Journal of Current Research in Oral Surgery
2021, Volume 1, Page No: 98-109

Available online at: www.tsdp.net

ISSN: 3062-3480

Original Article

YOLOv11-MobileNetV2 Two-Stage AI Framework for Lesion Localization and Differentiation of Oral Cancer and Precancerous Lesions

Thomas K. Nguyen[1]*, Sarah J. Bennett[2], William J. Carter[3]

[1]Department of Oral and Maxillofacial Surgery, School of Dentistry, University of Melbourne, Melbourne, Australia.
[2]Oral Surgery Unit, Royal Dental Hospital of Melbourne, Melbourne, Australia.
[3]Department of Oral and Maxillofacial Surgery, Monash Health, Melbourne, Australia.

***E-mail** ✉ thomas.nguyen@outlook.com

Received: 04 August 2021; **Revised:** 24 October 2021; **Accepted:** 26 October 2021

ABSTRACT

This work sought to construct and assess an artificial intelligence workflow that merges object-detection and image-classification models to support early recognition and distinction of oral lesions. A retrospective cross-sectional design was applied, using clinical photographs of oral potentially malignant disorders and oral squamous cell carcinoma. The primary dataset consisted of 773 images from the Faculdade de Odontologia de Piracicaba, Universidade Estadual de Campinas (FOP-UNICAMP), and an independent validation set included 132 images from the Federal University of Paraíba (UFPB). All images were captured before biopsy, each paired with histopathological confirmation. For lesion localization, ten YOLOv11 variants employing different augmentation schemes were trained for 200 epochs with pretrained COCO weights. For classification, three MobileNetV2 networks were trained on crops generated according to expert bounding boxes, each adopting distinct learning rate and augmentation configurations. After identifying the top-performing detection–classification pair, both components were linked in a two-stage pipeline in which the detector-generated crops were forwarded into the classifier. The optimal YOLOv11 model achieved an mAP50 of 0.820, precision of 0.897, recall of 0.744, and an F1-score of 0.813. The strongest MobileNetV2 model reached an accuracy of 0.846, precision of 0.871, recall of 0.846, F1-score of 0.844, and an AUC-ROC of 0.852. On the external set, the same classifier obtained an accuracy of 0.850, precision of 0.866, recall of 0.850, an F1-score of 0.851, and an AUC-ROC of 0.935. The integrated two-step framework, tested on the baseline dataset, achieved an accuracy of 0.784, precision of 0.793, recall of 0.784, F1-score of 0.784, and an AUC-ROC of 0.811. When applied to the independent dataset, it produced an accuracy of 0.863, a precision of 0.879, a recall of 0.863, F1-score of 0.866, and an AUC-ROC of 0.934. Visual review of the YOLO outputs showed consistent lesion localization across varied oral images, though 17.4% were not detected. The t-SNE map revealed partial clustering of OPMD and OSCC embeddings, suggesting the model captured relevant discriminative signals despite some overlap. This proof-of-concept investigation indicates that a coupled detection–classification AI framework can feasibly support early screening of oral lesions. Nonetheless, caution is necessary when interpreting two-stage results, since images not detected by YOLO do not advance to classification, potentially influencing the final metrics.

Keywords: Oral cancer, Precancerous lesions, Two-Stage AI, YOLOv11-MobileNetV2

How to Cite This Article: Nguyen TK, Bennett SJ, Carter WJ. YOLOv11-MobileNetV2 Two-Stage AI Framework for Lesion Localization and Differentiation of Oral Cancer and Precancerous Lesions. J Curr Res Oral Surg. 2021;1:98-109. https://doi.org/10.51847/kMIYO81bAR

Introduction

Timely recognition of oral malignancies and their precursor lesions is crucial for lowering the frequency of late-stage diagnoses, thereby improving treatment success, prognosis, and slowing malignant progression [1]. Despite this, early-stage OPMDs are often overlooked because they commonly appear as asymptomatic, flat lesions that do not alert patients.

These early-stage OPMDs may resemble subtle or "incipient" oral squamous cell carcinomas, which manifest as plaques in nearly 80% of cases [2]. Limited public and practitioner awareness of early signs contributes to delayed consultation, and restricted access to specialists in remote areas further reduces timely evaluation.

At present, conventional clinical oral examination remains the most widely applied method for early detection [3], yet it strongly depends on the examiner's judgment. Tarakji *et al.* [4] note that OPMD diagnosis requires competent clinical evaluation and histopathology, with specialists typically outperforming general practitioners. Addressing this gap requires ongoing training and supportive diagnostic tools. Many adjunctive tests have been proposed to enhance early recognition and risk assessment. Among these, AI-based approaches using clinical images provide a promising means of streamlining diagnostic processes by using computer-vision tools to distinguish OPMD from malignant lesions [5, 6], speed referrals [7], support biopsy decisions, and mitigate limitations of traditional examinations that rely on clinical indicators like erythema and ulceration [8, 9], which can also be present in OPMDs. Consequently, CNN-based systems trained on white-light imagery commonly explore two essential tasks: lesion detection and image categorization [10]. Applying both elements is key to building reliable diagnostic systems.

Object detection is a computer-vision technique that identifies where objects appear in an image—typically through bounding boxes—and assigns each object to a class [11]. These methods are valuable in clinical contexts because they highlight suspicious locations within oral images, reducing the chance of missing subtle abnormalities. They also allow integration into workflows that first locate a lesion and then classify it. Prominent models used for oral-disease tasks include YOLO, Faster R-CNN, RetinaNet, and CenterNet2. They are particularly useful for recognizing smooth leukoplakias that might be missed in routine inspection. Still, several limitations persist, including scarce datasets, the absence of external testing, and difficulty identifying very small lesions [12-16].

