Vibrational Spectroscopy 117 (2021) 103309

Contents lists available at ScienceDirect

Vibrational Spectroscopy

journal homepage: www.elsevier.com/locate/vibspec

The impact of scan number and its preprocessing in micro-FTIR imaging when applying machine learning for breast cancer subtypes classification

Matheus del-Valle [a], Moisés Oliveira dos Santos [a,b], Sofia Nascimento dos Santos [c], Pedro Arthur Augusto de Castro [a], Emerson Soares Bernardes [c], Denise Maria Zezell [a,*]

[a] Centro de Lasers e Aplicações, Instituto de Pesquisas Energéticas e Nucleares, IPEN - CNEN, 05508-000, Brazil
[b] Escola Superior de Tecnologia, Universidade do Estado do Amazonas, 69050-030, Brazil
[c] Centro de Radiofarmácia, Instituto de Pesquisas Energéticas e Nucleares, IPEN - CNEN, 05508-000, Brazil

ARTICLE INFO

Keywords:
FTIR imaging
Scan number
Preprocessing
Breast cancer subtype
Machine learning
Classification

ABSTRACT

The breast cancer molecular subtype is an important classification to outline the prognostic. Gold-standard assessing using immunohistochemistry adds subjectivity due to interlaboratory and interobserver variations. In order to increase the diagnosis confidence, other techniques need to be examined, where the FTIR spectroscopy imaging allied with machine learning techniques may provide additional and quantitative information regarding the molecular composition. However, the impact of co-added scans acquisition parameter into machine learning classifications still needs better evaluation. In this study, FTIR images of Luminal B and HER2 subtypes were acquired varying the scan number and preprocessing techniques. It was demonstrated a spectral quality improvement when the scan number was increased, decreasing the standard deviation and outliers. Six machine learning models were used to classify the subtypes: Linear Discriminant Analysis, Partial Least Squares Discriminant Analysis, K-Nearest Neighbors, Support Vector Machine, Random Forest and Extreme Gradient Boosting. Best mean accuracy of 0.995 was achieved by Extreme Gradient Boosting model. It was found that all models achieved similar high accuracies with groups b256_064 (256 background and 064 scans), b256_128 and b128_128. Besides assessing the performance of different models, the b256_064 was established as the optimal group due to the minimum acquisition time. Therefore, this work indicates b256_064 for breast cancer subtype classification and also as a basis for other studies using machine learning for cancer evaluation.

1. Introduction

The breast cancer molecular subtypes classification plays an import role in its treatment, sorting patients with divergent prognoses and helping to select an appropriate and specific therapy [1]. Subtypes are mainly classified according to its expression of three hormone receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). There are four subtypes: luminal A (ER/PR positive, HER2 negative); luminal B (ER/PR positive, HER2 variable); HER2 (ER/PR negative, HER2 positive); and Triple Negative (ER/PR/HER2 negative) [2].

Subtypes are widely classified by immunohistochemistry semi-quantitative techniques, however, several issues affect its assessment quality, such as interlaboratory (antibodies, detection systems, and protocols used) and interobserver variations [3]. Fourier Transform Infrared (FTIR) spectroscopy has been studied as a further cancer evaluation technique in the past years, not only to overcome the variations, but also to provide additional information regarding the biochemical content, contributing to improve assessment quality [4,5].

With the consolidation of FTIR spectroscopy imaging, which provides thousands of spectra in a single acquisition, machine learning approaches stood out as powerful tools for many diagnostics, including cancer classification [6,7]. Several protocols for using FTIR to analyze biological samples have been published [8–10], standardizing acquisition and processing parameters, although number of co-added scans is minimally commented, without considering its effects in machine learning classifications.

Tahtouh et al., 2007 [11], studied parameters optimization, including three scans numbers (4, 16 and 64), although no biological samples or machine learning techniques were applied. Recently, Sacharz

* Corresponding author.
E-mail addresses: matheusdv@gmail.com (M. del-Valle), mosantos@uea.edu.br (M.O. dos Santos), snsantos85@gmail.com (S.N. dos Santos), pedroarthur@usp.br (P.A.A. de Castro), emerson.bernardes@gmail.com (E.S. Bernardes), zezell@usp.br (D.M. Zezell).

https://doi.org/10.1016/j.vibspec.2021.103309
Received 11 June 2021; Received in revised form 17 September 2021; Accepted 15 October 2021
Available online 19 October 2021
0924-2031/

et al., 2020 [12], evaluated acquisition parameters in an empirical study with brain tissue samples, varying scans from 1 to 128. Nonetheless, the background scans number were kept the same of the sample scan number, and only k-means clustering was applied. In this way, there is still a lack of consensus for a systematic approach to define an optimal number of co-added scans regarding a classification task.

