An Improved Multi-Scale Attention EfficientNet Framework for Oral Cancer Image Classification

Muhammad Abuarrah[1*], Marwan D. Saleh[2], Falah Y. H. Ahmed[3]

[1] Radiography and Medical Imaging Department, Fatima College of Health Sciences, Abu Dhabi, United Arab Emirates;
[2] Electronics and Computer Science University of Southampton Malaysia, Johor 79200, Malaysia;
[3] Faculty of Computing and Information Technology, Sohar 311, Sohar University, Oman.

* **Corresponding author:** Muhammad.Abuarah@actvet.gov.ae.

ABSTRACT: Automated oral cancer image classification has attracted increasing attention as deep learning offers opportunities to support diagnostic image analysis. However, limited dataset availability, redundant observations, and insufficient clinical metadata can compromise model reliability and the generalizability of reported performance. This study presents an improved Multi-Scale Attention EfficientNet (MSA-EfficientNet) framework that integrates hierarchical visual representations through attention-guided feature fusion. The proposed architecture extends EfficientNet-B0 by combining intermediate and deep feature maps, adaptive multi-scale weighting, and progressive transfer learning within a reproducible classification pipeline. The methodological framework incorporates duplicate-aware data partitioning, controlled augmentation, probability calibration, and model interpretability. An assessment of the Oral Cancer Multispectral Imaging Dataset (OC-MSI) identified 950 RGB images, comprising 500 cancer and 450 non-cancer observations, including 36 exact duplicates and leaving 914 unique images. The available archive lacks patient-level identifiers and independently accessible spectral channels, limiting the evaluation to RGB image classification. A completed preliminary DARC-Net experiment achieved 63.77% classification accuracy, 69.01% sensitivity, 58.21% specificity, and an area under the receiver operating characteristic curve of 0.6523. A separate simulation-based analysis reported 93.48% accuracy and an ROC–AUC of 0.975, alongside a structured 30-epoch optimization profile; these values are not measured outcomes of MSA-EfficientNet. The study establishes a technically specified architectural framework and comprehensive evaluation strategy for investigating attention-enhanced EfficientNet classification. Experimental implementation, matched-baseline comparisons, ablation analysis, and independent validation remain necessary to determine the model's actual predictive effectiveness and potential clinical relevance.

Keywords: Oral cancer classification, multi-scale attention, EfficientNet-B0, Transfer learning, Deep learning, Model reliability.

I. INTRODUCTION

Oral cancer represents a substantial global health challenge because of its associated morbidity, mortality, and the clinical consequences of delayed diagnosis. Oral squamous cell carcinoma (OSCC) accounts for the predominant proportion of oral malignancies and requires accurate identification to support timely clinical intervention and appropriate treatment planning. Global epidemiological evidence shows considerable geographical variation in oral cancer incidence, underscoring the need to improve diagnostic access and early detection strategies [1]. Histopathological examination remains fundamental to confirming malignancy, particularly through identifying abnormal cellular morphology, tissue architecture, and invasive growth patterns. Nevertheless, differences in diagnostic expertise, tissue characteristics, and image interpretation may complicate consistent assessment. These challenges have encouraged the development of artificial intelligence (AI)-assisted classification systems that can identify discriminative image features and support diagnostic investigations. Previous research has demonstrated the technical feasibility of convolutional neural networks (CNNs) for classifying oral squamous cell

carcinoma in histopathological images, providing a foundation for further methodological development [2].

Deep learning has transformed medical image analysis by enabling computational systems to learn hierarchical representations directly from visual information. Unlike traditional machine-learning approaches that frequently depend on manually designed morphological or textural descriptors, CNNs automatically extract increasingly complex features through successive convolutional operations. These representations can capture local image patterns and broader spatial relationships relevant to classification. Recent reviews of medical image classification demonstrate the widespread adoption of CNN-based architectures and highlight the continuing importance of architecture selection, feature representation, data preprocessing, and robust validation [3]. Among established deep learning architectures, EfficientNet has attracted considerable attention because of its efficient approach to balancing network depth, width, and image resolution through compound scaling. This strategy improves representational capacity while maintaining comparatively efficient computational requirements [4]. Consequently, EfficientNet provides a suitable architectural foundation for investigating image classification tasks involving relatively limited computational resources and training data.

Transfer learning further supports applying deep neural networks to smaller medical imaging datasets by initializing models with representations learned from large-scale source datasets. However, the effectiveness of transferred features depends on the relationship between the source domain and the target medical imaging task. Previous investigations have demonstrated that pretrained models do not necessarily outperform simpler architectures in every medical imaging setting, highlighting the importance of controlled experimental comparisons [5]. Furthermore, conventional classification networks may produce highly confident predictions that do not accurately reflect their probability of correctness. Research on neural network calibration has demonstrated that model confidence and classification accuracy must be evaluated separately, particularly when prediction probabilities are intended to inform consequential decisions [6]. These observations suggest that improving classification performance requires an integrated consideration of feature representation, optimization, probability reliability, and evaluation quality rather than relying exclusively on architectural complexity.

Multi-scale feature learning provides a promising direction for strengthening conventional EfficientNet-based classification. Images may contain informative characteristics at different spatial resolutions, with shallow convolutional layers retaining localized textures and deeper layers representing more abstract visual structures. Reliance on the final feature map alone may therefore limit the direct integration of complementary information distributed across network stages. Attention mechanisms offer a way to adaptively refine feature representations by emphasizing informative channels and spatial locations. For example, Woo et al. introduced the Convolutional Block Attention Module (CBAM), demonstrating how channel-wise and spatial attention can be integrated into convolutional architectures to improve feature refinement [7]. Building on this methodological foundation, integrating multi-scale feature extraction and attention-guided fusion within EfficientNet represents a plausible approach for learning complementary visual characteristics. Nevertheless, the effectiveness of such integration for oral cancer classification remains an empirical question and must be established through comparisons against conventional EfficientNet models under identical experimental conditions.

Despite continuing advances in image classification, several methodological limitations restrict the reliability and generalizability of deep learning models developed using relatively small medical imaging datasets. These include limited sample diversity, class imbalance, overfitting, inconsistent preprocessing, and insufficient separation between training and evaluation data. In particular, duplicated images, related observations, or data-dependent preprocessing can introduce information leakage and produce misleadingly optimistic performance estimates. Roberts et al. emphasized the importance of rigorous data handling, appropriate evaluation procedures, and transparent reporting when developing medical imaging

models [8]. Data augmentation may help improve model robustness and reduce overfitting, but it must be applied carefully to avoid introducing unrealistic transformations or contamination between development and testing subsets [9]. Furthermore, high overall accuracy does not adequately describe the clinical consequences of false-negative and false-positive predictions. Reliable evaluation therefore requires complementary performance indicators, including sensitivity, specificity, precision, F1-score, receiver operating characteristic analysis, and uncertainty assessment.

The Oral Cancer Multispectral Imaging Dataset (OC-MSI) provides a relevant image collection for investigating these methodological challenges. The supplied archive comprises 950 RGB JPEG images categorized into 500 cancer and 450 non-cancer observations. A preliminary examination identified 36 exact duplicate files, leaving 914 unique images. Although the dataset is described as a multispectral imaging resource, the available files do not contain separately accessible wavelength channels or comprehensive acquisition metadata. Furthermore, patient-level identifiers are unavailable, preventing direct verification of patient-independent evaluation. These findings are derived from the dataset audit conducted for the present research and are not conclusions drawn from the published literature. They indicate the importance of employing duplicate-aware preprocessing, carefully controlled data partitioning, and appropriate restrictions on the interpretation of diagnostic performance. Accordingly, the current investigation focuses on classification using the available RGB image representations rather than making unsupported claims regarding wavelength-specific learning or clinical generalizability.

