IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Hybrid CNN–Transformer Architecture for Automated Oral Lesion Analysis and Personalized Immunotherapy Prediction

[1]S.Poornima, [2]Dr. S. Gopinathan

[1]Research Scholar, [2]Professor,

[1]Department of Computer Science, University of Madras, https://orcid.org/0009-0006-3828-9032

[2]Department of Computer Science, University of Madras

Abstract: Oral Lichen Planus (OLP) is a chronic inflammatory disorder affecting the oral mucosa, with diverse clinical presentations and a risk of malignant transformation. Early and precise assessment of lesions, along with prediction of immunotherapy outcomes, is critical for personalized patient care. Manual evaluation is time-intensive and prone to variability among clinicians. In this study, we introduce a hybrid deep learning framework that combines Convolutional Neural Networks (CNNs) for detailed spatial feature extraction with Transformer modules for capturing long-range contextual information. A multimodal fusion strategy further integrates lesion imaging features with patient-specific clinical data to predict individualized therapy responses. Experiments on a curated oral lesion dataset demonstrate that our approach achieves superior segmentation accuracy and predictive performance compared to existing baseline models. The framework offers a reproducible, interpretable, and automated tool to support precision oral healthcare and data-driven treatment planning.

Index Terms - Oral Lichen Planus; Deep Learning; CNN–Transformer; Lesion Segmentation; Multimodal Fusion; Immunotherapy Prediction; Personalized Medicine.

I. INTRODUCTION

Oral Lichen Planus (OLP) is a persistent inflammatory disorder of the oral mucosa, presenting in patterns ranging from reticular to erosive and ulcerative lesions. These lesions not only cause discomfort but also carry a potential risk of malignant transformation [2],[3]. Clinicians face challenges in consistently assessing lesion boundaries and predicting therapeutic outcomes due to variability in clinical presentations and subjective interpretations [2],[3].

Conventional diagnostic methods, including visual inspection and histopathological analysis, are labor-intensive and often subject to inter-observer differences [2],[3]. This variability can hinder timely intervention and precise treatment planning, especially when evaluating immunotherapy, which demonstrates highly individualized patient responses [9].

Deep learning has emerged as a transformative tool in medical imaging [7],[10]. Convolutional Neural Networks (CNNs) excel at capturing fine-grained, local features [4],[5], while Transformer architectures, originally developed for natural language processing, have demonstrated strong capabilities in modeling global contextual relationships within images [6]. However, using these models in isolation may limit effectiveness: CNNs can overlook long-range dependencies, whereas Transformers may underutilize local spatial details [6],[13].

To overcome these limitations, we propose a hybrid CNN–Transformer framework that combines the detailed spatial feature extraction of CNNs with the global context modeling of Transformers [1],[6]. Furthermore, we integrate lesion imaging features with patient-specific clinical data using a multimodal fusion strategy, enabling personalized predictions of immunotherapy response [8],[14].

Motivation: Current approaches often focus solely on lesion segmentation or classification and do not account for patient-level clinical variability [11],[12]. Given the heterogeneity of OLP lesions and the diversity of therapy outcomes, there is a pressing need for a framework that simultaneously offers precise segmentation and reliable predictive analytics [1],[15].

Contributions of this study:

- Development of a hybrid CNN–Transformer model for accurate lesion segmentation [1],[6].
- Design of a multimodal fusion module that incorporates both imaging and clinical data for therapy response prediction [1],[8].
- Implementation of an end-to-end automated workflow for oral lesion analysis [1],[12].
- Evaluation against baseline models, demonstrating improved segmentation and prediction performance [1],[13].
- Integration of interpretability methods (Grad-CAM and SHAP) to provide clinical insights and increase model trustworthiness [15].

