Hindawi
Journal of Spectroscopy
Volume 2021, Article ID 5572782, 11 pages
https://doi.org/10.1155/2021/5572782

Research Article

Raman Microspectral Study and Classification of the Pathological Evolution of Breast Cancer Using Both Principal Component Analysis-Linear Discriminant Analysis and Principal Component Analysis-Support Vector Machine

Heping Li,[1] Yu Ren,[2] Fan Yu,[1] Dongliang Song,[1] Lizhe Zhu,[2] Shibo Yu,[2] Siyuan Jiang,[2] and Shuang Wang[1]

[1]*State Key Laboratory of Photon-Technology in Western China Energy, Institute of Photonics and Photon-Technology, Northwest University, Xi'an, Shaanxi 710069, China*
[2]*Department of Breast Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi 710061, China*

Correspondence should be addressed to Shuang Wang; swang@nwu.edu.cn

Received 28 February 2021; Revised 22 March 2021; Accepted 9 April 2021; Published 22 April 2021

Academic Editor: Alessandra Durazzo

To facilitate the enhanced reliability of Raman-based tumor detection and analytical methodologies, an *ex vivo* Raman spectral investigation was conducted to identify distinct compositional information of healthy (H), ductal carcinoma *in situ* (DCIS), and invasive ductal carcinoma (IDC). Then, principal component analysis-linear discriminant analysis (PCA-LDA) and principal component analysis-support vector machine (PCA-SVM) models were constructed for distinguishing spectral features among different tissue groups. Spectral analysis highlighted differences in levels of unsaturated and saturated lipids, carotenoids, protein, and nucleic acid between healthy and cancerous tissue and variations in the levels of nucleic acid, protein, and phenylalanine between DCIS and IDC. Both classification models were principal component analysis-linear discriminant analysis to be extremely efficient on discriminating tissue pathological types with 99% accuracy for PCA-LDA and 100%, 100%, and 96.7% for PCA-SVM analysis based on linear kernel, polynomial kernel, and radial basis function (RBF), respectively, while PCA-SVM algorithm greatly simplified the complexity of calculation without sacrificing performance. The present study demonstrates that Raman spectroscopy combined with multivariate analysis technology has considerable potential for improving the efficiency and performance of breast cancer diagnosis.

1. Introduction

Breast cancer is the most common cancer experienced by women worldwide [1]. In 2018, the number of new breast cancer cases and deaths recorded around the world reached 2,088,849 and 626,679, respectively [1]. In China, more than 50,000 women under the age of 40 are diagnosed with breast cancer every year [2], the incidence and mortality in young breast cancer patients rising continually. A preinvasive form of breast cancer, ductal carcinoma *in situ* (DCIS), is observed in approximately 20% of all tumors detected by screening mammography [3, 4]. Without the appropriate treatment, 20%–30% of DCIS cases progress to invasive ductal carcinoma (IDC) [3, 5–7], which invades the blood and lymph nodes, ultimately spreading to other body organs. Therefore, it is imperative to use reliable cancer screening techniques to identify suspected cases of DCIS from healthy and other malignant lesions so as to spare unnecessary treatment and prevent further misdiagnosis of IDC precursor, and the compositional information needs to be elucidated for understanding and predicting the transition and progression processes [8].

Currently, breast cancer screening is conducted principally with a triple assessment using imaging examination that integrates X-ray mammography and ultrasound, clinical tests, and histological assessment [9]. After cancer screening,

various methods of biopsy (i.e., needle or surgical biopsy) are used for additional histopathological assessment, aiming to identify the type and stage of cancer and to determine appropriate treatment protocols. However, current histopathological assessment is inevitably limited due to discrepancies in methods of fixation, staining, and antigen retrieval, in addition to the experiences of pathologists [9]. Additionally, genomic profiling has also been proposed in order to identify the subtype-specific classification of the breast cancer and provide a more accurate diagnosis [10, 11].

Raman microspectroscopy allows a qualitative and quantitative analysis of the chemical nature of biological samples which requires minimal sample preparation and does not require a staining process. After years of development, it has been widely accepted by clinicians and research communities for the early diagnosis of cancer, the identification of cancer progression, and intraoperative guidance [12–18]. Despite there being clear Raman bands probably related to the abundance of a variety of different biomolecules, it is still necessary to use multivariate statistical techniques to more reliably correlate the spectra with tissue pathological characteristics, especially for clinicians or biomedical researchers without a solid background in physics [19].

In the present study, we characterized the spectral variations in healthy (H), DCIS, and IDC tissues so as to identify the features in spectra caused by cancer progression and to facilitate the development of Raman-based tumor detection algorithms. Two multivariate analysis models, principal component analysis (PCA) followed by linear discriminate analysis (LDA) and support vector machine (SVM) analysis, respectively, were further utilized to analyze and classify Raman spectra in the three types of tissue. Following a comparison of the performance of PCA-LDA and PCA-SVM models, an effective algorithm was verified to further bridge the knowledge gap in identifying the appropriate model for Raman spectroscopy in the breast cancer diagnosis.

2. Experimental Section

2.1. Sample Preparation. A total of twelve healthy breast samples, which contain both collagenous and adipose tissue, from four female patients were purchased from Alenabio (Xi'an, Shaanxi, China), and biopsies were performed using protocols approved by the IRB (Institutional Review Board) and the HIPAA (Health Insurance Portability and Accountability Act). It was additionally approved as commercial product development. IDC ($n = 6$) and DCIS ($n = 6$) samples from twelve female patients with an average age of 50 years were obtained from clinical breast-conserving surgery in the Department of Breast Surgery, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China. The study was approved by the Ethics Committee of the First Affiliated Hospital of Xi'an Jiaotong University (Xi'an, China). This was a retrospective study, for which formal consent was not required and the lesion type is verified by the pathologist in the First Affiliated Hospital of Xi'an Jiaotong University. All research procedures, including sample collection, tissue section preparation, and spectral analysis, have complied with current laws in China.

Immediately after lesion excision, the samples were embedded in optimal cutting temperature medium (OCT, Surgipath® FSC 22®, Leica Biosystems, USA) and frozen in liquid nitrogen for better preservation of native morphology. 12 μm thick longitudinal sections were placed on gold-coated glass substrates (BioGold® 63479-AS, Electron Microscopy Sciences, USA) for spectroscopic analysis, which is used to eliminate the background fluorescence from the microscope slides and optics for spectroscopic measurement [20, 21]. Consecutive 5 μm thick sections were stained with hematoxylin and eosin (H&E) to facilitate a comparison of spectral measurements with histopathological results. Frozen sections were maintained in an acetone cooling bath for dehydration and storage at −20°C until transportation to Northwest University, Xi'an, China, for spectroscopic studies. Tissue sections were thawed for less than 30 min at room temperature prior to spectroscopic analysis or additional histological processing.

2.2. Spectroscopic Acquisition. The equipment used for Raman spectroscopy has been described in detail previously [13, 14]. Briefly, a single spectrum was collected using a WITec Alpha 500 confocal micro-Raman spectroscopy system (WITec GmbH, Germany) using a 633 nm He-Ne laser source (35 mW, Research Electro-Optics, Inc.). A 100× microscope objective (NA = 1.25, EC Epiplan-Neofluar, Zeiss, Germany) was used for spectral excitation and measurement. For each sample, several areas were selected for spectral measurement by pathologists, and 30–50 spectra were randomly measured according to the size of the region, and each region was measured two to three times. In total, 100 spectra were randomly acquired from each samples (H, DCIS, and IDC), respectively. Although we did not observe any interference background from OCT which is composed of water-soluble glycols and resins, the tissue boundary was avoided for spectral measurement to prevent any contaminations from cutting medium. Each spectrum was recorded over a period of 1.5 seconds using a Raman spectrometer (UHTS300, WITec GmbH, Germany) incorporating a 600 mm^{-1} grating with a back-illuminated deep-depletion charge-coupled device camera (Du401A-BR-DD-352, Andor Technology, UK) at a resolution of approximately 3 cm^{-1}.

