Evaluation of Machine Learning Methods for Oral Cavity Histopathological Cancer Classification in a Brazilian Cohort

Matheus de Freitas Oliveira Baffa[1], **Luciano Bachmann**[2], **Denise Maria Zezell**[3]
Leandro Luongo Matos[4], **Joaquim Cezar Felipe**[1]

[1]Department of Computing and Mathematics
Faculty of Philosophy, Sciences and Letters at Ribeirão Preto
University of São Paulo – Ribeirão Preto, SP – Brazil

[2]Department of Physics
Faculty of Philosophy, Sciences and Letters at Ribeirão Preto
University of São Paulo – Ribeirão Preto, SP – Brazil

[3]Nuclear and Energy Research Institute
São Paulo, SP – Brazil

[4]Department of Head and Neck Surgery
University of São Paulo Medical School
São Paulo, SP 01246-000 Brazil

{mbaffa, l.b, zezell}@usp.br, l.matos@fm.usp.br, jfelipe@ffclrp.usp.br

Abstract. *Histopathological evaluation is the gold standard for oral cavity cancer diagnosis, but it is time-consuming and subject to inter-observer variability. In this study, we compare radiomics, convolutional neural networks, and a DINOv2-based transformer model for histopathological image classification using a Brazilian cohort. Radiomic features combined with a fully-connected neural network achieved the best overall performance, reaching 93.18% accuracy, 95.56% sensitivity, and 92.50% specificity, while requiring lower computational cost than end-to-end deep learning models. These findings suggest that radiomics provides an efficient alternative for oral cavity cancer classification, particularly in scenarios with limited data and computational resources.*

1. Introduction

Oral cavity cancer comprises malignant lesions affecting the oral mucosa, including the lips, tongue, gums, and palate [National Health Service 2021]. The most common subtype is oral squamous cell carcinoma (OSCC), which accounts for over 90% of cases and is associated with risk factors such as tobacco use, alcohol consumption, and HPV infection [Tan et al. 2023]. Diagnosis is typically established through clinical examination followed by biopsy, the gold standard method, enabling histopathological evaluation to distinguish benign from malignant lesions and guide treatment decisions [Mayo Clinic 2023, National Cancer Institute 2021].

Although biopsies play an important role in cancer prognosis, manual histopathological analysis presents several challenges. The interpretation of tissue samples can vary significantly between experts, leading to inconsistencies in diagnosis and classification [Maran et al. 2022]. Additionally, the process is time-consuming, requiring extensive expertise and careful examination of microscopic features, which can result in

delays in diagnosis. Moreover, many countries lack access to specialists for certain tissue types, further prolonging the diagnostic process and potentially affecting patient outcomes [Komura et al. 2024].

Computer vision techniques for automatic pathology guidance have emerged as a potential solution to overcome the limitations of manual analysis. Most recent studies have focused on deep learning techniques, particularly models based on convolutional neural networks (CNNs) and vision transformers (ViT), to enhance diagnostic accuracy and support automated decision-making. For instance, AlexNet has been applied to classify OSCC using patch-based image analysis [Asnake et al. 2025]. CNNs have also been extended beyond histological slides, as demonstrated by their application to *in-vivo* confocal microscopy, enabling real-time oral cancer diagnosis with high precision [Ramani et al. 2025]. In addition, some approaches have incorporated active learning strategies to improve the efficiency of CNN training by selectively annotating the most informative samples [Folmsbee et al. 2018]. Transfer learning has also been explored, leveraging pre-trained CNN architectures such as AlexNet, VGG-16, VGG-19, and ResNet-50 to adapt to oral cancer datasets, often outperforming models trained from scratch [Das et al. 2020, Warin et al. 2021].

Recent literature on transformer-based models has reported promising results as well. Most researchers focus on hybrid approaches that extract features from pretrained CNNs combined with attention mechanisms, achieving competitive performance in oral cavity cancer classification [Asif et al. 2025, Deo et al. 2024, Pham 2025]. Other strategies have explored multiple instance learning frameworks that integrate transformer architectures with attention mechanisms and CNN-derived features, such as CLAM [Lu et al. 2021] and TransPath [Wang et al. 2021], which have demonstrated strong performance in histopathological image classification.

