scientific reports

Check for updates

OPEN

CLASEG: advanced multiclassification and segmentation for differential diagnosis of oral lesions using deep learning

Afnan Al-Ali[1], Ali Hamdi[2], Mohamed Elshrif[3], Keivin Isufaj[3], Khaled Shaban[1], Peter Chauvin[4], Sreenath Madathil[4], Ammar Daer[4], Faleh Tamimi[5] & Raidan Ba-Hattab[5]✉

Oral cancer has a high mortality rate primarily due to delayed diagnoses, highlighting the need for early detection of oral lesions. This study presents a novel deep learning framework for multi-class classification-based segmentation, enabling accurate differential diagnosis of 14 common oral lesions—benign, pre-malignant, and malignant—across various mouth locations using photographic images. A dataset of 2,072 clinical images was used to train and validate the model. The proposed framework integrates EfficientNet-B3 for classification and ResNet-101-based Mask R-CNN for segmentation, achieving a classification accuracy of 74.49% and segmentation performance with an average precision (AP50) of 72.18. The gradient-weighted class activation map technique was applied to the model outputs to enable visualization of the specific areas that were most influential for predictive decisions made by the model. This significantly improves upon the state-of-the-art, where previous models achieved lower segmentation accuracy (AP50 < 50%). The framework not only classifies the lesion type but also delineates the lesion boundaries with high precision, which is critical for early detection and differential diagnosis in clinical practice.

Keywords Oral lesion, Oral cancer, Classification, Segmentation, Deep learning, Early detection

Globally, oral cancer (OC) accounts for more than 300,000 new cases annually, ranking as the sixth most commonly diagnosed cancer. Its mortality rate varies between 1 and 15 per 100,000 individuals, with a 5-year survival rate for OC patients hovering around 50%. The delayed diagnosis of OC contributes significantly to its progression and resultant high mortality rate[1]. Advanced stage diagnoses are associated with a significantly lower five-year survival rate of 39%, compared to 84% for early stage diagnoses. Additionally, late-stage diagnoses incur higher healthcare costs, increase the risk of impairment and disability, and decrease patient quality of life[2]. A crucial aspect of effective OC management is the ability to perform a differential diagnosis.

This process involves distinguishing OC from other oral lesions that may exhibit similar symptoms but require different treatment approaches. The complexity and subtlety of these cases underscore the importance of accurate differential diagnosis in the early stages of the disease.

The majority of OC cases, particularly early carcinomas which constitute over 90% of OC, are asymptomatic, leading to their delayed detection typically at advanced stages. Regular check-ups with primary care providers, including dental exams, are among the most effective methods for early detection of OC. Hence, routine screenings through visual examination are highly recommended, especially in high-risk populations[3]. Dentists play a critical role in this process by identifying abnormal areas in the mouth or throat and, if necessary, consulting specialists for biopsy to confirm malignancy. Despite their importance in screening, many primary care physicians and dentists find diagnosing oral cancerous lesions difficult due to limited training, lack of expertise and experience, and the clinical similarity of various lesions. For example, Fig. 1 illustrates challenging

[1]Computer Science and Engineering Department, Qatar University, Doha, Qatar. [2]MSA University, Giza, Egypt. [3]Qatar Computing Research Institute, HBKU, Doha, Qatar. [4]Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, Canada. [5]Pre-Clinical Oral Health Sciences Department, College of Dental Medicine, QU Health, Qatar University, P.O. Box: 2713, Doha, Qatar. ✉email: rbahattab@qu.edu.qa

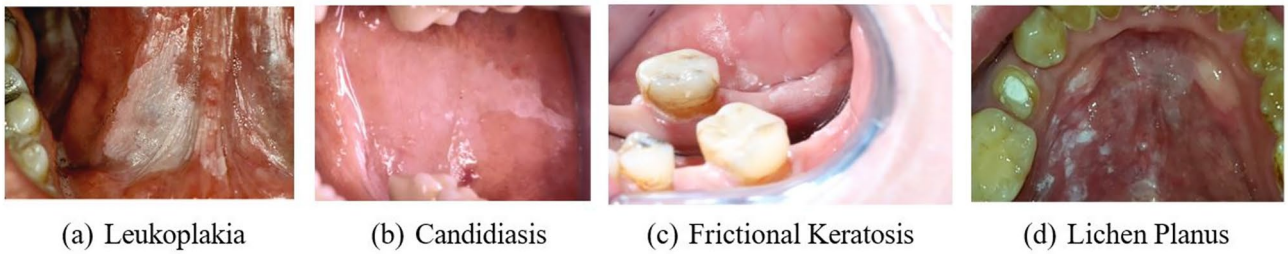

Fig. 1. Showcase of various white oral lesions, each with a unique diagnosis, illustrating the diagnostic challenges.

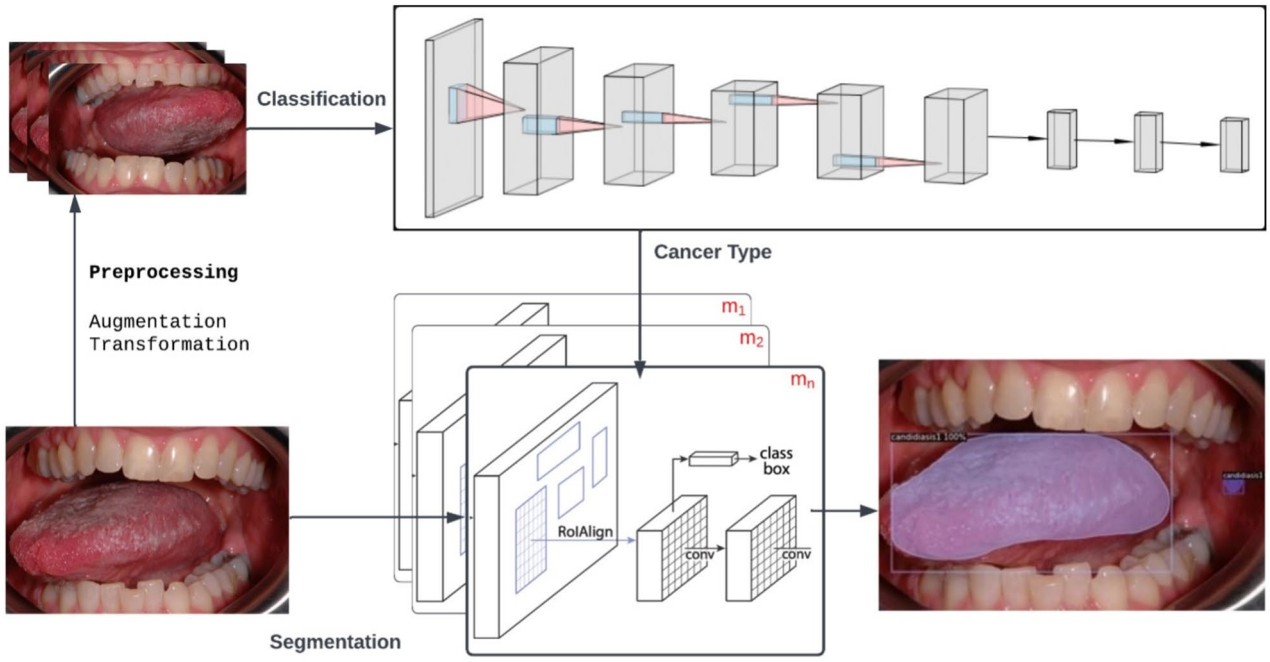

Fig. 2. Operational Workflow of CLASEG: This diagram shows the CLASEG process where an input image is first preprocessed (resizing, augmentation, transformation), then classified into a lesion type. Based on this classification, the image undergoes segmentation using one of the dedicated models (m_1 to m_n), tailored to the identified lesion type.

cases of different oral lesions with similar appearances. This underscores the need for simple, non-invasive, and automated methods to aid healthcare providers in diagnosing oral lesions through image processing.

