BioSpectralFormer: A Transformer-Based Architecture for FTIR Spectra Classification in Oral Cancer Diagnosis

Lucas. S. Procópio[1], **Robinson. S. da Silva**[2], **Murilo. G. Carneiro**[1], **Paulo. D. Souza**[1]

[1] Faculdade de Computação – Universidade Federal de Uberlândia (UFU)
Uberlândia – MG – Brazil

[2]Instituto de Ciências Biomédicas – Universidade Federal de Uberlândia (UFU)
Uberlândia – MG – Brazil

{lucas.janot,pdodonto,mgcarneiro,robinsonsabino}@ufu.br

Abstract. *Fourier-transform infrared (FTIR) spectroscopy is a promising non-invasive technique for oral cancer diagnosis, whose high mortality is largely attributable to late-stage diagnosis. This work proposes BioSpectralFormer (BSF), a Transformer-based architecture with two attention types for classification of salivary FTIR spectra. Evaluated on real spectra data under stratified 10-fold cross-validation against seven baselines, BSF achieved competitive balanced accuracy (Mean(SE,SP)* $= 0.67 \pm 0.15$*) with high sensitivity (*0.82 ± 0.20*), operating in the same statistical tier as state-of-the-art methods. Moreover, attention map analysis corroborated established oral cancer biomarkers, including Amide I and lipid C-H stretching, suggesting biological feature learning.*

1. Introduction

Oral cancer is a multifactorial genetic disease affecting millions worldwide, characterized by persistently high mortality rates. Between 1990 and 2021, the age-standardized mortality rate increased from approximately 1.83 to 2.64 per 100,000 individuals [Wu et al. 2025], with environmental risk factors such as tobacco use and alcohol consumption frequently associated with its development [Rivera 2015]. This stagnation is largely due to late diagnosis: approximately 50% of cases are detected at advanced stages, for which the five-year survival rate is below 50% [Swaminathan et al. 2024].

Fourier-transform infrared (FTIR) spectroscopy has emerged as a promising alternative for detecting biochemical alterations associated with cancerous activity. FTIR measures the absorption of infrared radiation by molecular bonds in a sample: each chemical bond vibrates at characteristic frequencies, producing a spectrum that functions as a biochemical fingerprint of the sample's molecular composition, capturing information on proteins, lipids, nucleic acids, and carbohydrates simultaneously. Owing to its rapid, non-invasive, and low-cost nature, it has been applied in diagnosing diseases such as Diabetes Mellitus, COVID-19, and Oral Cancer [Caixeta et al. 2023, Santos et al. 2023, Shree et al. 2024], with several studies demonstrating its capability of identifying early biochemical changes related to carcinogenesis [Su and Lee 2020].

Although traditional machine learning models are widely used for FTIR classification, spectra exhibit long-range dependencies between non-adjacent absorbance peaks, suggesting that Transformer architectures, based on self-attention mechanisms, may provide significant advantages. This article proposes and evaluates a novel Transformer-based architecture for classification of salivary FTIR spectra in oral cancer diagnosis,

with the following objectives: (i) adapt the Transformer architecture to the biochemical spectral domain; (ii) compare its performance with established and state-of-the-art methods (SVM-RBF, TabPFN2, CatBoost, XGBoost, TabM, LightGBM, RealMLP); and (iii) analyze interpretability through attention maps, seeking correlations with known molecular biomarkers.

2. Related Work

The application of computational techniques for analyzing FTIR spectra in oral cancer diagnosis has been explored in several recent works, with a predominant focus on traditional machine learning methods.

[Shree et al. 2024] conducted a study with 60 patients (30 with oral squamous cell carcinoma and 30 healthy controls), applying SVM for classification of salivary FTIR spectra. The model achieved 91.66% accuracy, 83.33% sensitivity, and 100% specificity and precision. The authors highlight notable alterations in protein secondary structure (Amide I and II bands) as distinctive biomarkers.

[Lima Filho et al. 2024] investigated high-level classification techniques based on complex network properties for salivary detection of oral cancer. The approach, using the Clustering Coefficient as a network measure, achieved 71% accuracy and 81% sensitivity, demonstrating that high-level classification can capture structural patterns not evident in low-level techniques.

Deep neural networks applied to FTIR spectroscopy have shown promising results in biomedical domains. For breast carcinoma, MLP models on ATR-FTIR spectra achieved 96.06% accuracy, surpassing traditional classifiers [Santos et al. 2020]. Yang et al. [Yang et al. 2022] applied 1D-CNNs to classify five stages of esophageal squamous cell carcinoma from 6,352 micro-FTIR spectra, achieving over 93% accuracy at each stage and 99% for distinguishing low-grade neoplasia, proliferation, and carcinoma.