Classification models, on the other hand, generate diagnostic outcomes for whole images or for selected regions without the need to outline the full margins of a lesion. By extracting visual cues from clinical photographs, these systems are capable of differentiating harmless findings from potentially malignant or overtly malignant conditions, helping prioritize referrals and supporting decision-making in settings with limited clinical expertise. Within the domain of oral pathology, CNN-based classification strategies have gained momentum for automating diagnostic triage and flagging cases that warrant specialist review [5, 17–23]. Nevertheless, these approaches still encounter obstacles related to the diversity of available datasets, the clarity of model outputs, and their practical deployment in routine care.

The purpose of this work is to design both detection and classification models that assist in recognizing OPMD and OSCC at early stages, and to merge these components into a two-stage framework. The detection module narrows the analysis to the relevant areas of the image, while the classification module interprets the localized regions from a diagnostic standpoint. This ordered workflow not only improves the reliability of the final prediction but also mirrors the clinical sequence in which practitioners first identify a lesion and then assess its malignant potential. As a result, both interpretability and accuracy of the proposed solution are enhanced.

Materials and Methods

Dataset

This retrospective cross-sectional project relied on a real-world collection of 773 clinical photographs obtained from individuals presenting with oral lesions at the Faculdade de Odontologia de Piracicaba, Universidade Estadual de Campinas (Piracicaba, São Paulo, Brazil) between 2000 and 2025. Images were separated into 380 OPMD and 393 OSCC cases. For external assessment, a supplementary group from the Federal University of Paraíba (UFPB) in João Pessoa, Paraíba, Brazil, contributed 53 OPMD and 79 OSCC images. Classification of both categories followed the criteria defined by the World Health Organization (WHO Classification of Tumours Editorial Board, 2022). The OPMD set included conventional leukoplakia—with or without oral epithelial dysplasia—and proliferative verrucous leukoplakia, while the OSCC group incorporated multiple clinical and histological variants (conventional, verrucous, and incipient) to broaden the range of presentations.

To maintain coherent diagnostic labeling and image quality, several exclusion rules were applied. Photographs with inadequate resolution were removed, as were samples linked to non-representative biopsies, defined as: (1) specimens diagnosed as OED despite clinical signs suggestive of OSCC, or (2) biopsies too small or technically compromised to establish a reliable diagnosis. If a patient underwent more than one biopsy due to notable clinical evolution, only images taken before each procedure were used, with at least a

three-month interval required between biopsies. All photographs were captured before any biopsy and were paired with histopathological confirmation.

The dataset was divided non-randomly into training, validation, and test sets. Images from the same patient were kept together in the training subset to eliminate leakage, and the proportional distribution of the two major classes was preserved in each partition **(Table 1).**

Table 1. Datasets.

Class	Training (80%)	Validation (10%)	Test (10%)	Total Baseline Dataset (FOP-UNICAMP)	External Validation Dataset (UFPB)	Overall Total
OPMD (Oral Potentially Malignant Disorders)	312	34	34	380	53	433
OSCC (Oral Squamous Cell Carcinoma)	316	39	38	393	79	472
Total	**628**	**73**	**72**	**773**	**132**	**905**

OPMD, oral potentially malignant disorder; OSCC, oral squamous cell carcinoma.

Bounding-box annotation was carried out by A.L.D.A., with C.S.S. consulted to achieve agreement. Labeling was performed using Aperio ImageScope (Leica Biosystems) with a Huion Inspiroy H1060P tablet, and annotators were kept blinded to the diagnostic category. Lesions were enclosed within rectangular regions.

The project followed the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [24] and the MAIC-10 criteria [25]. Ethical approval was provided by the Piracicaba Dental School Ethical Committee (Registration 42235421.9.0000.5418) and the Federal University of Paraíba Ethical Committee (Registration 72314323.0.0000.5188). Material Transfer Agreements were included to enable the exchange of image data between institutions.

Workstation

All processing took place in Google Colab within a uniform virtual environment. The system utilized an Intel(R) Xeon(R) CPU running at 2.00 GHz (2 threads, 1 physical core) with 39 MB L3 cache, and an NVIDIA Tesla T4 GPU providing 15,360 MiB of VRAM (CUDA 12.4, Driver 550.54.15).

Object detection task

YOLO (You Only Look Once) [26] is a high-speed, convolutional-network–driven framework for detecting objects within images. Instead of relying on multi-stage region proposals or sequential processing, YOLO reframes detection as a unified regression task in which the entire image is analyzed simultaneously. The image is partitioned into a grid, and each grid segment predicts bounding boxes, confidence values, and class likelihoods. This design allows the method to operate at real-time speeds while maintaining competitive accuracy. Its end-to-end training scheme also simplifies optimization, making it widely adopted in fields ranging from surveillance and autonomous navigation to medical imaging.

YOLOv11 [27], the newest member of the YOLO family, incorporates several architectural upgrades intended to boost performance in diverse computer-vision scenarios. Among the most notable additions are the C3k2 block (a Cross-Stage-Partial variant using kernel size 2), the SPPF (Spatial Pyramid Pooling—Fast), and the C2PSA module (a parallel spatial-attention convolutional block). Collectively, these components enhance representational strength and computational efficiency.

For this study, ten detector models were constructed across four YOLOv11 variants: YOLOv11n (2.6M parameters), YOLOv11s (9.4M), YOLOv11m (20.1M), and YOLOv11l (25.3M). Each version was paired with specific augmentation schemes to examine performance differences. All detectors were initialized using pretrained COCO weights [28] and trained for 200 epochs with images standardized to 640 × 640 pixels. The augmentation operations included hue shifts (±0.015), saturation changes (±0.7), translation (±10% of image dimensions), and scale adjustments (±50%). Horizontal flips were applied at a 0.5 probability. Additional variations incorporated mosaic augmentation, wherein one to four images are randomly fused during training.

Evaluation of the detection models used mean Average Precision at 50% IoU (mAP50) as the main performance metric, with emphasis on accurate lesion localization. A single class ("lesion"—combining OPMD and OSCC) was used, as preliminary experiments indicated that while the model localized

regions well, discrimination between the two lesion types was very poor; when treated as separate classes, mAP50 fell to roughly 22%. Because the task effectively becomes single-class detection, AUROC was not computed. Precision, Recall, and F1-Score were also obtained using the Ultralytics built-in evaluation tools.