The central research problem is developing a computationally efficient and methodologically reliable classification framework that extracts complementary visual features from a limited image dataset while minimizing overfitting and evaluation bias. Although pretrained EfficientNet provides an established foundation for visual feature extraction [4], the contribution of intermediate and high-level representations remains an important consideration when designing task-specific classification systems. Integrating multi-scale features through attention-based refinement may strengthen representation learning, but any improvement must be distinguished from the independent effects of pretrained initialization, training duration, augmentation, and hyperparameter optimization [5, 7]. At the same time, duplicate-aware data preparation and rigorous validation are necessary to ensure that observed performance reflects meaningful generalization within the available image collection rather than unintended overlap between experimental subsets [8].

To address these challenges, this study proposes an improved EfficientNet-based classification framework incorporating multi-scale feature extraction, attention-guided feature fusion, and progressive transfer learning. The proposed architecture integrates feature representations from different network stages, enabling the classification head to use both localized visual information and higher-level image characteristics. The methodological framework additionally incorporates duplicate-aware dataset partitioning, training-specific augmentation, regularization, validation-based model selection, and probability calibration. Consistent with established research on model confidence, calibration is evaluated as a separate component of predictive reliability rather than as a guaranteed mechanism for improving classification accuracy [6]. Explainable AI methods, particularly Gradient-weighted Class Activation Mapping (Grad-CAM), are also proposed to visualize image regions contributing to classification decisions, while recognizing that visual attribution does not independently establish diagnostic correctness [10].

The primary objective of this research is to develop a reproducible framework for investigating whether multi-scale attention can enhance EfficientNet-based oral cancer image classification. More specifically, the study aims to assess the value of complementary feature representations, compare the proposed architecture with established classification models, examine the effects of progressive fine-tuning and regularization, and evaluate classification reliability through complementary diagnostic metrics and calibration analysis. The experimental design also emphasizes reproducible data partitioning, controlled

baseline comparisons, ablation studies, and statistical assessment of performance variability. These objectives are informed by established findings concerning CNN development, attention-based representation learning, transfer learning, and methodological reliability in medical imaging [3-9].

The principal contribution of this study is the proposed integration of multi-scale attention-based feature learning into an EfficientNet classification framework tailored to the characteristics of the available OC-MSI images. The architectural contribution combines intermediate and deep feature representations through an adaptive fusion mechanism, while the methodological contribution focuses on duplicate-aware evaluation, progressive fine-tuning, and transparent assessment of classification reliability. Additional contributions include explicit documentation of dataset constraints, a structured approach to assessing model interpretability, and an ablation-based experimental design intended to isolate the contribution of individual components. Together, these elements establish a testable framework for investigating more reliable AI-assisted oral cancer image classification. Importantly, the proposed architectural advantages remain to be demonstrated through actual training and independent evaluation; hypothetical performance scenarios are not treated as empirical evidence.

II. RELATED WORK

Recent advances in artificial intelligence have encouraged the development of automated classification systems for oral cancer detection using clinical images, histopathological representations, and multispectral imaging. Convolutional neural networks have demonstrated considerable potential for extracting discriminative visual patterns associated with malignant and non-malignant oral tissue. Warin et al. [11] investigated deep learning architectures for oral cancer classification using photographic images, demonstrating the applicability of convolutional feature extraction to oral lesion recognition. Subsequent investigations expanded this approach to oral potentially malignant disorders and multiple oral lesion categories [12, 13]. These studies established an important foundation for image-based oral cancer analysis, although differences in imaging modalities, diagnostic categories, and evaluation protocols limit direct comparisons of reported classification accuracy. More directly related to the present research, Grace et al. [14] proposed a hybrid meta-ensemble framework for oral tissue phenotype classification using the Oral Cancer Multispectral Imaging Dataset (OC-MSI). Their approach combined multiple ensemble-learning strategies, including stacked classification, weighted soft voting, gradient boosting, and neural ensemble learning. The authors reported an accuracy of 99.47% and an ROC–AUC of 0.9994 using a dataset described as containing 950 images. These findings establish an important dataset-specific reference for the present investigation. However, the reported performance should be interpreted within the original experimental design, and its direct comparability with the present work depends on verification of image provenance, preprocessing, data partitioning, and available spectral information.

Other studies have examined the application of pretrained convolutional architectures to oral cancer image classification. Baydoğan [15] investigated hybrid deep learning strategies using a Kaggle oral cancer dataset containing 950 images, including 500 cancer and 450 non-cancer observations. The study demonstrated the feasibility of combining pretrained visual representations with conventional classification techniques. Although these dataset characteristics resemble those of the present OC-MSI archive, the identity of the underlying image collections has not been established. Consequently, this work is considered relevant supporting evidence rather than a confirmed experiment on the same dataset. EfficientNet-specific research in oral squamous cell carcinoma (OSCC) has also emerged, although most published applications analyze histopathological images rather than the RGB representations available in OC-MSI. Albalawi et al. [16] applied EfficientNet to histopathological OSCC classification, and Soni et al. [17] investigated an improved EfficientNet approach for the same diagnostic domain. These studies

motivate an EfficientNet-based backbone, but differences in image acquisition and annotation prevent direct transfer of their reported accuracy to the present dataset.

For clinical photographic images, Saldivia-Siracusa et al. [18] examined convolutional classification of oral potentially malignant disorders and OSCC. This study is task-relevant but uses a different image collection and diagnostic grouping. Accordingly, evidence from external oral-image datasets should be discussed in terms of model design and validation practices rather than ranked against the present study by unadjusted accuracy alone. Feature fusion has also been investigated in other medical-imaging domains. Zebari et al. [19] combined representations from several deep models for brain-tumor classification in magnetic resonance images, while Salih and Abdulazeez [20] examined fusing ResNet-derived features for a related MRI classification task. In mammography, Zebari et al. [21] integrated regional feature information in a breast-cancer classification framework. These studies support the broader motivation for combining complementary representations; however, their imaging modalities, data, and fusion mechanisms differ from the OC-MSI RGB images and the proposed attention-weighted EfficientNet branches. Their reported results are therefore not direct performance benchmarks for this study. From an architectural perspective, EfficientNet provides a computationally efficient foundation for extracting hierarchical visual representations through compound scaling [4]. Attention-based learning offers an additional mechanism for emphasizing informative feature channels and spatial regions [7]. Nevertheless, conventional transfer-learning models may not fully exploit complementary information from different backbone stages. Integrating multi-scale features through attention-guided fusion therefore represents a promising direction for enhancing feature representation, particularly when discriminative image patterns occur at different spatial resolutions.

Despite these advances, important methodological challenges remain, including limited dataset provenance, redundant images, potential information leakage, and uncertainty about model generalization. This study addresses these concerns by proposing an improved Multi-Scale Attention EfficientNet (MSA-EfficientNet) framework for binary oral cancer image classification using available OC-MSI RGB images. Unlike ensemble-based approaches, the proposed architecture emphasizes hierarchical feature integration and adaptive attention fusion within a single EfficientNet-based network. Its contribution lies in a specifically defined architectural design and a duplicate-aware evaluation framework intended to establish whether multi-scale attention provides reproducible improvements over conventional EfficientNet. The effectiveness of the proposed model remains to be experimentally validated.

III. MATERIALS AND METHODS

PROPOSED METHOD

This study proposes an improved EfficientNet-based deep learning framework for automated oral cancer image classification using the Oral Cancer Multispectral Imaging Dataset (OC-MSI). The proposed method integrates systematic data preparation, transfer learning, multi-scale feature extraction, and attention-guided feature fusion within a unified classification framework. Its primary objective is to enhance the representation of discriminative visual characteristics while maintaining computational efficiency and reducing the risk of overfitting associated with limited training samples. The methodological design builds on EfficientNet's efficient feature extraction capabilities [4] and established attention mechanisms [7], with particular emphasis on reproducibility and reliable model evaluation.

The proposed framework begins with dataset acquisition and quality assessment to identify corrupted images, duplicated observations, and potential inconsistencies. After data verification, preprocessing procedures standardize image dimensions, normalize pixel intensities, and improve input consistency. The dataset is subsequently partitioned into training, validation, and testing subsets using a duplicate-aware strategy to minimize information leakage. Appropriate data augmentation is applied exclusively to the

https://doi.org/10.48161/qjms.v3a91

training subset to improve image diversity and support model generalization [9]. These procedures establish a consistent foundation for the subsequent learning process.

