Taylor & Francis
Taylor & Francis Group

ORIGINAL RESEARCH

OPEN ACCESS | Check for updates

Advancing cancer diagnostics with artificial intelligence and spectroscopy: identifying chemical changes associated with breast cancer

Abdullah C.S. Talari[a], Shazza Rehman[b] and Ihtesham U Rehman[a]

[a]Department of Engineering, Lancaster University, Lancaster, UK; [b]Department of Medical Oncology, Airedale NHS Foundation Trust, Airedale General Hospital, Steeton, UK

ABSTRACT

Background: Artificial intelligence (AI) and machine learning (ML) approaches in combination with Raman spectroscopy (RS) to obtain accurate medical diagnosis and decision-making is a way forward for understanding not only the chemical pathway to the progression of disease, but also for tailor-made personalized medicine. These processes remove unwanted affects in the spectra such as noise, fluorescence and normalization, and help in the optimization of spectral data by employing chemometrics.
Methods: In this study, breast cancer tissues have been analyzed by RS in conjunction with principal component (PCA) and linear discriminate (LDA) analyses. Tissue microarray (TMA) breast biopsies were investigated using RS and chemometric methods and classified breast biopsies into luminal A, luminal B, HER2, and triple negative subtypes.
Results: Supervised and unsupervised algorithms were applied on biopsy data to explore intra and inter data set biochemical changes associated with lipids, collagen, and nucleic acid content. LDA predicted specificity accuracy of luminal A, luminal B, HER2, and triple negative subtypes were 70%, 100%, 90%, and 96.7%, respectively.
Conclusion: It is envisaged that a combination of RS with AI and ML may create a precise and accurate real-time methodology for cancer diagnosis and monitoring.

1. Introduction

Incidence of cancer has arisen enormously during the last decade, estimated to have 18.1 million new cases and 9.6 million deaths in 2018. According to latest global cancer data, 1 in 5 men and 1 in 6 women worldwide develop cancer during their lifetime, and 1 in 8 men and 1 in 11 women die from the disease. Worldwide, the total number of people who are alive within 5 years of a cancer diagnosis, called the 5-year prevalence, is estimated to be 43.8 million [1]. Cancers of the lung, female breast, and colorectum are the top three cancer types in terms of incidence, and are ranked within the top five in terms of mortality (first, fifth, and second, respectively). Together, these three cancer types are responsible for one-third of the cancer incidence and mortality burden worldwide. Female breast cancer ranks as the fifth leading cause of death (627,000 deaths, 6.6%) because the prognosis is relatively favorable, at least in more developed countries. In women, incidence rates for breast cancer far exceed those for other cancers in both developed and developing countries, followed by colorectal cancer in developed countries and cervical cancer in developing countries [1]. Breast cancer is the most commonly diagnosed cancer in women (24.2%, i.e. about 1 in 4 of all new cancer cases diagnosed in women worldwide are breast cancer), and the cancer is the most common in 154 of the 185 countries included in GLOBOCAN 2018 [2]. Breast cancer is also the leading cause of cancer death in women (15.0%). The major limitations of current breast cancer screening and diagnostic

approaches are false positives, time delays, needle and open surgical biopsies, pain and anxiety and harms such as physical, emotional, social, and financial [3].

Raman spectroscopy (RS) has attracted considerable attention in recent years due to several practical reasons. This method does not require any dyes or probes to investigate molecular and chemical changes in biological tissues. Raman spectra are information-rich and chemical fingerprint can be used to quantify multiple components of the sample using same data set [4]. The principle is based on inelastic light scattering, also known as Raman scattering, depends on the change in charge distribution that give rise to the molecular polarizability [5]. Raman scattered light, including stokes and anti-stokes transmitted through the optical filter is directed to a diffraction grating to obtain its composite wavelengths. Position and intensity of each dispersed, composite wavelength is translated into the Raman spectrum [6]. These characteristic Raman peaks interpret not only information about biological components of the cell but also their quality, quantity, symmetry, and orientation the compound. Recently, a combination of Raman microspectroscopy and machine learning techniques have been used successfully in the identification of various cancers such as lung, breast, cervix, stomach, brain, skin, oral, leukemia, and bladder cancer [7–24]. Initial RS studies used near-infrared and visible lasers to identify biochemical profiles of human breast biopsies. Lipid and protein content of breast tissues were found to be markers for benign

and malignant states [25–28]. Raman was used to detect different grades of breast cancer by using qualitative and quantitative chemical information. Micro-spectroscopic model was built using nine basic components of breast tissue. Modeling method has exactly mimicked the breast tissue and these models were used to predict and characterize the disease status of the breast biopsies. These studies have established correlation between biochemical composition and disease aggressiveness [29].

RS accompanied by enhanced AI and ML have become powerful tools in prediction of disease more accurately. Our previous studies have used AI- and ML-based systematic intelligent control-based approaches for identification of normal cell lines from breast cancer lines, hypoxic region from proliferation region, and even identification of single cell in 3D scaffold. These approaches have helped in visualization of multiple components of single cell and tissue biopsies by means of fuzzy control, genetic algorithm, PCA, LDA, and cluster analysis [30–32]. In our previous studies, RS and mathematical modeling approaches were employed to characterize and differentiate two breast cancer and one normal breast cell lines (MDA-MB-436, MCF-7, and MCF-10A). Spectral features were detectable in high resolution and RS also offered information at the cellular level. RS was used to measure the different cell lines without the use of an external label and potentially, non-invasively. Spectra of the cell lines have revealed differences in the concentration of biochemical compounds such as lipids, nucleic acids, and proteins. Raman peaks were found to differ in intensity and PCA was able to identify variations that lead to accurate and reliable separation of the three cell lines. LDA model of three cell lines was predicted with 100% sensitivity and 91% specificity. The specificity of this approach is shown to be sensitive enough to detect cellular components of within cells [30]. Our studies were extended from cell lines to multicellular spheroids. Tumor spheroid model was developed, and it mimics the characteristics of non-vascular in vitro tumor model. A combination of Raman and AI and ML approaches were used to identify chemical changes associated with normal proliferating, hypoxic, and necrotic regions of T-47D human breast cancer spheroid model. Raman markers such as low amide I and high tryptophan content were differentiating three regions of spheroids [31]. RS was also employed to identify chemical changes in ductal carcinoma in situ and invasive carcinoma using breast tissue biopsies. Spectral features of carcinoma biopsies have shown clear shift toward proteins and backwards to lipids and triacylglycerides. Different grades of ductal carcinoma in situ (DCIS) have shown spectral variations in asymmetric and symmetric vibrations of lipids at high-frequency region and phosphodiester vibrations in fingerprint region [33]. In our current study, for the first time, a combination of Raman and AI and ML approaches have been applied on different breast cancer biopsies of TMA to identify chemical changes associated with different subtypes of breast cancer. Characteristic Raman markers were used to differentiate luminal A, luminal B, HER2 (Human epidermal growth factor receptor 2) positive, and triple negative subtypes of breast cancer. Furthermore, intra and inter biochemical changes of within subtypes using different spectral features, was also identified.

2. Materials and methods

2.1. Sample quantity and preparation

TMA approach has gained much attention in pathological research analysis. High-throughput analysis of many biopsies in a single slide can be achieved in short time. It increases research effectiveness with minimal quantities of tissues [34]. TMA slide with different breast cancer patient biopsies was used for Raman analysis. TMA slide and ethical approval was provided by Clinical Trials Research Unit (CTRU), University of Leeds. Clinical trial name was Azure and trial Data Transfer Agreement (DTA) ID is Azure_07. TMA slide has provided information about patient trial number, ER (Estrogen receptor) status, PR (Progesterone receptor) status, HER2 status, histological grade, and tumor pathology. ER and PR status has described as 0–8, where 0–2 is regarded as negative, sometimes 3 is also considered as negative. HER status has descried as 0–3, where 0 and 1 is negative, and 3 is positive. Score 2 are equivocal and need to perform an in situ hybridization method to ascertain true status. Approximately 30% or slightly less of 2+ are amplified and therefore considered as positive. Histological grade is represented as low grade, moderate, and high grade [35]. Tumor pathology of breast biopsies were ductal No Special Type (NST), lobular, mixed (ductal/lobular), mucinous, tubular, and other types. Majority of breast biopsies belong to ductal NST. A total 132 biopsy samples were present on TMA slide and known ER, PR and HER2 status biopsies were used for spectral collection. Thirty spectra were collected from each biopsy and in total 3,960 spectra were collected from patient biopsies. TMA section was mounted onto a glass slide and firmly fixed. De-waxing was achieved using xylene treatment for 30 min, 50% alcohol for 5 min, 70% alcohol for 5 min, and 100% alcohol for 5 min. The aim of xylene treatment is to remove the wax from the sample and alcohol treatment rehydrates the sample [31].

