

PAPER • OPEN ACCESS

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

To cite this article: D L Peres *et al* 2026 *J. Phys.: Conf. Ser.* **3183** 012003

View the [article online](#) for updates and enhancements.

You may also like

- [Prediction of malignant transformation in oral epithelial dysplasia using machine learning](#)
James Ingham, Caroline I Smith, Barnaby G Ellis et al.
- [Analysis of Oral Squamous Cell Carcinoma into Various Stages using Pre-Trained Convolutional Neural Networks](#)
A. Ramalingam, P. Aurchana, P. Dhanalakshmi et al.
- [Using random forest for brain tissue identification by Raman spectroscopy](#)
Weiyi Zhang, Chau Minh Giang, Qingan Cai et al.

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

D L Peres¹, D F T Silva¹, G C M. Germano¹, T M Pereira²,
J C Felipe³, L L de Matos⁴ and D M Zezell¹

¹ Centro de Lasers e Aplicações, Instituto de Pesquisas Energéticas e Nucleares, Av. Prof. Lineu Prestes, 2272- São Paulo- SP, Brazil. Contact Phone: +551128105667

² Instituto de Ciências Tecnologia, Universidade Federal de São Paulo, Rua Talim, n° 330- São José dos Campos - SP - CEP: 12231-280, Brazil, São José dos Campos, Brazil.
Contact Phone: +551233099621

³ Faculty of Philosophy, Sciences and Letters at Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes, 3900 - CEP 14040-901 Bairro Monte Alegre - Ribeirão Preto - SP -Brazil, Ribeirão Preto, Brazil. Contact Phone: +551633150598

⁴ State Cancer Institute of São Paulo, Universidade de São Paulo, Av. Dr. Arnaldo, 251 – CEP 01246-000 Cerqueira César - São Paulo - SP -Brazil. Contact Phone: +551126610000

E-mail: zezell@usp.br

Abstract. Oral squamous cell carcinoma (OSCC) remains a major global health challenge, and its early detection is essential for improving patient prognosis. Fourier-transform infrared (FTIR) hyperspectral imaging offers a powerful, label-free approach for probing biochemical alterations in biological tissues, providing rich spectral information that can support computational diagnosis. In this study, we evaluated the performance of a Random Forest classifier applied to FTIR hyperspectral images of OSCC and control oral tissues. After rigorous preprocessing and restriction of the analysis to the 1500–1750 cm^{-1} region—the portion of the fingerprint band with highest discriminative variability—the model achieved strong pixel-level performance, with high precision, recall, and an area under the ROC curve of 0.986. Image-level classification, a clinically relevant metric, yielded an accuracy of 0.86 across 96 samples, demonstrating the model’s reliability in assigning whole-sample labels. Feature importance analysis identified key vibrational modes associated with malignant transformation, notably the Amide I and Amide II bands, as well as lipid-related C=O stretching near 1735 cm^{-1} . These findings indicate that Random Forest, combined with FTIR hyperspectral imaging, can robustly capture biologically meaningful spectral signatures and represents a promising complementary tool for OSCC screening and diagnostic support.

1. Introduction

Lip and oral cavity cancers represent a major global health challenge. In 2022, there were approximately 390,000 new cases and 190,000 deaths [1]. Among these, about 90% correspond to oral squamous cell carcinoma (OSCC), the most prevalent and aggressive malignancy within head and neck cancers, primarily affecting the oral cavity [2].

The primary risk factors for OSCC include tobacco use, excessive alcohol consumption, and high-risk human papillomavirus (HPV) infections, particularly subtypes HPV-16 and HPV-18 [3]. Clinically, OSCC may manifest as persistent oral ulcers, difficulty swallowing, or unexplained bleeding [4]. Traditional diagnostic methods rely on biopsy with histopathological evaluation,



often complemented by imaging techniques such as MRI and CT to assess tumor spread and severity. However, these methods can be time-consuming [5].

Recent advances in imaging technologies, particularly hyperspectral imaging (HSI), have provided non-destructive alternatives for detailed tissue characterization, capturing both spatial and biochemical information. When combined with machine learning, HSI enables the analysis of complex datasets and can improve diagnostic precision [6, 7, 8].