Specialized imaging methods such as optical coherence tomography and hyper-spectral microscopy have shown promising results in automated image-based OC diagnosis[4,5]. However, these methods require expensive equipment that is rarely available in everyday clinical practice.

Thus, recent research emphasizes the use of more accessible photographic images, although the diagnostic efficacy is limited by factors such as camera type, contrast, and resolution.

Modern devices such as smartphones and iPads facilitate telemedicine strategies for easy and quick remote diagnosis of oral lesions[6]. Telemedicine improves care quality but currently relies on specialist involvement for consultations. Artificial intelligence (AI) diagnosis technologies could address this limitation, allowing fully automated remote consultations through the use of readily available smartphones and empowering primary care providers to offer critical screening services[6,7].

Current attempts at screening oral lesions using AI have shown promising results. However, these technologies have primarily focused on differentiating between malignant and benign lesions and are not able to provide a differential diagnosis of the wide range of lesions that can appear in the oral cavity. To overcome these limitations, we propose the CLASEG model, a deep multiclass CLAssification-based SEGmentation architecture that combines classification with segmentation for diagnosing oral lesions. CLASEG classifies an image into one of 14 oral lesions, followed by segmentation using a corresponding pre-trained model. This approach is tailored to specific oral lesion types and uses annotated photographic images. As shown in the operational workflow diagram of Fig. 2, CLASEG pipeline starts with preprocessing an input image and then classifying it into a

particular oral lesion, e.g., Candidiasis. Following the classification phase, a fine-tuned segmentation model designated to this class of oral lesion is invoked to indicate the area of interest. The contributions of our work are summarized as follows:

- Development of CLASEG: We introduce an advanced deep learning (DL) framework named CLASEG, which uniquely integrates multi-class classification and segmentation. This innovation is specifically designed for the differential diagnosis of a wide variety of oral lesions, making a pioneering approach in the field of oral cancer diagnosis.
- Comprehensive Evaluation: Our study thoroughly evaluates the CLASEG framework on significant dataset, showcasing its superior effectiveness and precision over existing methods. This evaluation includes the fine-tuning and benchmarking of CLASEG against state of-the-art DL models, underscoring its robustness in oral lesion diagnosis.
- Enhancement of Early Diagnosis: CLASEG significantly enhances the early and differential diagnosis of oral lesions by its precise identification and segmentation capabilities.

Advances in AI-Based Telemedicine: Highlighting CLASEG's role in the evolution of telemedicine, our research demonstrates how AI can facilitate remote diagnosis and expand access to early detection of oral cancer. This is particularly beneficial in regions with limited resources, where access to specialist care is often challenging.

This paper is structured as follows: Section "Previous work" reviews related research, Section "Materials and methods" details our proposed methodology; Section "Results" presents the experimental results that validate the effectiveness of CLASEG, Section "Discussion" explores the advantages and potential limitations of CLASEG, and Section "Conclusion" concludes the paper with a discussion on future research directions.

Previous work

Recent advancements in DL have shown significant potential in automating the diagnosis of oral lesions using medical and non-medical imaging[8–17]. In this section, we categorize previous works into three primary themes: classification approaches, segmentation techniques, and integrated classification-segmentation frameworks, followed by a critical analysis of their limitations.

Classification approaches

DL-based classification models for oral lesions have predominantly focused on binary classification tasks, such as distinguishing malignant from benign lesions. EfficientNet-B0 has been used to classify lesions into 'malignant' and 'benign' categories, achieving promising accuracy rates[18]. Similarly, VGG19 and ResNet-101 models have been employed to differentiate between 'non-referral' and 'referral' cases[19]. Other studies have developed classifiers for squamous cell carcinoma and normal controls based on the Dense network architecture[20], and tongue lesion classification has been implemented using AlexNet, GoogLeNet, and SqueezeNet with accuracy rates ranging from 83 to 98%[21].

Beyond photographic images, some researchers have explored alternative imaging types. For example, hyperspectral medical images were classified with GoogLeNet and InceptionNet-v3[22], while 3D-CNN architecture was used on CT images to differentiate benign and pre-cancerous tongue lesions[23]. These studies demonstrated high accuracy (up to 91%), but their reliance on specialized imaging modalities limits their clinical applicability in resource-constrained settings[24].

Segmentation techniques

Segmentation plays a crucial role in delineating lesion boundaries for improved diagnosis. Several OC segmentation models have utilized region-based convolutional neural networks (CNNs)[25–28]. For instance, Faster R-CNN architecture has been successfully applied to colon polyp detection[29] and lesion detection on mammograms[30]. In the oral cavity, Faster R-CNN achieved a precision of 76.67% and recall of 82.14% for OC detection[31].

Semantic segmentation techniques, such as U-Net and Mask R-CNN, have also shown strong performance. U-Net with EfficientNet-B7 achieved a Dice score of 0.926 without test-time augmentation (TTA) and 0.929 with TTA for oral lesion segmentation[4]. Another approach used Mask R-CNN with ResNet-101 as a backbone to detect herpes labialis and aphthous ulcers, achieving Dice coefficient scores of 0.774 and 0.714, respectively[32]. However, most segmentation models are limited to one or two lesion types and often require extensive preprocessing or specialized imaging techniques[32–34].

Integrated classification-segmentation frameworks

Few studies have integrated classification and segmentation in a unified framework. Welikala et al. developed a two-stage model combining ResNet-101 for image classification and Faster R-CNN for object detection, achieving F1-scores of 87.07% and 78.30% for lesion identification and referral detection, respectively[5]. Tanriver et al. proposed a model for detecting oral lesions and classifying them into benign, potentially malignant, and carcinoma categories, with their best segmentation model reaching a Dice score of 0.929[4]. Song et al. employed a cascaded CNN framework for squamous cell carcinoma detection, reporting a sensitivity of ~ 95% and specificity of ~ 89%[33].

Tobias et al. presented a DL-based model using VGG Image Annotator for lesion segmentation and ResNet for classification, achieving an F1-score of 77%[34]. While these studies illustrate the potential of integrated frameworks, they are often constrained by small datasets and a limited range of lesion types, reducing their generalizability.

Critical analysis and justification for CLASEG

Despite the significant progress in DL-based classification and segmentation, several limitations persist:

Dataset Constraints: Many studies rely on small, unbalanced datasets, which limits the robustness and generalizability of the models.
Limited Lesion Diversity: Most models focus on a narrow range of lesion types, often binary or few-class classification, making them less useful in clinical practice.
Reliance on Specialized Imaging: Techniques involving hyperspectral imaging or CT scans are effective but impractical for widespread use due to equipment constraints.
Lack of Comprehensive Approaches: Integrated models that simultaneously classify and segment a diverse range of lesions are rare, leaving a significant gap in clinical applicability.

To address these limitations, we propose **CLASEG**, a DL framework that integrates multi-class classification and segmentation for diagnosing 14 types of oral lesions. Unlike prior works, CLASEG utilizes photographic images from everyday clinical settings, making it more accessible and practical for early diagnosis and treatment.

Materials and methods

We employ a data-centric approach to develop end-to-end DL models for OC classification and segmentation. This involves fine-tuning several state-of-the-art models, pre-trained on ImageNet, to adapt to our specific dataset, thereby the challenge of limited data availability. For classification, models including VGG (16, 19, 11-bn), EfficientNet (B0 to B4), MobileNet-v2, ResNet (50, 101,152), DenseNet, AlexNet, and SqueezeNet (1.0, 1.1) are fine-tuned by adjusting all parameters to our domain. Specifically, we resize the output layer of these models from 1000 to 16 nodes to match our classification needs. For segmentation, Mask R-CNN with a ResNet-101 backbone is fine-tuned for precise lesion demarcation.