2.2. Raman spectral measurements

High quality Raman spectra were collected from TMA breast biopsies using a non-invasive dispersive micro-Raman system dxr™ (Thermo Nicolet) equipped with a 532 nm laser. Before collecting spectral data from biopsies, the system was fully calibrated using polystyrene as standard. The sensitivity in terms of signal to noise ratio of the laser was 1000:1. Laser power of 10 mw was used and it has not damaged to biological materials, and it has demonstrated in previous communications. A 50× long working distance objective was used and a spectrograph aperture was set to 50-micron pinhole. Spectral collection exposure time was set to 50 s with five exposures. Twenty spectra were collected from each cell line over the spectral range 400–3200 cm^{-1}. Thermo Nicolet OMNIC™ software was employed for data acquisition.

2.3. Data processing and analysis

Spectral range was analyzed between 600 and 3400 cm^{-1}. Breast biopsies were examined using DXR Raman confocal microscope. A total of 400 spectra were collected for this study and mean spectrum of each subtype was collected for comparison studies. AI approach involved in noise filtering, fluorescence removal, and spectrum normalization. Substrate subtraction and peak

measurements were accomplished using OMNIC Atlμs™ software and TQ Analyst (Thermo Scientific, Madison, WI, USA). Data analysis was performed using Unscrambler X 10.2 software (Camo software, Oslo, Norway). Chemometric methods such as PCA, LDA, and cluster analysis (CA) were applied in data analysis. The data were first base line corrected and unit vector normalized (UVN). Every PCA was setup with a minimum of 10 orthogonal variables depending on the spectral region used, where the number of PCs chosen for each setup described >99% of the variation. PCA was performed on high-wavenumber (3200–2600 cm^{-1}), amide I (1750–1500 cm^{-1}), amide III (1400–1200 cm^{-1}), and nucleic acid (980–610 cm^{-1}) regions. CA was performed on full spectral range (3200–400 cm^{-1}) using Wards' method squared Euclidean distance. Linear discrimination analysis (LDA) was applied on classification and prediction accuracy. LDA model was setup over full spectral range. Spectral processing for all LDA models were baseline correction and unit vector normalization. Five samples from each group were left out at each pass until total number of 30 spectra of each subtype predicted.

3. Results

Four hundred spectra were collected from luminal A, luminal B, HER2, and triple negative subtypes and mean spectra of each subtype was extracted. The major peaks corresponding biological component assignments across the whole spectral range of each subtype are as –CH$_3$, –CH$_2$ and –CH=CH– of lipids and proteins, Amide I, II and III, C-S stretching, C-H bending, O-P-O vibrations, phenylalanine, tyrosine, C-C stretching, CH$_2$ twisting, CH$_2$ deformations, and C=C stretching of lipids. The other Raman peaks observed in these subtypes were nucleic acid bases such as adenine, thymine, uracil, guanine and cytosine (ring breathing mode), tryptophan (C-C twisting), proline (C-C stretching), glycogen, CH$_2$, CH$_3$ deformations of lipids. Collagen is the prominent contributor of amide peaks in breast cancer tissues. This is due to either C=C stretching of proteins, α-helical structure of proteins or carbonyl stretches of tumor collagen. Raman peaks of four subtypes have differed in peak intensities and peak shifts. Furthermore, chemometric approaches were used to classify four subtypes. These approaches have assisted in accurate and reliable classification with improved sensitivity and specificity [36].

3.1. Multivariate analysis

AI provides numerous algorithms to analyze Raman spectra such as hierarchical cluster analysis and linear discriminant analysis for disease subtype classification, whereas multivariate analysis technique such as PCA used to standard approach to analyze spectra data. AI approach has been useful in decision making between normal and cancer biopsies [37,38]. PCA was performed on luminal A, luminal B, HER2 positive, and triple negative subtype spectra in order to focus on variability existing among subtypes. In this process, we extracted loading plots from each score plot. These describe the amount of variation for each variable for a given principal component. Loading plots not only provide variations of data sets but also provide correlation of loading point and position of the spectra. Examination of loading plots will offer information

regarding origin of variability with in the data set and based on this information we can explore biochemical components and its contribution towards variations among subtypes.

3.2. High-wavenumber region

In the first score plot of PCA lipid region, PC2 has pretty much separated luminal A from triple negatives, whereas, HER2 positive and luminal B was dispersed. Notably, luminal A were grouped at positive sides on the PC-2 axis, whereas, triple negatives were on both axes. PC-3 has shown whole distribution subtype clusters on positive side of axis whereas a subset of luminal B has distributed on negative side of the axis. The second score plot has shown huge mixture of subtype spectra along PC-3 axis rather than clusters. A three-dimensional PCA plot of high-wavenumber range has shown good cluster formation in HER2 and triple negative subtypes and weak cluster formation in luminal A and B subtypes (Figure 1).

PCA loading plot analysis of high-wavenumber region has described biochemical differences among four subtypes. Firstly, PC-1 loading plot (blue) describes variations in protein specific symmetric CH$_3$ vibrations of all subtypes. PC-2 loading plot (red) has pretty much positively separated luminal A and majority of triple negatives subtype on 2991 and 2848 cm^{-1} (CH$_3$ symmetric vibration of lipids and proteins) compared to remaining subtypes. PC-2 has also negatively separated triple negative, luminal B subtypes and minority of HER2 subtype with 3060, 2914 cm^{-1} (CH stretching of proteins), and 2869 cm^{-1} (CH$_2$ symmetric and asymmetric stretch of lipids, and CH$_2$ asymmetric stretch of proteins). PC-3 loading plot (green) has shown that majority of all subtypes and except minority of luminal B has positively separated with 2918 cm^{-1} (CH stretching of proteins) and 2881 cm^{-1} (CH$_2$ symmetric and asymmetric stretching vibrations) [36].

3.3. Amide I

The first score plot of PCA Amide I region has shown overall distribution of four subtypes clusters. PC-2 describes a variation that is considerably different between luminal A and luminal B. PC-3 has seen almost four subtype cluster distribution both negative and positive side of the axis. The second score plot has also shown pretty different to first plot. PC-1 and PC-2 that are good for separating luminal A from luminal B, PC-3 does not really separate any subtype except may be triple negative from the rest and PC-4 certainly separates triple negative from the rest. Interestingly, PC-2 has pretty much separated luminal B from luminal A from luminal B. PC-3 does not really separate any subtype except may be triple negative from rest of the subtypes. PC-4 certainly separates triple negative from the rest of subtypes. A 3D PCA plot of amide I range has shown good cluster formation for luminal A, HER2 positive, and triple negative subtypes and weak cluster formation in luminal B subtype. A 3D PCA plot has shown good separation in luminal A, HER2 positive and triple negative subtypes whereas luminal B is overlapped with luminal A using the principal components PC-1, PC-3, and PC-4 (Figure 2).

PCA loading plot analysis of amide I region has described biochemical differences among four subtypes. Firstly, PC-1 loading plot (blue) has sensitive in protein conformation of
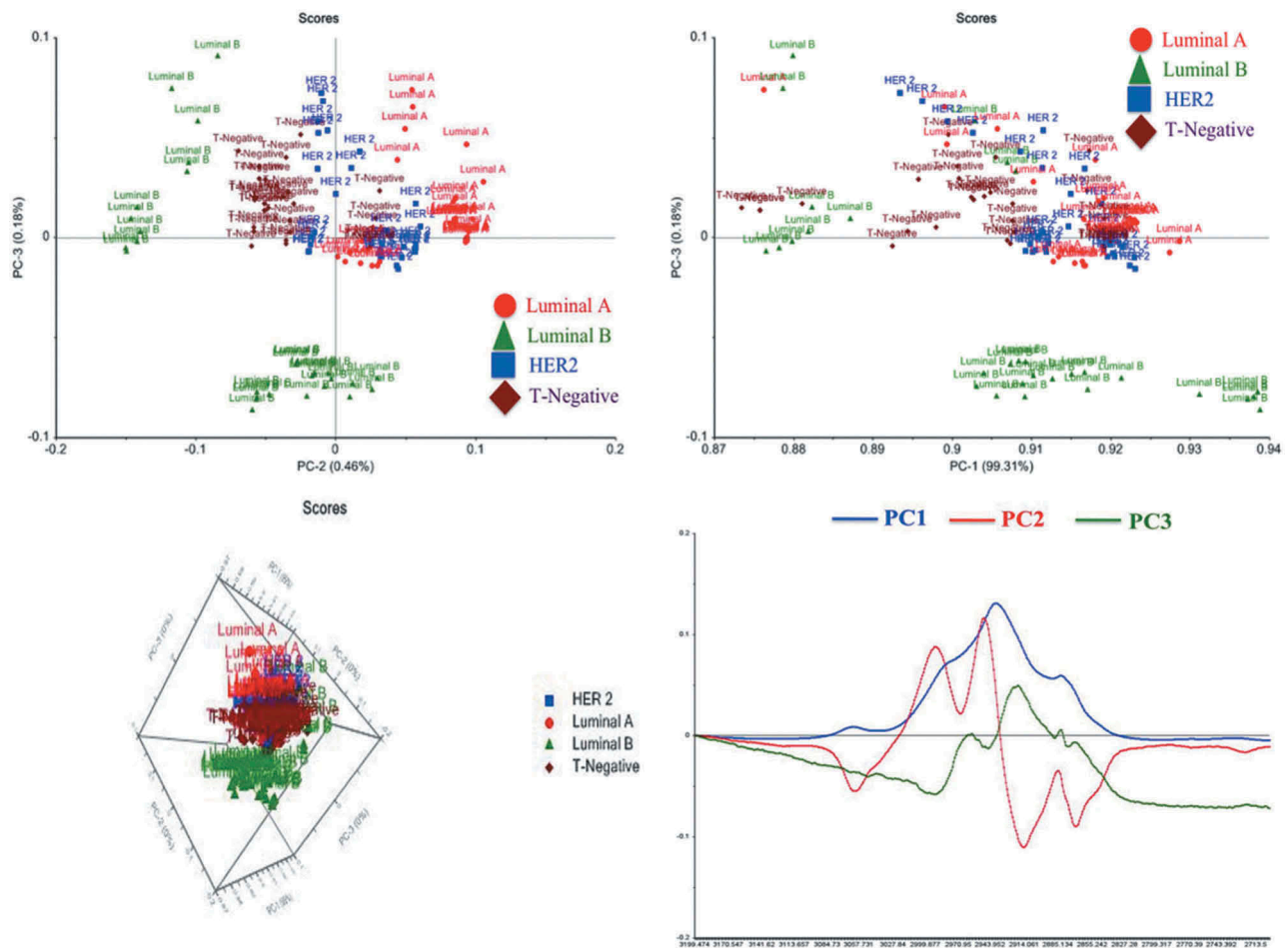
Figure 1. 2D and 3D PCA plots and loading plots of high wavenumber region for luminal A (red), B (green), HER2 (blue) and triple negative (brown) subtypes based on first, second and third principal components (PC).

