Brazilian Dental Science

unesp
UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Ciência e Tecnologia
Campus de São José dos Campos

Review

DOI: https://doi.org/10.4322/bds.2026.e5017

AI in early oral cancer detection: a systematic review of technologies and clinical impact

IA no diagnóstico precoce do câncer bucal: uma revisão sistemática de tecnologias e impacto clínico

Luigi de Almeida ALBERTONI[1], Letícia dos Santos Sales MARTINS[1], Cecília Helpes RODRIGUES[1], Lara Castor CUNHA[1], Henrique Souza MAGALHÃES[1], Bruno Alves FERREIRA[1], Lívia Garcia RESENDE[1], Quézia Soares de PAULA[1], Maria Clara Fernandes Ribeiro DANTAS[1]*, Gisele Maria Campos FABRI[1]

1 - Universidade Federal de Juiz de Fora, Faculdade de Odontologia, Departamento de Odontopediatria e Ortodontia. Juiz de Fora, MG, Brazil.

How to cite: Albertoni LA, Martins LSS, Rodrigues CH, Cunha LC, Magalhães HS, Ferreira BA, et al. AI in early oral cancer detection: a systematic review of technologies and clinical impact. Braz. Dent. Sci. 2026;29:e5017. https://doi.org/10.4322/bds.2026.e5017

ABSTRACT

Objective: To evaluate the effectiveness and clinical applicability of artificial intelligence (AI) in the early diagnosis of oral cancer compared to conventional methods. **Material and Methods:** This systematic review was prospectively registered in the PROSPERO database under the number CRD420251083609. A descriptive systematic review was conducted based on the PICO question: "In patients with suspected oral cancer, is AI more effective than traditional methods for early diagnosis?" Searches were performed between April and June 2025 in PubMed, LILACS, SciELO, Scopus, Web of Science, Embase, and Google Scholar using the keywords "artificial intelligence," "oral neoplasm," and "early diagnosis." Comparative studies reporting diagnostic metrics such as sensitivity, specificity, and accuracy were included. Methodological quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies – Artificial Intelligence extension (QUADAS-AI) and the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) system. **Results:** Of 1,704 identified studies, 12 met eligibility criteria. Most studies employed retrospective observational designs, primarily using convolutional neural networks (CNNs) or hybrid models on clinical and histopathological images. Reported diagnostic accuracy was generally above 80%, and some lightweight models demonstrated potential for remote screening. Common limitations included lack of external validation, methodological heterogeneity, and dependence on image quality. **Conclusion:** AI shows promising potential to support early diagnosis of oral cancer, improving diagnostic speed and accuracy. Broader clinical implementation will require multicenter validation, standardized datasets, and integration with clinical and histopathological evaluation.

KEYWORDS

Artificial Intelligence; Computer-aided diagnosis; Early diagnosis; Mouth neoplasms; Squamous Cell Carcinoma of Head and Neck.

RESUMO

Objetivo: Avaliar a eficácia e a aplicabilidade clínica da inteligência artificial (IA) no diagnóstico precoce do câncer bucal em comparação com métodos convencionais. **Material e Métodos:** Esta revisão sistemática foi registrada prospectivamente na base de dados PROSPERO sob o número CRD420251083609. Foi realizada uma revisão sistemática descritiva com base na questão PICO: "Em pacientes com suspeita de câncer bucal, a IA é mais eficaz do que os métodos tradicionais para o diagnóstico precoce?" As buscas foram realizadas entre abril e junho de 2025 nas bases PubMed, LILACS, SciELO, Scopus, Web of Science, Embase e Google Scholar, utilizando os termos "inteligência artificial", "neoplasia bucal" e "diagnóstico precoce". Estudos comparativos que relataram métricas de diagnóstico, como sensibilidade, especificidade e acurácia, foram incluídos. A qualidade metodológica foi avaliada utilizando os sistemas QUADAS-AI e GRADE. **Resultados:** De 1.704 estudos identificados, 12 atenderam aos critérios de elegibilidade. A maioria dos estudos utilizou desenhos observacionais retrospectivos, empregando principalmente redes neurais convolucionais (CNNs) ou modelos híbridos em imagens clínicas e histopatológicas. A acurácia diagnóstica relatada foi geralmente acima de 80%, e alguns modelos leves demonstraram potencial para triagem remota. Limitações comuns incluíram falta de validação externa, heterogeneidade metodológica e dependência da qualidade das imagens. **Conclusão:** A IA apresenta potencial promissor para apoiar o diagnóstico precoce do câncer bucal, melhorando a rapidez e a precisão diagnóstica. A implementação clínica mais ampla exigirá validação multicêntrica, conjuntos de dados padronizados e integração com avaliação clínica e histopatológica.

PALAVRAS-CHAVE

Inteligência Artificial; Diagnóstico auxiliado por computador; Diagnóstico precoce; Neoplasias bucais; Carcinoma de células escamosas de cabeça e pescoço.

