

Machine learning versus clinicians for detection and classification of oral mucosal lesions

Julia Schwärzler^{a,b}, Ekaterina Tolstaya^{a,b,c}, Antonin Tichy^{a,b,d}, Sebastian Paris^c, Ghazal Aarabi^e, Akhilanand Chaurasia^{b,f}, Yoana Malenova^g, David Steybe^g, Falk Schwendicke^{a,b,*}

^a Clinic for Conservative Dentistry, Periodontology and Digital Dentistry, LMU University Hospital, LMU Munich, Goethestraße 70, 80336 Munich, Germany

^b Topic Group Oral Health, ITU/WHO/WIPO Global Initiative on Artificial Intelligence for Health, CH-1211 Geneva 20, Switzerland

^c Department of Operative, Preventive and Pediatric Dentistry, Charité - Universitätsmedizin Berlin, Assmannshausen Straße 4-6, 14197, Berlin, Germany

^d Institute of Dental Medicine, First Faculty of Medicine, Charles University, Karlovo náměstí 32, 121 11 Prague, Czech Republic

^e Department of Periodontics, Preventive and Restorative Dentistry, Center for Dental and Oral Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

^f Department of Oral Medicine and Radiology, King George's Medical University, Lucknow 226003, India

^g Clinic for Oral and Maxillofacial Surgery, LMU University Hospital, LMU Munich, Lindwurmstr. 2a, 80337 Munich, Germany

ARTICLE INFO

Keywords:

Artificial intelligence
Squamous cell carcinoma
Deep learning
Computer-assisted diagnosis
Early detection of cancer
Mouth neoplasms
Oral potentially malignant disorders (OPMDs)

ABSTRACT

Objectives: The detection and classification of oral mucosal lesions is a challenging task due to high heterogeneity and overlap in clinical appearance. Nevertheless, differentiating benign from potentially malignant lesions is essential for appropriate management. This study evaluated whether a deep learning model trained to discriminate 11 classes of oral mucosal lesions could exceed the performance of general dentists.

Methods: 4079 intraoral photographs of benign, potentially malignant and malignant oral lesions were labeled using bounding boxes and classified into 11 classes. The data were split 80:20 for training ($n = 3031$) and validation ($n = 766$), keeping an independent test set ($n = 282$). The YOLOv8 computer vision model was implemented for image classification and object detection. Model performance was evaluated on the test set which was also assessed by six general dentists and three specialists in oral surgery. Evaluation metrics included sensitivity, specificity, F1-score, precision, area under the receiver operating characteristic curve (AUROC), and average precision (AP) at multiple thresholds of intersection over union.

Results: In terms of classification, the highest F1-score (0.80) and AUROC (0.96) were observed for human papillomavirus (HPV)-related lesions, whereas the lowest F1-score (0.43) and AUROC (0.78) were obtained for keratosis. In terms of object detection, the best results were achieved for HPV-related lesions (AP25 = 0.82) and proliferative verrucous leukoplakia (AP25 = 0.80; AP50 = 0.76), while the lowest values were noted for leukoplakia (AP25 = 0.36; AP50 = 0.20). Overall, the model performed comparable to specialists ($p = 0.93$) and significantly better than general dentists ($p < 0.01$).

Conclusion: The developed model performed as well as specialists in oral surgery, highlighting its potential as a valuable tool for oral lesion assessment.

Clinical significance: By providing performance comparable to oral surgeons and superior to general dentists, the developed multi-class model could support the clinical evaluation of oral lesions, potentially enabling earlier diagnosis of potentially malignant disorders, enhancing patient management and improving patient prognosis.

1. Introduction

Oral potentially malignant disorders (OPMDs) represent a diverse

group of clinically defined conditions, characterized by a variable risk of malignant transformation into oral squamous cell carcinoma (OSCC) [1, 2]. OPMDs present clinically as red, white or mixed red-white patches

* Corresponding author at: Clinic for Conservative Dentistry, Periodontology and Digital Dentistry, LMU University Hospital, LMU Munich, Goethestraße 70, 80336 Munich, Germany.

E-mail address: Falk.Schwendicke@med.uni-muenchen.de (F. Schwendicke).

<https://doi.org/10.1016/j.jdent.2025.105992>

Received 14 April 2025; Received in revised form 14 July 2025; Accepted 20 July 2025

Available online 20 July 2025

0300-5712/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

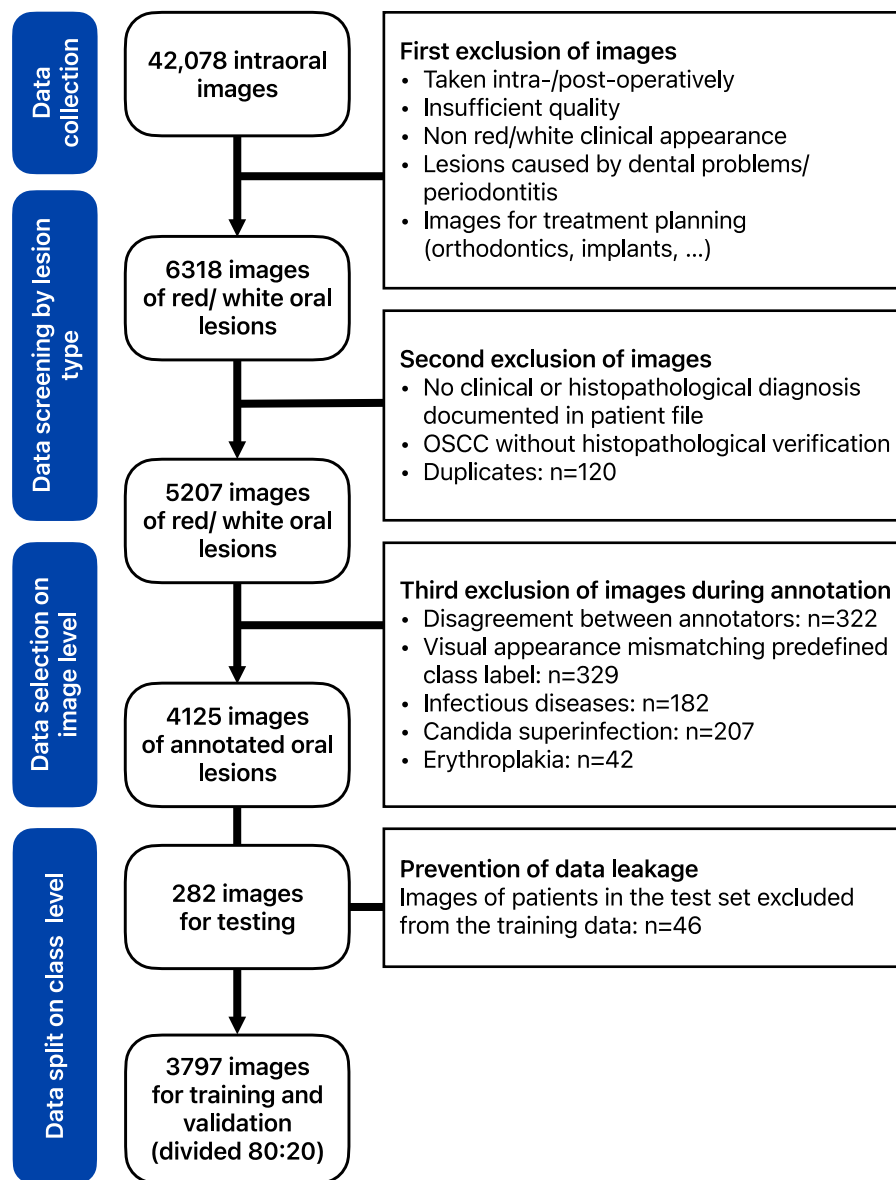


Fig. 1. Flowchart of the image selection process.

with a heterogenous appearance. Their estimated global prevalence is 4.5 % [3,4]. The primary challenge in the clinical context is to distinguish OPMDs and OSCC from non-transformative conditions with similar clinical appearance [5]. This is crucial, as the five-year overall survival rate for OSCC drops from 85 % when identified in stage I to <40 % when it has progressed to stage IV [3,6]. As of now, <30 % of patients with OSCC are diagnosed and treated in an early stage [7–10], and mortality remains high with 188,438 cases per year [11].

Detection and diagnosis of OPMDs and OSCC mainly relies on general dentists, who are an integral part of primary healthcare. During routine screening in general practice, oral lesions are diagnosed based on their clinical appearance, patient-related risk factors, medication and general illnesses. However, the wide spectrum and heterogenous clinical appearance of these lesions present a significant diagnostic challenge for dentists, with the accuracy of clinical evaluators varying considerably [12,13]. Absent or delayed targeted referrals due to diagnostic uncertainty negatively impacts on patient's clinical prognosis for OSCC [14]. Vice versa, unnecessary referrals or follow-up visits generate healthcare costs and additional burden to patients.

