

Article

Raman Spectroscopy Assisted by Machine Learning Algorithms for the Prediction of Different Types of Oral Cancer Cells

Maria Lasalvia, Vito Capozzi and Giuseppe Perna *

Department of Clinical and Experimental Medicine, University of Foggia, 71122 Foggia, Italy; maria.lasalvia@unifg.it (M.L.); vito.capozzi@unifg.it (V.C.)

* Correspondence: giuseppe.perna@unifg.it

Abstract

Oral squamous cell carcinoma (OSCC) cytology involves extracting a cell sample consisting of single cells or small clusters of cells from patients' head and neck area in order to identify abnormal morphological characteristics after staining it. This method is used to screen for early cancer and the formation of metastases within the oral cavity. OSCC diagnosis partly depends on pathologists' skills and also laboratories' instrumentation. The use of Raman spectroscopy could support diagnoses performed using traditional methods, providing information based on the cellular biochemical environment. Technical drawbacks related to low signal-to-noise ratios of Raman spectroscopy and the need to obtain diagnostic information within a reasonable time frame have recently led to the analysis of Raman spectra using machine learning (ML) methods in order to obtain reliable information about the correct attribution of unknown cellular spectra. So, we used Raman micro-spectroscopy combined with machine learning methods to build classification models, which allow the diagnosis of different grades of OSCC in cell samples. The Raman spectra were analysed in the 980–1800 cm^{-1} range by focusing the laser beam onto the nucleus and the cytoplasm regions of single cells from different cell lines modelling healthy (HaCaT) and cancer (Cal-27, SAS and HSC-3) cytological samples. We considered six classification algorithms (k-Nearest Neighbours, Logistic Regression, Naïve Bayes, artificial Neural Network, Random Forest and Support Vector Machine) to classify unknown Raman spectra. We report two classification tasks: a 4-level classification, which encompasses healthy cells, two different types of cancer cells, and one type of metastatic cells, and a 3-level classification, which includes healthy cells, non-metastatic cancer cells, and metastatic cancer cells. Our findings show that both Neural Network and Support Vector Machine algorithms applied to Raman spectra measured in the cytoplasm region can achieve sensitivity, precision and F1-score values larger than 90% in the 3-groups classifications, whereas Support Vector Machine performs better in the 4-groups classification with respect to a Neural Network. These results contribute to increasing confidence in the clinical translation of ML-assisted Raman spectroscopy as a tool to support conventional cytological techniques.

Keywords: oral cells; Raman spectroscopy; machine learning algorithms

Academic Editor: Georges Wagnieres

Received: 2 February 2026

Revised: 25 February 2026

Accepted: 27 February 2026

Published: 28 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Oral cavity cancers, collectively known as head and neck cancers (HNC), usually originate in the squamous cells that line the mucosal surfaces of the organs in the head and neck (e.g., lips, tongue, gums, cheeks, palate, throat, larynx, etc.). The Global Cancer Observatory reported in 2022 about $4.8 \cdot 10^5$ fatalities and $9.4 \cdot 10^5$ new cases of HNC, making it the seventh most prevalent type of cancer [1]. An accurate diagnosis is necessary to enable more precise and targeted surgery, which could reduce the mortality rate. The first step in making a diagnosis is a visual or endoscopic examination of the suspicious areas by the oral surgeon. This preliminary analysis is followed by the removal of a tissue or cell sample on which to perform histological or cytological analysis. The histological analysis of tissue or cytological inspection of cell samples is performed by microscope observation of the sample after it has been properly treated (stained with suitable dyes). However, such a diagnosis may be partially subjective and depend on the skill of the pathologist, the instruments, the staining procedure and the analysis procedures [2].

Recently, artificial intelligence (AI) techniques have become very popular in histological image analysis, achieving very high levels of accuracy and diagnostic efficiency [3,4]. As for the accuracy level in classifying cytological images of HNC samples using AI methods, a recent review [5] reports only one study from 2012, where the accuracy of 92.5% achieved in recognising normal samples and 89.4% in recognising suspicious ones is reported [6].

However, histological and cytological images may present some critical issues that can influence diagnoses. In fact, image recognition with AI methods requires a learning step with numerous datasets, and sometimes to achieve such numbers, information is needed from multiple institutes, where procedures are used that are not comparable with each other. Consider also the dependence of the parameters extracted from images obtained from scanners from different manufacturers, with the related problems of obsolescence. Therefore, it is desirable that traditional diagnostic techniques be assisted by other methodologies based on the biochemical characteristics of the investigated cell or tissue samples, rather than on images. In fact, the evolution of a normal cell toward a cancerous state is accompanied by changes in the cellular biochemical environment [7].

Vibrational techniques, such as Raman and Fourier Transform Infrared (FTIR) spectroscopies, allow the identification of specific functional groups within biomolecules characteristic of cellular components [8,9]. Hence, they can provide information about the chemical composition of a biological sample and its modifications as a consequence of the onset of a disease. In particular, Raman spectroscopy is able to obtain such information with micrometric resolution, enabling the biochemical analysis of subcellular components. Therefore, diagnostic tools based on this technique reveal the potential to assist traditional cytopathological techniques, determining more reliable results for patients. Furthermore, Raman spectroscopy allows the analysis of the samples under investigation (cells, tissues, biopolymers, etc.) without the use of expensive reagents and/or dyes. However, obtaining a rapid and reliable diagnosis remains difficult due to the weak Raman scattering signal from the samples. In fact, approximately one in every 10^6 incident photons interacts with the molecules within the sample by the Raman effect [10]. In addition, the differences in spectral signals of samples having a slightly dissimilar biochemical content (it can be the case when cytological samples from healthy and cancer patients are measured) are minimal and sometimes scarcely visible if few spectra are measured. Therefore, it is useful to obtain datasets comprising a sufficiently large number of spectra and analyse them using statistical techniques capable of highlighting minimal differences among the signals. Multivariate analysis techniques are suitable for this purpose, as they simultaneously analyse multiple variables of each spectrum in order to discover relationships between them [11].

In particular, machine learning (ML) methods have been used to analyse vibrational spectroscopy data and obtain diagnostic information. Such methods are based on algorithms that can first be trained on spectra whose class is known, and subsequently, they assign unknown spectra to the proper class [12,13]. ML algorithms, such as k-nearest-neighbours (kNN) [14,15], logistic regression (LR) [16,17], Naïve Bayes (NB) [18,19], neural networks (NN) [20,21], random forest (RF) [22,23] and support vector machines (SVMs) [24,25] have been recently used for the diagnostic classification of vibrational spectroscopy spectra. The characteristics of these algorithms are explained elsewhere (see [26] and references therein). These classification techniques are largely accessible, as they are included in several commercial [27,28] and free [29–32] software programs. Although, in some cases, they have been modified in order to improve classification performances, the free and basic versions offer the advantage of being usable even by users who do not have advanced mathematical and statistical skills.

However, Raman spectroscopy is not currently officially recognised as a diagnostic technique for use in clinical settings, mainly because there is currently no universally adopted protocol that standardises the measurement and data analysis procedure. Although some important review works have been published in order to overcome this limitation [12,33,34], it can currently only be proposed as a technique to support standard diagnostic methods.

The present study investigates the classification performance of several ML algorithms for the classifications of two distinct spectral datasets of Raman spectra measured in the nucleus and cytoplasm region of cell samples: (i) a 4-group dataset, including a healthy epithelial cell line and three different types of tongue cancer cell lines, one of which was established from a metastatic cancer; (ii) a 3-group dataset, encompassing a spectral dataset of a healthy epithelial cell line, a mixture of two different tongue cancer cell lines and a metastatic tongue cell line. Our findings show that both NN and SVM algorithms applied to datasets of Raman spectra measured in the cytoplasm region can achieve sensitivity, precision and F1-score values larger than 90% in the 3-group classifications, whereas SVM performs better in the 4-group classification with respect to NN.

2. Materials and Methods

2.1. Cell Culture and Preparation

Four cell lines were selected for this study:

- A spontaneously immortalised nontumorigenic human cell line, derived from normal adult epidermal keratinocytes (HaCaT), was used as a model of healthy cells [35]. The HaCaT cell line was purchased from CliniSciences (#T0020001, Guidonia Montecelio, Rome, Italy) and cultured in DMEM/F12 (Gibco, Waltham, MA, USA) supplemented with 10% heat-inactivated FBS and 10 mM HEPES, and 1% pen/streptococcus.
- Cal-27 cells are epithelial cancer cells isolated from oral squamous cell carcinoma tissue of the tongue. Therefore, this cell line can be considered a model of oral cancer cells [36]. These cells were provided by the laboratories of Prof. Lorenzo Lo Muzio (University of Foggia) and cultured in DMEM medium supplemented with 10% fetal bovine serum (FBS) and 1% pen/streptococcus.
- SAS is a human tongue squamous cell carcinoma cell line, originally isolated from a primary cancer of a patient with tongue cancer. Therefore, this cell line is also a model of oral cancer cells [37]. The SAS cell line was provided by the laboratories of Prof. Lorenzo Lo Muzio (University of Foggia, Foggia, Italy) and cultured in RPMI medium supplemented with 10% fetal bovine serum (FBS) and 1% pen/streptococcus.

- The HSC-3 cell line was created from metastatic lymph node cancer originating from squamous cell carcinoma of the tongue. The highly aggressive characteristics of this cell line make HSC-3 a model for the study of metastatic squamous cell carcinoma [38]. The HSC-3 cell line was provided by the laboratories of Prof. Lorenzo Lo Muzio (University of Foggia, Italy) and cultured in RPMI medium enriched with 10% fetal bovine serum (FBS) and 1% pen/streptococcus.

The three cancer cell lines are frequently used models [39] deriving from the same oral district (tongue): therefore, they are suitable for evaluating the pathology of a specific oral district. However, healthy cell lines derived from the tongue are not widely available. Therefore, HaCaT cells were used as a model of healthy cells because they have a relatively high protein content compared to lipids, a characteristic found in tongue tissue compared to other surrounding areas.

The four types of cell samples were grown on a polylysine-coated glass coverslips. Once approximately 80% confluence of cells was achieved on the coverslip surface, the adherent cells were fixed in 3.7% paraformaldehyde and stored in Petri dishes with phosphate-buffered saline (PBS) until Raman spectra acquisition. Each cell sample was rinsed in deionised water to remove PBS residues before the Raman measurements.