Transformer-based architectures have been recently adapted for spectral analysis. Dai et al. [Dai et al. 2024] applied Vision Transformers to classify FTIR spectra of materials with high transferability. Fcg-Former [Zhang et al. 2024] identifies functional groups in FTIR spectra using self-attention to overcome band overlap, while spectral patch-based models have been optimized for predicting molecular structures from infrared spectra [Chen et al. 2024]. The Analytical-Chemistry-Informed Transformer [Huang et al. 2025] represents the current state-of-the-art, combining chemometric processing with intra/inter-spectral attention and baseline recombination, reducing RMSEP by over 20% in pharmaceutical, chemical, and agricultural applications.

Despite advances, important gaps remain: (i) scarcity of studies exploring Transformers for biomedical FTIR problems; (ii) absence of systematic comparisons between traditional, state-of-the-art, and deep learning models; and (iii) limited interpretability analysis correlating attention patterns with known biomarkers. This work seeks to fill all these gaps by proposing, evaluating, and interpreting a specialized Transformer architecture for salivary FTIR spectra classification in oral cancer.

3. Materials and Methods

This section describes the dataset, preprocessing pipeline, proposed architecture, training strategy, baseline models, experimental protocol, and interpretability analysis method-

ology. Figure 1 illustrates the complete experimental pipeline. All experiments were conducted on an Intel Core Ultra 9 185H processor. Mean training time per fold was 37.57 ± 13.54 s, and mean inference time per sample was 1.593 ± 0.650 ms, a latency compatible with clinical screening workflows. The source code is publicly available at `https://github.com/Lucas-Sabbatini/BioSpectralFormer`.

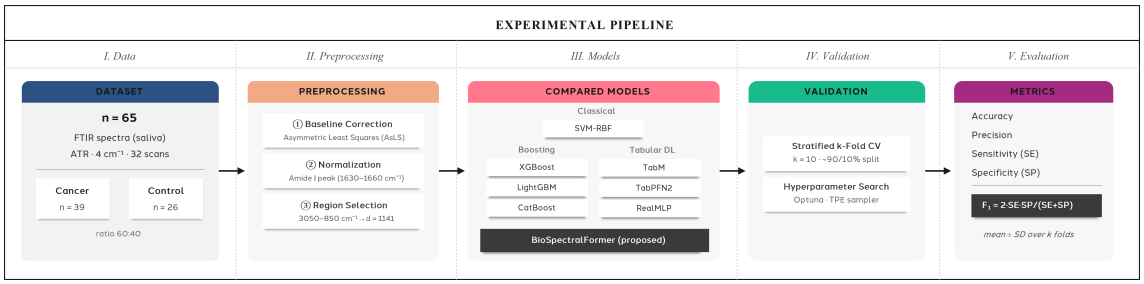

Figure 1. Experimental pipeline: from data acquisition through preprocessing, model comparison, validation, and evaluation.

3.1. Dataset

Data was collected with approval of the Research Ethics Committee of the Federal University of Uberlândia (UFU), under protocol 249.200.9, at the Clinical Hospital of UFU. It comprises 65 FTIR spectra from saliva samples: 39 from patients with histopathologically confirmed Oral Squamous Cell Carcinoma, staged according to the TNM classification of the International Union for Cancer Control, and 26 from healthy controls age- and sex-matched to the target group, with no prior history of other cancers. Samples were obtained following a standardized unstimulated salivary collection protocol. Spectra were acquired in the mid-infrared region (4000–400 cm^{-1}) using an FTIR spectrometer with ATR accessory, at 4 cm^{-1} resolution with 32 co-added scans.

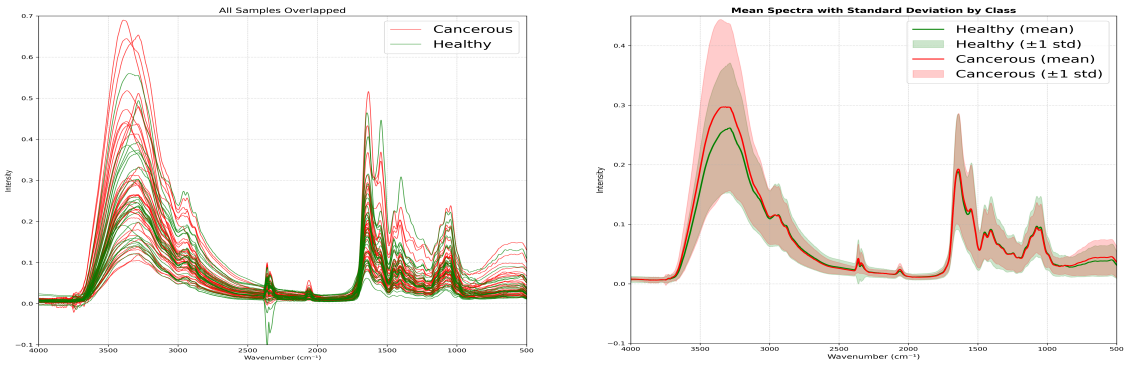

Figure 2. FTIR Samples: Overlapped (left) and Mean/Std Plot (right).