different subtypes. Every PC of this region is sensitive to protein conformation. Collagen at 1667 cm^{-1} (either unordered or β sheets) is positively separated triple negative and majority of HER2 subtypes. PC-3 loading plot (red) has pretty much sensitive in 1654 cm^{-1} (α-helical confirmation of collagen), 1614 cm^{-1} (tyrosine), 1604 cm^{-1} (phenylalanine and tyrosine), and 1576 cm^{-1} (guanine) and 1634 cm^{-1} (amide I). PC-4 loading plot (green) has shown that majority of triple negative, luminal A, and HER2 subtypes are positively separated at 1671 (β sheet structures of protein conformation) and 1575 (ring breathing mode of DNA bases), and some of the luminal Band HER2 and luminal A are negatively separated at 1601 (phenylalanine) and 1547 (proline) [36]. Collagen was important for separating luminal A and luminal B from rest of the subtypes but nucleic acids and amino acids were more important in separating triple negatives from rest of the subtypes. In a nut shell, PCA of amide I region is sensitive to protein conformation of different subtypes.

3.4. Amide III

The first score plot of PCA amide III region, PC-2 has pretty much separated luminal A from triple negative subtype, whereas, HER2 positive and luminal B are widely distributed. All four subtypes were randomly distributed along the PC-3

axis both positive and negative side. A 3D PCA plot of Amide III range has shown better cluster formation in triple negative subtypes and weak cluster formation in luminal A, luminal B, and HER2 positive subtypes. A 3D PCA plot has shown a weak separation in luminal A whereas luminal B, HER2 positive, and triple negative subtypes are overlapped using the principal components PC-1, PC-2, and PC-3 (Figure 3).

PCA-loading plot analysis of Amide III region has described biochemical differences among four subtypes. Firstly, PC-2 loading plot (red) describes a variation that is sensitive to protein conformation of different subtypes. It has positively separated majority of luminal B, HER2, and triple negative subtypes at 1269 and 1244 and negatively at 1336 cm^{-1} (purine bases and collagen assignment). PC-3 loading plot (green) has pretty much sensitive in 1377 cm^{-1} (lipid assignments), 1339 cm^{-1} (C-C stretch of phenylalanine), 1279 cm^{-1} (amide III of α-helix), and 1277 cm^{-1} (amide III of α-helix) [36].

3.5. Nucleic acid region

The first and third score plot of PCA nucleic acid region has shown clear separation between triple negative and luminal A subtypes, whereas, HER 2 and luminal B were distributed on both sides of PC-2 axis. Interestingly, each subtype can be observed as almost subsets of clusters except few spectra of
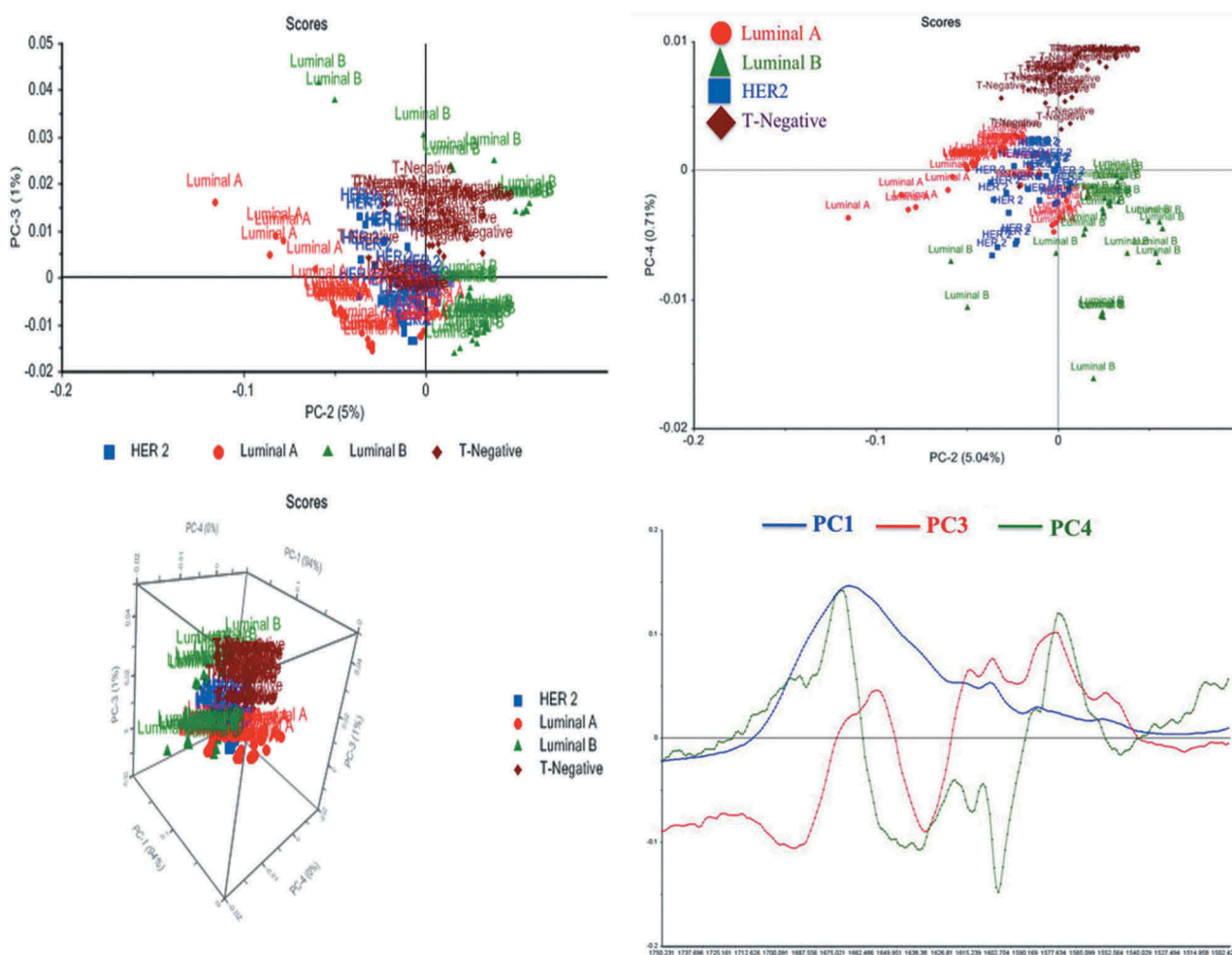
Figure 2. 2D and 3D PCA plots and loading plots of region for luminal A (red), B (green), HER2 (blue) and triple negative (brown) subtypes based on first, third and fourth principal components (PC).

luminal A and luminal B. The second score plot has shown indiscriminate separation of four subtypes across the PC-3 axis. A 3D PCA plot of nucleic acid range has shown weak two-subset cluster formation in HER2, three-subset cluster formation in triple negative subtype and weak cluster formation in luminal A and B subtype. A 3D PCA plot has shown good separation in luminal A subtype, whereas, luminal B, HER2, and triple negative subtypes are overlapped using the principal components PC-1, PC-2, and PC-3 (Figure 4).