2. Material and methods

2.1. Sample preparation

SK-BR-3 cells (ATCC number: HTB-30), a HER 2 subtype, ER/PR negative, HER2 positive, and BT-474 (ATCC number: HTB-20), a luminal B subtype ER/PR/HER2 positive [13], were cultured in DMEM (Gibco, Life technologies, MD, USA) supplemented with 10 % of fetal bovine serum (Gibco, Life technologies, MD, USA) and 50 μg/mL of gentamicin (Gibco, Life technologies, MD, USA). For *in vivo* studies, eight-week-old female Balb/c nude mice were subcutaneously injected with 1×10^6 BT-474 or SK-BR-3 cells and tumor growth was followed for 4 weeks. Balb/c nude mice were bred at the animal facility of Nuclear and Energy Research Institute, and all experiments complied with the relevant laws and were approved by local animal ethics committees (protocol number: 203/17). When tumors volume reached approximately 0.5 cm^3, biopsies were collected and processed by formalin fixation and paraffin embedding (FFPE) method. Twenty sections of 5 μm, ten for each cell line, were then obtained using a microtome and fixed in low-e microscope slides (MirrIR, Kevley Technologies, USA).

2.2. Image acquisition

Spectral images acquisition was accomplished using a Cary Series 600 system (Agilent Technologies, USA), composed by a Cary 660 FTIR spectrometer and a Cary 620 FTIR microscope. This system has a focal plane array (FPA) detector of 32 × 32 elements and 5.5 μm spatial resolution, providing 1024 spectra per acquisition. The system was set to operate between 3950 and 900 cm^{-1}, with 4 cm^{-1} spectral resolution in transflection mode, due to the use low-e slides.

In each histological section, the FTIR image was acquired by varying twice the number of co-added scans of the background, where each one was acquired before a batch of varying sample scans six times, totalizing 12 acquisition per section. Background scans were set to 128 and 256, while sample scans were set to 4, 8, 16, 32, 64 and 128. The groups were labeled as "b*x*_*y*", where *x* and *y* are the number of scans for the background and sample, respectively. Furthermore, adjacent paraffin regions images were acquired, varying scans from 4 to 128 as sample scans.

For a reproducibly purpose, the same histological section region was settled during the collection of all scans. Hence, single acquisitions were performed instead of mosaics (grouping several single acquisitions in one measure), resulting in 12,288 sample spectra acquired for each section. In addition, scans of the same section were collected in sequence, within the same spatial position, even for different background scans.

2.3. Data preprocessing and analysis

Data preprocessing was applied according to the protocols [8–10]. The pipeline was defined depending on the analysis, as describe in the next paragraph. General steps included fingerprint truncation; Savitzky–Golay (SG) filtering for smoothing with window size of 7 and for obtain the second derivative; extended multiplicative signal correction (EMSC) [14], with a zero degree polynomy, as second derivative does not need baseline correction, and digital de-waxing [15]; and outlier detection (quality test) using Hotelling's T^2 *versus* Q residuals with a fixed removal threshold of 95 % confidence interval.

The analysis was divided in two: a spectral analysis to better visualize the spectra distribution and variations; and a classification analysis to check how each group performs within the machine learning approaches.

For the spectral analysis, each scan group was preprocessed by three different pipelines, which originated three main groups division: RAW, PP and OUT. The Table 1 describes the applied preprocessing steps in each one.

Cell lines were not considered as a grouping feature for spectral analysis, therefore only scan number and preprocessing steps were compared in this part. The groups were analyzed using mean + standard deviation (SD) plots and principal component analysis (PCA) scores plots. The PCA scores and all the other analysis which used the fingerprint preprocessing refer to the biofingerprint region (1800 to 900 cm^{-1}), excepts the spectra for mean + SD plots, where the dead region (2000 to 1800 cm^{-1}) was added, hence showing spectra truncated from 2000 to 900 cm^{-1}.

For the classification analysis, PP and OUT groups division were used, besides the scan group division. In addition, another main division was also applied in this part, splitting the previous groups according to the cell line. In this way, the classification could be compared by scan number, preprocessing steps, and machine learning technique.

Six machine learning classifiers were modeled using default hyperparameters:

Linear Discriminant Analysis (LDA): number of components = 1; singular value decomposition solver with tolerance (significance threshold) = 1e-4.

Partial Least Squares Discriminant Analysis (PLS-DA): number of components = 10; nonlinear iterative solver with tolerance (convergence criteria) = 1e-6; maximum number of iterations = 500; discriminant threshold = 0.5.