INTRODUCTION

Oral cancer represents a significant public health challenge worldwide [1,2]. Global epidemiological data from the Global Burden of Disease indicate a marked increase in mortality from this neoplasm, with a 98.7% rise between 1990 and 2017, from 97,492 to 193,696 deaths [1,2]. In many countries, the average five-year survival rate remains below 50% [2,3]. Notable examples include India, with rates near 37%; Uganda and Egypt, around 20%; and Poland and Thailand, ranging from 36% to 39% [2,3]. These figures reflect challenges in early diagnosis and limited access to effective treatment [2,3]. In Brazil, five-year survival varies by region and health service but often remains below 50% [4]. While referral centers such as A.C. Camargo in São Paulo report survival rates close to 51.7%, other locations, such as Florianópolis, report significantly lower rates, with survival at 33.3% [4]. Factors contributing to these unfavorable outcomes include late diagnosis, unequal access to care, and structural limitations of the public health system [4].

The majority of oral cancer cases are oral squamous cell carcinoma (OSCC), accounting for over 90% of malignant neoplasms in the oral cavity [3,5]. OSCC is often preceded by oral potentially malignant disorders (OPMDs), such as leukoplakia, erythroplakia, and oral lichen planus, which carry variable risks of malignant transformation [2,6,7]. Early detection significantly improves prognosis, with cure rates up to 80% for early-stage diagnoses, compared to 20-30% in advanced stages [2,3]. However, timely diagnosis is often hindered by a shortage of experienced pathologists, interobserver and intraobserver variability in histopathological analysis, limited access to specialized services, and the low sensitivity of visual screenings performed by primary care professionals [5,8-11].

In this context, artificial intelligence (AI) has emerged as a promising tool to support the diagnosis of malignant and potentially malignant oral lesions [12,13]. Deep convolutional neural networks (CNNs), a type of AI inspired by human brain function, have been successfully applied to classification, segmentation, and analysis of both clinical and histopathological images [2,3,9,12]. Recent studies report diagnostic accuracy ranging from 72.1% to 98.6%, depending on the network architecture and dataset [2,8,11]. AI has also been applied to images obtained via smartphones in remote areas and to automated interpretation of digitized histological slides, demonstrating utility across diverse clinical scenarios [7,9,12].

Despite these advances, concerns remain regarding the diagnostic accuracy of AI compared to conventional methods, such as expert clinical examination and standard histopathological analysis [13,14]. Key limitations include variability in training datasets, risk of overfitting, limited generalizability to different populations, and the lack of robust external validation [6,10,11,13,15]. Moreover, many studies still involve small sample sizes or lack standardized methodological designs, which hinders direct comparison between algorithms [3,4,8].

METHODS

The strength of this review lies in its inclusion of the most recent (2023-2025) AI applications for oral cancer detection, the adoption of QUADAS-AI for methodological appraisal, and the comparison of diverse imaging modalities, offering a more holistic assessment of diagnostic potential. This systematic review was prospectively registered in the PROSPERO database under the number CRD420251083609

Research question

The research question was formulated according to the PICO strategy (P: population/patients; I: intervention; C: comparison/control; O: outcome), defined as follows:

- P: Patients with suspected oral cancer
- I: Artificial intelligence (AI)-assisted diagnosis
- C: Traditional diagnostic methods (clinical examination, conventional complementary tests)
- O: Higher diagnostic effectiveness in early detection (accuracy, sensitivity, specificity, and detection time)

The guiding question was:

"In patients with suspected oral cancer, does the use of artificial intelligence result in greater diagnostic accuracy than conventional methods, in terms of sensitivity, specificity, and accuracy?"

Search strategy

A comprehensive search was performed across the following electronic databases: PubMed, SciELO, LILACS, Scopus, Web of Science, Embase, and Google Scholar. The strategy included both indexed and gray literature to minimize publication bias due to non-indexed or unpublished studies.

Keywords were selected based on Medical Subject Headings (MeSH), using the National Library of Medicine database. Boolean operators ("AND", "OR"), quotation marks, and parentheses were applied to enhance search precision and specificity.

To align the strategy with the PICO framework, the following descriptors were used:

- P (Population): "Mouth Neoplasms", "Early Detection of Cancer", "Diagnosis", "Precancerous Conditions"
- I (Intervention): "Artificial Intelligence", "Machine Learning", "Deep Learning", "Computer-Assisted Diagnosis"
- C (Comparison): "Clinical Examination", "Clinical Image"
- O (Outcome): "Early Diagnosis", "Sensitivity and Specificity", "Diagnostic Accuracy"

In Embase, the strategy was refined using controlled vocabulary from the *Emtree* thesaurus, including the terms "oral cancer," "early diagnosis," and "artificial intelligence," as well as their relevant variations and synonyms, to maximize retrieval sensitivity and specificity.

Study selection and eligibility criteria

Titles and abstracts were screened independently and in duplicate by ten reviewers using the Rayyan QCRI platform [16], which supports blinded screening, tagging of inclusion and exclusion decisions, and conflict resolution. Discrepancies were resolved through discussion and consensus after reviewing the full text when necessary.

Eligible studies met the following inclusion criteria:

- Application of AI in the early diagnosis of oral cancer;
- Comparison with conventional diagnostic methods;
- Reporting of quantitative diagnostic metrics (example: sensitivity, specificity, or accuracy).

The selection process adhered to the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The detailed inclusion and exclusion criteria are summarized in Table I. Although the primary objective of this review was to evaluate AI in early oral cancer detection, the inclusion criteria also encompassed studies on precursor oral lesions (OPMDs), as these entities are biologically and clinically linked to malignant transformation and are routinely part of early diagnostic pathways.

Studies that lacked clinical evaluation, or those that did not compare AI performance with standard diagnostic approaches were excluded.