Although these methods achieve good performance, they are highly data-intensive, requiring large annotated datasets, high-performance hardware, and in some cases, pretrained foundation models. Radiomic features, in contrast, offer an alternative approach for extracting clinically relevant information from medical images. Through automated high-throughput feature extraction, radiomics enables the analysis of correlations between image characteristics and disease progression, supporting outcome prediction and classification tasks [Tomaszewski and Gillies 2021].

Therefore, this study presents a comparative evaluation of methods for oral cavity histopathological image classification, including radiomic-based approaches, combining traditional machine learning classifiers and a fully-connected neural network (FCNN), as well as CNN and a ViT model. This design enables a systematic assessment of accuracy, generalization, and computational requirements across different methodological paradigms. Importantly, all experiments were conducted on data collected at the Cancer Institute of the State of São Paulo, representing a Brazilian cohort. Given Brazil's highly diverse population and environmental context, this study contributes by establishing performance benchmarks based on locally derived data, addressing the underrepresentation of Brazilian cohorts in computational pathology research.

2. Materials and Methods

This section describes the dataset, preprocessing procedures, and classification strategies adopted in this study. First, we present the Brazilian cohort and the image preprocessing pipeline applied prior to model development. Subsequently, we detail the experimental protocol, including the patient-level cross-validation strategy used to ensure strict separation between training and testing samples.

Four classification paradigms were investigated: (i) a radiomics-based benchmark employing handcrafted features combined with traditional machine learning classifiers; (ii) an FCNN trained on selected radiomic features; (iii) a CNN based on EfficientNet for end-to-end image classification; and (iv) a ViT approach using DINOv2 embeddings followed by a supervised multilayer perceptron (MLP). All models were evaluated under the same 10-fold cross-validation scheme to ensure fair and consistent comparison.

2.1. Dataset Description and Preprocessing

The dataset comprised 151 oral cavity biopsy samples collected at the Cancer Institute of the State of São Paulo (ICES - HCFMUSP), as part of previous studies [Matos et al. 2020, Menderico Junior et al. 2021]. Among these, 102 samples were diagnosed as squamous cell carcinoma and 48 as healthy controls. To prevent data leakage, a patient-level grouping strategy was enforced during cross-validation using a stratified group-based scheme, ensuring that samples from the same patient were not simultaneously included in training and testing sets.

All tissues were stained with hematoxylin and eosin (H&E) and digitally scanned using an Olympus BX61VS microscope at $40\times$ magnification, generating high-resolution whole-slide images for computational analysis. For each biopsy core, a single representative patch was selected as the region of interest (ROI), prioritizing areas with preserved tissue morphology and high cellular density.

To reduce staining variability and enhance structural visualization, contrast-limited adaptive histogram equalization (CLAHE) was applied independently to each RGB channel. This preprocessing step was consistently adopted for all models.

Given the class imbalance, stratification at the patient level was enforced during cross-validation. No test data were used in any form of preprocessing, feature selection, or normalization, thereby strictly avoiding information leakage.

2.2. Radiomic-Based Approach

Following the preprocessing stage described above (Fig. 1, Steps 1-2), a radiomic-based pipeline was implemented. A binary mask was generated to isolate cellular regions and exclude background areas (Fig. 1, Step 3). The segmentation process involved extraction of the green channel, histogram equalization, thresholding with a cutoff value of 70, and a morphological opening operation to remove small artifacts. This ensured that feature extraction was restricted to diagnostically relevant tissue regions.