K-Nearest Neighbors (KNN): number of neighbors = 5; Euclidean distance metric.

Support Vector Machine (SVM): radial basis function kernel with gamma coefficient = one divided by number of features times feature variance; cost parameter (regularization) = 1; tolerance (convergence criteria) = 1e-3.

Random Forest (RF): number of trees = 100; split metric = Gini impurity; maximum features = 21 (square root of the number of features); minimum impurity decrease = 0; no pruning.

Extreme Gradient Boosting (XGB): gbtree booster; learning rate = 0.3; maximum depth = 6; L2 regularization = 1; no L1 regularization; minimum loss decrease = 0;

Model training was performed by a stratified 5-fold cross-validation (CV), varying the fold split seed 10 times, resulting in 50 trainings per model. Preprocessing steps were applied after each fold split, where test data were only transformed according to the train fitting, preventing information leakage between train and test data. Models were assessed independently, where intra-groups test accuracies, *i.e.*, different scans performances from a same model and same PP or OUT group, were evaluated using Friedman + Nemenyi test [16] with a significance level of 5%.

All the study was accomplished using in house algorithms in Python, except by Friedman + Nemenyi test, written in R language. The main source codes can be found at https://github.com/delvallem/specDS and the usage at the supplementary material.

Table 1
Preprocessing steps applied in each group: RAW, PP, OUT.

Step	RAW	PP	OUT
Outlier	×	×	✓
Fingerprint	✓	✓	✓
Derivative	✓	✓	✓
Smoothing	×	✓	✓
EMSC	×	✓	✓

3. Results and discussion

3.1. Spectral analysis

The Fig. 1 shows a representative spectra comparison of lowest and highest scans for both background and sample (b128_004 and b256_128). The RAW plots demonstrate greater SD of b128_004 group in comparison to b256_128, as evidenced by its wider range of shades in relation to its mean, especially at Amide I and II region (1700 to 1500 cm^{-1}) [8,17]. The PP group presented a decrease in b256_128 SD, enabling a better visualization of the mean spectrum shape, while b128_004 did not present visual improvements due to the high SD. These findings suggest more outliers in b128_004, as the preprocessing techniques so far were not able to improve its quality. This tendency also occurs in OUT group plots, as the implementation of outlier removal method decreased b128_004 SD and made possible to distinguish its mean spectrum shape, which became like to b256_128. When compared to its PP, the OUT b256_128 group presented a slight decrease of its SD and a similar mean spectrum, indicating less impact of outliers, since there was less improvement than b128_004 after the outlier removal.

Comparing all groups (please check Supp. Fig. S1 to S6 in Supplementary Information), it is possible to verify the same changes, SD decrease and better definition of the mean spectrum shape, progressively from lowest to highest sample scan (004–128). The OUT plots (Fig. S5 and S6) evidence a SD decrease not only in Amide I and II region, but also within 1350 to 900 cm^{-1}, which are mainly related to amide III (1350 to 1200 cm^{-1}), DNA and RNA (1235 to 1080 cm^{-1}), and carbohydrates (1200 to 900 cm^{-1}) content. Lower SD in this region is expected as amide III is not reported with intense variation in breast cancer [17], and as BT-474 and SK-BR-3 present similar DNA features [18]. Besides, tumors from xenograft cells exhibit a less heterogenous tissue than a regular human tissue [19], thus decreasing features variation in this analysis.

The dead region (2000 to 1800 cm^{-1}) does not contain tissue information, thus the SD of this region is only affected by environmental fluctuations [7]. These fluctuations are mainly related to water vapor content, as its absorption region is characterized from 2072 to 1205 cm^{-1} [20]. The dead region did not exhibit SD in any group, thus indicating that there was no water vapor contribution to the different SD range in the biofingerprint region (1800 to 900 cm^{-1}) among the groups. This fact is mainly related to the fast single scan acquisitions in a sequential way. A 128 scan took an average of 2 min to be acquired, where halving the scan almost halves the acquisition time, hence obtaining scans variations in similar ambient conditions.

PCA plots (Fig. 2) corroborates with spectra comparison, where b128_004 presented sparse scores distribution in RAW and PP groups, which can also be evidenced by the scale range, and as observed in b256_128 the distribution of same groups was concentric with few sparse points. Both OUT plots show clustered scores, however b256_128 scale range is lower than b128_004. As in spectra plots, when comparing all PCA plots (Fig. S7 to S12), it is possible to visualize the detailed changes along the increase of scan number.