Fourier Transform Infrared (FTIR) spectroscopy, widely employed in medical and biochemical applications [9, 10, 11], provides hyperspectral imaging data with high sensitivity to molecular composition and structural changes. This modality is particularly effective for analyzing heterogeneous biological samples, including OSCC tissues, where subtle biochemical variations can serve as early indicators of malignancy [12, 13]. Its ability to detect alterations in proteins, lipids, and nucleic acids makes FTIR a valuable tool for supporting early and accurate diagnosis, a key factor in improving patient prognosis [14].

In this study, we leverage Random Forest (RF) models, an ensemble learning method [15, 16], to classify FTIR hyperspectral data from OSCC tissues. RF constructs multiple decision trees trained on different random subsets of the data, and combines their predictions through a majority voting mechanism. This structure, based on bagging, introduces diversity among trees and improves model robustness when dealing with complex and variable spectral datasets. An illustrative diagram of this process is shown in Figure 1.

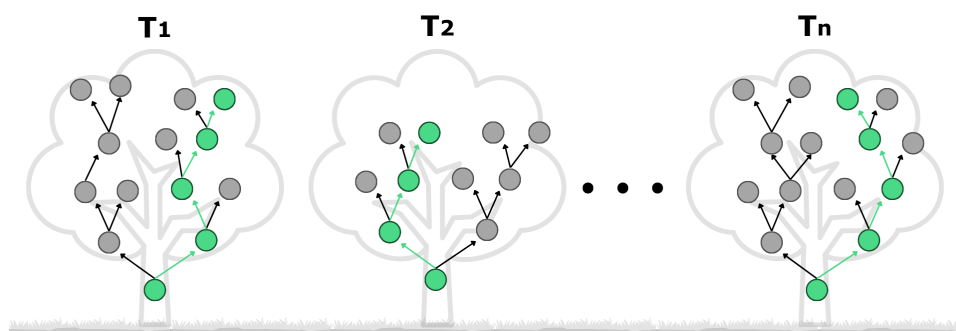


Figure 1. Illustrative diagram of a Random Forest model composed of n decision trees. Each tree acts as an independent classifier trained on a unique subset of the data. The Random Forest aggregates the tree predictions using a majority voting mechanism.

Beyond its straightforward architecture, Random Forest offers several properties that make it particularly well suited for FTIR hyperspectral analysis. The algorithm naturally handles high-dimensional and heterogeneous spectral data without requiring prior feature reduction, while capturing nonlinear relationships between biochemical signatures and tissue pathology. Its ability to mitigate overfitting through ensemble averaging is especially valuable in biomedical scenarios where datasets tend to be small and noisy. Furthermore, Random Forest provides interpretable measures of feature importance, allowing the identification of wavenumbers associated with OSCC-related biochemical alterations. These advantages have contributed to its widespread adoption in infrared spectroscopic imaging studies.

2. Material and methods

This study used 96 specimens arranged on a low-emission tissue microarray (TMA) slide, including 48 control samples of healthy human oral tissue and 48 human OSCC samples, obtained in collaboration with the Cancer Institute of Sao Paulo State (ICESP). All samples were approved by the Institutional Review Board under protocol 228/14 (CAAE

32884214.5.0000.0065). A mirrored-glass TMA slide stained with hematoxylin and eosin (H&E) provided optical microscopy images for pathologist evaluation, which served as the gold standard for labeling the samples used in model training, as shown in Figure 2.

Hyperspectral images (Figure 3) were acquired using a Cary 660 FTIR spectrometer in transfection scanning mode, coupled to a microscope equipped with a 32×32 element focal plane array (FPA) detector and a spatial resolution of $5.5 \mu\text{m}$. Each image acquisition consisted of 128 background scans, followed by 64 co-added scans per pixel during sample measurement, with a spectral resolution of 4 cm^{-1} . Reference optical microscopy images were obtained using a Nikon Eclipse Ti inverted microscope.

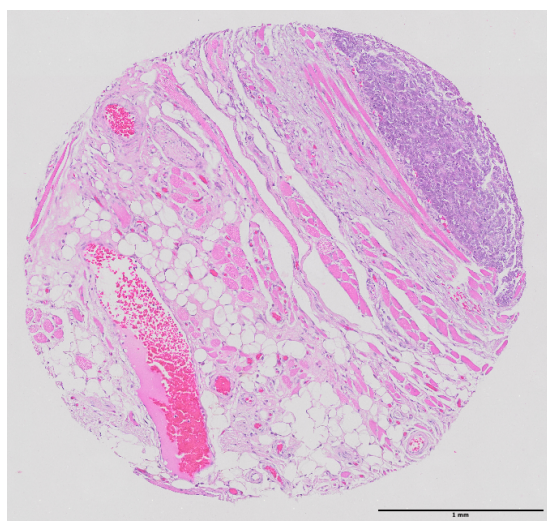


Figure 2. Biopsy image from the TMA slide stained with H&E, acquired using optical microscopy with a Nikon Eclipse Ti inverted microscope.

