubMed]
25. Arévalo, L.A.; Antonova, O.; O'Brien, S.A.; Singh, G.P.; Seifert, A. Detection of Alzheimer's by machine learning-assisted vibrational spectroscopy in human cerebrospinal fluid. *J. Phys. Conf. Ser.* **2022**, *2407*, 012026. [CrossRef]
26. Lasalvia, M.; Gallo, C.; Capozzi, V.; Perna, G. Discrimination of Healthy and Cancerous Colon Cells Based on FTIR Spectroscopy and Machine Learning Algorithms. *Appl. Sci.* **2023**, *13*, 10325. [CrossRef]
27. Menges, F. *Spectragryph—Optical Spectroscopy Software*, Version 1.2.16; Spectragryph: Oberstdorf, Germany, 2023. Available online: http://www.effemm2.de/spectragryph/(accessed on 1 February 2024).
28. Perna, G.; Lasalvia, M.; Capozzi, V. Raman microspectroscopy discrimination of single human keratinocytes exposed at low dose of pesticide. *J. Mol. Struct.* **2012**, *1010*, 123–129. [CrossRef]
29. Beton, K.; Brożek-Płuska, B. Biochemistry and Nanomechanical Properties of Human Colon Cells upon Simvastatin, Lovastatin, and Mevastatin Supplementations: Raman Imaging and AFM Studies. *J. Phys. Chem. B* **2022**, *126*, 7088–7103. [CrossRef] [PubMed]
30. Brozek-Pluska, B.; Beton, K. Oxidative stress induced by tBHP in human normal colon cells by label free Raman spectroscopy and imaging. The protective role of natural antioxidants in the form of β-carotene. *RSC Adv.* **2021**, *11*, 16419–16434. [CrossRef]
31. Brozek-Pluska, B.; Jarota, A.; Kania, R.; Abramczyk, H. Zinc Phthalocyanine Photochemistry by Raman Imaging, Fluorescence Spectroscopy and Femtosecond Spectroscopy in Normal and Cancerous Human Colon Tissues and Single Cells. *Molecules* **2020**, *25*, 2688. [CrossRef]

32. Talari, A.C.S.; Movasaghi, Z.; Rehman, S.; ur Rehman, I. Raman Spectroscopy of Biological Tissues. *Appl. Spectrosc. Rev.* **2015**, *50*, 46–111. [CrossRef]
33. Chowdary, M.V.P.; Kumar, K.K.; Thakur, K.; Anand, A.; Kurien, J.; Krishna, C.M.; Mathew, S. Discrimination of Normal and Malignant Mucosal Tissues of the Colon by Raman Spectroscopy. *Photomed. Laser Surg.* **2007**, *25*, 269–274. [PubMed]
34. Fousková, M.; Vališ, J.; Synytsya, A.; Habartová, L.; Petrtýl, J.; Petruželka, L.; Setnička, V. In vivo Raman spectroscopy in the diagnostics of colon cancer. *Analyst* **2023**, *148*, 2518–2526. [CrossRef] [PubMed]
35. Brozek-Pluska, B. Statistics assisted analysis of Raman spectra and imaging of human colon cell lines—Label free, spectroscopic diagnostics of colorectal cancer. *J. Mol. Struct.* **2020**, *1218*, 128524. [CrossRef]
36. Dong, L.; Sun, X.; Chao, Z.; Zhang, S.; Zheng, J.; Gurung, R.; Du, J.; Shi, J.; Xu, Y.; Zhang, Y.; et al. Evaluation of FTIR spectroscopy as diagnostic tool for colorectal cancer using spectral analysis. *Spectrochim. Acta Part. A Mol. Biomol. Spectrosc.* **2014**, *122*, 288–294. [CrossRef] [PubMed]
37. Kaznowska, E.; Depciuch, J.; Szmuc, K.; Cebulski, J. Use of FTIR spectroscopy and PCA-LDC analysis to identify cancerous lesions within the human colon. *J. Pharm. Biomed. Anal.* **2017**, *134*, 259–268.
38. Fang, J.; Swain, A.; Unni, R.; Zheng, Y. Decoding Optical Data with Machine Learning. *Laser Photon Rev.* **2021**, *15*, 2000422.
39. Karnachoriti, M.; Stathopoulos, I.; Kouri, M.; Spyratou, E.; Orfanoudakis, S.; Lykidis, D.; Lambropoulou, M.; Danias, N.; Arkadopoulos, N.; Efstathopoulos, E.; et al. Biochemical differentiation between cancerous and normal human colorectal tissues by micro-Raman spectroscopy. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2023**, *299*, 122852. [CrossRef]
40. Blake, N.; Gaifulina, R.; Griffin, L.D.; Bell, I.M.; Rodriguez-Justo, M.; Thomas, G.M.H. Deep Learning Applied to Raman Spectroscopy for the Detection of Microsatellite Instability/MMR Deficient Colorectal Cancer. *Cancers* **2023**, *15*, 1720.
41. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern. Med.* **2013**, *4*, 627–635.