Although spectra and PCA plotting make possible to visualize the changes as sample scan number is increased from 004 to 128, it is hard to

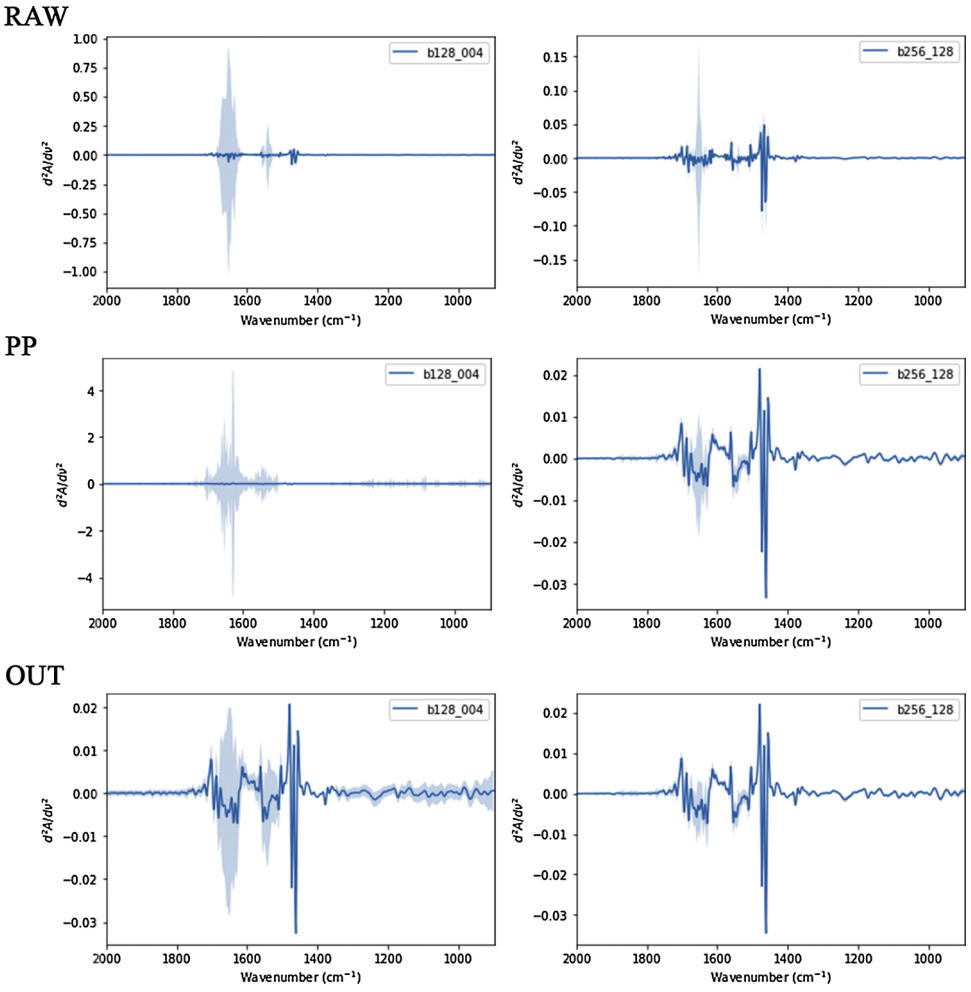

Fig. 1. Spectra comparison of b128_004 and b256_004 for groups RAW, PP, and OUT. Mean spectrum (solid line) and standard deviation (shades).

