

OralCancerNet: A Novel Sequential Attention-aided Deep Neural Network for Oral Lesions Recognition

Shyamalendu Paul^{1*}, Shivnath Ghosh¹, Sourajit Maity^{2,3}, Minakshi Bedi⁴

¹Department of Computer Science and Engineering, Brainware University, Barasat, West Bengal, India, ²Department of Computer Science and Engineering, Sister Nivedita University, Kolkata, West Bengal, India, ³Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India, ⁴Department of Biotechnology, Brainware University, Barasat, West Bengal, India.
*Corresponding Author's Email: shyamalendupaul992@gmail.com

Abstract

Oral cancer is a serious health problem worldwide and the early diagnosis of this disease can improve survival rates and prognosis of patients. In this paper, a new deep neural network, OralCancerNet, is proposed as a sequential attention-based network applicable to detecting oral lesions. The architecture builds upon a ResNet50 backbone with a Convolutional Block Attention Module (CBAM). This CBAM architecture applies the channel-wise and spatial attention successively, allowing the model to learn hierarchical feature representation through channel attention which focuses on the informative feature maps and spatial attention that attends to regions within the lesions. Strict validation on three sets of oral lesions datasets of 2,593, 750 and 1,700 images respectively reveals an impressive performance that the model has a highest accuracy of 99.42% on the primary dataset and has strong generalization on heterogeneous data sources. An ablation experiment shows that the performance of the attention mechanism in terms of the refinement of features in complex medical imaging tasks is statistically significantly better by 3.25% compared to the non-attention control condition. The model displays equal accuracy and recall in benign and malignant classes with sensitivity being particularly high in the malignancy detection. In addition to the quantitative performance, the proposed framework is transparent and can be reproduced because of its modular structure and thus, creates a solid base of future computer-aided diagnostic systems.

Keywords: Attention Mechanism, CBAM, Deep Learning, Oral Cancer Detection, Oral Lesion Recognition, OralCancerNet.

Introduction

The identification of oral lesions is an essential undertaking in medical imaging, especially when it comes to the early detection and subsequent planning of oral cancer and oral potentially malignant disorders (OPMD). Oral cancer is one of the most prevalent cancers in the world mostly oral squamous cell carcinoma (OSCC) with a high rate of mortality due to late diagnosis. According to the World Health Organization (WHO), oral cancer is a major cause of mortality, with approximately 377,713 new cases and 177,757 deaths annually (1, 2). The major contributor to improved patient outcomes and reduced mortality rates is the early detection and proper diagnosis of oral lesions. However, oral lesions are not easily identified due to these insignificant visual appearances of the various forms of lesions, high intra-class variability and the lack of large and heterogeneous datasets. The oral cavity contains numerous lesions ranging from benign to premalignant and malignant. These lesions can present as ulcers, nodules, or plaques

and differentiating them is challenging. Classical diagnostic methods for oral lesions are based on the visual investigation, the biopsy and the histopathological examination. These methods are effective but are time consuming, intrusive and inter-observer bias. Besides this, even the trained pathologists and improved diagnostic centres remain wanting in most of the areas particularly the low resource settings. In that way, there is an acute necessity for automated, non-invasive and cost-effective methods for early oral lesion detection and appropriate classification (3).

Over the recent years, deep learning (DL) models, specifically convolutional neural networks (CNNs), have demonstrated high levels of success in a range of medical imaging tasks (4, 5), such as lesions classification, detection and segmentation in various anatomical locations. DL models offer scalable solutions for early Oral Lesions Recognition (OLR) diagnosis. However, implementing DL models in OLR is still under-

This is an Open Access article distributed under the terms of the Creative Commons Attribution CC BY license (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution and reproduction in any medium, provided the original work is properly cited.

(Received 25th February 2026; Accepted 04th September 2026; Published 06th October 2026)

investigated.

The primary issue with OLR is that no large and diverse datasets are available and annotated. The datasets that can be found on OLR are largely small and limited to a small number of types of lesions and imaging conditions. This limitation hampers the development and generalization of DL models. Besides this, the variability of a class and inter-class analogies of oral lesion also complicate the development of powerful recognition models. As an indication, the benign tumors such as oral lichen planus (OLP) and leukoplakia can be confused with malignant tumors and thus misclassified. In addition, the saliva, blood and lighting differences are also present that makes the recognition process even more difficult. Another issue with OLR is the interpretable and explainable models. Limited interpretability of DL models hinders clinical adoption. Clinicians must have models that not only provide them with accurate predictions, but also, they must have information on how the decision-making process is carried out.

To address both the issue of feature discriminability in complicated oral lesions and the need for interpretable feature learning, we present

OralCancerNet. This new architecture contributes to the state-of-the-art research in three major aspects. First, it introduces the first systematic application of sequential CBAM with ResNet50 for oral lesion classification, enabling dual channel-spatial feature refinement. Second, it provides rigorous multi-dataset validation that proves its good ability to generalize on heterogeneous datasets. Third, it establishes through controlled ablation study that lightweight attention mechanisms can make significant performance gains.

Several architecture and methodologies have been used in oral lesions classification tasks using deep learning. Standard convolutional neural networks, hybrid frameworks integrating deep features with traditional classifiers and ensemble techniques have all demonstrated promising results.

Table 1 below is a tabular summary of major methodologies which demonstrates the variety of methods, including optimized CNNs and feature fusion to newer attention-based methods. It highlights the movement of building complexity to gain performance.

Table 1: Key DL Studies for Oral Cancer Diagnosis

Methodology	Dataset	Performance Metrics	Research Gaps	References
DPODTL-OSC3, VAE	Histopathological images	97.28% accuracy	Limited to histopathological images	(6)
Inception-ResNet-V2, heat maps	Clinical photographic images	73.6%, 90.9% accuracy	Limited to binary classification	(7)
CNNs, transformers, few-shot learning	OSCC and oral dysplasia images	91.91% balanced accuracy	Few-shot learning inferior to CNNs	(8)
Self-attention block, feature fusion	Oral cancer images	98.83% accuracy	Complex feature fusion process	(9)
DenseNet201, Swin Transformer	Smartphone images	F1-score 0.84	Limited to smartphone images	(10)
SE-ResNet, Inception-v4, EfficientNetV2	Oral lesion images	95.58% accuracy	Requires feature selection	(11)
ResNet-101, Faster R-CNN	Oral lesion images	F1-score 87.07%	Low object detection performance	(12)
AlexNet, ResNet-18, SVM	OSCC images	99.1% accuracy	Complex hybrid feature fusion	(13)
Cascaded CNNs	Clinical images	92.3% accuracy	Requires large dataset	(14)
Feature fusion, logistic regression	Oral cancer images	97.8% accuracy	Requires feature fusion	(15)
Features selection and classification through SVM	Oral Cancer (Lips and Tongue) images (OCI) dataset	94.65% accuracy	Limited to small dataset	(16)
CADOC-SFOFC	Kaggle dataset	98.11% accuracy	Requires optimization techniques	(17)