While mAP50 was used for consistency with standard detection literature, it should be recognized that for a single-class task, mAP50 is identical to AP at IoU = 0.5. Since AP corresponds to the area under the Precision–Recall curve (AUC-PR), AP and AUC-PR are mathematically equivalent. We therefore report AP only, following established norms in the field.

Classification task

Three MobileNetV2-based models [29, 30] were trained, each using different learning rates and augmentation configurations. Rather than full images, the classifiers received crops derived from the expert-annotated bounding boxes. Learning rate values were selected through structured hyperparameter searches, guided by loss-curve behavior observed in pilot trials. All classifiers were initialized with ImageNet pretrained weights [31, 32] and trained for 200 epochs, with image inputs resized to 224 × 224 pixels. Augmentations included adjustments to brightness, contrast, and saturation (±0.2), hue changes (±0.1), and translations of up to ±10% of the image dimension. Random flips—both horizontal and vertical—were applied in multiple variants. Because the two classes were perfectly balanced, no class-imbalance corrections were required.

Accuracy, precision, recall, F1-score, and AUC-ROC were computed using the scikit-learn toolkit [33]. After determining the optimal pairing of detector and classifier, the models were combined into a sequential framework in which the detector's cropped outputs were forwarded directly to the classification stage.

Results and Discussion

Object detection

Across the lesion-detection experiments, the YOLOv11 configurations displayed a range of performance, with mAP50 values spanning 0.718–0.820 depending on the augmentation scheme. The most effective setup incorporated Albumentations with mild blur, grayscale conversion, CLAHE [34], and restrained geometric modifications—specifically an 80° rotation and a very small 0.001 perspective adjustment**.** This combination reached the top mAP50 **(**0.820), precision (0.897**)**, and F1-score (0.813). Although its recall (0.744) was marginally below that of a few alternative variants, the notable improvement in precision and the resulting balanced F1-metric suggest an advantageous compromise **(Table 2)**. Qualitative review of the detections demonstrated stable lesion localization across various oral images, shown through bounding boxes and confidence estimates. Overall, tailored augmentation—particularly rotation and subtle perspective changes—enhanced detection outcomes on the baseline-derived test set **(Figure 1).**

Table 2. summarizes YOLOv11 one-class detection metrics using the baseline dataset. Definitions:

Model	Albumentations	Close_mosaic	Degrees	Perspective	Flipud	mAP50	Precision	Recall	F1-Score
YOLOv11n	True	0	0.0	0.0	0.0	0.767	0.811	0.718	0.761
YOLOv11n	True	0	80.0	0.0	0.5	0.784	0.841	0.748	0.792
YOLOv11n	True	0	0.0	0.0	0.5	0.743	0.763	0.782	0.772
YOLOv11n	True	50	0.0	0.0	0.5	0.776	0.831	0.758	0.792
YOLOv11n	True	100	0.0	0.0	0.5	0.765	0.775	0.718	0.745
YOLOv11s	True	0	80.0	0.0	0.5	0.766	0.724	0.740	0.732
YOLOv11m	True	0	80.0	0.0	0.5	0.752	0.720	0.692	0.705
YOLOv11l	True	0	80.0	0.0	0.5	0.718	0.713	0.679	0.695
YOLOv11n	False	0	80.0	0.0	0.5	0.788	0.766	0.754	0.760
YOLOv11n	True	0	80.0	0.001	0.5	0.820	0.897	0.744	0.813

Albumentations indicates the augmentation library; *Close_mosaic* describes a multi-image composite technique; *Degrees* corresponds to the rotation range; *flipud* denotes vertical flips; *mAP* is mean average precision; *Perspective* refers to the degree of geometric distortion. Bolded entries mark the best scores.

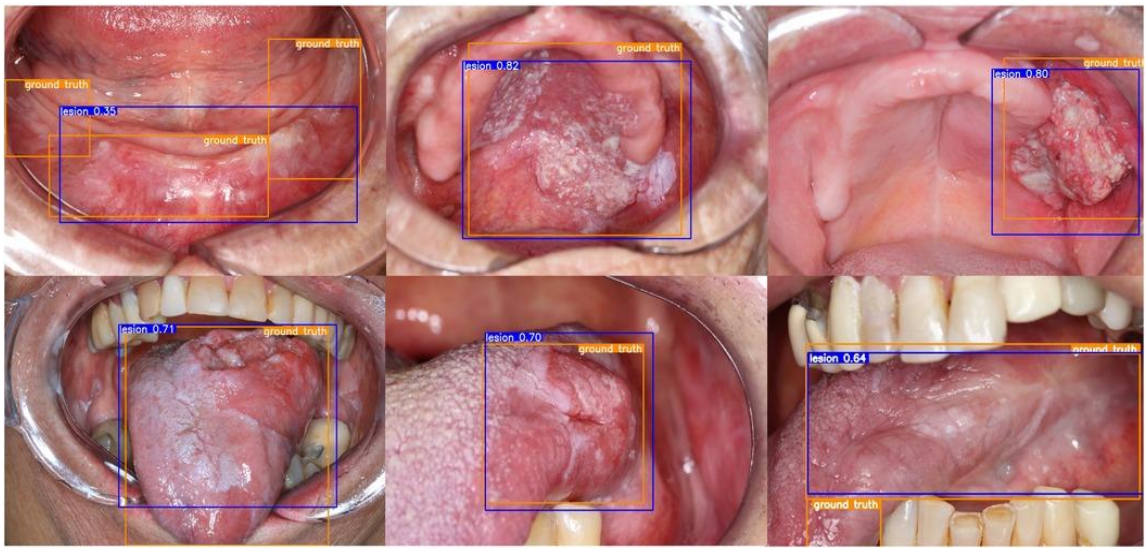


Figure 1. presents YOLO-based lesion detection examples, with orange boxes indicating ground truth and blue boxes showing predictions with associated confidence values.