2.3. Data Preprocessing and Analysis. WITec Project FOUR software (WITec GmbH, Germany) was used to preprocess all datasets that were obtained for band range selection, cosmic ray removal, background subtraction, and spectral smoothing, using the same parameters in each case. The background subtraction is achieved by a nine-order polynomial fit and we use a five-order Savitzky–Golay smoothing to noise removal. All Raman spectra were normalized using an area under the curve method over the ranges 600–1800 cm^{-1} and 2800–3000 cm^{-1} to minimize the effects of sample and instrument variability.

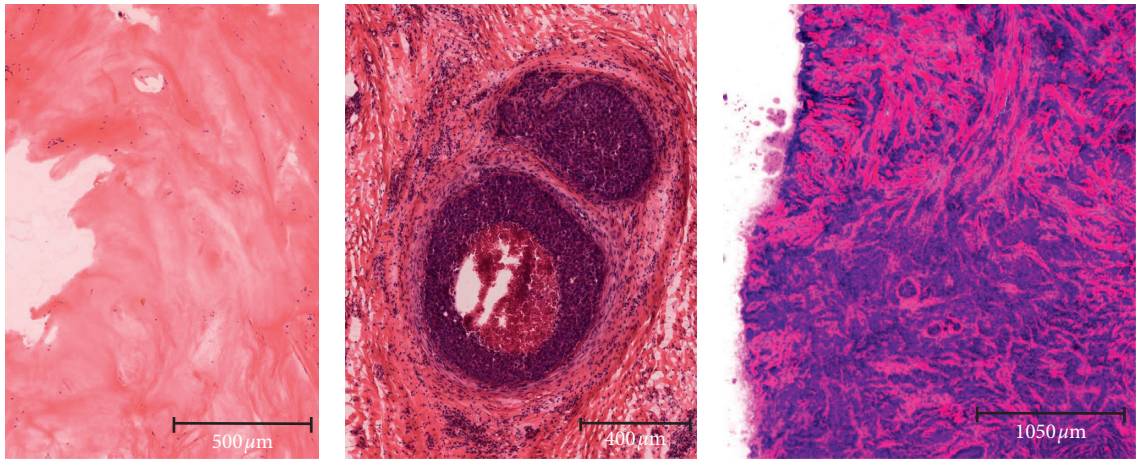


Figure 1: H&E-stained images of healthy breast tissue (a), ductal carcinoma *in situ* (b), and invasive ductal carcinoma (c) tissue with 500 μm, 400 μm, and 1050 μm scale bar, respectively.

The spectral datasets were mean-centered and then used to conduct additional analysis. PCA was used to simplify complexity and identify key variables in the multidimensional datasets [22, 23]. As a supervised classification method, LDA can be used in combination with PCA to improve the performance of classification. In addition, one-way ANOVA was used to extract the most diagnostically significant PCs ($P < 0.01$) for better classifying the three different types of tissue [24]. Significant PC scores were input into the LDA to generate a classification algorithm [25]. A leave-one-out cross-validation (LOOCV) technique was used to verify the performance of the diagnostic model based on the PCA-LDA algorithm for the classification of different tissue types.

Using Vapnik–Chervonenkis (VC) theory and the principle of structural risk minimization, an SVM algorithm was also adopted using PC scores as input variables to construct a PCA-SVM model. In the present study, three kernel types were tested in the PCA-SVM model, namely, a linear kernel, polynomial kernel, and Gaussian radial basis function (RBF). All acquired spectral data were divided into either a training (80%) or a testing set (20%) during testing. To obtain a model with the best performance, grid search combined with 10-fold cross-validation was employed to determine the most appropriate combination of parameters for each kernel. Finally, those parameters and the trained algorithms were used to construct the final PCA-SVM model and identify the unknown spectra. All statistical analyses were performed using Matlab R2015b software (Mathworks, Inc., Natick, MA, USA).

3. Results

3.1. Pathological Analysis. Using H&E-stained tissue sections, significant morphological differences were observed among the H, DCIS, and IDC tissues, representing pathological progression (Figure 1). The form of cancer type can be ascertained by the degree of infiltration of cancer cells represented by black particles in the H&E-stained images. Compared with healthy breast tissue, shown in Figure 1(a), cancer cells in DCIS tissue were distributed around the duct without breaking through the basement membrane, shown in Figure 1(b). Conversely, in IDC (Figure 1(c)), cancer cells did break through the basement membrane of the breast duct, exhibiting an anisotropic distribution pattern.

3.2. Raman Spectral Analysis. As shown in Figure 2(a), all three tissue types revealed characteristic peaks at 868 cm^{-1} (C-C stretching, collagen) [26], 1076 cm^{-1} (C-C stretching, lipid) [27], 1302 cm^{-1} (CH_2 twisting, wagging, phospholipids) [28], 1450 cm^{-1} (CH_2 deformation) [26], and 1654 cm^{-1} (C=C lipid stretching) [29]. However, compared with H and DCIS tissues, IDC displayed additional peaks at 669 cm^{-1} (T and G in nucleic acids) [30], 754 cm^{-1} (symmetric ring breathing in tryptophan, in protein) [27], 1552 cm^{-1} (C=C stretching in tryptophan) [28], and 1608 cm^{-1} (C=C stretching in phenylalanine) [31]. Furthermore, H and DCIS tissues exhibited some overlapping of the peak positions at 1267 cm^{-1} (lipids) and 2854 cm^{-1} (CH_2 symmetric stretch, lipids) [32], 2900 cm^{-1} (CH stretching, lipids and proteins) [32], and 2934 cm^{-1} (CH_2 anti-symmetric stretching in lipids) [33], while IDC exhibited identifiable peaks at 1243 cm^{-1} (CH_2 wagging, C-N stretching, amide III of collagen) [34] and 2878 cm^{-1} (CH_2 asymmetric stretch in lipids and proteins) [32].

To better identify the underlying compositional information for the different stages of breast cancer invasion, differential spectra were calculated by subtracting the acquired featured spectra from each tissue type, as shown in Figure 2(b). From the differential spectra for DCIS and H, positive peaks were visible at 1002 cm^{-1} (phenylalanine) [35, 36] and 2934 cm^{-1} (lipids and proteins) [37, 38], while

