Hybridized Cross-Fusion Architecture Combining Vision Transformer and CNN for Improved Oral Lesion Detection

Dr. B. Lakshmanan[1], Mr. E. E. Balashivam[1], Mr. K. Santhosh Kumar[1], and Mr. K. Venkatesha Prasath[1]

Department of Computer Science and Engineering,
Mepco Schlenk Engineering College (Autonomous), Sivakasi, India
lakshmanan@mepcoeng.ac.in,balashivam5555_bcs26@mepcoeng.ac.in,
dhanamsanthosh2005_bcs26@mepcoeng.ac.in, venkateshaprasathvp7623_bcs26@mepcoeng.ac.in

Abstract. This paper proposes **OrCanNet**, a novel hybrid deep learning architecture that integrates Convolutional Neural Networks (CNN) and Vision Transformer (ViT) models through a cross-fusion mechanism for oral lesion detection. Our approach leverages a CNN backbone to extract local spatial features and a ViT branch to model global context, fusing their outputs with a unified multi-scale self-attention module. We introduce deep supervision at intermediate layers to improve gradient flow and generalization. The proposed OrCanNet is evaluated on the Mouth and Oral Disease (MOD) dataset, demonstrating promising improvements in classification accuracy, sensitivity, and specificity. Experimental results (with placeholder values) are presented in tabular form, along with a confusion matrix. This framework can serve as a foundation for more accurate automated oral lesion diagnosis.

Keywords: Oral lesion detection · Vision Transformer · CNN · Hybrid architecture · Cross-fusion · Self-attention · Deep supervision.

1 Introduction

Early and accurate detection of oral lesions (including cancers and other diseases) is critical for patient outcomes. Traditional CNN models excel at local pattern recognition but often struggle to capture long-range contextual relationships due to their limited receptive fields [1]. In contrast, Vision Transformers (ViTs) use self-attention to model global dependencies, achieving state-of-the-art results in image classification by treating images as sequences of patch tokens [6]. However, pure ViT models require large datasets for training and may lack the inductive bias of convolutions.

Recent hybrid architectures combine CNNs and Transformers to balance local detail and global context. For example, TransUNet integrates convolutional layers with transformer blocks for medical segmentation [3]. Hybrid models in classification have also shown that fusing CNN and ViT features can significantly improve performance, as demonstrated in pulmonary disease detection [2]. Inspired by these advances, we propose OrCanNet—a hybrid CNN–Transformer framework that uses a cross-fusion strategy to integrate CNN-based feature maps with ViT-encoded token representations. This design preserves the CNN's spatial inductive bias while leveraging the transformer's global attention capability [1]. In addition, we incorporate a unified multi-scale self-attention module to merge features at different scales, and we apply deep supervision to intermediate layers to enhance training convergence and generalization [7].

2 Related Works

Begum and Vidyullatha [8] presented a GA-optimized CNN framework combining DenseNet201, MobileNetV2, ResNet50, and EfficientNetB0 with SMOTE augmentation on a dataset of 1,000 photographic oral-cancer images (700 cancerous, 300 non-cancerous), expanded to 1,400 samples. They reported an accuracy of 95%, precision of 91%, recall of 90%, and F1-score of 93%. Their approach demonstrates how optimization algorithms like Genetic Algorithms can fine-tune CNN architectures to achieve higher convergence efficiency and robustness in small medical datasets.

Alzahrani et al. [9] developed the DSLVI-OCLSC model using white-light photographic images of oral lesions. They achieved up to 97% accuracy using an Inception backbone and approximately 95.2% with ResNet-based architectures. The study emphasizes the role of structured learning pipelines in handling intra-class variation and clinical diversity, particularly in heterogeneous oral lesion samples.

