Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 310 (2024) 123941

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.journals.elsevier.com/spectrochimica-acta-part-a-molecular-and-biomolecular-spectroscopy

Recognition of breast cancer subtypes using FTIR hyperspectral data

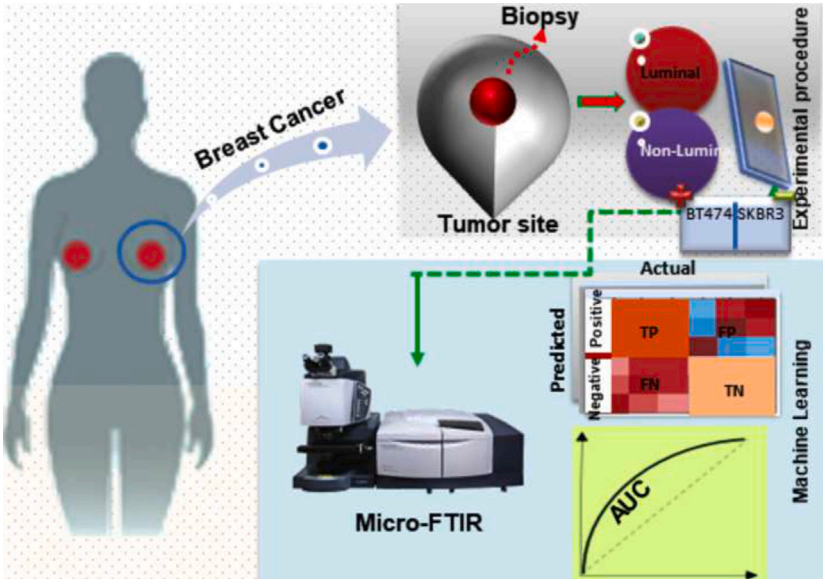
Sajid Farooq [a], Matheus del-Valle [a], Sofia Nascimento dos Santos [b], Emerson Soares Bernardes [b], Denise Maria Zezell [a,*]

[a] Center for Lasers and Applications, Instituto de Pesquisas Energeticas e Nucleares, IPEN—CNEN, Address One, Sao Paulo, 05508-000, Sao Paulo, Brazil
[b] Center for Radiopharmaceutics, Instituto de Pesquisas Energeticas e Nucleares, IPEN—CNEN, Address One, Sao Paulo, 05508-000, Sao Paulo, Brazil

HIGHLIGHTS

- Breast cancer subtypes were evaluated by micro-FTIR and 3D discriminant analysis.
- 3D discriminant analysis algorithm significantly improved discrimination performance.
- Proposed method accuracy is over 98%, in opposition to 85% conventional unfold one.
- 3D-DA provided biochemical signatures in BC that may be used for tailored treatments.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
FTIR
Breast cancer
Hyperspectral data
Discriminant analysis

ABSTRACT

Fourier-transform infrared spectroscopy (FTIR) is a powerful, non-destructive, highly sensitive and a promising analytical technique to provide spectrochemical signatures of biological samples, where markers like carbohydrates, proteins, and phosphate groups of DNA can be recognized in biological micro-environment. However, method of measurements of large cells need an excessive time to achieve high quality images, making its clinical use difficult due to speed of data-acquisition and lack of optimized computational procedures. To address such challenges, Machine Learning (ML) based technologies can assist to assess an accurate prognostication of breast cancer (BC) subtypes with high performance. Here, we applied FTIR spectroscopy to identify breast cancer subtypes in order to differentiate between luminal (BT474) and non-luminal (SKBR3) molecular subtypes. For this reason, we tested multivariate classification technique to extract feature information employing three-dimension (3D)-discriminant analysis approach based on 3D-principle component analysis-linear discriminant analysis (3D-PCA-LDA) and 3D-principal component analysis-quadratic discriminant analysis (3D-PCA-QDA), showing an improvement in sensitivity (98%), specificity (94%) and

* Corresponding author.
E-mail address: zezell@usp.br (D.M. Zezell).

https://doi.org/10.1016/j.saa.2024.123941
Received 22 September 2023; Received in revised form 22 December 2023; Accepted 20 January 2024
Available online 24 January 2024

accuracy (98%) parameters compared to conventional unfolded methods. Our results evidence that 3D-PCA-LDA and 3D-PCA-QDA are potential tools for discriminant analysis of hyperspectral dataset to obtain superior classification assessment.

1. Introduction

Breast cancer (BC) is the one of the most common heterogeneous disease with tumors exhibiting changeable morphology, behavior, molecular-profiles and response to therapy [1]. According to International Agency for Research on Cancer Worldwide (GLOBOCAN 2020), BC among women has excelled than lung-cancer as the leading diagnosed disease, with 2.30-million predicted more new cases (11.70%) than lung-cancer (11.40%), colorectal-cancer (10.00%), prostate-cancer (7.30%) and stomach-cancer (5.60%) [2]. Rising evidence presents that BC depicts variable growth rates that has important medicolegal and clinical consequences [3]. Thus, death rates for women due to BC disease were significantly soaring in transitioning versus transitioned countries i.e. 15.00 versus 12.80 per 100,000 as well as 12.40 versus 5.20 per 100,000, respectively. As a consequence, the worldwide cancer implication is estimated to be 28.40-million cases till 2040, a 47% increase from 2020, with a rapid rise in transitioning (64% to 95%) vs. transitioned (32% to 56%) countries because of demographic variations [2]. Albeit, this could be further worsened by growing risk factors related to globalization and challenging economy. However, early detection of disease can enhance treatment, decrease mortality and reduce economic burden.

In the complex landscape of BC, characterized by a diversity of molecular cell lines, it becomes important for researchers to precisely differentiate between the BT474 and SKBR3 cell lines. The accurate distinction between these cell lines is pivotal for advancing both preclinical and clinical research. BT474 embodies the luminal subtype, identified by the presence of hormone receptors, while SKBR3 stands as the HER2-positive subtype, marked by amplified human epidermal growth factor receptor 2 (HER2). This distinction holds immediate therapeutic relevance and extends beyond theoretical significance, emphasizing the practical implications for improved treatment strategies [4]. Exploring the therapeutic significance of subtype identification, tailored treatment approaches like hormone therapy for luminal subtypes and HER2-targeted treatments for HER2-positive subtypes underscores the practical value [5,6]. These cell lines not only contribute to drug discovery but also serve as valuable models for comprehending the intricate biology of BC. Their role extends beyond the laboratory, facilitating the ongoing quest for customized and effective treatment interventions. To delve into the complexities of BT474 and SKBR3 cell lines and explore their potential in advancing precision medicine in oncology, thus demanding advanced techniques. Therefore, there is an urgent demand of reliable, non-invasive techniques for monitoring, diagnosis and prognosis of BC.