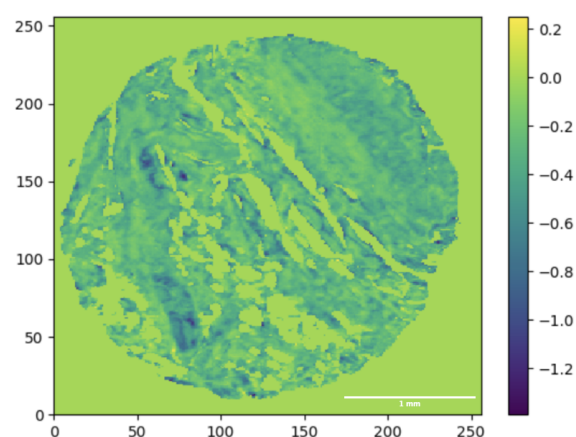


Figure 3. Hyperspectral image representation of the same sample shown in Figure 2, where the intensity of each pixel corresponds to the integrated value of its associated spectrum. The image was acquired from the Low-E TMA slide using transfection-mode FTIR.

A Random Forest model was implemented to classify the hyperspectral images, using a maximum tree depth of 3 and a total of 100 trees. Model training and evaluation were performed with the scikit-learn library [17], which provided both the implementation framework and the tools for computing performance metrics.

Random Forest was selected for its robustness when handling high-dimensional and heterogeneous spectral data, its capacity to model nonlinear relationships between biochemical signatures and tissue classes, and its reliable performance in small and noisy biomedical datasets, qualities that naturally aligned with the goals of the investigation. To ensure that the classifier operated under optimal conditions, a grid search was conducted to identify the most suitable hyperparameters, including the maximum tree depth and the number of estimators. This procedure resulted in a well-tuned Random Forest configuration tailored to the characteristics of the FTIR dataset.

Although machine learning models are designed to handle large datasets, applying preprocessing steps can significantly enhance their performance by reducing noise and limiting the number of spectra. In this study, the dataset underwent the following preprocessing procedures: (i) a spectral cut to restrict the analysis to the region of interest between 1500 cm^{-1} and 1750 cm^{-1} [18]; (ii) a quality test to remove noisy spectra with low or no information, typically originating from paraffin in the image background [18]; (iii) spectral smoothing and baseline correction using a Savitzky–Golay filter with a second-order polynomial, second

derivative, and a window size of seven points [18]; (iv) extended multiplicative signal correction (EMSC) to reduce water and paraffin contributions [18]; and (v) Standard Normal Variate (SNV) normalization, in which each point of the signal is transformed into a standardized score [19].

The evaluation was performed using a 10-fold cross-validation scheme, in which the dataset was partitioned into 10 folds, each containing 10 images for testing and 86 for training. This procedure ensures that every image is used for both training and testing at least once, providing a comprehensive assessment of model performance. Class balance was maintained across all folds to preserve consistent representation of each group. This structured validation strategy reduces the risk of biased performance estimates and increases the reliability of the results [16]. All preprocessing and data processing were performed in a Python environment using Python 3.10.9 [20].

The fingerprint region, shown in Figure 4, corresponds to the spectral range between 900 cm^{-1} and 1800 cm^{-1} , this region is essential for the detailed characterization of biological tissues. Examination of the standard deviation of the mean spectrum reveals that the upper portion of the fingerprint region, between 1500 cm^{-1} and 1750 cm^{-1} , displays markedly higher variability. Such increased spectral variance indicates stronger contrast between classes, thereby enhancing the separability of the data and facilitating the classification task.

By restricting the analysis to this region of interest, shown in Figure 5, the model achieved a notable improvement in performance. Limiting the spectral range also reduced the computational cost associated with processing the full fingerprint region, thereby accelerating the training process.