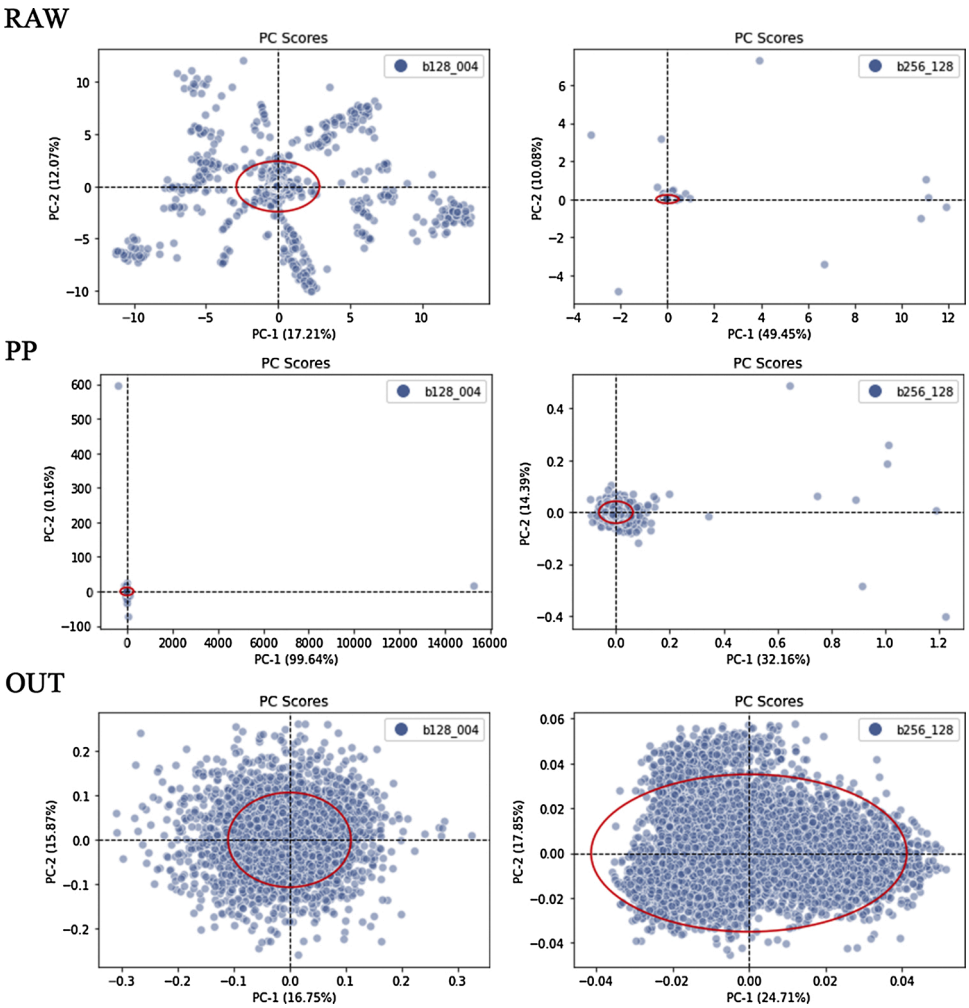

Fig. 2. Principal Component Analysis (PCA) of b128_004 and b256_004 for groups RAW, PP, and OUT. Dots denote PC-scores and red line the Hotelling's T^2 95 % confidence ellipse. PC-1 and -2 cumulative variance between parentheses.

perceive major changes among highest scans, such as 064 and 128, especially for OUT groups, where the preprocessing techniques, mostly the outlier removal, approximate lower and higher scan samples. It is even harder to distinguish between b128 and b256 for same sample scans. In this way, a classification quantitative analysis can give an additional perspective of these different scans and how they perform when modeled into machine learning techniques, which is the main objective for many cancer evaluations.

3.2. *Classification analysis*

The Fig. 3 presents the best model test accuracy boxplot, XGB, where

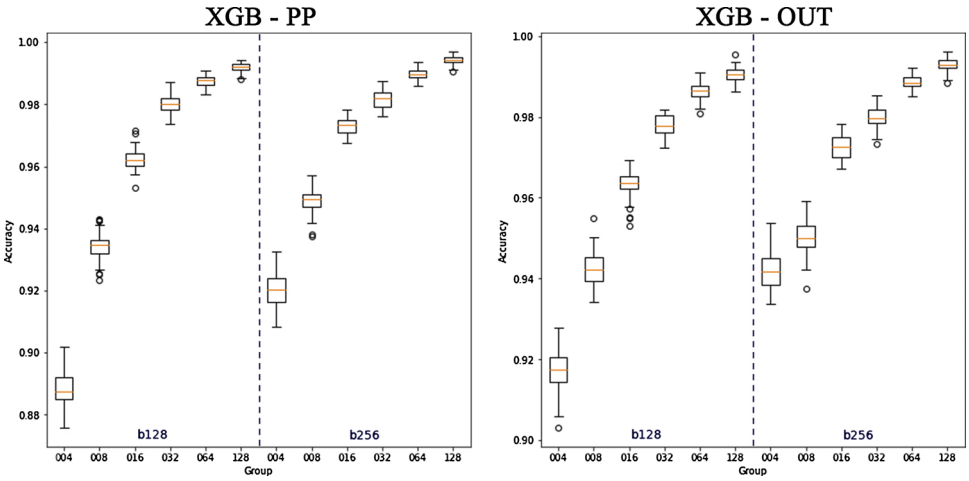

Fig. 3. Best model (XGB, Extreme Gradient Boosting) test accuracy boxplot by groups. Dashed blue line splits different background scans (b128 and b256).

the highest mean accuracy of 0.995 was achieved by OUT b256_128 group. It is possible to see an accuracy augmentation tendency in relation to higher sample scans, as well as for higher background scans, but in a less evident way. Lower sample scans also demonstrate higher accuracy SD, indicating a model generalization difficulty along the CV, possibly due to noisy data varying during the folds. All models' boxplots can be seen in Supp. Fig. S13 and S14.