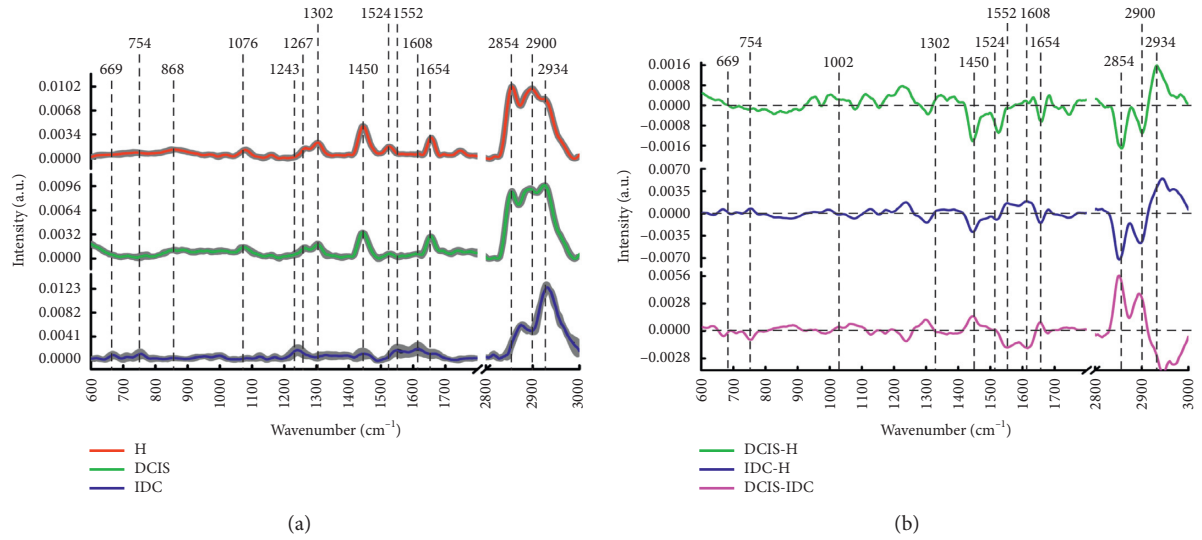


FIGURE 2: (a) The mean ± standard deviations (SD) of normalized spectra in H, DCIS, and IDC tissues; shading area represents standard deviations. (b) The differential spectra calculated from the normalized Raman spectra among different tissues.

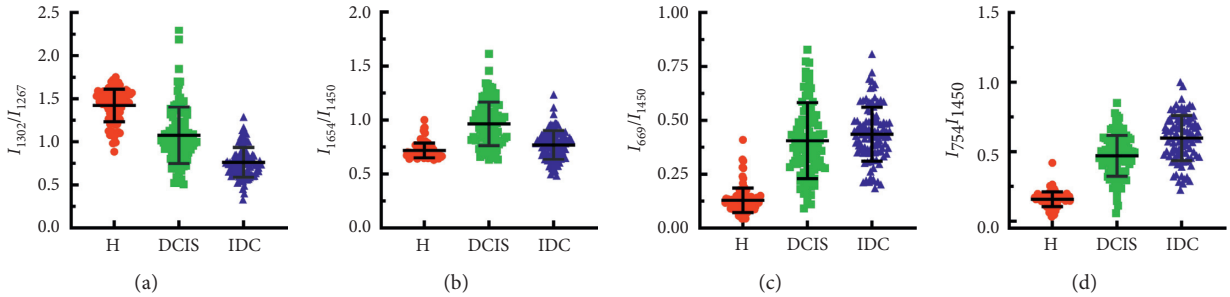


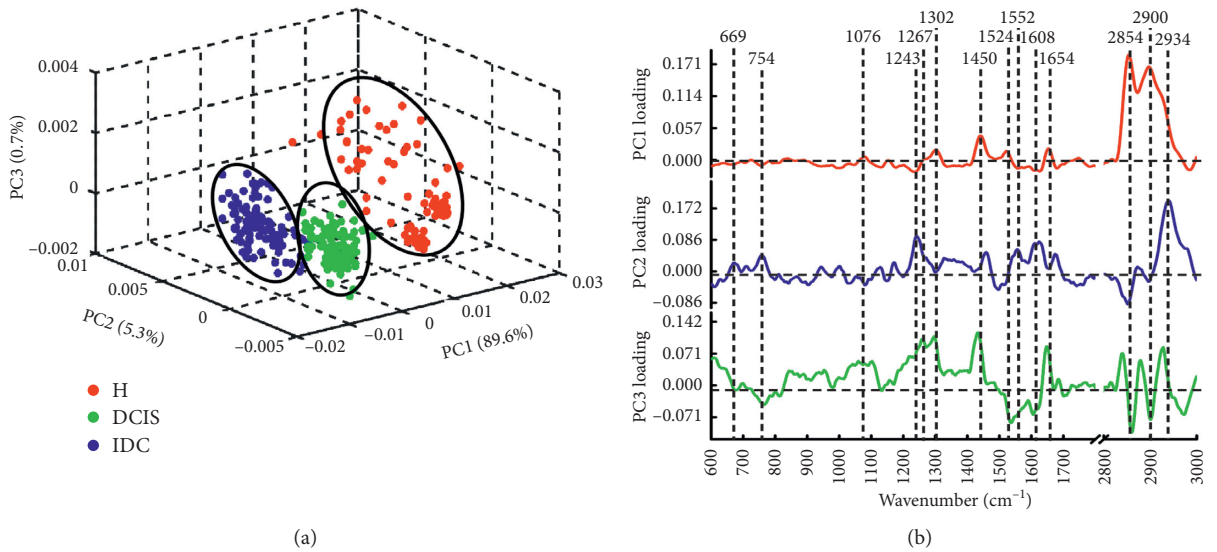
FIGURE 3: Comparisons among relative intensity ratios of the selected Raman bands with the corresponding tentative biochemical assignments of the tissue samples. All data are represented as mean ± standard deviation values. (a) Ratio for saturated lipid. (b) Ratio for unsaturated lipid. (c) Ratio for nucleic acid to lipid. (d) Ratio for protein to lipid.

negative peaks were observed at 1302, 1450, 1654, 2854, and 2900 cm^{-1} (lipids) and 1524 cm^{-1} (carotenoids), indicating that phenylalanine content increased while lipid and carotenoid levels decreased during the evolution of cancer from healthy tissue to DCIS.

Meanwhile, in the subtractive spectra for IDC and H tissues, positive peaks were observed at 669 cm^{-1} (nucleic acids), 754, 1243, 1552, and 1608 cm^{-1} (protein), and 2934 cm^{-1} (lipids and proteins), while there were negative peaks at 1302, 1450, 1654, 2854, and 2900 cm^{-1} (lipids) [39] and 1524 cm^{-1} (carotenoids) [40], indicating that nucleic acid and protein content in IDC was higher than in H tissue, but that of lipids and carotenoids was lower than in H tissue. Furthermore, in the differential spectrum between DCIS and IDC, negative peaks were observed at 669 cm^{-1} (nucleic acids), 754 cm^{-1}, 1243, and 1552 cm^{-1} (protein), 1608 cm^{-1} (phenylalanine), and 2934 cm^{-1} (lipids and protein), and positive peaks at 1302, 1450, 2854, and 2900 cm^{-1} (lipids). This suggests that the nucleic acid, protein, and phenylalanine levels were higher in IDC than in DCIS, but lipid levels in IDC were lower than in DCIS.

A ratio plot of Raman intensity of relevant specific wavenumbers is depicted in Figure 3. In accordance with the research of Zheng et al. [41], the intensity of the Raman peak ratios 1302 cm^{-1}/1267 cm^{-1} and 1654 cm^{-1}/1450 cm^{-1} can be used to evaluate levels of saturated and unsaturated lipids of breast tissue *in situ* and invasive cancers. The intensity of Raman peak ratios 669 cm^{-1}/1450 cm^{-1} and 754 cm^{-1}/1450 cm^{-1} indicates the change in proteins, nucleic acid, and lipid content as cancer progresses. The level of saturated lipids (Figure 3(a)) in cancerous tissues was found to be lower than in healthy tissues, while the level of unsaturated lipids (Figure 3(b)) was higher. The ratio of nucleic acid to lipid (Figure 3(c)) and protein to lipid (Figure 3(d)) gradually increased as the cancer in the breast invaded.