Several attempts have been made to identify BC subtypes via mammography, magnetic resonance imaging (MRI), and ultrasonography. Contrary to optical conventional imaging techniques, hyperspectral imaging (HSI) allows to capture both spatial and chemical information, where each spectrum is composed of a pixel in original image [7]. The HSI data are depicted by 3D arrays, where spatial-coordinates are represented in x-axis, *y*-axis, and wavenumbers in *z*-axis. These wavenumbers are processed to stack up on one another to construct a 3D object, termed as a data cube [8]. To deal HSI data, there are several chemometric data mining techniques to manipulate such kind of data, including statistical analytical methods, for instance, partial least squares (PLS), principal components analysis (PCA), and multivariate curve resolution.

One of those HSI techniques that has been found an emerging tool for potential applications is Fourier-transform infrared (FTIR) spectroscopy [9,10]. FTIR is a powerful tool for analyzing cellular and tissue samples at a microscopic level, evaluating valuable chemical signatures into their molecular structure as well as chemical composition [11]. Unlike other instrumental techniques, micro-FTIR leads for label-free and non-invasive analysis [12], eradicating the requirement for exogenous markers and complex sample preparation. By exploiting the intrinsic vibrational properties of molecules, micro-FTIR provides the investigation of biomolecular changes, such as protein conformation [13], lipid composition [14], and nucleic acid structures [15]. Due to its high spatial resolution, micro-FTIR allows the examination of specific regions of interest within the samples, improving the accuracy and precision of the analysis [16]. Furthermore, recent progress in micro-FTIR instrumental and data analytic methods have further enhanced its susceptibility, allowing the recognition of certain molecular changes attributed to several cellular processes.

Further, FTIR has been exclusively applied in diagnosis [17], microbiological analysis [18], drug screening analysis [19], monitoring glucose [20], forensic study [21], environmental studies [22], and sugar-quantification [23,24]. Moreover, as a versatile and non-destructive technique, FTIR has been randomly used for medical/biological purposes, for instance, for the detection of breast cancer, BC molecular subtypes, brain tissues, ovarian tumors, skin cancer, and liver diseases. However, there are several major challenges of protocols for FTIR utilization in order to use for multivariate classification such as standardization procedures and pre-processing parameters [25]. Further, because of big resolution as well as co-added scans acquire excessive acquisition time, that make it impossible for acquisition of numerous samples in short intervals [26].

In this work, we propose to employ 3D-discriminant analysis approach for identification of two important BC subtypes (BT474 and SKBR3). The discriminant analysis is based on 3D-principal component analysis-quadratic discriminant analysis (3D-PCA-QDA) and 3D-principal component analysis-linear discriminant analysis (3D-PCA-LDA). The hyperspectral images of BC subtypes samples were obtained from FTIR spectroscopy.

2. Materials and methods

2.1. Samples preparations

2.1.1. Cell culture

Two samples i.e. BT474 (ATCC: HTB-20), a luminal B subtype (ER/PR/HER2 +Ve) and SKBR3 cells (ATCC : HTB-30), a non-luminal (HER2) subtype (ER/PR -Ve and HER2 positive) [6] were cultured in DMEM and supplemented with 10% of fetal bovine-serum and 50 μg/mL of gentamicin (Gibco, Life technologies, MD, USA) at 5% CO_2 and 37 °C.

2.1.2. Tumor growth

For this study, Balb/c nude mice were bred at the animal facility (IPEN-SP/CNEN) and all the experiments were complied with the relevant laws and approved by local animal ethics committee (protocol No.: 203/17). For tumor induction, Balb/c nude mice were subcutaneously injected with 1×10^6 of BT474 or SKBR3 cells (5/group). The tumors growth were assessed 3-times per week by caliper measurement. As the tumors reached around 500 mm^3, mice were euthanized. Then tumors were collected and processed by formalin-fixation and paraffin embedding. The 20 sections of 5 μm for each tumor sample, were obtained applying a microtome. Further, each section was fixed in a low-e microscope slide (MirrIR, Kevley Technologies, USA) and 10-fields per section were analyzed as well as averaged.