RAW groups were excluded from this analysis as EMSC application is necessary. Most models are negatively affected when using non-normalized data, and all of them would use paraffin bands variation to misclassify the data. The fixed threshold for outlier removal in OUT groups assisted to not result in biased data, such as removing more data from lower scan numbers than from higher ones. On average, 10 % of the data were removed by the outlier algorithm.

Tree-based models, XGB and RF, presented similar PP and OUT groups metrics, endorsing their robustness to outliers [21]. Even though XGB tends to be more prone of being affected by outlier than RF, since each individual tree learns from the previous, its techniques, mainly the regularization [22], enabled to perform similar with PP and OUT. The others models exhibited better accuracies for OUT groups than for PP, specially the SVM model, once outliers affect its hyperplane separation and make it a fragile model for this case [23].

SVM and KNN models presented the lowest accuracies, that may be related to their difficulties when dealing with large features number [23, 24], while tree-based models apply subsampling techniques, and LDA and PLS employ dimensionality reduction. Using feature selection and extraction approaches, such as feeding the models with *n* PCs instead of the whole spectra, could help to overcome this issue, although it was not tested in this study for the sake of comparison. Additionally, to standardize the comparison, default hyperparameters were chosen for all the models to avoid better results due to improved optimization in one model in comparison to the others. Still, feasible hyperparameters were used in relation to the data, making a general evaluation of the models.

PCA noise reduction [8] was also tested, using *n* PCs until 99 % of explained variance was obtained, but no accuracy improvement was presented by the models, corroborating with the clustering analysis performed by Sacharz et al., 2020 [12], thus this result was omitted for simplicity.

The Friedman test indicated significant difference in all the models, for both PP and OUT, hence the Nemenyi test could be performed, where its results are shown in Fig. 4. It is possible to see a pattern of better critical values (closer to one) when increasing the number of sample scans for both backgrounds. All the models presented no significant statistical difference between b256_128, b256_64 and b128_128, except for SVM-PP model, which had its shortcomings already discussed and exhibited difference for b128_128. Therefore, these three scan groups demonstrated similar prediction performance, especially when applying the outlier removal algorithm.

While it is plausible in terms of accuracy to choose 128 sample scans with any of the backgrounds, b256 or b128, the b256_64 choice brings the additional of time optimization. In real clinical tasks, larger areas of the biopsies have to be evaluated, requiring mosaic acquisition to cover a region in order of centimeters. In this way, halving the sample scan, almost halves the time acquisition, improving the clinical applicability of the technique. As the background can be collected in a single scan, even if sample mosaics are performed, using b256 scans instead of b128 does not imply in a significant impact to acquisition time.

Luminal B and HER2 subtypes were chosen for this study to focus on ER/PR classification, as up to 80 % of breast cancer are ER positives and up to 65 % are PR positives, besides presenting the best treatment outcome when they are both positive and diagnosed in early stage [25]. Hence, a better evaluation between them may provide important prognostic and therapeutic information. In addition, this scan evaluation can be used as a basis for other studies using machine learning with cancer samples.

4. Conclusion

The analysis of co-added scans number for FTIR spectroscopy images demonstrated great impact on the acquired spectra, where higher sample scans decreased the standard deviation and the outlier impact. Preprocessing techniques can help to improve the quality of data,