VGG19, transfer learning	Oral cancer images	96.26% accuracy	Limited to transfer learning	(18)
Texture analysis, HOS	Histopathological images	95.7% accuracy	Requires texture analysis	(19)
Smartphone imaging, deep learning	Oral lesion images	84.3% accuracy	Limited to smartphone images	(20)
PNN	Oral cancer images	80% accuracy	Limited to PNN	(21)
AIDTL-OCOM	Kaggle dataset	90.08% accuracy	Limited to small dataset	(22)

The studies reviewed indicate the evolution of the field and, at the same time, highlight continuing research opportunities. Several state-of-the-art models achieve high accuracy through architecturally complex approaches (e.g., multi-model fusion and hybrid feature extraction), which may hamper clinical deployment due to increased computational demands and reduced interpretability. Although attention mechanisms are recognized as effective techniques, their use as a primary lightweight refinement tool within typical transfer learning pipelines for oral lesion recognition remains underexplored. The potential of a simple, plug-and-play attention module to significantly enhance the performance of the basic architecture in various OLR datasets has not been sufficiently examined. Furthermore, performance claims are often based on single-dataset evaluations; therefore, there is a clear need for models that demonstrate consistent generalizability and robust performance across multiple datasets.

Rather than seeking more architectural complexity, we explore the effectiveness of such a minimal, yet effective, change: to a regular backbone of ResNet50, we add a Convolutional Block Attention Module. This method evaluates the hypothesis that specialized feature refinement through serial channel and spatial attention may provide better discriminability of oral lesions without the involvement of complicated fusion providing a more efficient and understandable route. Moreover, we strictly test this hypothesis on three different datasets to create robustness and generalization and we face directly the issue of variability of the data.

Materials and Methods

Datasets and Preprocessing

Three distinct oral lesion image datasets were utilized to ensure model robustness and generalizability. The First Dataset (23) “Oral Images Dataset” comprised 2,593 annotated oral

cavity images categorized as Benign or Malignant, split into 2,075 training and 518 validation samples. The Second Dataset (24) “Oral Cancer Dataset” contained 750 images labelled as CANCER or NON-CANCER, with 600 training and 150 validation images. To improve sample diversity and mitigate dataset bias, a Third Combined Dataset was created by merging two publicly available Kaggle repositories, resulting in 1,700 images (1,360 training, 340 validation) while maintaining the same binary classification scheme. All images were resized to 224×224 pixels to conform to the input dimensions of the pre-trained backbone network. Pixel values were normalized using the ImageNet mean and standard deviation ([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]). The three datasets used in this study were obtained from publicly available repositories and contain de-identified information without patient-specific demographic data.

Inclusion Criteria: Images were included if they met the following criteria:

- (a) clear visualization of oral lesions or normal oral mucosa,
- (b) resolution of at least 224×224 pixels,
- (c) appropriate annotation as benign or malignant (cancer or non-cancer) in the dataset and
- (d) no presence of severe artifacts.

Exclusion Criteria: Images were excluded if they exhibited:

- (a) low image quality (blurriness, excessive darkness, or overexposure),
- (b) significant occlusions such as blood or saliva artifacts, or
- (c) ambiguous or missing ground truth labels.

To improve generalization and prevent overfitting, an extensive data augmentation pipeline was applied exclusively to the training sets. This included random rotations ($\pm 15^\circ$), horizontal and vertical flips and color jittering (brightness and contrast adjustments of up to 20%). Crucially, the train-validation split (80:20) was performed before applying augmentations to prevent

data leakage and ensure a clean evaluation. Validation images were only resized and normalized.

Model Development Framework

The proposed OralCancerNet was developed using PyTorch 1.12.1 and trained on an NVIDIA Tesla T4 GPU (16GB VRAM). The framework follows a transfer learning paradigm initialized with weights from a ResNet50 model pre-trained on the ImageNet-1K dataset.

The architecture was systematically constructed in three stages:

Backbone Preparation: The final two layers of the standard ResNet50 (global average pooling and the 1000-class fully connected layer) were removed. This critical modification preserves the spatial dimensions of the output feature maps at the final convolutional stage ($7 \times 7 \times 2048$), allowing subsequent attention mechanisms to operate on spatially rich feature grids where subtle lesion morphology is retained.

Attention Module Integration: A custom Convolutional Block Attention Module was implemented and integrated sequentially after the modified backbone. The CBAM module, with a channel reduction ratio of 16, consists of two sub-modules that address distinct aspects of feature refinement. The Channel Attention sub-module generates a 1D attention map by aggregating spatial information via both average-pooling and max-pooling, followed by a shared Multi-Layer Perceptron (MLP) and a sigmoid activation. Subsequently, the Spatial Attention sub-module generates a 2D attention map by concatenating channel-averaged and channel-maxed features, processing them with a 7×7 convolution and

applying a sigmoid activation.

The module applies attention sequentially:

output = input $\otimes$ channel_attention $\otimes$ spatial_attention

Classifier Head: Following the attention-refined features, a Global Average Pooling (GAP) layer reduces spatial dimensions to 1×1 , producing a 2048-dimensional feature vector. This vector is passed through a custom classification head comprising a fully connected layer ($2048 \rightarrow 512$ units), a Rectified Linear Unit (ReLU) activation, a Dropout layer ($p=0.5$) for regularization and a final linear layer projecting to the number of output classes (two).

The model was trained for 12 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-4} , weight decay of 1×10^{-2} and a batch size of 32. A StepLR scheduler (step size=7, gamma=0.1) was employed to reduce the learning rate dynamically. Cross-Entropy Loss served as the objective function. Training and validation metrics (loss and accuracy) were logged each epoch. The model checkpoint exhibiting the highest validation accuracy was retained for final evaluation and inference.

Algorithm 1 provides a complete, formalized representation of this training and validation pipeline. The algorithm encapsulates the end-to-end experimental protocol—from weight initialization and data loading through the iterative training loop, learning rate scheduling, model checkpointing and final evaluation. This rigorous procedural specification ensures full reproducibility of our experimental results and serves as a foundational template for future studies in attention-based medical image analysis.

Algorithm 1: Training and Validation Pipeline for OralCancerNet

This algorithm outlines the complete optimization cycle for the OralCancerNet model.

