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

Identifying Breast Cancer Cell Lines Using High Performance Machine Learning Methods

Sajid Farooq,[1] Matheus Del-Valle[1] Sofia Santos,[2] Emerson Soares Bernandes [2] and Denise Maria Zezell,[1,*]

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

[2] *Center for Radiopharmaceutics, Instituto de Pesquisas Energéticas e Nucleares, Av. Prof. Lineu Prestes 2242, São Paulo, 05508-000, Brazil*

[*] *zezell@usp.br*

Abstract: We present a computational framework based on machine learning classifiers K-Nearest Neighbors and Neighborhood Component analysis for breast cancer (BC) subtypes prognostic. Our results has up to 97% accuracy for prognostic stratification of BC subtypes.

1. Introduction

Breast cancer (BC) is the leading cause of cancer-related fatality among women worldwide [1]. Because of its complex tumour micro-environment, this disease remains a huge challenge of informed therapeutic decision making [2]. Thereby, the BC classification based on molecular cell lines owns an important role during treatment, helping not only to select a specific therapy but also sorting patients with prognostication. Moreover, these BC cell lines are evolved based on the expression of receptors, and can be further categorized into distinct molecular subtypes such as luminal A, luminal B, HER2, and triple-negative [3]. Given the erratic classification, even inconsistent molecular characterization and nomenclatures, we are deluged with BC cell lines nonexistent organized characteristics documentation and consistent molecular subtypes [4].

To overcome such hurdles, we propose to develop a ML model for BC molecular cell lines prognostication. A ML algorithms is developed to make a computational pipe-line based on Neighborhood Component Analysis (NCA) couples with K-Nearest Neighbors (KNN) [5]. The ability to achieve BC prognostication successfully with NCA-KNN method evidences that de novo learning method is an important factor to classify the BC subtypes.

2. Methods and Materials

We selected BC cell lines such as BT-474(ATCC N0.: HTB-20), a luminal B (ER/PR/HER2 positive), and SKBR-3 (ATCC NO.: HTB-30), a HER2 (ER/PR negative and HER2 positive). For input data, we used a Fourier transform Infrared Spectroscopy (FTIR) for spectral image acquisitions. In order to obtain computational analysis, data was preprocessed through Savitzky-Golay(SG) and then, extended multiplicative signal correction (EMSC) for normalization purpose as well as to perform a digital de-waxing [6]. To apply NCA-KNN method, two BC subtypes were selected i.e. SKBR3 and BT474 and NCA-KNN method was used with cross-validation (k fold = 10). The receiver operating characteristics (ROC) curve was used to estimate the model performance. For comparison purpose, several individual classifiers such as Support Vector Machine (SVM), AdaBoost and KNN classifiers were introduced.

Table 1. The performance of different methods to achieve accuracy.

Computational ML methods	Accuracy
$NCA - KNN$	97.5
SVM	88.3
KNN	78.1
AdaBoost	76.7

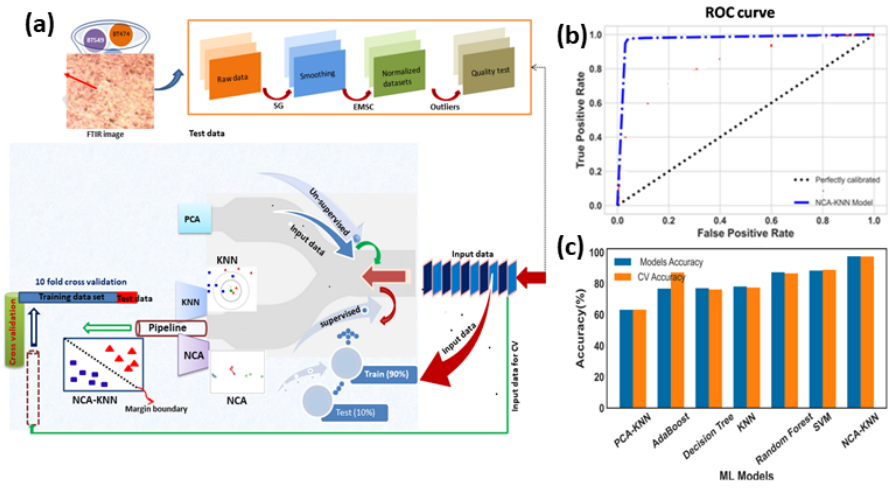

Fig. 1. (a)A workflow diagram of data for computational analysis based on supervised and unsupervised learning.(b) The receiver operating characteristic curve (ROC) and (c) evaluation of accuracy using different classifiers.

3. Results and Discussion

To generalize our BC cell lines findings and assist translation of the outcomes towards clinical implications, we cleaned the outliers using pre-processing procedure such as 3113 outliers (20.6 % of the total spectra) were removed due to bad quality, obtaining by a Hotelling T^2 vs. Q residuals. To apply NCA-KNN method to data sets that contain two classes we employed ROC curve to determine the prognostic stratification of BC cell lines. performance of the model is depicted by the area under curve (AUC), as shown in Fig. 1b. The obtained accuracy ROC curve by NCA-KNN method is $\sim$ 97.5%. Fig. 1c shows the comparison of our model to the other individual methods. The accuracy obtained by the proposed model is 97.5%, which is higher than all other single methods, as presented in table 1.

4. Conclusion

We proposed a machine learning model based on two superivsed classifiers to predict the breast cancer cell lines in stead of traditional models using single classifier. Our model present the performance upto 97% higher, which is around 10% higher as compared to the second best SVM method, showing a potential for the breast cancer subtypes prognostication for clinical uptake.

5. Acknowledgements

This work was supported by FAPESP [17/50332-0] and CNPq [INCT-465763/2014-6, PQ-31451712021-9, 142229/2019-9].

References

1. Amiri Souri, E., Alicia Chenoweth, Anthony Cheung, Sophia N. Karagiannis, and Sophia Tsoka, "Cancer Grade Model: a multi-gene machine learning-based risk classification for improving prognosis in breast cancer," British Journal of Cancer **125**, 748–758, no. 5 (2021).
2. Sammut et al., "Multi-omic machine learning predictor of breast cancer therapy response," Nature 601 **601**, 623–629, (2022).
3. Dai, Xiaofeng, Hongye Cheng, Zhonghu Bai, and Jia Li, "Breast cancer cell line classification and its relevance with breast tumor subtyping," Journal of Cancer **8**, no. 16, 3131–(2017)
4. Horr, Christina, and Steven A. Buechler., "Breast Cancer Consensus Subtypes: A system for subtyping breast cancer tumors based on gene expression," NPJ breast cancer **7**, 1-13 (2021)
5. Goldberger, Jacob, Geoffrey E. Hinton, Sam Roweis, and Russ R. Salakhutdinov, "Neighbourhood components analysis," Advances in neural information processing systems **17**, (2004)
6. del-Valle, et al., "The impact of scan number and its preprocessing in micro-FTIR imaging when applying machine learning for breast cancer subtypes classification," Vibrational Spectroscopy **117**,103309– (2021)