3.3. PCA-LDA Analysis. To identify the important variations within the acquired Raman data, multivariate analysis was conducted to distinguish spectral features characteristic of the different tissues. Raman spectra of



(a) (b)

Figure 4: (a) A scatter plot of the first three principal components acquired from the dataset consisting of all the collected spectra from three tissue types. (b) The corresponding PCA loading spectra of PC1, PC2, and PC3.

the low-wavenumber region (600–1800 cm^{-1}) and high-wavenumber region (2800–3000 cm^{-1}) were obtained from the three tissue sample types and categorized by PCA to obtain corresponding PC scores and loading values. The first PC accounted for the largest variance within the spectral dataset (PC1, 89.6%), while PC2 and PC3 represented 5.3% and 0.7% of the total variance, respectively. In order to visualize the spectral distribution of different tissues, Figure 4(a) represents a scatter plot of the first three PCs scores obtained from the spectral datasets of H, DCIS, and IDC groups following the PCA procedure. This demonstrates a clear separation between the spectra from the three tissue types, although the spectra of the DCIS and IDC groups overlapped slightly. In addition, PCA was able to not only distinguish the spectra of different tissues, but also extract molecular feature information related to their classification, depending on the corresponding loading spectra of each PC [42]. The loading for PC1, PC2, and PC3 is illustrated in Figure 4(b). The loading for PC1 was essentially above the zero line, with clear peak positions for carotenoid (1524 cm^{-1}) and lipid components (1076, 1302, 1450, 1654, 2854, and 2900 cm^{-1}); it can be seen that the PC1 contained more lipids and carotenoids. Compared to the loading of PC1 with the single spectrum of Figure 2, the features of PC1 loading were extremely similar to the spectral characteristics of H tissue (Figure 2(a)), suggesting that PC1 can be used primarily to differentiate the H tissue from the IDC and DCIS groups.

For the positive peaks of PC2, the corresponding loading can be assigned to biochemical components such as nucleic acids at 669 cm^{-1}, tryptophan at 754 and 1552 cm^{-1}, phenylalanine at 1608 cm^{-1}, collagen at 1243 cm^{-1}, and lipids at 2934 cm^{-1}, while negative peaks can be attributed to lipids at 1076, 2854, and 2900 cm^{-1}. In a comparison of PC2 loading with characteristic spectra of the three tissues, the characteristics of positive peak 669, 754, 1552, 1608, and 1243 cm^{-1} were obvious in IDC tissues, while the positive peak at 2934 cm^{-1} can be observed in all three tissue types. Therefore, positive features extracted by PC2 were principally derived from IDC tissue, while the negative features were mainly derived from the contribution of DCIS.

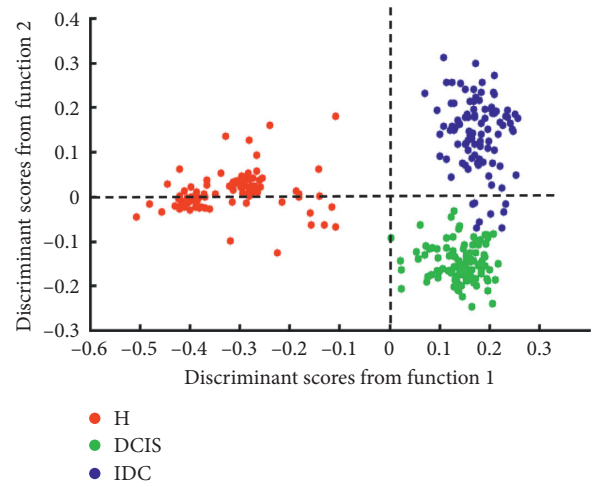


Figure 5: The scatter plot of linear discriminant scores for three types of tissue.

The loading of PC3 was evenly distributed on both sides of the zero line, but principal characteristic information in the negative loading appeared at 669, 754, 1552, 1608, 2854, and 2900 cm^{-1}, while positive peaks were observed at 1267, 1302, 1450, and 1654 cm^{-1}, peaks representing the spectral contribution of nucleic acids, proteins, and lipids. Component PC3 was rather noisy, displaying a mixture of spectral characteristics of both PC1 and PC2.

All three significant PCs were loaded into the LDA model for developing effective breast tissue diagnostic model. Figure 5

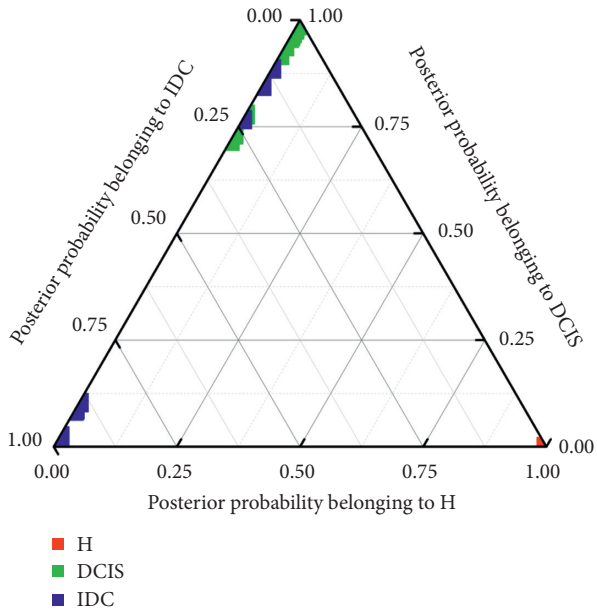


Figure 6: A two-dimensional ternary plot of the posterior probabilities belonging to the investigated H, DCIS, and IDC samples calculated from the acquired dataset consisting of all the collected spectra from three tissue types, using the PCA-LDA discriminant model combined with LOOCV method.

demonstrates the linear discriminant scores of all obtained spectral data in both fingerprint (600–1800 cm^{-1}) and high-wavenumber range (2800–3000 cm^{-1}) from investigated tissue types obtained by the PCA-LDA algorithm. The scatter plot of LDA discrimination distinguishes the spectra of the three tissue types, in which the zero line of the first discriminant function distinguishes the spectral feature of H group from that of cancerous tissue. The spectra of H group all distributed on the negative axis, while that of cancerous tissue appears on the positive axis. The spectra of the DCIS group were all represented on the negative axis of the second discriminant function, and the spectra of the IDC group on the positive axis. Thus, the zero line of the second discriminant function can separate the IDC group from the DCIS group. The posterior probabilities of H, DCIS, and IDC groups were also calculated and shown as a two-dimensional ternary scatter plot in Figure 6. The final diagnostic category of each data point is determined by the nearest proximity of data to the diagnostic category related to the vertex of the ternary plot. LOOCV was used to verify the performance of the classification model for the three different tissues. The overall accuracy of the PCA-LDA classification model was 99% (Table S1). For the fingerprint spectral region (600 to 1800 cm^{-1}) and the high-wavenumber region (2800–3000 cm^{-1}), the classification of three different types of breast tissues using PCA-LDA diagnostic model is shown in Figures S1 and S4 and Tables S5 and S9.