Artificial intelligence (AI), especially convolutional neural networks

(CNNs), hold a great potential for supporting dentists in the diagnostic screening processes for OPMDs and OSCC using photographic or histopathological images, computed tomography (CT) or optical coherence tomography (OCT) [15–20]. Multiple deep learning models have been employed for classification tasks, such as ResNet50 [21,22], VGG16 [23], LeNet-5, MobileNetV2 [24] or Inception V3 [25], sometimes combined with segmentation [26–30] or object detection tasks [29,31]. The ability of algorithms such as You Only Look Once (YOLO) to simultaneously detect and classify suspicious lesions in clinical images in real time is a significant advancement in the field [32]. A range of AI models trained on intraoral photographs have been developed recently, often with accuracies of 80 % and higher for distinguishing OSCC from a negative control (healthy mucosa) [24,28,33], suspicious versus non-suspicious mucosa [34,35], OSCC versus OPMDs [36], benign versus malignant tumors [37,38], referral cases versus non-referral cases [39] or oral mucosal ulcer versus normal mucosa [40]. Other studies used subjective labels, such as referral decisions based on assessments from one to seven experts, which could introduce bias associated with individual experts' judgment [41], and several were restricted to diagnosing only one site within the oral cavity, such as the tongue [38,

42–44]. Only a few studies developed and tested AI to distinguish multiple diagnoses, e.g. benign lesions, OPMDs and OSCC [45–50]. As guidelines for managing OPMDs and OSCC are based on a detailed differentiation of lesion types, AI models should ideally reflect most relevant diagnoses rather than grouping them together, as this reduces their clinical applicability [36]. In this aspect, a model distinguishing 16 different lesion types was recently published, although the lesions were mostly benign and restricted to characteristic sites within the oral cavity [51].

Eleven types of oral mucosal lesions with red or white clinical manifestations were selected in this study. These included OSCC and important OPMDs, such as proliferative verrucous leukoplakia (PVL) [52], low-risk human papillomavirus (HPV)-related lesions (i.e. papilloma, verruca vulgaris, condyloma acuminatum) [53], leukoplakia, and oral lichen planus (OLP) which was further subdivided into “white” (reticular, plaque-like) and “erosive” (erosive, atrophic) [54]. OPMDs must be distinguished from similar-looking lesions, including autoimmune bullous diseases such as pemphigus vulgaris [55] or lesions caused by trauma or infection, such as ulcerous lesions (including aphthous lesions) and candidiasis, both of which have been shown to increase the risk of carcinogenesis [56,57]. Lesions with a leukoplakic appearance that are linked to specific habits were also included, such as frictional keratosis, morsicatio buccarum/linguarum, and smoking-related hyperkeratosis of the palate (nicotine stomatitis). Additionally, lingua geographica was included for the associated diagnostic challenges and its prevalence.

The objective of the present study was to develop a CNN-based AI model for the detection and classification of the eleven types of oral mucosal lesions. To test the model’s performance in real-world settings, its predictions were compared with assessments by nine independent clinical examiners, including specialists in oral surgery and general dentists. The null hypothesis tested was that the diagnostic performance of the AI model would not be significantly different from clinicians.

2. Methods

This study was approved by the Ethics Committee of Charité – University hospital Berlin (EA1-277-22). Its reporting followed the Checklist for Artificial Intelligence in Dental Research [58].

2.1. Data acquisition, selection, labeling and preprocessing

A total of 42,078 intraoral photographs of oral lesions were collected over 20 years in the department of oral medicine and oral surgery of Charité – University hospital Berlin. The photographs were taken using a professional digital single-lens reflex (DSLR) camera (Canon EOS 100D, Nikon D3100, Nikon D50 or Canon EOS 5D Mark II) with a circular flash and saved in the .jpeg format. The images were not used in any previous studies.

Images were systematically screened as depicted in Fig. 1. Initially, images were excluded if they were taken for treatment planning, intra- or post-operatively, if they did not display lesions with red and/or white appearance, and if lesions were related to dental conditions or periodontitis. Furthermore, images of insufficient quality were excluded. If there were multiple photographs from a specific site and angle taken on a single screening date, only the highest-quality image was selected. For the remaining 6318 images, clinical diagnoses were collected from the patient management system High Dent Plus (CompuGroup Medical, Koblenz, Germany), and if such diagnosis was not available, the image was excluded. Then, patient-related metadata (age upon documentation, gender) were retrieved for the remaining images, and corresponding histopathological findings were identified if available. Non-histopathologically confirmed cases of OSCC were excluded. By removing 120 duplicates, the dataset was eventually reduced to 5207 images.

At this stage, the dataset was reviewed and subsequently labeled by

three annotators: a general dentist with 10 years of experience who received extensive training in the detection of oral lesions (JS), an oral surgeon with 5 years of experience (YM), and a maxillofacial surgeon with 9 years of experience (DS). YM and DS have been holding regular patient consultations regarding oral lesions in the department of maxillofacial surgery of a university hospital. Prior to the annotation process, annotators were calibrated on 129 randomly selected images which were not part of the final dataset (see below). Calibration was done in three rounds using the Redbrick AI annotation software (Zantula, Claymont, DE, USA), in which each annotator independently labeled lesions by placing bounding boxes and assigned each box to one of the 11 classes. Clinical diagnoses retrieved from the patient management system were visible to the annotators. Bounding boxes of different annotators were considered as in agreement if their intersection over union (IoU) was ≥ 0.5 . Fleiss kappa equaled 0.94 for inter-rater and 0.97 (JS), 0.93 (YM) and 0.96 (DS) for intra-rater reliability, indicating excellent agreement and consistency among annotators. In cases where they disagreed, annotators engaged in a consensus review to ensure labeling consistency and accuracy.

In the main dataset, lesions were either classified according to the histopathological finding, which was available in approximately 30 % of cases, or according to the clinical diagnosis, which had been documented by the treating physician based on clinical inspection, palpation and patient history. Each lesion was labeled using a bounding box (JS) and subsequently reviewed by one of the two surgeons (YM; DS). If the annotators agreed, the labels were used as the reference standard. If they disagreed ($n = 322$), the image was excluded from the dataset. Annotators further excluded: 329 images due to mismatch between the documented clinical diagnosis and the visual appearance of the lesion, i. e. if the image was not representative of the diagnosis; 182 images of infectious diseases; 207 cases of candidiasis which manifested as superinfections of other visually discernible lesions, and 42 images erythroplakia and erythroleukoplakia due to underrepresentation. Lastly, 46 images were excluded to prevent data leakage, as they stemmed from patients included in the test set. The final dataset of 4079 images was then randomly split into training data (3031 images), validation data (766 images) and test data (282 images).

To compensate for the imbalanced distribution of lesion types, oversampling and undersampling were employed for the training dataset. The overrepresented classes of OLP “white”, OLP “erosive” and leukoplakia were undersampled by excluding random images from the training set, whereas the underrepresented classes were oversampled by duplicating random images in the training set. As a result, the number of images per class was equivalent ($n = 450$). In addition, several image augmentation techniques were applied on-the-fly, including random rotation, flipping, scaling, and mosaicking in every epoch. Finally, all images were resized to 640×640 pixels.

2.2. Deep learning

We applied a state-of-the-art CNN for object detection, YOLO version 8 (YOLOv8) [59], as implemented by Jocher et al. [60]. The network utilized a DarkNet backbone to extract features at three levels with varying resolutions, thereby ensuring the detection of objects of differing sizes. The feature maps were then aggregated into a feature pyramid network, enhancing multi-scale object detection capabilities. Unlike previous versions, YOLOv8 features anchorless detection, where potential object locations are predicted directly for each pixel without relying on predefined anchor boxes. Additionally, the architecture includes a decoupled head, which separates the tasks of classification and localization (bounding box prediction). Allowing the model to optimize both tasks independently can improve its overall accuracy.

Model’s predictions were compared with the reference standard to compute the loss function, which comprises three components: classification, localization and objectness. Objectness refers to a confidence score representing the probability that an object is present in the

Table 1
Final dataset - distribution of oral lesion types (split into sets for training, validation and testing) and patient demographics.

Lesion type	Annotated images (n)			Median age (inter-quartile range) [years]	Sex (n)	
	Training	Validation	Testing		Male	Female
OSCC	161	43	24	62 (55–74)	104 (48 %)	113 (51 %)
PVL	92	25	9	60 (55–74)	44 (42 %)	62 (57 %)
Bullous diseases	187	47	25	70 (61–76)	66 (31 %)	139 (68 %)
HPV-related lesions	107	26	14	54 (40–63)	89 (67 %)	42 (32 %)
Leukoplakia	507	126	27	58 (48–66)	297 (47 %)	321 (52 %)
OLP "erosive"	636	166	44	64 (56–74)	183 (23 %)	588 (76 %)
OLP "white"	826	204	71	60 (51–68)	302 (29 %)	720 (70 %)
Ulcerous lesions	170	41	24	59 (39–70)	123 (53 %)	106 (45 %)
Candidiasis	179	42	21	63 (50–73)	103 (47 %)	114 (53 %)
Keratosis	102	28	13	49 (35–61)	79 (59 %)	56 (40 %)
Lingua geographica	64	18	10	60 (40–69)	50 (54 %)	42 (45 %)

Abbreviations: OSCC – oral squamous cell carcinoma; PVL – proliferative verrucous leukoplakia; HPV – human papilloma virus; OLP – oral lichen planus – "erosive" (erosive, atrophic) and "white" (reticular, plaque-like).