The small sample size and the strong overlap between groups reflects the challenges of biomedical spectroscopy studies (Figure 2), motivating investigation of models capable of effectively generalizing in limited-data scenarios.

3.2. Spectral Preprocessing Pipeline

The preprocessing pipeline consists of three sequential stages:

Baseline Correction: Asymmetric Least Squares (AsLS) was adopted after preliminary experiments demonstrated it provided the most robust results compared to polynomial fitting and Rubberband correction [Boelens et al. 2005].

Normalization: Amide I band normalization divides each spectrum by the maximum value in the 1630–1660 cm^{-1} region, correcting variations in total protein concentration. This is biologically grounded, as the Amide I region is highly conserved in salivary samples [Filho et al. 2023].

Truncation: The 3050–850 cm^{-1} region was selected for containing information about major macromolecules, including lipids (C-H stretches), proteins (Amide I/II bands), nucleic acids, and carbohydrates [Balan et al. 2019].

3.3. BioSpectralFormer Architecture

We propose BioSpectralFormer, a Transformer architecture adapted to the spectral domain (Figure 3), composed of the following components:

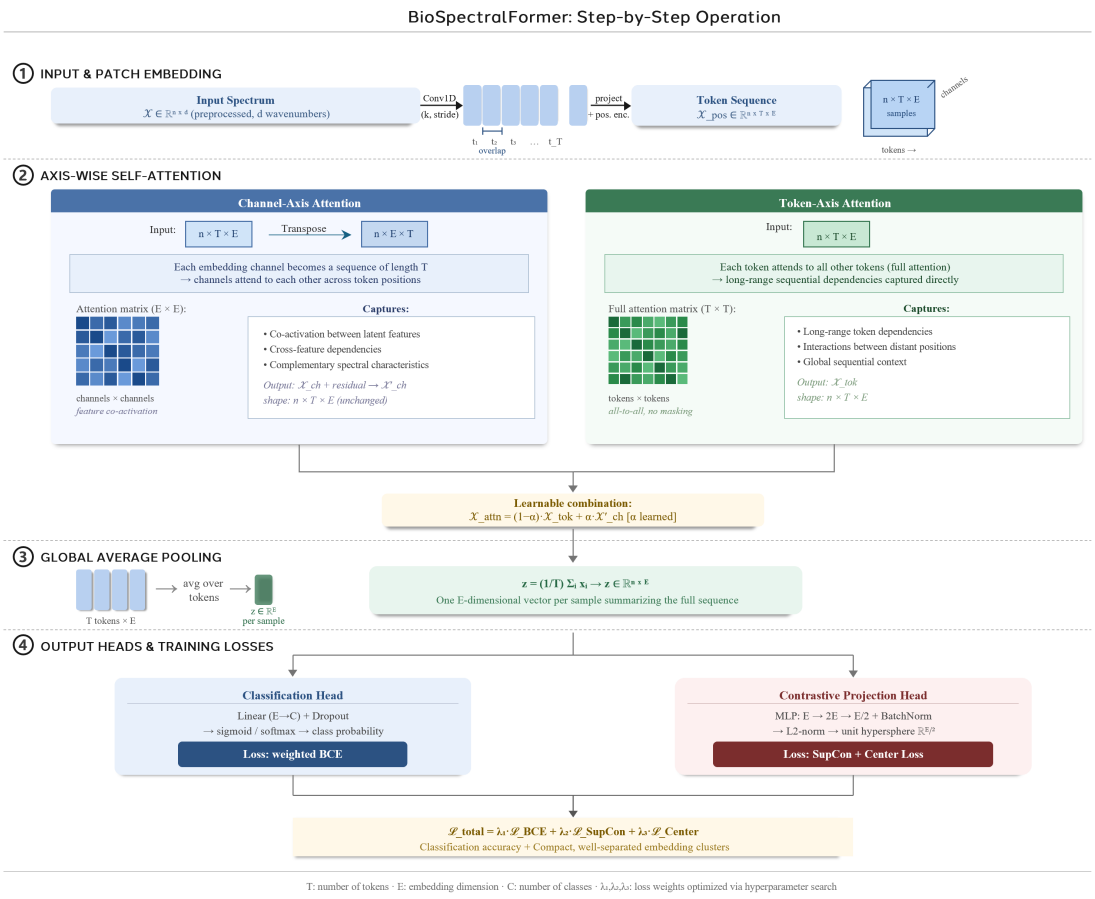

Figure 3. BioSpectralFormer: step-by-step operation.

Patch Embedding Layer: Preprocessed spectra $\mathcal{X} \in \mathbb{R}^{n \times d}$ (n=65, d=1141) are segmented into overlapping patches via 1D convolution:

$$\text{PatchEmbed}(x) = \text{Conv1D}(x; \text{kernel} = p, \text{stride} = p/2) \ ,$$

where p=16 with 50% overlap, producing $T = 141$ tokens projected to dimension E. The output is scaled by $\sqrt{E}$ with dropout applied, yielding $\mathcal{X}^{(c)} \in \mathbb{R}^{n\times T\times E}$ where T=141 and E=32.