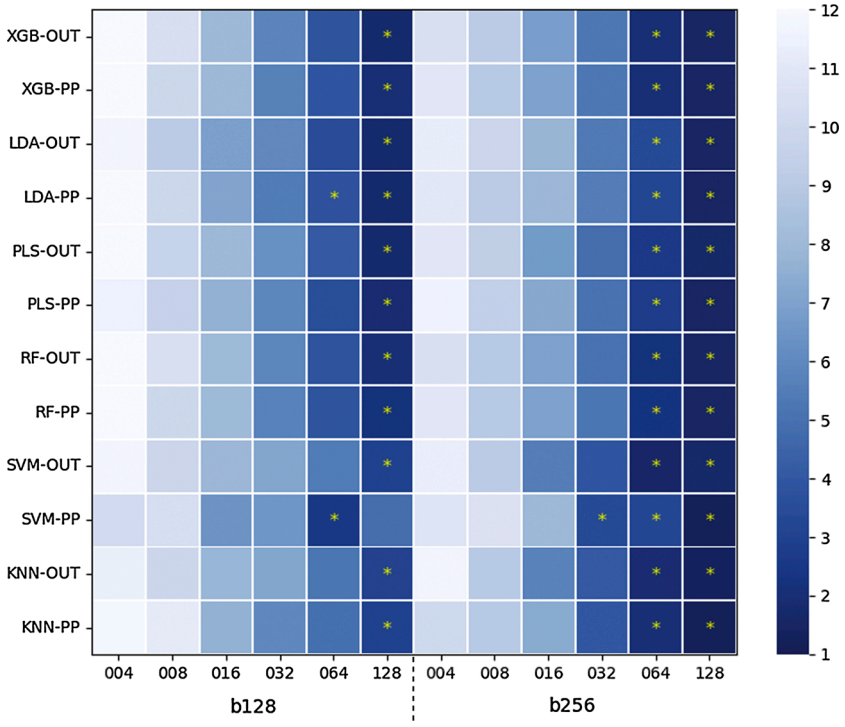

Fig. 4. Critical values heatmap for the Nemenyi test. Yellow stars denote the best tied accuracies within each model (scans where the critical distance presented no significant statistical difference).

approaching the characteristics of lower scan groups to higher ones. The 256 background and 064 sample scans group (b256_064) showed up as the best cost-benefit for machine learning classification tasks, presenting the best classifications together with b256_128 and b128_128, but with approximately half the acquisition time, thus a better clinical translational potential.

Improving the machine learning classification of breast cancer subtypes may lead to a better prognostic, where the ER/PR positive subtype, assessed in this work, present the best treatments responses among the subtypes when diagnosed in early stage. This work also suggests b256_064 as a basis for other studies using machine learning for cancer evaluation.

In order to test the code, measurements in other equipments should be evaluated, since the scans numbers performance may vary from different spectrometer manufactures. Other preprocessing steps, models' parameters optimization and different machine learning models may also be evaluated in other studies.

CRediT authorship contribution statement

Matheus del-Valle: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Moisés Oliveira dos Santos:** Conceptualization, Methodology, Validation, Investigation, Writing - review & editing. **Sofia Nascimento dos Santos:** Resources, Writing - review & editing. **Pedro Arthur Augusto de Castro:** Writing - review & editing. **Emerson Soares Bernardes:** Supervision. **Denise Maria Zezell:** Supervision, Funding acquisition.

Declaration of Competing Interest

The authors reported no declarations of interest.

Acknowledgements

This work was supported by FAPESP [17/50332-0], CAPES [Finance Code 001], CAPES/PROCAD [88881.068505/2014-01], and CNPq [INCT-465763/2014-6, PQ-309902/2017-7, 142229/2019-9, 141946/2018-0].

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.vibspec.2021.103309.

References

[1] A. Hennigs, F. Riedel, A. Gondos, P. Sinn, P. Schirmacher, F. Marmé, D. Jäger, H.-U. Kauczor, A. Stieber, K. Lindel, J. Debus, M. Golatta, F. Schütz, C. Sohn, J. Heil, A. Schneeweiss, Prognosis of breast cancer molecular subtypes in routine clinical care: a large prospective cohort study, BMC Cancer 16 (2016) 734, https://doi.org/10.1186/s12885-016-2766-3.

[2] N. Eliyatkin, E. Yalcin, B. Zengel, S. Aktaş, E. Vardar, Molecular classification of breast carcinoma: from traditional, old-fashioned way to a new age, and a new way, J. Breast Health 11 (2015) 59–66, https://doi.org/10.5152/tjbh.2015.1669.

[3] H.G. Russnes, O.C. Lingjærde, A.-L. Børresen-Dale, C. Caldas, Breast cancer molecular stratification, Am. J. Pathol. 187 (2017) 2152–2162, https://doi.org/10.1016/j.ajpath.2017.04.022.

[4] S. Kalmodia, S. Parameswaran, W. Yang, C.J. Barrow, S. Krishnakumar, Attenuated Total Reflectance Fourier transform Infrared Spectroscopy: an analytical technique to understand therapeutic responses at the molecular level, Sci. Rep. 5 (2015) 16649, https://doi.org/10.1038/srep16649.

[5] S. Kumar, A. Srinivasan, F. Nikolajeff, Role of infrared spectroscopy and imaging in cancer diagnosis, Curr. Med. Chem. 25 (2018) 1055–1072, https://doi.org/10.2174/0929867324666170523121314.

[6] K.-Y. Su, W.-L. Lee, Fourier transform infrared spectroscopy as a cancer screening and diagnostic tool: a review and prospects, Cancers (Basel) 12 (2020) 115, https://doi.org/10.3390/cancers12010115.