Methodological quality assessment

The methodological quality of the included studies was evaluated using the QUADAS-AI tool, complemented by the GRADE system for assessing the certainty of evidence (Tables II and III).

The QUADAS-AI framework extends the original QUADAS-2 domains to include AI-specific aspects such as dataset representativeness, model training and validation protocols, algorithm transparency, and reproducibility (Table II). The GRADE approach was applied to classify the overall certainty of the evidence based on five domains: risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Table I - Inclusion and exclusion criteria

Criteria	Inclusion	Exclusion
Study type	Full-text articles, randomized controlled trials, clinical trials, meta-analyses	Narrative reviews, letters, opinion pieces, monographs, case reports
Topic	AI studies focused on the diagnosis of oral lesions	AI applied to other anatomical regions
Comparator	Comparison with gold standard (clinical examination and/or biopsy)	Studies without a comparator group
Diagnostic measures	Studies reporting sensitivity, specificity, or other diagnostic metrics	Studies that do not present diagnostic measures
Population	Studies conducted on human subjects	Studies involving animal models or in vitro experiments

Table II - QUADAS-AI assessment of included studies

Study	Patient Selection	Index Test (AI Model)	Reference Standard	Flow and Timing	Applicability Concerns	Overall Risk of Bias
Ahmad et al. [11]	Low – large multicenter dataset with manual annotation	Low – external validation; hybrid deep network	Low – biopsy/histopath confirmation	Low – complete, consistent data	Low – broad representativeness	Low
Alabi et al. [17]	Low – clinical dataset well characterized	Low – supervised ML with internal validation	Low – recurrence confirmed histologically	Low – consistent timing	Low	Low
Alhazmi et al. [18]	Moderate – small single-center cohort	Moderate – ANN, no external validation	Low – biopsy reference	Moderate – incomplete follow-up	Moderate	Moderate
Bashir et al. [3]	Moderate – non-consecutive OED sampling	Moderate – limited validation; small dataset	Low – histopath verified	Moderate – missing data	Moderate	Moderate
Fati et al. [13]	Moderate – limited public dataset	Low – hybrid model (deep+handcrafted features)	Low – histopath gold standard	Moderate – class imbalance	Moderate	Moderate
Jubair et al. [5]	Moderate – heterogeneous sample	Low – CNN architecture, internal validation	Low – expert consensus + histopath	Moderate – partial data loss	Moderate	Moderate
Liyanage et al. [2]	Moderate – geographically restricted	Moderate – CNNs (MobileNet, EfficientNet) no external test	High – unclear reference verification	Moderate	High	High
Pruthi et al. [6]	Moderate – small cohort; unclear inclusion	Moderate – ML on miRNA, limited reproducibility	Low – biopsy gold standard	Moderate	Moderate	Moderate
Shephard et al. [10]	Low – multicenter design, external validation	Low – Transformer model with calibration	Low – histopath confirmed	Low	Low	Low
Talwar et al. [7]	Low – large smartphone image dataset	Low – CNN and Transformer models validated	Moderate – visual reference only	Moderate	Moderate	Low–Moderate
Warin et al. [8]	Low – 2-center data collection	Low – CNN with independent validation	Low – biopsy confirmed	Low	Low	Low
Wuttisarnwattana et al. [12]	Moderate – regional dataset	Moderate – CNN segmentation model	Low – biopsy standard	Moderate	Moderate	Moderate

Table III - GRADE assessment of the certainty of evidence

Study	Risk of Bias	Inconsistency	Indirectness	Imprecision	Publication Bias	Certainty of Evidence
Ahmad et al. [11]	Low	Not serious	Not serious	Serious	Undetected	Moderate
Alabi et al. [17]	Low	Not serious	Not serious	Not serious	Undetected	High
Alhazmi et al. [18]	Moderate	Not serious	Serious	Serious	Undetected	Low
Bashir et al. [3]	Moderate	Serious	Not serious	Serious	Possible	Low
Fati et al. [13]	Moderate	Not serious	Not serious	Serious	Undetected	Low
Jubair et al. [5]	Moderate	Not serious	Serious	Serious	Undetected	Low
Liyanage et al. [2]	Moderate	Serious	Not serious	Serious	Possible	Low
Pruthi et al. [6]	Moderate	Serious	Not serious	Serious	Undetected	Low
Shephard et al. [10]	Low	Not serious	Not serious	Not serious	Undetected	High
Talwar et al. [7]	Low	Not serious	Not serious	Serious	Undetected	Moderate
Warin et al. [8]	Low	Not serious	Not serious	Serious	Undetected	Moderate
Wuttisarnwattana et al. [12]	Moderate	Not serious	Serious	Serious	Undetected	Low

To ensure accessibility and color-perception neutrality across all figures, the Coblis – Color Blindness Simulator was used to evaluate whether contrasts, heatmaps, and categorical colors remained distinguishable for individuals with common forms of color vision deficiency (protanopia, deuteranopia, and tritanopia). This step followed current recommendations for inclusive scientific visualization.

RESULTS

The PRISMA flow diagram is presented in Figure 1. A total of 1,704 records were retrieved from seven databases. After removing 32 duplicates, 1,672 studies were screened by title and abstract, resulting in 1,637 exclusions for not meeting the thematic criteria. Thirty-five full-text articles were assessed for eligibility, and 23 were excluded.