From the segmented ROIs, a total of 217 quantitative descriptors were extracted, combining classical radiomic features and complementary texture descriptors (Fig. 1, Step 4). Radiomic features were computed using the PyRadiomics framework and included

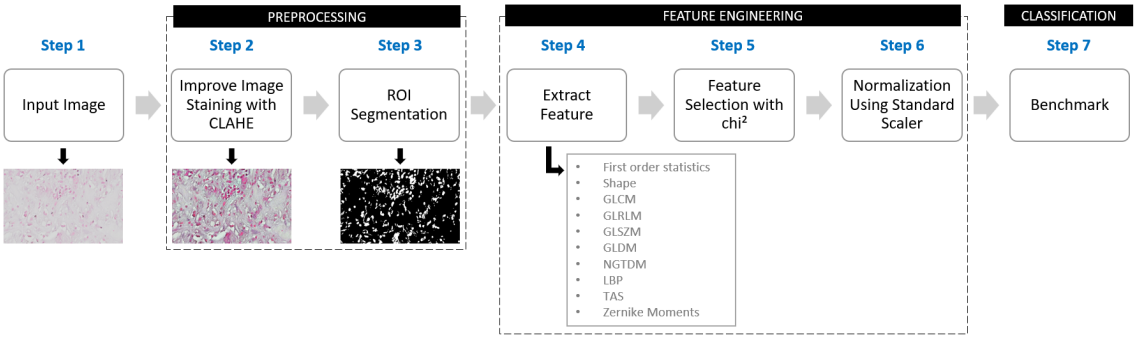

Figure 1. Overview of the radiomic-based classification pipeline.

first-order statistics, shape descriptors, and texture matrices such as the gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), gray level dependence matrix (GLDM), and neighboring gray tone difference matrix (NGTDM). Additional descriptors were extracted using the Mahotas framework, including local binary patterns (LBP), threshold adjacency statistics (TAS), and Zernike moments. Table 1 summarizes the extracted feature categories.

Table 1. Number of extracted features by categories and their respective frameworks

Feature Category	Quantity	Framework
First Order	18	PyRadiomics
Shape	9	PyRadiomics
GLCM	24	PyRadiomics
GLDM	14	PyRadiomics
GLRLM	16	PyRadiomics
GLSZM	16	PyRadiomics
NGTDM	5	PyRadiomics
Zernike Moments	25	Mahotas
LBP	36	Mahotas
TAS	54	Mahotas
Total	**217**	

To reduce dimensionality while preserving discriminative information, feature selection was performed using the chi-square (χ^2) statistical test (Fig. 1, Step 5). Since this method requires non-negative inputs, features were first scaled to the $[0, 1]$ range using a MinMax normalization. Subsequently, the 100 highest-ranked features were selected using *SelectKBest*. Importantly, this entire procedure (MinMax scaling and feature selection) was performed exclusively on the training data within each cross-validation fold to avoid information leakage.

After feature selection, the standard scaler method was applied to standardize the selected features before classification (Fig. 1, Step 6). As with feature selection, normalization parameters were learned from the training data and applied to the corresponding test fold.

To evaluate the discriminative capacity of the selected radiomic features, a benchmark including multiple traditional machine learning algorithms and a fully-connected

neural network (FCNN) was conducted (Fig. 1, Step 7). The traditional classifiers comprised logistic regression, random forest, support vector machines (linear, polynomial, and RBF kernels), k-nearest neighbors (KNN), XGBoost, AdaBoost, bagging, decision tree, extra trees, Gaussian naïve Bayes, stochastic gradient descent (SGD), and multilayer perceptron (MLP). Default hyperparameters were adopted except for KNN, the number of neighbors was fixed at $k = 46$ based on preliminary validation.

The FCNN architecture was set up as follows. The network received the 100 selected radiomic features as input and consisted of 10 hidden layers with 86 neurons each, using ELU activation functions and a dropout rate of 20% between layers. The output layer comprised a single neuron with sigmoid activation for binary classification. The network was trained for up to 5,000 epochs using binary cross-entropy loss and the RMSprop optimizer with a batch size of 128. Early stopping with a patience of 150 epochs was employed to mitigate overfitting. Model training and evaluation were performed independently within each cross-validation fold.

2.3. Convolutional Neural Network

An end-to-end convolutional neural network based on EfficientNetB0 was employed to directly learn discriminative representations from the histological patches. Unlike handcrafted feature-based approaches, this model operates directly on RGB image inputs, enabling hierarchical feature learning through deep convolutional layers pretrained on ImageNet.

