Discriminating Oral Squamous Cell Carcinoma via μFTIR Spectroscopy Imaging and Tree-Based Models

Daniella L. Peres
Center for Lasers and Applications
Nuclear and Energy Research Institute &
University of São Paulo
São Paulo, Brazil
0009-0001-6762-5422

Daniela Teixeira da Silva
Center for Lasers and Applications
Nuclear and Energy Research Institute
São Paulo, Brazil
0000-0002-7228-6146

Joaquim C. Felipe
Faculty of Philosophy,
Sciences and Letters at Ribeirão Preto
University of São Paulo
Ribeirão Preto, Brazil
0000-0002-3411-3551

Luciana Corrêa
Faculty of Dentistry
University of São Paulo
São Paulo, Brazil
0000-0002-5774-0750

Leandro L. de Matos
State Cancer Institute of São Paulo
University of São Paulo
São Paulo, Brazil
0000-0002-5068-8208

Thiago Martini Pereira
Institute of Science and Technology
Federal University of São Paulo
São José dos Campos, Brazil
0009-0007-1101-5354

Denise Maria Zezell
Center for Lasers and Applications
Nuclear and Energy Research Institute
São Paulo, Brazil
0000-0001-7404-9606

2025 SBFoton International Optics and Photonics Conference (SBFoton IOPC) | 979-8-3315-9497-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/SBFOTONIOPC66433.2025.11218421

***Abstract*—This study evaluates the classification of oral squamous cell carcinoma (OSCC) using micro-FTIR hyperspectral imaging and two tree-based models: Random Forest and XGBoost. After preprocessing and PCA, class imbalance was addressed with AllKNN and Tomek Links. Both models achieved high accuracy, and key spectral bands—Amide I, II, and III—were consistently highlighted, supporting their relevance in oral cancer detection.**

***Index Terms*—FTIR hyperspectral imaging, Oral squamous cell carcinoma, XGBoost, Random Forest.**

I. Introduction

Head and neck cancers encompass several malignancies; oral squamous cell carcinoma (OSCC) is the most common and aggressive subtype, primarily affecting the oral cavity [1]. In Brazil, about 15,100 new OSCC cases are expected annually (2023–2025) [4]. Risk factors include tobacco, alcohol, HPV, and poor oral hygiene [1]; the diagnostic gold standard remains histopathology, often supported by CT/MRI [2].

Fourier-transform infrared (FTIR) spectroscopy offers a non-destructive, label-free way to probe tissue biochemistry [5]. FTIR hyperspectral imaging yields high-dimensional data (thousands of spectra per image), motivating machine-learning approaches to extract robust diagnostic patterns.

Clinical deployment of μFTIR is challenging due to sample-preparation variability, optical/spectral artefacts, batch effects and domain shift, throughput and data logistics, workflow integration, and regulatory/human factors. These motivate harmonized SOPs, rigorous QA/QC, domain-shift mitigation, and aggregation from pixel-level outputs to slide-/patient-level decisions.

Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are tree ensembles suited to high-dimensional FTIR spectra. RF reduces variance via bagging and random feature subsets, while XGBoost uses second-order boosting with L1/L2 regularization and shrinkage. Both rely on univariate splits; we summarize importance as mean decrease in impurity (RF) and gain/cover/frequency (XGBoost). Because neighboring wavenumbers are highly collinear, we smooth/group importance curves and interpret biochemical peaks (e.g., amide I/II) with caution.

II. Material and Methods

The hyperspectral dataset comprised 96 core human biopsies from tissue microarrays (TMAs), including 48 healthy oral mucosa and 48 OSCC samples, mounted on Low-E IR-compatible slides. Samples were collected in collaboration with the State Cancer Institute of São Paulo (ICESP), under ethical approval (CAAE 32884214.5.0000.0065).

This work was supported by CNPq (INCT-INTERAS 406761/2022-1), INCT-INFO (465763/2014-6); Sisfoton (440228/2021-2); CAPES (88887.854461/2023-00–PROEX/USP); CAPES Finance code 001 and FAPESP (21/00633-0).

To capture biochemical information, spectral images were acquired in transmission mode using a Cary 660 FTIR spectrometer with a 32×32 FPA detector, providing 5.5 μm spatial resolution. Spectra were recorded from 400–4000 cm^{-1} at 4 cm^{-1} resolution, using 128 background scans and 64 scans per sample.

Pixels were treated independently (i.e., without spatial context), and split-based feature importances can be unstable under strong wavenumber collinearity. The μFTIR modality also poses challenges—sensitivity to sample preparation, scattering and baseline distortions, atmospheric absorptions, instrument drift, and cross-batch variability—which can induce domain shift and class imbalance; in this study, we mitigated these effects with standardized preprocessing and quality control.

For supervised learning, histological validation was essential. Thus, parallel slides were stained with hematoxylin and eosin (H&E) and imaged with a Nikon Eclipse Ti microscope. An expert pathologist annotated the slides, providing the ground truth labels used in model training and evaluation.