[7] R. Gautam, S. Vanga, F. Ariese, S. Umapathy, Review of multidimensional data processing approaches for Raman and infrared spectroscopy, EPJ Tech. Instrum. 2 (2015) 8, https://doi.org/10.1140/epjti/s40485-015-0018-6.

[8] M.J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H.J. Butler, K.M. Dorling, P. R. Fielden, S.W. Fogarty, N.J. Fullwood, K.A. Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G.D. Sockalingum, J. Sulé-Suso, R.J. Strong, M. J. Walsh, B.R. Wood, P. Gardner, F.L. Martin, Using Fourier transform IR spectroscopy to analyze biological materials, Nat. Protoc. 9 (2014) 1771–1791, https://doi.org/10.1038/nprot.2014.110.

[9] C.L.M. Morais, M. Paraskevaidi, L. Cui, N.J. Fullwood, M. Isabelle, K.M.G. Lima, P. L. Martin-Hirsch, H. Sreedhar, J. Trevisan, M.J. Walsh, D. Zhang, Y.-G. Zhu, F. L. Martin, Standardization of complex biologically derived spectrochemical datasets, Nat. Protoc. 14 (2019) 1546–1577, https://doi.org/10.1038/s41596-019-0150-x.

[10] C.L.M. Morais, K.M.G. Lima, M. Singh, F.L. Martin, Tutorial: multivariate classification for vibrational spectroscopy in biological samples, Nat. Protoc. 15 (2020) 2143–2162, https://doi.org/10.1038/s41596-020-0322-8.

[11] M. Tahtouh, P. Despland, R. Shimmon, J.R. Kalman, B.J. Reedy, The application of infrared chemical imaging to the detection and enhancement of latent fingerprints: method optimization and further findings, J. Forensic Sci. 52 (2007) 1089–1096, https://doi.org/10.1111/j.1556-4029.2007.00517.x.

[12] 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 (2020) 4334–4342, https://doi.org/10.1039/C9AY01200A.

[13] X. Dai, H. Cheng, Z. Bai, J. Li, Breast cancer cell line classification and its relevance with breast tumor subtyping, J. Cancer 8 (2017) 3131–3141, https://doi.org/10.7150/jca.18457.

[14] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, Chemometr. Intell. Lab. Syst. 117 (2012) 92–99, https://doi.org/10.1016/j.chemolab.2012.03.004.

[15] F.A. de Lima, C. Gobinet, G. Sockalingum, S.B. Garcia, M. Manfait, V. Untereiner, O. Piot, L. Bachmann, Digital de-waxing on FTIR images, Analyst 142 (2017) 1358–1370, https://doi.org/10.1039/C6AN01975G.

[16] Janez Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 17 (2006) 1–10.

[17] V. Balan, C.-T. Mihai, F.-D. Cojocaru, C.-M. Uritu, G. Dodi, D. Botezat, I. Gardikiotis, Vibrational spectroscopy fingerprinting in medicine: from molecular to clinical practice, Materials (Basel) 12 (2019) 2884, https://doi.org/10.3390/ma12182884.

[18] P. Meksiarun, P.H.B. 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 (2018) 3850–3858, https://doi.org/10.1039/C8AN00345A.

[19] T. Murayama, N. Gotoh, Patient-derived xenograft models of breast cancer and their application, Cells 8 (2019) 621, https://doi.org/10.3390/cells8060621.

[20] S.W. Bruun, A. Kohler, I. Adt, G.D. Sockalingum, M. Manfait, H. Martens, Correcting attenuated total reflection—Fourier transform infrared spectra for water vapor and carbon dioxide, Appl. Spectrosc. 60 (2006) 1029–1039, https://doi.org/10.1366/000370206778397371.

[21] V.J. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2004) 85–126, https://doi.org/10.1007/s10462-004-4304-y.

[22] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, 2016, https://doi.org/10.1145/2939672.2939785.

[23] T. Kanamori, S. Fujiwara, A. Takeda, Breakdown Point of Robust Support Vector Machine, 2014.

[24] M.M.R. Khan, R.B. Arif, M.A.B. Siddique, M.R. Oishe, Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository, 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT) (2018), https://doi.org/10.1109/CEEICT.2018.8628041.

[25] S.M. Fragomeni, A. Sciallis, J.S. Jeruss, Molecular subtypes and local-regional control of breast cancer, Surg. Oncol. Clin. N. Am. 27 (2018) 95–120, https://doi.org/10.1016/j.soc.2017.08.005.