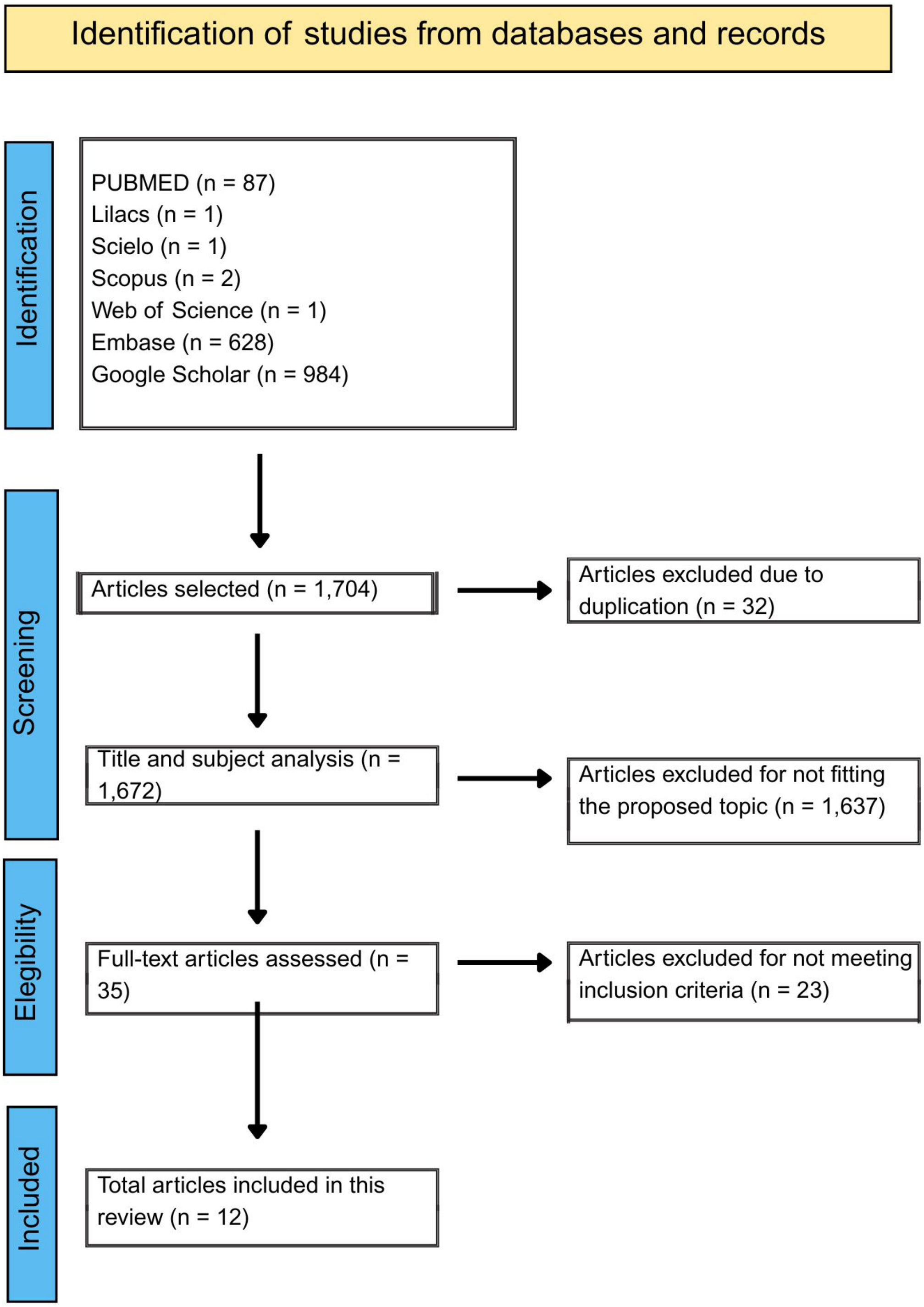

Figure 1 - Prisma Flowchart

Finally, twelve studies [3,5-8,10,11,13,17,18] fulfilled the eligibility criteria and were included in the qualitative synthesis. Table I presents the characteristics of the included studies, AI models, and datasets.

The included studies primarily employed retrospective designs to evaluate AI performance in diagnosing oral squamous cell carcinoma (OSCC) or oral epithelial dysplasia (OED). Sample sizes ranged widely, from 30 to 5,192 images/patients, with diverse modalities including clinical photographs, histopathological slides, and microRNA profiles. Convolutional neural networks (CNNs) were the most commonly used architectures, applied in eight of the twelve studies [3,5,7,8,10,11,13,17], while hybrid CNN models and Transformer-based models demonstrated the highest diagnostic performance. Artificial neural networks (ANNs) and traditional machine learning models generally showed lower accuracy and limited generalizability. Externally validated models, such as ODYN and DenseNet201, exhibited higher generalizability (AUC = 0.88-0.94), whereas single-center models often reported inflated accuracies without reproducibility.

Performance metrics varied across studies. Hybrid CNN models achieved accuracies between 94-99% [11,13], while Transformer-based models showed superior predictive performance for malignant transformation [10]. Smartphone-based models demonstrated diagnostic accuracy above 85%, highlighting feasibility for remote screening in underserved populations [7,12]. In contrast, ANN and traditional ML models achieved lower sensitivity and specificity [17,18]. Table IV summarizes the sample, AI model, and outcomes of each study.