3.4. PCA-SVM Classification Model. To achieve an optimized classification performance in our study, SVM with three kernel functions (linear, polynomial, and RBF) was also implemented in the present study. In addition, PC1 and PC2 scores were used as input variables in the SVM model for visual classification. The optimal parameters of each kernel type were determined from the training set using a grid search program combined with cross-validation. In order to observe the influence of different parameters on classification accuracy in the training of the SVM model, three-dimensional surface maps of the different parameters and corresponding classification accuracy were constructed, as shown in Figures 7(a) and 7(b). For the RBF kernel in the PCA-SVM model, two parameters (C and γ) required optimization. In Figure 7(a), both C and parameter γ were plotted over the range 2^{-5} to 2^{5}, with a step of power of two. It can be observed that the accuracy of the RBF kernel in the PCA-SVM classification model gradually increased with increasing values for parameters C and γ. The greatest accuracy for the model was observed when $C = 0.5$ and $\gamma = 0.25$, at 99.38%. Using a polynomial kernel, two parameters also required optimization for the PCA-SVM model. In Figure 7(b), where parameter $C = 0.0313$ and polynomial order $d = 2$, the model achieved the highest classification accuracy, of 98.75%. For a linear kernel PCA-SVM model, only one parameter, C, required optimization, so the range 2^{-5} to 2^{5} was selected, with a step of power of two. Figure 7(c) displays the dependence of classification accuracy on parameter C. It was found that the highest accuracy of 98.75% was obtained with parameter $C = 0.0625$.

These optimized parameters were used to build the final SVM classification model to classify the spectra in the test set. The classification accuracy of the RBF kernel PCA-SVM model in the test set was 96.7%, while those of the linear and polynomial kernel PCA-SVM models were 100% and 100%, respectively. PCA-SVM diagnostic model was used to classify the test set data of three breast tissues, as shown in Tables S2, S3, and S4. This demonstrates that both linear and polynomial kernel PCA-SVM models perform better than the RBF kernel. The performance of the three kernel PCA-SVM models was slightly different compared with that of the training set, although differences were within the allowable range, demonstrating that the three types of SVM classification model displayed no apparent overfitting phenomena. In order to compare the performance of PCA-LDA classification model, PC1 and PC2 scores are used as input variables for SVM classification. The classification results of PCA-SVM with three different kernels of all spectroscopy data from all tissue types are shown in Figure 8. It can be seen that the different tissues are clearly separated, with some misclassified plots. In Figures 8(a) and 8(b), the linear and polynomial kernel PCA-SVM models separate the spectra of the DCIS group from the H and IDC groups. However, some spectra of the H group were misclassified as IDC, and partial IDC spectra were classified as DCIS. In Figure 8(c), the spectra of the H and DCIS groups were separated from the IDC group by RBF PCA-SVM. Some of the IDC spectra were mistakenly classified as DCIS. Additional details of PCA-SVM model for fingerprint region (600–1800 cm^{-1}) and high-wavenumber (2800–3000 cm^{-1}) region could be found in Figures S2, S3 and Figures S5, S6, and Tables S6–S8 and S10–S12 in the supplementary material.

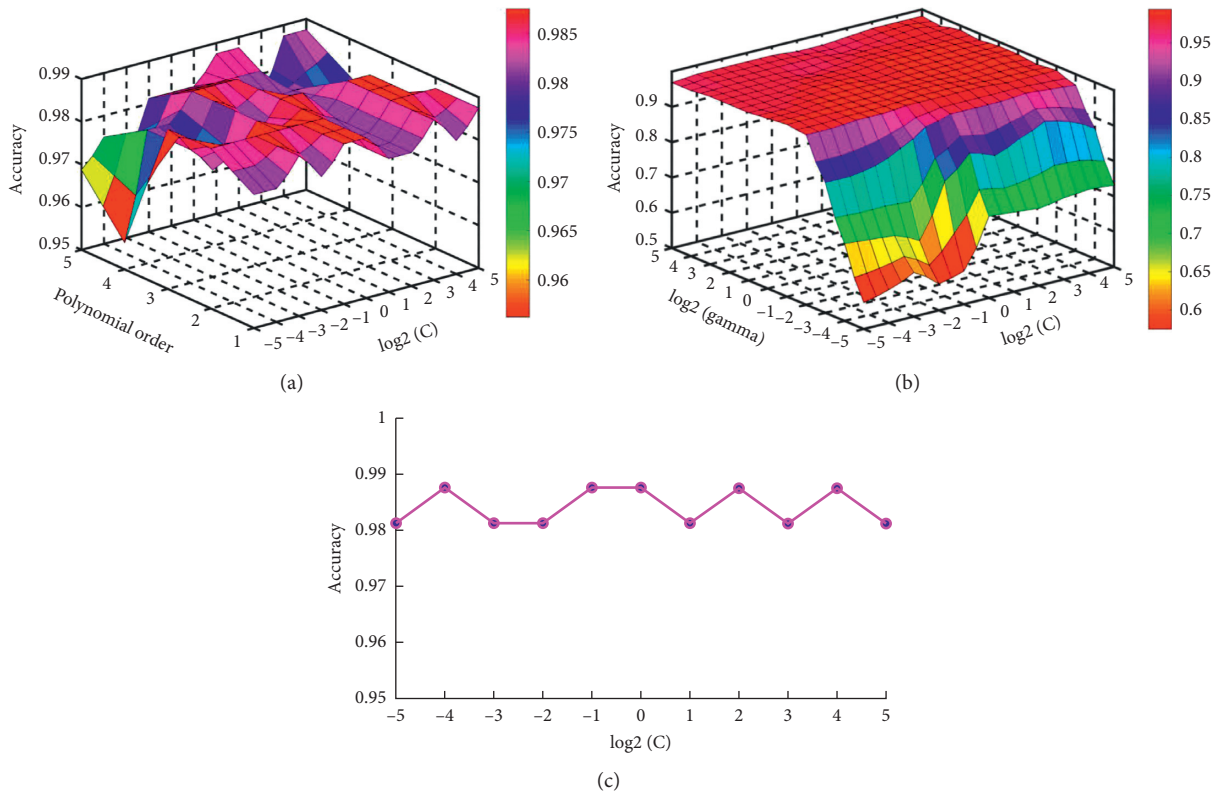


FIGURE 7: (a, b) The 3D map of classification accuracy as a function of parameter C and polynomial order d, parameter C and gamma γ, respectively. (c) Dependence of classification accuracy on parameter C for a linear PCA-SVM algorithm.

4. Discussion

Based on the calculated differential spectra in Figure 2(b), it is apparent that the intensity of lipids (1302, 1450, 1654, 2854, and 2900 cm^{-1}) was significantly lower in the DCIS and IDC tissues than in the H group. These results indicate that lipid content declined in the DCIS and IDC groups, possibly related to the high rate of cell division and the thinning of the lipid cell membranes in the process of cancer cell invasion and migration [43]. In addition, lipid peroxidation by reactive oxygen species (superoxide anion radicals, O^{-2}) or iron complexes in cancer tissues may also reduce lipid content [44]. In addition, the intensity of phenylalanine (1002, 1608 cm^{-1}) was higher in DCIS and IDC tissues than that in the H group, which may be associated with the large quantity of protein synthesized by cancer cells during uncontrolled growth and thence leads to an increase in phenylalanine levels [43]. In addition, the Raman peak at 1524 cm^{-1} (carotenoids) exhibited decreased intensity in DCIS tissue compared with healthy tissue, possibly attributable to the free radical oxidation of carotenoids [44]. As shown in Figure 2(b), higher nucleic acid (669 cm^{-1}) and protein (754 cm^{-1}) levels in the IDC group compared with the H group may be associated with the large quantity of protein synthesized by cancer cells during uncontrolled growth [43]. Furthermore, the uncontrolled proliferation of cancer cells increases intracellular DNA, and more nucleic acid molecules are contained in daughter cells due to augmented DNA replication, possibly explaining the increase in nucleic acid content in tumor tissues [45]. Similarly, the differential spectrum for DCIS and IDC displayed higher protein (754, 1552 cm^{-1}), phenylalanine (1608 cm^{-1}), and nucleic acid (669 cm^{-1}) content and lower lipid (1302, 1450, 1654, 2850, and 2900 cm^{-1}) levels in the IDC group compared with the DCIS group, consistent with features of cancer progression.