2.2. Raman Measurements

A confocal micro-spectrometer (Labram from Jobin-Yvon Horiba, Montpellier, France) was used to obtain Raman spectra from about 40 single cells of each type. A monochromatic beam from an Ar⁺ laser (514.5 nm line with 6 mW power) was focused by an Olympus 100x oil-immersion objective (Olympus, Hachioji-shi, Japan) within each investigated cell volume. The acquisition time was 10 s, and 3 acquisitions were averaged for each spectrum. The obtained spectra point out that the 514.5 nm excitation line avoids excessive heating or thermal damage to the sample and allow one to obtain quite well-resolved Raman peaks within the above acquisition time. Before each Raman measurement, a cell volume including the nucleus and a cell volume excluding it (i.e., within the cytoplasm region) was selected by means of optical images obtained with a charge-coupled device camera. Typical images of the investigated cells are shown in Figure 1: the nucleus compartments, indicated by arrows, correspond to the dark region because they are optically more opaque than the surrounding cytoplasm. The Raman signal collected from two cell compartments (defined as “nucleus region” and “cytoplasm region” hereafter) of each single cell in backscattering geometry was analysed by a diffraction grating with 600 grooves/mm, and it was detected by a charge-coupled device. The Raman spectra were measured in the 600–1800 cm⁻¹ range, with a spectral resolution ~5 cm⁻¹/pixel. The background signal was measured by focusing the laser beam within volumes where no cell was located.

2.3. Spectra Analysis

The preprocessing for each measured Raman spectrum consists of three sequential steps described in the following: (i) subtraction of the background signal; (ii) subtraction of the fluorescence and stray light signals by means of the adaptive algorithm (which creates a baseline signal to remove the broad and featureless signals underlying the Raman peaks) using the Spectragryph software (version 1.2.16.2023) [40], with a coarseness value of 30; (iii) area normalisation of each background—and baseline-subtracted spectrum in the 980–1800 cm⁻¹ range. The latter procedure scales the intensity values of the spectrum by dividing each individual intensity value at each wavenumber value by the sum of all intensity values: the result is that each spectrum has the sum of the intensity values equal to 1. The area normalisation procedure reduces the Raman signal intensity differences re-

lated to the sampling of cells with different thicknesses. The decision to narrow the analysis of the Raman spectra to the range 980–1800 cm^{-1} is related to the presence of a strong signal at approximately 940 cm^{-1} , due to the presence of the glass coverslip [41]: it is not completely removed from the spectra after background subtraction, and it has a slightly variable intensity from spectrum to spectrum. Therefore, the presence of this spurious signal could have influenced the results of the area normalisation.

The second derivative of the mean Raman spectra of normalised cells measured in the nucleus and cytoplasm regions was calculated by means of Spectragryph software (version 1.2.16.2023) [40]. This procedure allows the detection of the spectral positions of the peaks of the Raman spectra, corresponding to the negative minima of the second derivative signals.

An exploratory data analysis was performed by Principal Component Analysis (PCA) with Orange software (Orange data mining, version 3.39.0), in order to visualise the possibility of discriminating the different types of spectra samples in the score plots. Such a discrimination is a preparatory step to the classification analysis.

Then, each set of Raman spectra of the four cell lines was divided to obtain two subsets: a calibration set (70% of the whole set) for training the algorithms and a test set (the remaining 30% of the whole set) for testing them. The calibration and test sets were obtained for both the sampled nucleus and cytoplasm regions. A proper Orange software (version 3.39) widget was used to randomly select the spectra of the calibration set. Six classification models included in the Orange software were used for training and classification procedures: kNN, LR, NB, NN, RT and SVM. The intensity values corresponding to the Raman peaks identified by second derivative spectra were selected as features on which to build the classification model during the training and test steps. This choice was made because Raman spectra are characterised by a low signal-to-noise ratio. Therefore, it is essential to base classification on Raman signals due to vibrational modes of the main and best-resolved cellular biochemical components, rather than on signals whose intensity values are similar to those of noise.

During the training step, the values of some parameters that the algorithms use to perform the learning process were optimised in order to obtain maximum sensitivity 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. Both the full cross-validation procedure and the splitting of the whole spectral dataset into calibration and test set, with each one containing randomly selected spectra for 25 times, were carried out to reduce overfitting issues that might arise during the model development and classification processes.

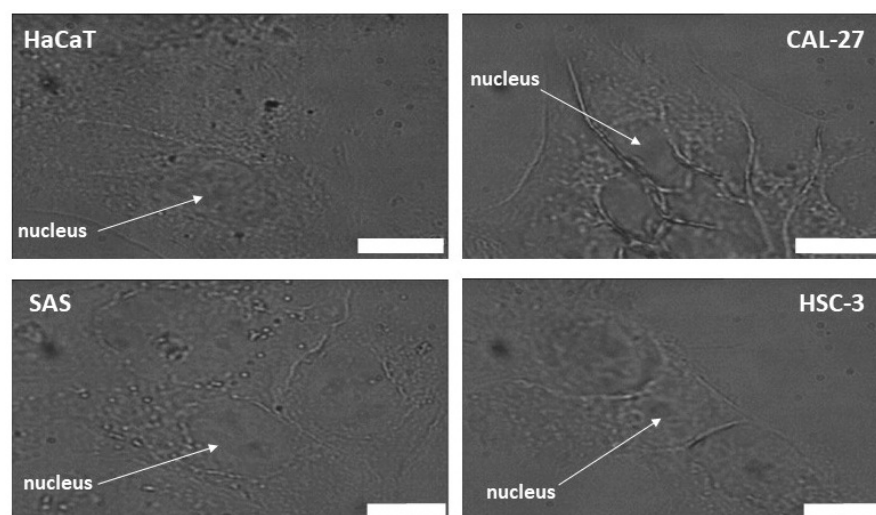


Figure 1. Microscope optical images of typical HaCaT, CAL-27, SAS and HSC-3 cells. The optical magnification is 100×. The scale bars correspond to 100 µm.

2.4. Classification Metrics

In the classification issues of spectral data, the first step to perform is deciding on the target, i.e., the categorical variable to be predicted. This is carried out by dividing the whole dataset of the measured spectra into different classes, based on the samples from which they were acquired. The simplest case is when there are two classes, one consisting of a group of spectra measured for a healthy sample and another comprising spectra measured for a diseased sample. The categorical variable to be predicted coincides with the class type. Then, unknown spectra are classified by the algorithm as belonging or not belonging to the class chosen as target and the classification performance is specified by suitable metrics: the most commonly used metric is accuracy, which indicates the proportion of spectra that are correctly classified (the ratio between the number of correct classification predictions and the total number of predictions). It is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP is the number of correct positive classifications (spectra correctly classified as belonging to the target), TN is the number of correct negative classifications (spectra correctly classified as not belonging to the target), FP corresponds to the number of spectra predicted as belonging to the target class but which actually do not and FN corresponds to the number of target spectra that were not predicted to belong to the target class.

If the number of spectra in each class of the dataset is approximately equal and the dataset includes more than two classes, the system is unbalanced, and this imbalance affects the values of the classification metrics. In fact, the number of spectra not belonging to the target class will be significantly larger than the number of spectra belonging to that class. In this case, accuracy is not a reliable classification parameter [42,43]. Indeed, the accuracy value would be greatly influenced by the number of correct negative classifications (TN) related to the target, which would be much greater than the number of correct positive classifications (TP). Therefore, it is appropriate to use classification parameters that do not depend on the number of TN classifications.

In particular, sensitivity (known also as “recall”) is the fraction of the target spectra that were correctly predicted. It is defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

In medical diagnostics, an analysis method with high sensitivity values allows for the correct identification of a high percentage of samples (cells, tissues, people) as belonging to the target class (which could be related to a healthy or pathological condition) [44]. This is particularly important when the target class refers to a pathological condition, because an incorrect identification (FN) could have significant negative consequences for the patient's future.

Another useful classification metric is precision, which is the ratio of the number of spectra correctly predicted as belonging to the target class to the total number of spectra assigned to the target class. This parameter is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

A high accuracy value indicates that the model is reliable in its diagnosis, minimising the number of samples that are incorrectly predicted to belong to the target class. Therefore, in case the target class corresponds to a pathological condition, a high precision assignment provides a decrease in unnecessary, expensive, or even harmful treatments for the patient.

Sensitivity and precision depend not only on the number of spectra correctly assigned to the target class (TP) but also on the number of spectra of the target class that were not predicted to belong to that class (FN) or the number of spectra assigned to the target class but actually did not belong to it (FP), respectively. A classification metric that takes into account both previous incorrect assignments is the F1-score, defined as the harmonic mean of sensitivity and precision:

$$\text{F1-score} = 2 \frac{\text{Sensitivity} \cdot \text{Precision}}{\text{Sensitivity} + \text{Precision}}$$

A high F1-score value indicates a low number of both FN and FP predictions. In medical diagnostics, when the target class corresponds to a pathological condition, a high F1-score reduces both the possibility that healthy people will undergo unnecessary treatments or be burdened by the stress of a positive result and the possibility that those with the disease will be led into false optimism [45].

3. Results and Discussion

Figure 2 shows the normalised mean and standard-deviation Raman spectra of the HaCaT and HNC cells, measured by focusing the laser beam on the nucleus (left side) and cytoplasm (right side) regions. In particular, the spectral positions of the main Raman signals are labelled in the HaCaT cell spectra: they were obtained by the minima of the second derivative signals of the average spectra [46], as shown in Figure 3. The attributions are reported in Table 1, in agreement with the article by Talari et al. [47].

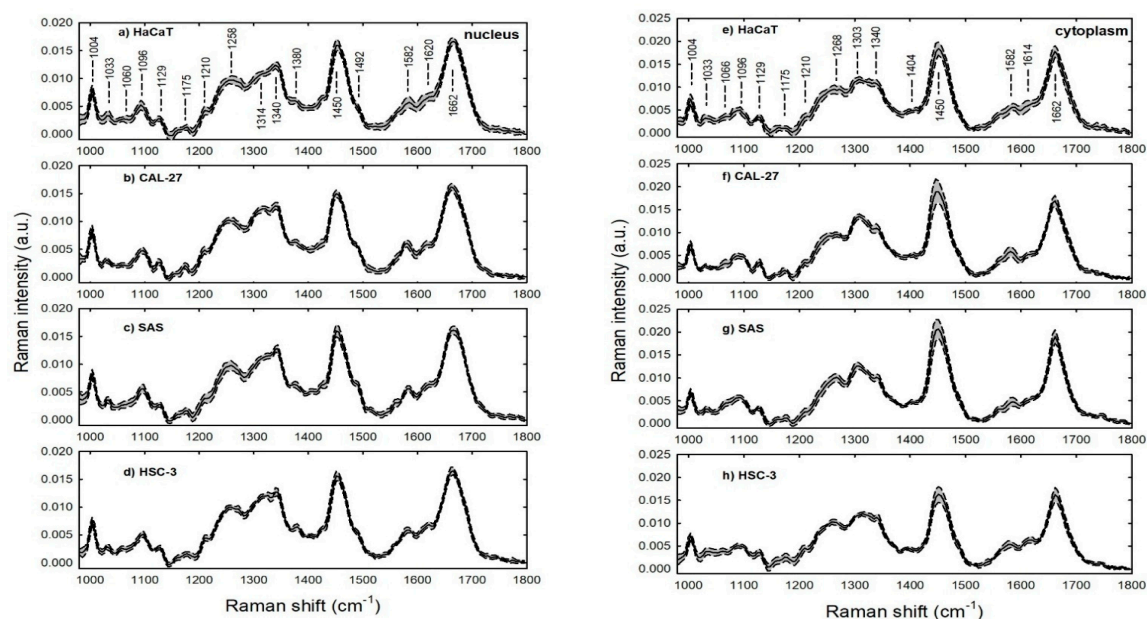


Figure 2. Mean Raman spectra of normalised healthy (HaCaT), cancer (CAL-27 and SAS) and metastatic (HSC-3) cells. The spectra were measured within a cell volume including the nucleus (**left panels**) and excluding it (**right panels**). Standard deviation spectra are shown as dashed lines. The labels indicate the spectral position of the main Raman peaks.