Dataset collection and annotation

This study received ethical approval from the Institutional Review Board (or Ethics Committee) of McGill University, Montreal, Canada (protocol code A07-M40-21B, 25th August 2021). We utilize a dataset of 2,072 high-resolution JPEG images of various oral lesions, captured under standard dental office conditions, provided by Prof. P.C. from McGill University. These images, ranging in resolution from 600 600 × 450 to 1000 × 750 pixels, were obtained with digital cameras and smart phones, ensuring high-quality visual data. Patient consent was secured for the use of these deidentified images for research purposes. The dataset includes 14 types of oral lesions: Amalgam Tattoo, Aphthous Ulcer, Candidiasis, Geographic Tongue, Tongue/ Cheek Chewing, Mucocele, Melanotic Macule, Hemangioma, Papilloma, Fibroma, Pyogenic Granuloma, Leukoplakia/Hyperkeratosis, Lichen Planus, and Squamous cell Carcinoma. Histopathological and clinical examinations were performed by an oral pathologist with 30 years of experience for definitive diagnosis. Masking the areas of interest for each image is done by a specialized dentist with 15 years of clinical experience using LabelMe annotation tool, facing challenges due to variations in size, location, color, and shape of lesions (Fig. 3), which introduces significant complexity to the classification and segmentation tasks as depicted in Fig. 1.

Data preprocessing

We consider each image as a two-dimensional function F (x, y), where x and y represent spatial coordinates and F of any pairs of (x, y) is the intensity of that point, i.e., pixel of the image. Images with (w x h), where w is the width of the image and h is its height, are resized to fit the input requirements of each model, maintaining aspect ratios to avoid distortion. For example, MobileNet requires the images to be 224 × 224x3, and EfficientNet-B4 to be 380 × 380x3. To accomplish this, consider r = w/h as the original aspect ratio and r = wnew/hnew as the new required aspect ratio.

Resizing results in the height and width calculated as follows[35]:

$$\begin{aligned} h_{new} &= \frac{W_{new}}{r}, if r > r \\ w_{new} &= h_{new} * r, if r < r \end{aligned} \quad (1)$$

Images that are larger than expected sizes are truncated, while smaller images are interpolated to increase their size by estimating the intensity and color of the pixel based on the values of neighbors. Given the dataset's imbalance with the number of images per class varies between 37 and 199 (Fig. 4). we perform augmentation techniques such as resizing, random cropping, and horizontal flipping to enrich the training set. The image tensors are normalized to scale pixel values between 0 and 1, further facilitating model training by standardizing input data. The images are shaped as (c, h, w), where c is the number of channels and (h, w) are the height and width. We normalize the images with the mean (μ) and the standard deviation (SD) by first calculating the channel mean, then subtracting it from each of the three input channels. The result is divided by the standard deviation of the channel as follows[36]:

$$output\,[channel] = \frac{input\,[channel] - mean\,[channel]}{std\,[channel]} \quad (2)$$

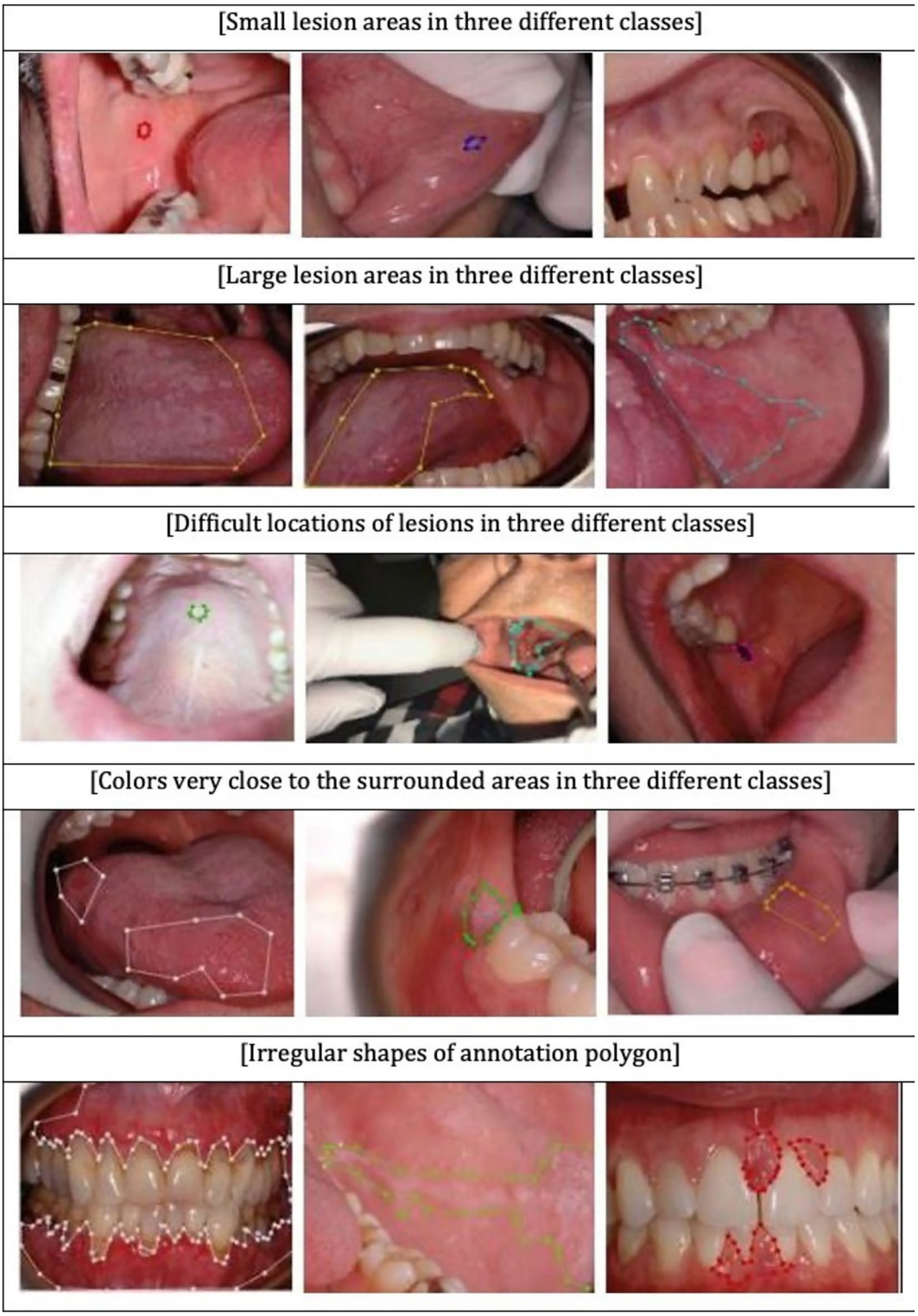

Fig. 3. Sample of the dataset images with different size, location, color, and shape of lesions.