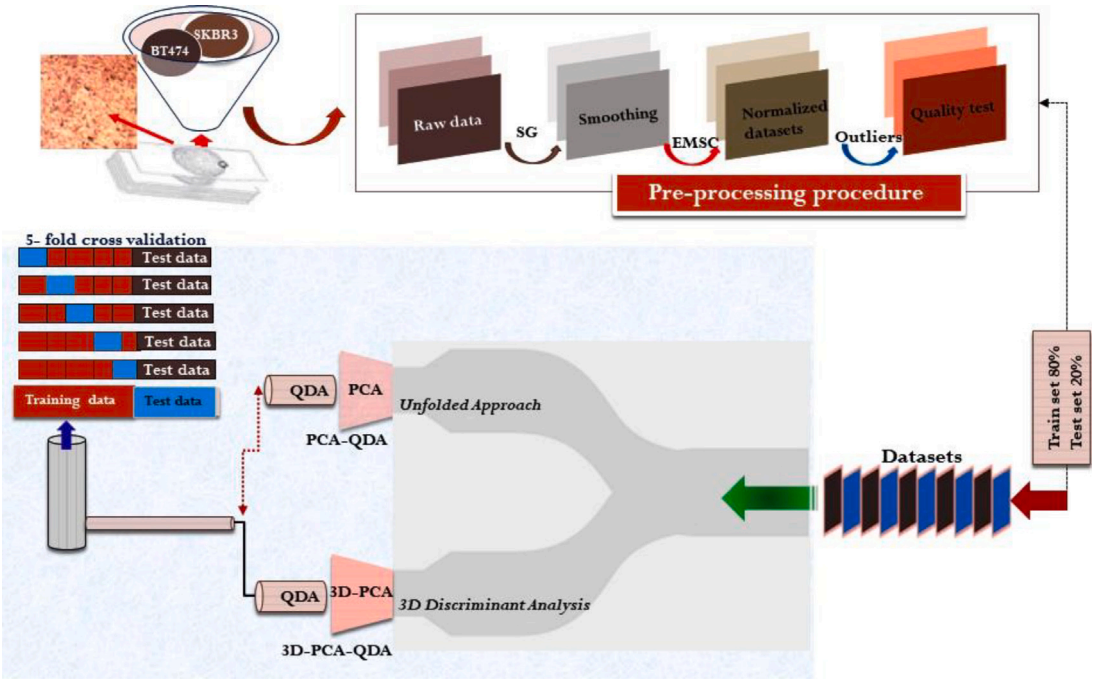


Fig. 1. A schematic diagram contains both pre-processing procedure and computational simulated analysis: the pre-processing method depends on smoothing (Savtizky-Golay), scattering correction using extended multivariate scattering correction (EMSC) and removal of outliers with quality test. The computational analysis is based on unfolded versus 3D discriminant analysis based approach.

2.2. Data acquistion and software

The input data for computational modeling was acquired applying a Cary-Series 600 system (Agilent Technologies, USA), combined by a Cary-660 FTIR-spectrometer and a Cary-620 FTIR-microscope. This setup consists of a focal plane array detector (32 × 32 elements) with spatial resolution of 5.5 μm, enabling 1024 spectra per acquisition. Furthermore, the optical configuration of FTIR composed of an objective (15x, NA = 0.62), with about 21-mm working distance and data collected in high transflection mode. The system was managed to function with data-spacing 3950–900 cm^{-1}, with 4 cm^{-1} per resolution.

The FTIR hyperspectral images were processed in Python 3.0. All the samples were preprocessed using Savitzky-Golay (SG) (window 15, second order polynomial fitting). Then, the extended multiplicative signal correction (EMSC) was applied for normalization and to perform a digital de-waxing. In addition, quality test (Hotelling's T^2 vs. Q-residuals) [27], was performed with the help of partial least regression using confidence interval (95%). The analysis based on unfolded and 3D-discriminant analysis were performed in Jupyter Notebook.

2.3. Computational modeling analysis

In order to evaluate the model performance, we used unfolded (PCA-LDA, PCA-QDA) and 3D-discriminant (3D-PCA-LDA, 3D-PCA-QDA) analysis approaches. Thereby, 3D-PCA procedure used in this study, a regular PCA is applied to each point of the hyperspectral image surface using the nonlinear iterative partial least squares (NIPLAS) algorithm Supplementary Material).

In order to deal 3D-discriminant (3D-PCA-LDA and 3D-PCA-QDA) analysis approaches, we used two important classifiers i.e., the quadratic discriminant analysis (QDA), and linear discriminant analysis (LDA) algorithms to the mean-scores of 3D-PCA. Therefore, the scores of 3D-PCA-QDA (Q_{ij}) as well as 3D-PCA-LDA (L_{ij}) are calculated as given below [28,29]:

$$Q_{ij} = (x_i - \overline{x_j})^T C_j^{-1}(x_i - \overline{x_j}) + log_e|C_j| - 2log_e\pi_j \quad (1)$$

$$L_{ij} = (x_i - \overline{x_j})^T C_{pooled}^{-1}(x_i - \overline{x_j}) - 2log_e\pi_j \quad (2)$$

where x_i represents 1× N is the mean-scores of T for sample i, and x_j is a row-vector of 1× N presenting the mean-scores of class j for their respective PCs. Moreover, the variables such as C_{pooled}^{-1} and C_j represent pooled co-variance matrix and variance–covariance matrix of class j. In order to evaluate the above relations (eqs. (2), (3)), C_{pooled}^{-1} and C_j are calculated as reported [28];

$$C_{pooled} = \frac{1}{n}\Sigma_{j=1}^{J} n_j C_j \quad (3)$$

$$C_j = \frac{1}{n_j - 1}\Sigma_{i=1}^{n_j}(x_i - \overline{x_j})^T \quad (4)$$

$$\pi_j = \frac{n_j}{n} \quad (5)$$

where n shows the total number of samples in training set and J depicts total number of classes and n_j is the samples of class j.

2.4. Performance evaluation

The 3D discriminant analysis approaches were employed to calculate the performance parameters such as sensitivity, specificity and accuracy. Those parameters were given as follow;

$$Sensitivity = \left[\frac{TP}{(TP + FN)}\right] \times 100 \quad (6)$$

$$Specificity = \left[\frac{TN}{(TN + FP)}\right] \times 100 \quad (7)$$

$$Accuracy = \left[\frac{TP + TN}{(TP + FN + TN + FP)}\right] \times 100 \quad (8)$$

where TP, TN, FP and FN are represented as true positive, true negative, false positive and false negative, respectively. Furthermore, in order for exploring correct classification rates in cross-validation, training and testing sets, the confusion matrices were evaluated using *Python3.0*.

3. Results

The BC subtypes dataset used in this study were composed on BT474 and SKBR3 hyperspectral imaging samples. Their HSI images were