Classification

For lesion-classification tasks, MobileNetV2 models benefited most from applying both horizontal and vertical flipping. The configuration trained with a learning rate of 0.0001 alongside these dual flips produced the strongest results: accuracy 0.846 (95% CI: 0.756–0.923), precision 0.871 (95% CI: 0.817–0.933)**,** recall 0.846 (95% CI: 0.756–0.923)**,** F1-score 0.844 (95% CI: 0.756–0.923)**,** and an AUC-ROC of 0.852 (95% CI: 0.759–0.941)**,** which was the second-highest. Although not the top AUC-ROC, the gap was minimal **(**0.004**)**, indicating negligible loss in discriminatory capability. Excluding vertical flips led to uniformly lower results, showing that combining both orientations improved generalization on the baseline test set. Lowering the learning rate to **0.00001** produced slightly reduced scores, implying that a moderately small learning rate, together with richer augmentation is preferable **(Table 3)**.

After choosing this optimal MobileNetV2 setup (Model 2); **(Table 3)**, external validation yielded accuracy 0.850 (0.798–0.902)**,** precision 0.866 (0.822–0.912), recall 0.850 (0.798–0.902), F1-score 0.851 (0.799–0.902)**,** and AUC-ROC 0.935 (0.900–0.968)**.**

Table 3. reports MobileNetV2 performance on expert-guided crops using the baseline dataset, with values shown as mean (95% CI) derived via 1,000 bootstrap runs. LR = learning rate. Best metrics are bold.

Model	Learning Rate	Horizontal Flip	Vertical Flip	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)	AUC-ROC (95% CI)
1	0.0001	True	False	0.743 (0.654 – 0.846)	0.750 (0.656 – 0.852)	0.743 (0.654 – 0.846)	0.742 (0.647 – 0.846)	0.823 (0.725 – 0.913)
2	0.0001	True	True	0.846 (0.756 – 0.923)	0.871 (0.817 – 0.933)	0.846 (0.756 – 0.923)	0.844 (0.756 – 0.923)	0.852 (0.759 – 0.941)
3	0.00001	True	True	0.820 (0.731 – 0.897)	0.844 (0.772 – 0.909)	0.820 (0.731 – 0.897)	0.818 (0.727 – 0.897)	0.856 (0.755 – 0.928)

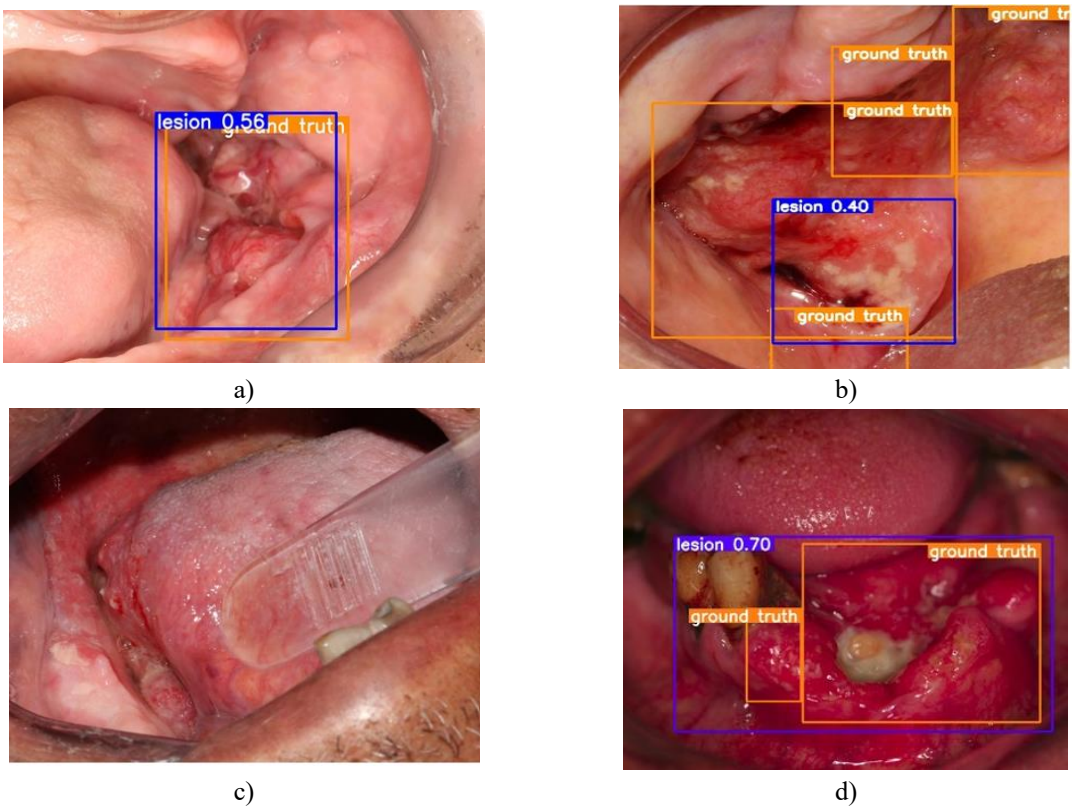
For the integrated two-stage method, we adopted the best MobileNetV2 classifier (Model 2), **(Table 3)** and applied it to cropped images obtained from the highest-performing YOLOv11n model **(Table 2)**, following an approach comparable to that described by Fu *et al*. [35]. Performance of this combined pipeline was also computed for the external validation set **(Table 4).**

Table 4. presents the resulting metrics, expressed as mean (95% CI) using **1,000 bootstrap** samples. Bold values highlight the strongest results.

Dataset	Accuracy (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)	AUC-ROC (95% CI)
Baseline dataset (FOP-UNICAMP)	0.784 (0.696 – 0.863)	0.793 (0.712 – 0.872)	0.784 (0.696 – 0.863)	0.784 (0.696 – 0.863)	0.811 (0.712 – 0.889)
External validation dataset (UFPB)	0.863 (0.806 – 0.914)	0.879 (0.831 – 0.928)	0.863 (0.806 – 0.914)	0.866 (0.807 – 0.916)	0.934 (0.883 – 0.975)

[a]Test set corresponds to the baseline dataset (FOP-UNICAMP).