Figure 2 points out that the spectra measured in both subcellular compartments are quite similar, regardless of cell type. Therefore, the Raman spectra can detect only slight and not highly visible spectral changes that differentiate healthy cells from cancerous or metastatic cells. In particular, the largest intensity peaks are related to the vibrational mode of the amide I bond at 1662 cm^{-1} and the CH deformation bond at 1450 cm^{-1} in lipids and proteins. Other well-resolved contributions due to vibrational modes present in proteins are the amide III peaks at 1258 and 1268 cm^{-1} , the phenylalanine ring breathing peak at 1004 cm^{-1} , the C-N stretching peaks at 1096 and 1129 cm^{-1} , and the aromatic ring vibrations of several aminoacids at 1033 , 1175 , and 1210 cm^{-1} . The contributions to the Raman spectrum related to nucleic acid components (mainly visible in the nucleus spectra) are found at 1098 cm^{-1} (phosphodioxo bond PO_2), 1314 and 1340 cm^{-1} (CH_3/CH_2 vibrational modes in DNA bases and ring mode of DNA/RNA bases, respectively) and 1582 cm^{-1} (vibrational ring breathing mode of adenine and guanine). On the other hand, peaks related to vibrational modes of lipid components are located at 1066 , 1096 , 1129 , 1303 , 1340 , and 1450 cm^{-1} , mainly in the cytoplasm spectra, and they overlap with protein peaks.

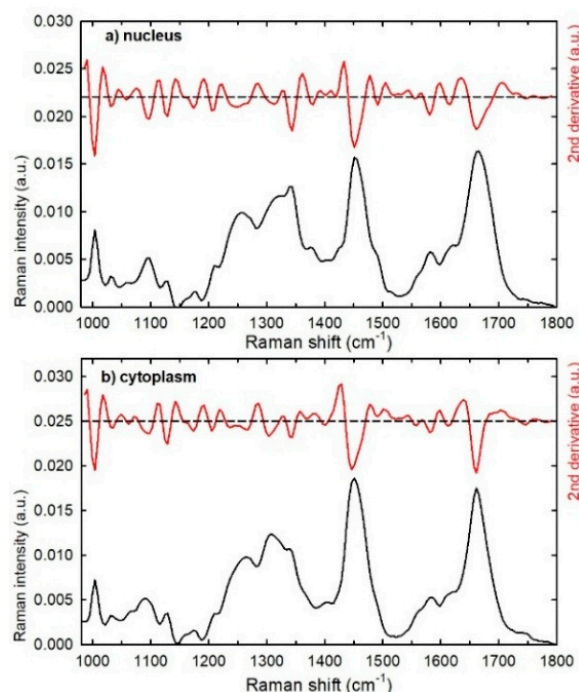


Figure 3. Mean Raman spectra (black lines) of normalised HaCaT cells measured in the nucleus (a) and cytoplasm (b) regions. The negative minima of the second derivative spectra (red lines) correspond to the peaks of the Raman spectra.

Table 1. Assignment of Raman spectral bands, according to previous results reported in the literature [47] and in the present investigation. Abbreviations: p: proteins; n.a.: nucleic acids; l: lipids; c: carbohydrates.

Spectral Position (cm ⁻¹)	Assignment
1004	C-C symmetric ring breathing of Phenylalanine (p.)
1033	C-H in-plane bending of Phenylalanine (p.)
1060 (nucleus)	C-C stretching (l.)
1066 (cytoplasm)	
1096	C-N stretching (p.) and C-C stretching (l.); Symmetric PO ₂ ⁻ stretching of DNA (n.a.)
1129	C-N stretching (p.), C-O stretching (c.), C-C stretching (l.)
1175	C-H bending aminoacids (p.)
1210	C-C ₆ H ₅ stretching aminoacids (p.)
1268 (cytoplasm)	Amide III (p.)
1258 (nucleus)	
1314 (nucleus)	CH ₃ /CH ₂ bending (n.a.)
1303 (cytoplasm)	CH ₃ /CH ₂ bending (l.)
1340 (nucleus)	Nucleic acid modes
1340 (cytoplasm)	CH ₃ /CH ₂ wagging mode in collagen (p.) and lipids (l.)
1380 (nucleus)	Ring breathing modes of DNA bases (n.a.)
1404 (cytoplasm)	Bending modes of methyl groups (p.)
1450	CH ₂ bending modes (p., l.)
1492 (nucleus)	Ring breathing modes of DNA bases (n.a.)
1582	Ring breathing modes in DNA bases (n.a.); bending modes of phenylalanine (p.)
1614 (cytoplasm)	C=C tryptophan (p.)
1620 (nucleus)	
1662	Amide I (p.); C=O stretching (l.); Ring breathing modes of DNA bases (n.a.)

If the aim of measuring the Raman spectra of an unknown cell sample is to identify the type of sample in order to perform reliable classification, Figure 2 suggests that it is not advisable to rely on the intensity values of specific peaks, as these values slightly vary from one cell type to another: consequently, the classification would not be reliable. Instead, it would be useful to consider multivariate analysis techniques, such as ML algorithms, which take into account many spectral intensity values simultaneously in order to classify an unknown spectrum. A multivariate analysis technique that is useful for exploring the possibility of classification is PCA, as it can highlight similar characteristics of spectra whose type is not given. In fact, PCA is able to visualise the whole spectral dataset to understand its structure and how well it can be separated into groups. So, a preliminary step before the use of classification techniques is that spectra of the same type cluster together in PCA score plots. Therefore, we analysed the whole dataset of Raman spectra by PCA.

Figure 4 shows the PCA score plots of the most informative PCs which are able to discriminate the spectra of the Raman dataset according to the cell type. In particular, the PC3 vs. PC5 is reported for the nucleus spectra in Figure 4a and the PC1 vs. PC4 is displayed for the cytoplasm spectra in Figure 4b. It is clearly evident that the discrimination among spectra of different cell types is larger for the cytoplasm dataset than for the nucleus dataset. In particular, the HaCaT cell spectra in Figure 4a,b are mainly characterised by negative PC4 score values, whereas the CAL-27 spectra mainly assume positive PC5 and PC4 score values, respectively. The differentiation of these spectra is based on a larger relative content of cellular components related to nucleic acids and lipids in the CAL-27 cells with respect to HaCaT cells (Figure S1). The SAS and HSC-3 spectra present predominantly positive and negative PC1 score values, respectively. However, some overlaps of positive and negative score values are visible in Figure 4 among the spectra of different cell lines. The score value overlaps are more pronounced in Figure 4a, related to nucleus spectra, although a clustering of spectra measured for each cell type is still visible. These results suggest the suitability to develop classification models capable of reliably assigning unknown spectra to the proper class, especially taking into account cytoplasm spectra.

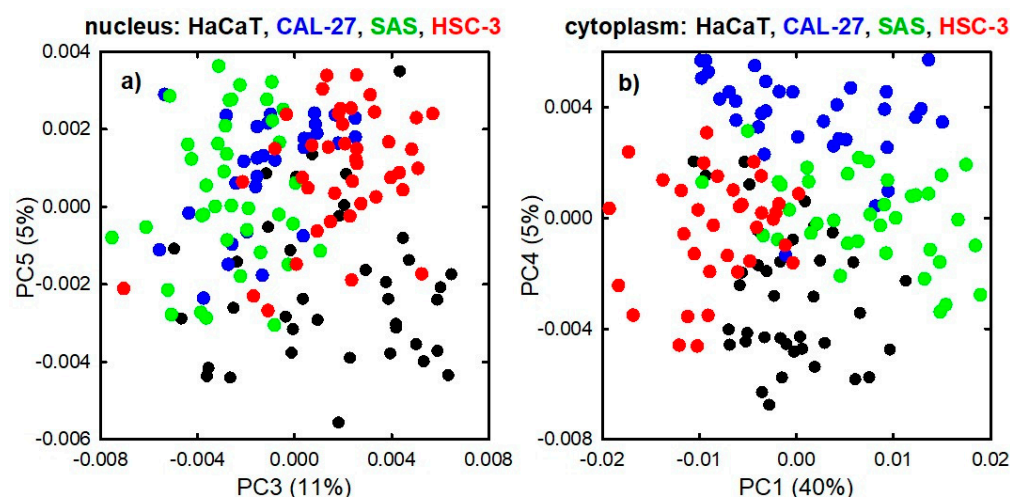


Figure 4. PC3 vs. PC5 (left side) and PC1 vs. PC4 (right side) score plot for the Raman spectra of HaCaT (black dots), CAL-27 (blue-dots), SAS (green dots) and HSC-3 (red dots) cells measured in the nucleus (a) and cytoplasm (b) region.

The clustering of spectra belonging to the same cell type and the distance between spectra of different cell types in the score plots are due to biochemical similarities and

differences, respectively. The PCA technique allows to identify the spectral and, consequently, biochemical differences between cell types by analysing the loading spectra corresponding to the main PCs responsible of discrimination: if the Raman spectra have positive score values for a certain PC, the positive peaks in the loading spectrum of such PC indicate increased Raman signals in the original spectra, whereas negative loading peaks denote a reduced contribution of the corresponding spectral signals of the measured Raman spectra [48]. However, this information extracted from loading spectra is reliable only if there is a clear separation in the score plots, that is, when the score values of one cell type are predominantly positive and those of the other ones are negative, and vice versa. Instead, the overlap of positive and negative score values visible in Figure 4 for each of the four cell types and for each of the plotted PCs does not make the loading spectra analysis reliable.

Therefore, we proceeded to evaluate the possibility of classifying unknown spectra randomly chosen from the complete dataset. The rationale of this investigation is to evaluate the use of a “spectroscopic biomarker” to correctly classify healthy, cancer or metastatic cells, regardless of the biochemical characteristics that characterise these different cell types. Thus, we want to evaluate the possibility of considering the whole spectrum as a marker of cell status, rather than the differences in intensity, shape, and spectral position of specific peaks (molecular markers).