Ratio plots of saturated lipids (1302 cm^{-1}/1267 cm^{-1}), as displayed in Figure 3, displayed a decreasing trend during the invasion of breast cancer, suggesting that lipid metabolism may have a significant effect on the development of cancer [46]. However, the level of unsaturated lipids (1654 cm^{-1}/1450 cm^{-1}) in cancerous tissue was higher than that in healthy tissues, possibly related to lipid peroxidation during the development of breast cancer. As cancer invasion occurred, the ratio of nucleic acid to lipid (669 cm^{-1}/1450 cm^{-1}) and protein to lipid (754 cm^{-1}/1450 cm^{-1}) gradually increased, consistent with changes in nucleic acid and lipid levels for DCIS and IDC in Figure 2(b). It is difficult to distinguish DCIS tissue from healthy tissue just from a comparison of Raman peaks since their spectral characteristics are similar, with only slight differences in peak position and intensity. Meanwhile, the biological tissue provides a relative complex system to identify the peak assignments correctly, because of the overlapped spectral features in different molecules and the different spectral measurement apparatus. Different types of multivariate analysis algorithms are needed to extract a more reliable correlation of spectra with pathology [47].

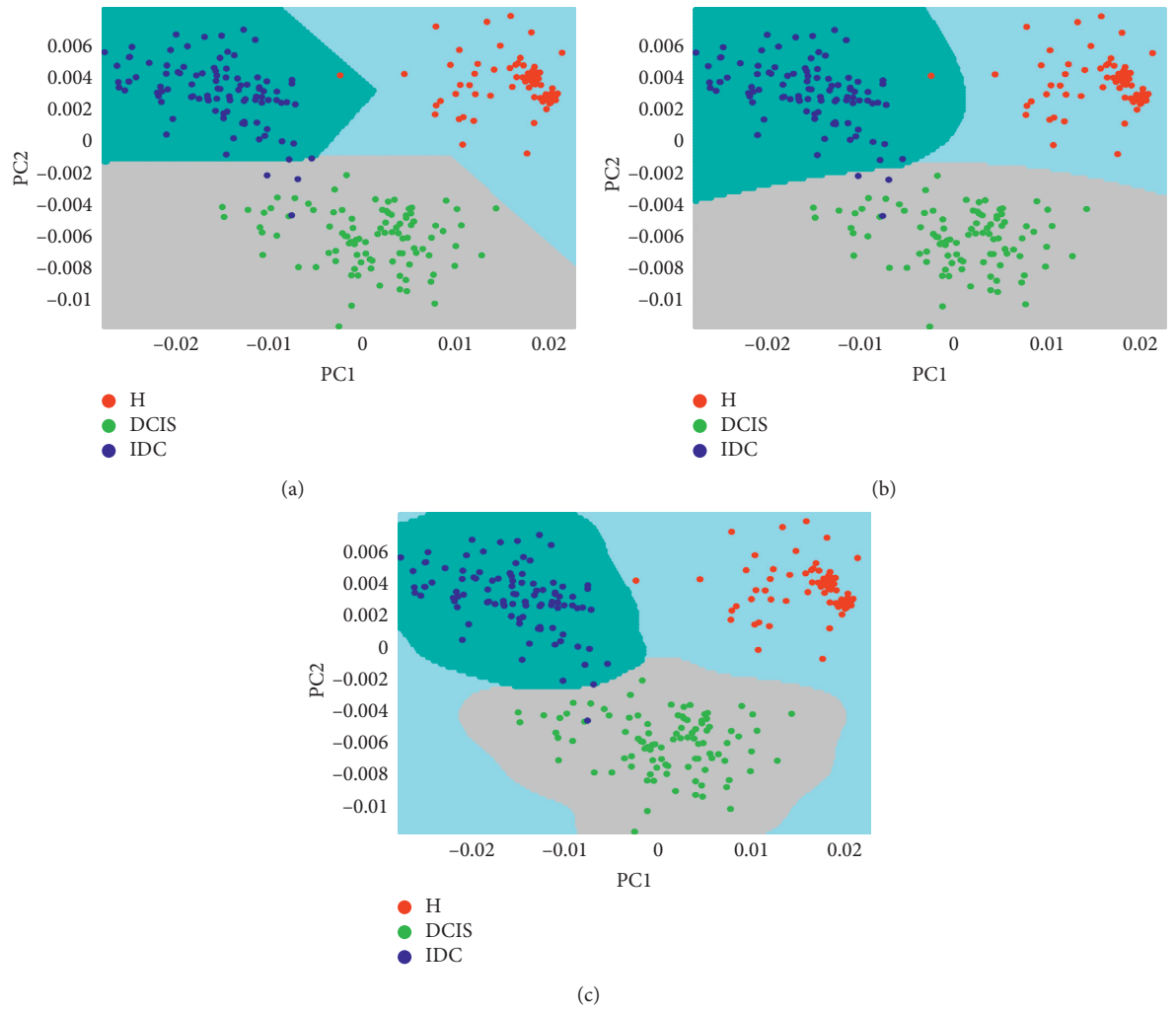


Figure 8: Scattering plots of PCA-SVM algorithm based on three kernel functions. (a) PCA-SVM with linear kernel, (b) PCA-SVM with polynomial kernel, and (c) PCA-SVM with RBF kernel. Points in different colors represent different tissue types; background color represents class domain created by SVM (a, linear kernel; b, polynomial kernel; c, RBF kernel).

The above analysis only employed a limited number of Raman peaks for tissue classification; however, many biochemical species would be involved in cancer evolution and progression. Therefore, multivariate statistical analysis method (such as PCA-LDA), which identifies the most significant spectral features from the whole spectrum, improves the diagnostic efficiency of Raman-based tissue analysis and classification. The advantage of PCA-LDA is that the modeler can query the spectral variables using principal components selected in the model to provide a source for classification [48]. By comparing the loading of PC1 with the normalized spectra of H tissue in Figure 2(a), it was found that the two spectra maintained a high degree of consistency, so it is believed that PC1 can distinguish H and cancer tissues. The loading of PC2 contained more nucleic acid and protein, consistent with the results of the differential spectra of DCIS and IDC in Figure 2(b). Therefore, PCA was effective in analyzing the molecular composition and biochemical differences within the spectral dataset, similar to the results presented by the differential spectrum. In the LOOCV confusion matrix (Table S1), a partial misclassified spectrum of the three groups was apparent, possibly due to similar spectral features of H, DCIS, and IDC tissues.

SVM is an additional multivariable analysis technique able to manage linear and nonlinear separable data. For SVM, the most important operation is to choose an appropriate kernel function and parameter optimization strategy, critical for the development of a robust model. When choosing the kernel function, it should first be considered whether the data is linearly separable, and the optimization should maximize the accuracy and minimize the complexity of the model. Compared with other multivariate statistical methods, SVM can deal with class boundaries under complex condition by replacing kernel functions. In the present study, an SVM model with three traditional kernels was developed by using PCA algorithm to reduce the dimension of spectral data, which greatly simplifies the SVM algorithm and improves its performance. The results indicated that the PCA-SVM model with linear

and polynomial kernels had the best classification performance, followed by the PCA-LDA method. The performance of the linear and polynomial kernel PCA-SVM models was slightly higher than that of PCA-LDA, possibly due to the use of a hyperplane to separate the classes [49]. This result also confirms that PCA combined with SVM not only simplifies the computational complexity of SVM, but also ensures that unknown spectra can be identified effectively. Conversely, where SVM is used for direct spectral classification, it is likely to fail to achieve acceptable performance due to considerable redundant information in the high-dimensional spectral datasets, with real-time applications unable to utilize the model because of excessive computation required to perform SVM.