A. Spectral Preprocessing Pipeline

Before model training, spectra were preprocessed to improve quality and comparability (for more details, see [3]). Analysis focused on the fingerprint region (900–1800 cm^{-1}), using SNV normalization and Savitzky–Golay second derivatives (7-point window) to correct baseline shifts and noise. EMSC was applied to reduce paraffin and vapor interference, and low-quality spectra were excluded.

PCA reduced dimensionality to 7 components, preserving 95% of variance. Class imbalance and redundancy were addressed using AllKNN and Tomek Links. All steps were implemented in Python 3.12.4 with standard scientific libraries.

B. XGBoost and Random Forest Classifiers

Despite their similarities, XGBoost and Random Forest differ in ensemble construction. Random Forest builds multiple independent trees on bootstrap samples, using random feature subsets at each split to reduce overfitting via decorrelation [7]. XGBoost, by contrast, builds trees sequentially, each correcting previous errors, and leverages gradient-based optimization for faster convergence and regularization [6].

These differences are particularly relevant for hyperspectral data, which are high-dimensional and complex. Hyperparameters were optimized separately via grid search with cross-validation. As summarized in Table I, XGBoost required shallow trees and low learning rates to avoid overfitting, while Random Forest performed better with deeper trees and minimal constraints—reflecting their distinct training paradigms.

To assess model performance, several evaluation metrics were employed, including precision, recall, F1-score, accuracy, and the area under the Receiver Operating Characteristic curve (ROC AUC).

III. Results and Discussion

This study emphasizes image-wise evaluation, where pixel-level predictions are aggregated via majority voting—mimicking how pathologists assess entire tissue slides.

TABLE I
BEST HYPERPARAMETERS FOR EACH MODEL

Parameter	XGBoost	Random Forest
`n_estimators`	100	50
`max_depth`	3	10
`learning_rate`	0.1	–
`subsample`	1.0	–
`colsample_bytree`	0.8	–
`min_samples_split`	–	2
`min_samples_leaf`	–	2

TABLE II
CLASSIFICATION REPORT (XGBOOST)

Class	Precision	Recall	F1-score	Support
Cancer	1.00	0.82	0.90	11
Control	0.82	1.00	0.90	9
Accuracy	0.90			
Macro avg	0.91	0.91	0.90	20
Weighted avg	0.92	0.90	0.90	20

This approach better reflects clinical decision-making and is key to assessing diagnostic applicability.

XGBoost achieved 90% accuracy (Table II), with balanced F1-scores for both classes. Notably, cancer precision was 1.00, indicating no false positives, while recall was 0.82, reflecting two missed cancer cases (Table III). All control samples were correctly classified. These results highlight the model's reliability in identifying healthy tissue, while its conservative cancer detection minimizes false positives—a desirable feature in clinical settings.

TABLE III
CONFUSION MATRIX (XGBOOST)

	Predicted: Cancer	Predicted: Control
Actual: Cancer	9	2
Actual: Control	0	9

When comparing image-wise performance, XGBoost outperformed Random Forest in both overall accuracy (90% vs. 85%) and recall for cancer detection (0.82 vs. 0.73), as seen in Tables II and IV. While both models achieved perfect precision (1.00) for cancer cases, XGBoost demonstrated greater sensitivity, misclassifying fewer cancer images. In contrast, Random Forest showed excellent specificity for the control class (recall = 1.00), but at the expense of missing more malignant cases. This trade-off highlights XGBoost's more balanced performance, making it a more favorable candidate in clinical scenarios where early and accurate cancer detection is essential.

A. Feature weights analisys

Figure 1 shows the spectral feature importance for XGBoost and Random Forest, projected back into the original infrared domain. Both models rely on key biochemical bands to distinguish cancerous from control tissue, particularly those linked to protein and nucleic acid alterations in OSCC.