PCA-loading plot analysis of nucleic acid region has described biochemical differences among four subtypes. Firstly, PC-1 loading plot (blue) is sensitive in 938, 921, 877, and 855 cm^{-1} regions. PC2 loading plot (red) is pretty much sensitive in 939 and 827 cm^{-1}. PC-3 loading plot (green) has shown positive separation at 960 cm^{-1} (phosphate vibrations), 903 cm^{-1} (C-C skeletal stretching), 829 cm^{-1} (O-P-O stretching of DNA and RNA), 783 cm^{-1} (thymine, uracil and cytosine), 758 cm^{-1} (tryptophan and ethanolamine of phosphatidylethanolamine), 643 cm^{-1} (C-C twisting mode of tyrosine), and 621 cm^{-1} (disulfide bridges of cysteine content). It has negatively separated for 920 cm^{-1} (proline/glucose/lactic acid), 870 cm^{-1} (proline/hydroxyproline/valine and polysaccharides), 814 cm^{-1} (proline/hydroxyproline/tyrosine/phosphodiester of RNA), and 770 cm^{-1} (sugar backbones or cyclic ring of nucleic acid bases) [36].

3.6. Cluster analysis

Figure 5 has shown the dendrogram of luminal A (L-A), luminal B (L-B), HER2 (Her 2), and T-Negative (T-N) subtype (four sets of each subtype) samples over full spectral range (3200–400 cm^{-1}). Four main clusters were formed. Some of the luminal B spectra were mixed with triple negative and HER2 subtypes. Each region has formed pretty much good cluster and distance measurements has suggested that biochemical snapshot of full spectral region has shown more similarity between luminal A and subset of luminal B, and HER2 and triple negative subtype. One subset of luminal B has shown similarity with triple negative subtype. Distance measurements have shown luminal A and B chemically more distant from HER2 and triple negative subtypes.

3.7. Linear discrimination analysis (LDA)

LDA, one of the AI was used in the classification of breast cancer subtypes A total 120 spectra (30 spectra each subtype) was selected and processed using baseline correction and normalization for predictive classification with LDA. LDA models were setup over the full spectral range 3200–600 cm^{-1}. Five samples from each group were left out at each pass until
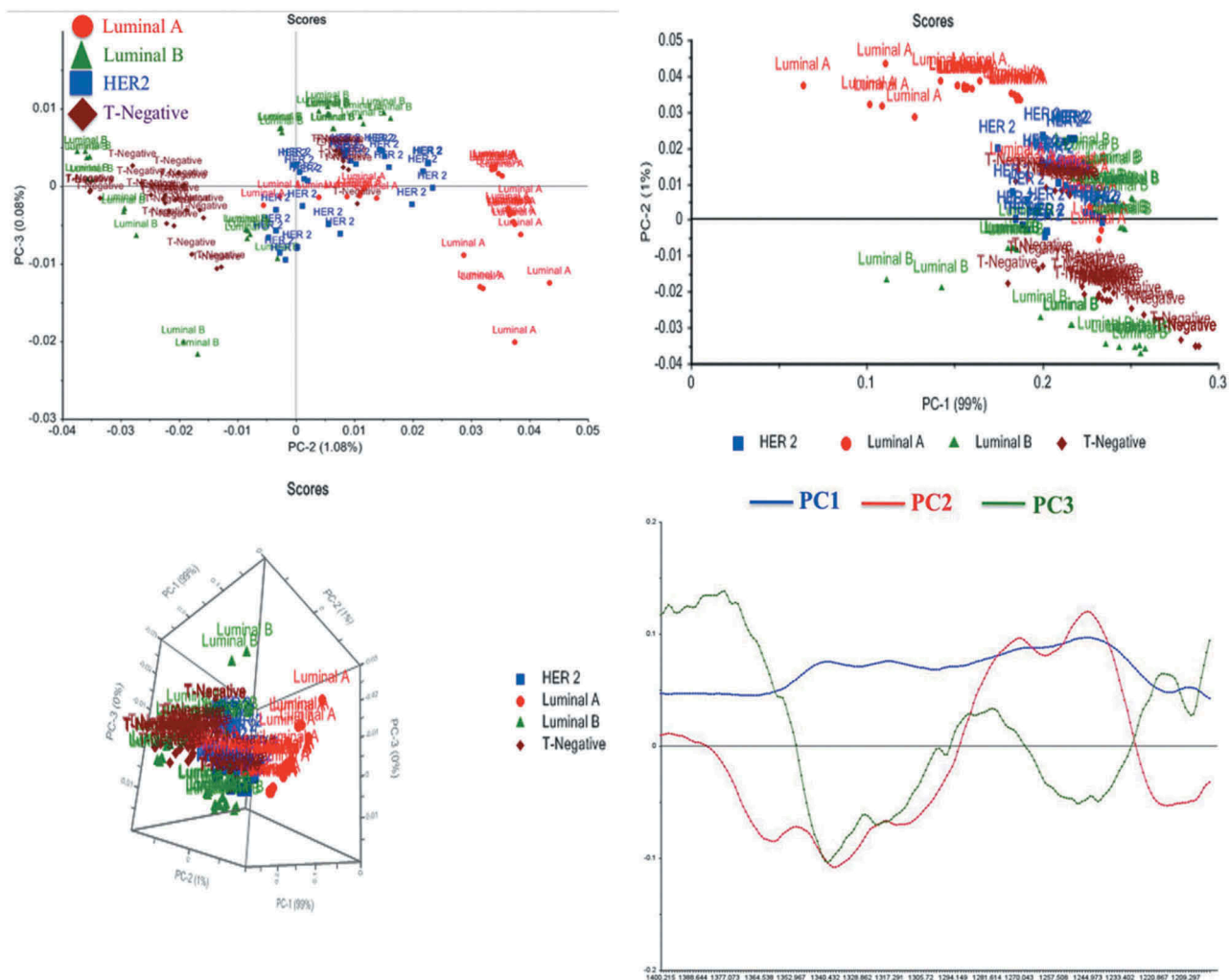
Figure 3. 2D and 3D PCA plots and loading plots of region for luminal A (red), B (green), HER2 (blue) and triple negative (brown) subtypes based on first, second and third principal components (PC).

a total number of 30 spectra from each group were tested against the training group. LDA fusion matrix of four subtypes of TMA breast biopsies were mentioned in Table 1.

Luminal A subtype was predicted with a sensitivity of 70% where eight spectra were predicted as Luminal B and one spectrum predicted as HER2 positive subtype. None of the Luminal A subtype spectra were misclassified in the triple negative subtype. Luminal B subtype was predicted with sensitivity of 100% where none of the luminal B spectra were predicted wrongly. HER2 positive subtype was predicted with sensitivity of 90% with three spectra were misclassified as luminal B. Triple negative subtype was predicted with sensitivity of 96.67% where only one spectrum was predicted as luminal B.

4. Discussion

Current screening and diagnosis approaches such as mammography, ultra sound and magnetic resonance imaging (MRI) have their own limitations. Mammogram generally provides 10–14% false negatives whereas MRI approach is expensive and lack of potential in visualization of both ductal carcinoma

in situ and invasive ductal carcinoma. RS not only provides real-time biochemical profile of tissues but also understanding of the disease as it progresses. Breast cancer biopsies that were analyzed in our study were mainly ductal carcinoma NST. Multiple genetic mutations and protein dysfunction are the main cause of breast cancer [39]. It is a well-known fact that cancer tissues have higher cell proliferation and metabolic activity that result in changes of concentration and oxidation states of different chemical species. The major biological activities observed in cancer cells are loss of differentiation, nuclear enlargement, especially increasing in genetic material [40]. Spectral changes were identified especially at lipids, amide I and III regions among luminal A, B, HER2, and triple negatives subtypes. Furthermore, subtypes were also classified using supervised and unsupervised algorithms. Significant spectral variations in terms of peak shifts, shapes, and intensities were observed among four subtypes.