Table 2
Diagnostic accuracy of the final model for the 11 types of lesions on the test set.

	Sensitivity (Recall)	Specificity	F1-score	Precision	AUROC	AP25	AP50	AP75
OSCC	0.54	0.98	0.63	0.76	0.84	0.53	0.36	0.23
PVL	0.89	0.97	0.64	0.50	0.91	0.80	0.76	0.16
Bullous diseases	0.64	0.95	0.59	0.55	0.87	0.61	0.48	0.14
HPV-related lesions	0.86	0.99	0.80	0.75	0.96	0.82	0.49	0.22
Leukoplakia	0.59	0.94	0.54	0.50	0.86	0.36	0.20	0.15
OLP "erosive"	0.48	0.95	0.54	0.62	0.86	0.44	0.34	0.20
OLP "white"	0.65	0.89	0.66	0.67	0.84	0.54	0.37	0.15
Ulcerous lesions	0.54	0.98	0.62	0.72	0.85	0.47	0.38	0.13
Candidiasis	0.86	0.96	0.73	0.64	0.81	0.57	0.37	0.21
Keratosis	0.38	0.98	0.43	0.50	0.78	0.44	0.44	0.39
Lingua geographica	0.70	0.99	0.67	0.64	0.90	0.73	0.58	0.35

Abbreviations: AUROC – area under the receiver operating characteristic curve; AP25, AP50, AP75 – average precision at intersection over union (IoU) > 0.25, 0.5, and 0.75, respectively. Abbreviations: OSCC – oral squamous cell carcinoma; PVL – proliferative verrucous leukoplakia; HPV – human papilloma virus; OLP – oral lichen planus: "erosive" (erosive, atrophic) and "white" (reticular, plaque-like). The highest and lowest values for each metric are indicated in bold.

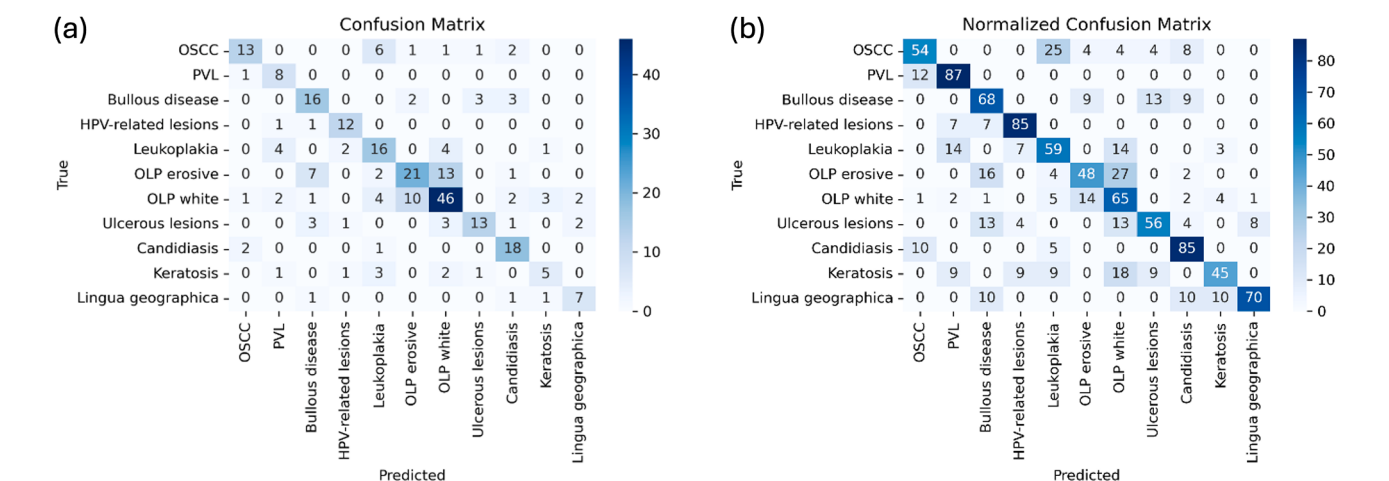


Fig. 2. The confusion matrix summarizes the performance of the final 11-class model. It displays overlap of the model's predicted labels (x-axis) and the reference standard labels (y-axis) on the test set. No lesion was detected in two images of the test set. Given the imbalanced data, the confusion matrices display both absolute (a) and normalized (b) values. Abbreviations: OSCC – squamous cell carcinoma; PVL – proliferative verrucous leukoplakia; human papilloma virus (HPV)-related (papilloma, verruca vulgaris, condyloma acuminatum); OLP – oral lichen planus: "erosive" (erosive, atrophic), "white" (reticular, plaque-like).

particular grid point, whereas the classification score assigns the grid point to a specific class. Model's weights were adjusted accordingly using backpropagation. With each training epoch, the networks were fine-tuned to predict with increasing accuracy based on the provided annotations. The batch size of the final model was 32. The Adam optimizer [61] was used with an initial learning rate of 0.01, which

decreased during training. The dropout rate was set to 0.0, and the model was trained for 300 epochs with a weight decay rate of 0.0005. The hardware used for training included a graphics card NVIDIA RTX A4000 with NVIDIA Ampere GPU Architecture, 16 GB GDDR6 memory size and 19.2 TFLOPS of FP32-performance.

The final model generated bounding boxes for the detected objects.

