Superior Machine Learning Method for breast cancer cell lines identification

Sajid Farooq
Center for Lasers and Applications(CLA)
Instituto de Pesquisas Energéticas e Nucleares (IPEN)
Sao Paulo, Brazil
sajiddahar@gmail.com

Amanda Caramel-Juvino
Center for Lasers and Applications(CLA)
Instituto de Pesquisas Energéticas e Nucleares (IPEN)
Sao Paulo, Brazil
caramel@usp.br

Matheus Del-Valle
Center for Lasers and Applications(CLA)
Instituto de Pesquisas Energéticas e Nucleares (IPEN)
Sao Paulo, Brazil
matheus.valle@usp.br

Sofia Santos
Center for Radiopharmaceutics
Instituto de Pesquisas Energéticas e Nucleares (IPEN)
Sao Paulo, Brazil
snsantos85@gmail.com

Emerson Soares Bernandes
Center for Radiopharmaceutics
Instituto de Pesquisas Energéticas e Nucleares(IPEN)
Sao Paulo, Brazil
emerson.bernardes@gmail.com

Denise Maria Zezell
Center for Lasers and Applications(CLA)
Instituto de Pesquisas Energéticas e Nucleares (IPEN)
Sao Paulo, Brazil
zezell@usp.br

2022 SBFoton International Optics and Photonics Conference (SBFoton IOPC) | 978-1-6654-5273-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/SBFotonIOPC54450.2022.9992467

***Abstract*—We propose an artificial intelligence platform based on machine learning (ML) algorithm using Neighborhood Component analysis and K-Nearest Neighbors for breast cancer cell lines recognition. Our model presents up to 97% accuracy for identification of breast cancer cell lines.**

***Index Terms*—Machine learning, Cell lines, FTIR, Accuracy, Breast Cancer**

I. Introduction

Breast cancer (BC) is the leading cause of cancer-related fatality globally among women [1]. Due to complex tumour micro-environment system, this cancer is a big challenge of clinical decision making procedures [2]. Therefore, the breast cancer stratification based on molecular cell lines plays a remarkable role during treatment. Moreover, appropriate classification also assists not only to choose an accurate therapy methodology but also sort patients with prognostic diagnosis [5]. Further, breast cancer cell lines are up to 92 and are evolved based on the expression of receptors. Although, these cell lines can be classified into distinct subtypes: as describe by luminal A, luminal B, HER2, and triple-negative [3]. Using the erratic stratification of these cell lines, we are overwhelmed with BC cell lines nonexistent with a lot of characteristics documentation and consistent BC cell lines [5].

Therefore, models of identifying BC cell lines are urgently required to avoid the under-diagnosis of tumours before metastasis and to reduce the over-treatment of low risk disease, that could assist to reduce the demand of aggressive systemic therapy.

To cope with the challenge, we aim to develop a ML algorithm for the recognition the BC cell lines accurately and rapidly. A ML algorithm is designed to use a semi-supervised and supervised classifiers i.e. Neighborhood Component Analysis (NCA) and K-Nearest Neighbors (KNN) [6]. The performance to achieve accurate prognostication perfectly with NCA-KNN algorithm that evidences de novo learning algorithm, is an important factor to identifiy BC cell lines.

II. Methods and Materials

We choose breast cancer cell lines such as BT-474 i.e. a luminal B (ER/PR/HER2 positive), and SKBR-3, a HER2 (ER/PR negative and HER2 positive). To compute the model, we obtained an input data using Fourier transform Infrared Spectroscopy (FTIR). The data is collected in the shape of spectral image acquisitions and number of spectrums are $>$10k. The aforementioned data is preprocessed using Savitzky-Golay(SG) and then, extended multiplicative signal correction (EMSC). Later on, it is normalized to perform a digital de-waxing [7]. To apply NCA-KNN method on the BC cell lines, the NCA-KNN method is employed with cross-validation (k fold = 10). The receiver operating characteristics (ROC) curve is explored to evaluate accuracy. Others different individual classifiers were used as Support Vector Machine (SVM), AdaBoost and KNN for comparison purpose.