Deep multiclassification-based segmentation (CLASEG)

Our approach integrates classification and segmentation in a two-part model. Initially, images undergo classification to identify lesion types. Subsequently, a dedicated segmentation model for each class isolates the lesion area. The dataset is split into 50% training, 30% validation, and 20% testing subsets through stratified sampling, ensuring proportional representation of each lesion type. Classification models are trained to minimize cross entropy loss across C classes, where, in our case, $C = 14$, as follows[37]:

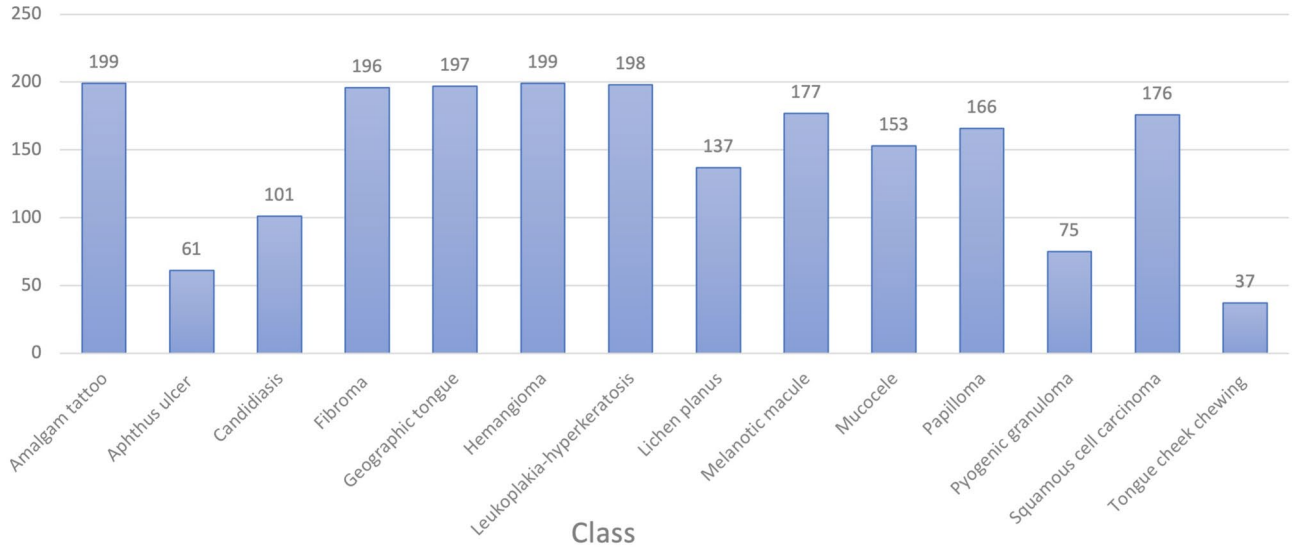

Fig. 4. The distribution of images per category.

$$L_{class} = -\frac{1}{m}\sum [y_i ln\,(\widehat{y_i}) + (1 - y_i) ln(1 - \widehat{y_i})] \tag{3}$$

where m is the number of training samples, $i = 1 : m$, y_i is the target class, $\widehat{y_i}$ is the predicted class, and ln is the natural logarithm.

The output of the trained model is a vector of equal size that represents the number of classes, C, in the form of scores; here a Softmax function is applied to make these scores sum to 1 and each individual class boundary is between 0 and 1, denoting the probability of prediction. The Softmax equation is defined by taking the exponential of the output vector divided by the sum of the exponential of all outputs as follows[38]:

$$Softmax\,(\widehat{y_i}) = \frac{e^{\widehat{y_i}}}{\sum_{i=1}^{C} e^{\widehat{y_i}}} \tag{4}$$

For segmentation, separate models $k_i, i = 1{:}\,C$ are trained for each class, utilizing a combination of classification, bounding box prediction, and mask prediction to accurately delineate lesion areas. That is, during testing, each image is passed through the classifier, and based on the predicted class, the segmentation model of the same class is utilized for segmenting the area of interest, i.e., OC lesion, related to the specific category. For each class i, the segmentation model k_i produces three probabilities, one for predicting the class of the OC lesion (P_{cls}), one for predicting the bounding box (P_{box}) and one for predicting the mask (P_{mask}) as follows[39]:

$$k_i\,(P_{cls}, P_{box}, P_{mask}) = backbone_{model} + RPN + RoI\_Align \tag{5}$$

where backbone model represents a pre-trained CNN model for image classification task, RPN is the region proposal network, and RoI_Align is the region of interest alignment layer.

We use ResNet101-FPN for $backbone_{model}$. The RPN is used to identify multiple objects within the image by generating "proposals" for the regions where these objects are located. The RPN consists of a classifier and a regressor. The classifier assesses whether a proposal likely contains the target (assigning a probability score), while the regressor adjusts the coordinates of the proposals for better alignment with the target objects. The backbone network produces a feature map, over which a small network slide. The central point of this sliding window is termed an "Anchor".

Given the image's scale $(size(w * h))$ and aspect ratio r, with a standard count of 3 for both w and h, up to 9 proposals can be generated for each pixel, resulting in a total of $(w * h * 9)$ Anchors across the image. An Anchor achieving an intersection over union (IoU) greater than 0.7 with a ground truth object is labeled as containing that object. The RoI Align accurately locates the bounding box for the pixels of each object by calculating their floating-point positions using the bilinear interpolation method[28].

The model's total loss function is a sum of the individual loss functions: classification loss (Lc_{class}) localization loss (L_{box}), and segmentation mask loss (L_{mask})[39]:

$$L_{total} = L_{class} + L_{box} + \ldots L_{mask} \tag{6}$$

L_{class} is defined earlier in Eq. 3.

$$L_{box} = \sum_{i \in \{x,y,w,h\}} L_i^{smooth_{(t_i^u - v_i)}} \tag{7}$$

where μ is the true class ∈ 1, 2, …, C, is the predicted bounding boxes, and vi is the actual bounding boxes. Lbox measures the difference between tu and vi. The L_{smooth} function is preferred for bounding box regression in some object detection frameworks due to its reduced sensitive to outliers[28] compared to other regression loss functions, such as L2, and is defined as:

$$L_1^{smooth}(x) = \{|x| - 0.5, if|x| \in 10.5\left(x^2\right), otherwise \tag{8}$$

where x is the difference between the predicted and actual bounding box coordinates. The segmentation model generates a mask of dimension nxn, for each region of interest across C classes, leading to a total output size of $C * n^2$. The mask L_{mask} for the c^{th} mask, corresponding to the ground truth class c, is defined by[39]:

$$L_{mask} = -\frac{1}{m^2}\sum_{i \geq 1, j \geq n} [y_{ij} lnln\left(\widehat{y}_{ij}^c\right) + (1 - y_{ij}) lnln\left(1 - \widehat{y}_{ij}^c\right)] \tag{9}$$

Evaluations

The primary metric for assessing the classification stage's performance is the accuracy, which quantifies the model's ability to correctly predict the test images' classes. In multi-class classification scenarios, accuracy is calculated as follows[40]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

where *(TP)* and *(TN)* represent the number of true positives and true negatives (correctly predicted classes), while *(FP)* and *(FN)* denote false positives and false negatives (incorrect predictions), respectively.

For this study, we evaluated 16 pre-trained DL models for the classification tasks, including VGG (16, 19, 11-bn), EfficientNet (B0-to-B4), MobileNet-v2, ResNet (50, 101, 152), DenseNet, AlexNet, and SqueezeNet (1–0, 1–1).

The segmentation performance was assessed using the precision-recall curve, from which the average precision (AP) and the IoU metrics were derived. IoU was evaluated across a reange of thresholds from 0.5 to 0.95 (in 0.05 increments), denoted as *AP@* [0.5:0.05:0.95], to provide a comprehensive measure of localization accuracy[41]:

$$AP_{class} = \frac{1}{\#thresholds}\sum_{IoU \in thresholds} AP[class, IoU] \tag{11}$$

where:

$$AP\left[class, IoU\right] = \int_0^1 pr\left(r\right) dr \tag{12}$$

The AP_{50}, AP_{75}, and AP large extensions were included to highlight performance across different thresholds and object sizes. AP_{50} indicates that the model employed a 0.5 threshold value to eliminate extraneous bounding boxes. This is the standard threshold value for most models. AP_{75} indicates that the model has employed a 0.75 threshold value. *AP* large indicates that the model has assigned a mean *AP* score based on the large objects. Given our focus on segmentation, we opted not to consider classification metrics such as sensitivity or specificity, finding accuracy sufficient for progressing to segmentation stages[41].

Visualization and interpretation

We applied a visualization technique called gradient-weighted class activation map (Grad-CAM)[42] to the model outputs. This technique depicts the detailed locations on the photographic image that influenced the prediction.