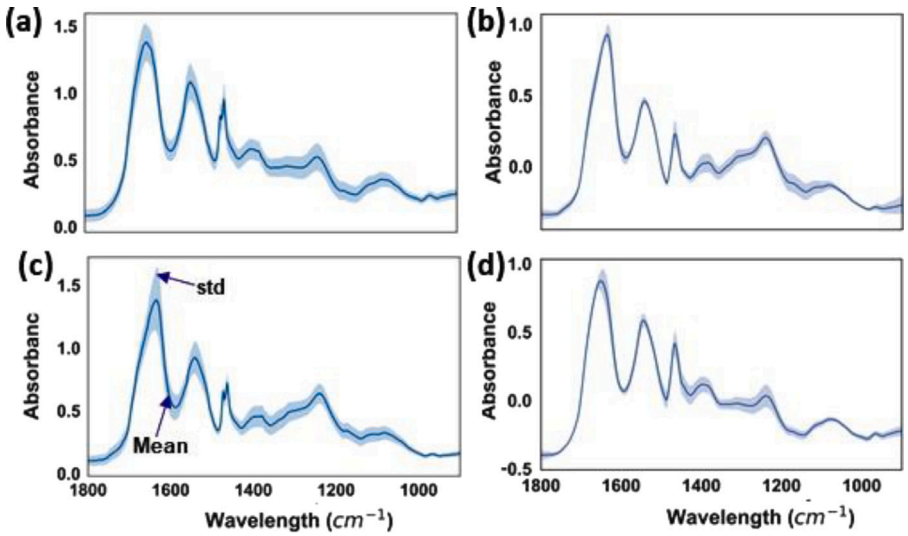


Fig. 2. The dataset used in this study: Raw (left-side), and Pre-processed (right-side). The data is obtained using FTIR hyperspectra images for BT474 (a, b) and SKBR 3 (c,d). Pre-processing is involved using Savitzky–Golay (SG), smoothing (window 15, polynomial fitting employing order 2), extended multivariate scattering corrections (EMSC) and quality test with PLS regression.

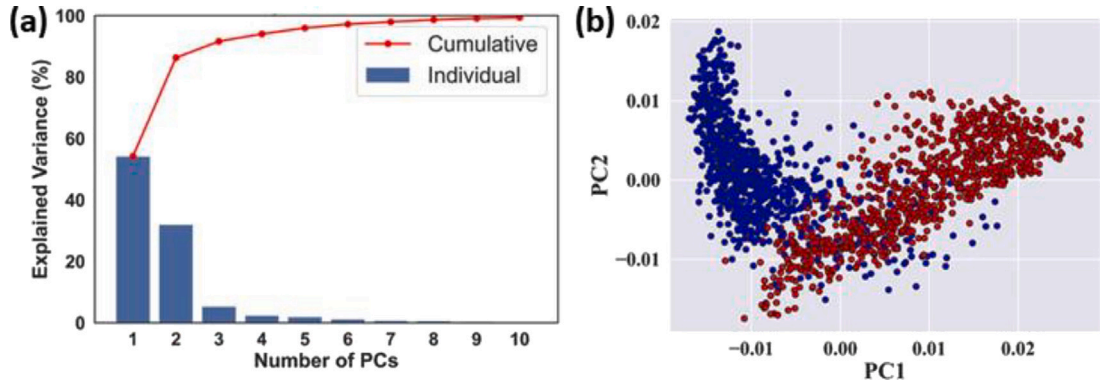


Fig. 3. The exploratory analysis of dataset: (a) explained variance versus number of PCs and (b) 2D scores plot of PCA of BT474 (red dots) and SKBR3 (blue dots).

processed to spectral dataset. The data comprising of 1024 spectra of each sample (totaling 2048 spectra). For study purpose, these spectra include the fingerprint regime (900–1800 cm^{-1}) to obtain biochemical signatures of the main molecular information within the samples. Fig. 1 shows a whole procedure that presents the pre-processing steps in detailed as well as the computational analysis approach. The pre-processing steps involve SG filter, EMSC, and quality test for smoothing, baseline correction, normalization, scattering correction and outlier handling. One can see the distinct difference between the raw (left), and pre-processed (right) data, as shown in Fig. 2. The three different methods were applied to the raw data: the first based on smoothing and vector normalization (SG followed by base-line correction) that removes the random noise and corrects the base-line. The second which removes the light scattering using EMSC and third approach (quality test+ PLS) which removes the outliers.

To differentiate between classes (i.e. BT474 and SKBR3), pattern recognition algorithms were applied. Fig. 3 presents exploratory analysis employing PCA procedure to reduce the dimensions of multivariate data. It can be seen that explained variance with respect to numbers of PCs (Fig. 3a) which shows the data variations exclusively in three dimensions (up to 90%). Fig. 3b shows the association of data variance and separation based on PC1 and PC2 scores. The major data variance can be seen on PC1 (>56%) and then on PC2 (>%30), respectively. Further, the scores plot also depicts each spectrum as a point in the space. The spectrum that are similar to each other are closer, and dissimilar spectra are spaced further apart (Fig. 3b). This kind of visualization assists for patterns of similar dataset to be identified quickly.

Fig. 4 presents the classification using 3D-PCA-QDA (PCs = 3). Fig. 4a shows the discriminant function of based on the score plot of 3D-PCA-QDA using pre-processed dataset. As seen, there is a distinctive separation between 2 classes with high accuracy. Further, the receiver operating characteristic curve is used to evaluate the accuracy by calculating area under the curve (AUC). As a result, the obtained accuracy is up to 98%, presenting by shaded AUC.

In order to analyze the 3D-discriminant analysis approach (3D-PCA-LDA, 3D-PCA-QDA) versus unfolded method (PCA-LDA, PCA-QDA), we used pre-processed data using 2-PCs. Fig. 5a and Fig. 5c present the unfolded approaches to calculate the classification decision boundaries between BT474 (red dots) and SKBR3 (blue dots) samples. To evaluate the response, the scores obtained from PCA were averaged per sample, indicating every point as a sample. One can observe the superposition pattern over the samples is very poor, reflecting imperfect classification employing conventional unfolded approaches (Fig. 5a,c).