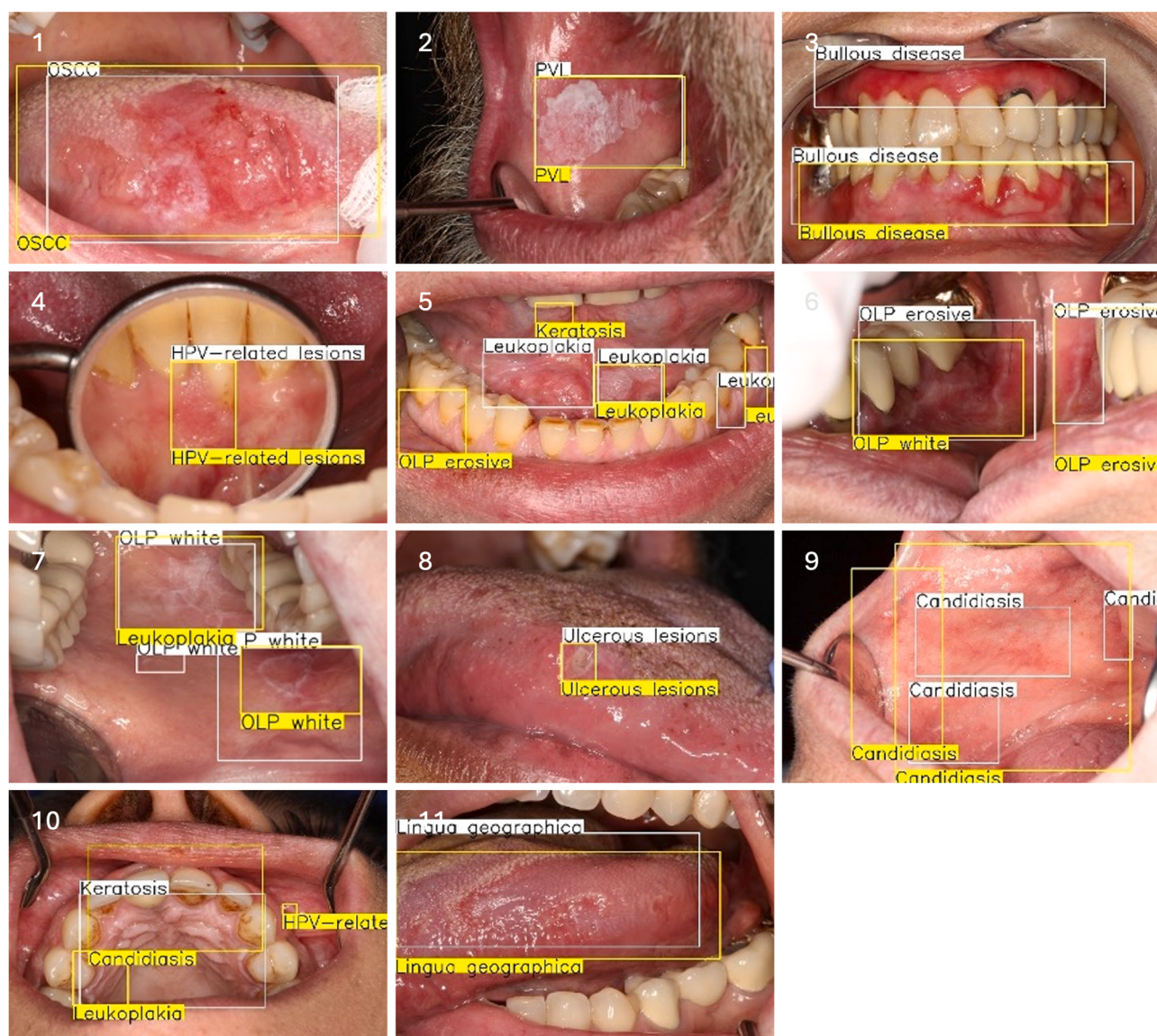


Fig. 3. Comparison of model's predictions (yellow) with the reference standard (white) for 11 classes of oral lesions: OSCC – oral squamous cell carcinoma (1), PVL – proliferative verrucous leukoplakia (2), Bullous disease (3), HPV (human papilloma virus)-related lesion: papilloma (4), Leukoplakia (5), OLP – oral lichen planus "erosive" (erosive, atrophic) (6), OLP "white" (reticular/plaque-like) (7), Ulcerous lesion (8), Candidiasis (9), Keratosis: nicotine stomatitis (10) and Lingua geographica (11). The model's ability to accurately classify and detect the extent of oral lesions depends on their clinical characteristics, clarity of boundaries, and visual complexity of the information presented. Image 1 presents correct classification and localization, but the predicted bounding box does not fully match the reference standard. Images 2, 4, 8 demonstrate perfect visual detection and classification. Image 3 shows correct detection for one of the two oral sites visible in the image, while the other was not detected by the model. Images 6, 7 illustrate the presence of mixed lesions, especially for different types of OLP, which were challenging for the model. Images 5, 10 illustrate the difficulties of the model when complementary information, such as extraoral structures, is presented within the image frame.

Non-maximum suppression was applied to eliminate redundant boxes, retaining only the most accurate and relevant predictions.

2.3. Comparison with clinicians

The performance of the final model was compared with a panel consisting of three specialists in oral surgery and six general dentists. The test set of 282 images was divided into three batches, each comprising 94 images and annotated by one specialist and two general dentists. The annotators received instructions on how to use the annotation software (RedBrick AI) and were asked to independently select a classification label and draw a bounding box for each lesion. There was no calibration of the general dentists and specialists, nor was any further information provided about the clinical diagnosis.

2.4. Evaluation

The classification performance of the model was evaluated using a confusion matrix and several statistical metrics: sensitivity (recall), specificity, F1-score, precision (positive predictive value), and the area under the receiver operating characteristic curve (AUROC). It should be noted that specificity and AUROC may be inflated by using the one-vs-the-rest approach for calculating true negatives in this multiclass scenario with an imbalanced dataset. Therefore, precision-recall curves were also constructed, as they are unaffected by true negatives and provide a more reliable evaluation under these conditions. Multiple thresholds (0.25, 0.5, 0.75) of intersection over union (IoU) were used for the precision-recall curves and for the calculation of average precision (AP), which corresponds to the area under the interpolated precision-recall curve.

Table 3
Comparison of the final model performance on the test set with human annotators.

	Accuracy	Precision	Recall	F1-score
Macro-averaged results (95 % CI)				
General dentists	0.41 (0.37–0.45)	0.42 (0.38–0.46)	0.42 (0.38–0.46)	0.39 (0.35–0.43)
Specialists	0.59 (0.53–0.65)	0.58 (0.52–0.63)	0.53 (0.46–0.59)	0.53 (0.47–0.59)
Final model	0.62 (0.56–0.68)	0.58 (0.51–0.66)	0.6 (0.54–0.68)	0.57 (0.51–0.65)
Weighted results (95 % CI)				
General dentists	0.46 (0.41–0.5)	0.49 (0.44–0.54)	0.41 (0.37–0.45)	0.41 (0.37–0.46)
Specialists	0.58 (0.5–0.64)	0.64 (0.59–0.7)	0.59 (0.53–0.65)	0.6 (0.55–0.66)
Final model	0.65 (0.59–0.71)	0.64 (0.58–0.7)	0.62 (0.56–0.68)	0.62 (0.56–0.68)

Macro-averaged and weighted metrics are presented to account for class imbalance; CI: confidence interval. The McNemar test indicated a significant difference between general dentists and the final model ($p < 0.01$), as well as between general dentists and specialists ($p < 0.01$). There was no significant difference between specialists and the final model ($p = 0.93$).

To account for class imbalance, we reported macro-averaged metrics, i.e. unweighted averages of metrics across all classes, as well as weighted averages that take class sizes into account. The 95 % confidence intervals were calculated using bootstrapping, employing resampling with replacement over 1000 iterations. Model predictions were compared with clinicians using the McNemar test at a significance level of 0.05.

3. Results

3.1. Model performance

Table 1 summarizes the distribution of oral lesion types in the final dataset and basic demographic information of the patients. The median age of the individuals ranged from 49 to 70 years; the youngest patient group presented with keratosis, while the oldest suffered from bullous diseases. The distribution of sex was relatively balanced, except for OLP and bullous diseases, which were more prevalent in women (≥ 68 %), and HPV-related lesions which prevailed in men (67 %).

The metrics of the final model on the test set are summarized in Table 2, its learning curves are shown in the appendix. In terms of classification, the highest F1-score (0.80) and AUROC (0.96) were observed for HPV-related lesions, whereas the lowest F1-score (0.43) and AUROC (0.78) were obtained for keratosis. Confusion matrices with absolute and normalized counts are presented in Fig. 2. Keratosis was most likely to be confused with leukoplakia or OLP "white". The most frequently observed classes of confusion were "white" and "erosive" forms of OLP, which can occur simultaneously. This is also illustrated in Fig. 3, which presents examples of model's predictions for each of the eleven classes, showing both correct detections and common errors.

In terms of object detection, i.e. considering the sizes and locations of bounding boxes, the best results were achieved for HPV-related lesions (AP25 = 0.82) and PVL (AP25 = 0.8; AP50 = 0.76). The lowest values were noted for leukoplakia (AP25 = 0.36; AP50 = 0.2), and at the IoU threshold of 0.75, low values were recorded for all lesion types (AP75 < 0.4).

3.2. Comparison with clinicians

Table 3 shows that the overall performance of the final model was comparable to that of specialists ($p = 0.93$); the model outperformed general dentists ($p < 0.01$). Results for individual classes are presented in Figs. 4 and 5, which display the min-max range of clinicians in comparison with the model's ROC and precision-recall curves, respectively. The ROC curves summarize classification accuracy of the model and show that it was outperformed by specialists in diagnosing OSCC, ulcerous lesions, candidiasis and keratosis. General dentists had lower sensitivity for most lesion types, and their variability was higher than that of specialists (Fig. 4). Unlike ROC curves, precision-recall curves consider IoU of the bounding boxes as well, which contributed to the inferior performance because correctly classified lesions were

considered as incorrect if IoU was < 0.5 .

4. Discussion

In this study, we developed a deep learning model to distinguish 11 types of oral lesions with various transformative potential, based on the official WHO classification system of OPMDs [2,62]. The diagnostic performance of the AI model was found to be similar to that of specialists, but significantly superior to that of general dentists. Our null hypothesis was therefore rejected.

The visual characteristics of OPMDs are often heterogeneous, and features of different OPMDs may overlap. Additionally, mixed forms of lesions are prevalent, which is why the diagnosis is clinically determined not only based on visual inspection, but also considering palpation, subjective symptoms and patient history, including risk factors [63] such as age, gender, smoking and alcohol consumption or viral, bacterial and mycotic infections. The development of the lesions over time is also an important factor, as some oral lesions are reactive and heal shortly after removing the cause, while others may show gradual progression in the long term; for instance, early PVL has been reported to mimic homogenous leukoplakia or OLP [64,65].

The final model showed the highest precision (0.76) in classifying OSCC, while its recall was suboptimal – almost half of OSCC cases were misclassified, most commonly as leukoplakia (Fig. 2). The specificity (0.98) and AUROC (0.84) were high, but these metrics were inflated by the one-vs-the-rest approach; so they need to be interpreted cautiously. The highest recall (0.89) and AP50 (0.76) were observed for PVL, while HPV-related lesions exhibited the second highest recall (0.86) and the highest values of F1 score (0.80) and AP25 (0.82). As AP accounts for the overlap of predicted bounding boxes with the reference standard, this score was higher for lesions with well-defined margins than for those with indistinct boundaries like OLP "erosive" or keratosis (0.44). The lowest AP25 (0.36) and precision (0.50) were seen for leukoplakia, reflecting the WHO definition of leukoplakia as "a predominantly white patch or plaque that cannot be characterized clinically or pathologically as any other disorder". Leukoplakia visually overlaps with several oral pathologies including keratosis, which had the lowest recall (0.38) and F1 score (0.43) (Table 2).