Results

Through various experiments, we demonstrated our methodology's efficacy. Employing the Detectron2 framework and Mask R-CNN with a ResNet-101 FPN backbone, we processed a dataset of 2,163 images, dividing them as 50% for training, 30% for validation, and 20% for testing using stratified sampling to maintain class proportionality. Optimal segmentation performance was observed with specific parameters achieving an AP50 of 32.7 after parameters tuning across 3,000 epochs, with a batch size of 64 and a learning rate of 0.0001 (Table 1).

Model	Epochs	*AP*	AP_{50}	AP_{75}	*APl*
Detectron2	3000	16.277	32.693	14.808	21.324

Table 1. The results of applying Detectron2 (based on mask R-CNN R-101-FPN-3x) on full dataset. The model parameters: batch size 64, no. of epochs 3000, and learning rate 0.0001.

Class	# test images	Correctly classified	Accuracy
Hemangioma	40	37	92.5%
Geographic tongue	40	37	92.5%
Amalgam tattoo	40	33	82.5%
Fibroma	40	31	77.5%
Leukoplakia-hyperkeratosis	39	26	66.6%
Squamous cell carcinoma	38	26	68.4%
Melanotic macule	36	31	86.1%
Papilloma	34	18	52.9%
Mucocele	31	20	64.5%
Lichen planus	27	22	81.4%
Candidiasis	21	7	33.3%
Pyogenic granuloma	16	8	50.0%
Aphthus ulcer	13	3	23.1%
Tongue/cheek chewing	8	3	37.5%

Table 2. The results of per class classification accuracy using EfficientNet-B3 classification model.

#	Model	Image	Batch	Epochs	Accuracy	*AP*	AP_{50}	AP_{75}	*APl*
1	ResNet-50	224×224	64	30	63.35%	39.94	64.67	43.52	40.28
2	**ResNet-101**	224×224	64	30	64.07%	**44.05**	**72.18**	**48.27**	**44.67**
3	ResNet-152	224×224	64	30	66.90%	41.55	67.68	46.35	42.36
4	VGG16	224×224	64	30	53.19%	38.32	64.67	39.64	38.62
5	VGG11-bn	224×224	64	30	63.36%	41.67	67.67	45.70	42.24
6	VGG19	224×224	64	30	51.53%	37.43	64.59	38.88	38.34
7	DenseNet-121	224×224	64	30	68.32%	39.53	66.28	41.95	43.56
8	AlexNet	224×224	64	30	47.99%	30.43	53.50	31.56	30.77
9	Squeeze-Net1-0	224×224	64	30	42.31%	29.09	51.69	30.37	29.46
10	Squeeze-Net1-1	224×224	64	30	43.97%	35.78	59.90	38.02	36.28
11	EfficientNet-B0	224×224	64	30	69.97%	41.56	66.06	44.95	42.47
12	EfficientNet-B1	240×240	64	30	73.52%	42.81	70.02	47.74	43.57
13	EfficientNet-B2	260×260	64	30	68.55%	42.51	69.30	46.63	43.21
14	**EfficientNet-B3**	300×300	64	30	**74.49%**	41.30	67.88	45.03	41.96
15	EfficientNet-B4	380×380	64	30	71.16%	43.74	69.97	48.24	44.52
16	MobileNet-v2	224×224	64	30	70.21%	41.28	69.36	43.38	42.10

Table 3. The results of segmentation based on the predicted class; Eps refers to no. of epochs; using a learning rate of 0.0001. Bold values indicate the highest segmentation performance, representing the value of Intersection over Union(IoU), Average Precision (AP), or accuracy percentage

We used the Detectron2 framework, an open-source, high-performance library designed for object detection and segmentation, to implement Mask R-CNN with a ResNet-101 backbone and Feature Pyramid Network (FPN). This combination was chosen for its well-documented success in medical image analysis. ResNet-101 enables robust feature extraction due to its deep architecture, capturing intricate lesion patterns across different oral lesion types. The FPN further enhances the model's ability to detect lesions at multiple scales, ensuring consistent performance across lesions of various sizes. Mask R-CNN provides pixel-level segmentation, which is crucial for accurately delineating lesion boundaries, especially in complex and heterogeneous lesion types. Additionally, Detectron2's optimized implementation improves computational efficiency, enabling faster training and evaluation. This framework's modularity also offers flexibility for future improvements, such as multi-label segmentation and integration with other backbone architectures.

We further refine our approach by training 14 class-specific segmentation models, then integrating these with 16 different pre-trained classification models, including EfficientNet-B3, which showed promising classification accuracy without overfitting. This hybrid model significantly improved segmentation accuracy, demonstrating that class-specific segmentation models enhance performance (Table 2).

After training, each image was classified, and based on the predicted class, the corresponding segmentation model from the saved models was used to segment the area of interest. This procedure was applied to each of the 16 classification models. Despite the common characteristics among different lesion types, the baseline segmentation model's accuracy was low, as shown in Table 2. Therefore, we observed that the AP and AP50 values doubled the baseline model's results, as indicated in Table 3. These findings confirm that using a specific

Model	Epochs	*AP*	AP_{50}	AP_{75}	*APl*
16-Detectron2	3000	45.23	74.22	49.73	45.87

Table 4. The results of segmentation considering the 100% classification accuracy using 16- Detectrons based on image classes. The model parameters: batch size 64, no. of epochs 3000, and learning rate 0.0001.

Model	Epochs	*AP*	AP_{50}	AP_{75}	*APl*
EfficientNet-B3	30	30.45	50.61	32.53	30.87

Table 5. The results of segmentation after filtering the results based on the matching of predicted class and ground truth for EfficientNet-B3 with image size 300×300.. The model parameters: batch size 64, no. of epochs 30, and learning rate 0.0001.

segmentation model matched to the input image's class significantly improves segmentation accuracy, with EfficientNet-B3 achieving the best classification results and ResNet-101 the best segmentation results.

In addition to evaluating the segmentation performance under realistic conditions, we conducted a theoretical experiment assuming 100% classification accuracy, the results of which are shown in Table 4. While this scenario is not practically achievable, it serves as an upper-bound benchmark to explore the full potential of our segmentation models. This experiment provides critical insights into the segmentation model's performance when isolated from classification errors.

Under this idealized condition, the segmentation model achieved an AP50 of 74.22, demonstrating that classification accuracy has a direct impact on segmentation performance. This analysis helps in understanding how much of the segmentation error is attributable to misclassification versus limitations in the segmentation model itself. Moreover, it offers a theoretical target for future improvements, emphasizing the importance of refining classification accuracy to enhance the overall system performance.

This approach is consistent with practices in multi-stage machine learning pipelines, where the performance of each component is analyzed independently to identify optimization opportunities. Table 4 helps highlight these insights and supports a deeper understanding of the interplay between classification and segmentation.

The segmentation results listed in Table 3 represent the average evaluation metrics, considering both true positives (when the predicted class matches the actual class and the segmented area correctly corresponds to that class) and false positives (when the predicted class does not match the ground truth but segmentation of the area of interest or a part of it still occurs). By focusing solely on cases where the predicted class matched the ground truth, we found that our CLASEG methodology yields promising results compared to the baseline model. This approach was validated further by recalculating these metrics using only the best-performing pre-trained classification model, EfficientNet-B3, with the results displayed in Table 5.

Figure 5, showcases some segmented images based on different classification models, highlighting the high confidence of each class and the accuracy of the segmented areas. In a final experiment, we evaluated the impact of correct class prediction on segmentation accuracy by passing the input image through all 14 segmentation models based on Mask R-CNN with a ResNet-101 backbone. The model with the highest confidence was selected based on the segmentation results for each one. The average evaluation metrics for those samples across the entire test set are presented in Table 6.

