Oral Cancer Diagnosis *Via* Machine Learning and Micro-FTIR Hyperspectral Imaging

D L Peres[1], D F T Silva[1], G C M. Germano[1], T M Pereira[2], J C Felipe[3], and D M Zezell[1]

[1]*Centro de Lasers e Aplicações, Instituto de Pesquisas Energéticas e Nucleares, Av. Prof. Lineu Prestes, 2272- Sao Paulo- Sao Paulo- Brazil, São Paulo, Brazil. Contact Phone: +551128105667*
[2]*Instituto de Ciências Tecnologia, Universidade Federal de São Paulo, Rua Talim, n°330- São José dos Campos - São Paulo - CEP: 12231-280, Brazil, São José dos Campos, Brazil. Contact Phone: +551233099621*
[3]*Computação e Matematica, Universidade de São Paulo, Av. Bandeirantes, 3900 - CEP 14040-901 Bairro Monte Alegre - Ribeirão Preto - SP -Brazil, Ribeirão Preto, Brazil. Contact Phone: +551633150598*
Contact Email: zezell@usp.br

Throat and neck cancer account for approximately 2% of cancer cases globally, with 90% of these being squamous cell carcinoma of the oral cavity, which is more treatable when detected early. Therefore, a highly sensitive method capable of identifying subtle changes in tissue composition is crucial for a favorable prognosis. FTIR (Fourier Transform Infrared) spectroscopy effectively serves this purpose by providing a hyperspectral image of the sample, capturing a spectrum of biochemical information for each pixel. Classifying this data using machine learning algorithms is a promising approach.

This study was approved by the Institutional Review Board under protocol number 228/14 (CAAE 32884214.5.0000.0065). We utilized 24 samples of oral squamous cell carcinoma and 24 samples of healthy tissue from patients at the Cancer Institute of São Paulo State, which were previously examined by a pathologist and served as the gold standard for diagnosis. The Random Forest method was employed to classify the hyperspectral images of human oral cavity squamous cell carcinoma as 'cancer' or 'healthy'. The hyperspectral images were preprocessed using spectral smoothing with a Savitzky-Golay filter with an 11-point window, extended multiplicative signal correction, normalization, and quality testing. The model was trained with 100 decision trees.

Machine learning methods typically require extensive datasets for accurate predictions, which is challenging with biological human samples. To address this, a traditional machine learning method that uses a 2-dimensional dataset was employed. Given that hyperspectral images are 3-dimensional, each pixel (each spectrum) was dissociated and labeled individually, resulting in thousands of samples per image. To avoid overfitting, all spectra from a single image were kept in the same group, ensuring that no image was used for both training and testing. The "leave one out" method was applied: eight of the 48 available samples were successively designated as the test sample, while the remaining 40 were used for training. The test and control groups were balanced with an equal number of samples in each. Each tested image was evaluated individually, yielding an accuracy per image.

This method resulted in each image being classified individually, achieving an accuracy per image ranging between 0.98953 and 0.99993, with all images correctly classified. This demonstrates excellent predictive capability.

The successful classification of all images indicates that the Random Forest method is highly effective for this type of analysis. The model has proven to be highly accurate, sensitive, precise, and specific. Additional samples are being collected for further testing in the future.