First, we considered the whole dataset consisting of spectra from four groups of cells, i.e., HaCaT, CAL-27, SAS, and HSC-3, in order to estimate the possibility of classifying unknown spectra belonging to any of these four groups. In particular, we investigated six classification algorithms: k-NN, LR, NB, NN, RF, and SVM. For each of them, the training step was carried out using the spectra of the calibration set, separately for spectra measured in the nucleus and cytoplasm regions. During the training step, each algorithm determines the relationships between the input data necessary to create a model that allows data belonging to different classes to be differentiated. Therefore, the choice of input data (feature selection) to be provided to the algorithm to process the model is crucial. In our case, the input data provided were the intensity values of the Raman peaks shown in Figure 3 and reported in Table 1 (16 and 15 spectral features for the spectra measured in the nucleus and cytoplasm regions, respectively).

The median, 25th and 75th percentile values of sensitivity, precision, and F1-score obtained through the learning process are shown in Table 2. These values, expressed as a percentage and obtained with a consideration of calibration sets chosen randomly 25 times, were achieved by setting the model parameters of the above six algorithms to the following values:

- k-NN: the number of neighbours equal to two, by using the Euclidean metric and weights by distances;
- LR: non-regularisation type;
- NB: This algorithm does not need any parameter setting;
- NN: 30 neurons in the hidden layer, ReLu activation, Adam solver, and 300 maximum iterations;
- RF: number of trees equal to 10;
- SVM: radial basis function (RBF) kernel, SVM with cost 1.0 and regression loss epsilon 0.1, tolerance 0.001, and maximum 100 iterations.

The values of the three classification parameters that correctly attribute more than 90% of the spectra to their class have been highlighted in bold in Table 2. In particular, each percentage value was obtained by setting the cell type to which it refers as the classification target. The values reported in Table 2 suggest some remarks. First of all, during the building step of the models for assigning spectra to the proper classes, all algorithms achieve better results with cell lines describing cancer and metastatic cells than with cell

lines representing healthy cells. In addition, spectra measured in the cytoplasmic region are slightly better attributed to the right class than spectra measured in the nucleus region. Moreover, NN and SVM perform better than other algorithms. Therefore, NN and SVM models are more suitable than others for correctly classifying cancer and metastatic cell spectra using the signal measured in the cytoplasmic region.

Table 2. Median values of sensitivity, precision and F1-score for the investigated classification algorithms applied to 25 calibration sets of Raman spectra of HaCaT, CAL-27, SAS and HSC-3 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. Bold values correspond to results larger than 90%.

Sensitivity (%) (Calibration Set)	Target HaCaT	Target CAL-27	Target SAS	Target HSC-3
k-NN (nucleus)	63.0 (59.3, 68.5)	81.0 (76.2, 88.1)	81.5 (72.2, 85.2)	89.7 (81.0, 91.4)
k-NN (cytoplasm)	74.1 (70.4, 79.6)	87.0 (82.6, 93.5)	65.4 (61.5, 73.1)	95.8 (91.7, 95.8)
LR (nucleus)	74.1 (68.5, 79.6)	81.0 (81.0, 85.7)	81.5 (76.0, 87.1)	79.3 (75.9, 79.3)
LR (cytoplasm)	77.8 (70.4, 77.8)	95.7 (91.3, 95.7)	76.9 (73.1, 76.9)	83.3 (83.3, 87.5)
NB (nucleus)	70.4 (63.0, 74.1)	81.0 (73.8, 85.7)	66.7 (66.7, 74.1)	65.5 (62.1, 69.0)
NB (cytoplasm)	70.4 (64.9, 74.1)	73.9 (69.6, 78.3)	76.9 (73.1, 80.8)	87.5 (87.5, 91.7)
NN (nucleus)	81.5 (74.1, 85.2)	85.7 (85.7, 88.1)	92.6 (85.2, 92.6)	79.3 (75.9, 82.8)
NN (cytoplasm)	81.5 (77.8, 81.5)	91.3 (91.3, 95.7)	76.9 (76.9, 82.7)	95.8 (89.6, 95.8)
RF (nucleus)	74.1 (70.4, 77.8)	66.7 (61.9, 76.2)	74.1 (68.5, 77.8)	72.4 (65.5, 75.9)
RF (cytoplasm)	74.1 (70.4, 77.8)	73.9 (69.6, 78.3)	69.2 (61.5, 73.1)	87.5 (83.3, 89.6)
SVM (nucleus)	85.2 (81.5, 92.6)	85.7 (81.0, 90.5)	88.9 (85.2, 92.6)	82.8 (82.8, 86.2)
SVM (cytoplasm)	85.2 (81.5, 88.9)	87.0 (87.0, 91.3)	80.8 (80.8, 84.6)	95.8 (91.7, 95.8)
Precision (%) (calibration set)				
k-NN (nucleus)	86.4 (81.5, 94.4)	86.4 (81.0, 90.0)	80.0 (76.0, 88.0)	65.8 (63.4, 70.7)
k-NN (cytoplasm)	78.6 (74.1, 82.3)	87.0 (82.6, 90.9)	75.0 (67.9, 78.0)	82.1 (79.3, 84.9)
LR (nucleus)	77.3 (73.2, 80.4)	85.0 (78.6, 90.0)	86.4 (81.8, 88.3)	73.3 (69.9, 77.6)
LR (cytoplasm)	73.9 (71.9, 76.9)	91.7 (91.3, 95.5)	77.8 (75.0, 81.7)	84.0 (80.7, 87.0)
NB (nucleus)	74.1 (68.2, 76.9)	65.4 (61.5, 68.9)	76.9 (73.9, 79.2)	67.7 (61.8, 69.8)
NB (cytoplasm)	82.6 (79.6, 87.0)	81.0 (72.4, 84.2)	68.8 (64.5, 71.4)	81.5 (76.0, 82.7)
NN (nucleus)	81.5 (79.6, 85.7)	85.7 (82.6, 90.0)	89.3 (86.6, 95.9)	75.9 (71.2, 78.4)
NN (cytoplasm)	85.2 (80.4, 88.0)	88.0 (85.5, 91.3)	83.3 (78.9, 87.3)	91.7 (88.3, 95.6)
RF (nucleus)	75.9 (69.1, 79.6)	71.4 (67.5, 77.4)	76.9 (73.0, 80.8)	67.9 (63.9, 71.2)
RF (cytoplasm)	72.4 (67.3, 77.3)	80.0 (74.8, 86.1)	72.7 (65.9, 75.5)	82.1 (78.6, 87.5)
SVM (nucleus)	81.5 (76.8, 84.6)	94.7 (90.3, 97.6)	96.0 (92.0, 96.2)	80.6 (75.0, 85.7)
SVM (cytoplasm)	82.8 (81.5, 85.5)	94.7 (90.9, 95.4)	83.3 (80.8, 87.5)	92.0 (88.7, 92.1)
F1-score (%) (calibration set)				
k-NN (nucleus)	73.5 (68.1, 77.6)	85.0 (79.6, 88.4)	80.0 (76.2, 85.1)	74.6 (71.3, 79.1)
k-NN (cytoplasm)	76.9 (72.6, 80.7)	87.5 (83.9, 89.1)	69.4 (66.7, 72.8)	88.0 (85.7, 89.3)
LR (nucleus)	77.2 (70.0, 79.2)	84.2 (80.0, 87.5)	80.8 (78.9, 86.5)	75.4 (72.2, 78.6)
LR (cytoplasm)	75.0 (72.6, 76.4)	93.6 (91.3, 93.6)	76.0 (74.8, 79.6)	83.3 (82.1, 85.7)
NB (nucleus)	70.4 (65.3, 74.6)	72.3 (68.1, 75.3)	72.0 (70.6, 75.5)	66.7 (62.0, 69.0)
NB (cytoplasm)	76.0 (72.2, 78.8)	75.0 (72.7, 77.3)	71.7 (70.2, 75.4)	84.6 (80.4, 87.7)
NN (nucleus)	81.5 (79.2, 84.4)	85.7 (84.3, 87.8)	90.9 (86.5, 93.4)	78.0 (74.0, 81.0)
NN (cytoplasm)	82.4 (79.2, 84.9)	91.3 (88.2, 91.7)	80.0 (78.4, 83.5)	92.0 (89.6, 95.0)
RF (nucleus)	75.4 (71.6, 77.8)	70.0 (63.6, 75.6)	75.5 (70.6, 79.2)	70.0 (66.7, 72.4)
RF (cytoplasm)	72.7 (69.0, 76.9)	78.3 (71.2, 81.8)	70.8 (64.1, 74.0)	84.6 (82.0, 88.0)

SVM (nucleus)	83.6 (80.0, 87.3)	90.0 (87.5, 92.7)	92.3 (88.7, 94.3)	82.8 (78.8, 86.5)
SVM (cytoplasm)	84.6 (81.8, 85.7)	90.9 (88.7, 93.3)	83.0 (80.4, 85.2)	93.9 (91.9, 94.0)

Next, we tested the built machine learning algorithms on the test set in order to assess their sensitivity and precision in classifying unknown Raman spectra. The achieved median (and 25th and 75th percentiles) values of the classification parameters related to the class attribution of 25 sets of spectra randomly chosen from the original whole dataset are shown in Table 3. It is evident that SVM allows sensitivity and precision values of over 90% to be achieved when applied to test sets of spectra measured in the cytoplasm region of cancer and metastatic cells. However, although this algorithm performs better than others, it is not as effective in correctly assigning spectra of healthy cells to the right class: in this case, sensitivity and precision values of just over 80% are achieved. The F1-score values, which mediate the sensitivity and precision values, also confirm the superiority of the SVM algorithm over the others, but highlight that the NN algorithm also allows good results to be obtained. On the other hand, the other algorithms are less reliable in classifying unknown spectra.

Table 3. Median values of sensitivity, precision and F1-score for the investigated classification algorithms applied to 25 test sets of Raman spectra of HaCaT, CAL-27, SAS and HSC-3 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. Bold values correspond to results larger than 90%.

