Attention to FTIR: A Transformer 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 detection. 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 biologically feature learning.*

1. Introduction

Oral cancer is a multifactorial genetic disease affecting millions worldwide, characterized by persistently high mortality rates. 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]. Therefore, Fourier-transform infrared (FTIR) spectroscopy has emerged as a promising alternative for detecting biochemical alterations associated with cancerous activity. Owing to its rapid, non-invasive, and low-cost nature several studies demonstrate its capability of identifying early biochemical changes related to carcinogenesis [Su and Lee 2020].

Although traditional machine learning models are widely used in FTIR, spectra exhibit long-range dependencies between non-adjacent peaks, suggesting that Transformer architectures may provide significant advantages. This work proposes BioSpectralFormer (BSF), a Transformer-based architecture adapted to the biochemical spectral domain for salivary FTIR classification in oral cancer diagnosis. BSF is benchmarked against seven baselines, with attention maps validating known biomarker correlations.

2. Related Work

Computational approaches to FTIR-based oral cancer diagnosis have predominantly relied on traditional machine learning and deep neural networks. Previous studies using SVM and complex network methods on salivary FTIR achieved 71-92% accuracy with sensitivities ranging from 81-83% [Shree et al. 2024, Lima Filho et al. 2024]. These works identified protein structural changes and lipid alterations as key biomarkers.

Recent Transformer adaptations for spectroscopy include Fcg-Former for functional group identification [Zhang et al. 2024] and chemical-informed architectures [Huang et al. 2025], though biomedical FTIR applications remain limited. This work addresses gaps in Transformer-based biomedical FTIR classification, systematic method comparisons, and interpretability linking attention to cancer biomarkers.

3. Materials and Methods

3.1. Dataset

We analyzed 65 salivary FTIR spectra (39 oral squamous cell carcinoma, 26 healthy) acquired at 4000–400 cm^{-1} with 4 cm^{-1} resolution using ATR-FTIR. Approved by the Research Ethics Committee of the Federal University of Uberlândia under protocol 249.200.9, control patients were age- and sex-matched to the cancer group. Figure 1 shows strong overlap reflecting the challenges of biomedical spectroscopy. Code is available at `https://github.com/Lucas-Sabbatini/BioSpectralFormer`.

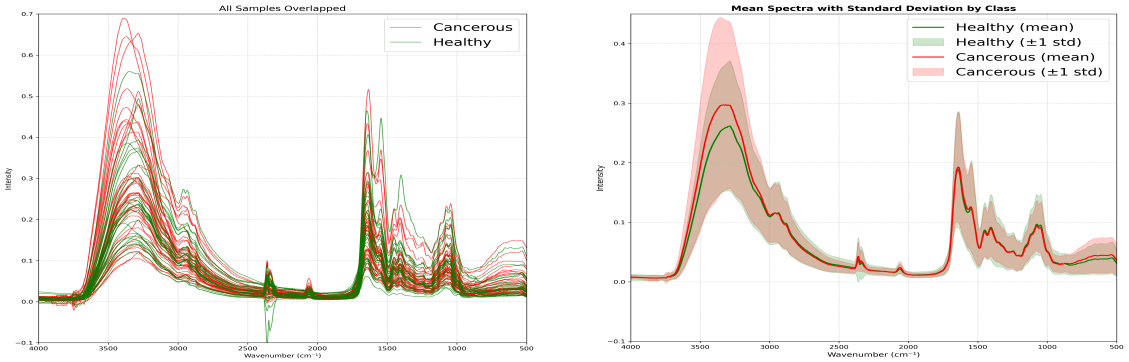

Figure 1. FTIR Samples: Overlapped (left) and Mean/Std Plot (right).

3.2. Spectral Preprocessing Pipeline

Three sequential preprocessing stages were applied: (i) **Baseline Correction** using Asymmetric Least Squares (AsLS); (ii) **Normalization** by the Amide I band maximum (1630–1660 cm^{-1}) to correct protein concentration variations; and (iii) **Truncation** to the 3050–850 cm^{-1}) region containing lipid, protein, nucleic acid, and carbohydrate information [Balan et al. 2019].

3.3. BioSpectralFormer Architecture

BSF adapts Transformer architecture to the spectral domain (Figure 2) through the following components:

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 with kernel size p=16 and stride $p/2$, producing T=141 tokens projected to embedding dimension E=32.

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

Dual Axis-Wise Attention: The architecture employs dual axis-wise attention on $\mathcal{X}_{\text{pos}} \in \mathbb{R}^{n \times T \times E}$. The Channel-axis acts as a feature gate, suppressing uninformative dimensions, while the token-axis captures long-range dependencies between non-adjacent regions.