5. Conclusions

In conclusion, the present study demonstrated that significant biochemical differences in breast cancer can be observed by Raman spectroscopy. Compared with H tissue, the content of protein and nucleic acid in DCIS and IDC tissue was higher, while the composition of lipids and carotenoids was lower or had even disappeared. Combined with multivariate analysis, the spectral characteristics in the H, DCIS, and IDC groups were further extracted by PCA loading and score plots. We also confirmed that the tissue classification model based on the PCA-LDA algorithm, together with LOOCV, was able to distinguish three different breast tissue types. In addition, a PCA-SVM diagnostic technique was developed with different kernel functions and comprehensive evaluation and comparison of the diagnostic performance were performed. This method greatly simplified the complexity of calculation without sacrificing the performance of the algorithm. The linear and polynomial PCA-SVM algorithm was superior to the PCA-LDA algorithm for classification of the spectra in breast tissue, indicating that it has great diagnostic potential in future applications. Therefore, the study confirmed the feasibility of Raman spectroscopy combined with multivariate analysis for the diagnosis of breast cancer.

Although our presented work or other groups' achievements has already demonstrated that Raman spectroscopy benefits early cancer diagnosis and pathological studies in a noninvasive way or without sample preparation procedures, there are still some practical issues that should be noted. Firstly, individual spectral diversity is a particularly prominent factor for making appropriate final diagnostic decisions; therefore, it is necessary to adopt machine learning method for accurately classifying the spectral features among different tissue types, and provide a reference for treatment practice. In this context, continued efforts are highly required to facilitate the transition from a Raman benchtop (micro)spectroscopy to bedside by developing advanced detection methodologies for bridging the gap between experimental studies and clinical practices. The improvement of newly developed Raman instrument would be symbolized by a high signal-to-noise ratio with automatic data analysis techniques, allowing fast, earlier, and more accurate diagnosis. Meanwhile, more *ex vivo* spectra-pathology studies are still needed to enrich our understanding of spectral information among different lesion types and grades, as well as invasive processes. To accomplish this, multidisciplinary collaboration between researchers and practitioners is required to establish the standardization of spectral data and increasing the clinical confidence.

Data Availability

The spectroscopic data used to support the findings of this study were supplied by Dr. Shuang Wang under license and so cannot be made freely available. Requests for access to these data should be made to Dr. Shuang Wang (swang@nwu.edu.cn).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61911530695) and Science Development Foundation of Shaanxi Province, China (2020KW-055).

Supplementary Materials

The supplementary documents explain the mathematical principal of PCA-SVM model and provide its evaluation results in both fingerprint and high-wavenumber region. (*Supplementary Materials*)

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] R. Guo, J. Si, J. Si et al., "Changing patterns and survival improvements of young breast cancer in China and SEER database, 1999–2017," *Chinese Journal of Cancer Research*, vol. 31, no. 4, pp. 653–662, 2019.

[3] R. Lesurf, M. R. Aure, H. H. Mørk et al., "Molecular features of subtype-specific progression from ductal carcinoma in situ to invasive breast cancer," *Cell Reports*, vol. 16, no. 4, pp. 1166–1179, 2016.

[4] H. G. Welch, S. Woloshin, and L. M. Schwartz, "The sea of uncertainty surrounding ductal carcinoma in situ—the price of screening mammography," *JNCI Journal of the National Cancer Institute*, vol. 100, no. 4, pp. 228-229, 2008.

[5] M. E. Sanders, P. A. Schuyler, W. D. Dupont, and D. L. Page, "The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up," *Cancer*, vol. 103, no. 12, pp. 2481–2484, 2005.

[6] C. R. Correa, P. Mcgale, C. Taylor et al., "Overview of the randomized trials of radiotherapy in ductal carcinoma in situ of the breast," *Journal of The National Cancer Institute Monographs*, vol. 2010, no. 41, pp. 162–177, 2010.

[7] A. Renshaw, "Rubin's pathology. Clinicopathologic foundations of medicine," *Advances in Anatomic Pathology*, vol. 15, no. 2, p. 125, 2008.

[8] C. F. Cowell, B. Weigelt, R. A. Sakr et al., "Progression from ductal carcinoma in situ to invasive breast cancer: revisited," *Molecular Oncology*, vol. 7, no. 5, pp. 859–869, 2013.

[9] S. Akira, S. Uematsu, and O. Takeuchi, "Pathogen recognition and innate immunity," *Cell*, vol. 124, no. 4, pp. 783–801, 2006.

[10] M. Alizart, J. Saunus, M. Cummings, and S. R. Lakhani, "Molecular classification of breast carcinoma," *Diagnostic Histopathology*, vol. 18, no. 3, pp. 97–103, 2012.

[11] L. Pusztai, C. Mazouni, K. Anderson, Y. Wu, and W. F. Symmans, "Molecular classification of breast cancer: limitations and potential," *The Oncologist*, vol. 11, no. 8, pp. 868–877, 2006.

[12] G. W. Auner, S. K. Koya, C. Huang et al., "Applications of Raman spectroscopy in cancer diagnosis," *Cancer and Metastasis Reviews*, vol. 37, no. 4, pp. 691–717, 2018.

[13] D. Song, F. Yu, S. Chen et al., "Raman spectroscopy combined with multivariate analysis to study the biochemical mechanism of lung cancer microwave ablation," *Biomedical Optics Express*, vol. 11, no. 2, pp. 1061–1072, 2020.

[14] D. Song, T. Chen, S. Wang et al., "Study on the biochemical mechanisms of the micro-wave ablation treatment of lung cancer by ex vivo confocal Raman microspectral imaging," *The Analyst*, vol. 145, no. 2, pp. 626–635, 2020.

[15] I. P. Santos, E. M. Barroso, T. C. Bakker Schut et al., "Raman spectroscopy for cancer detection and cancer surgery guidance: translation to the clinics," *The Analyst*, vol. 142, no. 17, pp. 3025–3047, 2017.

[16] C. A. Lieber, S. K. Majumder, D. D. Billheimer, L. M. D. D. Ellis, and A. Mahadevanjansen, "Raman microspectroscopy for skin cancer detection in vitro," *Journal of Biomedical Optics*, vol. 13, no. 2, 024013 pages, 2008.

[17] E. Kaznowska, J. Depciuch, K. Łach et al., "The classification of lung cancers and their degree of malignancy by FTIR, PCA-LDA analysis, and a physics-based computational model," *Talanta*, vol. 186, no. 15, pp. 337–345, 2018.

[18] J. Depciuch, E. Kaznowska, S. Golowski et al., "Monitoring breast cancer treatment using a Fourier transform infrared spectroscopy-based computational model," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 143, no. 5, pp. 261–268, 2017.

[19] V. Garla, C. Taylor, and C. Brandt, "Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 869–875, 2013.

[20] S. Wang, Z. Liang, Y. Gong et al., "Confocal raman microspectral imaging of ex vivo human spinal cord tissue," *Journal of Photochemistry and Photobiology B: Biology*, vol. 163, pp. 177–184, 2016.