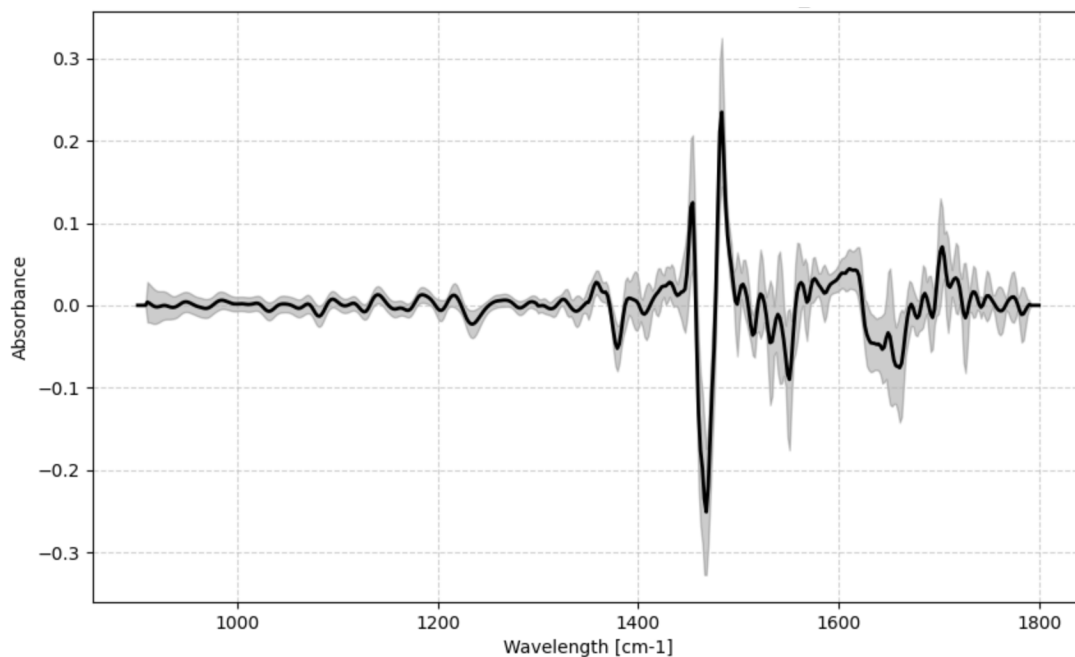


Figure 4. Average spectrum of the fingerprint region, ranging from 900 cm^{-1} to 1800 cm^{-1} . The x-axis represents the wavenumber (cm^{-1}), and the y-axis corresponds to absorbance. The bold line shows the mean spectrum, while the shaded area indicates the standard deviation.

Given the limited availability of only 96 samples, and the fact that machine learning methods require a larger number of training instances, each hyperspectral image was unfolded so that every spectrum was treated as an individual sample. This procedure resulted in a substantially