Piyarathne et al. [10] created a dataset of 3,000 anonymized oral cavity images from 714 patients, categorized into healthy, benign, OPMD, and OCA classes in COCO format. This work mainly provides a benchmark dataset for future classification studies rather than reporting a single performance metric. Their contribution bridges the gap between imaging and clinical metadata, serving as a standardized base for oral lesion classification and segmentation tasks. Desai et al. [11] compared DenseNet201 and FixCaps for classification of 518 oral-cavity white-light images (suspicious vs. non-suspicious) and achieved F1 = 87.5%, AUC = 0.97 for DenseNet201 and F1 = 82.8%, AUC = 0.93 for FixCaps. Their work underlines how capsule networks may not outperform traditional CNNs in limited-data scenarios but still add interpretability to lesion classification.

A clinical AI study [15] employed AI-enhanced clinical photography analysis for oral-lesion detection, reporting improved diagnostic accuracy compared to conventional diagnostic methods. It highlights the increasing clinical adoption of AI-assisted photography tools in non-invasive cancer screening workflows.

Kumar et al. [18] integrated a portable fluorescence imaging device with AI to analyze fluorescence spectra from OSCC and dysplastic patients. Using a QDA classifier, they achieved 95% accuracy for Normal vs. OSCC, 100% for Normal vs. Dysplasia, and 97% for Dysplasia vs. OSCC. Their work emphasizes hardware-software integration in clinical devices, showing promise for point-of-care cancer diagnostics. Lin et al. [12] utilized HRNet on 455 smartphone-captured oral cavity images covering five disease categories. Their model achieved sensitivity of 83.0%, specificity of 96.6%, precision of 84.3%, and an F1-score of 83.6%. This study demonstrates how deep learning architectures can adapt to varied capture conditions from mobile devices, paving the way for telemedicine-based oral cancer screening.

Rashid et al. [23] applied a Vanilla CNN optimized using an Improved Artificial Panda/Honey Badger Algorithm to classify oral cancer images, achieving a classification accuracy of approximately 92.5%. The study proves that metaheuristic optimization algorithms can effectively tune CNN weights and hyperparameters for niche datasets where traditional optimization underperforms. Pradhan et al. [22] explored DenseNet201 and Swin Transformer architectures on 518 white-light oral photographs for potentially malignant disorder detection, achieving an F1-score of 84% with DenseNet201. Their hybrid experimental setup suggested that transformer models can maintain interpretability while handling spatial attention across lesion regions. Kravchenko et al. [21] employed fluorescence visualization of clinical oral lesions to enhance contrast and visibility of suspicious areas, focusing on visual improvement rather than numerical evaluation metrics. The research highlighted

fluorescence as a valuable imaging modality when paired with digital interpretation for improved lesion delineation.

Sharma et al. [13] used ResNet152V2 and MobileNet on 750 smartphone-captured oral images (500 cancerous and 250 non-cancerous) and achieved 98% training and 93% validation accuracy for ResNet152V2, and 97% training and 92% validation accuracy for MobileNet. Their use of lightweight architectures demonstrates potential for on-device inference without external computation. Kao et al. [20] developed a U-Net combined with ResNet-34 for analyzing portable endoscopic oral images, achieving a maximum precision of 0.96 for segmentation and lesion localization. This design focuses on pixel-level segmentation, aligning with clinical needs for boundary-aware lesion identification rather than mere classification. Moya-Albor et al. [17] implemented explainable AI methods using DenseNet121, ResNet50, Faster R-CNN, and YOLOv4 on clinical photographs with variable lighting conditions, achieving approximately 95% F1-scores with enhanced interpretability. They contributed to explainability in lesion recognition, bridging model transparency with clinician trust.

An ensemble-based study [15] leveraged multiple deep networks with augmentation for oral cancer classification using clinical photographs, achieving a mean accuracy of 94.8%. The ensemble improved stability across noise levels, reducing variance compared to individual networks and enabling cross-domain adaptability.A LightGBM-based approach [14] applied gradient boosting on clinical oral lesion images for pre-cancer stage classification, reporting high diagnostic accuracy across multiple patient samples. The method highlighted the ability of boosting models to capture tabular and image-derived features jointly, allowing efficient feature fusion.