Previously developed models trained to classify just two different classes or visually distinct lesions, such as OSCC vs. healthy [33], ulcer vs. leukoplakia vs. healthy [47], or leukoplakia vs. OLP vs. OSCC vs. healthy [49], performed considerably better than our 11-class model. It is known that increasing the number of classifiable categories [15] decreases diagnostic performance of AI models, especially for multifaceted lesions. This mirrors the clinical reality, in which dentists are often uncertain in distinguishing between various oral lesions with similar clinical appearance [13]. While models differentiating two or slightly more classes of oral lesions may achieve high metrics, their clinical applicability is limited for differential diagnosis of multiple lesion types. To our knowledge, only one highly accurate multi-class model has been reported to date, using attention-guided classification to distinguish 16

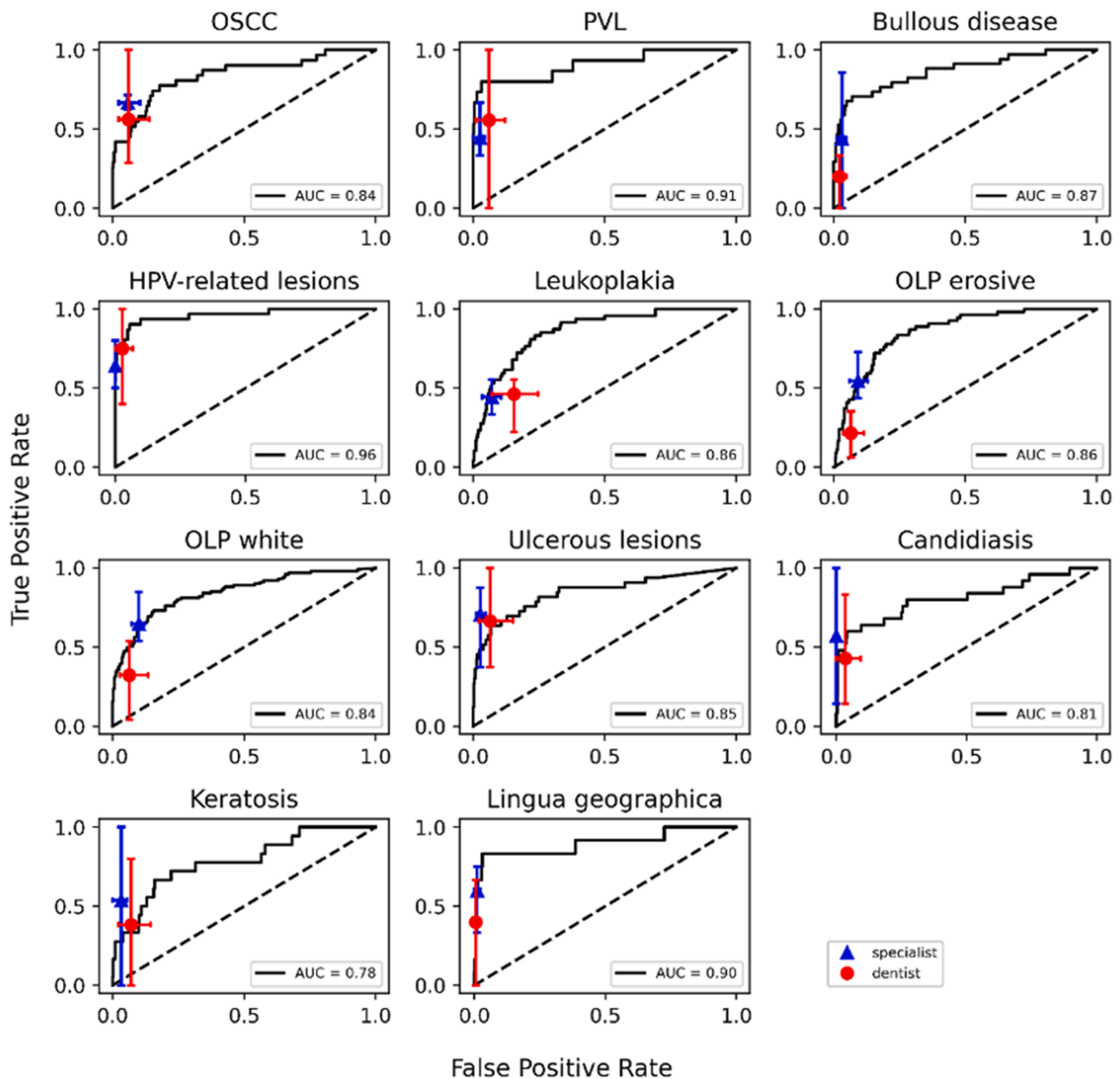


Fig. 4. Receiver operating characteristic curves of the final model for each of the 11 classes compared with specialists (blue) and dentists (red). The performance of human annotators is displayed as mean and Min-Max range. Abbreviations: OSCC – oral squamous cell carcinoma; PVL – proliferative verrucous leukoplakia; HPV – human papilloma virus; OLP – oral lichen planus – “erosive” (erosive, atrophic) and “white” (reticular, plaque-like); AUC – area under the curve.

classes while considering the specific anatomical site [51]. However, most of the lesions in that study were benign, which makes the comparison to our study difficult. In addition, the lesions may be visually more distinct and restricted to a specific region of the oral cavity, making the classification task easier.

In our study, approximately 30 % of lesions were confirmed by histopathology, including all OSCC cases, and the remaining photographs were clinically classified by dental professionals privy to patient-related information such as age, gender and habits. This information was available to the annotators of the training dataset, which contributed to their high inter- and intra-rater agreement. In contrast, the clinical information was available neither to the model, which was trained and tested only based on the visual features, nor to the specialists and dentists. Combined with the high number of lesion types, this could contribute to their lower overall diagnostic accuracy on the test set. Notably, the overall performance of the model was similar to that of

specialists, highlighting the difficulty of diagnosing OPMDs with similar features only based on visual inspection, but also demonstrating the value of having an AI-based diagnostic support available to general practitioners.

The use of AI-based diagnostic tools could help in timely diagnosis of OPMDs and OSCC, especially when combined with other clinical data. It could also help reduce the subjectivity and variability of the dentist’s assessment, possibly resulting in a faster referral or treatment, ultimately improving the prognosis. However, models need to cover a wide spectrum of lesions and achieve high accuracy, as false positive results may lead to unnecessary concerns and costs, while false negatives may be critical in case of malignant lesions [66]. Beyond performance metrics, a critical challenge lies in the opaque, “black box” nature of many deep learning models, which limits the ability to understand how predictions are made. This lack of transparency undermines clinician trust, especially in sensitive domains such as oncology, where explainability