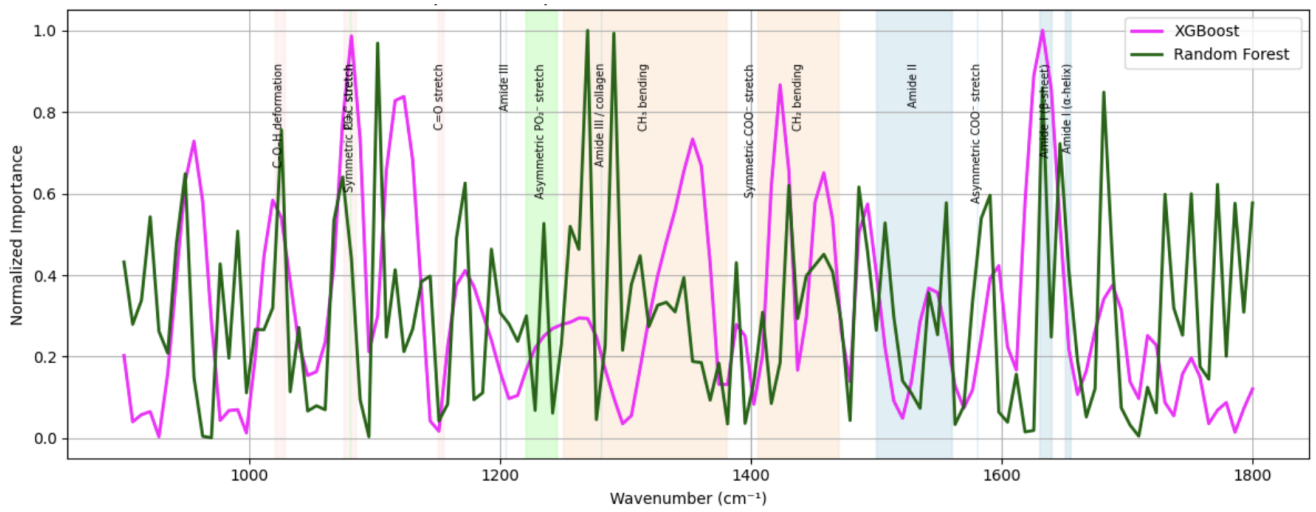

Fig. 1. Spectral importance curves for XGBoost and Random Forest models. Each curve represents the relative contribution of each spectral band (900–1800 cm^{-1}) to the classification task. Shaded areas indicate known biochemical bands associated with biological constituents: light blue for proteins, peach for lipids, light green for nucleic acids, and light pink for carbohydrates.

TABLE IV
CLASSIFICATION REPORT (RANDOM FOREST)

Class	**Precision**	**Recall**	**F1-score**	**Support**
Cancer	1.00	0.73	0.84	11
Control	0.75	1.00	0.86	9
Accuracy	0.85			
Macro avg	0.88	0.86	0.85	20
Weighted avg	0.89	0.85	0.85	20

TABLE V
CONFUSION MATRIX (RANDOM FOREST)

	Predicted: Cancer	**Predicted: Control**
Actual: Cancer	8	3
Actual: Control	0	9

High importance is observed in the Amide I (1650–1666 cm^{-1}), Amide II (1500–1560 cm^{-1}), and Amide III (1200–1300 cm^{-1}) regions—markers of protein structure and folding, often altered in malignant tissue. Both models also highlight the symmetric phosphate stretch at 1080 cm^{-1}, associated with increased nucleic acid content in proliferating cells.

XGBoost emphasizes Amide A and the phenolic ring (1468 cm^{-1}), while Random Forest shows broader sensitivity in the CH_2/CH_3 deformation range (1400–1450 cm^{-1}), often linked to lipid and membrane changes [8]. These patterns confirm that both classifiers capture spectroscopic markers relevant to OSCC, supporting their interpretability and clinical potential.

IV. CONCLUSION

From a spectral perspective, both XGBoost and Random Forest consistently highlighted key biochemical regions—such as Amide I, II, and III—which are closely linked to protein structural changes characteristic of malignant transformation. This alignment with known molecular signatures enhances the interpretability and biological relevance of the models' decisions.

In terms of diagnostic performance, XGBoost and Random Forest achieved image-level accuracies of 90% and 85%, respectively—figures comparable to those reported by experienced pathologists when evaluating conventional histological slides. Importantly, XGBoost demonstrated slightly higher sensitivity for cancer detection, a critical factor for minimizing false negatives in clinical applications.

REFERENCES

[1] M. I. Bellantoni, G. Picciolo, I. Pirrotta, N. Irrera, M. Vaccaro, F. Vaccaro, *et al.*, "Oral cavity squamous cell carcinoma: an update of the pharmacological treatment," *Biomedicines*, vol. 11, no. 4, p. 1112, 2023. Available: https://doi.org/10.3390/biomedicines11041112

[2] C. C. Torras, C. Gay-Escoda, "Techniques for early diagnosis of oral squamous cell carcinoma: Systematic review," in Medicina oral, patologia oral y cirugia bucal 3 (2015), vol. 20, pp. e305, Medicina Oral SL.

[3] T. Pereira, E. Almeida, D. Peres, D. F. T. Silva, G. Germano, D. M. Zezell, and L. Bachmann, "Mathematical method for enhancing biochemical information in micro-FTIR images of histopathological slides," in *Optical Biopsy XXIII: Toward Real-Time Spectroscopic Imaging and Diagnosis (SPIE)*, San Francisco, CA, USA, 2025, p. 46.

[4] INCA, Estatísticas de câncer, available at: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros, accessed on August 28, 2024.

[5] D. Max, "Modern vibrational spectroscopy and micro-spectroscopy: theory, instrumentation, and biomedical applications," John Wiley & Sons, 2015.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[7] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[8] A. Paraskevaidi, R. M. Ashton, K. M. Stringfellow, M. A. Lima, M. F. L. Lemos, P. L. Martin-Hirsch, and F. L. Martin, "Diagnostic applications of vibrational spectroscopy in the detection of cancer: A review," *Appl. Spectrosc. Rev.*, vol. 56, no. 8, pp. 621–652, 2021, doi: 10.1080/05704928.2020.1866571.