Visual examination plays a crucial role because relying only on YOLO's numerical metrics may not fully capture how well the detector behaves in practice. Reviewing the model's actual predictions offers an additional layer of understanding that improves interpretability **(Figure 2).**



a) b) c) d)

Figure 2. Qualitative assessment of YOLO's output. (a) Illustration of a highly accurate prediction, where the model's bounding box almost perfectly aligns with the reference annotation (ground truth shown in orange). (b) In certain instances, experts annotated lesions with multiple bounding boxes to represent their full extent. YOLO, however, tends to generate a single box, which may not include all annotated regions. This indicates that the detector may struggle with large or irregular lesions that cannot be captured using only one bounding region. As a result, only the area enclosed by the model-generated blue box is passed on to the MobileNetV2 classifier, which in some situations can deprive the classifier of essential contextual information, leading to mislabeling. (c) Example of a case where the model failed to detect any lesion. Within the external validation set of 132 images, 23 (17.4%) produced no detection box, and such images were consequently excluded from the classification step. (d) Illustration of YOLO predicting one extensive bounding box that covers the affected region containing multiple expert annotations. Although the detection is clinically accurate, the mAP metric is penalized because the model outputs fewer boxes than the ground truth, demonstrating how visual correctness may diverge from quantitative scoring.

.

t-SNE plot

We additionally generated a t-SNE projection of the feature embeddings derived from the two-stage workflow (YOLOv11n + MobileNetV2) on the external validation images, mapping them into two dimensions to qualitatively explore the learned representations. As shown in **Figure 3**, the visualization reveals partial differentiation between

OPMD and OSCC images. Distinct clusters dominated by one class emerge, suggesting that the model successfully recognizes meaningful, class-related information. Nonetheless, a considerable overlap region—particularly in the central and lower areas of the plot—shows points from both categories interspersed. This pattern indicates that while MobileNetV2 captures discriminative features, the two classes still share substantial visual similarity in many samples. The overlapping clusters likely represent diagnostically ambiguous lesions rather than any substantial flaw in the model, reflecting the reality that borderline cases often present nearly identical visual patterns. Consequently, classification mistakes tend to occur within this mixed region, highlighting the natural limit imposed by the dataset's complexity. This t-SNE distribution, therefore, helps explain the model's performance constraints: some errors are rooted in the inherent challenge of the problem rather than inadequacies in feature learning.

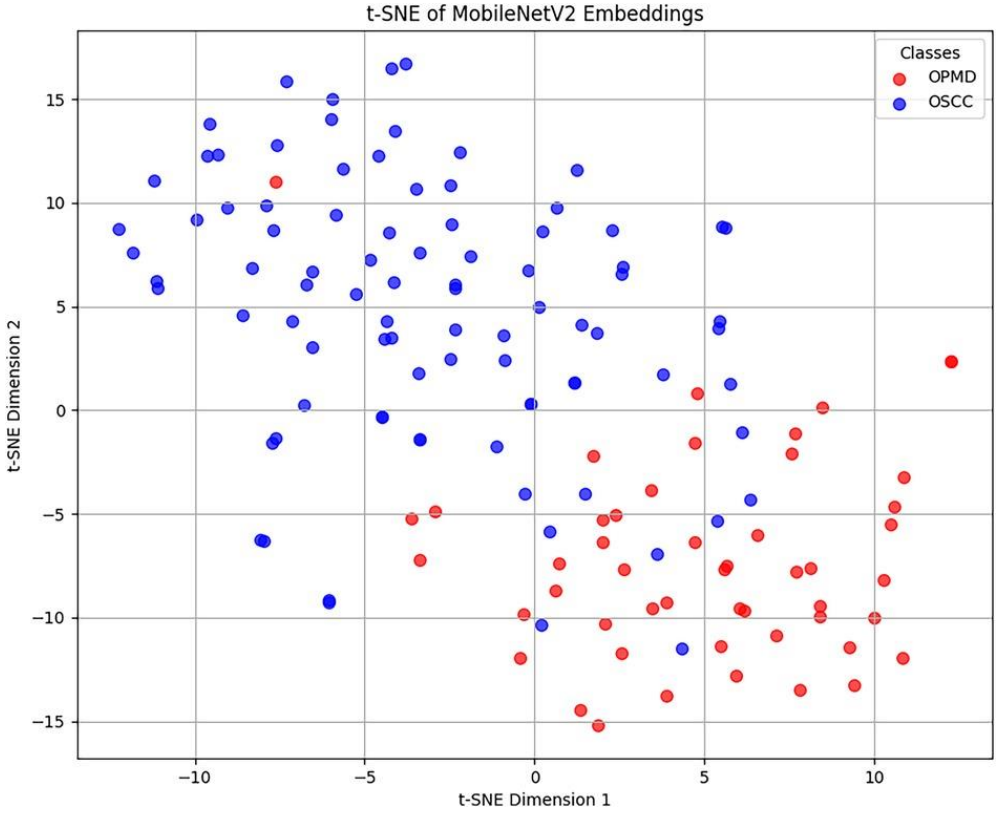


Figure 3. t-SNE mapping of embeddings from the external validation set. Each dot corresponds to an image's feature representation in two dimensions, with nearby points indicating similar learned features. Color coding reflects the true class labels: red indicates oral potentially malignant disorders (OPMD), and blue corresponds to oral squamous cell carcinoma (OSCC). The overlap of blue and red points in the central and lower zones illustrates OSCC samples that share close visual resemblance to OPMDs.