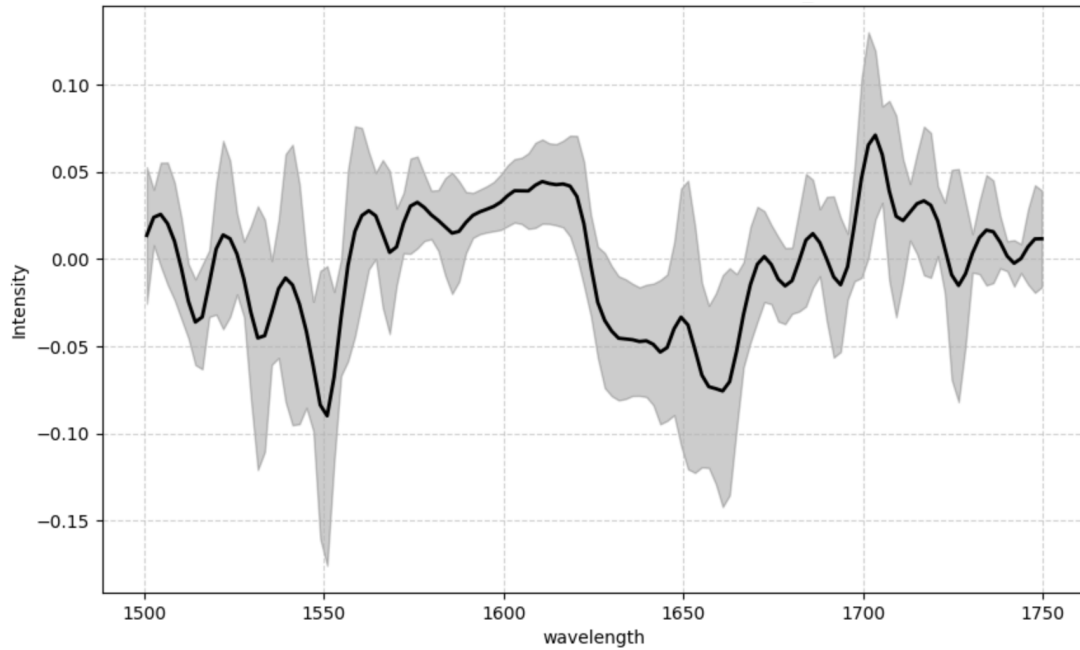


Figure 5. Average spectrum of the spectral region with the highest standard deviation, spanning from 1500 cm^{-1} to 1750 cm^{-1} . The x-axis represents the wavenumber (cm^{-1}), and the y-axis corresponds to absorbance. The bold line depicts the mean spectrum, while the shaded area illustrates the standard deviation.

larger dataset of nearly 2.5 million spectra.

$$96im \times (5 \times 5) \frac{aq}{im} \times (32 \times 32) \frac{spec}{aq} = 2457600spec \quad (1)$$

here, *image* refers to a full hyperspectral image, *spec* denotes an individual spectrum, and *aq* corresponds to a single acquisition. The term $(32 \times 32) \frac{spec}{aq}$ indicates the detector size, while $(5 \times 5) \frac{aq}{image}$ represents the area scanned—i.e., the number of acquisitions typically required to cover an entire sample, assuming the samples have an approximately regular size.

To minimize bias when working with unfolded hyperspectral images, it is essential to ensure that all spectra from the same image remain within a single group during cross-validation. This prevents correlated spectra—originating from the same image—from being split across training and testing sets. For instance, if an image is used for training, none of its spectra should appear in the test set; otherwise, the evaluation would be compromised, as the model would be tested on data it has effectively already seen, leading to overfitting and inflated performance estimates. This precaution was enforced during the labeling and group assignment stages to preserve the integrity of the analysis.

Model performance was assessed using a combination of image-level and pixel-level evaluation metrics to capture both the diagnostic reliability of the classifier and the biochemical relevance of its decisions. At the image level, accuracy was determined by assigning a label to each hyperspectral image through majority voting over its pixel predictions, reflecting the model's ability to correctly classify complete samples. At the pixel level, precision, recall, and F1-score were computed to quantify the classifier's sensitivity and specificity to local spectral variations, complemented by confusion matrices that summarize correct and incorrect assignments for each class. Receiver operating characteristic (ROC) analysis was performed to evaluate discriminative

ability across varying thresholds, with the area under the curve (AUC) providing a threshold-independent summary of performance. Additionally, the distribution of predicted probabilities was examined to assess the confidence and separability of class predictions. Feature importance values were also computed to identify the spectral regions contributing most to the classification, linking model decisions to underlying biochemical signatures. Together, these metrics offer a comprehensive evaluation of both predictive accuracy and interpretability.

3. Results and discussion

The pixel-level classification metrics, summarized in Table 1, provide a detailed assessment of the model's performance across the two classes (cancer and control). Precision measures the proportion of predicted positives that are correctly classified. Class 0 (cancer) achieved a precision of 0.92, indicating that most pixels predicted as cancer truly belonged to this class, while Class 1 (control) reached a precision of 0.96, reflecting a low rate of false positives.

Table 1. Per-pixel classification report showing precision, recall, F1-score, and support for each class.

Class	Precision	Recall	F1-score
0 (Cancer)	0.92	0.97	0.95
1 (Control)	0.96	0.89	0.93
Accuracy		0.94	

Recall quantifies the ability of the model to correctly identify all samples belonging to a given class. Class 0 exhibited a high recall of 0.97, meaning that the large majority of cancer pixels were correctly detected. Class 1 obtained a recall of 0.89, showing slightly more difficulty in identifying all control pixels. The F1-score, defined as the harmonic mean of precision and recall, summarizes the balance between both metrics, with values of 0.95 for Class 0 and 0.93 for Class 1. The overall accuracy was 0.94, demonstrating that 94% of all spectra in the dataset were correctly classified.

To further evaluate discriminative performance, a receiver operating characteristic (ROC) curve was generated (Figure 6). The ROC curve plots the true positive rate against the false positive rate across different classification thresholds, illustrating the model's ability to distinguish between classes. The area under the ROC curve (AUC) was 0.9864, indicating excellent separability and confirming that the classifier rarely confuses cancer and control spectra.

