Latin America Optics and Photonics (LAOP) Conference 2022 © Optica Publishing Group 2022

Breast Cancer Subtype Classification Using a One-Dimensional Convolutional Neural Network in Hyperspectral Images

Matheus del-Valle[1], Moises Oliveira dos Santos[1,2], Emerson Soares Bernardes[3], Denise Maria Zezell[1,*]

[1]*Center for Lasers and Applications, Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN, Av. Prof. Lineu Prestes, 2242, CEP: 05508-000, São Paulo-SP, Brazil*

[2]*Escola Superior de Tecnologia, Universidade do Estado do Amazonas, 69050-030, Brazil*

[3]*Centro de Radiofarmcacia, Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN, Av. Prof. Lineu Prestes, 2242, CEP: 05508-000, São Paulo-SP, Brazil*

**zezell@usp.br*

Abstract: FTIR spectroscopy imaging in addition to deep learning is a potential tool for breast cancer subtype classification, where accuracies higher than 86% can be achieved to predict among all subtypes.

1. Introduction

The breast cancer subtypes classification plays an import role in its treatment, helping to select an appropriate therapy [1]. Subtypes are defined using the expression levels of Ki67 biomarker and three hormone receptors: estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2). There are four breast cancer subtypes [2]: Luminal A (ER/PR+, HER2−, Ki67 low); Luminal B (ER/PR+, HER+/−, Ki67 high); HER2 subtype (ER−, PR−, HER2+, Ki67 usually high); Triple-Negative Breast Cancer − TNBC (ER−, PR−, HER2−, Ki67 usually high).

Subtypes are usually classified by immunohistochemistry techniques, where interobserver variations are present [3]. Fourier Transform Infrared (FTIR) spectroscopy imaging has been studied as a further cancer evaluation technique, providing additional information and helping to improve assessment quality [1, 4], particularly when applied machine learning approaches [5, 6]. In this way, this study aims to classify breast cancer subtypes using a deep learning model and FTIR hyperspectral images.

2. Material and Methods

It was used the human breast cancer microarray BR804b (Biomax, Inc., USA) in a calcium fluoride slide. In this study, eight cores were imaged, two of each subtype: Luminal A, LuminaL B, HER2, and TNBC. Image acquisition was performed by a Cary Series 600 micro-FTIR imaging system (Agilent Technologies, USA). About 100 thousand spectra were acquired for each sample.

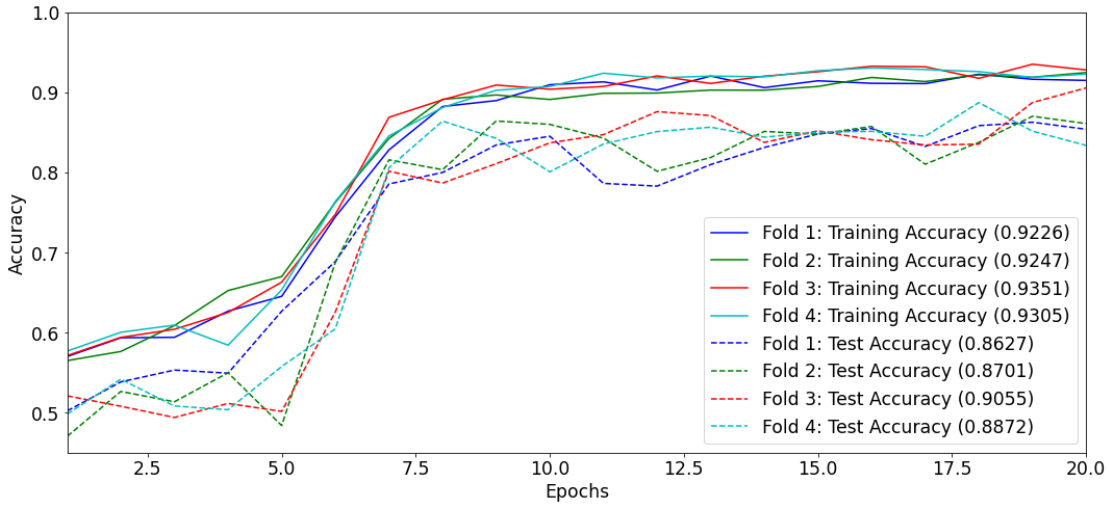
The regions of interest were selected by two k-means clustering with k = 2. The first one using the amide I/II (1700 to 1500 cm^{-1}), aiming to select tissue spectra; and the second one selecting highest paraffin intensity (1480 to 1450 cm^{-1}) bands, identifying paraffin spectra. Data were preprocessed by five steps: outlier removal using Hotelling's T2 versus Q residuals; biofingerprint truncation; Savitzky–Golay filtering for smoothing and second derivative; Extended multiplicative signal correction (EMSC) with digital de-waxing; another outlier removal.