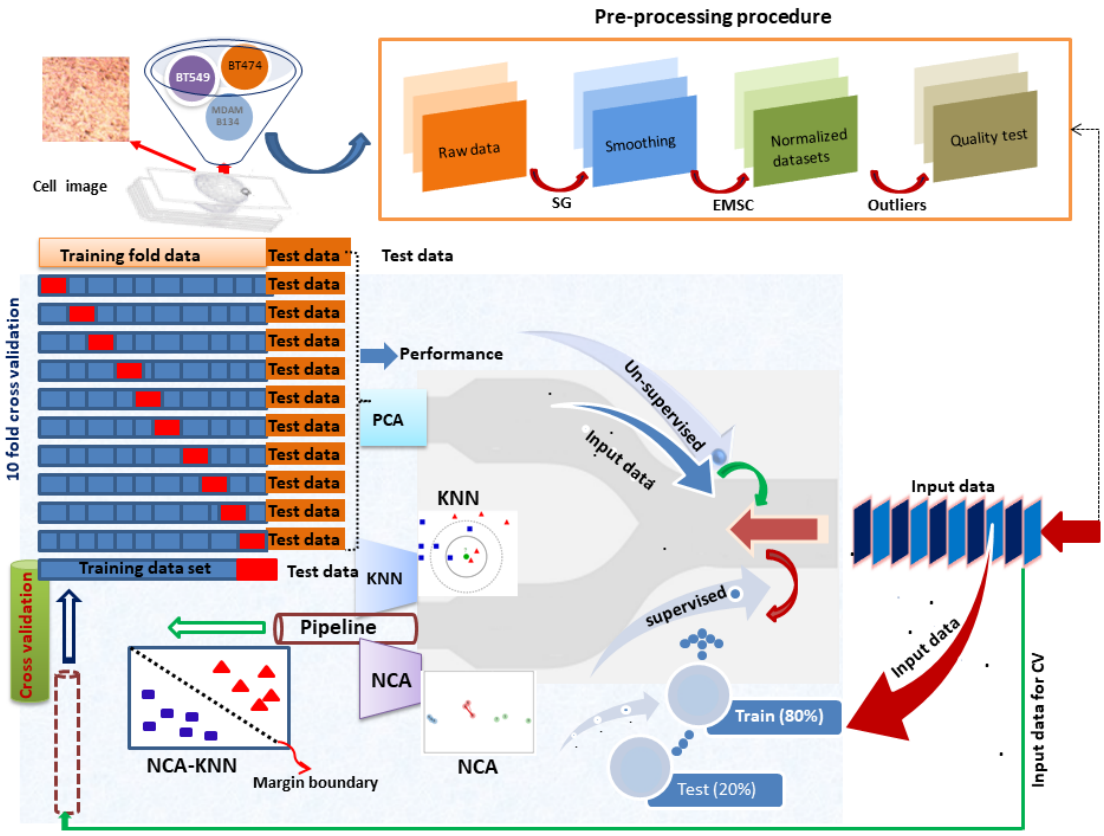

Fig. 1. A schematic diagram for computational modeling framework based on ML learning.

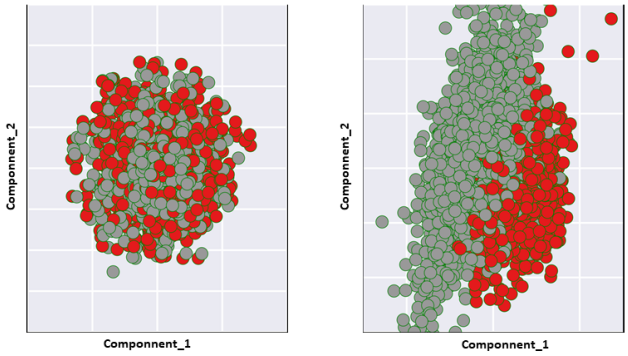

Fig. 2. A comparison of unsupervised (PCA) versus supervised ML algortihms (KNN−NCA).

TABLE I
THE PERFORMANCE EVALUATION USING SEVERAL KIND OF CLASSIFIERS TO COMPARE ACCURACY.