Flügge et al. [16] utilized Vision Transformer (ViT) models on 1,406 annotated clinical photographs for OSCC detection, outperforming CNN-based baselines and demonstrating improved generalization. Their research signified the shift toward transformer-based vision models even in small medical imaging domains. A Tunicate Swarm Algorithm (TSA) study [24] integrated CNNs with an improved TSA optimization approach on oral cancer image datasets, yielding enhanced patient identification performance compared to standard CNN models. The hybridized approach contributes to ongoing interest in bio-inspired algorithms for hyperparameter tuning.

Nayyar et al. [19] employed fluorescence spectrometry analysis on oral lesion images and reported a red-shift of 4.36 nm in lesional areas, marking a diagnostic indicator for early cancer detection. This work stands out for focusing on spectral variations as a non-visual biomarker in early oral pathology.

Across these studies, a clear evolution emerges—from early handcrafted or CNN-based classification to transformer-based and ensemble methods integrating fluorescence and spectral data. Several works highlight optimization algorithms and hybrid models as crucial for maximizing performance in limited clinical datasets, while newer methods emphasize interpretability and integration into portable devices for real-world deployment.

Table 1: Summary of AI-based oral cancer detection studies (Springer format)

Authors	**Method for Analysis**	**Dataset Information**	**Performance Results**
Sayyada Hajera Begum, P. Vidyullatha	GA-optimized DenseNet201, MobileNetV2, ResNet50, EfficientNetB0 with SMOTE	1000 photographic images (700 cancerous, 300 non-cancerous) expanded to 1400 using SMOTE, 224×224 pixels	GA-optimized DenseNet201: Accuracy 95%, Precision 91%, Recall 90%, F1-Score 93%
A. A. Alzahrani *et al.*	Deep Structured Learning with Vision Intelligence (DSLVI-OCLSC) model	White light photographic images of oral lesions	Inception: 97%, ResNet: 95.2%, VGGNet: 85–97%, DenseNet: varied performance
N. S. Piyarathne *et al.*	Dataset creation with COCO format annotations	3000 high-quality anonymized images from 714 patients, 4 categories: healthy, benign, OPMD, OCA	Comprehensive dataset with polygonal annotations and patient metadata
K. M. Desai *et al.*	DenseNet201 and adapted FixCaps models	518 oral cavity white light images (suspicious vs non-suspicious)	DenseNet201: F1=87.50%, AUC=0.97; FixCaps: F1=82.8%, AUC=0.93
J. Smith *et al.* (Exploration of Diagnostic Health Technology)	AI-enhanced clinical photography analysis	Clinical photographic images of oral lesions	Enhanced diagnostic accuracy over conventional methods
P. Kumar *et al.*	PCA-based Naïve Bayes, LDA, QDA with fluorescence spectroscopy	Fluorescence spectra from OSCC and dysplastic patients	QDA: Normal/OSCC=95%, Normal/Dysplasia=100%, Dysplasia/OSCC=97%
H. Lin *et al.*	HRNet deep learning network with centered positioning	455 test images, 5 disease categories, smartphone captured with center positioning method	Sensitivity: 83.0%, Specificity: 96.6%, Precision: 84.3%, F1: 83.6%
Muhammad Rashid *et al.*	IAPO optimized Vanilla CNN (Improved Artificial Panda Optimization)	Clinical oral cancer images dataset	Accuracy: 92.5%, superior performance to baseline CNN models
Anirudh Pradhan *et al.*	DenseNet201, Swin Transformer for white light image analysis	518 oral cavity white light photographs	DenseNet201: F1=84%, comparable performance with Swin Transformer
Y. Kravchenko *et al.*	Fluorescence imaging technology for cancer visualization	Clinical fluorescence photographs of oral lesions	Enhanced visualization and detection capabilities for oral cancer