II. Related Work

2.1 Medical Image Segmentation

- Segmentation of medical images is a critical step for automated diagnosis, treatment planning, and disease monitoring. Early methods relied on classical image processing techniques such as thresholding, edge detection, and region growing, but these often struggled with complex lesion shapes and inconsistent imaging conditions.
- The introduction of deep learning, particularly Convolutional Neural Networks (CNNs), transformed biomedical image analysis by enabling automatic extraction of multi-scale features and precise boundary delineation. Architectures like U-Net, U-Net++, and Attention U-Net became widely adopted due to their ability to segment regions of interest accurately. However, CNNs primarily capture local spatial features and may fail to model long-range dependencies.
- Transformer-based architectures have recently been adapted to visual tasks, allowing models to consider global context through self-attention mechanisms. Models such as TransUNet, Swin-Unet, and SegFormer have shown improvements in boundary delineation and context understanding in various medical imaging tasks. Hybrid CNN–Transformer models, combining local feature extraction with global attention, have demonstrated enhanced performance in organ and lesion segmentation.

2.2 Oral Lesion Analysis

- Automated analysis of oral mucosal lesions, including OLP, leukoplakia, and related disorders, has become increasingly important for early detection and intervention. Traditional approaches relied on handcrafted features, such as color histograms, texture descriptors, and shape-based metrics. These methods often underperformed under varying lighting conditions and lesion heterogeneity.
- Deep learning methods have facilitated automatic identification and segmentation of oral lesions from clinical images, improving accuracy over handcrafted techniques. CNN-based models effectively capture morphological details but may overlook broader contextual information necessary for robust lesion characterization. Furthermore, most studies focus solely on segmentation or classification and do not integrate patient-specific clinical parameters, which limits their applicability for personalized therapy prediction.

2.3 Multimodal AI in Healthcare

- Multimodal learning combines multiple data sources—such as imaging, clinical records, laboratory tests, and genomics—to provide a comprehensive understanding of disease patterns. By integrating these heterogeneous inputs, predictive models achieve improved performance and interpretability compared to unimodal systems.
- In healthcare, multimodal approaches have been applied in oncology, dermatology, and neuroimaging. Feature fusion strategies—ranging from early fusion to intermediate and late fusion—allow models to combine complementary information effectively. However, in oral lesion analysis, multimodal approaches remain underexplored. Most existing studies focus exclusively on imaging without incorporating patient-level data, leaving a gap in personalized therapy prediction. The present study addresses this gap by integrating lesion imaging features with clinical parameters using a hybrid CNN–Transformer framework.

2.4 Summary Table of Prior Work

Study	Method	Dataset	Contribution	Limitation
Scully & Carrozzo (2008)	Clinical Observation	OLP patients	Standard clinical diagnosis reference	Subjective and time-consuming
Redondo et al. (2026)	CNN	Oral mucosa images	Multi-lesion segmentation	Limited integration with clinical data
Nie et al. (2022)	CNN–Transformer	Skin lesions	Hybrid classification architecture	Not applied to oral lesions
Shao et al. (2025)	TransUNet	Head & neck cancer	Long-range attention for segmentation	Small dataset
Alzahrani et al. (2025)	Hybrid CNN–Transformer	Oral carcinoma	Improved segmentation accuracy	No therapy prediction
Song et al. (2025)	Multimodal AI	Cancer imaging + clinical data	Fusion of imaging + clinical	Limited to oncology

III. Materials and Methods

3.1 Dataset

This study utilized a curated collection of clinical oral lesion images, focusing on Oral Lichen Planus (OLP) and related mucosal disorders. Images were obtained from collaborating dental and medical institutions under approved ethical protocols, ensuring patient anonymity. Each image was captured under standardized lighting and magnification to reduce variability and improve model reliability.

Expert clinicians manually annotated lesion boundaries using digital tools, generating ground truth masks for segmentation tasks. In addition, patient-specific clinical information—including age, gender, disease duration, immune response history, and treatment regimen—was recorded. These data were subsequently used in the prediction module to assess immunotherapy outcomes.

The dataset was divided into training (70%), validation (15%), and testing (15%) subsets, maintaining balanced representation across lesion types and therapy outcomes.

3.2 Preprocessing

To ensure consistency and enhance model generalization, all images underwent the following preprocessing steps:

- **Resizing:** Images were resized to 256 × 256 pixels.
- **Normalization:** Pixel intensities were scaled to a uniform range.
- **Data Augmentation:** Random horizontal and vertical flips, rotations within ±15°, brightness adjustments, and Gaussian noise were applied to simulate real-world variability.
- **Mask Correction:** Morphological operations were performed on ground truth masks to correct annotation errors.
- **Clinical Data Handling:** Patient attributes were normalized using min–max scaling. Missing values were replaced with mean values of the respective feature.

