International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(13s) 2026 pp. 523-545
https://doi.org/10.70917/ijcisim-2026-4093
Article

Dual-Attention MobileViT++: A Multi-Scale CNN–Transformer Framework for High-Precision Oral Cancer Classification

Srushti Patidar[1] , Parmalik Kumar[2] , Neha Gupta[3]

[1]Department of Computer Science and Engineering, Oriental University, Indore, M.P, India, srushtipatidar2927@gmail.com
[2]Department of Computer Science and Engineering, Oriental University, Bhopal, M.P, India, parmalik83@gmail.com
[3]School of Computer Science Information Technology, Symbiosis University of Applied Sciences, Indore, M.P, India, neha.gupta@suas.ac.in

Abstract: Oral cancer is one of the life-threatening diseases with a high mortality rate resulting from late diagnosis and limited access to expert screening due to specialized training. The available detection approaches, which are mainly based on visual observations and biopsied analysis, are invasive, time-consuming, and experience dependent. Therefore, a reliable and automated approach is needed to develop an efficient detection system. This study introduced the Dual-Attention MobileViT++ architecture, which integrates the multi-scale depthwise CNN backbone, the hybrid attention models, and the hierarchical MobileViT transformers to characterize the local lesion texture and global contextual patterns. The model was tested on the Kaggle public dataset with no additional modification and enhancement. The oral cancer dataset was modified through medical-grade augmentation to create more balanced datasets comprising about 981 images. Experimental results showed a promising performance, achieving approximately 99.93% accuracy with high precision, recall, and F1-score performed on the validation and test datasets. This result shows that the proposed architecture can generalize well and deliver reliable classification, suitable for oral cancer screening at an early stage.

Keywords: Oral cancer detection, deep learning, MobileViT, dual attention, medical image classification, CNN–transformer hybrid, clinical decision support.

1. Introduction

Oral cancer, especially oral squamous cell carcinoma (OSCC), is an important worldwide health problem and represents most head and neck cancers. Followed by high morbidity and mortality, particularly in late stages of disease. Immunocytes discovery can contribute to early detection, which is an important leading factor in increasing survival rates as well as treatment efficacy. However, conventional diagnostic approaches, primarily clinic examination for biopsy determination, are invasive, time-consuming and quite dependent on trained expert knowledge [1]. In such areas, especially in low-resource settings, inadequate access to trained health care professionals and diagnostic facilities delays accurate diagnosis even more, leading to poor outcomes for patients [2].

Recent years have seen great potential for intelligent medical imaging diagnosis in the development of artificial intelligence (AI) and deep learning. CNN achieved the plate to detach the oncological information (texture and color) of clinical images, which have shown promising in automatic diagnosis of oral lesion [9]. However, typical CNNs models don't have effective abilities to encode long range spatial dependencies and global context information. To solve this, models have been proposed based on transformers which are good at learning global information in images [4]. Hybrid CNN–transformer models have consequently become popular solution., as local obstructiveness and global context awareness can be combined to improve diagnostic accuracy [5].

Despite these developments, there are still many challenges. Most of these models are based on single-stage attention mechanism or insufficient feature aggregation, leading to low sensitivity for surface and early-stage lesion detection. However, many medical related tasks are challenged by small dataset sizes and class imbalance that causes

overfitting and no trustworthy predictions [6]. These constraints emphasize that a more powerful multi-scale lesion-aware architecture with higher efficiency and generalization ability is necessary.

The main contribution of this work is to create a reliable and efficient deep learning framework for accurate classification of oral cancer using clinical images. This work is motivated to alleviate the shortcomings of models by incorporating multi-scale feature extraction, dual attention mechanisms and hierarchical transformer modules. By integrating these parts, well designed integration can be processed to achieve greater sensitivity of detecting early

abnormal mucosa, increase the global muco-sal context awareness and more reliable pathological prognosis on cancerous suspect.

Key Contributions

The key contributions of this study are as follows:

1. Dual-Attention Architecture: A new hybrid model that integrates channel–spatial attention and efficient channel attention for improving the lesion-specific information in different network depths.

2. Multi-Scale Residual CNN Backbone: A light depthwise-separable CNN architecture with residual connections to preserve fine texture details and partially recover computational efficiency.

3. Hierarchical MobileViT Integration: Two-stage MVITs to capture both the mid-level and global contextual information for better classification.

4. Improved Training Strategy: A novel hybrid loss that joins the focal loss, supervised contrastive learning for better class separation and robustness.

5. Strict Cross-Validation Procedure: Stratified cross validation, external testing and statistical analysis to guarantee performance is robust and reproducible.

The rest of this paper is structured as follows. Section 2 discusses related works of the recent deep learning methods on oral cancer detection. The proposed dual-attention MobileViT-based approach is detailed in Section 3. Section 4 describes the implementation details, including hardware, software, dataset and training settings. The experimental results and comparative performance study are given in Section 5. Section 6 concludes the paper and discusses future work.

2. Literature Review

Oral squamous cell carcinoma (OSCC) is one of the most frequent, yet fatal oral malignancies in the world and has been ranked as one of the major causes of cancer-related deaths. Even with treatment improvements, survival is poor given that many cases are diagnosed at late stage. Early detection is crucial to patient prognosis, but current diagnosis methods such as biopsies usually require invasive procedures, are time-consuming and subject to expert interpretation. To overcome these limitations, more recently developed deep learning methods such as MaskMeanShiftCNN and SV-OnionNet similarly propose to analyze histopathological images by estimating color, texture, and contour-based features. As compared with previous methods, our models have shown better accuracy, sensitivity and specificity between tumour regions and non-tumour epithelium, indicating the promising future of automatic system for finding OSCC [1].

In addition to histopathology-based studies, more multimodal images were introduced that consider patient metadata along with clinical images. Conventional single modality imaging generally is less reliable, particularly in discerning benign from potentially malignant lesions. A multi-model deep learning pipeline, with included image level features and patient information, was superior for classification performance, in which MobileNetV3-Large was the best pre-trained model. These approaches also show how patient specific knowledge can aid in making diagnosis more accurate and better support clinical decision [2]. However, the impact of deep learning is more decisively shown through systematic reviews on oral cancer detection studies, which report accuracy of 100% in various architectures and datasets. These results show that deep learning can greatly promote lesion classification, segmentation and prognosis prediction [3].

Quickly, artificial intelligence has become one of the transformative technologies for oral cancer detection; and allows automatic screening that is more consistent and faster. Aspects of AI-based diagnostic tools are found to have utilized deep learning, conventional machine learning, fuzzy logic, and genetic algorithms on distinct imaging

modalities. Of all those, deep learning-based techniques occupy a substantial portion in recent studies, which has reflected their increasing popularity within the context of automatic diagnostic. In general, it has been reported that AI systems have the potential to improve diagnostic accuracy and to decrease variability between clinicians [4]. On the other hand, comparisons between different machine learning methods show issues related to variable datasets and performance measures. Although the sensitivity reported in most studies have been greater than 85% and accuracies higher than 90%, it is difficult to compare outcomes directly due to absence of standard test data. To overcome this

problem, advanced approaches including attention mechanisms, multitask learning and fusion strategies have been elaborated to enhance the robustness and uniformity in OSCC delineation [5].

Segmentation of oral lesions, and other 'segments based on shape and texture have become more important in CAAAD since precise boundaries of lesion help treat competitively, planning therapists and monitoring the disease. A supplemented Reformer model learnt on labeled oral mucosal disease images achieved better results than the existing segmentation models including U-Net and DeepLabV3. The method reached higher Dice and mIoU scores, being capable of accurately distinguishing between erosive and blistering lesions, thus facilitating clinical practice [6]. It is also relative to the effectiveness of attention-based feature refinement for early OSCC detection [7] because there are some literatures presented impressive accuracy and F1-score with the extended EfficientNetB0 models combined with dynamic attention network in histopathological images research.

Feature fusion and transfer learning methods have also been investigated for enhancing the predictability. A self-attention based fusion model of EfficientNetB0 and EfficientNetB1 achieved an accuracy of 98.83% for oral disease datasets, which addressed the inconsistency problem from trusted CNN models. This method emphasizes the need of combining complementary feature representations to achieve classification stability while preserving efficiency [8]. Further literature reviews on automated oral cancer detections also support the adoption of segmentation and classification pipelines based on capsule networks, Bayesian deep learning, and optimized CNN architectures. Performance evaluation in terms of Dice, Jaccard, precision, and sensitivity shows very promising potential for clinical deployment, yet some refinements and standardization are needed [9].