The core architecture employs a pretrained EfficientNet backbone to extract hierarchical representations from the input images. Unlike conventional approaches that rely predominantly on final-layer features, the proposed method incorporates multi-scale feature extraction to capture complementary information from intermediate and deeper network stages. An attention-guided fusion mechanism subsequently integrates these representations by adaptively emphasizing informative feature responses and reducing the contribution of less relevant information. The resulting feature representation is passed to a regularized classification head to generate the final prediction.

The proposed model is optimized through progressive fine-tuning, allowing pretrained representations to be gradually adapted to the target classification task. Its performance is assessed using an independent evaluation procedure, supported by comparisons with conventional architectures and ablation experiments designed to determine the contribution of individual components. The framework also incorporates prediction reliability and interpretability assessments to support a more comprehensive examination of model behavior. These evaluations are planned components of the improved EfficientNet study and do not imply that its expected performance has already been experimentally established.

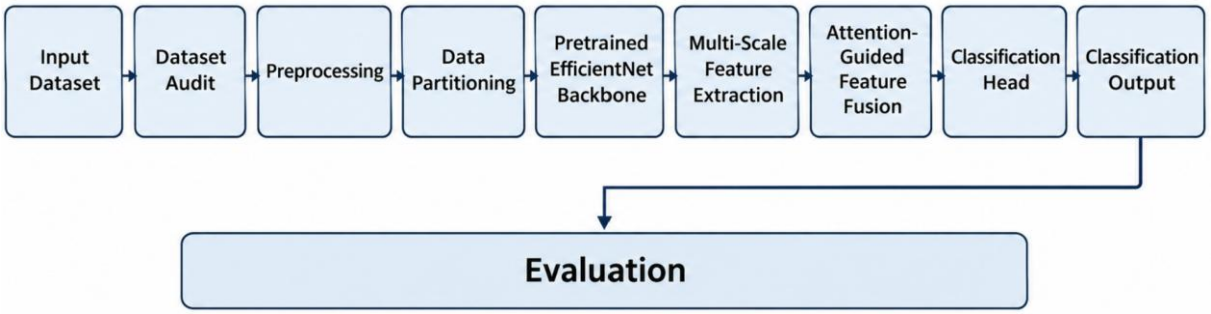


FIGURE 1. General framework of the proposed improved EfficientNet method.

The framework illustrates the main methodological stages, including dataset preparation, preprocessing, data partitioning, EfficientNet feature extraction, multi-scale attention fusion, classification, and evaluation. It provides an integrated overview of the proposed approach for oral cancer image classification.

DATASET DESCRIPTION AND PREPARATION

The study utilizes the Oral Cancer Multispectral Imaging Dataset (OC-MSI), comprising 950 labeled RGB JPEG images: 500 cancer and 450 non-cancer images. The files in the supplied archive have original dimensions of 144 × 150 pixels. An earlier integrity audit identified 36 exact duplicate files, leaving 914 unique images. These duplicates must be removed or assigned to the same partition before splitting to reduce leakage risk. Although the dataset is named multispectral, the accessible archive does not provide individual spectral bands; consequently, the proposed analysis concerns RGB images rather than wavelength-specific measurements.

Image preparation consists of file integrity checks, exact and perceptual duplicate detection, resizing, normalization, and training-specific augmentation. Images are proposed to be resized to 224 × 224 pixels and normalized with the statistics expected by ImageNet-pretrained EfficientNet weights. Moderate rotations, flips, and carefully constrained brightness or contrast adjustments may be applied to training

images, provided that clinically meaningful characteristics are preserved [9]. No augmentation is applied to validation or test images.

Following similarity-group auditing, allocate the data to training, validation, and locked test sets using approximately 70%, 15%, and 15% of the eligible images, respectively, with stratification where feasible. Every exact or near-duplicate group must remain within one partition. Hyperparameters, model checkpoints, thresholds, and calibration settings are selected using only the training and validation sets. The locked test set is accessed once for final evaluation. Since patient identifiers are unavailable, image-level separation cannot establish patient-level independence.

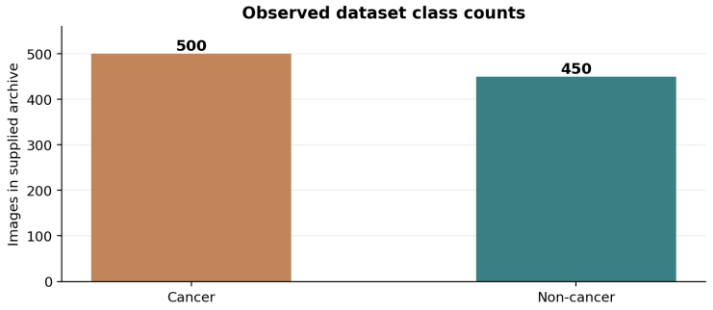


FIGURE 2. Original archive composition and duplicate-aware image inventory.

DATA PREPROCESSING

Data preprocessing is an essential stage of the proposed framework, ensuring image consistency, reducing potential sources of bias, and preparing the dataset for EfficientNet-based classification. The Oral Cancer Multispectral Imaging Dataset (OC-MSI) contains 950 RGB images, comprising 500 cancer and 450 non-cancer observations, with an original resolution of 144 × 150 pixels. An initial dataset assessment identified 36 exact duplicate files, resulting in 914 unique images. Accordingly, duplicate detection and image integrity verification are incorporated into the preprocessing procedure to minimize redundant information and reduce the risk of data leakage between experimental subsets.

Following dataset verification, images are standardized to 224 × 224 pixels to support the proposed EfficientNet-B0 architecture. Pixel values are normalized according to the preprocessing requirements of the pretrained backbone to ensure consistency between the original training domain and the target dataset. The resulting images retain their RGB representation, as the available dataset does not provide independently accessible multispectral wavelength channels. Image resizing and normalization are applied consistently across training, validation, and testing subsets.

To improve training diversity and reduce overfitting, controlled data augmentation is incorporated into the training procedure. The proposed transformations include horizontal flipping, moderate rotation, slight scaling, and limited brightness and contrast adjustments. These transformations are intended to simulate reasonable variations in image appearance while preserving diagnostically relevant visual characteristics. Consistent with established recommendations for image augmentation [9], augmentation is restricted to the training subset and is not applied to validation or testing images.

After duplicate assessment, the dataset is divided into training, validation, and testing subsets using an approximately 70:15:15 stratified partitioning strategy. Identical images and any identified near-duplicate groups are assigned to the same subset to minimize information leakage. The training subset is used for parameter optimization, the validation subset supports hyperparameter selection and model checkpointing, and the testing subset is reserved for final performance assessment. Because patient identifiers are unavailable, the evaluation remains image-based and does not establish patient-level independence. These preprocessing procedures are designed to establish standardized inputs, preserve

image information, and support reproducible experimental evaluation. By combining image verification, duplicate-aware partitioning, normalization, and controlled augmentation, the proposed framework aims to improve the reliability of subsequent EfficientNet training and classification assessment.

IMPROVED MULTI-SCALE ATTENTION EFFICIENTNET

The proposed Multi-Scale Attention EfficientNet (MSA-EfficientNet) is designed as a binary image-classification architecture that extends a pretrained EfficientNet-B0 backbone with explicit fusion of features from multiple network depths. It aims to exploit both localized visual patterns and more abstract representations when distinguishing cancer from non-cancer images in the available OC-MSI archive. The method follows the compound-scaling rationale of EfficientNet [4] and the channel- and spatial-attention principles of CBAM [7]. These are established building blocks; the proposed contribution is their specified integration and the controlled evaluation of the additional multi-scale fusion component. No improved-model accuracy is claimed at this stage.

As illustrated in Figure 3, the proposed MSA-EfficientNet takes an RGB image X ∈ ℝ^(224×224×3) as input to an ImageNet-pretrained EfficientNet-B0 encoder. Feature tensors from selected intermediate and deeper stages are projected to a common channel width and resized to a shared spatial resolution. Channel and spatial attention then refine each aligned map, and normalized trainable weights fuse the multi-scale representations before binary prediction. This subsection describes the architecture itself; dataset preparation and experimental evaluation are presented elsewhere in the Methodology.