Table IV - Studies

Author	Sample	Result	Outcome
Liyanage et al. [2]	342 oral lesion images with histopathologic diagnosis. Three groups: Neoplasms, Benign, Premalignant	MobileNetV3 and EfficientNetV2: accuracy 76% and 75%, AUC 0.88	Poor performance in lesion detection, but limited accuracy for benign lesions.
Bashir et al. [3]	37 cases of oral epithelial dysplasia (OED), 50% evaluated for carcinoma.	IDaRS achieved an area under the receiver operating characteristic curve (AUROC) of 0.78 and an F1-score of 0.69.	AI is superior to histopathology to assess progression risk.
Shephard et al. [10]	463 whole slide images (WSI): 358 oral epithelial dysplasia (OED), 105 controls; 105 with external centers.	F1-score 0.96; AUROC 0.93 (internal), 0.71 (transformation); C-index 0.63	ODYN is comparable to pathologists in predicting malignant transformation.
Jubair et al. [5]	716 oral lesion images from 543 patients (480 benign, 236 suspected SCC/dysplasia).	AI classification accuracy 85%, sensitivity 86.7%, specificity 84.5%, area under the curve (AUC) 0.928	Practical and comparable alternative to specialists for early screening.
Wuttisarnwattana et al. [12]	2591 oral lesion images via smartphones (normal, OPMDs, OSCC).	DeepLab v3+ with ResNet-50: accuracy = 91%.	Demonstrated high accuracy in rural populations using mobile phones.
Alabi et al. [17]	311 OSCC clinical images (254 training, 57 testing).	Boosted Decision Tree: accuracy 81%, F1-score = 0.85; AUC 0.83; better than PFI	Better than invasion depth in predicting recurrence.
Alhazmi et al. [18]	73 cancer patients (51 malignant, 22 benign).	Artificial neural network (ANN): sensitivity 85.71%, specificity 60%, precision 78.95%	Predicts risk based on age, gender, and clinical characteristics.
Pruthi et al. [6]	30 OSCC samples (30 SCC, 30 non-cancer) + 54 OPMDs for validation.	Accuracy 0.91, F1-score = 0.88; Precision 0.92; Recall 0.91	Prediction based on microRNAs without visible histological changes.
Talwar et al. [7]	1,252 smartphone images	DenseNet201: F1-score = 0.86–0.91; Transformer: F1-score 0.88	Mobile phones help screen in underserved settings.
Ahmad et al. [11]	5192 histopathological images (2494 malignant, 2698 benign, manually annotated).	DenseNet201 + GLCM + HOG + LBP: accuracy 94.1%	Reduced workload for pathologists and improved precision.
Fati et al. [13]	1400 images (300 OSCC, 300 OPMDs, 800 normal).	AlexNet + DWT + LBP + FCH + GLCM: AUC 0.98	High performance with hybrid methods, ideal for low-resource settings.
Warin et al. [8]	980 images (365 OSCC, 315 OPMDs, 300 normal), collected from 2 Thai universities.	DenseNet-169: AUC 1.00 (OSCC), 0.9 (OPMDs); Faster R-CNN: AUC 0.88	CNN models outperform clinical experts, suitable for early diagnosis and efficient triage.

The risk of bias, assessed using QUADAS-AI (Table II), ranged from low to moderate for most studies. Shephard et al. [10], Ahmad et al. [11], Warin et al. [8], Alabi et al. [17] and Talwar et al. [7] were classified as low risk, whereas Bashir et al. [3], Fati et al. [13], Jubair et al. [5], Pruthi et al. [6] and Alhazmi et al. [18] showed moderate risk, mainly due to limited sample sizes, lack of external validation, or incomplete reporting. Liyanage et al. [2] was the only study rated as high risk due to unclear reference verification and limited generalizability.

Reporting adherence according to STARD-AI criteria (Table V and Figure 2) varied from 50% to 90%. Shephard et al. [10] achieved the highest adherence by implementing a complete end-to-end diagnostic pipeline, whereas Wuttisarnwattana et al. [12] demonstrated the lowest adherence due to limited methodological transparency. While most studies reported accuracy, AUC, sensitivity, and specificity, few included explainability analyses such as Grad-CAM.

The certainty of evidence, evaluated using GRADE (Table III), ranged from low to high. High certainty was achieved by Shephard et al. [10] and Alabi et al. [17], supported by robust validation and consistent results. Moderate certainty was assigned to Ahmad et al. [11], Talwar et al. [7], and Warin et al. [8]. The remaining studies [2,3,5,6,13,18] were rated as low due to methodological limitations, indirectness, or imprecision related to small sample sizes or incomplete reporting. Overall, the evidence suggests that AI is a promising tool for early oral cancer diagnosis, but broader clinical implementation requires multicenter validation, standardized datasets, and integration with clinical evaluation.

DISCUSSION

Unlike previous systematic reviews, this study incorporates research published up to mid-2025, including both histopathological and mobile-based diagnostic applications, and applies the QUADAS-AI framework. This represents the first structured methodological assessment of AI-based diagnostic tools specifically in oral oncology. Although methodological heterogeneity and the absence of quantitative synthesis limit the strength of conclusions, the absence of language restrictions reduces the risk of language bias. By providing a comprehensive evaluation of study quality and reporting, these findings can guide future research toward more robust, reproducible, and clinically translatable AI applications.

