Identification of basal cell carcinoma skin cancer using FTIR and Machine learning

Daniella Lúmara Peres
Instituto de Pesquisas Energéticas e Nucleares (IPEN)
São Paulo, Brasil
daniellalumara@usp.br

Sajid Farooq
Instituto de Pesquisas Energéticas e Nucleares (IPEN)
São Paulo, Brasil
sajid.f@ipen.br

Rocío Raffaeli
Universidad Nacional de La Plata
La Plata, Argentina
rocioraffaeli@gmail.com

Maria Virginia Croce
Universidad Nacional de La Plata
La Plata, Argentina
mariavirginiacroce@gmail.com

Adela E. Croce
Universidad Nacional de La Plata
La Plata, Argentina
acroce@inifta.unlp.edu.ar

Denise Maria Zezell
Instituto de Pesquisas Energéticas e Nucleares (IPEN)
São Paulo, Brasil
zezell@usp.br

***Abstract*—Here we applied ATR-FTIR spectroscopy combined with computational modeling based on 3D-discriminant analysis (3D-PCA-QDA). Our results present an exceptional performance of 3D-discriminant algorithms to diagnose BCC skin cancer, indicating the accuracy up to 99%.**

***Index Terms*—ATR-FTIR, accuracy, BCC, discriminant analysis**

I. Introduction

Non melanoma skin cancer (NMSC) accounts for the majority of all cases of malignant skin neoplasms, with basal cell carcinoma (BCC) being the most common type (85.2%) [1]. BCC is typically characterized by slow growth, low metastatic potential, and is primarily caused by exposure to UV radiation [2]. Hence, early detection and treatment of BCC is important to prevent the cancer from spreading and causing more serious health problems.

Fourier transform infrared (FTIR) spectroscopy is a potent and sensitive method, identifying cancer diseases to revolutionize personalized treatment management [3]. Due to minimal sample preparation and rapid experimental analysis FTIR is a potential candidate to measure the biological samples analysis [4]. Albeit, the methods of measuring samples using FTIR can be time-consuming and computationally demanding, the use of ATR-FTIR spectroscopy could lead to faster and more accurate BCC diagnosis, thereby improving clinical outcomes [5]. With the continual improvement of machine learning (ML) techniques, the accuracy of ATR-FTIR spectroscopy in disease detection is expected to increase further [6].

In this study, we examine BCC and healthy tissues using ATR-FTIR spectroscopy combined with 3D discriminant analysis algorithm. This method may assist in finding data patterns to identify different cancer disease.

II. Materials and methods

The study was performed using 4 samples of basal cell carcinoma, and 4 samples of healthy skin as a control, with all specimens embedded in paraffin. The ATR-FTIR technique was employed to obtain spectral data between frequencies of 4000–$400cm^{-1}$ using a diamond crystal ATR accessory coupled to a Thermo Nicolet 6700 FTIR system. The spectra obtained were recorded with a spectral resolution of $4cm^{-1}$ and with 100 scans by spectrum.

Before data analysis, ATR-FTIR spectra were smoothed by the Savitzky-Golay filter (window =11, poly = 2^{nd} order), later baseline corrected and vector normalized as pre-processing. Dispersion correction was obtained using the extended multiple dispersion correction (EMSC) algorithm. The input dataset to the computational modeling framework was the fingerprint region of $1800 - 900cm^{-1}$ with a step of 4 cm^{-1}. The preprocessing and analysis steps were made using Python 3.0.

A. Computational modeling framework

The 3D-PCA procedure uses a regular PCA approach to decompose each point of the HsI surface, with the nonlinear iterative partial least squares (NIPLAS) algorithm. We applied 3D-principal component analysis, given as:

$$X^{*}_{lm} = T_{lm}P^{T}_{lm} + E_{lm} \qquad (1)$$

where the temporary spectral matrix at position (l, m) are represented by T_{lm} and P_{lm}, respectively.

To evaluate the 3D-discriminant analysis approach, termed 3D-PCA-QDA (Q_{ij}). The calculation of scores for those is as follow [5]:

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

979-8-3503-0402-2/23/$31.00 ©2023 IEEE

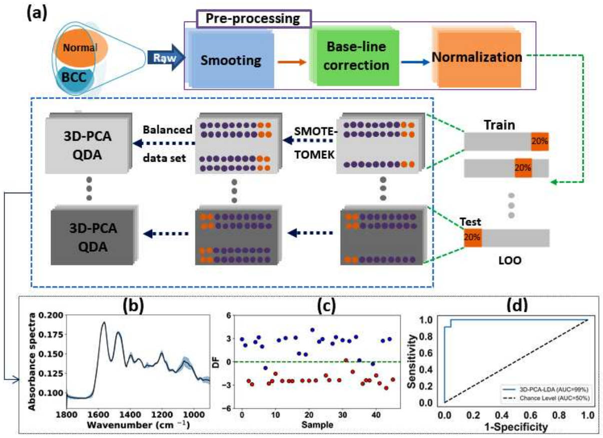

Fig. 1. A schematic diagram for computational modeling framework based on ML learning. (a) Flow diagram, (b) pre-processed data, (c)Discriminant Function (DF) and (d) receiver operating characteristics (ROC) curve with accuracy using area under the curve (AUC) of 99%.