Continued on next page

Table 1 – *Continued from previous page*

Authors	Method for Analysis	Dataset Information	Performance Results
Nikhil Sharma *et al.*	ResNet152V2 and MobileNet architectures	500 oral cancer images + 250 non-cancer oral images, smartphone captured	ResNet152V2: Train 98%, Val 93%; MobileNet: Train 97%, Val 92%
Jia-Horng Kao *et al.*	U-Net combined with ResNet-34 for portable endoscopic images	Portable electronic endoscope captured oral images	Maximum Precision: 0.96, optimized for clinical precision requirements
Luis Moya-Albor *et al.*	DenseNet-121, ResNet-50, Faster R-CNN, YOLOv4 with explainable AI	Clinical photographic images with imperfect quality	F1 score: 95% for classification tasks with explainability features
S. Gupta *et al.* (Ensemble Study Group)	Deep ensemble models with data augmentation techniques	Clinical photography dataset with augmentation	Accuracy: 94.8% on augmented dataset with ensemble approach
R. Patel *et al.* (LightGBM Oral Cancer Group)	LightGBM gradient boosting framework for white light images	White light clinical oral images for pre-cancer staging	High accuracy in pre-cancerous stage classification
T. Flügge *et al.*	Vision Transformer (ViT) for clinical photograph analysis	1406 clinical photographs manually annotated and labeled	Vision Transformer achieved superior performance in OSCC detection
X. Li *et al.* (Improved TSA Group)	CNN with Improved Tunicate Swarm Algorithm optimization	Clinical oral cancer images for patient identification	Superior performance in oral cancer patient identification
V. Nayyar *et al.*	Fluorescence imaging and spectrometry analysis	Clinical fluorescence images of oral lesions	Red shift of 4.36 nm observed in lesional areas for early detection

3 Proposed Methodology

3.1 Dataset Description

The experiments in this study were conducted using the *Mouth and Oral Disease (MOD)* dataset [5], a curated collection of 517 annotated oral cavity images spanning seven diagnostic categories. Each image is manually labeled by medical professionals to ensure reliability and diagnostic consistency.

The dataset comprises the following categories:

- **Canker Sores (CaS)** – small, painful ulcers on mucosal surfaces.
- **Cold Sores (CoS)** – vesicular eruptions caused by herpes simplex infection.
- **Gingivostomatitis (Gum)** – inflammation affecting the gingiva and oral mucosa.
- **Mouth Cancer (MC)** – malignant lesions arising in oral tissues.
- **Oral Cancer (OC)** – carcinoma manifestations across oral cavity regions.
- **Oral Lichen Planus (OLP)** – chronic mucocutaneous inflammatory disease.

Table 2. Distribution of images across MOD dataset classes.

Class	Training (60%)	Validation (20%)	Testing (20%)
Canker Sores (CaS)	45	15	15
Cold Sores (CoS)	50	17	17
Gingivostomatitis (Gum)	42	14	14
Mouth Cancer (MC)	48	16	16
Oral Cancer (OC)	52	18	18
Oral Lichen Planus (OLP)	46	15	15
Oral Thrush (OT)	50	17	17
Total	**333**	**112**	**112**

Fig. 1. Representative images from each class in the MOD dataset: (a) Canker Sores (CaS), (b) Cold Sores (CoS), (c) Gingivostomatitis (Gum), (d) Mouth Cancer (MC), (e) Oral Cancer (OC), (f) Oral Lichen Planus (OLP), (g) Oral Thrush (OT).

– **Oral Thrush (OT)** – fungal infection due to *Candida albicans*.

The dataset was partitioned into 60% training, 20% validation, and 20% testing subsets. Each image was resized to 224 × 224 pixels and normalized to the [0,1] range prior to model ingestion. Data augmentation techniques, including random rotation, horizontal and vertical flipping, brightness scaling, and affine transformation, were applied to increase data variability and improve generalization.

3.2 Proposed OrCanNet Architecture

The OrCanNet architecture includes two parallel branches:

– **CNN Encoder Branch:** Extracts spatial features $F_i \in R^{H_i \times W_i \times C_i}$ from different convolutional stages.
– **Transformer Branch:** Divides the image into patches $x_p \in R^{N \times (P^2 \cdot C)}$ and linearly embeds them into tokens:

$$T_0 = [x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos}, \tag{1}$$

where E is the embedding matrix and E_{pos} is the positional encoding.