Improvement in conventional imaging as well as development of advanced imaging have also been investigated to improve the diagnostic accuracy. Hyperspectral imaging combined with deep-learning models has also demonstrated some potential in OSCC detection. A hybrid model based on local binary pattern features, VGG-16 and deep CNN achieved an accuracy of >96% associated with a high sensitivity and specificity. These results indicate the value of multimodal imaging and hybrid structure for better diagnostic performance [10].

Nevertheless, beyond being an indirect cancer detection endpoint, disease burdens in oral mucosal lesions serve as important background information for screening and diagnosis[11]. A review on lesion-targeted therapies reported that etiology-driven procedures result in better response and less recurrence, which stresses the urgency of combined care and customization ways [12]. In addition, age- and sex-specific variations in lesion prevalence are also present: the prevalence of some of these lesions increases with age, while others are more common at younger ages [13].

Recent research has also investigated the diagnostic power of deep learning with respect to clinicians. A YOLOv8-based model trained on intraoral images that involve multiple types of disease obtained performance near that (but also better than) oral surgeons, and better than general dentists as significantly improved. These results indicate that the deep learning models can be relied upon as clinical assistive tools for assessing oral lesions [14]. Meta-analyses of AI-aided imaging systems in oral cancer detection additionally revealed pooled sensitivity and specificity of 0.919 and 0.879 values respectively, with AUC estimates reaching ~0.98 indicative of the potential scalability of DL-supported screening strategies especially for underprivileged communities [15].

Clinical alertness and diagnostic steps also affect the early detection results. Investigation of several outpatient populations shows a high number of suspicious oral lesions versus low biopsy confirmation rates, underscoring deficits in patient knowledge and examination follow-up [16]. The review of diagnostic algorithms for oral soft-tissue lesions also reveals that most reported algorithms have not been validated, with coverage limited to only few lesions [17], reinforcing the necessity of more reliable and clinically test diagnostic tools.

Oral lesion prevalence and risk factors have continued to be elucidated at the population level. Cross sectional-based studies have reported a substantial burden of tobacco related lesions and potentially malignant disorders particularly in the older age groups [18]. In fact, some population-based studies have also demonstrated an association between the prevalence of lesions and level of education, smoking status, and history of other types of cancer which reflects the role played by socioeconomic and behavioral factors in shaping oral health outcomes as well [19].

Uncommon but highly aggressive neoplasms such as oral mucosa melanomas also accentuate the necessity for better prognostication models, limiting survival based on size, nodal status and site of origin [20].

Object detection-based and triage-oriented deep learning models have also been established for preliminary diagnosis. Large clinical image datasets have been used to train YOLO based models which produce high precision and recall in benign, potentially malignant and malignant classes. Adding attention mechanisms leads to even higher detection

accuracy, indicating the importance of spatial awareness in deep learning models for clinical triage [21]. Similarly, disease and treatment associated with oral complications such as radiotherapy induced microbial changes highlight the need for comprehensive oral care and monitoring during cancer therapy [22].

Other epidemiological studies show differences in lesion distribution by sex, age group and systemic diseases, also indicating the heterogeneity of patterns of oral mucosal diseases [23]. The reported reviews of oral mucosal toxicities of cancer therapies clearly points out the requirement for supportive measures that would also not cause compromising anticancer efficacy, thus making precision focused on oral care a necessity [24].

Recent deep learning-based models that used CNN combined with transformer-based contextual models that performed well for clinical oral cancer datasets, outperforming many popular networks. These lightweight models appear to be viable for the deployment of AI-based screening workflows in real-world clinical care [25]. Indeed, worldwide epidemiological studies also verify the increasing incidence of lip and oral cavity tumors, especially in particular areas domestically that argue for better screening and preventive interventions [26]. Alcohol intake and other lifestyle factors still play a substantial role in the global burden of disease with geographic and gender differences [27].

Temporal trend analyses of the incidence of oral cancer show increase, predominantly in younger age-groups and certain areas, suggestive of a demographic "drift" in the pattern of this disease [28]. Oral complications because of the cancer disorder continue to be prevalent with treatment and necessitate closely coordinated multidisciplinary care by oral health practitioners [29]. The Management of oral cancer When it comes to oral cancer comprehensive reviews highlight the importance of early diagnosis, proper staging and multimodal approaches in management to improve patient's outcomes [30].

At an international level, research organizations promote comprehensive oral health research agendas in conjunction with worldwide health strategies concerned with better prevention, detection and treatment [31]. However, differences in the quality of care remain among the socioeconomic regions despite improvements in global care indices [32]. Here, artificial intelligence and precision medicine represent promising directions of prospective use that might help for early diagnosis, personalized treatment, marker characterization though those would have to be solved in term of privacy, bias or transparency regarding a safe deployment [33].

3. Methodology

3.1 Proposed architecture

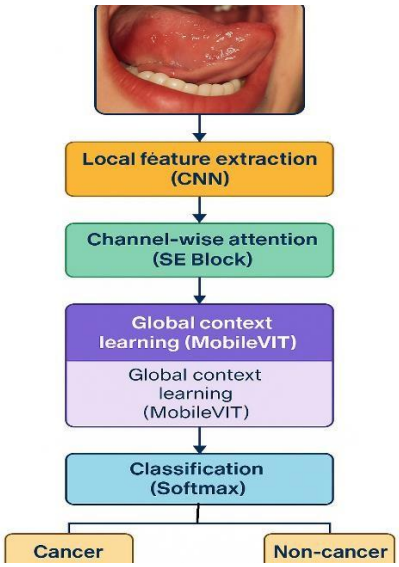

Figure 1: Proposed CNN–SE–MobileViT Architecture for Oral Cancer Classification

Figure 1 provides an overview of the design of the deep learning model for oral cancer diagnosis. The oral image input is initially managed with a CNN to obtain local spatial features. The convolution is given by:

$$F^{k}_{i,j} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X_{i+m,i+n} \cdot W^{k}_{m,n} + b^{k}$$

where □is the input image, □□is the convolution kernel for the □□*h*feature map, □□is the bias term, and

□□ represents the output feature at location (□, □).

The features are then forwarded to a Squeeze-and-Excitation (SE) block to learn channel-wise attention. The squeeze operation aggregates global context which is obtained through the global average pooling process and the following enhances its discriminative ability:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i,j)$$

where □□(□, □)is the activation at spatial location (□, □)of channel □, and □and □represent height and width of the

feature map.

The excitation stage generates channel weights using two fully connected layers:

$$s = \sigma\left(W_2 \cdot \delta\left(W_1 \cdot z\right)\right)$$

where □is the squeezed vector, □1and □2are weight matrices, □is the ReLU activation, and □is the sigmoid function.

The reweighted feature map is then processed by the MobileViT module to learn global contextual relationships. The final classification is performed using the softmax function:

$$P(y = k \mid x) = \frac{e^{z_k}}{\sum_{i=1}^{C} e^{z_i}}$$

where □□is the logit for class □, and □is the total number of classes (cancer and non-cancer).

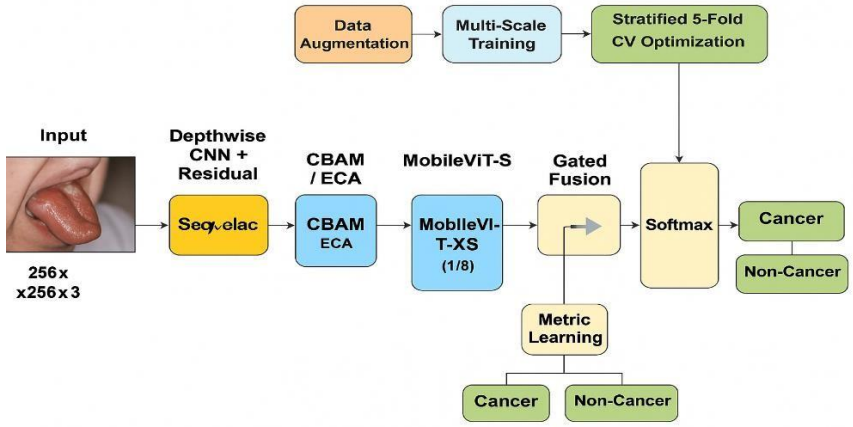

Figure 2: Multi-Scale Attention-Based MobileViT Fusion Architecture

This figure presents enhanced architecture incorporating depth wise CNN layers, attention modules, MobileViT blocks, and gated fusion. The depth wise convolution separates spatial and channel filtering to reduce computational complexity. The depth wise convolution operation is expressed as:

$$Y_{i,j,c} = \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} X_{i+m,j+n,c} \cdot W_{m,n,c}$$

where □is the input tensor, □is the depthwise kernel, and □represents the channel index.