High-wavenumber region of Raman spectra epitomizes stretching vibrations of lipids and proteins. These vibrations represent biochemical profile of lipid metabolism and are helpful in assessing lipid peroxidation in biological systems [40]. Lipid biosynthesis is correlated with saturated lipid
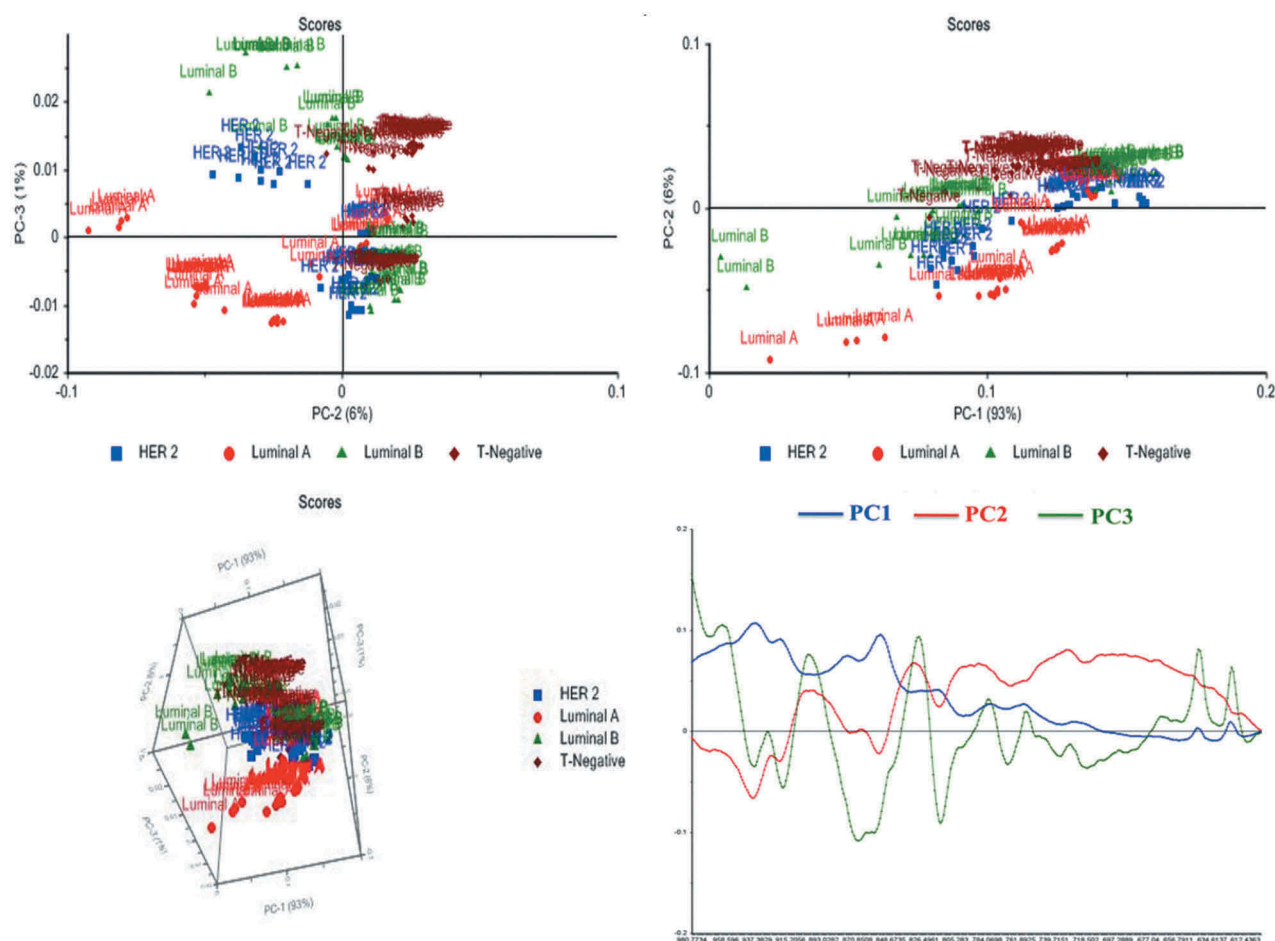
Figure 4. 2D and 3D PCA plots and loading plots of region for luminal A (red), B (green), HER2 (blue) and triple negative (brown) subtypes based on first, second and third principal components (PC).

content in cells. Lipid peroxidation is directly proportional to lipid degradation and synthesis of lipids is decreased in low number saturated fatty acids. Deregulated lipid biosynthesis is a characteristic feature of cancerous tissues. Lipid profiles of triple negative and luminal A were similar compared to HER2 positive and luminal B. Changes in peak intensities and shifts are a result of differences in structural conformations and concentration of lipids in different subtypes of cancerous biopsies. Lipid synthesis plays a vital role in cancer development and progression [41]. Lipid requirements of mammalian tissues are fulfilled by uptake of fatty acids and lipoproteins from the blood stream. Fatty acids and cholesterol production is restricted to a subset of tissues such as lactating breast, liver, and adipose tissue. Nevertheless, reactivation of lipid synthesis has been reported in tumorous tissues [42]. Initial studies demonstrated that cancerous tissues use *de novo* lipogenesis to produce lipids, including fatty acids and phospholipids [43,44]. In recent years, combination of gene expression and genome scale metabolic studies of breast cancer has revealed two important concepts of lipid synthesis in cancerous tissues [45]. Firstly, biosynthesis of fatty acids is a characteristic feature of early stages of tumor formation because these play important role in rapid proliferation of cancer cell. Secondly, advanced stages of cancer have shown antioxidant nature for detoxification of reactive oxygen species (ROS) [46]. Low

amount of saturated lipids also represents slow growth behavior of cell. Unsaturated fatty acid content is directly proportional to the cancer progress. High-unsaturated fatty acid content is present in metastatic stages [47]. This Raman approach has even identified stromal lipid peak intensity differences in individual TMA biopsies. These lipid intensity differences and structural information were helpful in understanding these four subtypes of breast cancer as well classification. The cancerous breast tissue is subjugated by protein content and high-wavenumber is useful to study protein vibrations in terms of amino acid nature. Increase in protein content and its contribution toward carcinogenesis is evident in breast cancer and nearly 30% cases are having amplified gene as protein product in HER2 positive subtypes [48]. The main difficulty in this region is the interpretation of Raman bands. The major problem is that most of the lipid and fatty acid bands overlap with protein bands. Overlap is generally between CH of lipids with CH_3, CH_2, and CH of amino acid side chains. Previous Raman spectroscopic studies based on breast cancer studies have reported that this region includes not only aliphatic and aromatic amino acids but also other amino acids such as histidine, threonine, and proline [49].

Protein metabolism plays a key role in cancer biology in terms of cell differentiation and cell proliferation, hence,
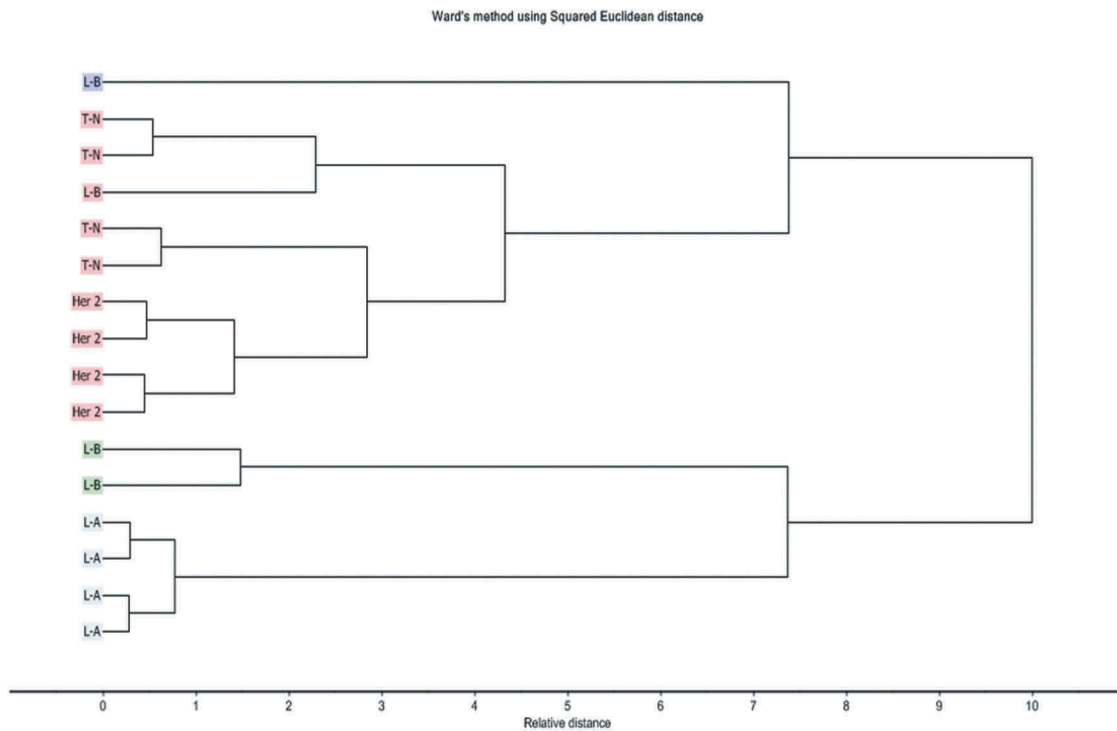
Figure 5. Dendrogram of Cluster Analysis results for luminal A (L-A), luminal B (LB), HER2 (Her 2) and T-Negative (T-N) subtype (four sets of each subtype) samples. CA was performed over the full spectral range using Wards method's Squared Euclidean distance.

Table 1. LDA fusion matrix of Luminal A, B, HER2 positive and triple negative subtypes.