All input patches were resized to a fixed spatial resolution and normalized to the $[0, 1]$ range. Prior to model training, contrast-limited adaptive histogram equalization (CLAHE) was applied independently to each RGB channel to enhance local contrast and reduce staining variability.

The EfficientNetB0 backbone was initialized with pretrained ImageNet weights and configured without its original classification head. The convolutional backbone was followed by a global average pooling layer and a fully connected classification head composed of dense layers with ReLU activation and dropout regularization (20%) to mitigate overfitting. The final output layer consisted of a single neuron with sigmoid activation for binary classification.

Data augmentation was applied exclusively to the training folds to improve model generalization. The augmentation strategy included random horizontal flipping, rotations of up to 10 degrees, and zoom variations of up to 10%. No augmentation was applied to validation data.

The model was optimized using the Adam optimizer with binary cross-entropy loss. Training was performed for up to 5000 epochs with early stopping configured to interrupt training after 40 consecutive epochs without improvement in validation loss, restoring the best-performing weights. As in the radiomic-based experiments, all training and evaluation procedures were conducted within a stratified patient-level cross-validation framework to ensure strict separation between training and testing samples.

2.4. Vision Transformers

To provide a transformer-based alternative while mitigating overfitting on the limited dataset, we adopted a foundation-model strategy based on DINOv2 [Oquab et al. 2023].

Instead of training a vision transformer from scratch, a pretrained DINOv2 ViT-S/14 backbone (`dinov2_vits14`) was used as a fixed feature extractor, and a lightweight MLP was trained on top of the resulting embeddings.

All patches were resized to 224×224 pixels to match the backbone input requirements. To maintain consistency with the CNN experiments, the same CLAHE preprocessing step was optionally applied to each RGB channel before normalization. Images were then normalized using ImageNet statistics and forwarded through the DINOv2 backbone to obtain a single D-dimensional embedding vector per sample. Embeddings were extracted in batches of 16 with the backbone kept frozen throughout all experiments.

For classification, an MLP head was trained on the extracted embeddings. The head consisted of a fully connected layer with 256 hidden units, followed by GELU activation and dropout (20%), and a final linear layer producing a single logit for binary prediction. The classifier was optimized using AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-4}) with binary cross-entropy applied to logits (`BCEWithLogitsLoss`). During inference, logits were converted to probabilities using the sigmoid function, and predictions were obtained using a 0.5 decision threshold.

To avoid using the test fold for model selection, an internal validation split was created *within each training fold* using a patient-level partition (20% of training patients). Early stopping with a patience of 25 epochs was applied based on validation loss, with a maximum of 200 epochs. As in the previous experiments, training and evaluation were performed under the same stratified patient-level 10-fold cross-validation protocol, ensuring strict separation between training and testing samples.

3. Experiments and Results

This study was implemented using Python 3.10.12. Deep learning experiments were conducted using TensorFlow 2.18.0 and Keras 3.8.0 for convolutional models, and PyTorch 2.5.1 (CUDA 12.4) for transformer-based experiments with DINOv2. Traditional machine learning models were implemented using Scikit-learn 1.6.1 and XGBoost 2.1.3. Radiomic feature extraction was performed using PyRadiomics 3.1.1, while data processing relied on NumPy 2.0.2 and Pandas 2.2.3. Image preprocessing was conducted using OpenCV 4.11.0. Experiments were executed on a Linux Ubuntu 22.04 server equipped with two Intel Xeon Silver processors, two NVIDIA RTX A4000 GPUs, 192 GB of RAM, and 2 TB of SSD storage.

We conducted experiments using a stratified patient-level 10-fold cross-validation protocol to ensure robust and unbiased evaluation. Instead of randomly partitioning individual samples, splits were generated using a stratified group-based strategy, where the grouping variable corresponded to the patient identifier. This approach guarantees that samples from the same patient don't appear simultaneously in training and testing sets, thereby preventing information leakage. The same predefined splits were used consistently across all methods (radiomics, EfficientNet, and DINOv2) to ensure a fair and controlled comparison. A fixed random seed (42) was adopted to ensure full reproducibility.