This misclassification can be seen clearly in Table 1, where BC samples were misclassified during the unfolded PCA-LDA model in the test datasets. Further, Table 2 shows the accuracy 84%, sensitivity 81% and specificity 89% for PCA-LDA, while the accuracy, sensitivity and specificity for PCA-QDA are respectively 85%, 84% and 85%. These results show that the unfolded methods i.e. PCA-LDA and PCA-QDA algorithms are substantially poor. Contrary to that, by applying

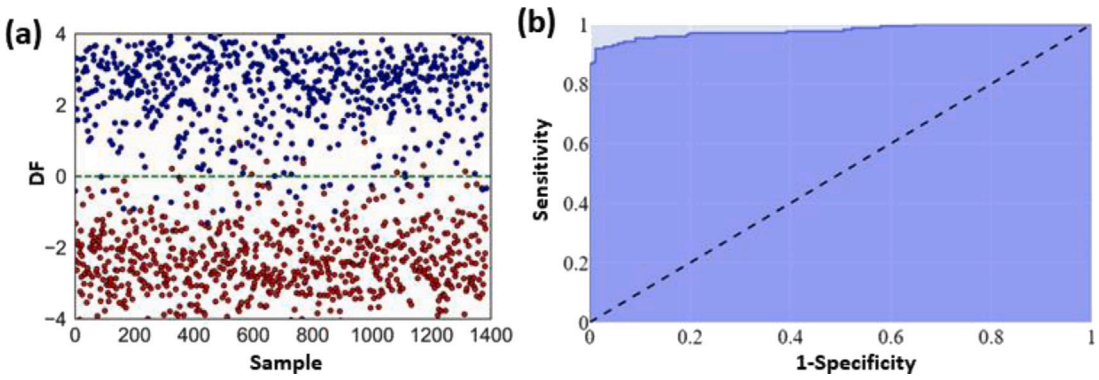


Fig. 4. The FTIR dataset used in the study to evaluate model performance using 3D discriminant analysis: (a) discriminant function of two BC subtypes (red dots: BT474, blue dots:SKBR3) (b) Receiver operating characteristics curve (ROC) of two subtypes and model performance is demonstrated by the area under curve (AUC).

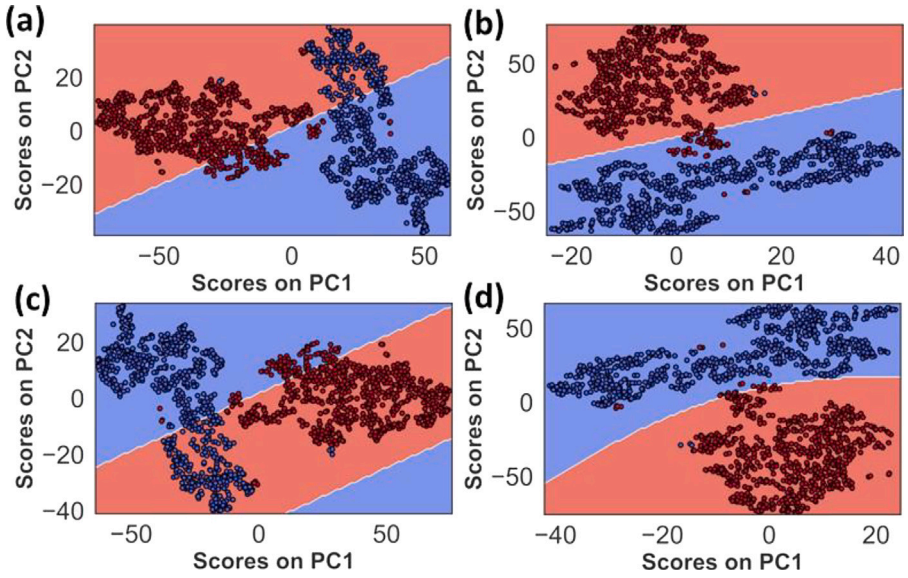


Fig. 5. The decision boundary of classification using unfolded (PCA-LDA and PCA-QDA) and 3D-discriminant analysis (3D-PCA-LDA and 3D-PCA-QDA): (a,c) demonstrate the classification decision boundaries between SKBR3 and BT 474 using unfolded approaches and (b,d) comparing with 2-D illustration where decision boundary reflects high separation between two classes using 3D discriminant analysis method.

Table 1
The confusion matrices for spliting dataset i.e. training, testing and cross-validation using unfolded and discriminant analysis approaches.

Methods	PCA-LDA		PCA-QDA		3D-PCA-LDA		3D-PCA-QDA	
Training	BT474	SKBR3	BT474	SKBR3	BT474	SKBR3	BT474	SKBR3
BT474	82%	18%	84%	16%	96%	4%	98%	2%
SKBR3	14%	86%	15%	85%	10%	90%	8%	92%
CV								
BT474	82%	18%	83%	17%	94%	6%	97%	3%
SKBR3	10%	90%	10%	90%	9%	91%	9%	91%
Test								
BT474	80%	20%	81%	19%	93%	7%	96%	4%
SKBR3	13%	87%	14%	86%	10%	90%	5%	95%

Table 2
Performance parameters (accuracy, sensitivity, specificity) of computational models using the unfolded (PCA-QDA, PCA-LDA) versus 3D-discriminant analysis for the BT474 and SKBR3 samples.

Data Analysis	Model	Accuracy	Sensitivity	Specificity
Unfolded	PCA-LDA	84%	81%	89%
	PCA-QDA	85%	84%	85%
3D	3D-PCA-LDA	93%	96%	91%
	3D-PCA-QDA	98%	98%	94%

3D-discriminant analysis based algorithms (3D-PCA-QDA and 3D-PCA-LDA), the performance of computational models to differentiate classes substantially enhanced. As shown in Fig. 5b and Fig. 5d, the 3D discriminant based algorithms showed the separation of BC samples between BT474 and SKBR3. One can see the clear distinct classification rates when using 3D-PCA-LDA (Fig. 5b) as well as 3D-PCA-QDA (Fig. 5d) discriminant algorithms approaches. As shown in Fig. 4b, the 3D-PCA-LDA algorithm depicts a few samples of BT474 subtype cross the decision boundary to SKBR3 dataset; whilst in 3D-PCA-QDA method, BT474 sample is projected on the class boundary. There is a clear substantial improvement of correct classification rate of 90% and 91%