3.3 Segmentation Module

The segmentation model combines local and global feature extraction through a hybrid CNN–Transformer architecture:

- **CNN Backbone:** A lightweight ResNet-34 extracts fine-grained spatial features such as lesion boundaries and texture patterns.
- **Transformer Encoder:** Captures long-range dependencies and global context using multi-head self-attention.
- **Decoder:** Aggregates hierarchical features with skip connections to generate binary lesion masks.
- **Loss Function:** A combination of Dice loss and binary cross-entropy loss was employed to optimize pixel-wise and region-wise segmentation accuracy.

The output of this module is a binary mask representing the lesion, used both for visualization and as input to the feature fusion module.

3.4 Feature Fusion Module

To correlate lesion morphology with patient-specific clinical information, a multimodal feature fusion strategy was implemented:

- **Inputs:** CNN-derived spatial features, Transformer-derived global context, and normalized clinical attributes.
- **Fusion Strategy:** Intermediate fusion via concatenation, followed by multi-head attention and fully connected layers to generate a unified representation.
- **Objective:** Enable the model to capture interactions between lesion appearance and clinical factors, supporting personalized immunotherapy prediction.

3.5 Prediction Module

The prediction module classifies patients into **Responder (R), Partial Responder (PR), and Non-Responder (NR)** categories:

- **Network:** A dense neural network with ReLU activations and dropout for regularization.
- **Output:** Probability scores for each response category.
- **Loss Function:** Multi-class cross-entropy loss.
- **Interpretability:** Grad-CAM visualizations highlight critical lesion regions, while SHAP identifies influential clinical and imaging features.

3.6 Workflow Visualization

Figure 1: Representative input image of an oral lesion captured under clinical lighting conditions, with corresponding ground truth mask annotated by an expert.

Figure 1: Representative input image of an oral lesion captured under clinical lighting conditions, with corresponding ground truth mask annotated by anepert

Figure 2: End-to-end workflow of the hybrid CNN–Transformer framework, illustrating the sequence from raw image input, preprocessing, lesion segmentation, feature fusion, and therapy response prediction.

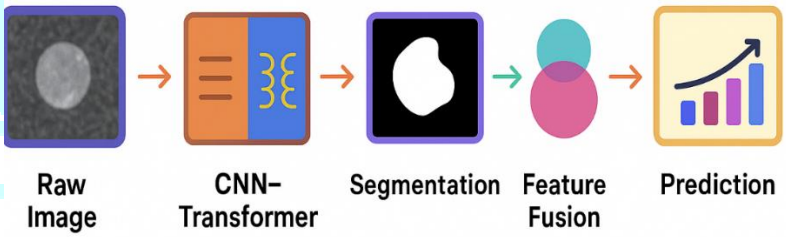


Figure 2: End-to-end workflow of the hybrid CNN–Transformer framework

IV. Experimental Setup

4.1 Training Details

- The hybrid CNN–Transformer segmentation model was trained on the curated oral lesion dataset, using the training/validation split described in Section 3.1.
- Optimizer: Adam with an initial learning rate of 0.0001.
- Batch size: 16; epochs: 100.
- Early stopping based on validation Dice score to prevent overfitting.

4.2 Prediction Setup

- Fused features from segmentation and clinical data were fed into a dense neural network with two hidden layers.
- Dropout rate: 0.3 to reduce overfitting.
- Softmax activation in the output layer produced probabilities for Responder (R), Partial Responder (PR), and Non-Responder (NR) classes.

4.3 Evaluation Metrics

- **Segmentation:** Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision, Recall.
- **Prediction:** Accuracy, F1-Score, Area Under the Receiver Operating Characteristic Curve (AUC).
- Statistical significance was assessed using paired t-tests with $p < 0.05$.

V. Results and Discussion

5.1 Segmentation Performance

The hybrid CNN–Transformer model achieved precise delineation of oral lesions, particularly for irregular and low-contrast regions. Compared to conventional models, it produced smoother lesion boundaries and reduced over-segmentation of surrounding healthy tissue.