Sensitivity (%) (Test Set)	Target HaCaT	Target CAL-27	Target SAS	Target HSC-3
k-NN (nucleus)	72.7 (54.5, 81.8)	88.9 (77.8, 88.9)	83.3 (66.7, 91.7)	83.3 (75.0, 91.7)
k-NN (cytoplasm)	72.7 (63.6, 81.8)	90.0 (85.0, 100)	63.6 (54.5, 72.7)	100 (90.0, 100)
LR (nucleus)	72.7 (68.1, 90.9)	88.9 (66.7, 88.9)	83.3 (66.7, 91.7)	66.7 (66.7, 83.3)
LR (cytoplasm)	72.7 (63.6, 81.8)	100 (90.0, 100)	81.8 (63.6, 90.9)	80.0 (80.0, 100)
NB (nucleus)	63.6 (54.5, 81.8)	77.8 (66.7, 88.9)	75.0 (50.0, 79.1)	58.3 (50.0, 70.9)
NB (cytoplasm)	63.6 (63.6, 81.8)	70.0 (65.0, 80.0)	72.7 (63.6, 86.3)	90.0 (80.0, 100)
NN (nucleus)	81.8 (72.7, 86.3)	88.9 (77.8, 88.9)	91.7 (83.3, 100)	75.0 (66.7, 83.3)
NN (cytoplasm)	72.7 (72.7, 81.8)	90.0 (90.0, 100)	81.8 (72.7, 90.9)	100 (90.0, 100)
RF (nucleus)	72.7 (63.6, 90.9)	77.8 (55.6, 77.8)	66.7 (58.3, 79.1)	58.3 (58.3, 75.0)
RF (cytoplasm)	72.7 (72.7, 86.3)	80.0 (70.0, 90.0)	72.7 (63.6, 81.8)	90.0 (80.0, 90.0)
SVM (nucleus)	90.9 (81.8, 100)	88.9 (77.8, 88.9)	91.7 (83.3, 100)	83.3 (70.9, 91.7)
SVM (cytoplasm)	81.8 (81.8, 90.9)	90.0 (90.0, 100)	90.9 (72.7, 90.9)	100 (90.0, 100)
Precision (%) (test set)				
k-NN (nucleus)	80.0 (75.0, 95.0)	87.5 (80.0, 100)	81.8 (73.8, 87.3)	68.8 (63.7, 77.8)
k-NN (cytoplasm)	75.0 (69.2, 81.8)	83.3 (78.4, 90.5)	77.8 (66.7, 84.5)	80.0 (73.2, 89.5)
LR (nucleus)	71.4 (66.7, 81.7)	90.0 (72.0, 100)	77.8 (66.9, 88.9)	73.3 (69.1, 85.1)
LR (cytoplasm)	72.7 (69.2, 84.5)	90.9 (85.4, 90.9)	81.8 (73.8, 88.2)	88.9 (81.8, 100)
NB (nucleus)	72.7 (62.5, 77.4)	62.5 (54.5, 71.4)	75.0 (64.6, 81.8)	66.7 (51.4, 71.4)
NB (cytoplasm)	83.3 (75.0, 100)	81.8 (66.7, 87.5)	69.2 (66.7, 77.8)	76.9 (69.2, 86.1)
NN (nucleus)	81.8 (75.0, 88.9)	87.5 (75.2, 100)	90.9 (84.6, 100)	78.6 (74.2, 80.0)
NN (cytoplasm)	85.7 (76.0, 90.0)	83.3 (81.8, 90.5)	83.3 (76.4, 90.9)	90.9 (83.3, 95.5)
RF (nucleus)	73.3 (66.7, 88.2)	71.4 (60.7, 83.3)	80.0 (67.9, 87.3)	70.0 (63.9, 72.0)
RF (cytoplasm)	75.0 (65.5, 80.0)	83.3 (73.8, 89.5)	75.0 (67.9, 87.5)	88.9 (76.0, 100)
SVM (nucleus)	78.6 (76.0, 82.5)	100 (88.2, 100)	91.7 (85.7, 100)	81.8 (71.6, 90.5)
SVM (cytoplasm)	83.3 (75.0, 90.0)	90.0 (90.0, 90.9)	90.0 (80.9, 91.3)	100 (89.5, 100)
F1-score (%)				

(test set)				
k-NN (nucleus)	76.2 (64.9, 84.2)	84.2 (78.9, 88.9)	78.3 (72.4, 87.5)	74.1 (69.2, 80.7)
k-NN (cytoplasm)	76.2 (71.4, 80.9)	88.9 (82.1, 90.9)	66.7 (59.4, 78.1)	87.0 (80.9, 90.9)
LR (nucleus)	76.2 (66.7, 82.9)	80.0 (76.0, 94.1)	78.3 (71.4, 85.7)	72.7 (69.6, 78.3)
LR (cytoplasm)	75.0 (68.3, 80.0)	95.2 (84.9, 95.2)	78.3 (71.4, 86.3)	85.7 (83.0, 89.9)
NB (nucleus)	66.7 (60.0, 78.3)	66.7 (61.3, 76.2)	70.0 (54.6, 77.6)	58.3 (53.0, 70.8)
NB (cytoplasm)	75.0 (68.3, 81.8)	73.7 (68.3, 82.1)	70.0 (69.4, 80.7)	81.8 (78.3, 87.0)
NN (nucleus)	81.8 (76.6, 84.9)	82.4 (74.4, 88.9)	91.7 (82.5, 94.0)	75.0 (72.7, 82.4)
NN (cytoplasm)	80.0 (72.7, 84.2)	90.0 (85.7, 92.8)	85.7 (73.2, 90.9)	94.7 (85.7, 95.2)
RF (nucleus)	75.0 (69.8, 80.0)	70.0 (62.9, 77.8)	73.7 (63.8, 78.3)	66.7 (59.6, 72.4)
RF (cytoplasm)	76.2 (69.8, 79.2)	80.0 (70.3, 84.9)	75.0 (69.6, 80.0)	87.0 (80.9, 92.3)
SVM (nucleus)	81.8 (76.6, 88.0)	87.5 (82.4, 94.1)	90.9 (84.5, 92.3)	81.5 (72.8, 86.3)
SVM (cytoplasm)	85.7 (79.2, 87.5)	90.0 (90.0, 95.0)	87.0 (80.0, 90.9)	94.7 (89.5, 100)

Many investigations have been recently carried out with the aim of the multi-classification task by means of several ML algorithms with the Raman spectra of HNC samples. Such works have mainly involved tissue samples rather than cytological samples. For example, Chang et al. investigated Raman spectra collected from fresh human tissue excised from the tongue, gums, and buccal mucosa of 90 subjects, including patients affected by squamous cell carcinoma and healthy people. Using a NN-based algorithm (ResNet50), they reported multi-classification of the spectra and declared overall sensitivity and precision values of 90% and 91%, respectively, in the test set by using a five-fold cross-validation method. However, although the authors report a confusion matrix, they do not specify the target to which such values refer to [49]. Many other studies about the classification of tissues and even biofluids using Raman spectroscopy and ML algorithms report accuracy, sensitivity, and specificity values [50]: however, these parameters are more suitable for characterising binary classification than multi-class classification.

Instead, there are few studies about the multi-classification of cytological samples from the oral cavity. The work by Ghosh et al. [21] is very interesting, reporting Raman and FTIR spectra of oral exfoliated cells from healthy (NRML), oral leucoplakia (OLPK) and OSCC patients. Using an NN-based algorithm (DRNN), the authors obtained classification sensitivity of 75% for NRML and OLPK and 100% for OSCC, precision of 87.5%, 93.3% and 92.3% for classification of NRML, OLPK and OSCC, respectively, and F1-score of 75%, 80% and 94% for classification of NRML, OLPK and OSCC, respectively; lower values were obtained with other ML algorithms. In another study [51], the same authors tested other classification algorithms and obtained the best results with LDA-SVM: sensitivity 67%, 100% and 67% for NRML OLPK and OSCC samples, precision 100%, 50%, and 100% for NRML, LKPK and OSCC samples and F1-score 80%, 67%, and 80% for NRML, OLPK and OSCC samples, respectively. However, such results were obtained by cross-validating the dataset rather than using a test set.

Many studies of Raman spectroscopy and multi-classification of biological samples from sources other than the oral cavity are also reported in the literature [50]. In particular, Yang et al. reported sensitivity, precision, and F1-score values of approximately 96% in five-class discrimination of Raman spectra of breast cells using an artificial NN algorithm, without specifying to which target these parameters refer [52]. Furthermore, these authors also report accuracy and sensitivity values which, as discussed above, are not reliable parameters in the case of multi-classification. Recently, M. Piazza et al. classified Raman spectra of saliva from patients affected by COVID-19, subjects negative to the SARS-CoV-2 test with an ascertained episode of COVID-19 and healthy subjects, using a robust version of SVM as the classification algorithm [53]. They report sensitivity values for each of the three classes ranging from 63% to 76%, depending on the type of kernel used in the

SVM algorithm. Also, Wen et al. investigated Raman spectra from six different living cell lines (cancer cell lines derived from different organs), reporting a classification accuracy of over 95% using SVM and discriminant analysis methods [54].

Considering our results, we are confident that the SVM technique applied to the multiclassification of Raman spectra measured in the cytoplasm region of cells can represent a valid method for confirming the diagnosis obtained with conventional cytological methods. This method has the advantage of using free software, for which the model parameters must be set. However, its disadvantage is that the classification of healthy samples is less reliable (sensitivity 81.8%, precision 83.3%, and F1-score 85.7%). The risk associated with classifying healthy samples as diseased is linked to unnecessary drug treatments and interventions, which pose health risks and high costs.

It is interesting to evaluate whether reducing the number of classes improves the classification ability of Raman spectra and ML algorithms. Therefore, we considered a dataset including three classes of Raman signals: spectra of healthy HaCaT cells, spectra of metastatic HSC-3 cancer cells, and spectra of non-metastatic CAL-27 and SAS cancer cells. To build the new calibration and test sets, we used the Raman spectra discussed above as for healthy and metastatic cells is concerned, while the cancer cell spectra were obtained by randomly selecting half of the spectra from the CAL-27 cells group and the other half from the SAS cells group, so that the number of cancer cells spectra was similar to those of healthy cell spectra and metastatic cell spectra. In a similar way to the classification procedure of 4 classes of spectra, the random selection of the calibration and test sets spectra from the whole spectral dataset measured in both the nucleus and cytoplasm regions was repeated 25 times.

The sensitivity, precision and F1-score results obtained through the learning process concerning the calibration set are shown in Table 4. Also, in this case, the algorithms with the best median values of sensitivity, precision and F1-score are NN and SVM. Therefore, these two algorithms are the most promising for constructing the classification model, especially when applied to Raman spectra measured in the cytoplasm region. Furthermore, Table 4 suggests that the median values of sensitivity, precision, and F1-score of the NN and SVM algorithms correspond to the attribution of about 90% of healthy cell spectra to the right class: these values are higher than the corresponding values in the case of four-class classification (reported in Table 2).