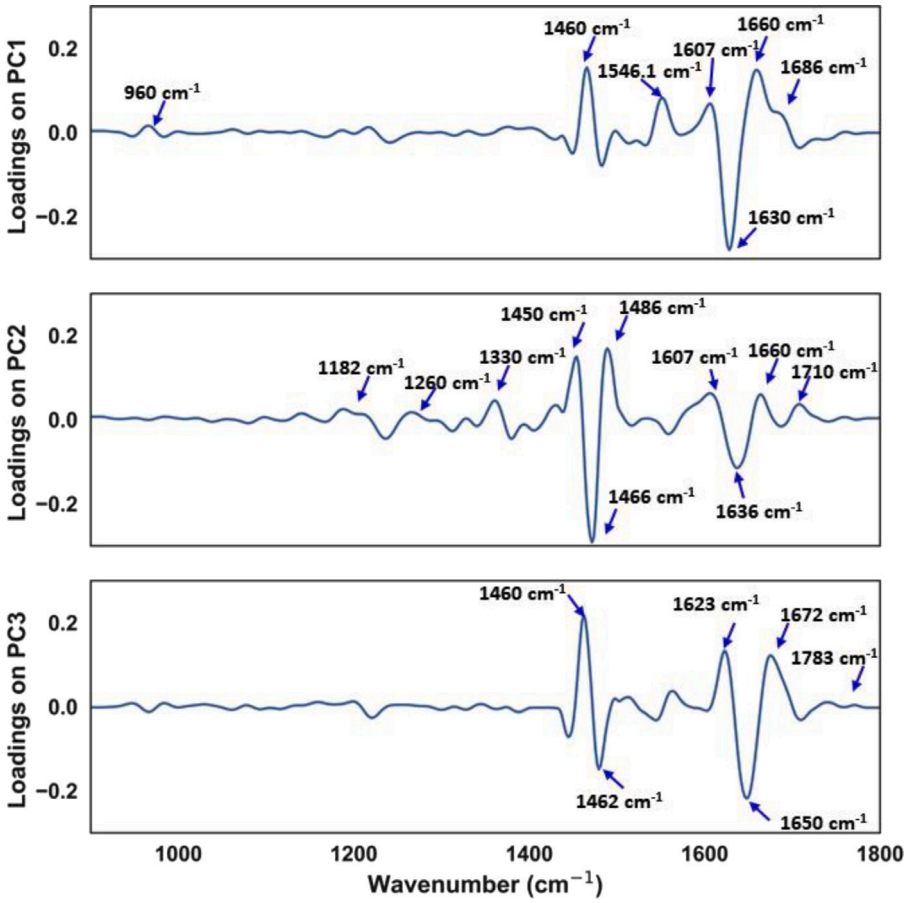


Fig. 6. 3D-PCA loadings: (a) loading on PC1, loading on PC2 and loading on PC3.

can be seen for SKBR3 in training (80%) and CV (KFold = 5) datasets for using 3D-PCA-LDA algorithm, respectively. However, for 3D-PCA-LDA method, the classification rate is 92% and 91% for SKBR3 sample in training and CV datasets respectively, as presented in Table 1. This distinct classification rate influences the model performance parameters, as seen in Table 2. In cross-validation dataset, the classification performance i.e. accuracy, sensitivity and specificity are 93%, 96% and 91%, respectively using 3D-PCA-LDA algorithm. For 3D-PCA-QDA model, the accuracy, specificity and sensitivity are improved i.e. 98%, 94% and 98%, respectively (Table 2). Our results present the clear advantage of using 3D discriminant analysis compared to unfolded algorithms approaches [28] (see Fig. 6).

Fig. 5 shows the 3D-PCA loading profiles (PC1, PC2 and PC3) to evaluate the chemical vibrational information. The loading on PC1 shows higher coefficient at several points: 960 cm^{-1} (PH_3^{-4}), 1460 cm^{-1} (CH-deformation), 1546.1 cm^{-1} (amide II), 1607 cm^{-1} (C–C phenylalanine, tyrosine), 1660 cm^{-1} (C=C), 1686 cm^{-1} (Amide I α-helix). The 3D-PCA loading on PC2 presents another aspect related chemical signatures and contains higher coefficients at: 1182 cm^{-1} (A, U, G, C out of plane ring deformation Phe.), 1260 (=CH deformation, amide III *alpha*–helix), 1330 cm^{-1} (G), 1486 cm^{-1} (A, G), 1607 cm^{-1} (C-C phenylalanine, tyrosine), 1660 cm^{-1} (C=C). Further, the loading due to 3D-PCA on PC3 contains prominent coefficients at 1460 cm^{-1} ((p), (l) CH-deformation), 1623 cm^{-1} (tryptophan/tyrosine), 1672 cm^{-1} (Amide I), 1743 cm^{-1} (C=O stretching).

4. Discussion

In this article, we conducted a computational analysis of BC subtypes in the context of FTIR hyperspectral dataset using semi-supervised learning 3D-discriminant approach i.e. 3D-PCA-LDA and 3D-PCA-QDA algorithms. Motivated by Ref. [28] computational model to manipulate new features for disease detection, we investigated the effectiveness of this model in molecular subtypes (BT474 and SKBR3). Our results depicted that there were highly divergence of these two subtypes by separation of FTIR data into discrete clusters. These clusters are associated with defined subtypes and could be validated employing 3D-discriminant analysis algorithm.

In this work, the 3D-discriminant analysis approach was trained on 80% of the entire dataset and then tested for the remaining 20% to conduct statistical algorithm. For instance, the distinct classification was obtained using discriminant analysis approach and clear separation can be observed from two classes through discriminant function (Fig. 4a). Moreover, the accuracy obtained using ROC curve is around 98%. Also the superposition pattern demonstrated the pure classification due to 3D-PCA-LDA as well 3D-PCA-QDA methods, comparing to unfolded approach (PCA-LDA, PCA-QDA). As a result, the classification performance in training, cross-validation and test dataset were remarkably improved employing 3D based algorithms. For example, the accuracy for 3D-PCA-QDA is up to 98% in comparison with 85% of PCA-QDA algorithm (Table 2). Ha et al. developed a neural network algorithm to predict BC subtypes based on MRI features, and found 70% test set accuracy [30], which is lower performance than our reported model.