Positional Encoding: Sinusoidal positional encodings are added to retain sequential order along the spectral axis, producing $\mathcal{X}_{\text{pos}} \in \mathbb{R}^{n\times T\times E}$.

Multi-Head Attention with Dual Axis-Wise Mechanism: Along the token axis, self-attention captures dependencies between non-adjacent absorption bands; along the channel axis, attention performs feature recalibration over the embedding dimensions, selectively amplifying diagnostically relevant features [Ding et al. 2022].

The architecture employs dual axis-wise attention on $\mathcal{X}_{\text{pos}} \in \mathbb{R}^{n\times T\times E}$:

Channel-Axis Attention operates on the embedding dimension. The input is transposed to $\mathcal{X}^T_{\text{pos}} \in \mathbb{R}^{n\times E\times T}$, treating each embedding channel across all tokens as a sequence. After linear projection $T \to E$, multi-head attention is applied and projected back, yielding $\mathcal{X}_{\text{ch}} \in \mathbb{R}^{n\times T\times E}$ with residual connection: $\mathcal{X}'_{\text{ch}} = \mathcal{X}_{\text{ch}} + \mathcal{X}_{\text{pos}}$.

Token-Axis Attention applies standard multi-head self-attention on $\mathcal{X}'_{\text{ch}}$ across the token dimension, producing $\mathcal{X}_{\text{tok}} \in \mathbb{R}^{n\times T\times E}$.

Outputs are combined via a learnable parameter $\alpha \in [0, 1]$ (sigmoid-constrained):

$$\mathcal{X}_{\text{attn}} = (1-\alpha) \cdot \mathcal{X}_{\text{tok}} + \alpha \cdot \mathcal{X}'_{\text{ch}} \in \mathbb{R}^{n\times T\times E} \ ,$$

Each attention head computes $\text{Attention}(Q, K, V) = \text{softmax}\left(QK^T/\sqrt{d_k}\right)V$ with $Q = \mathcal{X}W_Q$, $K = \mathcal{X}W_K$, $V = \mathcal{X}W_V$ and $d_k = E/h$.

Transformer Block: We employ one block containing: (1) pre-norm layer normalization, (2) multi-head dual axis-wise attention, (3) residual connection with dropout, (4) pre-norm layer normalization, (5) position-wise FFN (two-layer MLP, ReLU, d_{ff} projection), and (6) residual connection with dropout.

Global Average Pooling over the token dimension produces encoder representation $z \in \mathbb{R}^{n\times E}$, serving as input to both classification and contrastive heads.

Classification Head: Linear layer with dropout projecting to binary logits (sigmoid activation).

Projection Head: A 2-layer MLP with batch normalization maps representations to a lower-dimensional L2-normalized embedding space ($E \to 2E \to E/2$) for contrastive loss computation.

3.4. Training Strategy

To address the challenges of training a deep architecture on a small dataset, we combine a multi-objective loss function with class-balanced sampling, regularized optimization, and post-training threshold calibration.

Multi-Objective Loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{SupCon}}\mathcal{L}_{\text{SupCon}} + \lambda_{\text{Center}}\mathcal{L}_{\text{Center}}$$

Binary Cross-Entropy (weighted for class imbalance):

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[w_{\text{pos}} \cdot y_i \log(\sigma(f_i)) + (1-y_i)\log(1-\sigma(f_i))\right]$$

where $w_{\text{pos}} = N_{\text{neg}}/N_{\text{pos}}$.

Supervised Contrastive Loss (pulls same-class embeddings together, pushes different classes apart):

$$\mathcal{L}_{\text{SupCon}} = -\frac{1}{|P(i)|}\sum_{i\in I}\sum_{p\in P(i)}\log\frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p/\tau)}{\sum_{a\in A(i)}\exp(\mathbf{z}_i \cdot \mathbf{z}_a/\tau)}$$

where $P(i) = \{p \in A(i) : y_p = y_i\}$ and τ=0.07.

Center Loss (reduces intra-class variance):

$$\mathcal{L}_{\text{Center}} = \frac{1}{2N}\sum_{i=1}^{N}\|\mathbf{z}_i - \mathbf{c}_{y_i}\|_2^2 + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}}$$

where $\mathbf{c}_{y_i}$ is the learnable class center and $\mathcal{L}_{\text{sep}} = -\frac{1}{|\mathcal{C}|(|\mathcal{C}|-1)}\sum_{j\neq k}\|\mathbf{c}_j - \mathbf{c}_k\|_2$ encourages center separation. Centers are updated via a separate SGD optimizer at $10\times$ the main learning rate.