Variations in accuracy across studies can be attributed to differences in dataset size, heterogeneity of image acquisition (clinical vs. histopathological), annotation quality, preprocessing strategies, and the presence or absence of external validation. Lightweight CNNs trained on smartphone images often show lower sensitivity for benign lesions due to visual noise, while hybrid or Transformer-based architectures trained on large histopathological datasets tend to achieve higher accuracy due to richer feature representation and more stable patterns. Additionally, imbalanced datasets, common in OSCC research, can inflate performance metrics when minority malignant classes are underrepresented.

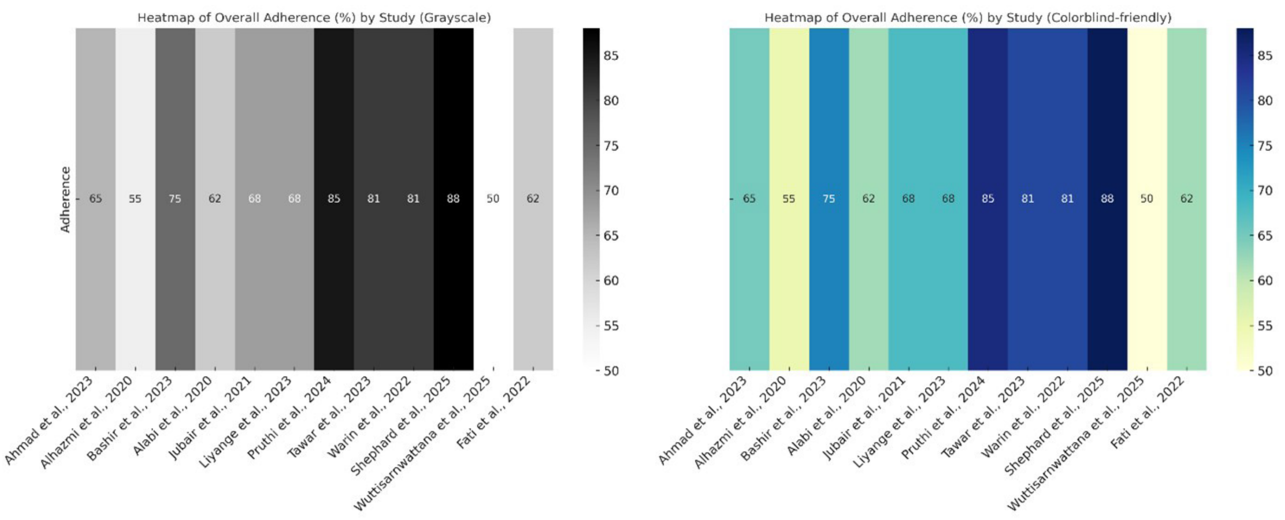

Figure 2 - Heatmap of overall adherence

Table V - STARD-AI

Study	Dataset Description & Availability	AI Model Reporting	Training / Validation	Performance Metrics	Explainability / Calibration	Code / Data Availability	Clinical Context & Deployment	Overall Adherence (%)
Ahmad et al. [11]	Multicenter histopathology (n = 5,192); manual annotations; publicly available	CNNs: Xception, InceptionV3, InceptionResNetV2, NASNetLarge, DenseNet201; architectures described	External validation performed; training details reported; data augmentation applied	Accuracy 90.47–94.44%; AUC 89.4–94.7%; Sensitivity 75.67–87.52%; Specificity 95.65–98.87%; F1-score 82.34–88.88%	Partial (feature maps visualized)	Yes	Diagnostic workflow discussed; preprocessing and ROI enhancement detailed	65
Alhazmi et al. [18]	Small single-center cohort; 73 patients	ANN, 1 hidden layer; inputs/outputs described	10,000 iterations; 10-fold CV	Sensitivity 85.7%; Specificity 60%; Accuracy 78.9%	Not reported	Not shared; available upon request	Screening application not fully discussed	~55
Bashir et al. [3]	163 WSIs (137 OED, 50 transformed); not publicly available	IDaRS pipeline described	Stratified 5-fold CV repeated 3x; parameters reported	AUROC 0.78 ± 0.07; F1 0.69 ± 0.05	Heatmaps & immune-cell maps reported	Not shared; available upon request	Translational context noted	~75
Alabi et al. [17]	311 OSCC patients; multicenter; 254 train/valid; 59 external test	SVM, NB, BDT, DF described; PFI applied	5-fold CV, 50:50 split, SMOTE; external test	Accuracy, specificity, F1, harmonic mean; BDT best	PFI only; no calibration	Not shared	Clinical context described	60–65
Jubair et al. [5]	716 clinical tongue images; not publicly available	Lightweight CNN (EfficientNet-B0), VGG19, ResNet101; 5-layer architecture detailed	Training 566 (79%), Validation 50 (7%), Test 100 (14%); weighted cross-entropy loss, early stopping	Accuracy, Sensitivity, Specificity, ROC/AUC	Not explicitly discussed	Implemented in Keras; code not shared	Early detection of oral lesions; adjunctive diagnostic	65–70
Liyanage et al. [2]	342 oral lesion images; not publicly available	MobileNetV3, EfficientNetV2; preprocessing & class merging described	Training 80%, validation/testing 20%; hardware reported	EfficientNetV2 test 75%, MobileNetV3 test 76%; F1 0.62–0.64; AUC 0.88	Standard metrics; no additional explainability	Code not mentioned; demo app	Classification of oral lesions; adjunctive tool	65–70
Pruthi et al. [6]	30 OSCC + 30 HNE for development; 54 OPMD unseen; FFPE; available upon request	Logistic regression on 6 miRNA features; hyperparameter tuning reported	70:30 split; 10,000 iterations; applied on unseen OPMD	Accuracy 0.859–0.894; ROC AUC 0.864–0.898; Precision 0.892; Recall 0.912	Coefficients interpreted; confusion matrices & ROC	Upon request	Adjunctive use for OSCC/OPMD risk stratification	~85
Talwar et al. [7]	2,178 intraoral smartphone images; not publicly available	CNN & Transformer (VGG19, DenseNet, Swin); detailed hyperparameters	Train 1,344, Val 412, Test 422; 5-fold CV; independent external test	Precision, Recall, Specificity, F1-score, AUC; 95% CI	Grad-CAM heatmaps; no calibration	Code/models on GitHub; data upon request	Early detection of OPMDs; point-of-care	~81