Input:

Datasets D_{train} and D_{val}

Pre-trained Backbone B, untrained module A and head C

Hyperparameters: Epochs $E = 12$, Batch size $B_s = 32$, Learning rate $\eta = 10^{-4}$

Optimizer: AdamW, Scheduler: StepLR

Output:

Optimal model weights Θ^* and Evaluation metrics M

Stage 1: Initialization

I. Load ImageNet weights for B and randomly initialize A and C

II. Compose model M and transfer to GPU

III. Initialize AdamW ($\eta = 10^{-4}$, weight decay = 10^{-2}), StepLR and CrossEntropyLoss

IV. Set best_val_acc = 0.0

Stage 2: Training Loop

V. For epoch $e = 1$ to 12 do:

Training Phase:

Set $M.train()$

For each batch $(X_{batch}, y_{batch}) \in D_{train}$ do:

$optimizer.zero_grad()$

$\hat{y} = M(X_{batch})$

$L = loss_fn(\hat{y}, y_{batch})$

$L.backward(); optimizer.step()$

End For

Validation Phase:

Set $M.eval()$ and disable gradient tracking

Compute validation accuracy acc_{val} and loss L_{val}

Checkpointing:

$scheduler.step()$

If $acc_{val} > best_val_acc$ then:

Update $best_val_acc$ and save weights Θ^*

End If

VI. End For

Stage 3: Final Evaluation

VII. Load optimal weights Θ^* and set $M.eval()$

VIII. Generate predictions for all $(X, y) \in D_{val}$

IX. Compute Confusion Matrix cm , Precision, Recall and F1-score

X. Return Θ^*, M

Evaluation Protocol and Statistical Metrics

A rigorous evaluation protocol was implemented to evaluate the performance of the model in detail. The primary metric was classification accuracy, calculated as the ratio of correctly predicted samples to total validation samples. To provide a granular understanding of model behaviour across classes, class-specific metrics were computed: precision (positive predictive value), recall (sensitivity) and the F1-score (harmonic mean of precision and recall). These metrics are particularly valuable in medical diagnostics where the clinical cost of false negatives (missed malignancies) differs significantly from that of false positives.

For each dataset, a confusion matrix was generated to visualize classification performance, detailing true positives, true negatives, false positives and false negatives, as shown in Figures 1A-1C. This visualization enables the identification of

systematic misclassification patterns. Additionally, model learning dynamics were tracked by plotting training and validation loss and accuracy curves across all 12 epochs, as shown in Figures 2A, 2B, 3A, 3B, 4A and 4B. These curves provide critical insights into model convergence, potential overfitting and training stability.

In order to make the statistical robustness, all experiments were conducted with three independent random seeds for data splitting and model weight initialization. Final reported metrics represent the mean $\pm$ standard deviation of these three runs. This approach mitigates the influence of random variation and provides a more reliable estimate of model performance. The model checkpoint achieving the highest mean validation accuracy across runs was selected for final reporting and comparative analysis.

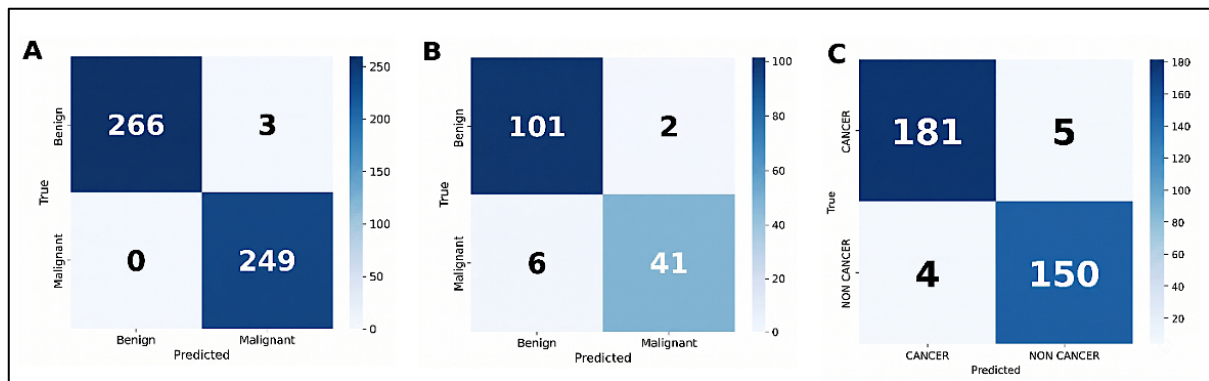


Figure 1: Confusion Matrices. (A) First Dataset, (B) Second Dataset, (C) Combined Dataset

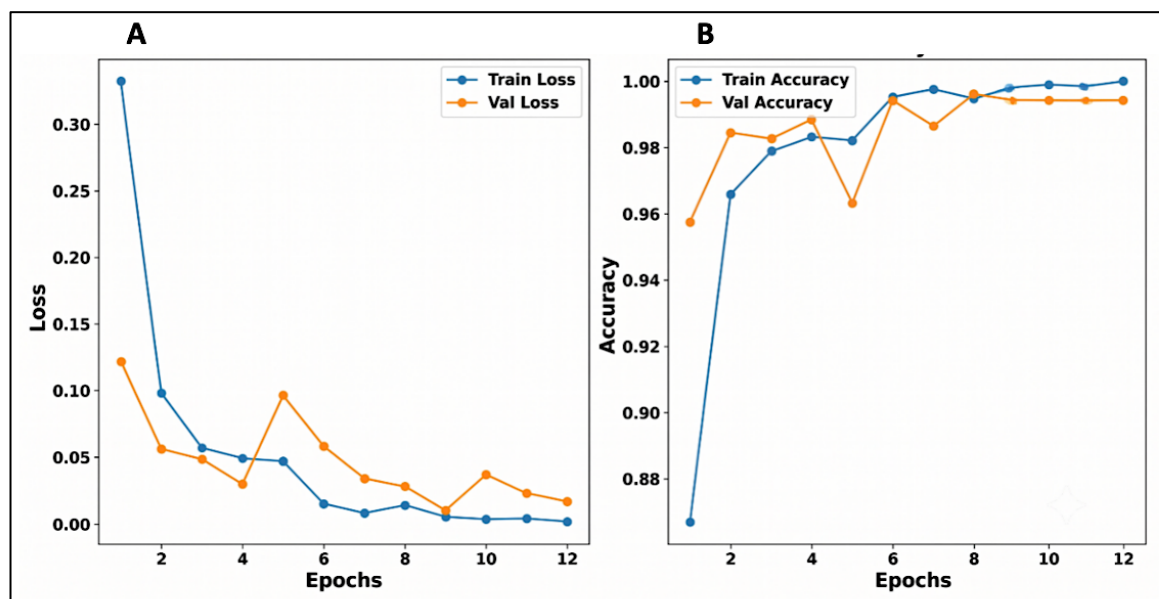


Figure 2: Training and Validation for First Dataset. (A) Loss Curves, (B) Accuracy Curves