The confusion matrix, shown in Table 2, provides additional insight. For Class 0 (cancer), 1,620,264 pixels were correctly classified, while 44,481 were misclassified as control. For Class 1 (control), the model correctly identified 1,108,626 pixels, with 134,535 misclassified as cancer. These values align with the recall metrics and reinforce the model's strong ability to distinguish between cancer and control spectra at the pixel level.

Table 2. Confusion matrix for pixel-level classification, showing the distribution of true and predicted labels.

	Predicted 0	Predicted 1
True 0	1,620,264	44,481
True 1	134,535	1,108,626

Feature importance analysis was performed to identify which wavenumbers within the selected spectral region contributed most to the classification task. In a Random Forest classifier,

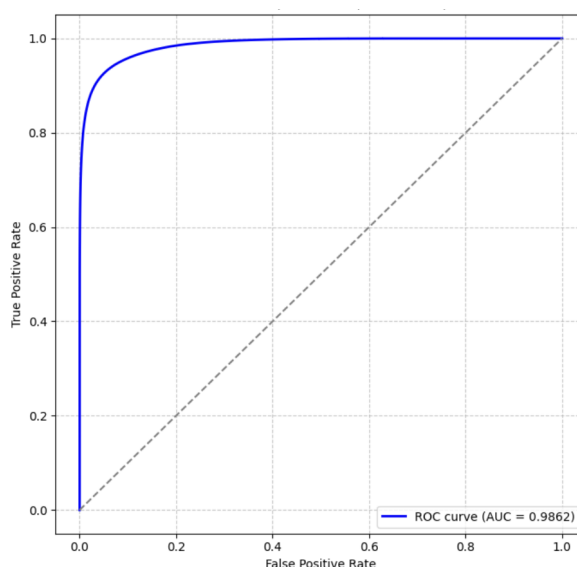


Figure 6. Receiver operating characteristic (ROC) curve for image-level classification using Random Forest with 10-fold cross-validation.

feature importance is computed by measuring how much each feature decreases impurity across all decision trees. At each split, the algorithm evaluates how effectively a given wavenumber separates the classes; features that consistently lead to larger reductions in Gini impurity receive higher importance scores. The final importance value for each wavenumber is obtained by averaging its contribution across all trees in the ensemble, normalized so that the values sum to one.

The feature importance curve, shown in Figure 7, reveals several prominent peaks between 1500 and 1750 cm^{-1} , indicating the spectral regions that most strongly contribute to the discrimination between cancerous and control tissues. In Random Forest models, feature importance quantifies how much each wavenumber reduces impurity during node splitting, averaged across all trees. By incorporating the mean importance values together with their standard deviations over the cross-validation folds, the figure highlights both the magnitude and the stability of these contributions.

The most prominent and consistent peaks occur near 1650 cm^{-1} and 1735 cm^{-1} [21], corresponding respectively to the Amide I band ($\text{C}=\text{O}$ stretching in proteins) and lipid-related $\text{C}=\text{O}$ stretching modes. These regions are well-established markers of biochemical alterations associated with malignant transformation, including changes in protein secondary structure and membrane lipid composition. Additional peaks in the 1550–1570 cm^{-1} [21] range correspond to the Amide II band, reflecting $\text{N}-\text{H}$ bending and $\text{C}-\text{N}$ stretching vibrations, which also undergo characteristic shifts in cancerous tissues.

The narrow standard deviation bands surrounding these peaks demonstrate that the most informative wavenumbers are consistently identified across all folds of the cross-validation procedure. Overall, the feature importance analysis emphasizes the biological relevance of the selected spectral window and provides insight into the molecular differences driving the classification.