Class-Balanced Sampling: A custom sampler guarantees each mini-batch contains at least m_{min}=2 samples from each class, preventing degenerate contrastive solutions.

Optimization: AdamW ($\eta = 5 \times 10^{-3}$, weight decay 5×10^{-5}); cosine annealing with warm restarts (T_0=20, T_{mult}=2, η_{min}=10^{-6}); gradient clipping (max norm 1.0); early stopping (patience 50) based on:

$$\text{Score}_{\text{val}} = 0.4 \cdot \text{Acc}_{\text{val}} + 0.5 \cdot \frac{S+1}{2}$$

where $S \in [-1, 1]$ is the Silhouette score.

Threshold Calibration: Post-training, we optimize the classification threshold on the validation set:

$$t^* = \arg\max_{t\in[0.1,0.9]}\frac{1}{2}\left(\text{Sensitivity}(t) + \text{Specificity}(t)\right)$$

Hyperparameters (optimized via Optuna): h=4 heads, E=32, d_{ff}=64, patch size p=16, dropout=0.3, batch size=8, n_{epochs}=200, λ_{BCE}=0.389, λ_{SupCon}=0.081, λ_{Center}=0.530.

3.5. Comparison Models

Seven models representing different paradigms were compared, all optimized via Optuna: **SVM-RBF**, widely established in FTIR literature [Shree et al. 2024]; **XGBoost** [Chen and Guestrin 2016], gradient boosting with regularization; **LightGBM** [Ke et al. 2017], leaf-wise gradient boosting for high-dimensional data; **CatBoost** [Prokhorenkova et al. 2018], gradient boosting with ordered boosting; **TabM** [Gorishniy et al. 2025], parameter-efficient MLP ensemble; **RealMLP** [Holzmüller et al. 2024], MLP with layer normalization and residual connections; and **TabPFN2** [Hollmann et al. 2022], Transformer pre-trained for few-example tabular classification.

3.6. Experimental Protocol

Data Splitting: Stratified 10-fold cross-validation, with 90% training (58–59 samples) and 10% testing (6–7 samples) per fold, maintaining the 39:26 class proportion.

Evaluation Metrics: Accuracy, Precision, Sensitivity (Recall), Specificity, and Mean(SE,SP) as the primary metric given class imbalance. Sensitivity is also prioritized given the clinical importance of minimizing false negatives.

Significance Analysis: Means and standard deviations over 10 folds, with paired t-tests for pairwise comparisons ($p < 0.05$).

3.7. Interpretability Analysis

For each classified test sample, attention weights are extracted from the last Transformer layer and averaged across all heads and samples to identify consistently prioritized spectral regions and verify correspondence with established oral cancer biomarkers.

4. Results

Table 1 presents the average results (±standard deviation) of 10-fold cross-validation for the 8 evaluated baseline models, within our proposed architecture.

Table 1. Average performance (±SD) of baseline models and BioSpectralFormer in 10-fold cross-validation.

Model	Accuracy	Precision	Recall	Specificity	Mean(SE,SP)
SVM-RBF	0.62±0.16	0.65±0.13	**0.83±0.16**	0.32±0.23	0.57±0.16
LightGBM	**0.74±0.17**	**0.84±0.18**	0.74±0.16	**0.75±0.27**	**0.75±0.18**
CatBoost	0.72±0.18	0.81±0.19	0.77±0.18	0.67±0.32	0.72±0.19
XGBoost	0.70±0.18	0.74±0.16	0.82±0.17	0.53±0.30	0.68±0.20
RealMLP	0.68±0.24	0.73±0.28	0.75±0.27	0.58±0.32	0.67±0.24
TabM	0.60±0.15	0.67±0.16	0.72±0.18	0.42±0.33	0.57±0.17
TabPFN2	0.52±0.16	0.59±0.18	0.75±0.22	0.18±0.32	0.47±0.17
BSF	0.70±0.15	0.72±0.13	0.82±0.20	0.52±0.25	0.67±0.15

LightGBM achieves the best overall performance, demonstrating excellent balance between sensitivity and specificity with high precision. **BioSpectralFormer** shows competitive performance with notably high sensitivity—valuable in cancer detection where false negatives carry serious clinical consequences. Its low standard deviations (lowest among deep learning models) suggest stable learning despite limited data. The dual axis-wise attention mechanism provides clear advantage over tabular deep learning baselines, though it does not yet match top gradient boosting methods in specificity.

4.1. Exploratory Attention Analysis

Attention weights were extracted separately for cancer and healthy samples across all heads and folds (Figure 4), enabling a richer analysis across both attention mechanisms.