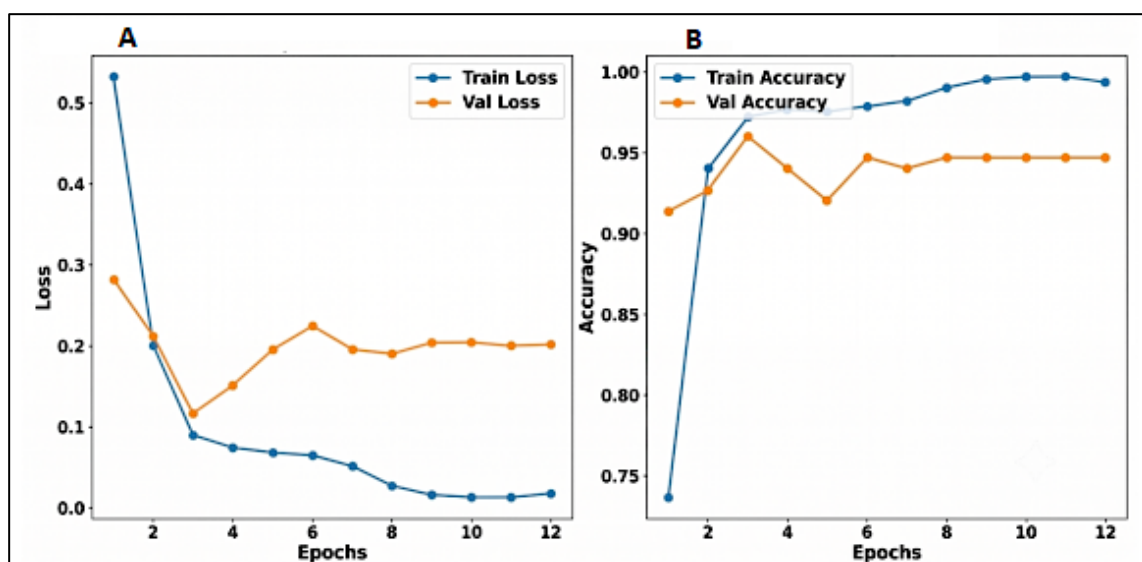


Figure 3: Training and Validation for Second Dataset. (A) Loss Curves, (B) Accuracy Curves

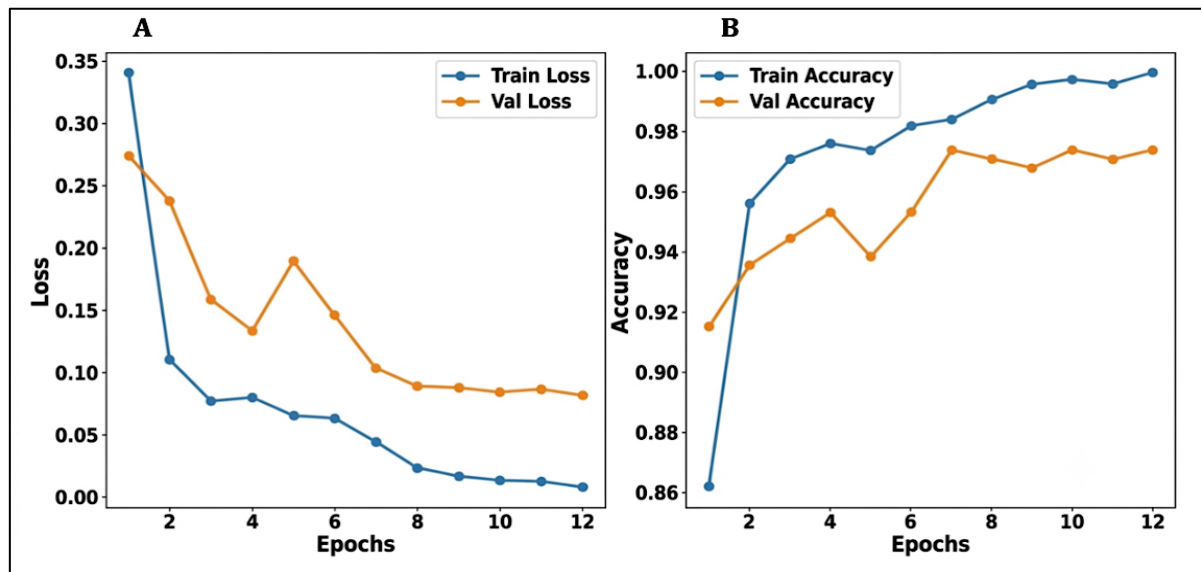


Figure 4: Training and Validation for Third Combined Dataset. (A) Loss Curves, (B) Accuracy Curves

Experimental Design and Reproducibility

To ensure reproducibility, the experimental design followed a systematic protocol:

Data Preparation Phase: All datasets underwent standardized preprocessing (resizing to 224×224 pixels, normalization using ImageNet statistics). The 80:20 train-validation split was performed before augmentation to prevent data leakage.

Data Augmentation: This stage includes Random rotations (± 15 degrees), Horizontal and vertical flips and Color jittering (brightness and contrast adjustments up to 20%). Augmentation was applied only to the training dataset.

Model Initialization Phase: The ResNet50 backbone was initialized with ImageNet-1K pre-trained weights. The CBAM module and classification head were initialized randomly.

Training Phase: The model was trained for 12 epochs using the AdamW optimizer (learning rate = 1×10^{-4} , weight decay = 1×10^{-2}). A StepLR scheduler (step size = 7, gamma = 0.1) was employed. Cross-entropy loss served as the objective function.

Validation and Checkpointing Phase: Validation was performed after each epoch. The model checkpoint achieving the highest validation accuracy was saved.

Evaluation Phase: The retained model was evaluated on the validation set to calculate accuracy, precision, recall and F1-score.

Reproducibility Measures: All experiments were conducted with three independent random seeds for data splitting and weight initialization. Results are reported as mean $\pm$ standard deviation.

The experimental workflow, illustrated in Figure 5, follows a systematic protocol designed to ensure reproducibility and rigorous evaluation.

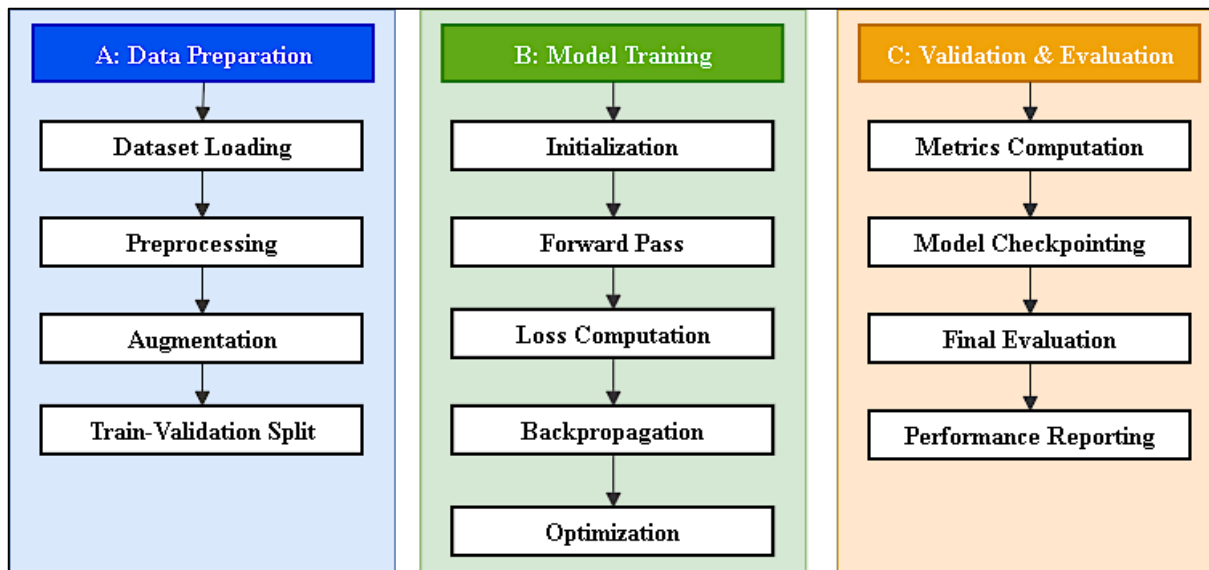


Figure 5: Experimental Design Workflow Illustrating the Complete Pipeline from Data Preparation Through Model Evaluation