where, the variables x_i and x_j are the mean-scores of T for sample *i* and the mean-scores of class *j* for their respective PCs and each is a $1 \times N$ row-vector. The variables C_{pooled}^{-1} and C_j represent the pooled co-variance matrix and the variance-covariance matrix of class *j*, respectively.

III. Results and Discuss

ATR-FTIR is an valuable analytical technique, it can be employed as a powerful tool to investigate and evaluate the accuracy when identifying the BCC and healthy groups. The dataset comprises a total of 44 samples, with 15 samples representing healthy skin, and 29 samples from BCC tumor tissue. In order to evaluate our model performance, we applied the Savitzky-Golay (SG) filter, to the fingerprint region ranging from 900 to 1800 cm^{-1} for smoothing, base line correction and normalization as shown in Fig.1a. Later on, with SMOTE-TOMEK technique, the unbalanced data were transformed in balanced data in order to achieve a higher precision. The pre-process spectral data were divided in two training set (80%) and test set (20%). In order to confirm validation of our model, we used Leave One Out (LOO) method in this present research.

Fig.1b shows the average spectra resulting from the pre-processing where the standard deviation (SD) indicates the variation in the mean spectra. Fig.1c depicts DF versus number of samples, indicating optimal separation between normal and BCC samples. The discriminant function is based on the score plot of 3D-PCA-QDA using the pre-processed data set as input data for computational modeling simulations.

The accuracy can be calculated by the AUC of the ROC curve, achieving from the receiver operating characteristics of two samples. Our results evidence that the model performance is highly precise with optimal accuracy up to 99% . Morais et al. studied ovarian cancer using 3D discriminant analysis approach to achieve accuracy 83% [7].

IV. Conclusion

In this work, we applied a computational approach based on 3D-discriminant analysis combined with ATR-FTIR spectroscopy to differentiate BCC from healthy tissues. With an accuracy of up to 99%, the discriminant analysis algorithms indicate the potential performance in comparison with unfolded algorithms, evidencing an outstanding performance of 3D-discriminant algorithms to identify cancer disease with ATR-FTIR data.

Acknowledgment

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

References

[1] Ciażyńska, Magdalena, Kamińska-Winciorek, Grażyna, Lange, Dariusz, Lewandowski, Bogumił, Reich, Adam, Sławińska, Martyna, Pabianek, Mart, Szczepaniak, Katarzyna, Hankiewicz, Adam, Ułańska, Małgorzata and others, "The incidence and clinical analysis of non-melanoma skin cancer". Scientific reports, vol. 11, 2021.

[2] Naik, Piyu Parth and Desai, Munaf B, "Basal cell carcinoma: a narrative review on contemporary diagnosis and management". Oncology and Therapy, p. 1-19, 2022.

[3] Amjad, M and Ullah, H and Andleeb, F and Batool, Z and Nazir, A and Gilanie, G, "Fourier-Transform Infrared Spectroscopy (FTIR) for Investigation of Human Carcinoma and Leukaemia". Lasers in Engineering (Old City Publishing), vol. 51, 2021.

[4] Farooq, Sajid and Del-Valle, Matheus and dos Santos, Moises Oliveira and dos Santos, Sofia Nascimento and Bernardes, Emerson Soares and Zezell, Denise Maria, "Rapid identification of breast cancer subtypes using micro-FTIR and machine learning methods", Applied Optics, vol. 62, pp. C80–C87, 2023.

[5] Farooq, Sajid and Del-Valle, Matheus and Santos, Sofia and Bernandes, Emerson Soares and Zezell, Denise Maria, "Identifying Breast Cancer Cell Lines Using High Performance Machine Learning Methods" Latin America Optics and Photonics Conference, pp. Tu5A–3, 2022.

[6] Morais, Camilo LM and Lima, Kássio MG and Singh, Maneesh and Martin, Francis L, "Tutorial: multivariate classification for vibrational spectroscopy in biological samples," Nature Protocols, vol. 15, pp. 2143–2162, 2020.

[7] Morais, C. L., Giamougiannis, P., Grabowska, R., Wood, N. J., Martin-Hirsch, P. L., and Martin, F. L. (2020). A three-dimensional discriminant analysis approach for hyperspectral images. Analyst, 145(17), 5915-5924.