Actual / Predicted	HER2	Luminal A	Luminal B	Triple Negative
HER2	27	1	0	0
Luminal A	0	21	0	0
Luminal B	3	8	30	1
Triple Negative	0	0	0	29
Total	**30**	**30**	**30**	**30**

considered as potential Raman markers in disease identification. The second and third steps of central dogma, such as, transcription (DNA to mRNA synthesis) and translation (mRNA to protein synthesis) signatures were captured in fingerprint region through molecular vibrations [50]. Aromatic amino acids such as tryptophan and tyrosine plays crucial role in various metabolic processes and are required for rapidly proliferating cancer cells. Previous studies have reported that high amount of tryptophan in cancerous tissues is also evident and tryptophan peak can be specifically observed in Raman spectrum at 1583 cm^{-1} [51]. Normal breast tissue has shown characteristic tryptophan between amide I and amide II and peak intensity is increased as cancer progresses in tissues. Peak intensity of tryptophan is increased in cancerous tissues. Tryptophan peak intensity differences and peak shifts were observed in four subtypes. Luminal B have shown tryptophan peak at 1583 cm^{-1}, whereas, luminal A, HER2 positive, and triple negative subtypes have observed in peak shift. Peak intensity is similar between HER2 and luminal B subtypes whereas luminal A has shown low intensity and triple negative is further lowest. Significant tryptophan changes were observed and luminal B and HER2 have 10-fold increase compared to triple negative and 5-fold increase compared to

luminal A. Protein content changes will be helpful in different subtypes of breast tissues. Normal breast tissue has shown less tryptophan content compared to cancerous tissues. Tryptophan Peak shift was observed in luminal B subtype and intensity difference identified in luminal A and triple negative subtypes because they are lower in intensity. Recent studies have confirmed that tryptophan metabolic pathway has significant role in cancer progression. Indoleamine 2,3-dioxygenase (IDO), rate-limiting enzyme of tryptophan metabolic pathway, expression profile is associated with advancement of breast cancer. Raman can be helpful in identification of tryptophan signatures in different subtypes of breast cancer.

Proteins peaks, majority of them amide peaks, result from carbonyl stretching from polypeptide backbones. The information provided by Amide I peak is mainly dependent on their position and shape. Triple negative subtype has shown amide I peak with high intensity next to Luminal B subtype whereas Luminal A and HER2 subtypes have shown less amide I peak intensities. The higher intensity of amide I represent higher amount of collagen. Ductal carcinoma *in situ* and fibroadenoma samples usually shows large number of cells compared to other lesions of the breast. Increasing cell nucleus size is one of the important factors in cancer. Pathologists consider higher nuclear to cytoplasm ratio is the best way to diagnose the disease. Amide I peak appeared to 1667 cm^{-1} in triple negative subtype and it represents β-sheets of collagen protein. The structural modes of malignant breast tissue proteins have shown this peak in previous studies [52]. Amide I vibrations of tumor proteins were shifted to HER2 subtype and then to luminal A and B. Surface Enhanced Raman Spectroscopy (SERS) studies of saliva have also described this

peak as C=O stretching vibrations, whereas, spectra of brain tumors have attributed this peak to C=C stretching vibrations [53]. Lung cancer studies have shown that Raman peak at 1666 cm^{-1} was only present in cancerous tissue and it is assigned to collagen. In this study it was suggested that collagen content plays important role not only in breast cancer but also other cancers as well.

Collagen content can be used as quantitative Raman markers in disease identification, especially in subtype classification [54] and our studies also confirmed the importance of collagen contents both in identification of cancer and classification of subtypes. Due to desmoplastic reaction, stroma of neoplastic breast tissue starts to accumulate dense fibrous tissue possessing newly formed ECM components, and predominantly, collagen. Previous studies has established that invasion sites of ductal infiltrating carcinoma have shown accumulation of type I and III collagens whereas protein extraction studies have revealed that type V collagen increases by nearly 10% in Ductal Invasive Carcinoma (DIC) compared to normal breast tissue. In recent years, gene expression studies have explored tumor-associated collagen structures (TACS) related to breast cancer progression [55,56]. Optical approaches such as Multiphoton and second harmonic generation microscopy show that TACS is important in cancer progression using mouse tumor explants and microarrays. TACS 1 is a limited density collagen and usually appears at small tumor foci. TACS 2 is tangentially oriented collagen toward the smooth boundary of tumor and TACS 3 is perpendicularly oriented toward irregular invasive tumor boundary. TACS 3 is described as a consistent and powerful marker in disease identification, especially in triple negatives [57]. This marker is going to be considered as an independent prognosis key irrespective of tumor grade, size, and receptor status. In recent years, second harmonic generation (SHG) microscopy is used to quantify collagen amount and evaluated abnormal collagen fibrils to study malignancy in different stages of breast cancer. Apart from SHG microscopy, X-ray scattering approaches have also proved that collagen content and its structural associated changes differ in different stages of breast cancer and also predicted invasion directions in terms of cancer spreading [58,59].

Multivariate approach (PCA) results on whole spectral range have shown much overlap in luminal A, HER2, and triple negative subtypes. Luminal B subtype has separated very well using PC2 and it formed very good cluster, and this might be due to less intervariation in biopsies. Overlapping of three subtypes might indicate that biopsies among these groups possess many similarities in whole spectral range and this led us to go for a closer look to explore small spectral regions. Lipids, amide I, amide III, and nucleic acid regions were explored using PCA. This is the first time; breast cancer biopsy data was explored using AI and ML approaches based on subtype classification. AI approach studies have identified malignant tissues from normal and benign in various cancers such as lung cancer [54,60], skin cancer [61,62], brain cancer [18,63], gastro intestinal cancer [64,65], and oral cancer [66].

Nuclear magnetic resonance studies have revealed that phospholipids and mitochondrial metabolism were affected in breast cancer cells and phosphocholine content was increased nearly 20-fold in primary breast cancer cell lines and nearly 30-fold in metastatic breast cell lines. In addition to that, genetic alterations increase choline transport and enhance the synthesis of phosphocholine and betaine and decline the choline derived ether lipids in breast cancer cells. Decreased levels of ATP, phosphocreatine, and influx of pyruvate have clearly shown that mitochondrial metabolism was damaged in breast cancer cells. Genetic alterations increased choline transport and enhanced the synthesis of phosphocholine and betaine and decline the choline derived ether lipids in breast cancer cells. Initially, Raman studies identified changes in fatty acid content and β carotene between normal cancerous biopsies. Later studies proved that protein changes in normal and ductal carcinoma were observed as peak shifts. Normal breast tissue has shown C-H protein vibration at 1439 cm^{-1}, whereas, ductal carcinoma NST samples have shown this peak at 1449 cm^{-1}. Here, all four subtypes have expressed this peak at 1449 cm^{-1} and it clearly shows that all of biopsies were in malignant state. This wavenumber shift is probably indicative of large lipid accumulation in the malignant state compared to normal breast.

The spectral region between 1400 and 800 cm^{-1} has represented six different amino acids such as glycine, alanine, proline, tyrosine, valine, and phenylalanine. These amino acids are basic backbones of primary structure of proteins. The principal component of breast tissue is collagen and it is mainly composed of glycine, valine, proline, and phenylalanine. It is clear that the intensity of these amino acids is increased in Triple negative subtype compared to the other subtypes. Alanine, glycine, proline, and tyrosine peak intensities were observed high in triple negative subtypes and low in luminal B subtypes. The order in which these amino acid peak intensities decrease is triple negative to luminal A to HER2 positive to luminal B. Collagen peak intensity variation is observed in amide I and amide III regions within individual biopsies. The intensity of collagen among different breast cancer subtype was much varied; this difference was due to various concentration of collagen among the investigated subtypes. Our Raman spectral data has shown that invasive triple negative cancer has the highest intensity of collagen among all other subtypes. Raman spectra also reported high collagen associated amino acids in fibroadenoma and invasive lobular carcinomas [29]. Based on this relative higher amount of collagen content in triple negative subtype compared to other subtypes have indicated that this is one of the most pathogenic subtype, Raman could be in useful in identification of new chemical pathogenic marker in terms of collagen content.

Amide I peak is shifted gradually toward lower wavenumber in HER2 and luminal subtypes. Maiti et al. reported that amide I peak at 1667 cm^{-1} has contributed from β sheet structures. A sharp peak also represents increased hydrogen bonding in β sheets of proteins [67,68]. Three shoulder peaks were observed in amide I region namely 1621, 1605, and 1584 cm^{-1}. The first two shoulder peaks attributed from amino acids and, later from amino acids and nucleic acids. Raman peak at 1605 cm^{-1} in triple negatives, 1604 cm^{-1} at HER2, and luminal subtypes have attributed to aromatic amino acids such as phenylalanine and tyrosine. CARS-based colorectal tissue studies have suggested that Raman peak at

1584 cm^{-1} is attributed to nucleic acids, whereas, RS-based nasopharyngeal tissue studies attributed to C=C bending mode of phenylalanine [69,70]. Previous studies have reported that these peak intensities are decreased in cancerous tissue compared to normal. HER2 and luminal B subtypes have shown increased intensities compared to luminal A and triple negative subtypes.