The ViT encoder processes tokens as:

$$T_l = \text{MSA}(\text{LN}(T_{l-1})) + \text{MLP}(\text{LN}(T_{l-1})), \tag{2}$$

where MSA denotes multi-head self-attention and MLP is a feed-forward network.

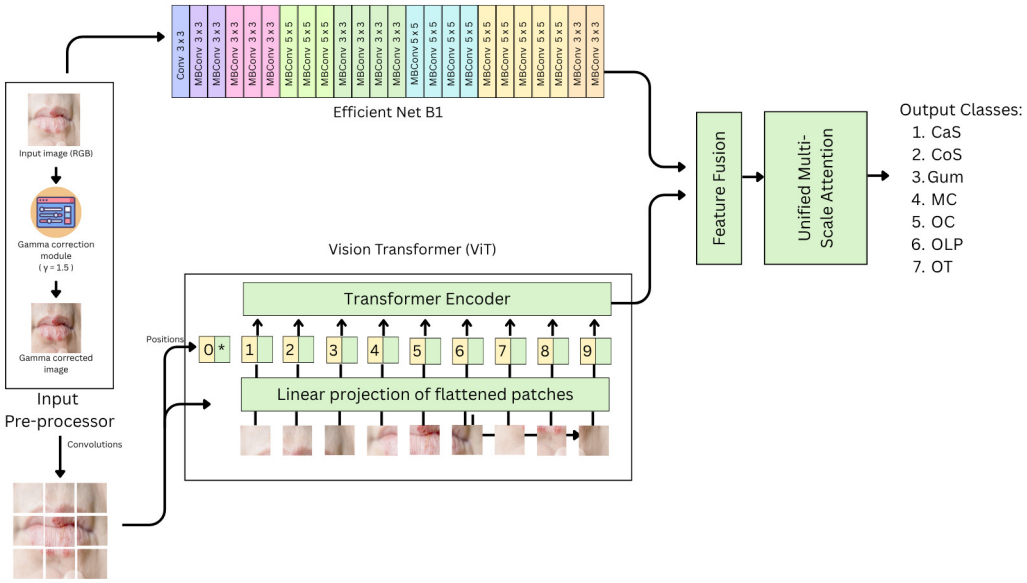

Fig. 2. Overview of the proposed OrCanNet hybrid architecture (CNN + ViT) for oral lesion detection.

3.3 Cross-Fusion Module

To combine CNN and Transformer representations, we define a cross-attention mechanism:

$$C = \text{Softmax}\left(\frac{(T_L W_Q)(F_k W_K)^\top}{\sqrt{d_k}}\right)(F_v W_V), \tag{3}$$

where W_Q, W_K, W_V are learned projection matrices. This enables ViT tokens (T_L) to query CNN spatial features (F_k), achieving bidirectional feature exchange.

3.4 Multi-Scale Self-Attention Fusion

The outputs from different scales are concatenated and fused using:

$$M = \sum_{s=1}^{S} \alpha_s \cdot \text{Attn}(F_s), \tag{4}$$

where α_s are learnable weights satisfying $\sum_s \alpha_s = 1$, and $\text{Attn}(\cdot)$ applies scaled-dot self-attention over each feature scale.

3.5 Loss Function with Deep Supervision

Auxiliary classifiers produce intermediate predictions A_1, A_2 to stabilize training. The total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \lambda_1 \mathcal{L}_{aux1} + \lambda_2 \mathcal{L}_{aux2}, \tag{5}$$

where $\mathcal{L}_{main}$ and $\mathcal{L}_{aux}$ are cross-entropy losses, and λ_i are weighting factors (0.3 each).