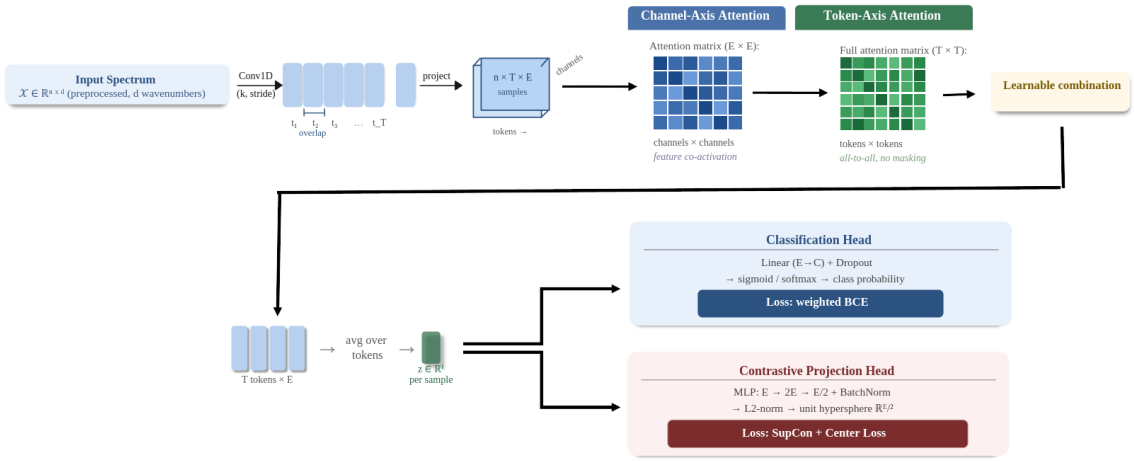

Figure 2. BioSpectralFormer: step-by-step operation.

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 learnable weighting: $\mathcal{X}_{\text{attn}} = (1-\alpha)\cdot\mathcal{X}_{\text{tok}} + \alpha\cdot\mathcal{X}'_{\text{ch}}$ where $\alpha \in [0,1]$ (sigmoid constrained).

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: Each block comprises pre-norm layer normalization, dual axis-wise multi-head attention (h=4 heads), dropout residual connections, layer normalization and position-wise FFN with dimension d_{ff}=64.

Output Heads: Global average pooling over tokens produces representations $z \in \mathbb{R}^{n\times E}$. A classification head projects to binary logits (sigmoid), while a projection head (2-layer MLP: $E \to 2E \to E/2$ with L2 normalization) generates embeddings for contrastive learning.

3.4. Training Strategy

Multi-objective loss combines weighted cross-entropy ($w_{\text{pos}} = N_{\text{neg}}/N_{\text{pos}}$), supervised contrastive loss (τ=0.07), and center loss with separation regularization. Class-balanced sampling ensures m_{min}=2 samples per class per batch. AdamW optimizer (η=5×10^{-3}), cosine annealing, gradient clipping, and early stopping with post-training threshold calibration ($t^* = \arg\max \frac{1}{2}(\text{SE}+\text{SP})$). Hyperparameters (Optuna): h=4, E=32, d_{ff}=64, p=16, dropout= 0.3, batch= 8, epochs= 200, λ_{BCE}=0.389, λ_{SupCon}=0.081, λ_{Center}=0.530. $\mathcal{L}_{\text{total}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{SupCon}}\mathcal{L}_{\text{SupCon}} + \lambda_{\text{Center}}\mathcal{L}_{\text{Center}}$

3.5. Comparison Models

Seven Optuna-optimized baselines: SVM-RBF [Shree et al. 2024], XGBoost [Chen and Guestrin 2016], LightGBM [Ke et al. 2017], CatBoost

[Prokhorenkova et al. 2018], TabM [Gorishniy et al. 2025], RealMLP [Holzmüller et al. 2024], TabPFN2 [Hollmann et al. 2022].

3.6. Experimental Protocol

Data Splitting: Stratified 10-fold CV (90% train, 10% test), maintaining 39:26 class ratio. **Metrics:** Accuracy, Precision, Sensitivity, Specificity, Mean(SE,SP) (primary). **Significance:** Paired t-tests ($p < 0.05$). **Interpretability:** Attention weights from final layer averaged across heads and samples to identify biomarker-relevant regions.

4. Results

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

Table 1. Performance of models in 10-fold CV ($\pm$SD)

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 with excellent sensitivity-specificity balance. **BSF** demonstrates competitive performance with notably high sensitivity (0.82±0.20) and robust stability (lowest SD among deep learning baselines), valuable in cancer detection where false negatives carry serious clinical consequences. The dual axis-wise attention mechanism provides clear advantage over tabular deep learning baselines, though specificity remains lower than top gradient boosting methods.