The goal of this investigation was to construct detection and classification models to support oral cancer screening within a mobile application environment using standardized photographic acquisition. The findings align with previous studies showing that initializing models with COCO pre-trained weights can speed up learning and enhance detection outcomes [12–15]. Consistent with the conclusions of Welikala *et al*. [16], our results also indicate that ImageNet-based initialization may offer improved transfer effectiveness for clinical imaging tasks, likely due to broader variability and richer low-level features. Even with the use of widely adopted frameworks such as YOLO and Faster R-CNN, together with accepted strategies like extensive augmentation and consensus labeling, the performance remains influenced by annotation subjectivity and the difficulty of identifying small or subtle lesions, particularly those associated with OPMDs. However, prior reports showing that AI models can perform on par with or even outperform clinicians suggest that such systems hold considerable promise in supporting early OSCC detection—especially when embedded in mobile, user-friendly tools designed to complement everyday clinical practice.

Our two-stage framework parallels the strategy described by Fu *et al*. [35], who employed backbone architectures for both detection and classification,

using bounding boxes to localize lesions in clinical photographs and then passing the cropped areas to the classifier. Direct comparison with their results is not feasible because detection outcomes were not provided in their work. Nevertheless, the performance of our combined pipeline on the external validation dataset—AUC = 0.934 (95% CI: 0.883–0.975)—closely matches the value reported by Fu *et al*. [AUC = 0.935 (95% CI: 0.910–0.957)], indicating that our method reaches a similar discriminative capability despite variations in datasets and preprocessing. The reproducibility of results in external validation strengthens the likelihood that our system can generalize to independent cohorts. Still, it is crucial to acknowledge that YOLO detections may introduce constraints; for example, bounding boxes may only partly capture the lesion, or the lesion may be missed entirely. Such limitations should be considered when interpreting the final performance of the two-step pipeline, as they may influence the classifier's evaluation.

Consistent with most existing publications [12–15], our work also employed COCO-derived weights for initialization in the detection stage, accelerating training and improving performance due to the broad prior knowledge embedded in that dataset. Only one earlier report [16] relied on ImageNet as the pre-training source, and their enhanced recall and F1-score indicate that ImageNet may offer stronger feature transfer for clinical imagery, potentially due to its wider visual diversity and richer low-level descriptors. Even so, pre-training cannot be assumed to guarantee superiority, since models, datasets, and target categories vary considerably across studies, and additional factors—including data augmentation protocols, balancing strategies, and IoU thresholds—can substantially influence outcomes.

It is well recognized that lesion labeling (segmentation or bounding-box definition) carries intrinsic subjectivity from clinical interpretation, which may propagate bias into the models when annotations come from a single reader. Many studies define ground truth using the region of maximal overlap among annotators [13-15], while others rely on aggregated labels from multiple experts [16]. In our methodology, two annotators reached agreement on each lesion to create a unified bounding box (except in cases where lesion size or configuration required more than one), thereby simplifying processing while maintaining annotation consistency.

A comprehensive review of the literature revealed five relevant investigations [12-16] that applied object detection architectures. These included various YOLO versions [12-14], Faster R-CNN [13-16], RetinaNet [13], and CenterNet2 [14]. YOLO emerged as the most common approach, used in all studies except [15, 16]. Its popularity likely reflects the model's direct bounding-box regression mechanism, compact architecture, and straightforward deployment—attributes that are advantageous for real-time or mobile environments [26]. However, an important drawback is the reduced sensitivity of some YOLO variants when dealing with very small targets [36].

Regarding sampling design, our baseline dataset was divided using an 80:10:10 scheme, and a distinct test subset derived from the same dataset was used to evaluate generalization. Additionally, we employed an external validation dataset to strengthen the assessment of model robustness, following guidance from Cerdá-Alberich *et al*. and Tejani *et al*. [24, 25]. The 80:10:10 division aligns with the strategy reported by Tanriver *et al*. [12], although it is not the prevailing approach in the object detection literature. When data volume is limited, five-fold cross-validation remains a strong option [13–15], and supplemental techniques such as bootstrapping, ensembling, label smoothing, stratified or nested cross-validation, and early stopping are recommended to mitigate heterogeneity, labeling inaccuracies, and overfitting, thus promoting more stable and transferable models [37].

For evaluating detection performance, commonly recommended metrics include precision, recall (sensitivity), F1-score, Intersection over Union (IoU), and mean Average Precision (mAP). Among these, mAP provides the broadest assessment because it examines performance across confidence thresholds, across classes, and across varying object complexities [28]. Despite its importance, only the present work and one earlier study [12] reported mAP. IoU is also noteworthy as a stringent measure—quantifying how well predictions overlap with their corresponding annotations—and often penalizes predictions that are directionally correct but not perfectly aligned, particularly in the presence of small, complex, or irregular lesions.

Warin *et al*. [14] evaluated how accurately AI systems detect OSCC compared with oral and maxillofacial surgeons. Interestingly, even their weakest-performing algorithm—CenterNet2—surpassed the surgeons in OSCC detection. In contrast, clinicians demonstrated superior sensitivity for identifying OPMD. This likely reflects the heterogeneous and often subtle nature of OPMDs, whose visual characteristics can resemble other oral conditions and pose challenges even for seasoned specialists. Prior research also indicates that fast object-detection architectures may struggle to

capture the finer details required to reliably detect OPMDs.

One advantage of the present study is that the baseline dataset spans a 25-year interval (2000–2025), introducing natural variation in imaging conditions (e.g., device type, illumination), which may contribute to more robust model performance when annotations are consistent. The dataset also maintains an approximate class balance, reducing the need for extensive reweighting or oversampling [38]. Nevertheless, we employed data augmentation procedures to reinforce the training process, an essential step given the relatively modest dataset size. Importantly, the dataset includes early and advanced OSCC as well as multiple OPMD subcategories, supporting clinical realism. Some subclasses contain few samples, which reflects the actual distribution of lesion types seen in practice.