Although each molecular subtype showed a particular FTIR spectra (mean spectrum), so specific chemical signature can be obtained from peak shifts around different wavenumbers. From Fig. 2b,d, our results evidence that STD reduce not only in Amide-I as well as amide-II region, but also from 1350–900 cm^{-1}, which are mainly attributed to amide III (1350–1200 cm^{-1}), RNA and DNA (1235 to 1080 cm^{-1}), as well carbohydrates (1200–900 cm^{-1}) contents [31]. The 3D-PCA loading profiles were developed to investigate bio-markers using wavenumbers. The PCA loadings present larger coefficients such as peaks at

1265 cm^{-1}, 1647–1672 cm^{-1} and 1744 cm^{-1} relate to higher lipids contents in ER/PR^{+}-group (BT474) [32]. This is in agreement with the expression of 'FOXA1' gene, ER/PR^{+} in comparison to ER/PR^{-} (SKBR3) [32] . Further, loading profiles at 1607 cm^{-1}, 1660 cm^{-1}, and 1686 cm^{-1} represent C–C Phy, Tyr, C=C stretching, and amino I $\alpha-helix$ respectively. Vibrations around 1450 cm^{-1} is amide I band. Moreover, SKBR3 and BT474 possess the similar DNA features [31].

Finally, the ability of vibrational spectroscopic (e.g. Raman, FTIR) techniques to manipulate breast cancer disease successfully established at the cellular and tissue levels to assess the chemical signatures of inflammatory BC [33], BC tissues and lymph nodes for detection of recurrence [34], MCF-7 BC cell line [35], and BC triple-negative cell line [36]. These studies also discovered specific markers for an objective diagnosis of breast cancer. However, combining vibrational spectroscopic techniques with ML can assess molecular changes to facilitate early identification of high risk cancers, ultimately enhancing clinical decision making uptakes [35,36]. Taken together, our findings can provide avenues of evaluations of breast cancer identification for giving of therapeutic decisions for patients.

5. Conclusion

In this paper, we report a computational approach using 3D-discriminant analysis algorithms for the identification of breast cancer subtypes. We compared the performance of these algorithms with unfolded algorithms methods and showed that the 3D-discriminant algorithms outperformed them in identifying the BT474 and SKBR3 molecular subtypes based on FTIR dataset. The foundation of the 3D-discriminant analysis approach resulted in increasing accuracy, sensitivity, and specificity, improving from 84% to 98%, 81% to 98%, and 89% to 94%, respectively. These findings show the effectiveness of 3D-discriminant based algorithms in rapid and precise classification of hyperspectral dataset, surpassing the performance of traditional simulated algorithms.

CRediT authorship contribution statement

Sajid Farooq: Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Matheus del-Valle:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing – review & editing. **Sofia Nascimento dos Santos:** Data curation, Investigation, Methodology, Validation, Writing – review & editing. **Emerson Soares Bernardes:** Conceptualization, Funding acquisition, Validation, Writing – review & editing. **Denise Maria Zezell:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by CNPq, Brazil (INCT-INTERAS 406761/2022-1), INCT-INFO (465763/2014-6); Sisfoton (440228/2021-2); PQ (314517/2021-9); CAPES, Brazil Finance code 001 and FAPESP, Brazil (17/50332-0) and FAPESP, Brazil (21/00633-0).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.saa.2024.123941.

References

[1] R.E. Hendrick, J.A. Baker, M.A. Helvie, Breast cancer deaths averted over 3 decades, Cancer 125 (9) (2019) 1482–1488.
[2] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin. 71 (3) (2021) 209–249.
[3] P. Gamble, R. Jaroensri, H. Wang, F. Tan, M. Moran, T. Brown, I. Flament-Auvigne, E.A. Rakha, M. Toss, D.J. Dabbs, et al., Determining breast cancer biomarker status and associated morphological features using deep learning, Commun. Med. 1 (1) (2021) 1–12.
[4] C.M. Perou, T. Sø rlie, M.B. Eisen, M. Van De Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, et al., Molecular portraits of human breast tumours, Nature 406 (6797) (2000) 747–752.
[5] D.J. Slamon, Studies of the HER-2/neu proto-oncogene in human breast cancer, Cancer Invest. 8 (2) (1990) 253–254.
[6] X. Dai, H. Cheng, Z. Bai, J. Li, Breast cancer cell line classification and its relevance with breast tumor subtyping, J. Cancer 8 (16) (2017) 3131.
[7] J. Sacharz, D. Perez-Guaita, M. Kansiz, S.S. Nazeer, A. Wesełucha-Birczyńska, S. Petratos, B.R. Wood, P. Heraud, Empirical study on the effects of acquisition parameters for FTIR hyperspectral imaging of brain tissue, Anal. Methods 12 (35) (2020) 4334–4342.
[8] D. Porro-Muñoz, R.P. Duin, I. Talavera, M. Orozco-Alzate, Classification of three-way data by the dissimilarity representation, Signal Process. 91 (11) (2011) 2520–2529.
[9] C. Petibois, B. Desbat, Clinical application of FTIR imaging: new reasons for hope, Trends Biotechnol. 28 (10) (2010) 495–500.
[10] D. Biswal, J. Hilt, Analysis of oxygen inhibition in photopolymerizations of hydrogel micropatterns using FTIR imaging, Macromolecules 42 (4) (2009) 973–979.
[11] L.J. Macedo, F.P. Rodrigues, A. Hassan, L.N. Máximo, F. Zobi, R.S. da Silva, F.N. Crespilho, Non-destructive molecular FTIR spectromicroscopy for real time assessment of redox metallodrugs, Anal. Methods 14 (11) (2022) 1094–1102.
[12] H. Yang, X. Li, S. Zhang, Y. Li, Z. Zhu, J. Shen, N. Dai, F. Zhou, A one-dimensional convolutional neural network based deep learning for high accuracy classification of transformation stages in esophageal squamous cell carcinoma tissue using micro-FTIR, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 289 (2023) 122210.
[13] D. Tocco, D. Chelazzi, R. Mastrangelo, A. Casini, A. Salis, E. Fratini, P. Baglioni, Conformational changes and location of BSA upon immobilization on zeolitic imidazolate frameworks, J. Colloid Interface Sci. 641 (2023) 685–694.
[14] C.E.E. Grace, P.K. Lakshmi, S. Meenakshi, S. Vaidyanathan, S. Srisudha, M.B. Mary, Biomolecular transitions and lipid accumulation in green microalgae monitored by FTIR and Raman analysis, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 224 (2020) 117382.
[15] H.M. Elsheikha, N.A. Elsaied, K.A. Chan, C. Brignell, M.S. Harun, K. Wehbe, G. Cinquee, Label-free characterization of biochemical changes within human cells under parasite attack using synchrotron based micro-FTIR, Anal. Methods 11 (19) (2019) 2518–2530.
[16] S. Farooq, M. Del-Valle, M.O. Dos Santos, S.N. Dos Santos, E.S. Bernardes, D.M. Zezell, Rapid identification of breast cancer subtypes using micro-FTIR and machine learning methods, Appl. Opt. 62 (8) (2023) C80–C87.
[17] G. Bellisola, C. Sorio, Infrared spectroscopy and microscopy in cancer research and diagnosis, Am. J. Cancer Res. 2 (1) (2012) 1.
[18] L.-C. Fengou, A. Lianou, P. Tsakanikas, E.N. Gkana, E.Z. Panagou, G.-J.E. Nychas, Evaluation of Fourier transform infrared spectroscopy and multispectral imaging as means of estimating the microbiological spoilage of farmed sea bream, Food Microbiol. 79 (2019) 27–34.
[19] P.S. Sampaio, C.R. Calado, Potential of FTIR-spectroscopy for drugs screening against helicobacter pylori, Antibiotics 9 (12) (2020) 897.
[20] D.C. Caixeta, C. Lima, Y. Xu, M. Guevara-Vega, F.S. Espindola, R. Goodacre, D.M. Zezell, R. Sabino-Silva, Monitoring glucose levels in urine using FTIR spectroscopy combined with univariate and multivariate statistical methods, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 290 (2023) 122259.
[21] J.M. Duarte, N.G.S. Sales, J.W.B. Braga, C. Bridge, M. Maric, M.H. Sousa, J. de Andrade Gomes, Discrimination of white automotive paint samples using ATR-FTIR and PLS-DA for forensic purposes, Talanta (2021) 123154.
[22] C.M. Simonescu, Application of FTIR spectroscopy in environmental studies, Adv. Aspects Spectrosc. 29 (1) (2012) 77–86.
[23] O. Anjos, M.G. Campos, P.C. Ruiz, P. Antunes, Application of FTIR-ATR spectroscopy to the quantification of sugar in honey, Food Chem. 169 (2015) 218–223.