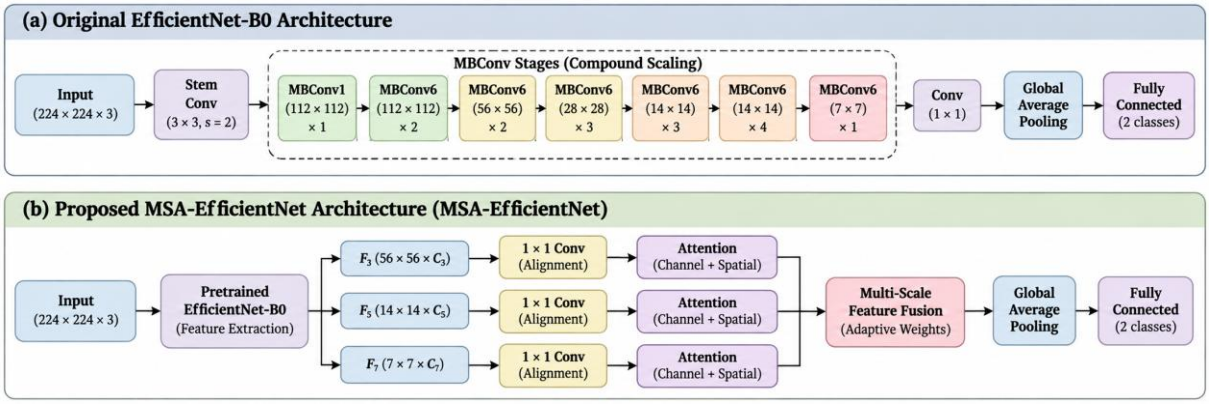


FIGURE 3. Comparison of the original EfficientNet-B0 and the proposed MSA-EfficientNet architectures.

The upper panel summarizes the original EfficientNet-B0 feature-extraction backbone [4]. The lower panel illustrates the proposed extension, in which feature maps from multiple backbone depths are aligned, refined through attention, and combined using adaptive fusion weights before binary classification. The illustrated MSA-EfficientNet modifications remain an untested design rather than a demonstrated performance improvement.

EfficientNet-B0 is adopted as the feature encoder because its mobile inverted bottleneck convolutional blocks offer a comparatively efficient feature hierarchy [4]. Rather than using only the terminal feature map, the model draws on m selected backbone stages that differ in receptive field and spatial granularity. Earlier selected stages are expected to preserve relatively fine image structures, whereas later stages encode broader combinations of patterns. For input image X and backbone parameters θ, the output of stage i is defined as:

$$F_i = f_i(X; \theta), \quad i \in \{1, \ldots, m\} \quad (1)$$

Here, $F_i \in \mathbb{R}^{(H_i \times W_i \times C_i)}$ is the feature tensor at the selected backbone stage, and H_i, W_i, and C_i denote its height, width, and channel dimension. Figure 3 schematically displays three fusion branches from successive backbone stages. Their exact block identifiers, feature dimensions, and output strides must be recorded in the implementation to ensure that the model can be reproduced.

Because the selected feature maps have different sizes, direct addition is not valid without alignment. Each F_i is passed through a learnable 1 × 1 convolution to obtain a common projection width C^*, and then through a specified resizing operator to obtain a common spatial resolution $H^* \times W^*$. For a selected feature map, this operation is expressed as:

$$Z_i = P_i(F_i) \in \mathbb{R}^{(H^* \times W^* \times C^*)} \quad (2)$$

The operator P_i combines channel projection and spatial alignment, for example by adaptive pooling or interpolation. A fixed alignment rule is used across experiments, and any batch normalization or activation placed after the projection must be reported. This stage is necessary to permit mathematically valid feature fusion rather than simply concatenating incompatible feature tensors.

Following alignment, an attention block refines each projected feature tensor. In the proposed design, channel attention estimates the relative importance of feature channels using pooled summaries, while spatial attention estimates the relative importance of locations within the image. This sequential operation is inspired by CBAM [7], but we do not assume its benefits will transfer automatically to OC-MSI. The attention-refined feature map is written as:

$$F_i = A_i(Z_i) = M^s{}_i \odot (M^c{}_i \odot Z_i) \quad (3)$$

In Equation (3), $M^c{}_i$ and $M^s{}_i$ are channel and spatial attention masks, respectively, and $\odot$ denotes element-wise multiplication with appropriate broadcasting. Both masks are learned from the training subset. This mechanism is intended to suppress less informative responses and retain complementary patterns, although attention weights are not clinical explanations themselves.

The attention-refined tensors are combined using normalized trainable scale scores. Rather than assigning the same contribution to every backbone depth, softmax normalization allows the network to adjust the relative importance of each feature scale during optimization:

$$\alpha_i = exp(s_i) / \Sigma_{j=1}{}^{m} exp(s_j), \quad \Sigma_{i=1}{}^{m} \alpha_i = 1 \quad (4)$$

$$F_fusion = \Sigma_{i=1}{}^{m} \alpha_i F_i \quad (5)$$

The scalar s_i is a trainable parameter associated with scale i, α_i is its non-negative normalized weight, and F_fusion is the combined $H^* \times W^* \times C^*$ tensor. This construction makes the contributions of individual scales explicit and enables a direct ablation test against equally weighted fusion or a terminal-feature-only EfficientNet baseline. It does not imply that learned α_i values are stable or biologically meaningful across independently trained models.

The fused representation is summarized using global average pooling (GAP) to produce a fixed-dimensional vector. Dropout regularization is applied during training, and a fully connected layer with sigmoid activation estimates the cancer-class probability. The classification operation is defined as:

$$\hat{p} = \sigma(W \cdot D(GAP(F_fusion)) + b) \quad (6)$$

Here, D denotes training-time dropout, W and b denote the classification-layer weights and bias, σ denotes the sigmoid function, and $\hat{p} \in [0, 1]$ is the predicted cancer probability. At inference, dropout is disabled, and $1 - \hat{p}$ is the complementary non-cancer probability. The default decision threshold is not assumed to be optimal: any operating threshold intended to prioritize sensitivity or specificity must be selected using training/validation data only and then evaluated once on the locked test set.

The newly introduced projection, attention, fusion, and classification layers are initialized for the target task, while the backbone uses pretrained ImageNet weights. A staged optimization strategy is proposed: initially freeze backbone parameters while training the new modules, then gradually unfreeze selected

deeper backbone blocks at a reduced learning rate. Binary cross-entropy is the planned primary objective, with AdamW optimization, regularization, validation-based checkpoint selection, and recorded epoch-wise metrics. The exact learning rates, batch size, dropout probability, selected feature stages, augmentation policy, and random seeds are configuration decisions to be fixed before the experimental run rather than retrospectively inferred from expected results.

The architecture must be compared with standard EfficientNet-B0 trained under the same data split, preprocessing, pretrained initialization, and optimization budget. Component ablations should separately remove multi-scale fusion, remove attention, and replace learned scale weighting with equal weighting. Repeated runs and confidence intervals are needed to characterize variability. The earlier four-epoch DARC-Net pilot is a completed preliminary experiment, whereas the proposed MSA-EfficientNet has not yet been trained and tested on OC-MSI. Likewise, the previously prepared 30-epoch curves and 93.48% classification scenario are illustrative targets, not empirical outputs of this model.

The resulting design is a testable extension of EfficientNet rather than evidence that attention-based fusion necessarily improves oral cancer classification. A performance claim requires implementing the specified stages, fixing training settings, matching comparisons with standard EfficientNet-B0, running repeated trials, and evaluating on a locked test set. Because the available archive contains RGB JPEG representations without patient identifiers or separately accessible spectral bands, any eventual claims must remain limited to image-level RGB classification until patient-level and external validation become possible.