[21] S. Wang, J. Zhao, H. Lui, Q. He, and H. Zeng, "A modular Raman microspectroscopy system for biological tissue analysis," *Spectroscopy*, vol. 24, no. 6, pp. 577–583, 2010.

[22] S. Feng, R. Chen, J. Lin et al., "Nasopharyngeal cancer detection based on blood plasma surface-enhanced Raman spectroscopy and multivariate analysis," *Biosensors and Bioelectronics*, vol. 25, no. 11, pp. 2414–2419, 2010.

[23] D. Song, Y. Chen, J. Li, H. Wang, T. Ning, and S. Wang, "A graphical user interface (NWUSA) for Raman spectral processing, analysis and feature recognition," *Journal of Biophotonics*, vol. e202000456, 2021.

[24] G. R. Lloyd, L. E. Orr, J. Christie-Brown et al., "Discrimination between benign, primary and secondary malignancies in lymph nodes from the head and neck utilising Raman spectroscopy and multivariate analysis," *The Analyst*, vol. 138, no. 14, pp. 3900–3908, 2013.

[25] Y. Li, J. Pan, G. Chen et al., "Micro-Raman spectroscopy study of cancerous and normal nasopharyngeal tissues," *Journal of Biomedical Optics*, vol. 18, no. 2, 027003 pages, 2013.

[26] G. Shetty, C. Kendall, N. Shepherd, N. Stone, and H. Barr, "Raman spectroscopy: elucidation of biochemical changes in carcinogenesis of oesophagus," *British Journal of Cancer*, vol. 94, no. 10, pp. 1460–1464, 2006.

[27] N. Huang, M. Short, J. Zhao et al., "Full range characterization of the Raman spectra of organs in a murine model," *Optics Express*, vol. 19, no. 23, pp. 22892–22909, 2011.

[28] Z. Huang, A. Mcwilliams, H. Lui, D. I. Mclean, S. Lam, and H. Zeng, "Near-infrared Raman spectroscopy for optical diagnosis of lung cancer," *International Journal of Cancer*, vol. 107, no. 6, pp. 1047–1052, 2003.

[29] N. Stone, C. Kendall, J. Smith, P. Crow, and H. Barr, "Raman spectroscopy for identification of epithelial cancers," *Faraday Discussions*, vol. 126, pp. 141–157, 2004.

[30] W. Huang, S. Wu, M. Chen et al., "Study of both fingerprint and high wavenumber Raman spectroscopy of pathological nasopharyngeal tissues," *Journal of Raman Spectroscopy*, vol. 46, no. 6, pp. 537–544, 2015.

[31] Z. Liu, C. Davis, W. Cai, L. He, X. Chen, and H. Dai, "Circulation and long-term fate of functionalized, biocompatible single-walled carbon nanotubes in mice probed by Raman spectroscopy," *Proceedings of the National Academy of Sciences*, vol. 105, no. 5, pp. 1410–1415, 2008.

[32] S. Koljenovic, T. C. B. Schut, A. J. P. E. Vincent, J. M. Kros, and G. J. Puppels, "Detection of meningioma in dura mater by Raman spectroscopy," *Analytical Chemistry*, vol. 77, no. 24, pp. 7958–7965, 2005.

[33] A. F. García-Flores, L. Raniero, R. A. Canevari et al., "High-wavenumber FT-Raman spectroscopy for in vivo and ex vivo measurements of breast cancer," *Theoretical Chemistry Accounts*, vol. 130, no. 4–6, pp. 1231–1238, 2011.

[34] L. M. Almond, J. Hutchings, G. Lloyd et al., "Endoscopic Raman spectroscopy enables objective diagnosis of dysplasia in Barrett's esophagus," *Gastrointestinal Endoscopy*, vol. 79, no. 1, pp. 37–45, 2014.

[35] D. W. Shipp, E. A. Rakha, A. Koloydenko, R. D. Macmillan, I. O. Ellis, and I. Notingher, "Intra-operative spectroscopic assessment of surgical margins during breast conserving surgery," *Breast Cancer Research*, vol. 20, no. 1, pp. 1–14, 2018.

[36] N. Stone, C. Kendall, N. Shepherd, P. Crow, and H. Barr, "Near-infrared Raman spectroscopy for the classification of epithelial pre-cancers and cancers," *Journal of Raman Spectroscopy*, vol. 33, no. 7, pp. 564–573, 2002.

[37] H. Fabian, N. A. N. Thi, M. Eiden, P. Lasch, J. Schmitt, and D. Naumann, "Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1758, no. 7, pp. 874–882, 2006.

[38] C. Krafft, L. Neudert, T. Simat, and R. Salzer, "Near infrared Raman spectra of human brain lipids," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 61, no. 7, pp. 1529–1535, 2005.

[39] G. Quintás, S. Garrigues, A. Pastor, and M. de la Guardia, "FT-Raman determination of Mepiquat chloride in agrochemical products," *Vibrational Spectroscopy*, vol. 36, no. 1, pp. 41–46, 2004.

[40] H. Abramczyk, B. Brozek-Pluska, J. Surmacki, J. Jablonska, and R. Kordek, "The label-free Raman imaging of human

breast cancer," *Journal of Molecular Liquids*, vol. 164, no. 1-2, pp. 123–131, 2011.

[41] Q. Zheng, J. Li, L. Yang et al., "Raman spectroscopy as a potential diagnostic tool to analyse biochemical alterations in lung cancer," *The Analyst*, vol. 145, no. 2, pp. 385–392, 2020.

[42] F. Bonnier and H. J. Byrne, "Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems," *The Analyst*, vol. 137, no. 2, pp. 322–332, 2012.

[43] T. Bhattacharjee, L. C. Fontana, L. Raniero, and J. Ferreira-Strixino, "In vivo Raman spectroscopy of breast tumors prephotodynamic and postphotodynamic therapy," *Journal of Raman Spectroscopy*, vol. 49, no. 5, pp. 786–791, 2018.

[44] B. Brozekpluska, I. Placek, K. Kurczewski, Z. Morawiec, M. Tazbir, and H. Abramczyk, "Breast cancer diagnostics by Raman spectroscopy," *Journal of Molecular Liquids*, vol. 141, no. 3, pp. 145–148, 2008.

[45] Y. Chen, J. Dai, X. Zhou, Y. Liu, W. Zhang, and G. Peng, "Raman spectroscopy analysis of the biochemical characteristics of molecules associated with the malignant transformation of gastric mucosa," *PLoS One*, vol. 9, no. 4, Article ID e93906, 2014.

[46] O. F. Kuzu, M. A. Noory, and G. P. Robertson, "The role of cholesterol in cancer," *Cancer Research*, vol. 76, no. 8, pp. 2063–2070, 2016.

[47] H. C. Mcgregor, M. A. Short, A. Mcwilliams et al., "Real-time endoscopic Raman spectroscopy forin vivoearly lung cancer detection," *Journal of Biophotonics*, vol. 10, no. 1, pp. 98–110, 2017.

[48] A. Maguire, I. Vega-Carrascal, J. Bryant et al., "Competitive evaluation of data mining algorithms for use in classification of leukocyte subtypes with Raman microspectroscopy," *The Analyst*, vol. 140, no. 7, pp. 2473–2481, 2015.

[49] S. Li, Q. Zeng, L. Li et al., "Study of support vector machine and serum surface-enhanced Raman spectroscopy for non-invasive esophageal cancer detection," *Journal of Biomedical Optics*, vol. 18, no. 2, 027008 pages, 2013.