[24] I.F. Duarte, A. Barros, I. Delgadillo, C. Almeida, A.M. Gil, Application of FTIR spectroscopy for the quantification of sugars in mango juice as a function of ripening, J. Agric. Food Chem. 50 (11) (2002) 3104–3111.
[25] C.L. Morais, K.M. Lima, M. Singh, F.L. Martin, Tutorial: multivariate classification for vibrational spectroscopy in biological samples, Nat. Protoc. 15 (7) (2020) 2143–2162.
[26] M. del Valle, M.O. Dos Santos, S.N. Dos Santos, P.A.A. de Castro, E.S. Bernardes, D.M. Zezell, The impact of scan number and its preprocessing in micro-FTIR imaging when applying machine learning for breast cancer subtypes classification, Vib. Spectrosc. 117 (2021) 103309.
[27] J. Cavaglia, D. Schorn-García, B. Giussani, J. Ferré, O. Busto, L. Aceña, M. Mestres, R. Boqué, Monitoring wine fermentation deviations using an ATR-MIR spectrometer and MSPC charts, Chemometr. Intell. Lab. Syst. 201 (2020) 104011.
[28] C.L. Morais, P. Giamougiannis, R. Grabowska, N.J. Wood, P.L. Martin-Hirsch, F.L. Martin, A three-dimensional discriminant analysis approach for hyperspectral images, Analyst 145 (17) (2020) 5915–5924.
[29] C.L. Morais, P.L. Martin-Hirsch, F.L. Martin, A three-dimensional principal component analysis approach for exploratory analysis of hyperspectral data: identification of ovarian cancer samples based on Raman microspectroscopy imaging of blood plasma, Analyst 144 (7) (2019) 2312–2319.
[30] R. Ha, S. Mutasa, J. Karcich, N. Gupta, E. Pascual Van Sant, J. Nemer, M. Sun, P. Chang, M.Z. Liu, S. Jambawalikar, Predicting breast cancer molecular subtype with MRI dataset utilizing convolutional neural network algorithm, J. Digital Imaging 32 (2) (2019) 276–282.
[31] P. Meksiarun, P.H. Aoki, S.J. Van Nest, R.G. Sobral-Filho, J.J. Lum, A.G. Brolo, A. Jirasek, Breast cancer subtype specific biochemical responses to radiation, Analyst 143 (16) (2018) 3850–3858.
[32] F. Slebe, F. Rojo, M. Vinaixa, M. García-Rocha, G. Testoni, M. Guiu, E. Planet, S. Samino, E.J. Arenas, A. Beltran, et al., Foxa and LIPG endothelial lipase control the uptake of extracellular lipids for breast cancer growth, Nature Commun. 7 (1) (2016) 1–11.
[33] H.T. Mohamed, V. Untereiner, I. Proult, S.A. Ibrahim, M. Götte, M. El-Shinawi, M.M. Mohamed, G.D. Sockalingum, S. Brézillon, Characterization of inflammatory breast cancer: A vibrational microspectroscopy and imaging approach at the cellular and tissue level, Analyst 143 (24) (2018) 6103–6112.
[34] J. Depciuch, A. Stanek-Widera, N. Khinevich, H.V. Bandarenka, M. Kandler, V. Bayev, J. Fedotova, D. Lange, J. Stanek-Tarkowska, J. Cebulski, The spectroscopic similarity between breast cancer tissues and lymph nodes obtained from patients with and without recurrence: A preliminary study, Molecules 25 (14) (2020) 3295.
[35] K. Iwasaki, A. Araki, C.M. Krishna, R. Maruyama, T. Yamamoto, H. Noothalapati, Identification of molecular basis for objective discrimination of breast cancer cells (MCF-7) from normal human mammary epithelial cells by Raman microspectroscopy and multivariate curve resolution analysis, Int. J. Mol. Sci. 22 (2) (2021) 800.
[36] I.P. Santos, C.B. Martins, L.A. Batista de Carvalho, M.P. Marques, A.L. Batista de Carvalho, Who's who? Discrimination of human breast cancer cell lines by Raman and FTIR microspectroscopy, Cancers 14 (2) (2022) 452.