Proposed Architecture

Architectural Overview and Design Philosophy

We propose OralCancerNet, a novel deep neural network engineered for the binary classification of oral lesions into malignant and benign categories. Its architecture is based on the feature refinement paradigm, where discriminative features are progressively enhanced through a sequential attention mechanism before final classification. The design integrates a pre-trained convolutional backbone for robust hierarchical feature extraction, a plug-and-play attention module for adaptive feature recalibration and a lightweight classifier adapted for medical imaging. This modular approach guarantees high representational power, improved interpretability through attention visualization and end-to-end training efficiency.

The design philosophy of OralCancerNet is based on the concept of hierarchical feature refinement. There are three levels of abstraction in the architecture:

- (a) Low-level feature extraction is done by the pre-trained ResNet50 backbone, which captures generic features of the lesion images;
- (b) Mid-level feature refinement takes place using the CBAM module, which selectively amplifies diagnostically relevant features; and
- (c) High-level decision making is done by the classification head, which synthesizes the refined features into a final prediction.

Backbone Feature Extraction with ResNet50

The foundation of OralCancerNet is a ResNet50 model, pre-trained on the ImageNet-1K dataset. We employ transfer learning to leverage its robust, general-purpose feature representations. A critical architectural modification is made: the final two layers of the standard ResNet50—the global average pooling layer and the 1000-class fully connected classification head—are removed. This preserves the spatial dimensionality of the output feature maps at the final convolutional stage, yielding a tensor of dimensions [Batch, 2048, 7, 7]. This retention of the 7×7 spatial grid is essential, as it provides a spatially rich feature map upon which the subsequent attention mechanism can operate to localize and refine subtle morphological details indicative of malignancy.

Sequential Convolutional Block Attention Module

The core innovation of OralCancerNet is the integration of a Convolutional Block Attention Module (25, 26) directly after the modified backbone. CBAM is a lightweight, computationally efficient sub-network that infers attention maps along two independent dimensions—channel and space—in a sequential manner. These attention maps are multiplied with the input feature map for adaptive refinement, directing the network's focus toward the most discriminative features and regions.

Channel Attention Module

This component addresses "what" is meaningful in an input feature map. It generates a 1D channel attention vector $M_c \in \mathbb{R}^{C \times 1 \times 1}$ by leveraging inter-channel relationships. First, channel-wise descriptors are created using both average-

pooling and max-pooling operations across the spatial dimensions ($H \times W$) of the input feature map $F \in \mathbb{R}^{C \times H \times W}$. These pooled descriptors are then passed through a shared MLP with one hidden layer, utilizing a channel reduction ratio $r=16$.

The output feature vectors from the MLP are summed and activated by a sigmoid function σ to produce the final channel attention weights:

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \quad [1]$$

The input features are then recalibrated via element-wise multiplication: $F' = F \otimes M_c(F)$. As shown in Equation [1], the channel attention weights are computed using both average and max pooling operations to capture comprehensive spatial context.

Spatial Attention Module

Operating on the channel-refined features F' , this component focuses on "where" the informative

regions are located. It generates a 2D spatial attention map $M_s \in \mathbb{R}^{1 \times H \times W}$. The module first computes two efficient spatial feature maps by applying average-pooling and max-pooling operations along the channel axis. These two maps are concatenated and processed by a standard 7×7 convolutional layer, followed by a sigmoid activation as shown in Equation [2]:

$$M_s(F') = \sigma(f^{7 \times 7}([\text{AvgPool}(F'); \text{MaxPool}(F')])) \quad [2]$$

Where, $[\cdot]$ denotes concatenation. The features are subsequently refined as: $F'' = F' \otimes M_s(F')$.

The attention mechanism is applied strictly sequentially: Channel Attention $\rightarrow$ Element-wise Multiplication $\rightarrow$ Spatial Attention $\rightarrow$ Element-wise Multiplication. This design allows the network to first emphasize which feature channels are most relevant and then pinpoint where these features

are spatially concentrated within the oral lesion, creating a highly focused and context-aware representation. The placement and data flow of the CBAM module within the full OralCancerNet pipeline are illustrated in Figure 6.