Token-Axis Attention. Per-head analysis reveals meaningful specialization: Head 2 consistently emphasizes 1623.9 cm^{-1} (Amide I shoulder), Head 4 prioritizes 991.3 cm^{-1} (nucleic acid backbone, PO_2^- modes), and Heads 1 and 3 jointly emphasize lipid C-H stretching (2734.8, 2919.9 cm^{-1}) and Amide II subregions, suggesting complementary

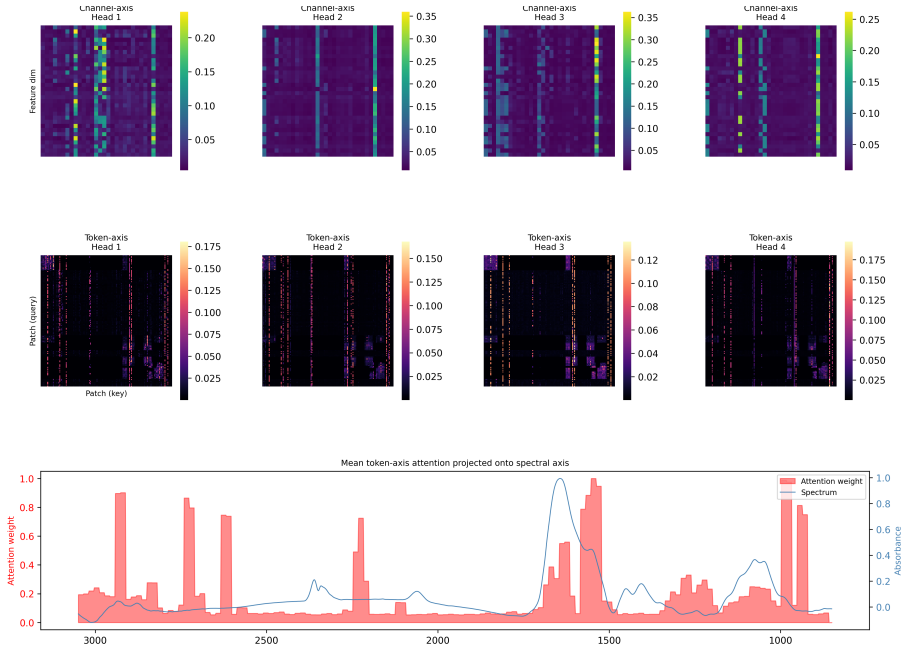

Figure 4. Average Attention over Test Set across all folds

spectral roles rather than redundant behavior. Class-separated analysis reveals a discriminative pattern: cancer samples consistently elevate 1546.7 cm^{-1} as the primary attended hub, while healthy samples shift emphasis to 1562.2 cm^{-1}, a subtle but biologically meaningful displacement within the Amide II band, associated with cancer-driven alterations in protein secondary structure. Regions 991.3, 2919.9, and 2734.8 cm^{-1} appear prominently in both classes, suggesting they function as shared spectral anchors rather than discriminative features.

Channel-Axis Attention. Unlike token-axis attention, channel dominance is distributed between all dimensions, which alternate in ranking across heads and classes without consistent hierarchy. This more diffuse pattern suggests the channel-axis mechanism functions as a broad feature gate rather than converging on a single dominant dimension, complementing the more focused relational behavior of token-axis attention. However, top attention dimensions were 3 and 27.

These regions strongly correlate with established oral cancer biomarkers in the literature [Su and Lee 2020, Shree et al. 2024], suggesting the model is learning biologically relevant features rather than artifacts or noise.

4.2. Ablation Study

Loss Components. Using only BCE yields the weakest performance (Accuracy: 58.8%, Mean(SE,SP): 56.3%). Adding $\mathcal{L}_{\text{SupCon}}$ improves accuracy to 68.3% and sensitivity to 81.7%, but specificity remains low (46.7%), indicating contrastive learning drives discriminative feature learning without sufficient class compactness. Adding $\mathcal{L}_{\text{Center}}$ alone yields a more balanced profile (Accuracy: 67.9%, Specificity: 55.0%). The full combination achieves the best results (Accuracy: 70.0%, Mean(SE,SP): 66.7%, Specificity: 51.7%), confirming the three losses are complementary rather than redundant.

Attention Mechanisms. Token-axis attention alone achieves 51% accuracy, channel-axis

alone 62%, and both combined 70%, confirming complementarity rather than redundancy.

4.3. Statistical Analysis

Table 2 presents the comparative analysis between BSF and LightGBM.

Table 2. Statistical Performance Analysis: BioSpectralFormer vs. LightGBM

Metric	BSF	LGBM	Diff	t-stat	p (t-test)	p (Wilcoxon)
Accuracy	0.7000	0.7429	-0.0429	-0.865	0.4094	0.3750
Precision	0.7171	0.8350	-0.1179	-2.115	0.0635	0.0703
Sensitivity	**0.8167**	0.7417	0.0750	1.406	0.1934	0.3750
Specificity	0.5167	**0.7500**	-0.2333	-2.409	0.0393	0.0625
Mean(SE, SP)	0.6667	0.7458	-0.0792	-1.541	0.1578	0.2109

Results averaged over 10 stratified folds.