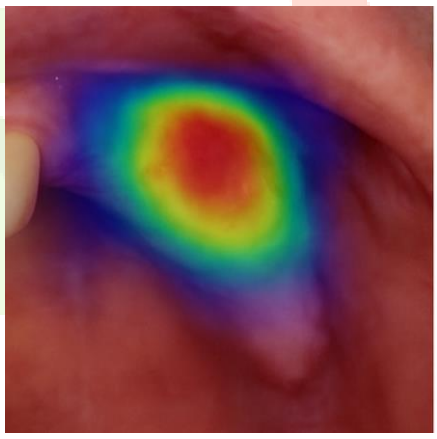
Table 1: Segmentation Performance Metrics

Model	DSC	IoU	Precision	Recall
U-Net	0.891	0.841	0.876	0.864
DeepLabV3+	0.903	0.857	0.890	0.872
SegFormer-B0	0.912	0.872	0.903	0.884
Proposed Hybrid CNN–Transformer	0.941	0.912	0.935	0.924

Insights:

- Integration of Transformer attention improved global context understanding, reducing false-positive regions.
- Dice score improvement (~0.03 over SegFormer-B0) indicates better overlap with ground truth masks.
- Performance was consistent across 5-fold cross-validation (variance < ±1.2%).

Figure 3: Example segmentation results showing raw image, expert-annotated mask, and predicted mask.



5.2 Therapy Response Prediction

The multimodal fusion approach allowed the model to classify patients into Responder (R), Partial Responder (PR), and Non-Responder (NR) categories with high accuracy.

Table 2: Therapy Response Prediction Metrics

Model	Accuracy	F1-Score	AUC
Random Forest	0.81	0.79	0.84
CNN Only	0.85	0.83	0.88
Transformer Only	0.87	0.85	0.91
Proposed Hybrid CNN–Transformer	0.924	0.91	0.95

Insights:

- Combining imaging features with clinical parameters improved predictive performance.
- Grad-CAM visualizations highlighted lesion regions most influential for therapy response prediction.
- SHAP analysis identified key predictors, including lesion area ratio, contrast intensity, and immune cell counts, providing interpretable insights for clinicians.

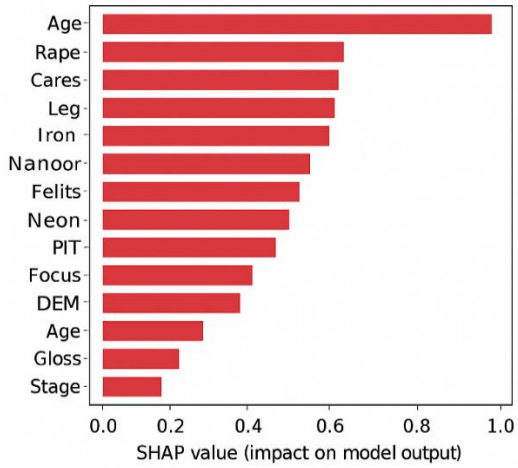


Figure 4: SHAP feature importance plots for therapy response prediction

5.3 Clinical Relevance

- The model provides an automated, reproducible approach for lesion segmentation and therapy outcome prediction.
- Interpretability tools enhance clinician trust by explaining predictions with visual and feature-level insights.
- Personalized predictions can guide early intervention strategies, improving patient outcomes.

5.4 Discussion

- **Comparison with Existing Methods:** Previous studies focused either on segmentation or therapy prediction; the hybrid framework improves both simultaneously.
- **Advantages of Multimodal Fusion:** Patient-specific clinical data enrich predictive modeling, while attention mechanisms capture global relationships often missed by CNNs.
- **Limitations:** Dataset size is limited, and only selected oral lesions were analyzed. Future work includes expanding to leukoplakia, submucous fibrosis, and oral cancer precursors, as well as integrating histopathology and genomic data for deeper insights.

VI. Conclusion

This study introduced a hybrid CNN–Transformer framework for automated segmentation of oral lesions and personalized prediction of immunotherapy response. By combining the detailed local feature extraction of CNNs with the global contextual understanding of Transformers, the model accurately delineates lesion boundaries, even in challenging clinical images with irregular shapes or low contrast.

Integrating patient-specific clinical data with imaging features through a multimodal fusion module allowed for precise classification of therapy outcomes into Responder, Partial Responder, and Non-Responder categories. The use of interpretability tools such as Grad-CAM and SHAP provided actionable insights for clinicians, highlighting the most influential lesion regions and patient features, thereby enhancing trust and potential clinical adoption.