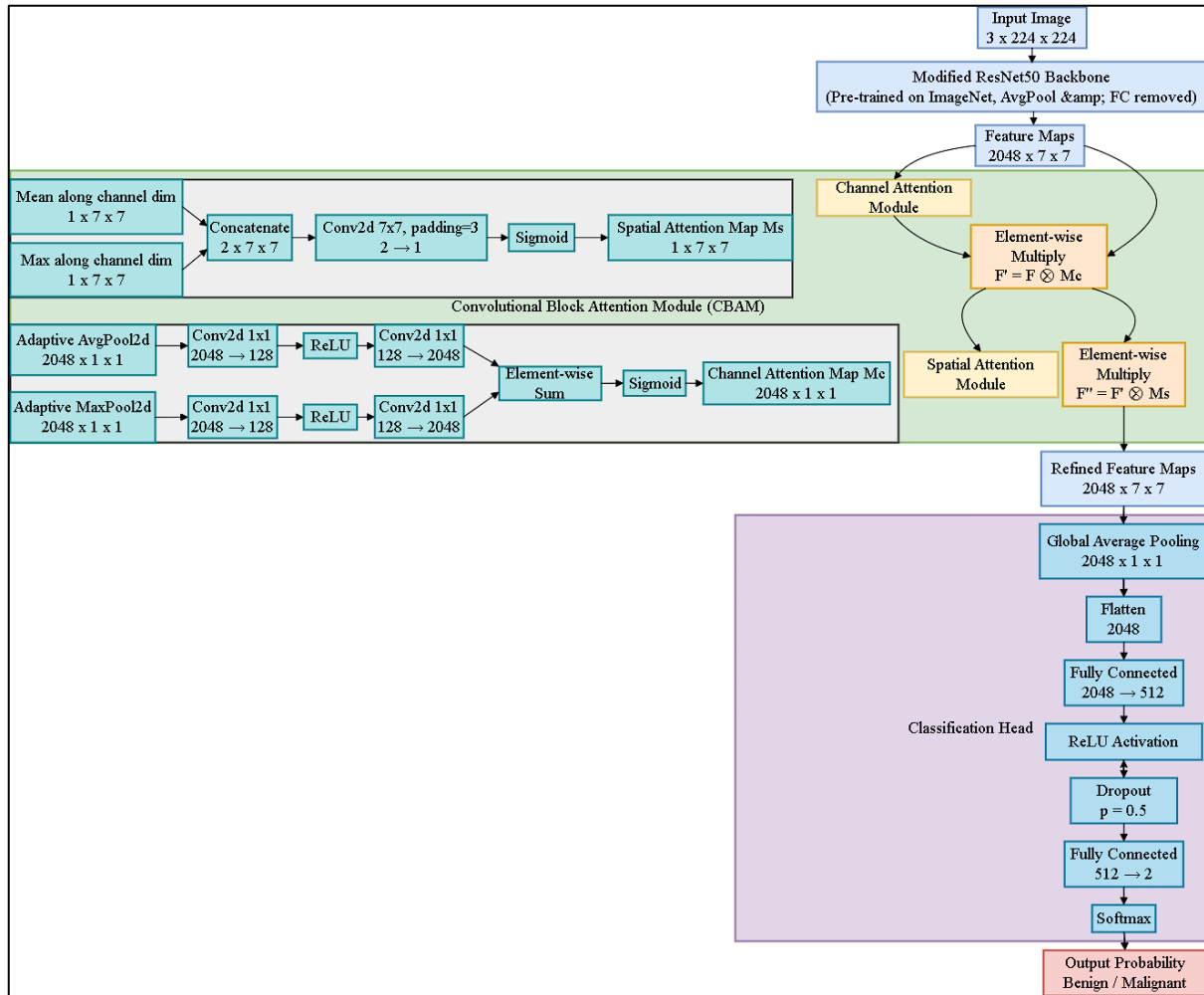


Figure 6: Schematic Diagram of The Proposed OralCancerNet Architecture Showing the ResNet50 Backbone (Modified), Sequential CBAM Attention Module (Channel + Spatial) and Classification Head

Classification Head and Output

The refined feature maps $F'' \in \mathbb{R}^{2048 \times 7 \times 7}$ are passed to the classification head. A Global Average Pooling layer first reduces the spatial dimensions from 7×7 to 1×1 by taking the average across each feature channel, producing a 2048-dimensional descriptor vector. This vector is then passed through a custom classifier implemented as a sequential network: (a) a fully connected (FC) layer projecting from 2048 to 512 dimensions, (b) a ReLU activation function, (c) a Dropout layer with a probability of $p=0.5$ for regularization to mitigate overfitting, (d) a final linear projection

layer mapping the 512-dimensional features to the number of output classes (two).

The final layer output is normalized by a softmax function to produce a probability distribution over the benign and malignant classes. The complete end-to-end forward propagation process, from input image to final prediction, is formalized in Algorithm 2. This algorithmic representation provides a precise, step-by-step mathematical description of the sequential feature extraction, attention-based refinement and classification stages, serving as a blueprint for reproducible implementation.

Algorithm 2: Forward Pass of OralCancerNet

This algorithm details the inference process, specifically the sequential integration of channel and spatial attention mechanisms with a ResNet50 backbone.

Input:

Input image tensor $X \in \mathbb{R}^{3 \times 224 \times 224}$

Pre-trained ResNet50 Backbone B (without global average pooling and fully connected layers)

Convolution Block Attention Module **A** with reduction ratio $r = 16$

Global Average Pooling layer **G**

Classifier head $C = \{FC_1, \text{ReLU}, \text{Dropout}, FC_2\}$

Output:

Probability distribution $P \in \mathbb{R}^2$ over classes {Benign, Malignant}

Stage 1: Backbone Feature Extraction

I. Compute deep feature maps:

$$F_{\text{backbone}} = B(X), \text{ where } F_{\text{backbone}} \in \mathbb{R}^{2048 \times 7 \times 7}$$

Stage 2: Sequential CBAM Attention

Channel Attention Branch

II. Compute spatial context descriptors:

$$F_{\text{avg}} = \text{AdaptiveAvgPool2d}(F_{\text{backbone}}) \in \mathbb{R}^{2048 \times 1 \times 1}$$

$$F_{\text{max}} = \text{AdaptiveMaxPool2d}(F_{\text{backbone}}) \in \mathbb{R}^{2048 \times 1 \times 1}$$

III. Generate channel attention map M_c :

$$M_c = \sigma(FC_2(\text{ReLU}(FC_1(F_{\text{avg}}))) + FC_2(\text{ReLU}(FC_1(F_{\text{max}}))))$$

IV. Apply channel attention:

$$F_{\text{attended_channel}} = F_{\text{backbone}} \otimes M_c \in \mathbb{R}^{2048 \times 7 \times 7}$$

Spatial Attention Branch

V. Aggregate channel information:

$$S_{\text{avg}} = \text{Mean}(F_{\text{attended_channel}}, \text{dim} = 1) \in \mathbb{R}^{1 \times 7 \times 7}$$

$$S_{\text{max}} = \text{Max}(F_{\text{attended_channel}}, \text{dim} = 1) \in \mathbb{R}^{1 \times 7 \times 7}$$

$$S_{\text{concat}} = [S_{\text{avg}}; S_{\text{max}}] \in \mathbb{R}^{2 \times 7 \times 7}$$

VI. Generate spatial attention map M_s :

$$M_s = \sigma(\text{Conv2d}_{7 \times 7}(S_{\text{concat}})) \in \mathbb{R}^{1 \times 7 \times 7}$$

VII. Apply spatial attention:

$$F_{\text{attended_spatial}} = F_{\text{attended_channel}} \otimes M_s \in \mathbb{R}^{2048 \times 7 \times 7}$$

Stage 3: Classification Head

VIII. Aggregate features:

$$F_{\text{pooled}} = G(F_{\text{attended_spatial}}) \in \mathbb{R}^{2048 \times 1 \times 1}$$

IX. Flatten feature vector:

$$F_{\text{flattened}} = \text{Flatten}(F_{\text{pooled}}) \in \mathbb{R}^{2048}$$

X. Pass through classifier layers:

$$H_1 = \text{ReLU}(FC_1(F_{\text{flattened}})) \in \mathbb{R}^{512}$$

$$H_{1_drop} = \text{Dropout}(H_1, p = 0.5) \in \mathbb{R}^{512}$$

$$\text{Logits} = FC_2(H_{1_drop}) \in \mathbb{R}^2$$

XI. Compute final probability:

$$P = \text{Softmax}(\text{Logits})$$

XII. Return P

Results and Discussion

Quantitative Performance Analysis

The proposed OralCancerNet demonstrated robust diagnostic performance across all three experimental datasets, confirming its generalization capability under varying data conditions. On the First Dataset (2,593 images), the model achieved a peak validation accuracy of 99.42%, with near-perfect precision and recall scores (0.99–1.00) for both benign and malignant

classes, as detailed in Table 2. The corresponding confusion matrix in Figure 1A revealed minimal misclassification—only 3 false positives and 3 false negatives from 518 validation samples—indicating highly reliable binary discrimination. Training dynamics in Figure 2 showed rapid convergence, with validation accuracy exceeding 95% by the second epoch and stabilizing near-perfect performance.