Most metrics show no statistically significant difference ($p > 0.05$). A distinct behavioral pattern emerges: BSF exhibits higher Sensitivity but lower Specificity. For Specificity, the t-test suggests significance ($p = 0.0393$) while the more conservative Wilcoxon test yields $p = 0.0625$, indicating that LightGBM's superiority in specificity is not entirely robust across all partitions.

Table 3. Pairwise Post-hoc Analysis: Rank Differences and p-values

	BSF	XGB	SVM	TPFN	CatB	RMLP	TabM
XGB	0.25 (.819)	–					
SVM	1.40 (.201)	1.65 (.132)	–				
TPFN	2.20* (.044)	2.45* (.025)	0.80 (.465)	–			
CatB	0.90 (.411)	0.65 (.552)	2.30* (.035)	3.10** (.004)	–		
RMLP	0.25 (.819)	0.00 (1.00)	1.65 (.132)	2.45* (.025)	0.65 (.552)	–	
TabM	1.40 (.201)	1.65 (.132)	0.00 (1.00)	0.80 (.465)	2.30* (.035)	1.65 (.132)	–
LGBM	1.60 (.144)	1.35 (.217)	3.00** (.006)	3.80** (.000)	0.70 (.522)	1.35 (.217)	3.00** (.006)

Cells: Rank Diff (p-value). * $p < 0.05$, ** $p < 0.01$.

The pairwise post-hoc analysis (Table 3) on Mean(SE,SP) shows that SVM, TabPFN2, and TabM consistently exhibit significant performance gaps against top-tier models. BioSpectralFormer operates in the highest statistical tier alongside LightGBM, CatBoost, XGBoost, and RealMLP, establishing it as a viable approach for this domain.

5. Discussion

Baseline results are consistent with biomedical spectroscopy studies facing sample size limitations. Mean accuracy of 66%, though lower than the 91.66% reported by Shree et al. [Shree et al. 2024], reflects real biological variability (salivary heterogeneity, cancer subtype diversity) and conservative preprocessing prioritizing reproducibility. Studies with larger datasets ($N \geq 150$) consistently achieve $\geq$90% accuracy [Yang et al. 2022, Leng et al. 2023], while comparable small-sample work reports 71–91% [Shree et al. 2024, Lima Filho et al. 2024], confirming sample size as the dominant limiting factor.

The observed SE/SP trade-off is a critical finding. Dataset imbalance introduces a bias toward the positive class. LightGBM achieved the best balance, while XGBoost and BSF maintained high sensitivity at moderate specificity cost. In clinical screening, prioritizing sensitivity is justifiable since false negatives carry more severe consequences. From this perspective, BSF could serve as an initial screening tool, though very low specificity implies unacceptable unnecessary biopsies. LightGBM's balanced profile represents a more viable compromise for clinical deployment.

Regarding architectural choices, the multi-objective loss creates gradient regularization: cross-entropy provides direct supervision, while supervised contrastive loss prevents task-specific overfitting by forcing generalizable features [Khosla et al. 2020], and center loss compacts class distributions [Wen et al. 2016]. The dual attention mechanism improves performance through complementarity, channel-axis performs feature selection while token-axis performs relational reasoning [Wang et al. 2017], as shown in the ablation study (Section 4). Transformers can operate on small data through constrained adaptation: the lightweight block limits hypothesis space while contrastive regularization prevents class collapse [Dosovitskiy et al. 2021, Touvron et al. 2021].

Future Directions: Promising paths emerge: (i) dataset expansion to 200–500 spectra through multicenter collaborations; (ii) pre-training on large public FTIR corpora followed by fine-tuning, analogous to NLP/CV transfer learning; (iii) ensemble combination of BSF and LightGBM, which have shown complementary strengths and weaknesses.

Limitations: The main limitations are: (i) **reduced sample size** – complex models like Transformers are especially susceptible to overfitting on small datasets, and regularization strategies mitigate but do not eliminate this risk; (ii) **absence of external validation** – all models were evaluated on the same dataset, requiring testing on independent cohorts for true generalization assessment; and (iii) **computational cost** – the $O(T^2 \times E)$ complexity remains efficient for our configuration, but larger datasets may require sparse or linear attention optimizations.

6. Conclusion

This work proposed and evaluated BioSpectralFormer (BSF), a Transformer-based architecture adapted to the biochemical spectral domain for classification of salivary FTIR spectra in non-invasive oral cancer diagnosis. Statistical analysis confirmed BSF operates in the highest performance tier alongside gradient boosting methods. Interpretability analysis through attention maps revealed biologically grounded behavior: the model consistently prioritized spectral regions corresponding to established oral cancer biomarkers. This correlation between attention patterns and known molecular markers validates that the Transformer architecture learns clinically relevant features rather than spurious correlations, addressing a critical gap in interpretability for deep learning-based diagnostic systems. Furthermore, this study establishes a methodological foundation for applying Transformers to biomedical spectroscopy, contributing a rigorously documented preprocessing pipeline, comparative benchmarking, and interpretable attention mechanisms while addressing the biggest challenge in this area: the lack of agreement between different methodologies.