In fact, the reduction from 4 to 3 classes allows for improved parameters for the classification of unknown spectra, as can be seen from the results reported in Table 5, related to the median values of sensitivity, precision, and F1-score obtained by providing the algorithms with the spectra of the test set. In particular, SVM and NN have median values for all three classification parameters for all cell types larger than 90% when referring to Raman spectra measured in the cytoplasm region. Overall, in the case of classifying unknown spectra into one of three classes, even the Raman spectra of healthy cells show optimal values of the classification parameters, unlike what was reported above in the case of classification into four classes. The LR algorithm also offers very good sensitivity in classifying the spectra of all three cell types measured in the cytoplasm region, but the precision in classifying healthy cells does not reach an optimal value.

Table 4. Median values of sensitivity, precision and F1-score for the investigated classification algorithms applied to 25 calibration sets of Raman spectra of healthy, cancer and metastatic 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. Bold values correspond to results larger than 90%.

Sensitivity (%) (Calibration Set)	Target Healthy	Target Cancer	Target Metastatic
k-NN (nucleus)	66.7 (61.1, 74.1)	66.7 (60.4, 70.8)	89.7 (86.2, 93.1)
k-NN (cytoplasm)	85.2 (79.6, 88.9)	75.0 (66.7, 83.3)	91.7 (87.5, 95.8)
LR (nucleus)	74.1 (70.4, 77.8)	70.8 (66.7, 77.1)	75.9 (72.4, 79.3)
LR (cytoplasm)	85.2 (81.5, 87.0)	87.5 (87.5, 91.7)	87.5 (87.5, 91.7)
NB (nucleus)	70.4 (61.1, 74.1)	62.5 (54.2, 66.7)	72.4 (67.2, 75.9)
NB (cytoplasm)	77.8 (74.1, 81.5)	83.3 (79.2, 87.5)	87.5 (83.3, 87.5)
NN (nucleus)	77.8 (74.1, 85.2)	75.0 (75.0, 79.2)	79.3 (74.1, 82.8)
NN (cytoplasm)	88.9 (85.2, 92.6)	91.7 (87.5, 95.8)	91.7 (87.5, 95.8)
RF (nucleus)	74.1 (70.4, 85.2)	58.3 (54.2, 66.7)	72.4 (69.0, 75.9)
RF (cytoplasm)	81.5 (77.8, 85.2)	79.2 (75.0, 87.5)	83.3 (83.3, 87.5)
SVM (nucleus)	85.2 (81.5, 88.9)	79.2 (75.0, 83.3)	82.8 (75.9, 86.2)
SVM (cytoplasm)	88.9 (88.9, 92.6)	91.7 (89.6, 95.8)	91.7 (91.7, 95.8)
Precision (%) (calibration set)			
k-NN (nucleus)	85.0 (83.0, 90.0)	76.2 (70.4, 80.0)	67.5 (62.5, 73.6)
k-NN (cytoplasm)	84.0 (78.9, 89.1)	88.9 (83.8, 90.7)	82.1 (77.8, 84.9)
LR (nucleus)	77.8 (75.0, 80.8)	72.7 (68.2, 75.0)	71.9 (69.5, 76.7)
LR (cytoplasm)	88.0 (83.4, 92.0)	91.7 (89.2, 95.5)	84.0 (81.5, 87.5)
NB (nucleus)	70.0 (66.0, 73.3)	61.5 (55.8, 65.9)	70.0 (64.2, 74.5)
NB (cytoplasm)	81.5 (79.2, 86.7)	80.0 (78.0, 82.6)	87.0 (82.4, 88.2)
NN (nucleus)	83.3 (80.4, 87.4)	78.3 (72.4, 81.8)	76.7 (73.0, 78.9)
NN (cytoplasm)	88.9 (85.7, 92.7)	91.3 (91.3, 95.7)	91.3 (88.0, 92.2)
RF (nucleus)	74.2 (69.3, 81.0)	65.4 (60.9, 70.7)	70.4 (67.7, 74.6)
RF (cytoplasm)	78.6 (75.9, 85.7)	85.7 (82.2, 90.0)	85.7 (80.0, 88.0)
SVM (nucleus)	82.8 (78.0, 85.2)	82.6 (77.1, 86.4)	81.5 (78.2, 84.0)
SVM (cytoplasm)	89.3 (88.2, 94.4)	95.5 (91.5, 95.8)	91.7 (89.9, 95.5)
F1-score (%) (calibration set)			
k-NN (nucleus)	75.0 (71.0, 79.2)	70.8 (66.0, 73.9)	77.1 (72.2, 81.5)
k-NN (cytoplasm)	83.3 (81.1, 88.1)	80.0 (76.2, 85.7)	86.3 (84.3, 90.0)
LR (nucleus)	74.5 (72.4, 78.9)	70.4 (66.7, 75.5)	75.4 (71.8, 76.9)
LR (cytoplasm)	85.2 (83.3, 88.5)	89.8 (89.1, 92.6)	87.5 (84.3, 88.2)
NB (nucleus)	69.0 (64.1, 72.7)	62.5 (56.6, 66.0)	70.2 (66.7, 73.9)
NB (cytoplasm)	80.0 (76.9, 83.0)	81.6 (80.0, 84.0)	85.7 (84.6, 88.7)
NN (nucleus)	81.5 (77.7, 84.4)	77.6 (73.5, 80.9)	78.0 (74.1, 80.0)
NN (cytoplasm)	88.9 (85.7, 92.3)	91.3 (89.4, 94.8)	91.7 (89.4, 92.9)
RF (nucleus)	76.4 (70.9, 79.7)	63.4 (59.3, 66.7)	72.4 (68.4, 75.2)
RF (cytoplasm)	80.7 (78.6, 84.9)	82.6 (77.9, 86.3)	84.4 (81.6, 87.5)
SVM (nucleus)	82.1 (80.7, 85.7)	82.6 (74.7, 83.3)	80.0 (77.8, 84.2)
SVM (cytoplasm)	90.9 (88.9, 92.8)	93.6 (91.3, 94.9)	92.0 (91.7, 93.9)

Table 5. Median values of sensitivity, precision and F1-score for the investigated classification algorithms applied to 25 test sets of Raman spectra of healthy, cancer and metastatic 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. Bold values correspond to results larger than 90%.

Sensitivity (%) (Test Set)	Target Healthy	Target Cancer	Target Metastatic
k-NN (nucleus)	72.7 (63.6, 81.8)	72.7 (63.6, 81.8)	91.7 (83.3, 91.7)
k-NN (cytoplasm)	90.9 (81.8, 100)	81.8 (72.7, 90.9)	90.0 (90.0, 100)
LR (nucleus)	72.7 (68.1, 81.8)	72.7 (50.0, 81.8)	75.0 (66.7, 83.3)
LR (cytoplasm)	90.9 (81.8, 100)	90.9 (86.3, 95.5)	90.0 (80.0, 100)
NB (nucleus)	72.7 (63.6, 86.3)	63.6 (45.5, 72.7)	75.0 (66.7, 83.3)
NB (cytoplasm)	81.8 (72.7, 90.9)	81.8 (81.8, 90.9)	90.0 (75.0, 90.0)
NN (nucleus)	90.9 (81.8, 90.9)	72.7 (68.1, 90.9)	83.3 (75.0, 91.7)
NN (cytoplasm)	90.9 (86.3, 100)	90.9 (90.9, 100)	100 (90.0, 100)
RF (nucleus)	81.8 (63.6, 90.9)	63.6 (54.5, 72.7)	75.0 (66.7, 83.3)
RF (cytoplasm)	90.9 (77.2, 90.9)	90.9 (81.8, 90.9)	80.0 (75.0, 90.0)
SVM (nucleus)	90.9 (81.8, 100)	81.8 (72.7, 90.9)	83.3 (79.1, 87.5)
SVM (cytoplasm)	90.9 (86.3, 100)	100 (90.9, 100)	100 (90.0, 100)
Precision (%) (test set)			
k-NN (nucleus)	88.9 (80.9, 100)	83.3 (72.7, 90.0)	68.8 (64.7, 73.3)
k-NN (cytoplasm)	84.6 (80.9, 90.0)	90.9 (88.9, 100)	81.8 (76.9, 100)
LR (nucleus)	75.0 (70.3, 88.2)	75.0 (68.3, 80.0)	69.2 (63.4, 80.0)
LR (cytoplasm)	84.6 (81.8, 91.7)	100 (86.7, 100)	90.0 (83.3, 100)
NB (nucleus)	72.7 (60.7, 80.0)	66.7 (61.3, 70.7)	70.0 (59.4, 76.0)
NB (cytoplasm)	80.0 (72.0, 87.3)	85.7 (79.3, 90.0)	88.9 (79.8, 100)
NN (nucleus)	88.9 (79.3, 91.3)	83.3 (78.2, 89.5)	83.3 (70.3, 85.7)
NN (cytoplasm)	91.7 (84.6, 100)	100 (91.3, 100)	100 (90.0, 100)
RF (nucleus)	73.3 (69.4, 79.8)	72.7 (66.7, 85.7)	71.4 (66.7, 80.9)
RF (cytoplasm)	83.3 (76.0, 90.9)	90.0 (81.8, 100)	88.9 (77.5, 100)
SVM (nucleus)	83.3 (76.0, 90.5)	85.7 (78.2, 90.9)	83.3 (79.3, 90.9)
SVM (cytoplasm)	91.7 (90.9, 100)	100 (91.7, 100)	100 (90.0, 100)
F1-score (%) (test set)			
k-NN (nucleus)	77.8 (71.8, 84.2)	76.2 (72.7, 83.7)	77.4 (74.1, 81.5)
k-NN (cytoplasm)	87.0 (80.9, 91.7)	85.7 (78.9, 90.9)	87.0 (85.7, 94.7)
LR (nucleus)	76.2 (71.4, 81.7)	72.0 (61.0, 78.3)	72.7 (66.7, 78.3)
LR (cytoplasm)	87.0 (82.5, 91.7)	91.7 (88.5, 95.2)	87.0 (84.2, 92.8)
NB (nucleus)	76.2 (66.7, 80.0)	63.6 (48.8, 71.4)	69.6 (65.1, 78.3)
NB (cytoplasm)	80.0 (74.4, 85.1)	85.7 (80.9, 88.0)	85.7 (79.8, 92.3)
NN (nucleus)	87.0 (77.2, 93.0)	80.0 (73.7, 87.0)	80.0 (72.0, 88.0)
NN (cytoplasm)	90.9 (90.0, 95.2)	95.2 (90.9, 100)	94.7 (90.0, 100)
RF (nucleus)	80.0 (69.6, 83.2)	66.7 (60.0, 75.0)	74.1 (67.8, 77.6)
RF (cytoplasm)	84.2 (77.6, 89.9)	87.0 (83.3, 95.2)	85.7 (77.8, 90.0)
SVM (nucleus)	85.7 (78.3, 91.3)	84.2 (78.3, 90.5)	83.3 (77.2, 87.5)
SVM (cytoplasm)	95.2 (90.0, 95.7)	95.7 (91.3, 100)	94.7 (90.5, 100)

4. Conclusions

In summary, ML techniques applied to Raman spectra of healthy and HNC cells are capable of discriminating between them and classifying different cell types with good sensitivity, accuracy, and F1-score values. The obtained values of such parameters are above 90% when using SVM and NN algorithms for classifying Raman spectra measured in the cytoplasm region. The ML techniques perform best in the case of classification into three classes (healthy, cancer and metastatic), while the classification of healthy cells is less optimal (sensitivity, precision, and F1-score equal to 81.8%, 83.3%, and 85.7%, respectively, using SVM) in the case of classification into four classes (healthy, two different types of cancer, and metastatic).