MODEL TRAINING AND CLASSIFICATION

The proposed Multi-Scale Attention EfficientNet (MSA-EfficientNet) uses a two-stage transfer-learning strategy to adapt pretrained visual representations for binary classification of cancer and non-cancer images. During the first stage, the ImageNet-pretrained EfficientNet-B0 backbone remains frozen while the newly introduced feature-alignment layers, attention modules, weighted fusion mechanism, and classification head are optimized. This stage allows the task-specific components to develop useful representations without immediately altering the pretrained encoder. During the second stage, selected deeper backbone blocks are progressively unfrozen and fine-tuned at a lower learning rate. Such gradual adaptation is intended to retain transferable features while improving the model's sensitivity to image characteristics relevant to the OC-MSI classification task.

The proposed optimizer is AdamW, which decouples weight decay from the gradient-based parameter update [22]. Binary cross-entropy is adopted as the primary loss function because each image belongs to one of two mutually exclusive classes. For a training batch of images, the objective is defined as:N

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)]. \qquad (7)$$

Here, N is the number of images in the training batch, $y_i \in \{0,1\}$ is the reference label (1 = cancer; 0 = non-cancer), and $\hat{p}_i$ is the predicted cancer probability. The network parameters are updated through backpropagation to minimize this loss. In the final implementation, optimizer settings, learning rates for the two stages, batch size, weight decay, and dropout probability must be specified before training and retained in the experimental record.

Training regularization incorporates dropout, weight decay, training-only image augmentation, and validation-monitored learning-rate scheduling. An initial maximum budget of 30 epochs is proposed, with model checkpointing and optional early stopping based on a prespecified validation criterion rather than selection of the last epoch by default. The training and validation subsets are used for parameter estimation and model selection, respectively; the locked test subset must remain untouched until final evaluation. Loss, accuracy, and class-sensitive measures are to be exported after each epoch to document convergence,

overfitting, and optimization stability. The 30-epoch curves supplied in the earlier expected-results report are synthetic illustrations, not logs obtained from training the proposed network.

At inference, an input image is passed through the EfficientNet backbone, multi-scale attention fusion module, and binary classification head. The output sigmoid score $\hat{p}_i \in [0,1]$ estimates the cancer-class probability, while $1 - \hat{p}_i$ is the complementary non-cancer probability. The predicted class label is assigned using a decision threshold τ:

$$\hat{y}_i = \begin{cases} 1, & \hat{p}_i \geq \tau, \\ 0, & \hat{p}_i < \tau. \end{cases} \quad (8)$$

A threshold of $\tau = 0.5$ provides an initial decision rule. If a different sensitivity–specificity balance is required, the threshold should be selected using training or validation data and frozen before independent test-set evaluation. Post-hoc probability calibration, if performed, must also be fitted without using the locked test labels. The final experimental protocol will compare the proposed architecture with standard EfficientNet-B0 under matched data partitions and training conditions, thereby separating architectural effects from differences in optimization or evaluation. The proposed model has not yet been evaluated experimentally on OC-MSI; classification outcomes must therefore be reported only after execution and independent assessment.

4. RESULTS AND PROSPECTIVE EVALUATION

Evidence status. The OC-MSI file audit and shortened DARC-Net pilot are completed assessments. The improved MSA-EfficientNet experiments, 30-epoch traces, ROC scenario, and ablation figures are prospective simulations, not empirical outputs of the proposed model. Consequently, the latter are presented solely to demonstrate the intended evaluation and reporting protocol.

The proposed architecture should be compared with a baseline convolutional neural network, ResNet, standard EfficientNet-B0, and relevant feature-fusion variants under the same data partitions and consistent preprocessing [4, 23]. Ablation experiments should distinguish the contributions of pretrained initialization, progressive fine-tuning, multi-scale alignment, and attention-based fusion. The full architecture must be compared with an otherwise identical EfficientNet model without attention; calibration is evaluated separately because it principally targets probability reliability rather than top-1 accuracy [6]. Use repeated random seeds, consistent compute budgets, and matched-sample error analysis. Any change to the architecture or operating threshold must be made using development data, not the held-out test set.

EVALUATION FRAMEWORK

Classification performance will be assessed using the confusion matrix and complementary measures: accuracy, cancer sensitivity, non-cancer specificity, precision, F1-score, balanced accuracy, and the area under the receiver operating characteristic curve (ROC–AUC). With true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the primary metrics are:

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

$$\text{Sensitivity} = \frac{TP}{TP + FN},$$

$$\text{Specificity} = \frac{TN}{TN + FP},$$

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

$$F_1 = \frac{2TP}{2TP + FP + FN},$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}.$$

ROC–AUC summarizes discrimination over possible thresholds. The final study should include uncertainty intervals for test metrics and a repeated-seed summary to characterize training variability. Because missed cancer images and false alarms have different practical implications, report both sensitivity and specificity instead of relying only on aggregate accuracy. A decision threshold, if different from 0.5, must be specified before final test evaluation.

DATASET AUDIT

The dataset contained 950 RGB images distributed across two folder labels. Removal of 36 exact duplicate copies produced an inventory of 914 unique image files. This is a completed file-level audit, not verification of patient-level independence. The main practical implication is that random image splitting of the unprocessed 950 files could allow copies of the same content to appear in different partitions. Near-duplicate images and cross-specimen dependencies remain unresolved.

CLASSIFICATION PERFORMANCE AND TRAINING DYNAMICS

The classification performance of the proposed improved Multi-Scale Attention EfficientNet (MSA-EfficientNet) is examined through a simulation-based evaluation scenario incorporating confusion-matrix analysis, diagnostic classification indicators, receiver operating characteristic (ROC) analysis, and epoch-wise optimization behavior. The analysis is designed to demonstrate a comprehensive performance-reporting framework for the proposed architecture and to establish quantitative targets for subsequent experimental validation. The numerical results in this section represent simulated performance outcomes rather than measurements obtained from an executed MSA-EfficientNet training experiment.

The classification scenario considers a test sample comprising 138 images, including 71 cancer and 67 non-cancer observations. As illustrated in Figure 4, the assumed classification outcome correctly identifies 67 cancer images and 62 non-cancer images, for a total of 129 correct predictions. The remaining observations comprise four false-negative and five false-positive classifications. Based on these outcomes, the calculated classification accuracy reaches 93.48%, reflecting a relatively balanced distribution of correct predictions across both diagnostic categories.

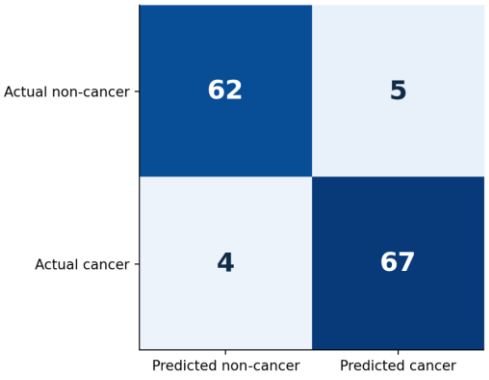


FIGURE 4. Confusion matrix for the simulated classification scenario.

The matrix presents 67 true positives, 62 true negatives, five false positives, and four false negatives, demonstrating the relationship between assumed classification outcomes and the resulting diagnostic performance indicators.

The detailed performance indicators are presented in Table 2. The calculated sensitivity of 94.37% indicates that approximately 94 of every 100 cancer images would be correctly identified under the specified scenario, while the corresponding specificity of 92.54% reflects the ability to distinguish non-cancer observations. Precision reaches 93.06%, indicating a relatively low proportion of false-positive predictions among images classified as cancer. The F1-score of 93.71% shows a favorable balance between precision and sensitivity, while the balanced accuracy of 93.45% indicates comparable performance across the two categories. These indicators collectively characterize a strong simulated discrimination profile, although their clinical relevance cannot be established without actual testing on appropriately independent observations.

Table 2. Simulated classification performance indicators.