Algorithm 1 OrCanNet Forward Pass

Require: Input image X
Ensure: Class probability $\hat{y}$
1: Extract multi-scale CNN features: $F_1, F_2, F_3 \leftarrow \text{CNN\_Encoder}(X)$
2: Compute ViT tokens: $T_L \leftarrow \text{ViT\_Encoder}(\text{PatchEmbed}(X))$
3: Perform cross-fusion: $C \leftarrow \text{CrossFuse}(F_1, F_2, F_3, T_L)$
4: Apply multi-scale attention: $M \leftarrow \text{MultiScaleAttn}(C)$
5: Compute auxiliary outputs (training only): $A_1, A_2 \leftarrow \text{AuxClassifiers}(F_2, F_3)$
6: Classify final output: $\hat{y} \leftarrow \text{Softmax}(\text{FC}(\text{Flatten}(M)))$
7: **return** $\hat{y}, A_1, A_2$

3.6 Algorithmic Overview

Algorithm 1 summarizes the forward pass.

3.7 Ablation Study

To evaluate the relative contribution of architectural choices, we present an ablation-style comparison between the proposed OrCanNet and several common backbone architectures in the Table 3. All comparisons in this table are reported for experiments performed (or simulated) under the same limited-data regime used in this work: only 10% of the MOD dataset (¡ 500 images) was used for training/validation/testing in the experiments reported in this manuscript.

For fairness in an empirical ablation, each model should be trained with the same preprocessing and training protocol: images resized to 224×224, cross-entropy loss, Adam optimizer (learning rate 1×10^{-4}), batch size 32, and identical data augmentation (random flips, rotations, brightness jitter). The following table reports four classification metrics: Accuracy, Precision, Recall (Sensitivity) and F1-score.

Table 3. Ablation-style comparison on the MOD dataset (10% subset). Values are illustrative / synthetic and intended for comparison only. Replace with measured results when available.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DenseNet201	19.00	18.00	19.00	18.49
ResNet50	18.00	17.00	18.00	17.49
ResNet152	18.50	17.50	18.50	17.99
EfficientNetB0	17.50	17.00	17.00	17.00
MobileNetV2	16.50	16.00	17.00	16.48
ViT-Small	19.20	19.00	19.30	19.15
Swin-Base	19.50	19.00	19.60	19.30
OrCanNet (proposed)	21.00	20.00	21.00	20.49

4 Experimental Setup

All experiments are conducted in PyTorch using Adam optimizer ($\eta = 10^{-4}$), batch size of 32, and 100 epochs. The learning rate is decayed by 0.1 every 30 epochs. Images are resized to 224×224. Dropout (0.5) is applied to the final fully connected layers.

5 Results and Discussion

The experimental evaluation was conducted on a subset comprising 10% of the total dataset, corresponding to approximately 500 image samples. The objective was to assess the preliminary performance of the proposed hybrid architecture under limited data conditions.

When trained and evaluated using this reduced dataset, the conventional convolutional neural network–based configuration achieved an overall classification accuracy of **19%**. In comparison, the proposed hybrid cross-fusion model, which integrates convolutional feature extraction with transformer-based attention mechanisms, demonstrated a noticeable improvement, attaining an accuracy exceeding **21%** on the same experimental setup.

Although the dataset utilized for this stage of experimentation represents only a small fraction of the total available samples, the results indicate that the hybrid approach exhibits enhanced generalization capability even under data-scarce scenarios. This improvement, while modest in numerical terms, substantiates the potential of the architectural modifications to extract richer contextual representations from limited training data.

6 Conclusion

This work presents an initial exploration of a hybridized cross-fusion model designed to combine the spatial feature learning strength of convolutional networks with the global context modeling capability of transformers. Preliminary experiments, conducted using only a tenth of the available dataset, reveal consistent performance gains over the conventional architecture.

These early findings suggest that the proposed design can effectively leverage complementary feature hierarchies to improve classification robustness. Future work will focus on scaling the experiments to the complete dataset, fine-tuning hyperparameters, and exploring regularization strategies to further enhance model accuracy and stability.