Computational ML methods	Accuracy
$NCA-KNN$	97.5
KNN	78.10
AdaBoost	76.70
SVM	88.30

III. Results and Discussion

In order to explore BC cell lines prognostication and help to translate the outcomes towards therapy decision, we removed unwanted outliers employing pre-processing procedure. For instance, 20.6 % outliers were removed because of bad quality. This procedure was achieved by a Hotelling T^2 versus Q residuals. Figure 1 presents the preprocess procedure for computational simulation analysis based on ML algorithm. Principal component analysis (PCA) shows the BC cell lines in blue and orange color (Fig.2-left). Using ML algorithm, the data is well separated and can be identified easily, as shown in Fig.2 (right). Further, NCA-KNN algorithm was performed to the pre-process data and the whole data is divided into train set and test set. Later on, ROC curve is used to identify the prognostic stratification of BC cell lines (Fig.3a). Using NCA-KNN based ROC cure, the accuracy is up to 97%. The performance based comparison of an individual classifiers to the coupled classifiers (KNN-NCA), as shown in Fig.3b. Table 1. shows that our model is up to 9% higher in comparison to the individual classifier like support vector machine and other classifiers. The obtained accuracy ROC curve by NCA-KNN method is $\sim 97.5\%$, as shown in Fig 3b. The accuracy obtained by the proposed model is 97.5%, which is higher than all other single methods, as presented in Fig.3b.

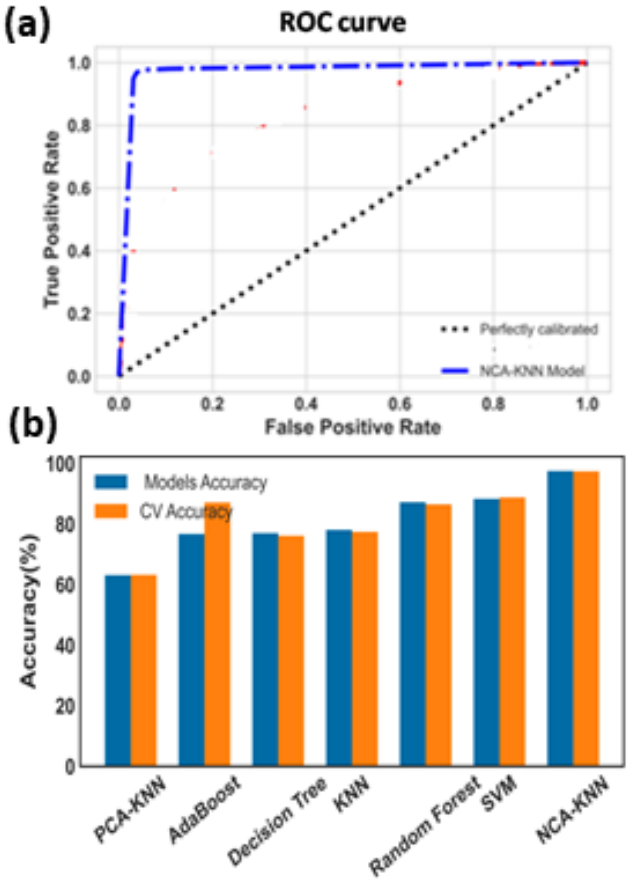

Fig. 3. The receiver operating characterisitics (ROC) curve to identify the accuracy (a) and different model performance (b) (KNN−NCA).

IV. Conclusion

We presented a ML algorithm employing two supervised classifiers to evaluate the stratification of breast cancer cell lines. Our proposed method shows the higher performance up to 97% , which is around 9% higher than that of second best SVM method, presenting a hidden potential for the cancer cell lines identification with superior performance for clinical uptake.

Acknowledgment

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

V. Conflict of Interest

The authors declare no conflict of interest.

References

[1] Sammut et al., "Multi-omic machine learning predictor of breast cancer therapy response," Nature 601 **601**, 623–629, (2022).

[2] 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).

[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 (2021): 1-13.

[5] 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)

[6] Goldberger, Jacob, Geoffrey E. Hinton, Sam Roweis, and Russ R. Salakhutdinov, "Neighbourhood components analysis," Advances in neural information processing systems **17**, (2004)

[7] 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)