These findings, although they should be considered as preliminary and need confirmation by increasing the number of spectra included in the dataset, are promising for promoting the development of auxiliary clinical diagnosis systems for oral cancer detection and screening activities based on Raman spectroscopy and ML techniques. In fact, Raman spectroscopy is a powerful and non-destructive means for the identification of biochemical components, so that it can distinguish healthy cells from cancer and metastatic cells. Therefore, this information may be useful in the initial phase of diagnosis. The high sensitivity and accuracy values of the proposed method can be considered an important starting point for encouraging the design of diagnostic systems based on Raman spectroscopy. Another positive issue of the proposed method is that the analysis was performed with free software, which is simple and easy to use even for people who do not have ML skills, after optimising the parameters of the ML model to be used.

However, the proposed procedure also presents some problems that delay and limit its applicability as a diagnostic technique in clinical practice. The relatively weak Raman signal and the need to measure a sufficient number of spectra to obtain reliable statistical information mean that the time required to obtain a cytological diagnosis can be as long as several hours, which is more than the time required for a cytologist to perform a qualitative analysis of stained cytological samples. Although the times of Raman-based diagnosis could be reduced, this technique is mainly intended as a way to supplement diagnostic information obtained from other techniques, as imaging methods. In fact, integrating biochemical and morphological information could lead to more reliable predictions.

Therefore, the proposed method can only be considered as a proof-of-concept to promote the use of Raman spectroscopy and ML algorithms for diagnosing the degree of malignancy of cytological samples. First, the number of spectra that allow for reliable classification within the shortest possible time should be optimised. Next, the topic needs to be further explored using *ex vivo* cytological samples extracted from patients, rather than cell lines, in order to enhance the sample diversity. Therefore, the applicability of the investigated method in clinical diagnostics can only be considered as a future prospect, and certainly it needs further investigations, with the main ones being related to the number of measurements to be carried out to obtain a reliable classification and the standardisation of the measurement and analysis procedures. Some issues to resolve when using *ex vivo* cytological samples extracted from patients are related to the presence of other cell types (e.g., inflammatory cells, bacteria) and the patient-to-patient variability that might complicate the classification of the Raman spectra. As for the exclusion of other cell types from the cytological sample to be classified, statistical techniques for removing outlier spectra could be used (in fact, the Raman spectra of different cells differ, although slightly in some cases, as a consequence of the relative intensities of the spectral peaks). Nonetheless, intrinsic intracellular variability would still be present (which is already present in cell lines, as visible in Figure 2). This variability could be reduced by measuring a few dozen spectra of the investigated cytological sample and then reducing the dataset by

selecting a portion of these spectra as representative of the cytological sample, using selection methods based on k-nearest neighbours techniques. The crucial problem to be solved is related to the patient-to-patient variability, which inevitably influences the classification of cytological spectral samples from different patients. That is, how can spectra measured for an unknown cytological sample be classified best, given the availability of reference datasets consisting of spectra whose classification is known (healthy, non-metastatic cancer, metastatic cancer, etc.) and which were measured for different patients? This is the main issue that needs to be addressed in order to state that Raman spectroscopy, assisted by ML classification techniques, can be used as a diagnostic support technique. The decrease in patient-to-patient variability involves several successive steps as the standardisation of the measurement protocol and spectrum normalisation procedures (to eliminate outliers and differences related to the cell volume sampled); in addition, it would be important to construct different spectral reference datasets related to age, gender, and comorbidity (in order to obtain more homogeneous datasets); eventually, the classification should be performed using several ML techniques, comparing the results to verify that they are not too discordant. The refinement of these points could allow the adoption of “clinical Raman spectroscopy” as an auxiliary diagnostic technique and support for traditional techniques currently recognised and accepted in the detection of oral cancer metastasis in the early stages.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16052380/s1>, Figure S1. Loading 5 and loading 4 plots from the PCA (Figure 4) of the Raman spectra of HaCaT, CAL-27, SAS and HSC-3 cells measured for the nucleus (left) and cytoplasm (right) regions. The HaCaT cell spectra in Figure 4 are mainly characterized by negative PC4 and PC5 score values, whereas the CAL-27 spectra mainly assume positive PC4 and PC5 score values. As positive peaks in loading plots indicate a higher content of the corresponding biochemical component for spectra with positive score values, it can be deduced that the cancer CAL-25 cells are characterized by a larger relative amount of nucleic acid (peaks at 1098, 1340, 1384, 1490, 1580 cm^{-1}) and lipid (1305 and 1440 cm^{-1}) components.

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

Funding: This research was funded by PRINPNRR 2022-P2022F4P8P-Bi-functioNal plasma-treated solutions as a new thErapeutiC Tool for cAnceR; acronym: NECTAR, CUP number: D53D23021610001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available upon request from the corresponding author.

Acknowledgments: The authors would like to thank Lorenzo Lo Muzio and Vittoria Perrotti for providing the investigated cell lines.

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

References

1. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2024**, *74*, 229–263. <https://doi.org/10.3322/caac.21834>.

2. Rashmi, R.; Prasad, K.; Udupa, C.B.K. Breast histopathological image analysis using image processing techniques for diagnostic purposes: A methodological review. *J. Med. Syst.* **2022**, *46*, 7.
3. Yu, H.; Yu, W.; Enwu, Y.; Ma, J.; Zhao, X.; Zhang, L.; Yang, F. Enhancing head and neck cancer detection accuracy in digitized whole-slide histology with the HNSC-classifier: A deep learning approach. *Front. Mol. Biosci.* **2025**, *12*, 1652144. <https://doi.org/10.3389/fmolb.2025.1652144>.
4. Dogan, R.S.; Yılmaz, B. Histopathology image classification: Highlighting the gap between manual analysis and AI automation. *Front. Oncol.* **2024**, *13*, 1325271. <https://doi.org/10.3389/fonc.2023.1325271>.
5. Bassani, S.; Santonicco, N.; Eccher, A.; Scarpa, A.; Vianini, M.; Brunelli, M.; Bisi, N.; Nocini, R.; Sacchetto, L.; Munari, E.; et al. Artificial intelligence in head and neck cancer diagnosis. *J. Pathol. Inform.* **2022**, *13*, 100153. <https://doi.org/10.1016/j.jpi.2022.100153>.
6. Macaulay, C.; Poh, C.F.; Guillaud, M.; Michele Williams, P.; Laronde, D.M.; Zhang, L.; Rosin, M.P. High throughput image cytometry for detection of suspicious lesions in the oral cavity. *J. Biomed. Opt.* **2012**, *17*, 086004. <https://doi.org/10.1117/1.JBO.17.8.086004>.
7. Qin, X.; Li, T.; Li, S.; Yang, H.; Wu, C.; Zheng, C.; You, F.; Liu, Y. The tumor biochemical and biophysical microenvironments synergistically contribute to cancer cell malignancy. *Cell. Mol. Immunol.* **2020**, *17*, 1186–1187. <https://doi.org/10.1038/s41423-019-0282-5>.
8. Xu, W.; Zhu, W.; Xia, Y.; Hu, S.; Liao, G.; Xu, Z.; Shen, A.; Hu, J. Raman spectroscopy for cell analysis: Retrospect and prospect. *Talanta* **2025**, *285*, 127283. <https://doi.org/10.1016/j.talanta.2024.127283>.
9. Al-Kelani, M.; Buthelezi, N. Advancements in medical research: Exploring Fourier Transform Infrared (FTIR) spectroscopy for tissue, cell, and hair sample analysis. *Ski. Res. Technol.* **2024**, *30*, e13733. <https://doi.org/10.1111/srt.13733>.
10. Yulchi, R.t.; Morgunov, V.; Junjuri, R.; Bocklitz, T. Artifacts and Anomalies in Raman Spectroscopy: A Review on Origins and Correction Procedures. *Molecules* **2024**, *29*, 4748. <https://doi.org/10.3390/molecules29194748>.
11. 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. <https://doi.org/10.1140/epjti/s40485-015-0018-6>.
12. 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. <https://doi.org/10.1038/s41596-020-0322-8>.
13. Behl, I.; Calado, G.; Vishwakarma, A.; Traynor, D.; Flint, S.; Galvin, S.; Healy, C.M.; Pimentel, M.L.; Malkin, A.; Byrne, H.J.; et al. Classification of cytological samples from oral potentially malignant lesions through Raman spectroscopy: A pilot study. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2022**, *266*, 120437. <https://doi.org/10.1016/j.saa.2021.120437>.
14. Zhang, H.; Chen, C.; Gao, R.; Yan, Z.; Zhu, Z.; Yang, B.; Chen, C.; Lv, X.; Li, H.; Huang, Z. Rapid identification of cervical adenocarcinoma and cervical squamous cell carcinoma tissue based on Raman spectroscopy combined with multiple machine learning algorithms. *Photodiagnosis Photodyn. Ther.* **2021**, *33*, 102104.
15. Zheng, Q.; Li, J.; Yang, L.; Zheng, B.; Wang, J.; Lv, N.; Luo, J.; Martin, F.L.; Liu, D.; He, J. Raman Spectroscopy as a Potential Diagnostic Tool to Analyse Biochemical Alterations in Lung Cancer. *Analyst* **2020**, *145*, 385–392.
16. Condino, F.; Crocco, M.C.; Pirritano, D.; Petrone, A.; Del Giudice, F.; Guzzi, R. A Linear Predictor Based on FTIR Spectral Biomarkers Improves Disease Diagnosis Classification: An Application to Multiple Sclerosis. *J. Pers. Med.* **2023**, *13*, 1596. <https://doi.org/10.3390/jpm13111596>.
17. 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.
18. Agbaria, A.H.; Rosen, G.B.; Lapidot, I.; Rich, D.H.; Mordechai, S.; Kapelushnik, J.; Huleihel, M.; Salman, A. Rapid diagnosis of infection etiology in febrile pediatric oncology patients using infrared spectroscopy of leukocytes. *J. Biophotonics.* **2020**, *13*, e201900215. <https://doi.org/10.1002/jbio.201900215>.
19. Mukta Sharma Chia-Lung Tsai, M.V. Arularasu, Shiang-Fu Huang, Yu-Li Hsieh, Liann-Be Chang, Combined feature-based classification of premalignant and malignant oral lesions using VELscope endogenous fluorescence imaging and dry serum Raman spectroscopy. *Photodiagnosis Photodyn. Ther.* **2025**, *57*, 105247. <https://doi.org/10.1016/j.pdpdt.2025.105247>.
20. Tomas, R.C.; Sayat, A.J.; Atienza, A.N.; Danganan, J.L.; Ramos, M.R.; Fellizar, A.; Notarte, K.I.; Angeles, L.M.; Bangaol, R.; Santillan, A.; et al. Detection of breast cancer by ATR-FTIR spectroscopy using artificial neural networks. *PLoS ONE* **2022**, *17*, e0262489. <https://doi.org/10.1371/journal.pone.0262489>.
21. Ghosh, A.; Chaudhuri, D.; Adhikary, S.; Chatterjee, K.; Roychowdhury, A.; Das, A.K.; Barui, A. Deep reinforced neural network model for cyto-spectroscopic analysis of epigenetic markers for automated oral cancer risk prediction. *Chemom. Intell. Lab. Syst.* **2022**, *224*, 104548, ISSN 0169-7439. <https://doi.org/10.1016/j.chemolab.2022.104548>.