As with most medical deep learning studies, our work is constrained by the limited number of annotated images, since DL models generally benefit from very large training sets. To mitigate this, we incorporated augmentation strategies and used a well-defined training/validation split to monitor loss progression and apply early stopping when appropriate. A separate test partition was then used to estimate generalization. In addition, an external dataset—captured with different equipment, acquisition conditions, and geographic origins—was used for validation. The strong performance on this independent dataset suggests that overfitting was unlikely. However, a subset of images (17.4%) yielded no detection box and therefore could not proceed to the classification stage. This restriction may introduce bias by decreasing the number of evaluable samples and possibly overlooking specific lesion forms. Another limitation is that augmentation parameters were adopted directly from the YOLO defaults rather than optimized for this dataset; refining these settings through ablation studies is planned for future investigation. Although we did not implement formal explainability tools, the YOLO detection stage naturally provides some interpretability via object localization. Because the MobileNetV2 classifier receives only cropped regions, methods such as Grad-CAM would offer limited additional insight. Nonetheless, explainability remains essential for clinical trust, and we intend to explore approaches like Grad-CAM [39, 40] and SHAP [41] moving forward.

Currently, no technology has conclusively demonstrated superior sensitivity or specificity for oral cancer screening compared with standard intraoral examination [42]. Screening, by definition, involves applying a test to individuals without symptoms to identify disease at an early, more treatable stage [3]. Therefore, AI tools that classify already-noticed lesions do not function as true screening modalities; they instead operate as decision-support systems that assist clinicians in evaluating findings detected during the clinical exam. Evidence is also insufficient to show that such systems modify disease progression or reduce mortality. Even with early diagnosis—whether AI-assisted or not—leukoplakia frequently reappears despite interventions such as surgery or CO_2 laser therapy [43]. Furthermore, whether accurately predicting malignant transformation risk would meaningfully alter outcomes remains uncertain.

Conclusion

This work presents an initial, proof-of-concept two-stage pipeline for detecting oral cancer and distinguishing between OPMD and OSCC. Future research will involve a prospective study in which clinicians use a mobile application to aid diagnosis, providing a real-time, accessible, low-cost, and non-invasive support tool for oral healthcare.

Acknowledgments: None

Conflict of Interest: None

Financial Support: The author(s) declare that financial support was received for the research and/or publication of this article. This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Numbers #2021/14585-7, #2022/13069-8 and #2024/20694-1.

Ethics Statement: This study was performed in accordance with the Declaration of Helsinki and was approved by the Piracicaba Dental School Ethical Committee (Registration number: 42235421.9.0000.5418) and by the Federal University of Paraíba Ethical Committee (Registration number: 72314323.0.0000.5188), which also comprised Material Transfer Agreements between co-participant Institutions to share images.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. (2024) 74:229–63. doi: 10.3322/caac.21834
2. Saldivia-Siracusa C, Araújo AL, Arboleda LP,

Abrantes T, Pinto MB, Mendonça N, et al. Insights into incipient oral squamous cell carcinoma: a comprehensive south- American study. Med Oral Patol Oral Cir Bucal. (2024) 29(4):e575. doi: 10.4317/ medoral.26551

3. Speight PM, Epstein J, Kujan O, Lingen MW, Nagao T, Ranganathan K, et al. Screening for oral cancer—a perspective from the global oral cancer forum. Oral Surg Oral Med Oral Pathol Oral Radiol. (2017) 123:680–7. doi: 10.1016/j.oooo.2016.08.021
4. Tarakji B. Dentists' perception of oral potentially malignant disorders. Int Dent J. (2022) 72:414–9. doi: 10.1016/j.identj.2022.01.004
5. Flügge T, Gaudin R, Sabatakakis A, Tröltzsch D, Heiland M, van Nistelrooij N, et al. Detection of oral squamous cell carcinoma in clinical photographs using a vision transformer. Sci Rep. (2023) 13:2296. doi: 10.1038/s41598-023-29204-9
6. Saldivia-Siracusa C, Carlos de Souza ES, Barros da Silva AV, Damaceno Araújo AL, Pedroso CM, Aparecida da Silva T, et al. Automated classification of oral potentially malignant disorders and oral squamous cell carcinoma using a convolutional neural network framework: a cross-sectional study. Lancet Reg Health Am. (2025) 47:101138. doi: 10.1016/j.lana.2025.101138
7. Lim JH, Tan CS, Chan CS, Welikala RA, Remagnino P, Rajendran S, et al. D'oraca: deep learning-based classification of oral lesions with mouth landmark guidance for early detection of oral cancer. In: Proceedings (2021). p. 408–22
8. Warnakulasuriya S. Oral potentially malignant disorders: a comprehensive review on clinical aspects and management. Oral Oncol. (2020) 102:104550. doi: 10.1016/j.oraloncology.2019.104550
9. Warnakulasuriya S, Kovacevic T, Madden P, Coupland VH, Sperandio M, Odell E, et al. Factors predicting malignant transformation in oral potentially malignant disorders among patients accrued over a 10-year period in South East England. J Oral Pathol Med. (2011) 40:677–83. doi: 10.1111/j.1600-0714.2011.01054.x
10. Araújo ALD, Pedroso CM, Vargas PA, Lopes MA, Santos-Silva AR. Advancing oral cancer diagnosis and risk assessment with artificial intelligence: a review. Explor Digit Heal Technol. (2025) 3:101147. doi: 10.37349/edht.2025.101147
11. Kaur J, Singh W. Tools, techniques, datasets and application areas for object detection in an image: a review. Multimed Tools Appl. (2022) 81:38297–351. doi: 10.1007/s11042-022-13153-y
12. 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). (2021) 13:2766. doi: 10.3390/cancers13112766
13. Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P. Performance of deep convolutional neural network for classification and detection of oral potentially malignant disorders in photographic images. Int J Oral Maxillofac Surg. (2022) 51:699–704. doi: 10.1016/j.ijom.2021.09.001
14. Warin K, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P, Vicharueang S. AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. PLoS One. (2022b) 17:e0273508. doi: 10.1371/journal.pone.0273508
15. 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. (2021) 50:911–8. doi: 10.1111/jop.13227
16. Welikala RA, Remagnino P, Lim JH, Chan CS, Rajendran S, Kallarakkal TG, et al. Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. IEEE Access. (2020) 8:132677–93. doi: 10.1109/ACCESS.2020.3010180
17. Camalan S, Mahmood H, Binol H, Araújo ALD, Santos-Silva AR, Vargas PA, et al. Convolutional neural network-based clinical predictors of oral dysplasia: class activation map analysis of deep learning results. Cancers (Basel). (2021) 13:1–18. doi: 10.3390/cancers13061291
18. Figueroa KC, Song B, Sunny S, Li S, Gurushanth K, Mendonca P, et al. Interpretable deep learning approach for oral cancer classification using guided attention inference network. J Biomed Opt. (2022) 27(1):015001–015001. doi: 10.1117/1.JBO.27.1.015001
19. Jubair F, Al-karadsheh O, Malamos D, Al Mahdi S, Saad Y, Hassona Y. A novel lightweight deep convolutional neural network for early detection