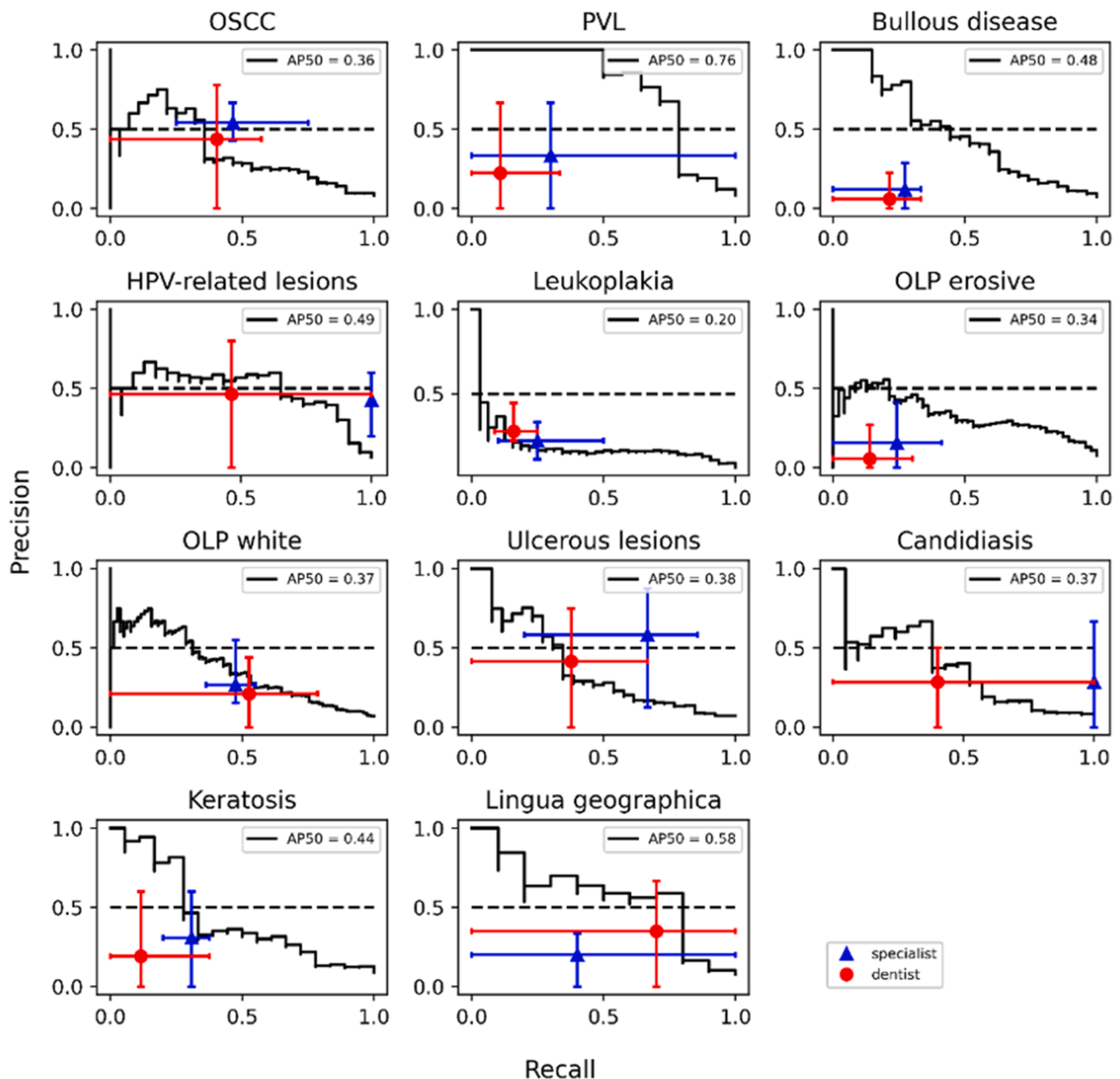


Fig. 5. Precision-recall curves of the final model for each of the 11 classes at the threshold of intersection over union – IoU = 0.5. The model was compared with specialists (blue) and dentists (red); their performance is displayed as mean and Min-Max range. Abbreviations: OSCC – oral squamous cell carcinoma; PVL – proliferative verrucous leukoplakia; HPV – human papilloma virus; OLP – oral lichen planus – “erosive” (erosive, atrophic) and “white” (reticular, plaque-like); AP50 – average precision at IoU $\geq$ 0.5.

and accountability are essential. Therefore, future research should explore explainable AI approaches and uncertainty quantification techniques, which may support clinical deployment of AI tools [67–70]. Moreover, ethical AI development must also address accessibility and equity, particularly in underserved regions where diagnostic resources are scarce [71,72]. In such contexts, AI tools could help reduce disparities and support earlier detection and referral.

This study comes with a number of limitations. First, both the model and the comparator dentists relied solely on image data when making their diagnoses. In clinical settings, further data would be available. Second, the data stemmed from one (albeit large) hospital, which may have an impact on generalizability. Notably, the data were collected over two decades by a wide range of practitioners using different sensors, increasing the heterogeneity of the data pool. Third, the underrepresentation of some classes either hindered model performance or

even led to their exclusion, e.g. in case of erythroplakia, an important OPMD. Future studies should aim at using additional patient-related information and balancing the dataset, possibly involving multiple centers to achieve sufficient sample size and to improve model generalizability. If data sharing is not possible for legal or ethical reasons, paradigms such as federated learning could be employed.

5. Conclusion

We developed a multi-class deep learning model distinguishing various oral lesions, which performed comparable to specialists in oral surgery and significantly outperformed general dentists. However, when relying solely on visual information, machine learning models encounter the same diagnostic challenges as clinicians in classifying various oral lesions with similar clinical appearance. This study highlighted the need

for diagnostic support in identifying OPMDs, particularly for general dentists. By integrating AI models into clinical practice, dentists may detect OPMDs at an earlier stage, ultimately contributing to enhanced prognosis and patient outcomes.

Funding

This work was supported by the Innovation Committee of the federal joint committee (G-BA) of Germany [grant number 01VSF22025 (KIDS)].

Ethical approval

This study was approved by the Ethics Committee of Charité – University hospital Berlin (EA1–277–22).

CRediT authorship contribution statement

Julia Schwärzler: Writing – original draft, Project administration, Methodology, Investigation, Data curation. **Ekaterina Tolstaya:** Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Antonin Tichy:** Writing – review & editing, Validation, Project administration, Formal analysis. **Sebastian Paris:** Writing – review & editing, Resources. **Ghazal Aarabi:** Writing – review & editing, Validation. **Akhilanand Chaurasia:** Writing – review & editing, Validation. **Yoana Malenova:** Writing – review & editing, Data curation. **David Steybe:** Writing – review & editing, Data curation. **Falk Schwendicke:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors have no conflict of interest to declare. Falk Schwendicke is co-founder of the startup dentalXrai GmbH. dentalXrai GmbH did not have any role in conceiving, conducting or reporting this study. The authors are solely responsible for the contents of this paper.

Acknowledgements

All clinical images were kindly provided by the department of oral surgery at Charité – University hospital Berlin, by Prof. Andrea Schmidt-Westhausen and Prof. Hendrik Dommisch. The authors would also like to express gratitude to the specialists and dentists who annotated the testing set for comparison with the developed model.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.jdent.2025.105992](https://doi.org/10.1016/j.jdent.2025.105992).

References

- [1] O. Iocca, T.P. Sollecito, F. Alawi, G.S. Weinstein, J.G. Newman, A. De Virgilio, P. Di Maio, G. Spriano, S. Pardiñas López, R.M. Shanti, Potentially malignant disorders of the oral cavity and oral dysplasia: a systematic review and meta-analysis of malignant transformation rate by subtype, *Head Neck* 42 (2020) 539–555, <https://doi.org/10.1002/hed.26006>.
- [2] S. Muller, W.M. Tilakaratne, Update from the 5th edition of the World Health Organization classification of head and neck tumors: tumours of the oral cavity and mobile tongue, *Head. Neck. Pathol.* 16 (2022) 54–62.
- [3] C. Grafton-Clarke, K.W. Chen, J. Wilcock, Diagnosis and referral delays in primary care for oral squamous cell cancer: a systematic review, *Br. J. Gen. Pract.* 69 (2019) e112–e126, <https://doi.org/10.3399/bjgp18X700205>.
- [4] F.W. Mello, A.F.P. Miguel, K.L. Dutra, A.L. Porporatti, S. Warnakulasuriya, E.N. S. Guerra, E.R.C. Rivero, Prevalence of oral potentially malignant disorders: a systematic review and meta-analysis, *J. Oral Pathol. Med.* 47 (2018) 633–640, <https://doi.org/10.1111/jop.12726>.