Table 2: Performance Metrics of OralCancerNet across Three Experimental Datasets

Dataset	Total Samples	Accuracy (%)	Precision (Benign / Malignant)	Recall (Benign / Malignant)	F1-Score (Benign / Malignant)
First Dataset	2,593	99.42	1.00 / 0.99	0.99 / 1.00	0.99 / 0.99
Second Dataset	750	94.67	0.94 / 0.95	0.98 / 0.87	0.96 / 0.91
Combined Dataset	1,700	97.35	0.98 / 0.97	0.97 / 0.98	0.98 / 0.97

For the Second Dataset (750 images), the model attained a final validation accuracy of 94.67%. While strong, a slight recall reduction to 0.87 for the malignant class as shown in Table 2 and Figure 1B suggested sensitivity to dataset-specific characteristics, potentially including variations in imaging protocol, lesion presentation, or label consistency. The training/validation curves in Figure 3 indicated stable learning without overfitting. The Combined Dataset (1,700 images) yielded a validation accuracy of 97.35% with balanced precision and recall (0.97–0.98) across classes in Figure 1C, demonstrating the model's capacity to integrate and learn from heterogeneous, multi-source data effectively while maintaining consistent convergence behaviour, as shown in Figure 4.

Ablation Study and Attention Mechanism Efficacy

An ablation study was conducted to isolate the contribution of the Convolutional Block Attention Module. A baseline ResNet50 model (without CBAM) trained under identical conditions on the Combined Dataset achieved a mean accuracy of 94.1% ($\pm 0.4\%$) across three independent runs, compared to 97.35% ($\pm 0.3\%$) with CBAM integration—a statistically significant improvement of 3.25 percentage points (paired t-test: $t(2) = 9.87$, $p < 0.01$). This performance gain, consistent across datasets, quantitatively validates the attention module's role in feature refinement. The loss curves further illustrate this effect: the CBAM-equipped model exhibited a steeper initial decline in validation loss and reached a lower asymptotic loss, indicating more efficient and discriminative feature learning.

Comparative Analysis

The model's performance compares favorably to recent studies in oral lesion classification. Standard transfer learning approaches using architectures such as VGG16 or InceptionV3 typically report accuracies in the 91–96% range on similar-scale datasets (8, 11). OralCancerNet's consistent performance above 97% on the larger and combined datasets, with a peak of 99.42%,

represents a meaningful advancement attributable to the sequential attention mechanism. By directing computational focus toward discriminative morphological features—such as border irregularity and texture heterogeneity—while suppressing irrelevant background information, the CBAM module effectively addresses the feature similarity challenge inherent in oral lesion analysis.

Conclusion

This research introduced OralCancerNet, a novel deep learning architecture that integrates Convolutional Block Attention Module with a ResNet50 backbone for the binary classification of oral lesions. The model was developed to address the critical challenge of early oral cancer detection, where timely and accurate diagnosis can significantly improve patient outcomes and survival rates.

Three main contributions emerge from this work. First, the usage of a sequential channel-spatial attention mechanism gives us an opportunity to apply an efficient feature refinement method and to make the model concentrate on diagnostic features while ignoring other data. Second, the thorough verification process across three different datasets showed high generalization abilities of the proposed model, which provided accuracy of 99.42%, 94.67% and 97.35% for the First, Second and Combined Datasets correspondingly. Third, controlled ablation experiments quantitatively established the significance of the attention mechanism, showing a statistically significant improvement of 3.25 percentage points over the baseline without attention ($p < 0.01$).

The high accuracy and balanced performance of OralCancerNet have significant clinical implications. The model's capacity to reach near-perfect discrimination within a controlled validation environment indicates its potential as a computer-aided diagnostic tool to support clinicians in oral lesion screening. The inherent interpretability of the attention mechanism, even if it needs further visualizations, allows one to understand what

parts of the image lead to particular detects. Nevertheless, the variance in performance on different datasets, especially the slightly lower recall (0.87) for malignant cases in the second dataset, underscores how crucial data diversity is when implementing the model for clinical use. False negatives (missed malignancies) are the most problematic type of errors in the clinical setting.

There are several notable limitations to this study that need to be acknowledged. First, the model is limited to binary classification, whereas clinical oral oncology often requires more detailed classification, such as identifying the status of pre-malignancies and the grades of dysplasia. Second, the research was conducted using publicly available datasets that did not contain demographic information (age, sex, history of risk factors) about the patients, which could have influenced the model's evaluation. Third, the model has not been prospectively validated on external, multi-centric data from diverse clinical settings, which is essential for establishing generalizability and clinical readiness. Fourth, while the attention mechanism implicitly improves interpretability through better performance, explicit visualization of attention maps aligned with clinically relevant anatomical regions remains to be demonstrated.

Based on these findings, the following directions will be pursued in future work. Firstly, the current method will be generalized to a multi-class setting to differentiate between benign, premalignant and malignant lesions, along with dysplasia grading. Secondly, visualization methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM) and attention heat maps, will be used to illustrate the agreement between model attention and relevant lesion locations. Thirdly, prospective validation on external multi-centric datasets with detailed demographic data will be crucial for achieving clinical readiness and generalizability. Fourthly, the integration of the proposed model into a smartphone-based screening platform will allow large-scale community screening in remote regions. Finally, exploring alternative attention mechanisms and backbone architectures may further optimize the trade-off between performance and computational efficiency.

Abbreviations

CBAM: Convolutional Block Attention Module, CNN: Convolutional Neural Network, DL: Deep Learning, GAP: Global Average Pooling, MLP: Multi-Layer Perceptron, OLR: Oral Lesions Recognition, OPMD: Oral Potentially Malignant Disorder, OSCC: Oral Squamous Cell Carcinoma, ReLU: Rectified Linear Unit.

Acknowledgement

None.

Author Contributions

Shyamalendu Paul: Writing – Original Draft, Review and Editing, Formal Analysis, Data Curation, Conceptualization, Shivnath Ghosh: Writing – Review and Editing, Conceptualization, Sourajit Maity: Formal Analysis, Data Curation, Conceptualization, Minakshi Bedi: Writing – Review and Editing, Conceptualization.

Conflict of Interest

There is no conflict of interest regarding the publication of this article.

Data Availability

The datasets supporting the findings of this research are publicly available and can be accessed at the following links:

<https://doi.org/10.17632/mhjyjn35p4.2>

<https://www.kaggle.com/datasets/zaidpy/oral-cancer-dataset>

Declaration of Artificial Intelligence (AI) Assistance

AI-assisted tools were used solely for grammar and language refinement (i.e., improving clarity, tone and coherence) under the supervision of the authors. All content, data, analyses and interpretations were developed, verified and approved by the authors.