To draw statistical conclusions on comparing the models' results, we train and test them 5 times each, recording their prediction accuracy. Next, we conduct Tukey's honest significance test (HSD) to determine the statistical significance of the differences among the models' results. We select our best model, EfficientNet-B3, for comparison against 7 others, as shown in Table 7. The results clearly show that EfficientNet-B3 significantly outperforms EfficientNet-B0, EfficientNet-B2, EfficientNet-B7, ResNet-101, and ResNet-152. However, when compared to DenseNet-121 and EfficientNet-B1, we do not reject the null hypothesis. From these outcomes, we conclude that EfficientNet-B3 is significantly superior to most models. For the two models where we fail to reject the null hypothesis, EfficientNet-B7 still outperforms them with mean differences being 1.37% and 2.15% higher respectively.

Figure 6 shows the activation maps of four representative oral lesions generated by Grad-CAM, alongside the original photographic images that highlight the lesion location.

Despite considering only the highest confidence for each class, the AP did not surpass the results from our second experiment, which depended on the predicted class to select the specific segmentation model. The outcomes of the last experiment closely matched those of the original baseline model, demonstrating that accurate image class prediction and subsequent lesion detection is a reliable method for achieving precise segmentation.

Discussion

In this study, we developed a deep multiclassification-based segmentation (CLASEG) framework capable of diagnosing 14 different types of oral lesions. We fine-tuned state-of-the-art image classification deep neural network architectures such as EfficientNet, MobileNet, ResNet, DenseNet, and SqueezeNet. These models demonstrated robust classification performance, with EfficientNet- B3 achieving the best results with a test accuracy of 74.49%. EfficientNet-B3, designed for high accuracy with low computational cost, employs efficient

Fig. 5. Samples of the segmented images from different detectors.

Model	Epochs	*AP*	AP_{50}	AP_{75}	*APl*
16-Detectron2	3000	20.90	32.46	24.13	21.23

Table 6. The results of 16-Detectros segmentation based on max confidence score. The model parameters: batch size 64, no. of epochs 3000, and learning rate 0.0001.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Reject
efficientnet_b0	efficientnet_b3	3.70%	0.0038	0.0111	0.0629	TRUE
efficientnet_b2	efficientnet_b3	6.07%	0.001	0.0348	0.0866	TRUE
efficientnet_b3	efficientnet_b7	−4.43%	0.001	−0.0702	−0.0183	TRUE
efficientnet_b3	resnet101	−4.81%	0.001	−0.0729	−0.0233	TRUE
efficientnet_b3	resnet152	−4.65%	0.001	−0.0725	−0.0206	TRUE
densenet121	efficientnet_b3	1.37%	0.7864	−0.0123	0.0396	FALSE
efficientnet_b1	efficientnet_b3	2.15%	0.2767	−0.0045	0.0474	FALSE

Table 7. Tukey's Honest Significant Difference (HSD) test for EfficientNet-B3.

and compound scaling to optimize accuracy and efficiency. This makes it an excellent choice for applications where computational resources are limited. The model's superior performance in our study can be attributed to its training on a substantial dataset, enabling it to learn complex features of oral lesions more effectively compared to other models fine-tuned on the same dataset. Table 2 illustrates the classification accuracy for each lesion type in our dataset, highlighting the trend that accuracy decreases with fewer training images per lesion type. This underscores the challenge posed by our unbalanced dataset, affecting both classification and segmentation models. To mitigate this, we initially compared our results with a baseline model, where segmentation was

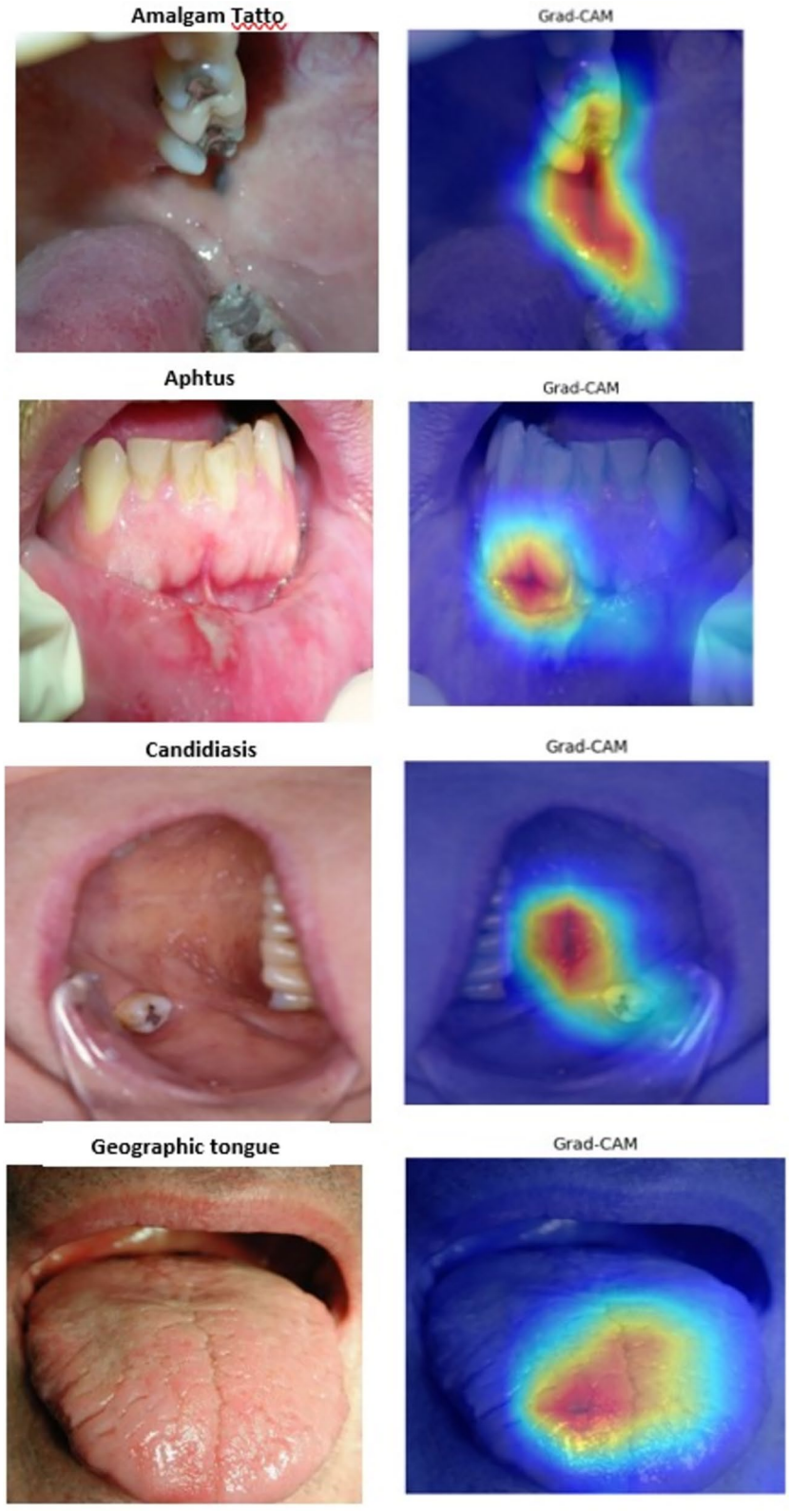

Fig. 6. Example of visualizations used for oral lesion detection, employing Grad-CAM explanation technique. These visualizations help identify the regions most influential in the model's predictions.

performed irrespective of class type (Table 1). We discovered that employing a specific segmentation model for each lesion type significantly enhanced detection accuracy, as evidenced by the improved AP50 scores from 32.69 to 72.18 (Table 3). Assuming 100% classification accuracy with EfficientNet-B3 model, we observed a slight increase in model performance, with AP50 reaching 74.22. While this improvement may seem modest, it indicates that even with perfect classification, performance gains are limited.