- [5] S. Warnakulasuriya, White, red, and mixed lesions of oral mucosa: a clinicopathologic approach to diagnosis, *Periodontol* 80 (2019) 89–104, <https://doi.org/10.1111/prd.12276>, 2000.
- [6] S. Abati, C. Bramati, S. Bondi, A. Lissani, M. Trimarchi, Oral cancer and precancer: a narrative review on the relevance of early diagnosis, *Int. J. Environ. Res. Public Health* 17 (2020) 9160, <https://doi.org/10.3390/ijerph17249160>.
- [7] P. Güneri, J.B. Epstein, Late stage diagnosis of oral cancer: components and possible solutions, *Oral Oncol.* 50 (2014) 1131–1136, <https://doi.org/10.1016/j.oraloncology.2014.09.005>.
- [8] R. Mehrotra, D.K. Gupta, Exciting new advances in oral cancer diagnosis: avenues to early detection, *Head. Neck. Oncol.* 3 (2011) 33, <https://doi.org/10.1186/1758-3284-3-33>.
- [9] M. Rutkowska, S. Hnitecka, M. Nahajowski, M. Dominiak, H. Gerber, Oral cancer: the first symptoms and reasons for delaying correct diagnosis and appropriate treatment, *Adv. Clin. Exp. Med.* 29 (2020) 735–743, <https://doi.org/10.17219/acem/116753>.
- [10] C. Saka-Herrán, E. Jané-Salas, A. Mari-Roig, A. Estrugo-Devesa, J. López-López, Time-to-treatment in oral cancer: causes and implications for survival, *Cancers (Basel)* 13 (2021) 1321, <https://doi.org/10.3390/cancers13061321>.
- [11] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA. Cancer J. Clin.* 71 (2021) 209–249, <https://doi.org/10.3322/caac.21660>.
- [12] M. Essat, K. Cooper, A. Bessey, M. Clowes, J.B. Chilcott, K.D. Hunter, Diagnostic accuracy of conventional oral examination for detecting oral cavity cancer and potentially malignant disorders in patients with clinically evident oral lesions: systematic review and meta-analysis, *Head Neck* 44 (2022) 998–1013, <https://doi.org/10.1002/hed.26992>.
- [13] M. Idrees, R. Halimi, S. Gadiraju, A.M. Frydrych, O. Kujan, Clinical competency of dental health professionals and students in diagnosing oral mucosal lesions, *Oral Dis.* 30 (2024) 3108–3116, <https://doi.org/10.1111/odi.14743>.
- [14] K. Hertrampf, M. Kunkel, S2k guideline “diagnostics and management of precursor lesions of oral squamous cell carcinoma in dental and oral medicine, *Dtsch. Zahnärztl. Z. int.* 2 (2020) 188–195.
- [15] J. De Chauveron, M. Unger, G. Lescaille, L. Wendling, C. Kurtz, J. Rochefort, Artificial intelligence for oral squamous cell carcinoma detection based on oral photographs: a comprehensive literature review, *Cancer Med.* 13 (2024) e6822, <https://doi.org/10.1002/cam4.6822>.
- [16] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, *Nat. Med.* 25 (2019) 24–29, <https://doi.org/10.1038/s41591-018-0316-z>.
- [17] T.D. Pham, M.-T. Teh, D. Chatzopoulou, S. Holmes, P. Coulthard, Artificial intelligence in head and neck cancer: innovations, applications, and future directions, *Curr. Oncol.* 31 (2024) 5255–5290, <https://doi.org/10.3390/currenol31090389>.
- [18] S. Xu, Y. Liu, W. Hu, C. Zhang, C. Liu, Y. Zong, S. Chen, Y. Lu, L. Yang, E.Y.K. Ng, Y. Wang, Y. Wang, An early diagnosis of oral cancer based on three-dimensional convolutional neural networks, *IEEE Access* 7 (2019) 158603–158611, <https://doi.org/10.1109/ACCESS.2019.2950286>.
- [19] A. Soni, P.K. Sethy, A.K. Dewangan, A. Nanthaamornphong, S.K. Behera, B. Devi, Enhancing oral squamous cell carcinoma detection: a novel approach using improved EfficientNet architecture, *BMC Oral Health* 24 (2024) 601, <https://doi.org/10.1186/s12903-024-04307-5>.
- [20] Z. Yang, H. Pan, J. Shang, J. Zhang, Y. Liang, Deep-learning-based automated identification and visualization of oral cancer in optical coherence tomography images, *Biomedicine* 11 (2023) 802, <https://doi.org/10.3390/biomedicine11030802>.
- [21] S.K.D. Sharma, Oral squamous cell detection using deep learning, (2024). <https://doi.org/10.48550/ARXIV.2408.08939>.
- [22] S. Panigrahi, B.S. Nanda, R. Bhuyan, K. Kumar, S. Ghosh, T. Swarnkar, Classifying histopathological images of oral squamous cell carcinoma using deep transfer learning, *Heliyon* 9 (2023) e13444, <https://doi.org/10.1016/j.heliyon.2023.e13444>.
- [23] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, (2014). <https://doi.org/10.48550/ARXIV.1409.1556>.
- [24] P. Kulkarni, N. Sarwe, A. Pingale, Y. Sarolkar, R.R. Patil, G. Shinde, G. Kaur, Exploring the efficacy of various CNN architectures in diagnosing oral cancer from squamous cell carcinoma, *MethodsX* 13 (2024) 103034, <https://doi.org/10.1016/j.mex.2024.103034>.
- [25] D. Pandiar, S. Choudhary, R. Poonthakulath Krishnan, Application of InceptionV3, SqueezeNet, and VGG16 convoluted neural networks in the image classification of oral squamous cell carcinoma: a cross-sectional study, *Cureus* (2023), <https://doi.org/10.7759/cureus.49108>.
- [26] R. Dharani, K. Danesh, Oral cancer segmentation and identification system based on histopathological images using MaskMeanShiftCNN and SV-OnionNet, *Intell.-Based Med.* 10 (2024) 100185, <https://doi.org/10.1016/j.ibmed.2024.100185>.
- [27] A. Albishri, S.J.H. Shah, Y. Lee, R. Wang, OCU-Net: a novel U-net architecture for enhanced oral cancer segmentation, (2023). <https://doi.org/10.48550/ARXIV.2310.02486>.
- [28] A.A. Alzahrani, J. Alsamri, M. Maashi, N. Negm, S.A. Askany, A. Alkharashi, H. Alkhiri, M. Obayya, Deep structured learning with vision intelligence for oral carcinoma lesion segmentation and classification using medical imaging, *Sci. Rep.* 15 (2025) 6610, <https://doi.org/10.1038/s41598-025-89971-5>.