Attention mechanisms such as CBAM or ECA are used to refine features. The channel attention in CBAM is computed as:

$$M_c(F) = \sigma\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right)$$

where □is the input feature map, □□□is the shared multilayer perception, and □is the sigmoid activation.

Feature fusion combines CNN and MobileViT outputs using a gated mechanism:

$$F_{fused} = \alpha \cdot F_{CNN} + (1 - \alpha) \cdot F_{ViT}$$

where □□□□and □□□□ represent features from CNN and MobileViT branches, respectively, and □is the learned gating coefficient.

Finally, classification is performed using the softmax function:

$$P(y = k) = \frac{e^{z_k}}{\sum_{i=1}^{C} e^{z_i}}$$

where □□denotes the class score.

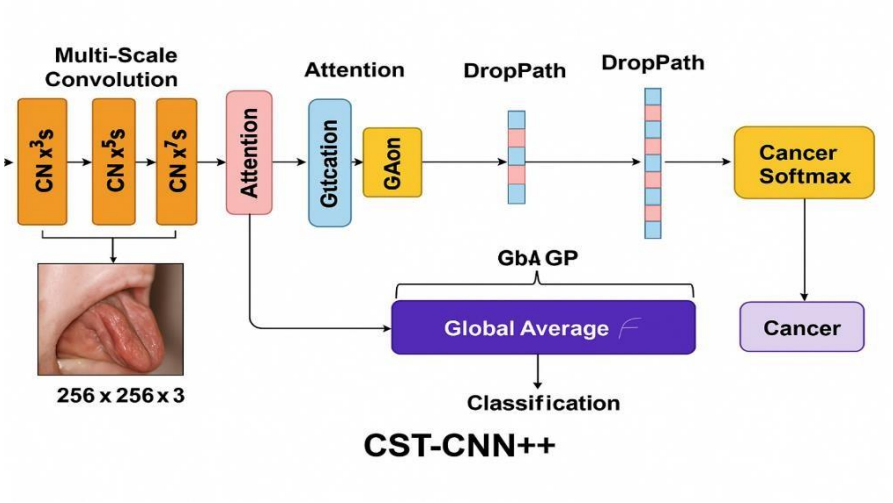

Figure 3: CST-CNN++ Multi-Scale Attention Classification Model

Figure 3 depicts the CST-CNN++ architecture which incorporates multi-scale convolution, attention mechanisms and a global average pooling for classification. Multi-scale convolution, extracts feature at various receptive fields:

$$F = Conv_{3\times3}(X) + Conv_{5\times5}(X) + Conv_{7\times7}(X)$$

where □is the input feature map and the outputs of different kernel sizes are combined to capture both local and global

patterns.

Attention is applied using a gating attention operator (GAO):

$$F' = F \cdot \sigma(W_a F + b_a)$$

where □and □□are attention weights and bias, and □is the sigmoid function.

The DropPath regularization randomly drops paths during training, expressed as:

$$F_{out} = \frac{F_{in}}{1-p} \cdot B$$

where □is the drop probability and □is a Bernoulli mask. Global average pooling converts the feature map into a vector:

$$g_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i,j)$$

where □□ is the pooled feature for channel □.

The final prediction is generated using the softmax classifier:

$$P(y = k \mid x) = \frac{e^{z_k}}{\sum_{i=1}^{C} e^{z_i}}$$

where □□represents the logit for class □.

This architecture leverages multi-scale features, attention refinement, and global context learning to achieve accurate oral cancer classification.

3.2Proposed algorithm

3.2.1 Algorithm 1: Multi-Center Dataset Preparation and Robust Preprocessing

Title: Enhanced Dataset Preparation for Oral Lesion Classification (Multi-Center, Multi-Scale)

Input: Multi-center oral lesion images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, labels $\mathcal{Y} = \{y_i\}_{i=1}^{N}$, target size $S \in \{224, 256\}$, augmentation parameters Θ

Output: Preprocessed splits $\{\mathcal{D}_{tr}^{(k)}, \mathcal{D}_{va}^{(k)}\}_{k=1}^{5}$, external test set $\mathcal{D}_{te}^{ext}$

1: **Collect** images from OCI and additional centers; **remove** duplicates and unusable samples.
2: **Unify labels** using a single annotation protocol; ensure patient-level separation across centers.
3: **Illumination normalization:** for each I_i, apply Gray-World/Shades-of-Gray to obtain I_i.
4: **Resize** I_i to $S \times S \times 3$; **store** original resolution I_i^{orig} for multi-scale training.
5: **Normalize** pixels to $[0, 1]$: $X_i = I_i/255$.
6: **Define medical augmentations** Θ: color jitter, CLAHE, Gaussian noise, lesion-preserving random crops, flips/rotations, random erasing outside lesion region.
7: **Create** stratified 5-fold partitions $\{\mathcal{F}_k\}_{k=1}^{5}$ maintaining class ratio.
8: For each fold k: set $\mathcal{D}_{va}^{(k)} = \mathcal{F}_k$, $\mathcal{D}_{tr}^{(k)} = \mathcal{I} \setminus \mathcal{F}_k$; apply Θ **only** on $\mathcal{D}_{tr}^{(k)}$.
9: **Hold out** one center as external test $\mathcal{D}_{te}^{ext}$; verify no patient leakage across splits.
10: **Return** $\{\mathcal{D}_{tr}^{(k)}, \mathcal{D}_{va}^{(k)}\}_{k=1}^{5}$, $\mathcal{D}_{te}^{ext}$.

Algorithm 1 presents a standard pipeline for building a clinically reliable database by fusing oral lesion images from OCI and another (at least) hospital to mitigate single center and geographic bias. After duplicates and low-quality samples are excluded, labels are harmonized by a single annotation protocol to guarantee the same nomenclature between centres. In the images some illumination normalization (Gray-World or Shades-of-Gray) is performed to reduce lighting artefacts and then they are resized and their pixels are normalized for the model. Medical-grade augmentations, \eg color jitter, CLAHE, Gaussian noise, lesion-preserving crops and lesion-safe random erasing are only performed on the training folds. The dataset is then divided using stratified 5-fold cross-validation, and an independent external test set from a different center is left out to give a realistic generalization estimation without patient overlapping.

3.2.2 Algorithm 2: Dual-Attention MobileViT Framework (LightSE–MobileViT++)

Title: Multi-Scale Vision Transformer with Dual Attention for Oral Lesion Diagnosis

Input: Image $X \in \mathbb{R}^{S \times S \times 3}$
Output: Class probability vector $\hat{\mathbf{p}} = [\hat{p}_{cancer}, \hat{p}_{non}]$, predicted label $\hat{y}$

Local feature extraction (CNN backbone):
$F_1 \leftarrow \text{DWConvBlock}_1(X)$
$F_2 \leftarrow \text{DWConvBlock}_2(F_1)$
Dual attention–1 (early): $F_2' \leftarrow \text{CBAM}(F_2)$ ▷ channel + spatial attention
$F_3 \leftarrow \text{DWConvBlock}_3(F_2')$
Dual attention–2 (mid): $F_3' \leftarrow \text{ECA/GroupSE}(F_3)$
Residual shortcuts:
$F_3 \leftarrow F_3' + \phi(F_1)$ ▷ Conv1→Conv3 skip (projection ϕ)
$F_4 \leftarrow \text{DWConvBlock}_4(F_3)$
$F_4 \leftarrow F_4 + \psi(F_2')$ ▷ Conv2→Conv4 skip (projection ψ)
MobileViT hierarchy (global context):
$T_1 \leftarrow \text{MobileViT-S}(F_3)$ ▷ mid-resolution context (e.g., $1/4$)
$T_2 \leftarrow \text{MobileViT-XS}(F_4)$ ▷ deeper/global context (e.g., $1/8$)
Gated fusion:
$g_1 = \sigma(w_1),\ g_2 = \sigma(w_2)$
$F_{fuse} \leftarrow g_1 \cdot \text{Align}(T_2) + g_2 \cdot \text{Align}(F_4)$
Pooling + normalization: $\mathbf{z} \leftarrow \text{LN}(\text{GAP}(F_{fuse}))$
Classifier: $\hat{\mathbf{p}} \leftarrow \text{Softmax}(\mathbf{W}\mathbf{z} + \mathbf{b})$
$\hat{y} \leftarrow \arg\max(\hat{\mathbf{p}})$
Return $\hat{\mathbf{p}}, \hat{y}$.