Within each training fold, an internal validation split was created at the patient level (20% of training patients) exclusively for early stopping and hyperparameter control. The test fold remained completely unseen during model training and selection.

Model performance was evaluated using six metrics: accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve (AUC). Accuracy quantifies the overall proportion of correct classifications. Sensitivity measures the ability to correctly identify malignant samples, while specificity evaluates the correct identification of healthy tissues. Precision reflects the proportion of predicted malignant cases that are truly malignant. The F1-score represents the harmonic mean of precision and sensitivity, balancing false positives and false negatives. AUC summarizes the discriminative capacity of the model across all possible decision thresholds. These evaluation metrics provide a comprehensive assessment of classification performance, particularly in a medical diagnostic setting where both false positives and false negatives carry significant clinical implications.

3.1. Results

The comparative analysis of traditional machine learning classifiers presented in Table 2 reveals a clear performance gradient across models. Among the evaluated approaches, the MLP achieved the highest overall performance, reaching 91.76% accuracy, with sensitivity of 95.56%, specificity of 88.75%, and the highest AUC of 98.09%. Logistic regression also demonstrated strong discriminative capability, achieving 89.50% accuracy and an AUC of 97.79%, indicating a well-balanced trade-off between sensitivity (93.44%) and specificity (86.25%). Linear SVM similarly showed competitive results, with 88.16% accuracy and robust specificity (85.00%).

KNN and Gaussian Naïve Bayes, on the other hand, exhibited the lowest accuracies (76.56% and 78.38%, respectively). KNN showed particularly low specificity (59.17%), indicating difficulty in correctly identifying healthy samples. In contrast, Gaussian Naïve Bayes achieved high specificity (89.17%) but comparatively lower sensitivity (75.11%), suggesting a conservative bias toward predicting healthy tissue.

The results of the FCNN model presented in Table 3 demonstrate consistently strong performance across the ten folds. Several folds achieved perfect classification (100% accuracy, sensitivity, specificity, precision, F1-score, and AUC), indicating that the network was capable of fully separating malignant and healthy samples in multiple partitions.

On average, the FCNN reached 93.18% accuracy, 95.56% sensitivity, and 92.50% specificity, with an AUC of 97.92%. These results indicate a high discriminative capability, particularly in terms of sensitivity, suggesting that the model is highly effective at identifying malignant cases while maintaining balanced performance in detecting healthy tissues. The relatively low standard deviation of AUC ($\pm$2.95) further indicates stable ranking performance across folds.

The EfficientNetB0 results presented in Table 4 demonstrate competitive performance, although with greater variability compared to the radiomic-based FCNN. The model achieved an average accuracy of 83.40%, with sensitivity of 83.72% and specificity of 77.92%. Precision remained relatively high (90.07%), while the mean F1-score reached 85.44%. The average AUC was 92.22%, indicating good overall discriminative capacity across decision thresholds.

Despite several folds reaching perfect classification (e.g., Fold 3 with 100% across all metrics), performance varied substantially across partitions. The lowest accuracies

Table 2. Benchmark results obtained using radiomic features. All values correspond to the mean over 10-fold cross-validation and are reported as percentages (%).

Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
KNN	76.56	87.89	59.17	81.49	82.77	89.23
Gaussian NB	78.38	75.11	89.17	91.81	81.47	86.92
SVM Poly	80.62	94.44	54.17	81.85	86.53	92.78
Decision Tree	81.91	85.50	79.17	87.51	85.31	82.33
Grad. Boosting	82.69	89.83	71.67	86.57	86.86	93.05
SGD	83.07	88.78	77.50	87.85	87.05	84.57
Random Forest	83.46	88.44	79.17	88.69	86.89	94.98
Extra Trees	83.55	89.44	74.17	87.03	87.09	96.44
AdaBoost	84.47	89.83	76.67	87.78	87.80	96.61
XGBoost	84.47	89.83	79.17	88.01	87.71	96.13
SVM RBF	85.51	89.28	82.92	90.06	88.33	93.52
Bagging	85.71	90.28	81.67	89.66	88.82	94.68
SVM Linear	88.16	90.94	85.00	92.30	90.63	97.31
Log. Reg.	89.50	93.44	86.25	92.57	91.96	97.79
MLP	91.76	95.56	88.75	92.89	93.40	98.09

Table 3. Results for the fully-connected neural network over 10-folds and its respective mean and standard deviation. All values are reported as percentages (%).