4.1. Significance Analysis

Pairwise t-tests and Wilcoxon signed-rank tests reveal BSF operates in the highest performance tier alongside LightGBM, CatBoost, XGBoost, and RealMLP, with no significant differences in Mean(SE,SP) ($p > 0.05$). BSF exhibits higher sensitivity but lower specificity than LightGBM (specificity: p=0.039 by t-test, p=0.063 by Wilcoxon), while SVM, TabPFN2, and TabM show significant performance gaps relative to top-tier models.

4.2. Exploratory Attention Analysis

Attention maps reveal biologically meaningful feature prioritization across two mechanisms (Figure 3). **Token-axis attention** exhibits head-specific specialization: Head 2 emphasizes the Amide I shoulder (1623.9 cm^{-1}), Head 4 targets nucleic acid modes (991.3 cm^{-1}), and Heads 1 and 3 prioritize lipid C-H stretching (2734.8 and 2919.9 cm^{-1}). A discriminative Amide II shift is observed between cancer (1546.7 cm^{-1}) and healthy (1562.2 cm^{-1}) samples, reflecting protein structural alterations. Conversely, **channel-axis attention** (notably dimensions 3 and 27) acts as a broad feature gate rather than a focused relational anchor. These regions align with established oral cancer biomarkers [Su and Lee 2020, Shree et al. 2024], validating that BSF learns clinically relevant features rather than noise.

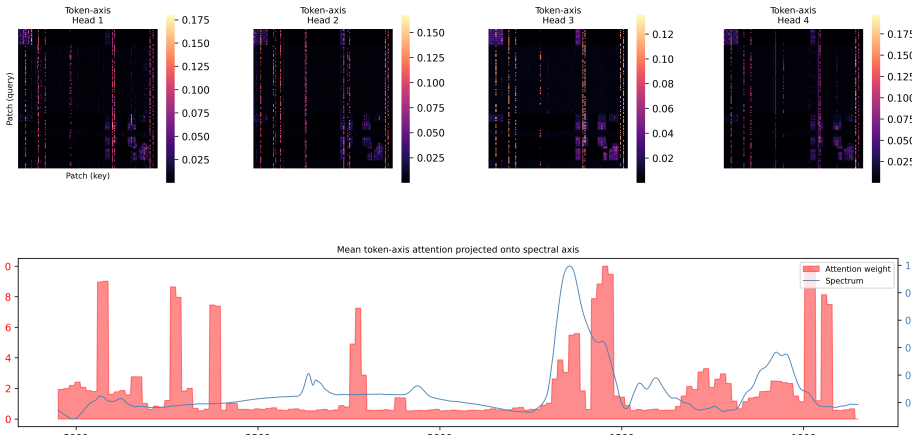

Figure 3. Average Attention over Test Set across all folds

5. Discussion

Results are consistent with biomedical spectroscopy under sample-size constraints, where studies with $N < 100$ typically report 60–75% accuracy compared to 85–95% for $N > 200$ [Shree et al. 2024, Yang et al. 2022]. The lower accuracy reflects biological variability and conservative preprocessing prioritizing reproducibility, suggesting sample size rather than technique inadequacy as the primary limitation.

The sensitivity-specificity tradeoff reveals distinct model behaviors: LightGBM achieves optimal balance, while BSF and XGBoost prioritize sensitivity at moderate specificity cost. In cancer screening contexts, high sensitivity is clinically preferable since false negatives carry severe consequences, positioning BSF as a potential initial screening tool pending specificity improvements.

BSF addresses small-data challenges through architectural design: dual axis-wise attention combines feature selection (channel-axis) with relational reasoning (token-axis), while multi-objective loss (cross-entropy + contrastive + center) prevents overfitting by balancing feature learning with intra-class compactness [Khosla et al. 2020]. This constraint directly mirrors clinical reality: prospective salivary FTIR datasets are inherently limited, making robustness under data scarcity a clinical requirement.

6. Conclusion

This work proposed BioSpectralFormer (BSF), achieving competitive performance on salivary FTIR oral cancer classification through dual axis-wise attention and multi-objective contrastive learning. BSF operates in the highest statistical tier alongside gradient boosting methods. Attention maps validate biological grounding, consistently prioritizing established biomarker regions (Amide I, lipid C-H), demonstrating clinically relevant feature learning. Key limitations include small sample size and absence of external validation. Future work should prioritize multicenter expansion (200–500 spectra), transfer learning from spectroscopic corpora, and prospective clinical validation.

Acknowledgment

Authors thank the financial support given by CNPq (grants n. 420212/2023-0 and 445027/2024-0), 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.

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.

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 pre-tuned 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*.

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.

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.

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.