POSTER 28

Classification of oral cancer using Random Forest: diagnosis by machinelearning

Daniella Lúmara Peres[1], Daniela F. T. Silva[1], Gleice Germano[1], Luciano Bachmann[3], Leandro L. de Matos[4], Joaquim C. Felipe[3], Thiago Martini Pereira[2], Denise Maria Zezell[1]*

[1]Nuclear and Energy Research Institute, IPEN-CNEN, São Paulo-SP - Brazil

[2]Federal University of São Paulo, UNIFESP, São José dos Campos-SP - Brazil

[3]University of São Paulo at Ribeirão Preto, USP – Ribeirão preto-SP, Brazil

[4]Cancer Institute of São Paulo State, ICESP, São Paulo-SP - Brazil

Introduction: Throat and neck cancer represent approximately 2% of cancer cases around the world, 90% of these cases are squamous cell carcinoma of the oral cavity, which is more easily treated when detected at an early stage. For this reason, a highly sensitive method capable of identifying changes in tissue composition that are invisible to the naked eye is crucial for a favorable prognosis and FTIR (Fourier Transform Infrared) spectroscopy serves this purpose effectively. With this technique, it is possible to obtain a hyperspectral image of the sample, which means one spectrum of biochemical information per pixel, so classifying this data with machine learning algorithms, is an interesting approach.

Methods: This study was approved by the Institutional Review Board under protocol number 228/14 (CAAE 32884214.5.0000.0065). 24 samples from patients from Cancer Institute of São Paulo State of oral squamous cell carcinoma and 24 samples of healthy tissue were used in this work, previously were examined by a Pathologist, which was used as the golden standard diagnosis. Random Forest method was used to classify human oral cavity squamous cell carcinoma images, labeled as ‘cancer’ and healthy ones labeled as ‘healthy’. The samples hyperspectral images were preprocessed using spectral smoothing with a Savitzky-Golay filter with an 11-point window, extended multiplicative signal correction, normalized, and had a quality test conducted, followed by training the model with 100 decision trees. Machine learning methods typically require a huge data set for accurate predictions, which is a challenge when dealing with biological human samples. To address this, a traditional machine learning method that requires a 2-dimensional dataset was employed. Since hyperspectral images are 3-dimensional, it is necessary to dissociate each pixel (each spectrum) of the image and label it as a sample by itself, causing the method to classify each spectrum individually, resulting in thousands of samples per image. By dissociating the pixels in this way, extra care is needed to avoid overfitting. This precaution was taken by keeping all spectra of an image always in the same group, so one image will not be used for both testing and training the algorithm. Additionally, the "leave one out" method was employed: With the 48 available samples, eight of them were successively designated as the test sample while the remaining 40 were used for training. The test and control groups were balanced, with the same number of samples in both groups, so this was not an issue. The tested images were evaluated one by one, with their own group of pixels. For this reason, we have an individual accuracy for each image.

Result: Using this method, each image is classified individually, which gives us an accuracy per image ranging between 0.98953 and 0.99993, with all images being correctly classified. This demonstrates an excellent ability to make correct predictions.

Discussion: As all images were correctly classified, we can conclude that the random forest method is highly recommended for this type of analysis. The model has proven to be highly accurate, sensitive, precise, and specific. Additional samples are being collected so that more tests can be conducted in the future.

Acknowledgement: 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 (21/00633-0).