Additionally, we conducted an experiment focusing solely on correctly classified cases, which were then segmented with their respective models. This approach still demonstrated an improvement over the baseline model (Table 5). In a further experiment based on the maximum confidence score for segmentation, the AP50 value of 32.46 was nearly identical to the baseline (32.69). highlighting the challenges posed by the high similarity among lesions (Table 6).

Several studies have explored models that segment before classifying oral cavity lesions. For example,[4] presented a model that classified lesions into benign, potentially malignant disorders, and carcinoma, achieving an AP50 of up to 78.92 through dataset augmentation and focused segmentation based on the Mask R-CNN with a ResNet-101 FPN + Test-Time Augmentation (TTA) backbone model. The model was built using a dataset of 652 images. Other studies,[28,32,43,44] have different approaches, such as using the VGG image annotator to segment 1,636 cancerous or precancerous photographic images, which were then classified using a deep ResNet model with 152 layers[28]. All the images were pre-processed to improve the generalization of the model. In another study by[43], 89 photographic images were sliced into 32×32 patches, which were then normalized and fed into a CNN-based architecture model for ensemble learning and classification.

A large number of CNN models were trained and then the ones achieved high performance are ensembled. Despite achieving over 85% classification accuracy, the study focused solely on oral ulcer lesion type and utilized a small dataset, limiting its generalizability. To improve the interpretability of the model, the authors of another study,[44], used gradient-weighted class activation mapping and attention inference network to improve both the classification and segmentation accuracy of the VGG19 DL model. The model was trained on an unbalanced dataset of "suspicious" and "non-suspicious" photographic images captured by mobile phones, which underwent random oversampling for balance and augmentation. Although it achieved ~85% classification accuracy, the study was constrained to a single lesion type with binary classification[4] explored semantic segmentation using the U-Net architecture with various convolutional backbones for lesion segmentation. The EfficientNet-B7 model, in particular, achieved a dice score of 0.926 without TTA and 0.929 with TTA[32] proposed a pixel-based, instance segmentation method using Mask R-CNN for two common types of oral lesions: herpes labialis (cold sore) and aphthous ulcer (canker sore). The model, trained on 40 color-augmented visible light images sourced from Google Images and captured by smartphones, reached a pixel-wise accuracy of ~74%. However, the small dataset size may not adequately represent all lesion variations, potentially affecting the accuracy of localization.

Segmentation is a critical component of lesion diagnosis, particularly for distinguishing the lesion boundary, which aids clinicians in accurately assessing lesion extent and planning further interventions. Unlike simple classification models, the segmentation capability of CLASEG provides a precise visual representation of the affected area, making it an indispensable tool for early diagnosis. The segmentation performance of our model, with an AP50 of 72.18, shows substantial improvement over existing studies, where AP50 scores were reported between 32.69 and 50 for oral lesion detection[4,32]. This improvement demonstrates the robustness of our approach in real-world clinical conditions, where variations in lesion shape, color, and texture are common challenges. Future work will focus on further refining segmentation accuracy and expanding its application to multi-label segmentation for complex cases with mixed lesion types.

Despite the promising outcomes of these studies, they are somewhat limited in their ability to detect and diagnose a broader spectrum of lesions observable in the oral cavity. Many focus on a single or, at most, two lesion types and do not prioritize the early detection of lesions or cancerous areas. To overcome these limitations, we introduced a new model, CLASEG, trained on a relatively larger dataset to identify up to 14 different types of benign, premalignant, and malignant oral cavity lesions. Unlike previous studies, CLASEG relies exclusively on photographic images without the aid of additional features, such as those extracted from medical data or other image types. Unfortunately, we could not compare our results with other state-of-the-art techniques due to the absence of a common/public dataset. While the dataset used in our study was adequate for validation and testing, expanding the dataset size would likely enhance our model's performance.

Conclusion

Oral cancer represents a significant global health challenge, with many regions lacking specialists to detect early symptoms and facilitate timely treatment. In this research, we introduced an effective multi-classification-based segmentation technique, CLASEG, that integrates classifications with segmentation. We fine-tuned up to 14 different pre-trained DL models to classify each input image.

On the basis of the predicted class, a specific detector matched to that class was then used to segment the oral lesion area. Our results show considerable improvement over the baseline model, with the segmentation AP50 score increasing from 32.69 to 72.18. These promising outcomes suggest that further enhancements could be achieved by expanding the dataset and incorporating more diverse images varying in contrast, lighting, and size. Clinically, CLASEG offers a non-invasive and cost-effective diagnostic tool for healthcare providers, potentially early oral cancer detection. This tool could significantly alleviate human suffering associated with oral cancer, improving healthcare access, optimize health outcomes, enhance communication within the healthcare teams, and reduce financial and administrative burdens for patients and providers alike. Despite these advantages, applying AI in oncology comes with challenges, including patient privacy and confidentiality. Additionally, the question of responsibility in case of diagnostic errors whether it falls on the doctor or the AI-remains unresolved. Moreover, the introduction of AI could potentially impact the clinician-patient relationship negatively[45].

Future research directions

Future research directions include enhancing CLASEG's segmentation accuracy by exploring new backbones for the Mask R-CNN architecture, enlarging the dataset with additional lesion types, and advancing towards multi-class classification and multi-label segmentation. Furthermore, training segmentation models on subsets of lesion types could offer a more tailored and effective diagnostic tool.

In addition to enhancing segmentation accuracy and expanding the dataset, future research will focus on addressing real-world challenges such as mixed lesion types and low-quality images. Specifically, we aim to develop multi-label classification and segmentation models to detect and delineate multiple lesion types within a single image. Furthermore, we plan to integrate image augmentation techniques and preprocessing algorithms to handle low-quality images captured under diverse clinical conditions, ensuring improved model robustness and generalizability.

Data availability

Data is available from the corresponding author upon reasonable request.