- [29] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, Automatic classification and detection of oral cancer in photographic images using deep learning algorithms, *J. Oral Pathol. Med.* 50 (2021) 911–918, <https://doi.org/10.1111/jop.13227>.
- [30] L. Li, C. Pu, J. Tao, L. Zhu, S. Hu, B. Qiao, L. Xing, B. Wei, C. Shi, P. Chen, H. Zhang, Development of an oral cancer detection system through deep learning, *BMC. Oral Health* 24 (2024) 1468, <https://doi.org/10.1186/s12903-024-05195-5>.
- [31] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, S. Vicharueang, AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer, *PLoS One* 17 (2022) e0273508, <https://doi.org/10.1371/journal.pone.0273508>.
- [32] N. Palanivel, D. S. L.P. G, S. B. S.M. M, The art of YOLOv8 algorithm in cancer diagnosis using medical imaging, in: 2023 Int. Conf. Syst. Comput. Autom. Netw. ICSCAN, IEEE, 2023, pp. 1–6, <https://doi.org/10.1109/ICSCAN58655.2023.10395046>.
- [33] Q. Fu, Y. Chen, Z. Li, Q. Jing, C. Hu, H. Liu, J. Bao, Y. Hong, T. Shi, K. Li, H. Zou, Y. Song, H. Wang, X. Wang, Y. Wang, J. Liu, H. Liu, S. Chen, R. Chen, M. Zhang, J. Zhao, J. Xiang, B. Liu, J. Jia, H. Wu, Y. Zhao, L. Wan, X. Xiong, A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: a retrospective study, *EClinicalMedicine* 27 (2020) 100558, <https://doi.org/10.1016/j.eclinm.2020.100558>.
- [34] V. Talwar, P. Singh, N. Mukhia, A. Shetty, P. Birur, K.M. Desai, C. Sunkavalli, K. S. Varma, R. Sethuraman, C.V. Jawahar, P.K. Vinod, AI-assisted screening of oral potentially malignant disorders using smartphone-based photographic images, *Cancers (Basel)* 15 (2023) 4120, <https://doi.org/10.3390/cancers15164120>.
- [35] R.D. Uthoff, B. Song, S. Sunny, S. Patrick, A. Suresh, T. Kolar, G. Keerthi, O. Spire, A. Anbarani, P. Wilder-Smith, M.A. Kuriakose, P. Birur, R. Liang, Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities, *PLoS One* 13 (2018) e0207493, <https://doi.org/10.1371/journal.pone.0207493>.
- [36] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, P. Jantana, S. Vicharueang, AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer, *PLoS One* 17 (2022) e0273508, <https://doi.org/10.1371/journal.pone.0273508>.
- [37] B.R. Nanditha, A. Geetha, H.S. Chandrashekar, M.S. Dinesh, S. Murali, An ensemble deep neural network approach for oral cancer screening, *Int. J. Onl. Biomed. Eng.* (2021) 121–134, <https://doi.org/10.3991/ijoe.v17i02.19207>.
- [38] F. Jubair, O. Al-karadsheh, D. Malamos, S. Al Mahdi, Y. Saad, Y. Hassona, A novel lightweight deep convolutional neural network for early detection of oral cancer, *Oral Dis.* 28 (2022) 1123–1130, <https://doi.org/10.1111/odi.13825>.
- [39] R.A. Welikala, P. Remagnino, J.H. Lim, C.S. Chan, S. Rajendran, T.G. Kallarakkal, R.B. Zain, R.D. Jayasinghe, J. Rimal, A.R. Kerr, R. Amtha, K. Patil, W. M. Tilakaratne, J. Gibson, S.C. Cheong, S.A. Barman, Fine-tuning deep learning architectures for early detection of oral cancer, in: G. Bebis, M. Alekseyev, H. Cho, J. Gevertz, M. Rodriguez Martinez (Eds.), *Math. Comput. Oncol.*, Springer International Publishing, Cham, 2020, pp. 25–31, https://doi.org/10.1007/978-3-030-64511-3_3.
- [40] J. Guo, H. Wang, X. Xue, M. Li, Z. Ma, Real-time classification on oral ulcer images with residual network and image enhancement, *IET. Image Process.* 16 (2022) 641–646, <https://doi.org/10.1049/ipr.12144>.
- [41] R.A. Welikala, P. Remagnino, J.H. Lim, C.S. Chan, S. Rajendran, T.G. Kallarakkal, R.B. Zain, R.D. Jayasinghe, J. Rimal, A.R. Kerr, R. Amtha, K. Patil, W. M. Tilakaratne, J. Gibson, S.C. Cheong, S.A. Barman, Automated detection and classification of oral lesions using deep learning for early detection of oral cancer, *IEEe Access.* 8 (2020) 132677–132693, <https://doi.org/10.1109/ACCESS.2020.3010180>.
- [42] A.R. Hassoon, A. Al-Naji, G.A. Khalid, J. Chahl, Tongue disease prediction based on machine learning algorithms, *Technologies* 12 (2024) 97, <https://doi.org/10.3390/technologies12070097>.
- [43] S.-J. Lee, H.-J. Oh, Y.-D. Son, J.-H. Kim, I.-J. Kwon, B. Kim, J.-H. Lee, H.-K. Kim, Enhancing deep learning classification performance of tongue lesions in imbalanced data: mosaic-based soft labeling with curriculum learning, *BMC Oral Health* 24 (2024) 161, <https://doi.org/10.1186/s12903-024-03898-3>.
- [44] M.Z.M. Shamim, S. Syed, M. Shiblee, M. Usman, S.J. Ali, H.S. Hussein, M. Farrag, Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer, *Comput. J.* 65 (2022) 91–104, <https://doi.org/10.1093/comjnl/bxaa136>.
- [45] G.A.I. Devindi, D.M.D.R. Dissanayake, S.N. Liyanage, F.B.A.H. Francis, M.B. D. Pavithya, N.S. Piyarathne, P.V.K.S. Hettiarachchi, R.M.S.G.K. Rasnayaka, R. D. Jayasinghe, R.G. Ragel, I. Nawinne, Multimodal deep convolutional neural network pipeline for AI-assisted early detection of oral cancer, *IEEe Access.* 12 (2024) 124375–124390, <https://doi.org/10.1109/ACCESS.2024.3454338>.
- [46] H. Lin, H. Chen, J. Lin, Deep neural network uncertainty estimation for early oral cancer diagnosis, *J. Oral Pathol. Med.* 53 (2024) 294–302, <https://doi.org/10.1111/jop.13536>.
- [47] R. Shah, J. Pareek, Non-invasive primary screening of oral lesions into binary and multi class using convolutional neural network, stratified K-fold validation and transfer learning, *Indian J. Sci. Technol.* 17 (2024) 651–659, <https://doi.org/10.17485/IJST/v17i7.2670>.
- [48] G. Tanriver, M. Soluk Tekkesin, O. Ergen, Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders, *Cancers (Basel)* 13 (2021) 2766, <https://doi.org/10.3390/cancers13112766>.
- [49] S. Vinayahalingam, N. Van Nistelrooij, R. Rothweiler, A. Tel, T. Verhoeven, D. Tröltzsch, M. Kesting, S. Bergé, T. Xi, M. Heiland, T. Flügge, Advancements in diagnosing oral potentially malignant disorders: leveraging vision transformers for multi-class detection, *Clin. Oral Investig.* 28 (2024) 364, <https://doi.org/10.1007/s00784-024-05762-8>.
- [50] Y. Dinesh, K. Ramalingam, P. Ramani, R.M. Deepak, Machine learning in the detection of oral lesions with clinical intraoral images, *Cureus.* 15 (2023) e44018, <https://doi.org/10.7759/cureus.44018>.
- [51] A. Patel, C. Besombes, T. Dillibabu, M. Sharma, F. Tamimi, M. Ducret, P. Chauvin, S. Madathil, Attention-guided convolutional network for bias-mitigated and interpretable oral lesion classification, *Sci. Rep.* 14 (2024) 31700, <https://doi.org/10.1038/s41598-024-81724-0>.
- [52] K. Staines, H. Rogers, Proliferative verrucous leukoplakia: a general dental practitioner-focused clinical review, *Br. Dent. J.* 236 (2024) 297–301, <https://doi.org/10.1038/s41415-024-7066-8>.
- [53] M.S. Lipsky, G. Wolfe, B.A. Radilla, M. Hung, Human Papillomavirus: a narrative review for dental providers in prevention and care, *Int. J. Environ. Res. Public. Health* 22 (2025) 439, <https://doi.org/10.3390/ijerph22030439>.
- [54] S.L. Roberts, R. Bhamra, V. Ilankovan, Malignant transformation rate of erosive oral lichen planus: a retrospective study, *Br. J. Oral Maxillofac. Surg.* 62 (2024) 788–793, <https://doi.org/10.1016/j.bjoms.2023.11.020>.
- [55] X. Feng, H. Zheng, M. Wang, Y. Wang, X. Zhou, X. Zhang, J. Li, Y. Xiao, M. Wei, X. Li, T. Hashimoto, J. Li, W. Li, Autoimmune bullous diseases: pathogenesis and clinical management, *Mol. Biomed.* 6 (2025) 30, <https://doi.org/10.1186/s43556-025-00272-9>.
- [56] G. Gilligan, E. Piemonte, J. Lazos, M.C. Simancas, R. Panico, S. Warnakulasuriya, Oral squamous cell carcinoma arising from chronic traumatic ulcers, *Clin. Oral Investig.* 27 (2022) 193–201, <https://doi.org/10.1007/s00784-022-04710-8>.
- [57] G. M. S.S.S. Ravi, D. Maheswary, K.V. Leela, R. Harikumar Lathakumari, K.S. LP, Role of Candida albicans in chronic inflammation and the development of oral squamous cell carcinoma, *Cancer Pathog. Ther.* (2025), <https://doi.org/10.1016/j.cpt.2025.03.002>, S2949713225000321.
- [58] F. Schwendicke, T. Singh, J.-H. Lee, R. Gaudin, A. Chaurasia, T. Wiegand, S. Uribe, J. Krois, Artificial intelligence in dental research: checklist for authors, reviewers, readers, *J. Dent.* 107 (2021) 103610, <https://doi.org/10.1016/j.jdent.2021.103610>.
- [59] R. Varghese, S. M. YOLOv8: a novel object detection algorithm with enhanced performance and robustness, in: 2024 Int. Conf. Adv. Data Eng. Intell. Comput. Syst. ADICS, 2024, pp. 1–6, <https://doi.org/10.1109/ADICS58448.2024.10533619>.
- [60] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLO. <https://github.com/ultralytics/ultralytics>, 2023 (accessed 2 September 2024).
- [61] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization (2017). <https://doi.org/10.48550/arXiv.1412.6980>.
- [62] D. Baumhoer, T.E. Reichert, Aktuelle WHO-klassifikation der oralen potenziell malignen erkrankungen (OPMD), *MKG-Chir.* 17 (2024) 72–81, <https://doi.org/10.1007/s12285-024-00470-4>.
- [63] A. Barsouk, J.S. Aluru, P. Rawla, K. Saginala, A. Barsouk, Epidemiology, risk factors, and prevention of head and neck squamous cell carcinoma, *Med. Sci.* (2023), <https://doi.org/10.3390/medsci11020042>.
- [64] A.M. Gillenwater, N. Vigneswaran, H. Fatani, P. Saintigny, A.K. El-Naggar, Proliferative verrucous leukoplakia: recognition and differentiation from conventional leukoplakia and mimics, *Head Neck* 36 (2014) 1662–1668, <https://doi.org/10.1002/hed.23505>.
- [65] G. Gilligan, F. Garola, E. Piemonte, N. Leonardi, R. Panico, S. Warnakulasuriya, Lichenoid proliferative leukoplakia, lichenoid lesions with evolution to proliferative leukoplakia or a continuum of the same precancerous condition? A revised hypothesis, *J. Oral Pathol. Med.* 50 (2021) 129–135, <https://doi.org/10.1111/jop.13133>.
- [66] R. Giddings, A. Joseph, T. Callender, S.M. Janes, M. Van Der Schaar, J. Sheringham, N. Navani, Factors influencing clinician and patient interaction with machine learning-based risk prediction models: a systematic review, *Lancet Digit. Health* 6 (2024) e131–e144, [https://doi.org/10.1016/S2589-7500\(23\)00241-8](https://doi.org/10.1016/S2589-7500(23)00241-8).
- [67] J. Petch, J.P. Tabja Bortesi, W. Nelson, S. Di, M.H. Mamdani, Should I trust this model? Explainability and the black box of artificial intelligence in medicine. *Artif. Intell. Med.*, Elsevier, 2024, pp. 265–273, <https://doi.org/10.1016/B978-0-443-13671-9.00015-6>.
- [68] K.C. Figueroa, B. Song, S. Sunny, S. Li, K. Gurushanth, P. Mendonca, N. Mukhia, S. Patrick, S. Gurudath, S. Raghavan, T. Imchen, S.T. Leivon, T. Kolar, V. Shetty, V. Bushan, R. Ramesh, V. Pillai, P. Wilder-Smith, A. Sigamani, A. Suresh, M. A. Kuriakose, P. Birur, R. Liang, Interpretable deep learning approach for oral cancer classification using guided attention inference network, *J. Biomed. Opt.* 27 (2022), <https://doi.org/10.1117/1.JBO.27.1.015001>.
- [69] T. Hulsen, Explainable artificial intelligence (XAI): concepts and challenges in healthcare, *AI* 4 (2023) 652–666, <https://doi.org/10.3390/ai4030034>.
- [70] M. Parola, F.A. Galatolo, G. La Mantia, M.G.C.A. Cimino, G. Campisi, O. Di Fede, Towards explainable oral cancer recognition: screening on imperfect images via Informed Deep Learning and case-based reasoning, *Comput. Med. Imaging Graph.* 117 (2024) 102433, <https://doi.org/10.1016/j.compmedimag.2024.102433>.
- [71] P. Badri, H. Lai, S. Ganatra, V. Baracos, M. Amin, Factors associated with oral cancerous and precancerous lesions in an underserved community: a cross-sectional study, *Int. J. Environ. Res. Public. Health* 19 (2022) 1297, <https://doi.org/10.3390/ijerph19031297>.
- [72] D. Adorno-Farias, S. Morales-Pisón, G. Gischkow-Rucatti, S. Margarit, R. Fernández-Ramires, Genetic and epigenetic landscape of early-onset oral squamous cell carcinoma: insights of genomic underserved and underrepresented populations, *Genet. Mol. Biol.* (2024), <https://doi.org/10.1590/1678-4685-gmb-2024-0036>.