of oral cancer. Oral Dis. (2022) 28:1123–30. doi: 10.1111/odi.13825

20. Shamim MZM, Syed S, Shiblee M, Usman M, Ali SJ, Hussein HS, et al. Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer. Comput J. 65(1):91–104. doi: 10.1093/comjnl/bxaa136
21. Sharma D, Kudva V, Patil V, Kudva A, Bhat RS. A convolutional neural network based deep learning algorithm for identification of oral precancerous and cancerous lesion and differentiation from normal Mucosa: a retrospective study. Eng Sci. (2022) 18:278–87. doi: 10.30919/es8d663
22. Song B, Zhang C, Sunny S, Kc DR, Li S, Gurushanth K, et al. Interpretable and reliable oral cancer classifier with attention mechanism and expert knowledge embedding via attention map. Cancers (Basel). (2023) 15:1421. doi: 10.3390/cancers15051421
23. Song B, Li S, Sunny S, Gurushanth K, Mendonca P, Mukhia N, et al. Classification of imbalanced oral cancer image data from high-risk population. J Biomed Opt. (2021) 26(10):105001. doi: 10.1117/1.JBO.26.10.105001
24. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell. (2024) 6(4):e240300. doi: 10.1148/ryai.240300
25. Cerdá-Alberich L, Solana J, Mallol P, Ribas G, García-Junco M, Alberich-Bayarri A, et al. MAIC–10 brief quality checklist for publications using artificial intelligence and medical images. Insights Imaging. (2023) 14:11. doi: 10.1186/s13244-022-01355-9
26. Redmon J., Divvala S., Girshick R., Farhadi A., 2016. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 779–88.
27. Khanam R, Hussain M. YOLOv11: An Overview of the Key Architectural Enhancements (2024).
28. Lin T.-Y., Maire M., Belongie S., Bourdev L., Girshick R., Hays J., Perona P., Ramanan D., Zitnick C.L., Dollár P., 2015. Microsoft coco: common objects in context. In European Conference on Computer Vision, pp. 740–55. Cham: Springer International Publishing, 2014.
29. Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv [Preprint]. arXiv:1704.04861 (2017). doi: 10. 48550/arXiv.1704.04861
30. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C., 2018. Mobilenetv2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 4510–20.
31. Deng J., Dong W., Socher R., Li L.-J., Li Kai, Fei-Fei Li, 2009. Imagenet: a large- scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–55.
32. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis. (2015) 115:211–52. doi: 10.1007/ s11263-015-0816-y
33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. (2011) 12:2825–30. doi: 10.5555/1953048.2078195
34. Khan SA, Hussain S, Yang S. Contrast enhancement of low-contrast medical images using modified contrast limited adaptive histogram equalization. J Med Imaging Health Inform. (2020) 10:1795–803. doi: 10.1166/jmihi.2020.3196
35. Fu Q, Chen Y, Li Z, Jing Q, Hu C, Liu H, et al. A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: a retrospective study. EClinicalMedicine. (2020) 27:100558. doi: 10.1016/j.eclinm. 2020.100558
36. Yan B, Li J, Yang Z, Zhang X, Hao X. AIE-YOLO: auxiliary information enhanced YOLO for small object detection. Sensors (Basel). (2022) 22(21):8221. doi: 10.3390/s22218221
37. Araújo ALD, Sperandio M, Calabrese G, Faria SS, Cardenas DAC, Martins MD, et al. Artificial intelligence in healthcare applications targeting cancer diagnosis—part II: interpreting the model outputs and spotlighting the performance metrics. Oral Surg Oral Med Oral Pathol Oral Radiol. (2025c) 140(1):89–99. doi: 10.1016/j.oooo. 2025.01.002
38. Araújo ALD, Sperandio M, Calabrese G, Faria SS, Cardenas DAC, Martins MD, et al. Artificial intelligence in healthcare applications targeting cancer diagnosis—part I: data structure, preprocessing and data organization.

Oral Surg Oral Med Oral Pathol Oral Radiol. (2025b) 140(1):79–88. doi: 10.1016/j.oooo.2025.01.004

39. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization (2017). Available online at: http://gradcam.cloudcv.org (Accessed June 4, 2025).
40. Zhang Y, Hong D, McClement D, Oladosu O, Pridham G, Slaney G. Grad- CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. J Neurosci Methods. (2021) 353:109098. doi: 10.1016/j.jneumeth.2021.109098
41. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. arXiv [Preprint]. arXiv:1705.07874 (2017). Available online at: https://arxiv.org/abs/ 1705.07874 (Accessed June 14, 2025).
42. Lingen MW, Kalmar JR, Karrison T, Speight PM. Critical evaluation of diagnostic aids for the detection of oral cancer. Oral Oncol. (2008) 44:10–22. doi: 10.1016/j.oraloncology.2007.06.011
43. van der Waal I. Oral leukoplakia: diagnosis and management revisited. J Dent Indones. (2023) 30:73–80. doi: 10.14693/jdi.v30i2.1507