Earlier breast cancer studies have reported significant and reproducible peak shifts in amide II region. Raman Spectra of normal breast tissue have expressed amide II peak at 1439 cm^{-1}, whereas, invasive ductal carcinoma biopsies at 1449 cm^{-1}. All four subtypes have expressed this peak at 1449 cm^{-1} with peak intensity differences. This is due to CH$_2$ deformations of fatty acid and proteins. Fatty acid composition varies in breast cancerous tissues. Triglyceride content is reduced nearly more than half in cancerous tissues compared to normal, whereas, phospholipid content is increased four folds in cancerous biopsies. Overall reduced lipid content in cancerous biopsies was reported. Triple negative and luminal B subtypes have shown increased peak intensities compared to luminal A and HER2 subtypes. Phospholipid metabolic studies have revealed that membrane lipids, such as, phosphatidylethanolamine and phosphatidylcholine plays vital role in breast cancers. Membrane lipid content might assist breast tumor invasion as well. Low levels phosphatidylethanolamine act as marker in early prediction of visceral metastasis. A later phase of metastasis was predominated by low levels of phosphatidylcholine [26]. Normal mice mammary gland has not expressed Raman peak at 1339 cm^{-1} but mammary tumors have expressed, and it is attributed to hydrated α-helix δ (N-H) and v(C-N) of elastin proteins. It might also be attributed to purine bases (adenine and guanine) of nucleic acid content [71]. Raman peak at 1317 cm^{-1} represents v (C-H), CH$_2$ of aliphatic amino acids. All four subtypes have shown amide III peak at 1245 cm^{-1} and shoulder peak at 1207 cm^{-1} attributed to hydroxyproline and tyrosine. Amide III region represents major vibrations such as N-H bending and C-N stretching and minor vibrations of C=O in-plane bending and C=C stretching vibrations [72]. Protein conformation in amide III region is more complex because it relies on side chain of amino acids, where N-H bending is responsible for many modes of vibrations in this region. Polypeptide backbone and amino acid side-chain vibrations significantly vary in this region making it difficult to interpret secondary structure of proteins [73]. Overall, higher intensity of amide III is observed in triple negative subtype and higher protein content with tyrosine and hydroxyproline was observed in HER2 subtype. Amide III region has shown triple negative and luminal B has higher protein content than HER2 and luminal A subtype.

A strong phenylalanine peak was observed in all four subtypes. Luminal B has showed the highest intensity of phenylalanine compared to other subtypes. The region between 1000 and 600 cm^{-1} has shown significant collagen associated amino acid content [26]. Proline and tyrosine are observed in higher quantities in triple negative subtypes compared to luminal and HER2 subtypes. It is well-known fact that an increase in collagen content is a key marker in carcinogenesis [29]. In the case of breast cancer, desmoplastic reaction leads to deposition of collagen known as reactive fibrosis. Collagen deposition is a stromal indicator to invasive carcinoma. Two prominent peaks observed around 642 and 620 cm^{-1} representing C-C twisting of tyrosine and C-S stretching vibrations. Peak intensities for these two peaks differed in all four subtypes of breast cancer. In high-wavenumber, amide I, III, and nucleic acid regions, luminal B spectra have distributed on both positive and negative sides of PCA plots. Luminal B has shown considerable spectral scattering in all regions and one small subset has shown similarity with triple negative subtype whereas majority of them have shown more similarity toward luminal A subtype.

5. Conclusions

A combination of RS, AI, and ML approaches have great potential in chemical assessment of cancerous tissues. These approaches used in this study provide a good cross check to validate spectral data, the popular method used to classify breast cancer subtypes spectra is PCA, which has shown excellent agreements with the results of cluster analysis and LDA. TMA breast biopsies have shown huge chemical heterogeneity not only in different subtypes but also within each individual subtype. Fatty acid, amide I, and III have contributed many variations among four subtypes of breast biopsies. Rapid mapping of large number of biopsies of different subtypes will be helpful to track changes in collagen and lipid profiles in each subtype and to establish large database based on these profiles. Finally, these differences observed in the four subtypes may also be useful in understanding cancer migration although this needs confirmation of larger panel of biopsies.

Declarations

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Reviewers Disclosure

Peer reviewers on this manuscript have no relevant financial relationships or otherwise to disclose.

Declaration of interest

No potential conflict of interest was reported by the authors.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.

- **This review describes the breast cancer incidences and stats around the world.**
2. International Agency for Research on Cancer (2018). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. [cited 2019 Sep 12]. Retrieved from https://www.who.int/cancer/PRGlobocanFinal.pdf.
3. National Research Council. Saving women's lives: strategies for improving breast cancer detection and diagnosis. Boca Raton (FL): National Academies Press; 2005.
4. Mahadevan-Jansen A, Richards-Kortum RR. Raman spectroscopy for the detection of cancers and precancers. J Biomed Opt. 1996;1(1):31–71.
5. Raman CV, Krishnan KS. A new type of secondary radiation. Nature. 1928;121(3048):501.
6. Rehman I, Movasaghi Z, Rehman S. Vibrational spectroscopy for tissue analysis. Boca Raton (FL): CRC Press; 2012.
- **This books provides history, Raman and FTIR analysis of Cancer in detail.**
7. Surmacki JM, Woodhams BJ, Haslehurst A, et al. Raman micro-spectroscopy for accurate identification of primary human bronchial epithelial cells. Sci Rep. 2018;8(1):12604.
8. Wang H, Zhang S, Wan L, et al. Screening and staging for non-small cell lung cancer by serum laser Raman spectroscopy. Spectrochim Acta A Mol Biomol Spectrosc. 2018;201:34–38.
9. Movasaghi Z, Rehman S, Rehman IU. Raman Spectroscopy Can Detect and Monitor Cancer at Cellular Level: Analysis of Resistant and Sensitive Subtypes of Testicular Cancer Cell Lines. Appl Spectrosc Rev. 2012;47(7):571–581.
10. Lyng FM, Traynor D, Nguyen TNQ, et al. Discrimination of breast cancer from benign tumours using Raman spectroscopy. PloS One. 2019;14(2):e0212376.
11. Zhang C, Winnard PT Jr, Dasari S, et al. Label-free Raman spectroscopy provides early determination and precise localization of breast cancer-colonized bone alterations. Chem Sci. 2018;9(3):743–753.
12. Brozek-Pluska B, Kopec M, Surmacki J, et al. Raman microspectroscopy of noncancerous and cancerous human breast tissues. Identification and phase transitions of linoleic and oleic acids by Raman low-temperature studies. Analyst. 2015;140(7):2134–2143.
13. Duraipandian S, Traynor D, Kearney P, et al. Raman spectroscopic detection of high-grade cervical cytology: Using morphologically normal appearing cells. Sci Rep. 2018;8(1):15048.
14. Shaikh R, Prabitha VG, Dora TK, et al. A comparative evaluation of diffuse reflectance and Raman spectroscopy in the detection of cervical cancer. J Biophotonics. 2017;10(2):242–252.
15. Feng S, Pan J, Wu Y, et al. Study on gastric cancer blood plasma based on surface-enhanced Raman spectroscopy combined with multivariate analysis. Sci China Life Sci. 2011;54(9):828–834.
16. Kawabata T, Kikuchi H, Okazaki S, et al. Near-infrared multichannel Raman spectroscopy with a 1064 nm excitation wavelength for ex vivo diagnosis of gastric cancer. J Surg Res. 2011;169(2):e137–e143.
17. Krafft C, Dochow S, Latka I, et al. Diagnosis and screening of cancer tissues by fiber-optic probe Raman spectroscopy. Biomed Spectrosc Imaging. 2012;1(1):39–55.
18. Anna I, Bartosz P, Lech P, et al. Novel strategies of Raman imaging for brain tumor research. Oncotarget. 2017;8(49):85290.
19. Shipp DW, Rakha EA, Koloydenko AA, et al. Intra-operative spectroscopic assessment of surgical margins during breast conserving surgery. Breast Cancer Res. 2018;20(1):69.
20. Larraona-Puy M, Ghita A, Zoladek A, et al. Discrimination between basal cell carcinoma and hair follicles in skin tissue sections by Raman micro-spectroscopy. J Mol Struct. 2011;993(1–3):57–61.
21. Li Y, Wen ZN, Li LJ, et al. Research on the Raman spectral character and diagnostic value of squamous cell carcinoma of oral mucosa. J Raman Spectrosc. 2010;41(2):142–147.
22. Park J, Hwang M, Choi B, et al. Exosome classification by pattern analysis of surface-enhanced Raman spectroscopy data for lung cancer diagnosis. Anal Chem. 2017;89(12):6695–6701.
23. Shapiro A, Gofrit ON, Pizov G, et al. Raman molecular imaging: a novel spectroscopic technique for diagnosis of bladder cancer in urine specimens. Eur Urol. 2011;59(1):106–112.
24. Abramczyk H, Brozek-Pluska B, Surmacki J, et al. Hydrogen bonds of interfacial water in human breast cancer tissue compared to lipid and DNA interfaces. J Biophys Chem. 2011;2(02):159.
25. Frank CJ, Redd DC, Gansler TS, et al. Characterization of human breast biopsy specimens with near-IR Raman spectroscopy. Anal Chem. 1994;66(3):319–326.
- **This study provides first comprehensive Raman analysis of breast biopsies.**
26. Frank CJ, McCreery RL, Redd DC. Raman spectroscopy of normal and diseased human breast tissues. Anal Chem. 1995;67(5):777–783.
27. Manoharan R, Shafer K, Perelman L, et al. Raman spectroscopy and fluorescence photon migration for breast cancer diagnosis and imaging. Photochem Photobiol. 1998;67(1):15–22.
28. Shafer-Peltier KE, Haka AS, Fitzmaurice M, et al. Raman microspectroscopic model of human breast tissue: implications for breast cancer diagnosis in vivo. J Raman Spectroscopy. 2002;33(7):552–563.
29. Haka AS, Shafer-Peltier KE, Fitzmaurice M, et al. Diagnosing breast cancer by using Raman spectroscopy. Proc Nat Acad Sci. 2005;102 (35):12371–12376.
30. Talari ACS, Evans CA, Holen I, et al. Raman spectroscopic analysis differentiates between breast cancer cell lines. J Raman Spectroscopy. 2015;46(5):421–427.
- **This study describes the combination of Raman and multivariate approach to differentiate breast cancer cell lines.**
31. Talari ACS, Raza A, Rehman S, et al. Analyzing normal proliferating, hypoxic and necrotic regions of T-47D human breast cancer spheroids using Raman spectroscopy. Appl Spectrosc Rev. 2017;52 (10):909–924.
- **This study describes and compares three different regions of spheroids using Raman spectral markers.**
32. Baldock SJ, Talari ACS, Smith R, et al. Single-cell Raman microscopy of microengineered cell scaffolds. J Raman Spectroscopy. 2019;50 (3):371–379.
33. Rehman S, Movasaghi Z, Tucker AT, et al. Raman spectroscopic analysis of breast cancer tissues: identifying differences between normal, invasive ductal carcinoma and ductal carcinoma in situ of the breast tissue. J Raman Spectrosc. 2007;38(10):1345–1351.
34. Moch H, Kononen J, Kallioniemi OP, et al. Tissue microarrays: what will they bring to molecular and anatomic pathology?. Adv Anat Pathol. 2001;8(1):14–20.
35. Rakha EA, Reis-Filho JS, Baehner F, et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. 2010;12(4):207.
36. Talari ACS, Movasaghi Z, Rehman S, et al. Raman spectroscopy of biological tissues. Appl Spectrosc Rev. 2015;50(1):46–111.
- **This article reviews the various cancers and their Raman spectral significance.**
37. Ye Z. Artificial-intelligence approach for biomedical sample characterization using Raman spectroscopy. IEEE Trans Autom Sci Eng. 2005;2(1):67–73.
38. Harris AT, Garg M, Yang XB, et al. Raman spectroscopy and advanced mathematical modelling in the discrimination of human thyroid cell lines. Head Neck Oncol. 2009;1(1):38.
39. Rakha EA, El-Sheikh SE, Kandil MA, et al. Expression of BRCA1 protein in breast cancer and its prognostic significance. Hum Pathol. 2008;39(6):857–865.
40. Kang DH. Oxidative stress, DNA damage, and breast cancer. AACN Adv Crit Care. 2002;13(4):540–549.
41. Hirsch HA, Iliopoulos D, Joshi A, et al. A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases. Cancer Cell. 2010;17(4):348–361.
42. Ozek NS, Tuna S, Erson-Bensan AE, et al. Characterization of microRNA-125b expression in MCF7 breast cancer cells by ATR-FTIR spectroscopy. Analyst. 2010;135(12):3094–3102.
43. Swinnen JV, Roskams T, Joniau S, et al. Overexpression of fatty acid synthase is an early and common event in the development of prostate cancer. Int J Cancer. 2002;98(1):19–22.
44. Milgraum LZ, Witters LA, Pasternack GR, et al. Enzymes of the fatty acid synthesis pathway are highly expressed in in situ breast carcinoma. Clin Cancer Res. 1997;3(11):2115–2120.