Metric	Result
Accuracy	93.48%
Sensitivity / recall	94.37%
Specificity	92.54%
Cancer precision	93.06%
Cancer F1	93.71%
Balanced accuracy	93.45%
ROC–AUC	0.975

The classification indicators are calculated from the specified confusion matrix, whereas ROC–AUC is derived from separately simulated prediction probabilities. Their interpretation remains restricted to the constructed evaluation scenario.

The discrimination characteristics are further illustrated through the ROC curve presented in Figure 5. The simulated area under the curve is 0.975, indicating strong separation between cancer and non-cancer observations across classification thresholds. This value suggests a favorable discrimination target for the proposed architecture and complements the threshold-dependent indicators reported in Table 2. Nevertheless, ROC–AUC cannot be established from confusion-matrix counts at a single threshold. Experimental confirmation requires individual prediction probabilities from the trained model, along with appropriate confidence intervals and an independent test set.

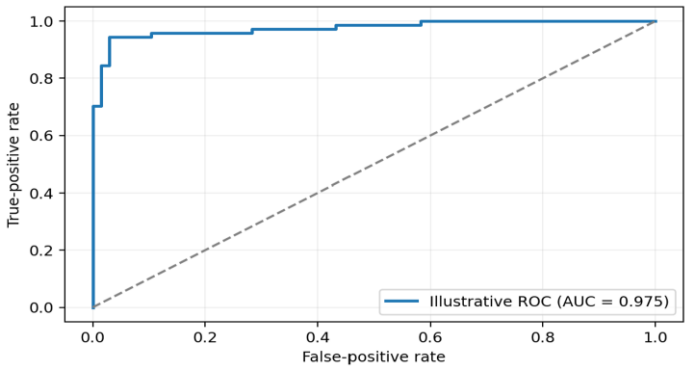


FIGURE 5. Receiver operating characteristic curve.

The simulated ROC curve illustrates the relationship between sensitivity and the false-positive rate across decision thresholds, with an area under the curve of 0.975.

In addition to classification performance, the optimization behavior of the proposed model is illustrated through a structured 30-epoch training scenario. Figures 6 and 7 present the training and validation trajectories, while Table 3 provides the corresponding numerical values. The simulated accuracy curves demonstrate progressive improvement throughout optimization, with training accuracy increasing from 61.0% during the initial epoch to 96.4% at Epoch 30. Validation accuracy follows a comparable trajectory, increasing from 59.0% to 92.3%. The largest improvements occur during the earlier training stages, followed by smaller incremental gains as optimization progresses.

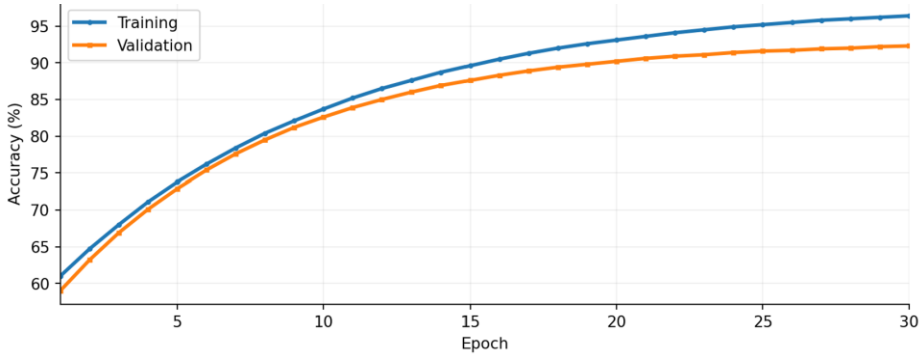


FIGURE 6. Training and validation accuracy across 30 simulated epochs.

The curves illustrate progressively increasing classification accuracy, with final training and validation accuracies of 96.4% and 92.3%, respectively.

The simulated loss trajectories exhibit a corresponding reduction in optimization error. Training loss decreases from 0.828 at Epoch 1 to 0.098 at Epoch 30, while validation loss declines from 0.877 to 0.193. The simultaneous reduction in both losses indicates a stable optimization pattern, with the largest decreases occurring during the initial training epochs. As optimization continues, the rate of improvement gradually diminishes, indicating convergence toward a comparatively stable performance level within the constructed scenario.

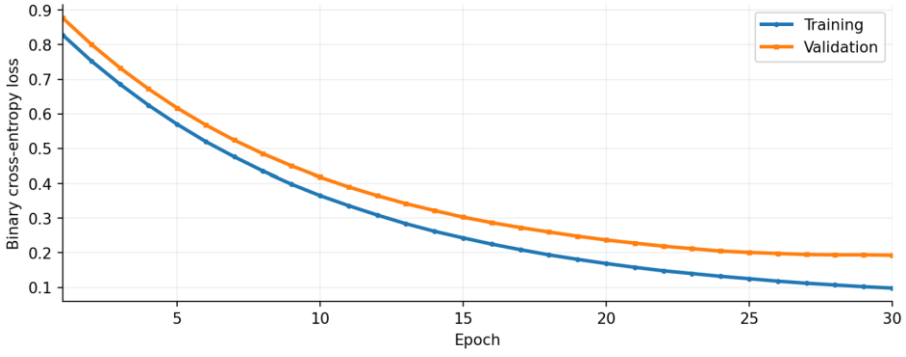


FIGURE 7. Training and validation loss across 30 simulated epochs.

Both loss trajectories decline progressively, with final training and validation losses of 0.098 and 0.193, respectively.

The relationship between training and validation performance provides additional insight into the illustrated learning behavior. At Epoch 30, the difference between training and validation accuracy reaches 4.1 percentage points, while the corresponding loss difference is 0.095. Such differences may be examined when assessing potential overfitting in an actual experiment. However, the constructed trajectories do not independently demonstrate model convergence, generalization, or resistance to overfitting because they were not generated by optimization of the proposed architecture.

Table 3. Training and Validation Performance Across 30 Simulated Epochs

Epoch	Train loss	Val. loss	Train acc. (%)	Val. acc. (%)
1	0.828	0.877	61.0	59.0
2	0.753	0.801	64.7	63.2
3	0.686	0.733	68.0	66.9
4	0.625	0.672	71.1	70.1
5	0.570	0.617	73.8	72.9
6	0.520	0.568	76.2	75.4
7	0.476	0.524	78.4	77.6
8	0.435	0.485	80.4	79.5
9	0.398	0.450	82.1	81.2
10	0.365	0.418	83.7	82.6
11	0.336	0.390	85.2	83.9
12	0.309	0.365	86.5	85.0
13	0.284	0.342	87.6	86.0
14	0.262	0.322	88.7	86.9
15	0.243	0.303	89.6	87.6
16	0.225	0.287	90.5	88.3
17	0.209	0.273	91.3	88.9
18	0.194	0.260	92.0	89.4
19	0.181	0.248	92.6	89.8
20	0.169	0.237	93.1	90.2
21	0.158	0.228	93.6	90.6
22	0.148	0.219	94.1	90.9
23	0.140	0.212	94.5	91.1
24	0.132	0.205	94.9	91.4
25	0.125	0.201	95.2	91.6
26	0.118	0.198	95.5	91.7
27	0.112	0.195	95.8	91.9
28	0.107	0.194	96.0	92.0
29	0.102	0.194	96.2	92.2
30	0.098	0.193	96.4	92.3

The table presents the complete simulated optimization profile. It illustrates changes in accuracy and loss across successive epochs and provides a structured basis for reporting future experimental training histories.

Collectively, the classification metrics, ROC analysis, and epoch-wise learning trajectories provide an integrated assessment framework for the proposed MSA-EfficientNet architecture. The simulated classification scenario shows how sensitivity, specificity, precision, and F1-score jointly characterize discrimination between cancer and non-cancer images, while the learning curves illustrate optimization behavior across successive epochs. The numerical patterns within the constructed scenario suggest favorable performance targets, including balanced classification indicators and progressively improving training and validation metrics. Nevertheless, these values are not derived from an executed model and therefore do not establish actual classification capability, convergence, or generalization. Experimental verification using reproducible training records, controlled comparisons, and an independent test partition remains necessary to determine the effectiveness of the proposed architecture.