Received: 17 December 2024; Accepted: 19 May 2025
Published online: 02 July 2025

References

1. Sankaranarayanan R, R. K. A. H. S. S. J. N. Oral Cancer: Prevention, Early Detection, and Treatment. in *Cancer: Disease Control Priorities* vol. 3 (Washington (DC), 2015).
2. Liao, L.-J. et al. Health-related quality of life and utility in head and neck cancer survivors. *BMC Cancer* **19**, 425 (2019).
3. Kar, A. et al. Improvement of oral cancer screening quality and reach: The promise of artificial intelligence. *J. Oral Pathol. Med.* **49**, 727–730 (2020).
4. Tanriver, G., Soluk Tekkesin, M. & Ergen, O. Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders. *Cancers Basel* **13**, 2766 (2021).
5. Welikala, R. A. et al. Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. *IEEE Access* **8**, 132677–132693 (2020).
6. Almubarak, H. The potential role of telemedicine in early detection of oral cancer. *J. Pharm. Bioallied. Sci.* **14**, S19–S23 (2022).
7. Flores, A. P. et al. Teledentistry in the diagnosis of oral lesions: A systematic review of the literature. *J. Am. Med. Inf. Assoc.* **27**, 1166–1172 (2020).
8. Jiang, H., Zhang, P., Che, C., Jin, B. & Zhu, Y. CariesFG: A fine-grained RGB image classification framework with attention mechanism for dental caries. *Eng. Appl. Artif. Intell.* **123**, 106306 (2023).
9. Turki, T. & Taguchi, Y. GENEvaRX: A novel AI-driven method and web tool can identify critical genes and effective drugs for lichen planus. *Eng. Appl. Artif. Intell.* https://doi.org/10.1101/2023.02.23.529678 (2023).
10. Islam, M. Z., Naqvi, R. A., Haider, A. & Kim, H. S. Deep learning for automatic tumor lesions delineation and prognostic assessment in multi-modality PET/CT: A prospective survey. *Eng. Appl. Artif. Intell.* **123**, 106276 (2023).
11. Mathew, T., Johnpaul, C. I., Ajith, B., Kini, J. R. & Rajan, J. A deep learning based classifier framework for automated nuclear atypia scoring of breast carcinoma. *Eng. Appl. Artif. Intell.* **120**, 105949 (2023).
12. Omeroglu, A. N., Mohammed, H. M. A., Oral, E. A. & Aydin, S. A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. *Eng. Appl. Artif. Intell.* **120**, 105897 (2023).
13. Abdi, M. J. & Giveki, D. Automatic detection of erythemato-squamous diseases using PSO–SVM based on association rules. *Eng Appl Artif Intell* **26**, 603–608 (2013).
14. Dayananda, C., Yamanakkanavar, N., Nguyen, T. & Lee, B. AMCC-Net: An asymmetric multi-cross convolution for skin lesion segmentation on dermoscopic images. *Eng. Appl. Artif. Intell.* **122**, 106154 (2023).
15. Qayyum, A., Mazher, M., Khan, T. & Razzak, I. Semi-supervised 3D-InceptionNet for segmentation and survival prediction of head and neck primary cancers. *Eng. Appl. Artif. Intell.* **117**, 105590 (2023).
16. Sunny, S. P. et al. Oral epithelial cell segmentation from fluorescent multichannel cytology images using deep learning. *Comput. Methods Programs Biomed.* **227**, 107205 (2022).
17. Rönnau, M. M., Lepper, T. W., Amaral, L. N., Rados, P. V. & Oliveira, M. M. A CNN-based approach for joint segmentation and quantification of nuclei and NORs in AgNOR-stained images. *Comput. Methods Programs Biomed.* **242**, 107788 (2023).
18. Jubair, F. et al. A novel lightweight deep convolutional neural network for early detection of oral cancer. *Oral Dis* **28**, 1123–1130 (2022).
19. Welikala, R. A. *et al.* Clinically Guided Trainable Soft Attention for Early Detection of Oral Cancer. in 226–236 (2021). https://doi.org/10.1007/978-3-030-89128-2_22.
20. Fu, Q. et al. A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. *EClinicalMedicine* **27**, 100558 (2020).
21. Shamim, M. Z. M. et al. Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. *Comput. J.* **65**, 91–104 (2022).
22. Jeyaraj, P. R. & Samuel Nadar, E. R. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *J. Cancer Res. Clin. Oncol.* **145**, 829–837 (2019).
23. Xu, S. et al. An early diagnosis of oral cancer based on three-dimensional convolutional neural networks. *IEEE Access* **7**, 158603–158611 (2019).
24. Das, N., Hussain, E. & Mahanta, L. B. Automated classification of cells into multiple classes in epithelial tissue of oral squamous cell carcinoma using transfer learning and convolutional neural network. *Neural Netw.* **128**, 47–60 (2020).
25. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. in *2014 IEEE Conference on Computer Vision and Pattern Recognition* 580–587 (IEEE, 2014). https://doi.org/10.1109/CVPR.2014.81.
26. Girshick, R. Fast R-CNN. in *2015 IEEE International Conference on Computer Vision (ICCV)* 1440–1448 (IEEE, 2015). https://doi.org/10.1109/ICCV.2015.169.
27. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017).
28. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. in *2017 IEEE International Conference on Computer Vision (ICCV)* 2980–2988 (IEEE, 2017). https://doi.org/10.1109/ICCV.2017.322.
29. Shin, Y., Qadir, H. A., Aabakken, L., Bergsland, J. & Balasingham, I. Automatic colon polyp detection using region based deep CNN and post learning approaches. *IEEE Access* **6**, 40950–40962 (2018).

30. Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Sci. Rep.* **8**, 4165 (2018).
31. Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S. & Jantana, P. Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. *J. Oral Pathol. Med.* **50**, 911–918 (2021).
32. Anantharaman, R., Velazquez, M. & Lee, Y. Utilizing Mask R-CNN for Detection and Segmentation of Oral Diseases. in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2197–2204 (IEEE, 2018). https://doi.org/10.1109/BIBM.2018.8621112.
33. Song, B. et al. Exploring uncertainty measures in convolutional neural network for semantic segmentation of oral cancer images. *J. Biomed. Opt.* https://doi.org/10.1117/1.JBO.27.11.115001 (2022).
34. Tobias, M. A. S. et al. Artificial intelligence for oral cancer diagnosis: What are the possibilities?. *Oral Oncol.* **134**, 106117 (2022).
35. Gonzalez, R. C. and R. E. *Digital Image Processing Second Edition*. (Pearson, 2009).
36. Reinhold, J. C., Dewey, B. E., Carass, A. & Prince, J. L. Evaluating the impact of intensity normalization on MR image synthesis. in *Medical Imaging 2019: Image Processing* (eds. Angelini, E. D. & Landman, B. A.) 126 (SPIE, 2019). https://doi.org/10.1117/12.2513089.
37. Murphy, K. *Machine Learning: A Probabilistic Perspective*. (MIT Press, 2012).
38. Wu, Y., Li, J., Kong, Y. & Fu, Y. Deep Convolutional Neural Network with Independent Softmax for Large Scale Face Recognition. in *Proceedings of the 24th ACM international conference on Multimedia* 1063–1067 (ACM, New York, NY, USA, 2016). https://doi.org/10.1145/2964284.2984060.
39. Weng, L. Object Detection for Dummies Part 3: R-CNN Family. (2017).
40. Grandini M, B. E. and V. G. Metrics for multi-class classification: an overview. (2020).
41. Padilla, R., Passos, W. L., Dias, T. L. B., Netto, S. L. & da Silva, E. A. B. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics* **10**(3), 279. https://doi.org/10.3390/electronics10030279 (2021).
42. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision 618–626 (2017).
43. Jain, M., Tanu & Rai, C. S. Early Detection of Oral Ulcers Using Photographic Evidence: A Novel Approach using Ensemble of Convolution Neural Network. in *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* 1–5 (IEEE, 2022). https://doi.org/10.1109/ACCAI53970.2022.9752570.
44. Figueroa, K. C. et al. Interpretable deep learning approach for oral cancer classification using guided attention inference network. *J. Biomed. Opt.* https://doi.org/10.1117/1.JBO.27.1.015001 (2022).
45. Al-Rawi, N. et al. The effectiveness of artificial intelligence in detection of oral cancer. *Int. Dent. J.* **72**, 436–447 (2022).

Acknowledgements
This publication was supported by Qatar University Internal Grant No. IRCC-2021-009. The findings achieved herein are solely the responsibility of the authors.

Author contributions
Conceptualization, F.T.,S.M., A.AA., A.H., K.S.; methodology, A.H.; software, A.AA., M.E., K, I.; validation, M.E., K.I.; formal analysis, A.AA., A.H., M.E. K.I, K.S.; investigation, A.AA.; resources, A.AA., A.H, S.M.,P.C., F.T., R.BH. ;data collection, P.C.; data curation, R.BH., F.T., P.C., A.D. and S.M.; writing—original draft preparation, A.AA.; writing—review and editing, M,E; A.H., K.S., R.BH., F.T.; visualization, A.AA.,A.H.; supervision, K.S.; project administration, R.BH., K.S., F.T. All authors have read and agreed to the published version of the manuscript.

Funding
This publication was supported by Qatar University Internal Grant No. IRCC-2021–009. The findings achieved herein are solely the responsibility of the authors.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval
This study received ethical approval from the Institutional Review Board of McGill University, Montreal, Canada (protocol code A07-M40-21B, 25 August 2021). All methods were carried out in accordance with relevant institutional guidelines and regulations.

Patient consent
A written informed consent was obtained from all participants before taking any pictures.

Additional information

Correspondence and requests for materials should be addressed to R.B.-H.