45. Hilvo M, Denkert C, Lehtinen L, et al. Novel theranostic opportunities offered by characterization of altered membrane lipid metabolism in breast cancer progression. Cancer Res. 2011;71(9):3236–3245.

46. Germain E, Chajès V, Cognault S, et al. Enhancement of doxorubicin cytotoxicity by polyunsaturated fatty acids in the human breast tumor cell line MBA-MB-231: Relationship to lipid peroxidation. Int J Cancer. 1998;75(4):578–583.

47. Bartsch H, Nair J, Owen RW. Dietary polyunsaturated fatty acids and cancers of the breast and colorectum: emerging evidence for their role as risk modifiers. Carcinogenesis. 1999;20(12):2209–2218.

48. Olayioye MA. Intracellular signaling pathways of ErbB2/HER-2 and family members. Breast Cancer Res. 2001;3(6):385.

49. Surmacki J, Musial J, Kordek R, et al. Raman imaging at biological interfaces: applications in breast cancer diagnosis. Mol Cancer. 2013;12(1):48.

50. Thomas GJ Jr. Raman spectroscopy of protein and nucleic acid assemblies. Annu Rev Biophys Biomol Struct. 1999;28(1):1–27.

51. Takeuchi H. Raman structural markers of tryptophan and histidine side chains in proteins. Biopolymers. 2003;72(5):305–317.

52. Hanlon EB, Manoharan R, Koo T, et al. Prospects for in vivo Raman spectroscopy. Phys Med Biol. 2000;45(2):R1.

53. Farquharson S, Shende C, Inscore FE, et al. Analysis of 5-fluorouracil in saliva using surface-enhanced Raman spectroscopy. J Raman Spectrosc. 2005;36(3):208–212.

54. Kaminaka S, Yamazaki H, Ito T, et al. Near-infrared Raman spectroscopy of human lung tissues: possibility of molecular-level cancer diagnosis. J Raman Spectroscopy. 2001;32(2):139–141.

55. Provenzano PP, Inman DR, Eliceiri KW, et al. Collagen density promotes mammary tumor initiation and progression. BMC Med. 2008;6(1):11.

56. Campagnola PJ, Clark HA, Mohler WA, et al. Second harmonic imaging microscopy of living cells. J Biomed Opt. 2001;6(3):277–287.

57. Provenzano PP, Eliceiri KW, Campbell JM, et al. Collagen reorganization at the tumor-stromal interface facilitates local invasion. BMC Med. 2006;4(1):38.

58. Conklin MW, Keely PJ. Why the stroma matters in breast cancer: insights into breast cancer patient outcomes through the examination of stromal biomarkers. Cell Adh Migr. 2012;6(3):249–260.

59. Campagnola PJ, Lewis A, Loew LM. High-resolution nonlinear optical imaging of live cells by second harmonic generation. Biophys J. 1999;77(6):3341–3349.

60. Huang Z, McWilliams A, Lui H, et al. Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. Int J Cancer. 2003;107 (6):1047–1052.

61. Cheng WT, Liu MT, Liu HN, et al. Micro-Raman spectroscopy used to identify and grade human skin pilomatrixoma. Microsc Res Tech. 2005;68(2):75–79.

62. Oliveira AFD, Santos IDDAO, Cartaxo SB, et al. Differential diagnosis in primary and metastatic cutaneous melanoma by FT-Raman spectroscopy. Acta Cir Bras. 2010;25(5):434–439.

63. Mehta K, Atak A, Sahu A, et al. An early investigative serum Raman spectroscopy study of meningioma. Analyst. 2018;143(8):1916–1923.

64. Lin K, Wang J, Zheng W, et al. Rapid fiber-optic Raman spectroscopy for real-time in vivo detection of gastric intestinal metaplasia during clinical gastroscopy. Cancer Prev Res. 2016;9 (6):476–483.

65. Wan QS, Wang T, Zhang KH. Biomedical optical spectroscopy for the early diagnosis of gastrointestinal neoplasms. Tumor Biol. 2017;39(7):1010428317717984.

66. Ghosh A, Raha S, Dey S, et al. Chemometric analysis of integrated FTIR and Raman spectra obtained by non-invasive exfoliative cytology for the screening of oral cancer. Analyst. 2019;144(4):1309–1325.

67. Maiti NC, Apetri MM, Zagorski MG, et al. Raman spectroscopic characterization of secondary structure in natively unfolded proteins: α-synuclein. J Am Chem Soc. 2004;126(8):2399–2408.

68. Mensch C, Bultinck P, Johannessen C. The effect of protein backbone hydration on the amide vibrations in Raman and Raman optical activity spectra. Phys Chem Chem Phys. 2019;21:1988–2005.

69. Krafft C, Dietzek B, Popp J. Raman and CARS microspectroscopy of cells and tissues. Analyst. 2009;134(6):1046–1057.

70. Lau DP, Huang Z, Lui H, et al. Raman spectroscopy for optical diagnosis in normal and cancerous tissue of the nasopharynx—preliminary findings. Lasers Surg Med. 2003;32(3):210–214.

71. Kast RE, Serhatkulu GK, Cao A, et al. Raman spectroscopy can differentiate malignant tumors from normal breast tissue and detect early neoplastic changes in a mouse model. Biopolymers. 2008;89(3):235–241.

72. Socrates G. Infrared and Raman characteristic group frequencies: tables and charts. Hoboken (NJ): John Wiley & Sons; 2004.

73. Mikhonin AV, Ahmed Z, Ianoul A, et al. Assignments and conformational dependencies of the amide III peptide backbone UV resonance Raman bands. J Phys Chem A. 2004;108(49):19020–19028.