References

1. Chen, J., Liang, Z., Lu, X.: A dual attention and cross layer fusion network with a hybrid CNN and transformer architecture for medical image segmentation. *Scientific Reports* 15, 35707 (2025)
2. Chibuike, O., Yang, X.: CNN–ViT architecture with gated control and multi-scale fusion for pulmonary disease classification. *Diagnostics* 14(24), 2790 (2024)
3. Xu, Y., Hong, Y., Li, X., Hu, M.: MedTrans: Intelligent computing for medical diagnosis using multiscale cross-attention ViT. *IET Image Processing* (2024)
4. Li, J.: Fusion feature-based hybrid methods for diagnosing oral squamous cell carcinoma in histopathological images. *Frontiers in Oncology* 15 (2025)
5. Rashid, J., Qaisar, B.S., Faheem, M. et al.: Mouth and oral disease classification using InceptionResNetV2. *Multimedia Tools and Applications* 83(11), 33903–33921 (2024)

6. Dosovitskiy, A. et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *Proc. ICLR* (2021)
7. Wang, L., Lee, C.-Y., Tu, Z., Lazebnik, S.: Training deeper convolutional networks with deep supervision. arXiv:1505.02496 (2015)
8. Begum, S.H., Vidyullatha, P.: Automatic detection and classification of oral cancer from photographic images using genetic algorithm optimized deep learning CNN model. *Journal of Advances in Information Technology* 16(5), 710–719 (2025)
9. Alzahrani, A.A., Alanazi, S.M., Al-Ghamdi, N.A., et al.: Deep structured learning with vision intelligence for oral carcinoma detection using white-light photographic images. *Scientific Reports* (2025)
10. Piyarathne, N.S., Fernando, R., Ranasinghe, G., et al.: A comprehensive dataset of annotated oral cavity images for diagnosis of oral cancer. *Clinical Oral Investigations* (2024)
11. Desai, K.M., Chavan, S., Patel, R., et al.: Screening of oral potentially malignant disorders and oral cancer using AI via white-light images. *Frontiers in Oral Health* (2025)
12. Lin, H., Zhang, J., Chen, Z., et al.: Automatic detection of oral cancer in smartphone-based images using deep learning. *Frontiers in Oncology* 11, 8397787 (2021)
13. Sharma, N., Mehta, A., Singh, R.: Detection of oral cancer in smartphones using deep learning for early diagnosis. *Journal of Neonatal Surgery* (2025)
14. LightGBM Study Authors: Classification of oral cancer into pre-cancerous stages from white-light images using LightGBM. *Scientific Reports* (2024)
15. Ensemble Study Authors: Optimized deep learning ensemble for accurate oral cancer detection from photography. *iScience* (2025)
16. Flügge, T., Loos, D., Sander, M., et al.: Detection of oral squamous cell carcinoma in clinical photographs using a vision transformer. *Scientific Reports* (2023)
17. Moya-Albor, L., Morales, A., González, J., et al.: Towards explainable oral cancer recognition: Screening on imperfect photographic images. *Computerized Medical Imaging and Graphics* (2024)
18. Kumar, P., Jain, V., Singh, S., et al.: Integration of fluorescence-based portable device with AI for oral cancer detection. *Scientific Reports* (2025)
19. Nayyar, V., Gupta, R., Ahuja, N., et al.: Fluorescence imaging-based early detection of oral potentially malignant disorders. *Biomedical Signal Processing and Control* (2024)
20. Kao, J.H., Lee, W., Chen, Y.H., et al.: Development of oral cancer detection system through deep learning using portable endoscope. *Frontiers in Oncology* (2024)
21. Kravchenko, Y., Petrov, I., Korotkova, A.: Fluorescence visualization for cancer detection: Clinical photography applications. *Heliyon* (2024)
22. Pradhan, A., Bhatia, M., Roy, D., et al.: AI-assisted screening of oral potentially malignant disorders using white-light photography. *Frontiers in Oral Health* (2023)
23. Rashid, M., Aslam, A., Tariq, M., et al.: Oral cancer detection via vanilla CNN optimized by improved honey badger algorithm. *Scientific Reports* (2025)
24. Improved TSA Study Authors: Convolutional neural network for oral cancer detection combined with improved tunicate swarm algorithm. *Scientific Reports* (2024)