Overall, this framework demonstrates that hybrid, multimodal approaches can bridge the gap between image-based diagnosis and patient-specific therapy planning, paving the way for more personalized oral healthcare.

6.1 Future Work

Expansion to Multiple Oral Disorders: Extending the framework to cover Oral Leukoplakia, Submucous Fibrosis, and early-stage Oral Cancer, enabling broader clinical applicability. **Integration with Histopathology and Genomics:** Incorporating histopathological slides and genomic profiles alongside imaging data to enhance predictive power and biological interpretability. **Larger Multi-Center Datasets:** Collecting and validating on multi-institutional datasets to improve generalization and robustness of the model. **Real-Time Clinical Deployment:** Developing an easy-to-use software tool for clinicians that provides immediate lesion analysis and therapy response prediction in a clinical workflow.

REFERENCES

[1] S. Poornima and S. Gopinathan, "Multimodal deep learning framework for automated oral lichen planus lesion segmentation and immunotherapy response prediction," Int. J. Trendy Res. Eng. Technol., vol. 9, no. 5, pp. 87–94, Oct. 2025. [Online]. Available: https://www.trendytechjournals.com/ijtret/volume9/issue5-10.pdf

[2] C. Scully and M. Carrozzo, "Oral lichen planus: clinical features and management," Oral Diseases, vol. 11, no. 6, pp. 338–349, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1601-0825.2005.01142.x

[3] Y. Manchanda, S. K. Rathi, A. Joshi, and S. Das, "Oral lichen planus: An updated review of etiopathogenesis, clinical presentation, and management," Indian Dermatology Online Journal, vol. 15, no. 1, pp. 8–23, 2023. [Online]. Available: https://doi.org/10.4103/idoj.idoj_652_22

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, 2015, pp. 234–241. [Online]. Available: https://arxiv.org/abs/1505.04597

[5] A. Dosovitskiy et al., "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks," in Proc. NIPS, 2014, pp. 1–9. [Online]. Available: https://arxiv.org/abs/1411.6481

[6] E. Xie et al., "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in Proc. NeurIPS, 2021, pp. 12077–12087. [Online]. Available: https://arxiv.org/abs/2105.15203

[7] S. K. Zhou et al., "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," Proc. IEEE, vol. 109, no. 5, pp. 820–838, May 2021. [Online]. Available: https://doi.org/10.1109/JPROC.2021.3054390

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, Feb. 2019. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2798607

[9] S. L. Topalian, C. G. Drake, and D. M. Pardoll, "Immune checkpoint blockade: A common denominator approach to cancer therapy," Cancer Cell, vol. 27, no. 4, pp. 450–461, Apr. 2015. [Online]. Available: https://doi.org/10.1016/j.ccell.2015.03.001

[10] M. Ghaffari et al., "A review of advancements of artificial intelligence in dentistry," Computers in Biology and Medicine, vol. 147, p. 105792, 2022. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2022.105792

[11] J. Schwärzler et al., "Machine learning versus clinicians for detection and classification of oral mucosal lesions," Oral Oncology, vol. 64, pp. 1–6, 2025. [Online]. Available: https://doi.org/10.1016/j.oraloncology.2025.104361

[12] Y. J. Ye et al., "Utilizing deep learning for automated detection of oral lesions," Computers in Biology and Medicine, vol. 147, p. 105792, 2024. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2024.105792

[13] M. Hesamian et al., "Deep learning techniques for medical image segmentation: Achievements and challenges," Journal of Digital Imaging, vol. 32, pp. 582–596, 2019. [Online]. Available: https://doi.org/10.1007/s10278-019-00227-x

[14] F. Prinzi et al., "Shallow and deep learning classifiers in medical image analysis," Multimodal Technologies and Interaction, vol. 8, no. 2, p. 28, 2024. [Online]. Available: https://doi.org/10.3390/mti8020028

[15] Q. Teng et al., "A survey on the interpretability of deep learning in medical diagnosis," Multimedia Systems, vol. 28, pp. 2335–2355, 2022. [Online]. Available: https://doi.org/10.1007/s00530-022-00960-4