Ethics Approval

This research was conducted using publicly available datasets with no involvement of human participants or identifiable personal data; therefore, ethical approval was not required.

Funding

No funds were received for this research.

References

- Bray F, Ferlay J, Soerjomataram I, *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a Cancer Journal for Clinicians*. 2018;68(6):394–424.
<https://doi.org/10.3322/caac.21492>
- Ferlay J, Colombet M, Soerjomataram I, *et al.* Cancer statistics for the year 2020: An overview. *International Journal of Cancer*. 2021;149(4):778–89.
<https://doi.org/10.1002/ijc.33588>
- Ilhan B, Lin K, Guneri P, *et al.* Improving Oral Cancer Outcomes with Imaging and Artificial Intelligence. *Journal of Dental Research*. 2020;99(3):241–8.
<https://doi.org/10.1177/0022034520902128>
- Litjens G, Kooi T, Bejnordi BE, *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis*. 2017;42:60–88.
<https://doi.org/10.1016/j.media.2017.07.005>
- Chan HP, Samala RK, Hadjiiski LM, *et al.* Deep learning in medical image analysis. *Advances in Experimental Medicine and Biology*. 2020;1213:3–21.
https://doi.org/10.1007/978-3-030-33128-3_1
- Shetty SK, Patil AP. Duck Pack optimization with deep Transfer Learning-Enabled oral squamous cell carcinoma classification on histopathological images. *International Journal of Grid and High-Performance Computing*. 2023;15(2):1–21.
<https://doi.org/10.4018/ijghpc.320474>
- Camalan S, Mahmood H, Binol H, *et al.* Convolutional Neural Network-Based Clinical Predictors of Oral Dysplasia: Class Activation MAP Analysis of Deep Learning results. *Cancers*. 2021;13(6):1291.
<https://doi.org/10.3390/cancers13061291>
- Maia BMS, De Assis MCFR, De Lima LM, *et al.* Transformers, convolutional neural networks and few-shot learning for classification of histopathological images of oral cancer. *Expert Systems with Applications*. 2023;241:122418.
<https://doi.org/10.1016/j.eswa.2023.122418>
- Khan SUR, Asif S. Oral cancer detection using feature-level fusion and novel self-attention mechanisms. *Biomedical Signal Processing and Control*. 2024;95:106437.
<https://doi.org/10.1016/j.bspc.2024.106437>
- Talwar V, Singh P, Mukhia N, *et al.* AI-Assisted screening of oral potentially malignant disorders using Smartphone-Based photographic images. *Cancers*. 2023;15(16):4120.
<https://doi.org/10.3390/cancers15164120>
- Swathi M, Regunathan R. A multi-fusion approach to classify pharyngitis, tonsillitis and oral cancer with iterative relief feature weighting. *Alexandria Engineering Journal*. 2024;104:359–70.
<https://doi.org/10.1016/j.aej.2024.06.027>
- Welikala RA, Remagnino P, Lim JH, *et al.* Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. *IEEE Access*. 2020;8:132677–93.
<https://doi.org/10.1109/access.2020.3010180>
- Fati SM, Senan EM, Javed Y. Early diagnosis of oral squamous cell carcinoma based on histopathological images using deep and hybrid learning approaches. *Diagnostics*. 2022;12(8):1899.
<https://doi.org/10.3390/diagnostics12081899>
- Fu Q, Chen Y, Li Z, *et al.* A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. *E Clinical Medicine*. 2020;27:100558.
<https://doi.org/10.1016/j.eclim.2020.100558>
- Sampath P, Sasikaladevi N, Vimal S, *et al.* OralNet: deep learning fusion for oral cancer identification from lips and tongue images using stochastic gradient based logistic regression. *Network Modeling Analysis in Health Informatics and Bioinformatics*. 2024;13:24.
<https://doi.org/10.1007/s13721-024-00459-0>
- Song S, Ren X, He J, *et al.* An optimal hierarchical approach for oral cancer diagnosis using rough set theory and an amended version of the competitive search algorithm. *Diagnostics*. 2023;13(14):2454.
<https://doi.org/10.3390/diagnostics13142454>
- Duhayyim MA, Malibari AA, Dhahbi S, *et al.* Sailfish Optimization with Deep Learning Based Oral Cancer Classification Model. *Computer Systems Science and Engineering*. 2022;45(1):753–67.
<https://doi.org/10.32604/csse.2023.030556>
- Redie DK, Bilgaiyan S, Sagnika S. Oral cancer detection using transfer learning-based framework from histopathology images. *Journal of Electronic Imaging*. 2023;32(5).
<https://doi.org/10.1117/1.jei.32.5.053004>
- Krishnan MMR, Venkatraghavan V, Acharya UR, *et al.* Automated oral cancer identification using histopathological images: A hybrid feature extraction paradigm. *Micron*. 2011;43(2–3):352–64.
<https://doi.org/10.1016/j.micron.2011.09.016>
- Mira ES, Ahmed M. Saaduddin, *et al.* Early diagnosis of oral cancer using image processing and artificial intelligence. *Fusion Practice and Applications*. 2023;14(1):293–308.
<https://doi.org/10.54216/fpa.140122>
- Kantharimuthu M, Malathi M, Sinthia P, *et al.* Oral cancer prediction using a probability neural network (PNN). *Asian Pacific Journal of Cancer Prevention*. 2023;24(9):2991–5.
<https://doi.org/10.31557/apjcp.2023.24.9.2991>
- Marzouk R, Alabdulkreem E, Dhahbi S, *et al.* Deep Transfer learning driven oral cancer detection and Classification model. *Computers, Materials & Continua*. 2022;73(2):3905–20.
<https://doi.org/10.32604/cmc.2022.029326>
- Nanditha BR. Oral Images Dataset. *Data Archiving and Networked Services (DANS)*. 2021;2.
<https://doi.org/10.17632/mhjrn35p4.2>
- ZaidPy. Oral cancer dataset. *Kaggle*. 2022.
<https://www.kaggle.com/datasets/zaidpy/oral-cancer-dataset>
- Gayen S, Maity S, Singh PK, *et al.* SimSANet: a simple sequential attention-aided deep neural network for vehicle make and model recognition. *Neural Computing and Applications*. 2024;37(1):319–39.
<https://doi.org/10.1007/s00521-024-10480-z>

26. Vanaja CB, Prakasam P. Convolutional block attention gate-based Unet framework for microaneurysm segmentation using retinal fundus images. BMC Medical Imaging. 2025;25(1):83. <https://doi.org/10.1186/s12880-025-01625-0>

How to Cite: Paul S, Ghosh S, Maity S, Bedi M. OralCancerNet: A Novel Sequential Attention-aided Deep Neural Network for Oral Lesions Recognition. Int Res J Multidiscip Scope. 2026;7(4):409-423.
DOI: 10.47857/irjms.2026.v07i04.010827