Acknowledgment

Authors thank the financial support given by the Brazilian National Council for Scientific and Technological Development - CNPq (grants n. 420212/2023-0 and 445027/2024-0), the Minas Gerais Research Foundation – FAPEMIG (grant n. BDT-00010-25), and the INCT in Theranostics and Nanobiotechnology (grant n. CNPq-465669/2014-0).

References

Balan, V. et al. (2019). Vibrational spectroscopy fingerprinting in medicine: from molecular to clinical practice. *Materials*, 12(18):2884.

Boelens, H. F. M. et al. (2005). New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and raman spectroscopic detection. *Journal of Chromatography A*, 1057:21–30.

Caixeta, D. C. et al. (2023). Salivary atr-ftir spectroscopy coupled with support vector machine classification for screening of type 2 diabetes mellitus. *Diagnostics*, 13:1396.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. ACM.

Chen, X. et al. (2024). Patch-based self-attention for molecular structure prediction from infrared spectroscopy. *Journal of Physical Chemistry A*, 128:5665–5678.

Dai, H. et al. (2024). Vision transformers for materials identification using x-ray diffraction and infrared spectroscopy. *Digital Discovery*, 3:234–245.

Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., and Yuan, L. (2022). DaViT: Dual attention vision transformers. In *Computer Vision – ECCV 2022*, pages 74–92.

Dosovitskiy, A. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Filho, A. M., Fernandes, J., Sabino-Silva, R., and Carneiro, M. (2023). Ocanspectra: an oral cancer detection system from salivary atr-ftir spectroscopy. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*, pages 984–996. SBC.

Gorishniy, Y., Kotelnikov, A., and Babenko, A. (2025). Tabm: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, volume 2025, pages 77899–77935.

Hollmann, N. et al. (2022). Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*.

Holzmüller, D., Grinsztajn, L., and Steinwart, I. (2024). Better by default: Strong pretuned mlps and boosted trees on tabular data.

Huang, S., Jin, Y., Jin, W., and Mu, Y. (2025). Analytical-chemistry-informed transformer for infrared spectra modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17440–17448.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30.

Khosla, P. et al. (2020). Supervised contrastive learning. In *NeurIPS*.

Leng, P. et al. (2023). Cnn-lstm neural network for ftir spectroscopy-based cancer detection. *Analytical and Bioanalytical Chemistry*, 415:3891–3901.

Lima Filho, R. B., Fernandes, J. M., Ji, D., Zhao, L., Sabino-Silva, R., and Carneiro, M. G. (2024). High-level network-based detection of oral cancer from atr-ftir spectroscopy. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Rivera, C. (2015). Essentials of oral cancer. *International Journal of Clinical and Experimental Pathology*, 8(9):11884–11894.

Santos, A. P., Filho, A. C. M., Sabino-Silva, R., and Carneiro, M. G. (2023). Convolutional neural networks for the molecular detection of covid-19. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 51–62, Cham. Springer Nature.

Santos, M. C. D. et al. (2020). Atr-ftir spectroscopy with chemometric algorithms of multivariate classification in the discrimination between healthy vs. dystrophic mammary tissues. *Analytical Methods*, 12:1385–1396.

Shree, P., Aggarwal, Y., Kumar, M., Majhee, L., Singh, N. N., Prakash, O., Chandra, A., Mahuli, S. A., Shamsi, S., and Rai, A. (2024). Saliva based diagnostic prediction of oral squamous cell carcinoma using FTIR spectroscopy. *Indian J Otolaryngol Head Neck Surg*, 76(3):2282–2289.

Su, K.-Y. and Lee, W.-L. (2020). Fourier transform infrared spectroscopy as a cancer screening and diagnostic tool: A review and prospects. *Cancers*, 12(1).

Swaminathan, D., George, N. A., Thomas, S., and Iype, E. M. (2024). Factors associated with delay in diagnosis of oral cancers. *Cancer Treat. Res. Commun*, 40:100831.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *ICML*.

Wang, F. et al. (2017). Residual attention network for image classification. In *CVPR*.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *ECCV*.

Wu, J., Chen, H., Liu, Y., Yang, R., and An, N. (2025). The global, regional, and national burden of oral cancer, 1990-2021: a systematic analysis for the global burden of disease study 2021. *J. Cancer Res. Clin. Oncol*, 151(2):53.

Yang, L. et al. (2022). Deep learning-based fourier transform infrared spectroscopy for identifying esophageal squamous cell carcinoma. *Spectrochimica Acta Part A*, 271:120891.

Zhang, Y. et al. (2024). Fcg-former: Functional group identification in ftir spectra using transformers. *Analytical Chemistry*, 96:7890–7899.