In addition to pixel-level evaluation, the model's performance was assessed at the image level, where each hyperspectral image receives a single predicted label based on the majority vote of its pixel classifications. This metric is particularly important for diagnostic applications, as clinical decisions are typically made at the sample level rather than at the level of individual pixels. The

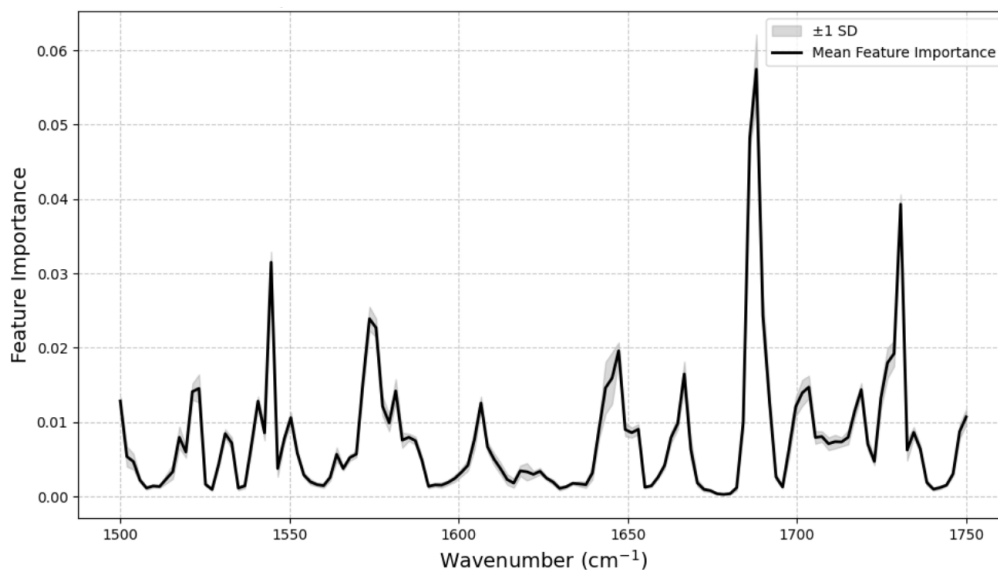


Figure 7. Mean feature importance values across the ten cross-validation folds, with shaded regions representing ± 1 standard deviation.

image-level classification report is presented in Table 3, and the corresponding confusion matrix is shown in Table 4.

Table 3. Image-level classification report showing precision, recall, F1-score, and support for each class.

Class	Precision	Recall	F1-score
0 (Cancer)	0.88	0.90	0.89
1 (Control)	0.83	0.81	0.82
Accuracy	0.86		

Table 4. Image-level confusion matrix showing the number of correctly and incorrectly classified samples for each class.

	Predicted 0	Predicted 1
True 0	53	6
True 1	7	30

The model achieved an accuracy of 0.86 across the 96 images, with Class 0 (cancer) obtaining a precision of 0.88 and a recall of 0.90, while Class 1 (control) reached a precision of 0.83 and a recall of 0.81. These results indicate that the classifier reliably distinguishes between cancerous and control samples when evaluated on complete images. Although pixel-level analysis provides finer biochemical insight into spectral heterogeneity within tissues, image-level performance is ultimately the most relevant metric for computational diagnostics, as it reflects the model's ability to assign a correct label to the entire specimen.

The confusion matrix further illustrates this behavior: out of 59 cancer images, 53 were correctly classified, and 30 out of 37 control images were identified correctly. The limited number of misclassifications suggests that, despite local spectral variability captured at the pixel scale,

the global biochemical signatures present in each sample are sufficiently distinct to support highly reliable image-level predictions.

4. Conclusion

The findings of this study demonstrate that Random Forest, when applied to FTIR hyperspectral imaging within a rigorously preprocessed spectral window, can accurately differentiate OSCC from healthy oral tissue at both the pixel and image levels. Pixel-level evaluation revealed strong precision, recall, and F1-scores for both classes, reflecting the model's ability to capture subtle biochemical variations embedded in the vibrational signatures of the tissue. The high-quality separation between cancer and control spectra was further supported by the ROC curve, which exhibited an AUC of 0.986, indicating excellent discriminative capacity.

Image-level classification—more directly aligned with clinical diagnostic needs—achieved an accuracy of 0.86, confirming that the model can generate reliable whole-sample predictions. While pixel-level analysis is valuable for biochemical interpretation and spatial mapping, the image-level result ultimately reflects the diagnostic outcome required in real-world pathology workflows.

The examination of feature importance revealed that the most discriminative wavenumbers were concentrated around the Amide I band (1650 cm^{-1}), associated with C=O stretching of proteins and highly sensitive to changes in secondary structure; the Amide II band ($1550\text{--}1570\text{ cm}^{-1}$), linked to N-H bending and C-N stretching; and the lipid-related C=O stretching band (1740 cm^{-1}). These spectral features are well established as biochemical markers of malignant transformation, reflecting alterations in protein conformation, nucleic acid content, membrane lipid composition, and metabolic activity. The consistency of these peaks across all cross-validation folds reinforces the robustness and biological relevance of the spectral signatures captured by the model.