Warin et al. [8]	980 oral photographic images; retrospective; ethics approved; data restricted	DenseNet-169, ResNet-101, SqueezeNet, Swin-S; Faster R-CNN, YOLOv5, RetinaNet, CenterNet2	5-fold CV; batch size, learning rate, epochs, optimizer reported	Classification & detection metrics; compared vs 20 clinicians	Grad-CAM heatmaps; no calibration curves	Code partially available; data restricted	Early detection OSCC/OPMD; benchmarked against clinicians	81
Shephard et al. [10]	Multi-centric retrospective WSIs; internal 358 OED + 105 controls; external 105 OED WSIs; not publicly available	Segmentation: Trans-UNet; Epithelium/nuclei: HoVer-Net+; Classification: REpith; Prognosis: MLP	Internal: train/test split; external validation; 5-fold CV repeated for ODYN-score	Segmentation F1 0.71–0.81; Classification F1 0.96; ODYN-score AUROC 0.71–0.73	Patch-level nuclear features; top 10 patches analyzed; REpith ratio correlated	Data restricted; Supplementary figures available	Predict OED vs non-dysplastic; risk of malignant transformation; clinical decision support	85–90
Wuttisarnwattana et al. [12]	2,591 photographic images; normal, OPMD, OSCC; ethically approved; not publicly available	DeepLab v3+ with ResNet-50; rationale mentioned	Train/validation split not reported	Mean accuracy 87.57%; lesion boundary precision qualitatively described	Model outlines lesion boundaries; no formal calibration	Code & dataset not reported	Aid clinicians in lesion type & boundary identification	~50
Fati et al. [13]	5,192 histopathological images (100× magnification); 2,494 normal, 2,698 malignant; dataset described as public, but no direct link provided	Hybrid AI models: (1) AlexNet + ResNet-18 features + SVM; (2) AlexNet + ResNet-18 + handcrafted features (DWT, LBP, FCH, GLCM) + ANN; preprocessing and fusion methods described	Train/val/test split (~80/20); augmentation (rotation, flipping, shifting); PCA for dimensionality reduction; external validation not performed	Hybrid ANN model: Accuracy 99.1%, Specificity 99.61%, Sensitivity 99.5%, Precision 99.71%, AUC 99.52	Not reported (no saliency maps, no calibration curves)	Data described as "public," but dataset source not clearly specified; code not shared	Framed as support for histopathology-based OSCC diagnosis; no prospective deployment or clinical validation	~60–65

Artificial intelligence (AI) is increasingly recognized as a valuable tool for early detection of oral cancer, addressing limitations of conventional diagnostics such as specialist scarcity, subjectivity, and restricted access in remote areas [13-21]. Several studies have demonstrated that AI can enhance screening accuracy and prognostic assessment of malignant and potentially malignant lesions [13,19]. CNN models, including EfficientNetV2 and MobileNetV3, achieved accuracy above 80% for malignant and premalignant lesions, while the IDaRS histological model reached an AUROC of 0.78 [2,3]. By analyzing large datasets, AI systems can detect complex patterns with consistency, minimizing interobserver variability [2,3] and expediting the diagnostic process [13,14]. These advantages are crucial for early detection, which is directly linked to improved treatment outcomes [4,18].

Nonetheless, several limitations persist. Model performance depends heavily on image quality, and many algorithms display reduced sensitivity for benign lesions or require multicenter validation (13-14). The Transformer-based ODYN model, for example, achieved an F1-score of 0.96 and outperformed the WHO grading system for oral epithelial dysplasia [10]. Yet, its reliance on manual annotations indicates that subjectivity remains a challenge (20). The lack of publicly annotated databases for OSCC also restricts model robustness and reproducibility [22,23].

Recent advances in technological approaches for oral lesion assessment have shown promising potential for improving early diagnosis and prognosis. Computational models and deep learning architectures have been applied to oral tongue lesions, enabling accurate classification and risk prediction [23,24]. Other imaging techniques have also been explored for detecting oral malignancies. Ex vivo fluorescence confocal microscopy has shown high sensitivity and specificity compared with histopathology [25].