Fold	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
1	93.75	91.67	100.00	100.00	95.65	100.00
2	100.00	100.00	100.00	100.00	100.00	100.00
3	93.75	100.00	75.00	92.31	96.00	97.92
4	81.25	75.00	100.00	100.00	85.71	91.67
5	100.00	100.00	100.00	100.00	100.00	100.00
6	85.71	100.00	75.00	75.00	85.71	93.75
7	100.00	100.00	100.00	100.00	100.00	100.00
8	91.67	88.89	100.00	100.00	94.12	100.00
9	85.71	100.00	75.00	75.00	85.71	95.83
10	100.00	100.00	100.00	100.00	100.00	100.00
Mean	**93.18**	**95.56**	**92.50**	**94.23**	**94.29**	**97.92**
Std. Dev.	$\pm$ **6.65**	$\pm$ **7.88**	$\pm$ **11.46**	$\pm$ **9.88**	$\pm$ **5.97**	$\pm$ **2.95**

were observed in Folds 5 and 9 (71.43%), and specificity dropped to 50.00% in Folds 1 and 4. This variability is reflected in the higher standard deviations, particularly for sensitivity ($\pm$15.49) and specificity ($\pm$19.72), suggesting that the model's performance is more sensitive to patient distribution across folds.

The performance of the DINOv2 model combined with an MLP classifier (Table 5) demonstrated strong discriminative capacity across the 10 folds. Five folds (1, 3, 8, 9, and 10) achieved perfect classification results, reaching 100% in accuracy, sensitivity, specificity, precision, F1-score, and AUC.

Table 4. Results for EfficientNetB0 over 10-fold cross-validation and its respective mean and standard deviation. All values are reported as percentages (%).

Fold	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
1	75.00	83.33	50.00	83.33	83.33	85.42
2	92.31	88.89	100.00	100.00	94.12	97.22
3	100.00	100.00	100.00	100.00	100.00	100.00
4	87.50	100.00	50.00	85.71	92.31	100.00
5	71.43	70.00	75.00	87.50	77.78	75.00
6	78.57	50.00	100.00	100.00	66.67	77.08
7	78.57	70.00	100.00	100.00	82.35	100.00
8	91.67	100.00	66.67	90.00	94.74	100.00
9	71.43	83.33	62.50	62.50	71.43	91.67
10	87.50	91.67	75.00	91.67	91.67	95.83
Mean	**83.40**	**83.72**	**77.92**	**90.07**	**85.44**	**92.22**
Std. Dev.	**± 9.28**	**± 15.49**	**± 19.72**	**± 11.08**	**± 10.40**	**± 9.23**

Across all folds, the model achieved a mean accuracy of 90.07%, sensitivity of 91.56%, specificity of 85.00%, precision of 92.31%, F1-score of 91.59%, and AUC of 94.19%. Compared to EfficientNetB0, DINOv2 showed higher average accuracy and improved overall balance between sensitivity and specificity. The standard deviation values, particularly for specificity ($\pm$20.00), indicate some variability across folds; however, this variability was primarily associated with a small number of partitions (e.g., Folds 4 and 5), where specificity decreased to 50.00%.

Besides, even in folds with lower specificity, sensitivity remained consistently high, demonstrating the model's robustness in detecting malignant cases. Overall, these findings suggest that self-supervised transformer embeddings extracted by DINOv2 provide highly informative representations for oral cavity histopathology, achieving performance close to the radiomic-based FCNN while maintaining the advantages of deep representation learning.