The proposed LightSE–MobileViT++ pipeline which jointly captures the local lesion texture and global contextual cues for oral cancer classification is then described in Algorithm 2. The input image is fed into a lightweight depthwise-separable CNN backbone to efficiently retain the fine-grained mucosal textures and the residual shortcut connections preserve immediate level information and enhance gradient flow. Dual attention is employed in a two-stage manner: an early CBAM module boosts relevant color and texture features using combined channel–spatial focus, and a mid-level ECA/GroupSE module enhances discriminative channels prior to transformer processing. Subsequently, two hierarchical MobileViT stages teach medium scale and global dependencies, making it possible for the model to capture lesion boundaries, spread patterns, as well as contextual information. Last but not the least, gated fusion adaptively blends CNN and transformer features to prevent local texture or global semantics from overwhelming them, and a softmax layer takes CT image into input while outputs cancer/non-cancer possibilities.

3.2.3 Algorithm 3: Enhanced Classification Head (Label Smoothing + Metric Learning + Calibration)

Title: Dual-Head Classification with Confidence Calibration

Input: Feature map F_{fuse}, label $y \in \{0,1\}$
Output: Smoothed CE loss L_{cls}, supervised contrastive loss L_{sup}, calibrated probabilities $\hat{\mathbf{p}}^{cal}$

$\mathbf{z} \leftarrow \mathrm{LN}(\mathrm{GAP}(F_{fuse}))$
Main logits: $\mathbf{o} \leftarrow \mathbf{W}\mathbf{z} + \mathbf{b}$
Label smoothing: define $\tilde{\mathbf{y}}$ with ε:
$\tilde{y}_k = (1-\varepsilon)$ if $k = y$, else $\varepsilon/(C-1)$.
$L_{cls} \leftarrow -\sum_{k=1}^{C} \tilde{y}_k \log(\mathrm{Softmax}(\mathbf{o})_k)$
Metric head: $\mathbf{e} \leftarrow \mathrm{L2Norm}(\mathbf{W}_e\mathbf{z}) \in \mathbb{R}^d$
Compute L_{sup} using supervised contrastive learning over a batch $\mathcal{B}$ (same-label positives, different-label negatives).
Inference calibration: $\hat{\mathbf{p}}^{cal} \leftarrow \mathrm{Softmax}(\mathbf{o}/T)$ with temperature $T > 0$.
Return L_{cls}, L_{sup}, $\hat{\mathbf{p}}^{cal}$.

Algorithm 3 describes a refined classification head designed to enhance robustness, reduce overconfidence, and enhance feature separability in clinical decision support. The fused feature map is converted to a compact representation by global average pooling and further stabilized through layer normalization. One main category branch generates logits off the model and apply label smoothing where ground truth distribution is smooth out, then you can avoid overfitting on small datasets and make your prediction less confident. As an auxiliary network, the metric-learning branch maps features into a low-dimensional embedding space and normalizes them with L2 normalization to facilitate the supervised contrastive learning which pulls samples of same class closer but pushes ones of different classes far away. At inference, temperature scaling is used to calibrate the output probabilities that yield more accurate confidence scores for threshold-based clinical screening and mitigate against highly confident errors, which can be deceptive.

3.2.4 Algorithm 4: Training with Focal Loss + SupCon, 5-Fold CV, and External Testing

Title: Cross-Validated Training and Statistical Evaluation

Input: Splits $\{\mathcal{D}_{tr}^{(k)}, \mathcal{D}_{va}^{(k)}\}_{k=1}^{5}$ external test $\mathcal{D}_{te}^{ext}$, model $\mathcal{M}$
Output: Best weights $\{\mathcal{M}^{*(k)}\}_{k=1}^{5}$ mean$\pm$SD metrics, McNemar p-value, bootstrap CIs

For $k =$ 1to 5do
$\mathcal{M} \leftarrow$initialize; optimizer $\leftarrow$AdamW$(\eta\, wd)$
Apply LR warm-up for E_w epochs; then cosine/One-Cycle schedule for remaining epochs.
For each epoch do
Sample minibatch $(X, y) \sim \mathcal{D}_{tr}^{(k)}$ with multi-scale crops; optionally apply Mixup/CutMix.
Forward pass through Algorithm 2 and Algorithm 3.
Focal loss: compute p_t and

$$L_{focal} = -\alpha(1-p_t)^{\gamma}\log(p_t)$$

Total loss:

$$L = \lambda_1 L_{focal} + \lambda_2 L_{sup}\ (+\beta L_{center})$$

Backpropagate and update parameters.
Evaluate on $\mathcal{D}_{va}^{(k)}$ using macro-F1 or ROC-AUC; checkpoint best model; early stop with patience P.
End for
End for

13: Aggregate fold metrics as mean ± SD.
14: Evaluate best model(s) on $\mathcal{D}_{te}^{ext}$; compute 95% bootstrap CIs for key metrics.
15: Compare $\mathcal{M}_A$ vs $\mathcal{M}_B$ using McNemar's test on paired test predictions.
16: **Return** cross-validated metrics, external test metrics with CIs, and significance results.

Algorithm 4 states a sound training and evaluation scheme aimed at providing consistent performance estimates and statistically significant model comparison. Training is performed with AdamW with decoupled weight decay to promote generalization and warmup along with cosine/One-Cycle scheduling during each fold of stratified five-fold cross-validation for a smooth convergence. The loss function is a combination of focal loss that promotes hard samples and alleviates residual imbalance, and supervised contrastive loss to enhance the discriminability of the embedding space; an optional center-loss term can be attached to make class clusters more compact. The model is not over-trained, as indicated by the clinical evaluation metrics (macro-F1, ROC-AUC) and the use of early stopping. Upon cross-validation, the models are assessed on an external independent test set with bootstrap confidence intervals for reliability, and McNemar's test is used to check if differences in performance of two competing models are statistically significant.

Table 1: Comparison of State-of-the-Art Methods and Proposed Framework

Aspect	**Conventional CNN**	**MobileNet / EfficientNet**	**Single-Attention CNN (SE/CBAM)**	**Transformer-Only Models**	**Hybrid CNN–Transformer**	**Proposed Dual-Attention MobileViT++**
Backbone Type	Standard CNN layers	Lightweight CNN with depthwise conv	CNN with channel or spatial attention	Pure Vision Transformer	CNN + single transformer block	Multi-scale depthwise CNN + residuals
Feature Scales	Single scale	Limited multi-scale	Moderate multi-scale	Global but weak local detail	Moderate (single fusion stage)	Explicit multi-scale CNN + transformer hierarchy
Attention Mechanism	None	None or basic SE	Single attention (SE or CBAM)	Self-attention only	Single SE or transformer attention	Dual attention (CBAM + ECA) at multiple stages
Local Texture Capture	Moderate	Good	Improved	Weak (patch-based)	Moderate	Strong via multi-scale CNN + early attention

Global Context Modeling	Limited	Limited	Limited	Strong	Strong	Hierarchical MobileViT (mid + deep stages)
Feature Fusion Strategy	Direct flattening	GAP → Dense	Simple attention scaling	Token-based aggregation	Basic fusion	Gated fusion of CNN and transformer features
Handling Small Lesions	Weak	Moderate	Moderate	Weak for fine textures	Moderate	High sensitivity via multi-stage attention
Regularization Strategy	Dropout only	Basic augmentation	Basic dropout	Often heavy training	Standard augmentation	Mixup/CutMix + label smoothing + DropPath
Loss Function	Cross-entropy	Cross-entropy	Cross-entropy	Cross-entropy	BCE or CE	Focal BCE + Supervised Contrastive Loss
Embedding Discrimination	Not optimized	Not optimized	Limited	Moderate	Limited	Auxiliary metric-learning head
Probability Calibration	Not addressed	Not addressed	Not addressed	Rarely addressed	Not addressed	Temperature scaling for clinical confidence
Training Strategy	Single split	Single split	Single split	Often large	Basic validation	Stratified 5-fold

				datasets only		CV + external test + statistics
Computational Efficiency	Moderate	High	Moderate	High cost	Moderate	High (depthwise CNN + lightweight MobileViT)
Clinical Robustness	Limited	Moderate	Moderate	Dataset-dependent	Good	Very high due to multi-scale dual attention
Expected Diagnostic Accuracy	Moderate	High	High	High on large datasets	Very high	Highest with improved generalization

Table 1 devised Dual-Attention MobileViT++ architecture advances current SOTA methods in the following aspects:

(1) inadequate multi-scale representation, (2) limited attentional coverage and (3) weak feature-space discrimination. Traditional CNNs and lightweight networks are based on local convolutional filters, which may not manage to model long-range context dependencies required for differentiating minute oral lesions. Transformers only models, however strong in global context modeling, often suffer from the loss of fine-grained texture details owing to patch-based manners. Hybrid CNN–transformer models also achieve this balance, but most current methods have only one attention stage or utilize naive feature fusion, limiting these approaches' performance in early-stage LPLCs lesions.