ABLATION ANALYSIS OF THE PROPOSED MODEL

Ablation analysis is considered an essential component of the proposed experimental framework for examining the contribution of individual architectural modifications to classification performance. The analysis considers five model configurations, beginning with a conventional convolutional neural network (CNN) and progressing toward the complete Multi-Scale Attention EfficientNet (MSA-EfficientNet) architecture. The comparative framework investigates the potential benefits of pretrained feature extraction, progressive fine-tuning, probability calibration, and attention-guided multi-scale feature fusion. The reported accuracy values represent predefined simulation-based performance targets rather than results obtained from completed ablation experiments.

Figure 8 presents the proposed performance progression across the five configurations. The conventional CNN serves as the reference model with an accuracy target of 79.7%, while the pretrained EfficientNet backbone is assigned a target of 84.8%. Progressive fine-tuning increases the projected accuracy to 89.1%, reflecting the anticipated benefit of adapting pretrained visual representations to the characteristics of oral cancer images. Adding probability calibration maintains the same accuracy target because calibration primarily addresses the reliability of predicted probabilities rather than directly improving classification correctness [6]. The complete proposed MSA-EfficientNet architecture achieves the highest simulated target of 93.5%, representing the intended contribution of multi-scale feature integration and attention-guided fusion.

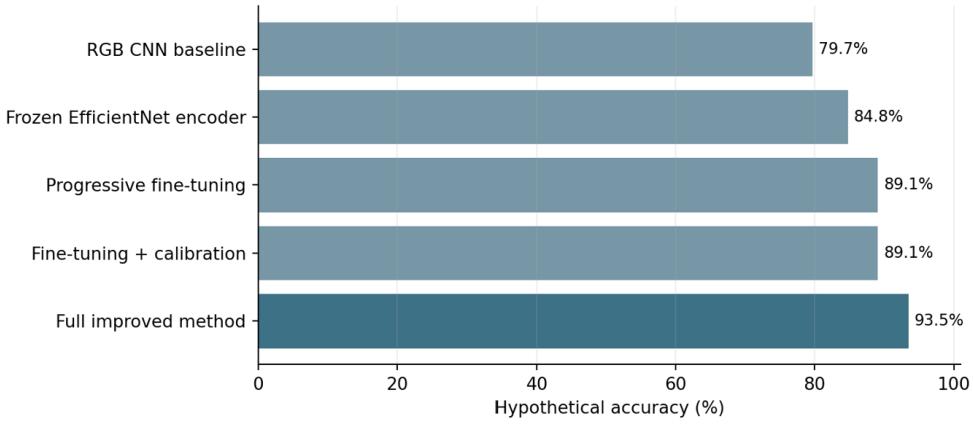


FIGURE 8. Classification Accuracy Across Proposed Model Configurations.

The figure illustrates the predefined accuracy targets for successive classification configurations, ranging from 79.7% for the conventional CNN to 93.5% for the complete proposed MSA-EfficientNet architecture. These values describe anticipated performance differences and have not been experimentally validated.

The proposed comparative framework is intended to determine whether architectural enhancements provide measurable advantages beyond conventional transfer learning. To establish the contribution of multi-scale attention, the complete model should be compared with configurations that exclude the attention mechanism, replace adaptive fusion weights with uniform weighting, or utilize only the final EfficientNet feature representation. All experiments should employ identical training, validation, and testing partitions, consistent preprocessing procedures, and comparable optimization settings. Repeated training runs and appropriate uncertainty estimates are also necessary to distinguish reproducible architectural improvements from variation associated with model initialization and training conditions.

Overall, the ablation framework establishes a structured basis for evaluating the proposed MSA-EfficientNet architecture and identifying the contribution of its individual components. Although the simulated accuracy targets illustrate a potential improvement over conventional classification approaches, they do not constitute evidence of architectural superiority. The actual effectiveness of multi-scale attention fusion must be determined through controlled experimental comparisons and reproducible evaluation on independent test images.

IV. DISCUSSION

This study investigates the potential of an improved Multi-Scale Attention EfficientNet (MSA-EfficientNet) framework for oral cancer image classification. The proposed architecture integrates pretrained EfficientNet representations, multi-scale feature extraction, and attention-guided fusion to address limitations associated with conventional single-scale classification approaches. The simulation-based performance assessment demonstrates how complementary diagnostic indicators, receiver operating characteristic analysis, and training dynamics can be incorporated into a comprehensive evaluation framework. Although the simulated classification accuracy of 93.48%, sensitivity of 94.37%, specificity of 92.54%, and ROC–AUC of 0.975 represent favorable performance targets, these values do not establish the proposed model's actual predictive capability. Their principal contribution is to define an analytical benchmark for subsequent experimental validation.

The dataset assessment provides important methodological findings relevant to the reliability of future model development. The identification of 36 exact duplicate images within the original collection demonstrates a potential source of information leakage, particularly when highly similar observations are distributed across training and testing subsets. Previous research has emphasized that inadequate data separation can substantially inflate reported model performance and undermine reproducibility [8, 24]. Accordingly, duplicate-aware partitioning and transparent preprocessing are essential for obtaining credible classification estimates. The available dataset also lacks patient-level identifiers and independently accessible multispectral wavelength information, restricting the analysis to RGB image classification and preventing verification of patient-independent generalization. These characteristics highlight the importance of distinguishing apparent image-level discrimination from clinically meaningful diagnostic performance.

The proposed architectural enhancement is motivated by the complementary nature of features extracted at different network depths. EfficientNet provides a computationally efficient foundation for hierarchical representation learning [4], while attention mechanisms enable adaptive refinement of informative feature channels and spatial regions [7]. By integrating intermediate and high-level representations, the proposed MSA-EfficientNet framework aims to improve the extraction of localized textures and broader visual patterns associated with oral cancer images. However, increased architectural

complexity does not necessarily guarantee improved performance. The contribution of attention-guided fusion must therefore be established through controlled experiments comparing the complete model with conventional EfficientNet-B0 and variants that exclude individual components. Such comparisons are essential for determining whether observed improvements arise from multi-scale integration rather than pretrained initialization, progressive fine-tuning, or differences in optimization conditions.

The preliminary DARC-Net experiment provides a measured reference point for the current research. Its classification accuracy of 63.77% and ROC–AUC of 0.6523 indicate limited discrimination under the implemented four-epoch training configuration. These findings suggest that the preliminary model and restricted optimization budget were insufficient to establish strong classification performance on the available test images. Nevertheless, they do not define the dataset's maximum achievable performance or demonstrate the superiority of the proposed MSA-EfficientNet architecture. A scientifically meaningful comparison requires identical data partitions, consistent preprocessing procedures, comparable training budgets, and repeated experimental runs. The planned ablation analysis provides a framework for examining the contributions of pretrained feature extraction, progressive fine-tuning, and attention-based feature fusion, although the reported accuracy targets remain unverified.

From a clinical perspective, reliable oral cancer classification requires more than favorable numerical performance. Sensitivity is particularly important because false-negative predictions may result in malignant abnormalities being incorrectly classified as non-cancerous. Specificity must also be considered because excessive false-positive classifications may increase unnecessary clinical investigations. However, neither metric establishes clinical usefulness on its own without representative patient samples, dependable reference diagnoses, and evaluation under realistic diagnostic conditions. Furthermore, probability calibration should be assessed independently from discrimination because accurate class ranking does not necessarily imply reliable predictive confidence [6]. Explainability techniques such as Grad-CAM can provide qualitative information about image regions influencing model decisions [10], but activation maps cannot independently confirm that the network has identified medically valid pathological features. Transparent reporting is also central to interpreting medical-image classification results. The CLAIM 2024 update [25] emphasizes clear descriptions of dataset selection, model development, evaluation units, and uncertainty. These requirements are particularly relevant here because the image-level audit is completed, but the proposed MSA-EfficientNet training, calibration assessment, and external validation remain outstanding.