4. Discussion

All models in this study were evaluated under an identical and controlled experimental protocol, using predefined stratified patient-level splits shared across radiomic, convolutional, and transformer-based approaches. This design ensured that performance differences were attributable to modeling strategies rather than variations in data partitioning or preprocessing. In particular, strict patient-level separation prevented information leakage and provided a realistic estimate of generalization performance in a clinical setting.

The three paradigms investigated differ fundamentally in how they represent histopathological information. Radiomics relies on handcrafted quantitative descriptors that explicitly encode intensity, morphology, and texture characteristics. EfficientNetB0 performs hierarchical convolutional feature learning through transfer learning from large-scale natural image datasets. DINOv2, in contrast, leverages self-supervised transformer-based representations trained on a large cohort, extracting high-level semantic embeddings without task-specific fine-tuning. These approaches therefore represent progressively increasing levels of representation abstraction.

Table 5. Results for DINOv2_vits14 with an MLP classifier over 10-fold cross-validation and its respective mean and standard deviation. All values are reported as percentages (%).

Fold	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC
1	100.00	100.00	100.00	100.00	100.00	100.00
2	84.62	88.89	75.00	88.89	88.89	94.44
3	100.00	100.00	100.00	100.00	100.00	100.00
4	87.50	100.00	50.00	85.71	92.31	95.83
5	78.57	90.00	50.00	81.82	85.71	82.50
6	71.43	66.67	75.00	66.67	66.67	79.17
7	78.57	70.00	100.00	100.00	82.35	90.00
8	100.00	100.00	100.00	100.00	100.00	100.00
9	100.00	100.00	100.00	100.00	100.00	100.00
10	100.00	100.00	100.00	100.00	100.00	100.00
Mean	**90.07**	**91.56**	**85.00**	**92.31**	**91.59**	**94.19**
Std. Dev.	**± 10.68**	**± 12.33**	**± 20.00**	**± 10.85**	**± 10.50**	**± 7.43**

Radiomic modeling combined with an FCNN achieved the highest and most stable performance across folds. The strong results suggest that structured quantitative descriptors capture diagnostically relevant morphological and textural differences in oral cavity tissues. Texture matrices (GLCM, GLRLM, GLSZM, GLDM, NGTDM), together with LBP and Zernike moments, likely contributed complementary information reflecting both local micro-architectural organization and global shape variability. This low variability in AUC indicates consistent ranking performance even under strict cross-validation.

EfficientNetB0 demonstrated moderate performance with higher variability across folds. Although transfer learning provides a good initialization, end-to-end convolutional fine-tuning appears more sensitive to limited dataset size and patient distribution. In contrast, the DINOv2-based approach achieved competitive performance, with multiple folds reaching perfect classification and strong average sensitivity and AUC. Notably, DINOv2 embeddings were extracted without backbone fine-tuning, suggesting that self-supervised transformer representations already encode transferable histopathological structure. However, variability in specificity across certain folds indicates that representation robustness does not fully eliminate the influence of limited sample size.

From a methodological perspective, these findings highlight an important distinction that radiomic features offer structured and interpretable representations that remain highly effective in small-data scenarios, while transformer's embeddings provide powerful, generalizable representations that reduce the need for handcrafted engineering. The results suggest that transformer-based foundation models narrow the performance gap with radiomics and may surpass them as dataset size increases.

Clinically, radiomic pipelines remain attractive due to their computational efficiency and interpretability, requiring limited hardware resources and offering explicit feature-level information. Meanwhile, foundation models such as DINOv2 represent a scalable alternative that can potentially generalize across institutions and staining variations. Future work should explore larger multicenter datasets, multimodal integration, and hybrid architectures combining radiomic descriptors with transformer-derived em-

beddings to further enhance robustness and clinical applicability.

5. Conclusion

This study compared radiomic descriptors, convolutional neural network (EfficientNetB0), and transformer-based foundation embeddings (DINOv2) for oral cavity histopathological image classification under strict patient-level cross-validation. The radiomic-based FCNN achieved the highest and most stable performance, while DINOv2 embeddings combined with a lightweight MLP delivered competitive results and demonstrated strong discriminative capability without backbone fine-tuning. EfficientNetB0 showed moderate performance with higher variability across folds. Overall, the findings indicate that radiomic modeling remains highly effective in limited-data settings, while foundation-model representations offer a promising and scalable alternative for histopathological image analysis.