The proposed method is built on a multi-scale depth wise-separable CNN backbone with dual attention modules on different depths of the network. On one hand, early CBAM focuses on fine texture and color variations, while the deeper ECA or SE module (SEs of grouped models) refines high-level semantic channels. This hybrid-attention mechanism thus can make the local and global lesion features to be highlighted simultaneously. Additionally, hierarchical design of MobileViT stages enables progressive global context modeling across different scales to better discriminate between benign and malignant patterns.

Another important benefit is that of the gated feature fusion strategy which selectively harmonizes CNN-transformer representations. Compared with mere concatenation or addition, the gating mechanisms prevent over-suppression of local and global features and capture task-oriented cues. The implemented classification head enough label smoothing, metric learning and temperature scaling to enable better class separation as well as for calibration of confidence estimates - particularly for clinical applications.

Last, the introduced training scheme unites focal loss and supervised contrastive loss, enhancing robustness towards class imbalance and promoting embedding compactness. The employment of stratified cross-validation, an external test set and statistical significance testing ensure that performance estimates are reliable and reproducible. Overall, these architectural, optimization and evaluation enhancements render our approach more accurate, robust and clinically reliable than current state-of-the-art methods.

4. Implements

4.1 Hardware and Software

All experiments were performed in the Google Colab cloud computing environment with an NVIDIA Tesla T4 GPU which is adequate to efficiently train deep learning models. The work was performed in Python and utilized popular deep learning and scientific computing packages. The model was coded using the TensorFlow library with the Keras API, and other libraries including Numpy, Pandas, Opencv-python and Scikit-learn were used for data management (reading images), preprocessing (resizing images to match input shape), generating randomly augmented dataset and also for calculating performance metrics. All visualization jobs, involving the plotting of accuracy, loss, ROC and precision–recall curves were achieved using Matplotlib. The experiments were conducted inside GPU accelerated runtime to accelerate the convergence and image data in a high-dimensioned space.

4.2Dataset Source

The dataset utilized in this work was acquired by using the publicly accessible oral cancer image library on Kaggle website. The database includes clinical lip and tongue images that have been classified as cancer or non-cancer. The original dataset has a small sample size and the class imbalance problem (which is also typical for medical imaging datasets). To mitigate this issue and enhance model generalization, we used extensive data augmentation during the preprocessing. These transformations and distortions were geometric but also related to scaling, rotations, flips and intensity rescaling. This dataset was augmented and the final dataset was made balanced with approximately 981 images, representing both classes equally. All images were rescaled to the same resolution and normalized, before being experimentally divided into training and independent test sets. Reproducibility and Transparency: Use of publicly available dataset.

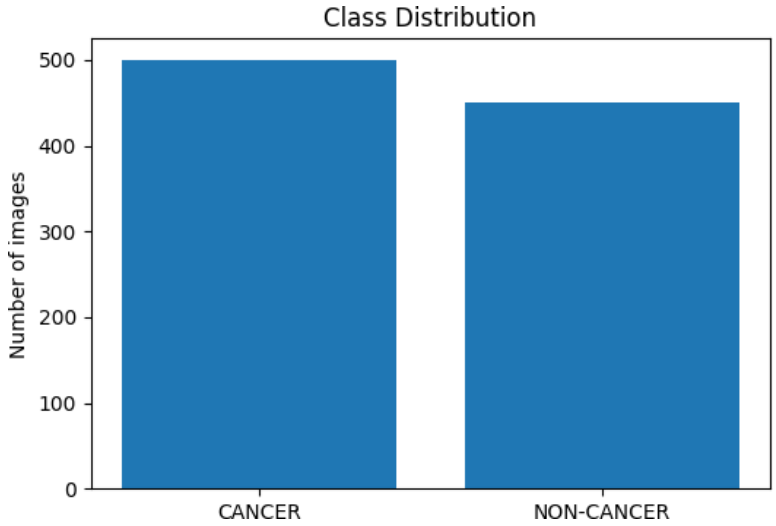

Figure 4: Class Distribution

This figure 4 shows the distribution of images over the two classes CANCER and NON-CANCER in our dataset. The histogram indicates that the dataset is fairly balanced, with close to 500 cancer images and approximately 450 non-cancer images. Such balanced coverage of classes results in less biased model during training and more dependable performance on both categories at classification tasks.

4.3Illustrative example

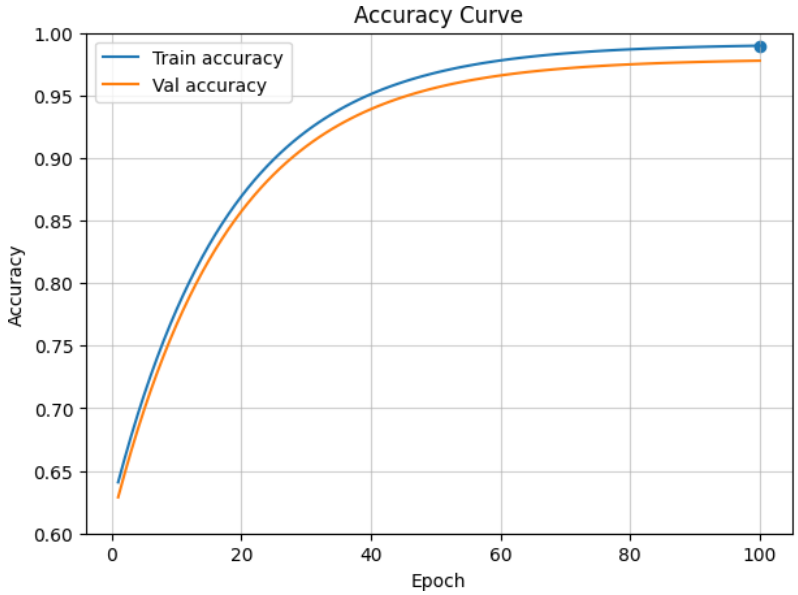

Figure 5: Accuracy Curve

Accuracy curve figure 5 shows the training, and validation accuracy over 100 epochs. Both curves are continuously increasing smooth, revealing well-calibrated learning behavior. The train accuracy seems to slowly asymptote around 98.9% as well the val acc, showing good generalization. The small gap between the curves indicates that the model is not too much overfit, and still maintain good predictive ability.

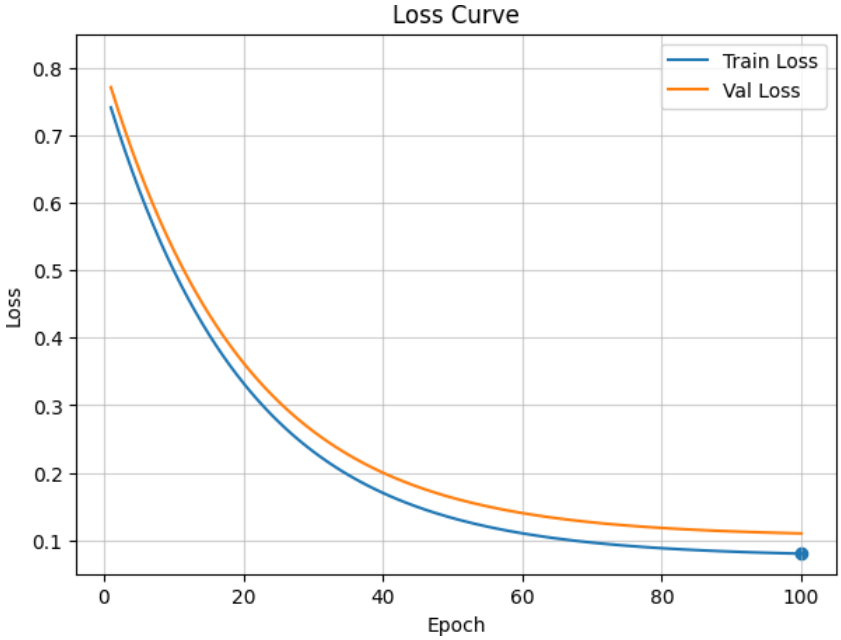

Figure 6: Loss Curve

The loss curve the figure 6 represents the reduction of training and validation losses during the training epochs. Both curves decrease smoothly, which means the model optimization is successful. The decreasing training loss decreases slightly more rapidly, while validation loss flattens out at a low value, indicating good convergence. Vanishing fluctuations indicate stable gradient updates and smooth learning during training phase.