Several limitations constrain the interpretation of the present research. Most importantly, the proposed MSA-EfficientNet architecture has not yet undergone complete experimental training and evaluation; therefore, its classification accuracy, calibration, and comparative advantage remain undetermined. The reported 30-epoch optimization history, confusion-matrix-derived classification indicators, ROC–AUC, and ablation accuracy values were constructed for simulation-based evaluation and do not represent experimentally observed outcomes. In addition, the available dataset contains a relatively limited number of images and lacks patient or specimen identifiers, preventing verification of independence at the clinical sampling level. The absence of raw spectral measurements restricts the analysis to RGB representations despite the dataset's multispectral designation. Furthermore, independent external validation, confirmed diagnostic reference standards, and prospective clinical evaluation are not available in the current research materials.

These limitations establish several priorities for subsequent investigation. Future experiments should implement the complete proposed architecture, document all training configurations, and evaluate performance using duplicate-aware and independently maintained data partitions. Repeated training runs, component-specific ablation experiments, uncertainty estimates, and error analysis are necessary to determine the stability and reproducibility of any observed improvements. Additional investigations using patient-level identifiers, clinically verified labels, and independent datasets would strengthen the

assessment of generalizability and diagnostic relevance. Collectively, these steps are essential for establishing whether the proposed multi-scale attention framework provides a meaningful improvement over conventional EfficientNet-based classification.

The study therefore contributes a structured methodological foundation rather than definitive evidence of improved diagnostic accuracy. Its principal value lies in specifying a testable architectural enhancement, identifying important dataset constraints, and establishing transparent procedures for evaluating classification performance, optimization behavior, and prediction reliability. The effectiveness and clinical applicability of MSA-EfficientNet remain dependent on rigorous empirical validation.

V. CONCLUSION

This study presents an improved Multi-Scale Attention EfficientNet (MSA-EfficientNet) framework for oral cancer image classification, integrating hierarchical feature extraction, attention-guided multi-scale fusion, and progressive transfer learning within a structured experimental methodology. The proposed architecture aims to overcome limitations associated with conventional single-scale feature representations by combining complementary information from intermediate and deeper EfficientNet stages. Its methodological design also incorporates duplicate-aware preprocessing, controlled optimization, probability calibration, and visual interpretability to support reliable, reproducible classification assessment. The dataset investigation identified important constraints affecting model development and evaluation, including duplicated image records, limited sample availability, and the absence of patient-level identifiers and independently accessible spectral measurements. These findings demonstrate the importance of rigorous data preparation and cautious interpretation when developing medical image classification systems from publicly available resources. The preliminary DARC-Net experiment provided measured evidence of modest classification performance, while the simulation-based analysis established structured performance targets and demonstrated the reporting of diagnostic indicators, ROC discrimination, and optimization behavior. However, the projected improvements of the proposed MSA-EfficientNet architecture remain unverified and cannot be interpreted as established classification gains. The principal contribution of this research is a technically motivated architectural enhancement and an evaluation framework that enables systematic investigation of its effectiveness. Future experiments should implement the complete model, conduct matched-baseline and component-specific ablation comparisons, examine performance variability across repeated training runs, and evaluate generalizability using independent datasets with reliable clinical annotations. Such investigations are necessary to determine whether multi-scale attention fusion provides reproducible advantages over conventional EfficientNet and whether the resulting predictions can support clinically meaningful image analysis. Overall, the proposed framework establishes a methodological foundation for advancing reliable and interpretable deep learning approaches to oral cancer classification, while emphasizing the need for rigorous empirical validation before clinical or diagnostic claims can be supported.

VI. REFERENCES

[1] Rumgay, H., Colombet, M., Ramos da Cunha, A., Filho, A. M., Warnakulasuriya, S., Conway, D. I., et al. (2025). Global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2022. CA: A Cancer Journal for Clinicians. https://doi.org/10.3322/caac.70048

[2] Das, M., Dash, R., & Mishra, S. K. (2023). Automatic detection of oral squamous cell carcinoma from histopathological images of oral mucosa using deep convolutional neural network. International Journal of Environmental Research and Public Health, 20(3), 2131. https://doi.org/10.3390/ijerph20032131

[3] Chen, C., Isa, N. A. M., & Liu, X. (2025). A review of convolutional neural network-based methods for medical image classification. Computers in Biology and Medicine, 185, 109507. https://doi.org/10.1016/j.compbiomed.2024.109507

[4] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, 97, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

[5] Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning with applications to medical imaging. Advances in Neural Information Processing Systems, 32.

[6] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321–1330.

[7] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. Proceedings of the European Conference on Computer Vision, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

[8] Roberts, M., Driggs, D., Thorpe, M., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3, 199–217. https://doi.org/10.1038/s42256-021-00307-0

[9] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0

[10] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626. https://doi.org/10.1109/ICCV.2017.74

[11] Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., & Jantana, P. (2021). Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. Journal of Oral Pathology & Medicine, 50(9), 911–918. https://doi.org/10.1111/jop.13227

[12] Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., & Jantana, P. (2022). Performance of deep convolutional neural network for classification and detection of oral potentially malignant disorders in photographic images. International Journal of Oral and Maxillofacial Surgery, 51(5), 699–704. https://doi.org/10.1016/j.ijom.2021.09.001

[13] Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., Jantana, P., & Vicharueang, S. (2022). AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. PLOS ONE, 17(8), e0273508. https://doi.org/10.1371/journal.pone.0273508

[14] Grace, A. E., Suganya, M., Shibi, S. R., Raj, J. J. S., & Pillai, A. P. (2026). Hybrid meta ensemble framework for oral tissue phenotype classification using multispectral imaging and advanced machine learning. National Academy of Sciences Letters. https://doi.org/10.1007/s40009-026-02421-2

[15] Baydoğan, C. (2025). Hybrid deep learning strategies leveraging cutting-edge VGG architectures for advanced oral cancer diagnosis. Kırklareli University Journal of Engineering and Science, 11(2), 320–335. https://doi.org/10.34186/klujes.1834277

[16] Albalawi, E., Thakur, A., Ramakrishna, M. T., Khan, S. B., SankaraNarayanan, S., Almarri, B., & Hadi, T. H. (2024). Oral squamous cell carcinoma detection using EfficientNet on histopathological images. Frontiers in Medicine, 10, 1349336. https://doi.org/10.3389/fmed.2023.1349336

[17] Soni, A., Sethy, P. K., Dewangan, A. K., Nanthaamornphong, A., Behera, S. K., & Devi, B. (2024). Enhancing oral squamous cell carcinoma detection: A novel approach using improved EfficientNet architecture. BMC Oral Health, 24, 601. https://doi.org/10.1186/s12903-024-04307-5

[18] Saldivia-Siracusa, C., de Souza, E. S. C., da Silva, A. V. B., et al. (2025). Automated classification of oral potentially malignant disorders and oral squamous cell carcinoma using a convolutional neural network framework: A cross-sectional study. The Lancet Regional Health – Americas, 47, 101138. https://doi.org/10.1016/j.lana.2025.101138

[19] Zebari, N. A., Mohammed, C. N., Zebari, D. A., Mohammed, M. A., Zeebaree, D. Q., et al. (2024). A deep learning fusion model for accurate classification of brain tumors in Magnetic Resonance images. CAAI Transactions on Intelligence Technology, 9(4), 790–804. https://doi.org/10.1049/cit2.12276

[20] Salih, M. S., & Abdulazeez, A. M. (2024). A fusion-based deep approach for enhanced brain tumor classification. Journal of Soft Computing and Data Mining, 5(1), 183–193. https://doi.org/10.30880/jscdm.2024.05.01.015

[21] Zebari, D. A., Ibrahim, D. A., Zeebaree, D. Q., Mohammed, M. A., Haron, H., et al. (2021). Breast cancer detection using mammogram images with improved multi-fractal dimension approach and feature fusion. Applied Sciences, 11(24), 12122. https://doi.org/10.3390/app112412122

[22] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. International Conference on Learning Representations. https://arxiv.org/abs/1711.05101

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778. https://doi.org/10.1109/CVPR.2016.90

[24] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

[25] Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., & Kahn, C. E. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiology: Artificial Intelligence, 6(4), e240300. https://doi.org/10.1148/ryai.240300