Acknowledgments

This work was supported in part by the National Council for Scientific and Technological Development (CNPq) through the National Institutes of Science, Technology and Innovation Program under grant INCT-INTERAS 406761/2022-1. M. F. O. Baffa is supported in part by the São Paulo Research Foundation (FAPESP).

Use of Generative AI

Generative AI tools, such as ChatGPT 5.2, were used exclusively for language revision (grammar and spelling).

References

Asif, S., Wang, V. Y., and Xu, D. (2025). Oraltransnet: A novel hybrid model integrating transformer attention and cnn features for accurate diagnosis of mouth and oral diseases. *Engineering Applications of Artificial Intelligence*, 159:111609.

Asnake, N. W., Ayalew, A. M., and Engda, A. A. (2025). Detection of oral squamous cell carcinoma cancer using alexnet on histopathological images. *Discover Applied Sciences*, 7:155.

Das, N., Hussain, E., and Mahanta, L. B. (2020). Automated classification of cells into multiple classes in epithelial tissue of oral squamous cell carcinoma using transfer learning and convolutional neural network. *Neural Networks*, 128:47–60.

Deo, B. S. et al. (2024). Supremacy of attention-based transformer in oral cancer classification using histopathology images. *International Journal of Data Science and Analytics*, pages 1–19.

Folmsbee, J., Liu, X., Brandwein-Weber, M., and Doyle, S. (2018). Active deep learning: Improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 770–773. IEEE.

Komura, D., Ochi, M., and Ishikawa, S. (2024). Machine learning methods for histopathological image analysis: Updates in 2024. *Computational and Structural Biotechnology Journal*, 27:383–400.

Lu, M. Y. et al. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570.

Maran, B., de Brot, L., Corassa, M., and de Paula, R. (2022). Disagreement in anatomopathological review reports in gynaecological pathology and its impact on treatment: the importance of the subspecialist pathologist in a cancer centre. In *VIRCHOWS ARCHIV*, volume 481, pages S115–S115. SPRINGER ONE.

Matos, L. L. et al. (2020). Cancer-associated fibroblast regulation by micrornas promotes invasion of oral squamous cell carcinoma. *Oral Oncology*, 110:104909.

Mayo Clinic (2023). Biopsy: Types of biopsy procedures used to diagnose cancer. Available at https://www.mayoclinic.org/diseases-conditions/cancer/in-depth/biopsy/art-20043922. Accessed:on 25 Feb. 2025.

Menderico Junior, G. M. et al. (2021). Microrna-mediated extracellular matrix remodeling in squamous cell carcinoma of the oral cavity. *Head & Neck*, 43(8):2364–2376.

National Cancer Institute (2021). Head and neck cancers: Fact sheet. Available at https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet. Accessed on 27 Feb. 2026.

National Health Service (2021). Head and neck cancer. Available at https://www.nhs.uk/conditions/head-and-neck-cancer/. Accessed on 27 Feb. 2026.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Pham, T. D. (2025). Integrating support vector machines and deep learning features for oral cancer histopathology analysis. *Biology Methods and Protocols*, 10(1):bpaf034.

Ramani, R. S., Tan, I., Bussau, L., et al. (2025). Convolutional neural networks for accurate real-time diagnosis of oral epithelial dysplasia and oral squamous cell carcinoma using high-resolution in vivo confocal microscopy. *Scientific Reports*, 15:2555.

Tan, Y. et al. (2023). Oral squamous cell carcinomas: state of the field and emerging directions. *International journal of oral science*, 15(1):44.

Tomaszewski, M. R. and Gillies, R. J. (2021). The biological meaning of radiomic features. *Radiology*, 298(3):505–516.

Wang, X. et al. (2021). Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer.

Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., and Jantana, P. (2021). Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. *Journal of Oral Pathology & Medicine*, 50(9):911–918.