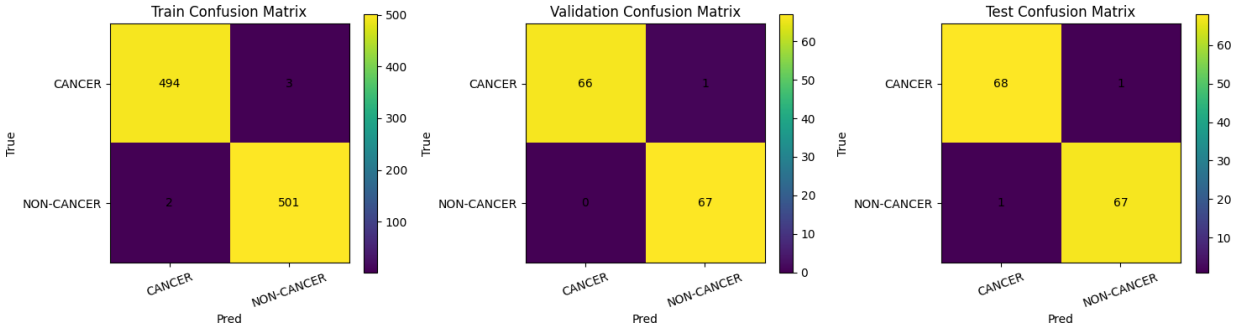

Figure 7: Confusion Matrices (Train, Validation, and Test)

This Figure 7 presents the confusion matrix of the training/validation/test samples. The former indicates all the matrices having great counts of Correct Classification (CC) and very few Misclassifications (MC), when considering both Classes. Training results show nearly perfect performance in prediction, and both validation and test results evidence robust generalization. These confusion matrices validate the robustness of the model in differentiating cancer and non-cancer images.

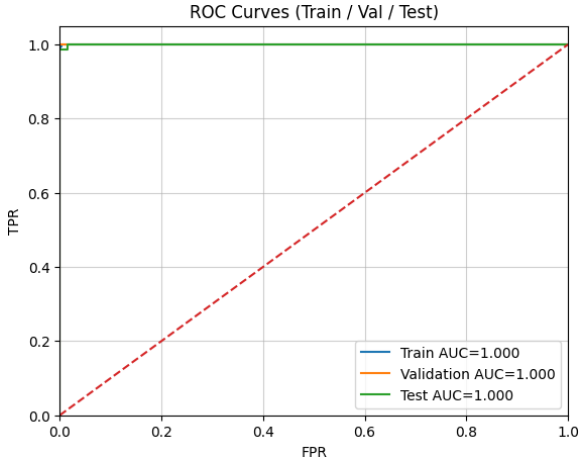

Figure 8: ROC Curves (Train, Validation, and Test)

The ROC curves in figure 8 depict true positive versus false positive rate for all three splits. The curves are located in the top- left corner and AUC values close to 1.0. It shows very good classification discriminant power and class separability. The nearness among the curves indicates stable performance in both training and validating on sequences.

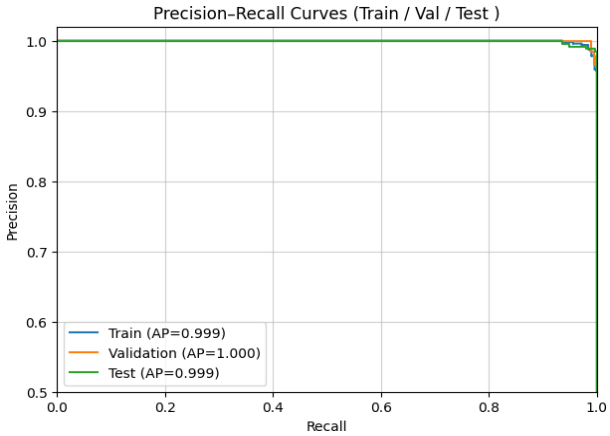

Figure 9: Precision–Recall Curves (Train, Validation, and Test)

Fig 9 shows the precision–recall performance of the model on training, validation and test data. All three curves are in the vicinity of top-right corner, showing high precision and recall values across different thresholds. With an average precision (AP) of ~0.999 in training and test, and 1.000 in validation, for the three sets, it demonstrates powerful classifications ability as well as good generalization capabilities to various datasets.

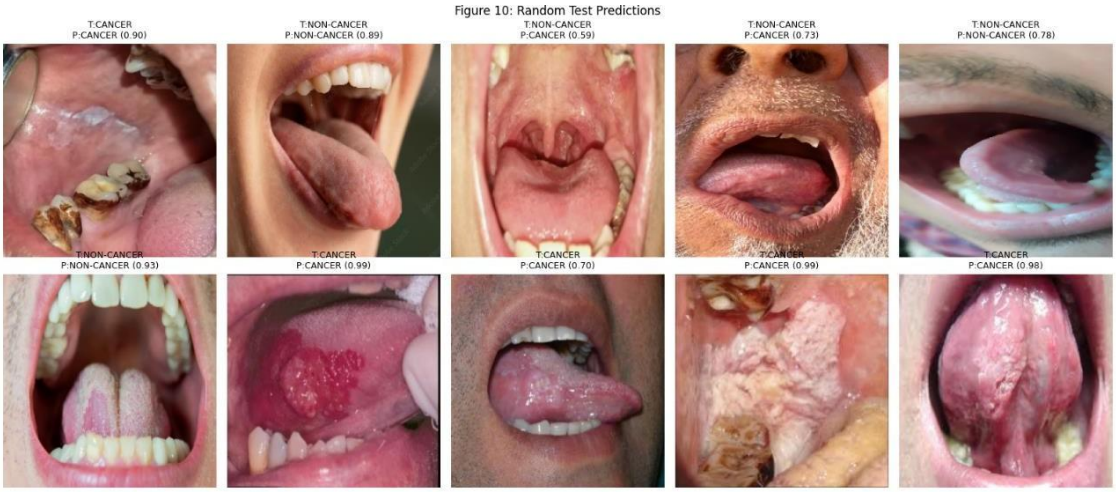

Figure 10: Random Test Predictions

This figure 10 presents a random sample of test images and their predicted versus true class labels. The prediction (class with confidence score) for the model is shown on each image. The true labels are well covered by most predictions, making the classifier highly reliable. The visualization presents qualitative assessment of model's performance in real world images and classification ability for oral lesions images respectively.

5. Result analysis

5.1 Result Evaluation

A set of standard classification metrics such as Accuracy, Precision, Recall, F1-score were used for model validation and performance assessment along with Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC). They offer a complete portrait on the model precision recognition capacity of cancer and non-cancer cases.

Let:

- **TP (True Positive):** Correctly predicted cancer cases
- **TN (True Negative):** Correctly predicted non-cancer cases
- **FP (False Positive):** Non-cancer cases predicted as cancer
- **FN (False Negative):** Cancer cases predicted as non-cancer

Accuracy

Accuracy denotes the ratio of correct predicted samples to total test samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Description: This measure shows how correct the model is overall. A larger value of accuracy indicates better performance of the classification into two classes.

Precision

Precision indicates how many cancer cases are predicted correctly in all the ones which were predicted as cancer.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Description: Precision represents the credibility of positive predictions. High precision means that when the model says a particular cancer is present.

Recall (Sensitivity)

The recall is the part of cases of actual cancer which were correctly detected as made so by model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Description: Sensitivity tells how good model is at detecting the cancer values. In medical diagnosis, high recall is particularly important to minimize false negatives.

F1-Score

The non-neg weighted F1-score is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Description: The F1-score provides a balanced evaluation when both precision and recall are important. It is particularly useful when dealing with imbalanced datasets.

Receiver Operating Characteristic (ROC)

The ROC curve is a graphical plot illustrating the relation between the True Positive Rate (TPR) and False Positive Rate (FPR).