Taken together, the results indicate that the combination of FTIR spectroscopy and Random Forest classification forms a reliable framework for OSCC tissue characterization. The approach is statistically robust, computationally efficient, and grounded in biologically meaningful vibrational features associated with carcinogenesis. These findings support the potential integration of FTIR-based machine learning tools into auxiliary diagnostic pipelines, offering interpretable and reproducible assessments that complement conventional histopathology.

Acknowledgments

This work was supported by CNPq (INCT-INTERAS 406761/2022-1, PQ 314517/2021-9); Sisfoton (440228/2021-2); CAPES (88887.176297/2025-00; Finance code 001) and FAPESP LOOPS (2022/0355-9).

References

- [1] International Agency for Research on Cancer (IARC) 2024 Cancer Today: Data Visualization Tool accessed: December 14, 2024 URL https://gco.iarc.who.int/today/en/dataviz/pie?mode=populationgroup_populations=0&cancers=1
- [2] Chen S H, Hsiao S Y, Chang K Y and Chang J Y 2021 *International journal of molecular sciences* **22** 2252
- [3] Mohan M and Jagannathan N 2014 *Oncology reviews* **8**
- [4] Celentano A and Pantaleo G 2023 Oral squamous cell carcinoma *Oral Medicine – A Clinical Guide* ed Balasubramaniam R, Yeoh S, Yap T and Prabhu S (Cham: Springer) URL https://doi.org/10.1007/978-3-031-36797-7_39
- [5] Pekarek L, Garrido-Gil M J, Sánchez-Cendra A, Cassinello J, Pekarek T, Fraile-Martínez O, García-Montero C, Lopez-Gonzalez L, Rios-Parra A, Álvarez-Mon M *et al.* 2023 *Oncology Reports* **50** 213
- [6] Urbanos G, Martín A, Vázquez G, Villanueva M, Villa M, Jimenez-Roldan L, Chavarrias M, Lagares A, Juárez E and Sanz C 2021 *Sensors* **21** 3827
- [7] Cui R, Yu H, Xu T, Xing X, Cao X, Yan K and Chen J 2022 *Sensors* **22** 9790

- [8] Lai C L, Karmakar R, Mukundan A, Natarajan R K, Lu S C, Wang C Y and Wang H C 2024 *APL bioengineering* **8**
- [9] del Valle M, Bernardes E S and Zezell D M 2023 *arXiv preprint arXiv:2310.15099*
- [10] Voronina L, Leonardo C, Mueller-Reif J B, Geyer P E, Huber M, Trubetskov M, Kepesidis K V, Behr J, Mann M, Krausz F *et al.* 2021 *arXiv preprint arXiv:2102.00765*
- [11] Baianu C, Costescu D, Hofmann N *et al.* 2004 *Microspectroscopy* **7** 450–455
- [12] Wang R and Wang Y 2021 *International journal of molecular sciences* **22** 1206
- [13] Talukdar P J, Bharti K, Banerjee S, Basu S, Das S K, Paul R R, Pal M, Mishra M P, Mukherjee S, Lahiri P *et al.* 2024 *Sensors & Diagnostics* **3** 1854–1865
- [14] Hashim D, Genden E, Posner M, Hashibe M and Boffetta P 2019 *Annals of Oncology* **30** 744–756
- [15] Prinzi F, Currier T, Gaglio S and Vitabile S 2024 *European Radiology Experimental* **8** 26
- [16] Goodfellow I 2016 Deep learning
- [17] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R and Dubourg Vincent *et al.* 2011 *Journal of Machine Learning Research* **12** 2825–2830
- [18] Ferreira M M C 2015 *Quimiometria: conceitos, métodos e aplicações* (Editora da UNICAMP)
- [19] Syvilay D, Wilkie-Chancellier N, Trichereau B, Texier A, Martinez L, Serfaty S and Detalle V 2015 *Spectrochimica Acta Part B: Atomic Spectroscopy* **114** 38–45
- [20] Python Software Foundation 2023 *Python 3.10.9 Documentation* available at <https://www.python.org/> URL <https://docs.python.org/3/>
- [21] Movasaghi Z, Rehman S and Rehman I U 2008 *Applied Spectroscopy Reviews* **43** 134–179