A convolutional neural network (CNN) was modeled to classify the data. The feature extraction part was built with two Conv1D-ReLU-MaxPooling1D-Dropout layers. The kernel size was set to 5 and dropout rate of 0.5. The classification part was built by two layers of neurons-BatchNorm-ReLU-Dropout, with 100 and 50 neurons, and dropout of 0.2. The output was a 4-neuron layer with softmax activation. Categorical cross-entropy loss function was adopted with Adam optimizer. Accuracy metric was calculated during the training, where a threshold of 0.5 was applied on the output predictions. Model was trained by a 4-fold cross-validation by 20 epochs and using a batch size of 250.

3. Results and Discussion

The two steps K-Means enabled spectra identification by clustering tissue spectra, paraffin spectra (regions with tissue absence and margins of the core section), and pure slide spectra, if present. Thus, it was possible to apply the digital de-waxing and to use only tissue related spectra in the training of the model.

The train accuracy was 0.9282 ± 0.0049 (Fig. 1), while the testing accuracy was 0.8814 ± 0.0165, demonstrating that the model was able to predict among the four subtypes available. Training and testing accuracies close to each other indicates that the model was able to prevent overfitting, being generalized to the input data. Hyper-parameters and architecture should be optimized along a greater samples number to increase the performance of the model. Training loops callbacks should also be tested during longer training epochs, such as learning rate monitoring.



4. Conclusion

The deep learning model was able to predict the breast cancer subtype with test accuracies higher than 85%, indicating the proposed model as a potential technique for the breast cancer subtype classification, where model optimization and a larger dataset should be tested.

5. Acknowledgements

This work was supported by FAPESP [21/00633-0, 17/50332-0], CNPq [INCT-465763/2014-6, PQ-31457/2021-9], [PhD-grant- 142229/2019-9], and CAPES [Finance Code 001].

6. References

1. S. Kalmodia, S. Parameswaran, W. Yang, C. J. Barrow, and S. Krishnakumar, “Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy: An analytical technique to understand therapeutic responses at the molecular level,” Sci. Reports 5, 16649 (2015).
2. Z.KosandD.J.Dabbs,“Biomarkerassessmentandmoleculartestingforprognosticationinbreastcancer,”Histopathol- ogy 68, 70–85 (2016).
3. H. G. Russnes, O. C. Lingjærde, A.-L. Børresen-Dale, and C. Caldas, “Breast Cancer Molecular Stratification,” The Am. J. Pathol. 187, 2152–2162 (2017).
4. S. Kumar, A. Srinivasan, and F. Nikolajeff, “Role of Infrared Spectroscopy and Imaging in Cancer Diagnosis,” Curr. Medicinal Chem. 25, 1055–1072 (2018).
5. K.-Y. Su and W.-L. Lee, “Fourier Transform Infrared Spectroscopy as a Cancer Screening and Diagnostic Tool: A Review and Prospects,” Cancers 12, 115 (2020).
6. R.Gautam,S.Vanga,F.Ariese,andS.Umapathy,“Reviewofmultidimensionaldataprocessingapproachesfor Raman and infrared spectroscopy,” EPJ Tech. Instrumentation 2, 8 (2015).