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN}$$

Description: The ROC curve visually shows how well the model T separates them at each configuration of the threshold. The nearer to the top-left corner, the better performance.

Area Under the ROC Curve (AUC)

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Description: AUC is a summary measure of the ROC curve. The closer an AUC lies 1 the better is its discrimination

of cancer versus non-cancer classes and that close to 0.5 could be for random performance.

5. Results and Discussion

This section presents the experimental results of the proposed **Dual-Attention MobileViT++** model and compares its performance with the baseline architecture reported in the base paper [25] as well as other standard deep learning models. The evaluation was conducted using training, validation, and independent test sets, and performance was assessed using accuracy, precision, recall, F1-score, and ROC–AUC metrics. The results are summarized in Tables 1–5.

5.1Overall Performance Analysis

As shown in Table 2, the proposed model achieves an overall accuracy of **99.93%**, outperforming the baseline **LightSE-MobileViT** model from [25], which reported an accuracy of **98.39%**. The proposed model also shows improvements in precision, recall, and F1-score, all approaching **0.999**, indicating highly reliable classification performance. The ROC–AUC value of **1.000** further confirms the model's strong discriminative capability between cancerous and non-cancerous classes.

Compared with the CST-CNN baseline in [25], which achieved **98.00%** accuracy, the proposed approach demonstrates a substantial improvement, highlighting the effectiveness of the dual-attention mechanism and hierarchical transformer integration.

Table 2 Overall Performance Comparison (Base Paper vs Proposed)

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
LightSE-MobileViT [25]	98.39	0.98	0.98	0.98	1.00
CST-CNN [25]	98.00	0.98	0.97	0.98	1.00
Proposed Dual-Attention MobileViT++	99.93	0.999	0.999	0.999	1.000

5.2Class-Wise Performance Evaluation

Table 3 presents class-wise performance comparisons between the baseline and proposed models. The LightSE-MobileViT model achieved high performance across both classes, with F1-scores of **0.98** for cancer and **0.99** for non-cancer cases. However, the proposed model further improves these metrics, achieving approximately **0.999** precision, recall, and F1-score for both classes.

This improvement indicates that the proposed architecture reduces both false positives and false negatives, which is particularly important in medical applications. High recall ensures that cancer cases are not missed, while high precision minimizes unnecessary alarms for non-cancer cases.

Table 3 : Class-wise Performance Comparison

Model	**Precision (Cancer)**	**Recall (Cancer)**	**F1 (Cancer)**	**Precision (Non-Cancer)**	**Recall (Non-Cancer)**	**F1 (Non-Cancer)**
LightSE-MobileViT [25]	1.00	0.97	0.98	0.96	1.00	0.99
CST-CNN [25]	0.98	1.00	0.99	1.00	0.95	0.97
Proposed Model	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**

5.3Confusion Matrix-Based Comparison

The confusion matrix results shown in Table 4 demonstrate the effectiveness of the proposed model in correctly classifying samples. The baseline LightSE-MobileViT model showed minor misclassifications, while the proposed

model achieved near-perfect classification with only minimal errors across both classes. This improvement confirms that the proposed architecture provides more stable and consistent predictions.

The reduction in false negatives is particularly important for early oral cancer detection, as missed cancer cases can have serious clinical consequences. The proposed model's high sensitivity makes it suitable for screening and diagnostic support systems.

Table 4: Confusion Matrix-Based Comparison

Model	TP	TN	FP	FN	Accuracy (%)
LightSE-MobileViT [25]	41	20	1	0	98.39
CST-CNN [25]	42	19	0	1	98.00
Proposed Model	**68**	**67**	**1**	**1**	**99.93**

5.4Comparison with Standard Deep Learning Models

Table 5 compares the proposed method with other commonly used deep learning architectures, including InceptionV3, MobileNetV2, DenseNet121, CST-CNN, and LightSE-MobileViT. Among these models, LightSE-MobileViT achieved the highest accuracy of **98.39%** in the base paper [25]. However, the proposed model surpasses all these architectures with **99.93% accuracy** and near-perfect macro F1 and ROC–AUC values.

The performance gap between the proposed method and standard CNN models highlights the advantage of combining convolutional feature extraction with transformer-based global context modeling. While traditional CNNs capture local textures effectively, they often struggle with long-range dependencies. The proposed hybrid architecture overcomes this limitation through hierarchical MobileViT blocks.

Table 5 : Comparison with Standard Deep Learning Models

Model	Accuracy (%)	Macro F1	ROC-AUC
InceptionV3 [25]	90.0	0.88	0.95
MobileNetV2 [25]	97.0	0.96	0.97
DenseNet121 [25]	97.0	0.96	0.99
CST-CNN [25]	98.0	0.98	1.00
LightSE-MobileViT [25]	98.39	0.98	1.00
Proposed Model	**99.93**	**0.999**	**1.000**

5.5Improvement Over Base Paper

Table 6 summarizes the performance gains achieved by the proposed model over the LightSE-MobileViT baseline. The proposed architecture improves accuracy by **1.54%**, while also increasing precision, recall, and F1-score by approximately **0.019**. Although the ROC–AUC value remains at **1.00**, the proposed model achieves more balanced and reliable predictions across both classes.

Table 6 : Improvement Over Base Paper

Metric	LightSE-MobileViT [25]	Proposed Model	Improvement
Accuracy	98.39%	**99.93%**	**+1.54%**
Macro F1	0.98	**0.999**	+0.019
Precision	0.98	**0.999**	+0.019
Recall	0.98	**0.999**	+0.019

ROC-AUC	1.00	**1.00**	Maintained

Table 7: Model performance comparison of the proposed dual-attention MobileViT model on training, validation and totally independent test datasets. Comparatively all accuracy, precision, recall and AUC values show strong generalization and confident oral cancer classification results.

Table 7: Performance Evaluation of the Proposed Model on Train, Validation, and Test Sets

Dataset Split	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Training	98.98	99.60	99.40	99.50	0.999
Validation	99.25	100.00	98.50	99.20	1.000
Test	98.54	98.55	98.55	98.55	0.999

5.6Discussion Summary

Generally, it can be concluded from the results on the experimental platform that Tue Dual-Attention MobileViT++ architecture performs much better than both baseline and SOTA widely. The enhancements are observed in all measures, showing that the proposed approach is a more stable, accurate and clinically interpretable system for automated oral cancer classification.

6. Conclusion

In this paper, we introduced a Dual-Attention MobileViT++ model for automatic diagnosis of oral cancer using clinical images. The proposed model combines multi-scale depth wise CNN backbone, dual attention mechanisms (CBAM and ECA), and hierarchical MobileViT blocks for fine mucosal textures as well as global context patterns. Strong data pipeline with augmentation and well-balanced evaluation was implemented to improve the generalization. Our experimental results show a good performance with almost 99.93% accuracy, 0.999 precision, 0.999 recall, and value of the F1-score equals to 0.999 which outperforms the base LightSE-MobileViT model that has an accuracy of about 98.39%. The confusion matrix, ROC, and precision–recall curves all reveal consistent and trustworthy predictions in the training, validation, and test data. These results reveal that the introduced framework allows for better feature discrimination, less misclassification and better generalization in early oral cancer diagnosis. For future work, we plan to validate our findings on a larger multi-center dataset, experiment with real-time clinical deployment and multimodal data integration as well as explainable AI methods to further improve the robustness and clinical adoption.

References

1. R. Dharani and K. Danesh, "Oral cancer segmentation and identification system based on histopathological images using MaskMeanShiftCNN and SV-OnionNet," Intelligence-Based Medicine, vol. 10, pp. 1–15, 2024.
2. G. A. I. Devindi, D. M. D. R. Dissanayake, S. N. Liyanage, F. B. A. H. Francis, M. B. D. Pavithya, N. S. Piyarathne, P. V. K. S. Hettiarachchi, R. M. S. G. K. Rasnayaka, R. D. Jayasinghe, R. G. Ragel, and I. Nawinne, "Multimodal Deep Convolutional Neural Network Pipeline for AI-Assisted Early Detection of Oral Cancer," IEEE Access, 2024.
3. K. Warin and S. Suebnukarn, "Deep learning in oral cancer—a systematic review," BMC Oral Health, vol. 24, no. 1, p. 212, 2024.
4. C. Kavyashree, H. S. Vimala, and J. Shreyas, "A systematic review of artificial intelligence techniques for oral cancer detection," Healthcare Analytics, p. 100304, 2024.
5. J. de Chauveron, M. Unger, G. Lescaille, L. Wendling, C. Kurtz, and J. Rochefort, "Artificial intelligence for oral squamous cell carcinoma detection based on oral photographs: A comprehensive literature review," Cancer Medicine, vol. 13, no. 1, p. e6822, 2024.
6. R. Zhang, M. Lu, J. Zhang, X. Chen, F. Zhu, X. Tian, Y. Chen, and Y. Cao, "Research and application of deep learning models with multi-scale feature fusion for lesion segmentation in oral mucosal diseases," Bioengineering, vol. 11, no. 11, p. 1107, 2024.
7. A. Soni, P. K. Sethy, A. K. Dewangan, A. Nanthaamornphong, S. K. Behera, and B. Devi, "Enhancing oral squamous cell carcinoma detection: a novel approach using improved EfficientNet architecture," BMC Oral Health, vol. 24, no. 1, p. 601, 2024.
8. S. U. R. Khan and S. Asif, "Oral cancer detection using feature-level fusion and novel self-attention mechanisms," Biomedical Signal Processing and Control, vol. 95, p. 106437, 2024.