22. Zhang, W.; Giang, C.M.; Cai, Q.; Badie, B.; Sheng, J.; Li, C. Using random forest for brain tissue identification by Raman spectroscopy. *Mach. Learn. Sci. Technol.* **2023**, *4*, 045053. <https://doi.org/10.1088/2632-2153/ad1349>.
23. Das, S.; Roy, S.; Bir, A.; Ghosh, A.; Bhattacharyya, T.K.; Lahiri, P.; Lahiri, B. FTIR-based molecular fingerprinting for the rapid classification of dengue and chikungunya from human sera using machine learning: An observational study. *Lancet Reg. Health—S. Asia* **2025**, *40*, 100630. <https://doi.org/10.1016/j.lansea.2025.100630>.
24. IRehan, S.; Khan, R. Ullah Raman spectroscopy assisted support vector machine: A steadfast tool for noninvasive classification of urinary glucose of diabetes mellitus. *Phys. Scr.* **2024**, *99*, 026004.
25. Duc, L.A.; Tung, N.T. Use of Raman Spectroscopy to Diagnose Diabetes with SVM. In *Nature of Computation and Communication*; Phan, C.V., Nguyen, T.D., Eds.; ICTCC 2022; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; Springer: Cham, Switzerland, 2023; Volume 473. https://doi.org/10.1007/978-3-031-28790-9_6.
26. Meza Ramirez, C.A.; Greenop, M.; Almoshawah, Y.A.; Martin Hirsch, P.L.; Rehman, I.U. Advancing cervical cancer diagnosis and screening with spectroscopy and machine learning. *Expert Rev. Mol. Diagn.* **2023**, *23*, 375–390. <https://doi.org/10.1080/14737159.2023.2203816>.
27. *The Unscrambler® X Software*, version 10.5; Aspen Technology, Inc.: Bedford, MA, USA, 2018.
28. *MATLAB*, version R2025b; The MathWorks Inc.: Natick, MA, USA, 2010.
29. R Core Team. *R: A Language and Environment for Statistical Computing*, version 4.5.2; R Foundation for Statistical Computing: Vienna, Austria, 2024.
30. 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.
31. Georgiev, D.; Pedersen, S.V.; Xie, R.; Fernández-Galiana, Á.; Stevens, M.M.; Barahona, M. RamanSPy: An Open-Source Python Package for Integrative Raman Spectroscopy Data Analysis. *Anal. Chem.* **2024**, *96*, 8492–8500. <https://doi.org/10.1021/acs.anal-chem.4c00383>.
32. Pavlou, E.; Kourkoumelis, N. PyFasma: An open-source, modular Python package for preprocessing and multivariate analysis of Raman spectroscopy data. *Analyst* **2025**, *150*, 3112–3122.
33. Butler, H.J.; Ashton, L.; Bird, B.; Cinque, G.; Curtis, K.; Dorney, J.; Esmonde-White, K.; Fullwood, N.J.; Gardner, B.; Martin-Hirsch, P.L.; et al. Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.* **2016**, *11*, 664–687.
34. Traynor, D.; Behl, I.; O’dea, D.; Bonnier, F.; Nicholson, S.; O’connell, F.; Maguire, A.; Flint, S.; Galvin, S.; Healy, C.M.; et al. Raman spectral cytopathology for cancer diagnostic applications. *Nat. Protoc.* **2021**, *16*, 3716–3735.
35. Boukamp, P.; Petrussevska, R.T.; Breitkreutz, D.; Hornung, J.; Markham, A.; Fusenig, N.E. Normal keratinization in a spontaneously immortalized aneuploid human keratinocyte cell line. *J. Cell Biol.* **1988**, *106*, 761–771. <https://doi.org/10.1083/jcb.106.3.761>.
36. Jiang, L.; Ji, N.; Zhou, Y.; Li, J.; Liu, X.; Wang, Z.; Chen, Q.; Zeng, X. CAL 27 is an oral adenosquamous carcinoma cell line. *Oral Oncol.* **2009**, *45*, e204–e207. <https://doi.org/10.1016/j.oraloncology.2009.06.001>.
37. Takahashi, K.; Kanazawa, H.; Akiyama, Y.; Tazaki, S.; Takahara, M.; Muto, T.; Tanzawa, H.; Sato, K. Establishment and Characterization of a Cell Line (SAS) from Poorly Differentiated Human Squamous Cell Carcinoma of the Tongue. *J. Jpn. Stomatol. Soc.* **1989**, *38*, 20–28.
38. Ribeiro, I.P.; Rodrigues, J.M.; Mascarenhas, A.; Kosyakova, N.; Caramelo, F.; Liehr, T.; Melo, J.B.; Carreira, I.M. Cytogenetic, genomic, and epigenetic characterization of the HSC-3 tongue cell line with lymph node metastasis. *J. Oral Sci.* **2018**, *60*, 70–81. <https://doi.org/10.2334/josnusd.16-0811>.
39. Wijaya, J.; Gunardi, I.; Julvyn, J.; Lim, C.; Johannis, B.N.; Hartanto, F.K.; Rajasa, A.S.; Amtha, R.; Sari, E.F. Selecting the Appropriate Oral Cancer Cell Line: Characteristic-Based Recommendations from a Systematic Review. *J. Cancer Res. Updates* **2025**, *14*, 279–289.
40. Menges, F. *Spectragryph—Optical Spectroscopy Software*, version 1.2.16; Spectragryph: Oberstdorf, Germany, 2023. Available online: <http://www.effemm2.de/spectragryph/> (accessed on 20 June 2025).
41. 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.
42. Wang, L.; Han, M.; Li, X.; Zhang, N.; Cheng, H. Review of Classification Methods on Unbalanced Data Sets. *IEEE Access* **2021**, *9*, 64606–64628. <https://doi.org/10.1109/ACCESS.2021.3074243>.
43. Song, G.; Chung, H.; Hernanda, R.A.P.; Lee, J.; Lee, H. Nondestructive Discrimination of Plant-Based Patty Containing Traditional Medicinal Roots Using Visible–Near-Infrared Hyperspectral Imaging and Machine Learning Techniques. *Chemosensors* **2025**, *13*, 158. <https://doi.org/10.3390/chemosensors13050158>.

44. Monaghan, T.F.; Rahman, S.N.; Agudelo, C.W.; Wein, A.J.; Lazar, J.M.; Everaert, K.; Dmochowski, R.R. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina* **2021**, *57*, 503. <https://doi.org/10.3390/medicina57050503>.
45. Lever, J.; Krzywinski, M.; Altman, N. Classification evaluation. *Nat. Methods* **2016**, *13*, 603–604. <https://doi.org/10.1038/nmeth.3945>.
46. Huzortey, A.A.; Anderson, B.; Owusu, A. Raman spectra recovery using a second derivative technique and range independent baseline correction algorithm. *OSA Contin.* **2021**, *4*, 2468–2480.
47. Talari, A.C.S.; Movasaghi, Z.; Rehman, S.; Rehman, I.U. Raman Spectroscopy of Biological Tissues. *Appl. Spectrosc. Rev.* **2015**, *50*, 46–111.
48. Brauchle, E.; Thude, S.; Brucker, S.Y.; Schenke-Layland, K. Cell death stages in single apoptotic and necrotic cells monitored by Raman microspectroscopy. *Sci. Rep.* **2014**, *4*, 4698. <https://doi.org/10.1038/srep04698>.
49. Chang, X.; Yu, M.; Liu, R.; Jing, R.; Ding, J.; Xia, J.; Zhu, Z.; Li, X.; Yao, Q.; Zhu, L.; et al. Deep learning methods for oral cancer detection using Raman spectroscopy. *Vib. Spectrosc.* **2023**, *126*, 103522. <https://doi.org/10.1016/j.vibspec.2023.103522>.
50. Conforti, P.M.; Lazzini, G.; Russo, P.; D’Acunto, M. Raman Spectroscopy and AI Applications in Cancer Grading: An Overview. *IEEE Access* **2024**, *12*, 54816–54852. <https://doi.org/10.1109/ACCESS.2024.3388841>.
51. Chaudhuri, D.; Ghosh, A.; Raha, S.; Chakraborty, A.; Chatterjee, K.; Barui, A. Machine learning algorithm ensembles for early oral cancer risk assessment using Raman cyto-spectroscopy. *Soft Comput.* **2023**, *27*, 13861–13875. <https://doi.org/10.1007/s00500-023-08995-z>.
52. Yang, C.; Aizezi, K.; Li, J.; Wang, X.; Li, F.; Lei, W.; Xia, J.; Maimaitijiang, A. The Identification of Breast Cancer Subtypes by Raman Spectroscopy Integrated with Machine Learning Algorithms: Analyzing the Influence of Baseline. *J. Raman Spectrosc.* **2025**, *56*, 556–566. <https://doi.org/10.1002/jrs.6799>.
53. Piazza, M.; Spinelli, A.; Maggioni, F.; Bedoni, M.; Messina, E. A robust support vector machine approach for Raman data classification. *Decis. Anal. J.* **2025**, *16*, 100595. <https://doi.org/10.1016/j.dajour.2025.100595>.
54. Wen, J.; Tang, T.; Kanwal, S.; Lu, Y.; Tao, C.; Zheng, L.; Zhang, D.; Gu, Z. Detection and Classification of Multi-Type Cells by Using Confocal Raman Spectroscopy. *Front. Chem.* **2021**, *9*, 641670. <https://doi.org/10.3389/fchem.2021.641670>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.