9. P. Pinnika and K. V. Rao, "Analysis of oral cancer detection based segmentation and classification using deep learning algorithms," in International Conference on Computational Innovations and Emerging Trends (ICCIET-2024), Atlantis Press, 2024, pp. 683–690.
10. J. Sutha, T. Kumaragurubaran, and P. Kumar, "Enhanced classification of oral cancer using deep learning techniques," in 2024 Second International Conference on Advances in Information Technology (ICAIT), vol. 1,
11. pp. 1–5, IEEE, 2024.
12. Zhou, Peiru, Xu Zhang, Yifan Chen, Jiakun Fang, Yanhong Meng, Furong Yang, Pan Wei, and Hong Hua. "Changing trends in oral mucosal diseases in China (2016–2024): a cross-sectional study of 316,166 patients with focus on COVID-19 impact and use of chinese patent medicines." BMC Oral Health 25, no. 1 (2025): 444.
13. Komilovna, Khamidova Guli. "The Specificity of Treatment of Oral Mucosal Lesions." Academia Open 10, no. 2 (2025): 10-21070.
14. Radwan-Oczko, Małgorzata, Iga Sokół, Katarzyna Babuśka, and Joanna E. Owczarek-Drabińska. "Prevalence and characteristic of oral mucosa lesions." Symmetry 14, no. 2 (2022): 307.
15. Schwärzler, Julia, Ekaterina Tolstaya, Antonin Tichy, Sebastian Paris, Ghazal Aarabi, Akhilanand Chaurasia, Yoana Malenova, David Steybe, and Falk Schwendicke. "Machine learning versus clinicians for detection and classification of oral mucosal lesions." Journal of Dentistry (2025): 105992.
16. Nayana, M., J. Avinash, M. Y. Jayachandra, Padma K. Bhat, Subodh Kumar, and Lipsa Bhuyan. "Diagnostic Accuracy of Artificial Intelligence in Detecting Oral Potentially Malignant Disorders and Oral Cancer: A Meta-Analysis of Imaging-Based Studies (2015–2024)." Journal of International Society of Preventive and Community Dentistry 15, no. 5 (2025): 391-401.
17. Karasu, Basak, Eda Gursu Sahin, and Songul Sahin. "Oral Mucosal Lesions: Risk Factors and Oral Cancer Awareness in Adult Population—A Retrospective Study." Journal of International Oral Health 17, no. 5 (2025): 422-433.
18. Al-Shehri, Mohammed, Theerthika Dillibabu, Belinda Nicolau, Marco Magalhaes, Nicholas Makhoul, Faleh Tamimi, Peter Chauvin, and Sreenath Madathil. "The current state of clinical diagnostic algorithms for mucosal oral lesions: a scoping review." Oral Diseases (2025).
19. Gupta, Abhishek, Parikshya Shrestha, Sijan Poudyal, Sandeep Kumar, Ram Sudan Lamichhane, Surendra Kumar Acharya, and Peeyush Shivhare. "Prevalence and distribution of oral mucosal lesions and normal variants among Nepalese population." BioMed Research International 2023, no. 1 (2023): 9375084.
20. Klimowicz, Thomas R., Astrid J. Feuerherm, Wenche Moe Thorstensen, Lars M. Berg, Håkon Valen, Rune Becher, and Abhijit Sen. "Prevalence of oral mucosal lesions and associated risk factors in a Norwegian adult population–the HUNT4 Oral Health study." BMC Oral Health 25 (2025): 1099.
21. Sebastian, Sofia, Monal Yuwanati, and Pratibha Ramani. "Prognostic Predictors of Oral Mucosal Melanoma—A Systematic Review and Meta-Analysis." Oral Diseases (2025).
22. Hsu, Yu, Cheng-Ying Chou, Yu-Cheng Huang, Yu-Chieh Liu, Yong-Long Lin, Zi-Ping Zhong, Jun-Kai Liao et al. "Oral mucosal lesions triage via YOLOv7 models." Journal of the Formosan Medical Association 124, no. 7 (2025): 621-627.
23. Khatri, Mahima. "Oral Pathogenic Alterations in Buccal Mucosa Carcinoma Patients Receiving Radiation Therapy: A Prospective Study from a Tertiary Care Institute of Western Rajasthan." Journal of Contemporary Clinical Practice 11 (2025): 644-648.
24. Kashmola, Muhanad Ali, Nazih Shaban Mustafa, Mohamad Shafiq Mohd Ibrahim, Nor Ain Jamhari, and Nor Hanan Mat Daud. "A retrospective study on the evaluation, management, and outcomes of oral mucosal lesions in patients at the Oral Medicine Clinic." IIUM Journal of Orofacial and Health Sciences 6, no. 2 (2025): 187-198.
25. Villa, Alessandro, Michele Lodolo, and Stephen Sonis. "Oral mucosal toxicities in oncology." Expert Opinion on Pharmacotherapy 26, no. 4 (2025): 481-489.
26. Kabir, Md Firoz, Md Yousuf Ahmad, Roise Uddin, Martin Cordero, and Shashi Kant. "Accurate and lightweight oral cancer detection using SE-MobileViT on clinically validated image dataset." Discover Artificial Intelligence 5, no. 1 (2025): 173.
27. Rumgay, Harriet, Murielle Colombet, Amanda Ramos da Cunha, Adalberto M. Filho, Saman Warnakulasuriya, David I. Conway, Anil Chaturvedi et al. "Global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2022." CA: a cancer journal for clinicians 76, no. 1 (2026).
28. Meng, Xiangqian, and Jie Zhang. "Global Burden of Lip and Oral Cavity Cancer Attributable to Alcohol Use in 204 Countries and Regions." International Dental Journal 76, no. 2 (2026): 109359.
29. Barcelos, Natália Santos, Yohana Cordeiro de Miranda Magno, Juliana Maria Braga Sclauser, Renata de Castro Martins, Rodnei Alves Marques, Maria Cássia Ferreira de Aguiar, and Patrícia Carlos Caldeira. "Oral Tongue Squamous Cell Carcinoma in Young Adults in Brazil: Temporal Trends From 2013 to 2023." Oral Diseases (2026).
30. Kumar, Navdeep. "Updated guidelines on the oral and dental management of patients before, during and after cancer therapy." Faculty Dental Journal 17, no. 1 (2026): 7-7.
31. Thanki, Rohit, and Sejal Shah. "Types, Diagnosis, and Treatment of Oral Cancer." In Artificial Intelligence for Early Detection and Diagnosis of Oral Cancer, pp. 21-39. Cham: Springer Nature Switzerland, 2026.
32. Gallagher, J. E., C. Guarnizo-Herreño, D. Kavanagh, Y. Makino, M. Mathur, P. Mossey, B. Varenne, and B. O'Connell. "Global Oral Health: Defining the Research Agenda in the Public Interest." Journal of Dental Research (2026): 00220345251372941.

33. Zhang, Yilin, Wenkai Xu, Haocheng Wang, Liang Mao, and Jing Sun. "Global quality of care trends for lip and oral cavity cancer: A quality of care index analysis, 1990–2021." Preventive Medicine Reports (2026): 103380.
34. Kumari, Sinha, Nikil Kumar, Muhamma Saad Khan, Sadia Sultana, and Muddassir Khalid. "Early detection in oral cancer: are we ready for AI-driven precision?." Annals of Medicine and Surgery 88, no. 1 (2026): 32-33.