Building on these foundations, mobile-based AI applications have achieved high diagnostic accuracy (>85%) [7], representing a significant step toward accessible screening in resource-limited settings. Lightweight convolutional neural networks, such as EfficientNet-B0, have reached expert-level performance in analyzing tongue lesion images [5], while DeepLab v3+ combined with ResNet-50 has demonstrated 87.57% accuracy in lesion segmentation tasks [12,26].

Despite these encouraging results, limitations such as lack of clinical validation and imbalanced datasets continue to affect the generalizability of these models, highlighting the importance of integrating AI tools into standardized diagnostic workflows and multicenter validation studies [27-29].

Beyond imaging, boosted decision trees (BDT) have surpassed conventional histopathological parameters, such as depth of invasion, in predicting recurrence of early tongue squamous cell carcinoma [16]. Yet, classical histopathological variables remain essential [18,22].

ANN models based on clinical and behavioral data reached sensitivity up to 85% but suffered from limited interpretability [9]. AI-driven analysis of microRNA profiles showed 89.4% accuracy [6], demonstrating potential for molecular-level prediction, though small sample sizes and invasive sampling restrict applicability [20,24]. In contrast, smartphone-based intraoral imaging, even when performed by non-specialists, produced encouraging F1-scores (~0.86), suggesting utility for population-level screening [7].

Hybrid AI models combining CNNs with classical feature extraction techniques such as FCH, DWT, LBP, and GLCM achieved some of the best performances. Fati et al. [13] reported 99.1% accuracy, while Ahmad et al. [11] obtained 97% accuracy, 90.9% sensitivity, and 98.9% specificity using DenseNet201 and Xception architectures. However, their use of controlled datasets limits direct clinical translation [11,13]. Real-world performance is influenced by variable image quality and lack of standardized acquisition [23].

Artificial intelligence (AI) represents a valuable complementary tool for early oral cancer diagnosis and can help expand access to care [16,27,28]. Broader clinical implementation will require investment in digital infrastructure, standardized public databases [23,24], multicenter validation, and integration of clinical, histological, and molecular data [6,20]. Hybrid models that combine medical expertise with AI's analytical capabilities appear to offer the most promising route for clinical translation [5,11,19].

Despite encouraging results, heterogeneity in datasets, annotation protocols, and validation methods remain considerable. Only three studies (Shephard et al., Ahmad et al., Alabi et al.) included multicenter or external validation, which is essential for ensuring generalizability.

Inconsistent reporting of metrics (AUC, F1-score, accuracy) further complicates comparison and underscores the need for standardized frameworks such as STARD-AI and QUADAS-AI. None of the included studies reported integration into clinical workflows or prospective validation, indicating that AI currently serves best as an adjunctive diagnostic aid rather than a replacement for conventional methods.

The STARD-AI assessment, together with visualization procedures such as heatmap generation using the Coblis color-blindness simulator [30,31], highlighted substantial variability in reporting quality. While most studies detailed datasets and model architecture, few provided public access to data or code, limiting reproducibility. Validation strategies were inconsistent, with some studies using cross-validation or external datasets and others providing minimal information on data partitioning. Accuracy, AUC, sensitivity, and specificity were frequently reported, but explainability analyses were rare. Only one study implemented a fully integrated diagnostic pipeline, emphasizing that most AI tools remain experimental and not yet deployed in clinical practice. These findings reinforce the need for standardized reporting to enhance transparency, support replication, and facilitate safe clinical application of AI in oral oncology.

Early diagnosis plays a critical role in improving the prognosis of patients with oral cancer. Research by Bandeira et al. [32] highlights that a substantial proportion of individuals receive their diagnosis at advanced stages, which often complicates treatment and diminishes survival outcomes. This underscores the significant potential of artificial intelligence (AI) in facilitating timely detection, thereby enabling more favorable prognoses. AI applications have already demonstrated promising results in related areas, such as the Oral ID device, a handheld tool that utilizes autofluorescence technology to delineate safe surgical margins in patients with oral squamous cell carcinoma [33]. Comparative analyses indicate that using Oral ID improves margin assessment accuracy, surpassing the 52.2% achieved by the traditional 1 cm margin method [33].

Future research should prioritize prospective multicenter trials, open-access annotated datasets, and consistent use of AI-specific quality frameworks. Integration of multimodal data clinical, histopathological, and molecular may further enhance predictive accuracy and facilitate clinical translation.

CONCLUSION

Artificial intelligence shows strong potential as a complementary tool for early oral cancer diagnosis, with performance comparable to or exceeding conventional methods. Convolutional neural networks and hybrid models facilitate efficient analysis of clinical and histopathological images, improving access in resource-limited settings. Implementation challenges remain, including limited external validation, dataset heterogeneity, and generalizability. Careful integration with standard clinical and histopathological evaluation, along with standardized methodologies, well-annotated public datasets, and multicenter validation, is essential to establish AI as a reliable and clinically applicable diagnostic tool.

Acknowledgements

The authors would like to thank the staff and colleagues at the Universidade Federal de Juiz de Fora (UFJF) for their support and valuable contributions during the preparation of this manuscript.

Data availability

The data is already available within the text.

Author’s Contributions

LAA, LSSM, CHR, LCC, HSM, BAF, LGR: Conceptualization, Investigation, Writing – Original Draft Preparation. QSP: Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing. MCFRD: Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing. GMCF: Formal Analysis, Supervision, Validation, Writing – Review & Editing.

Conflict of Interest

The authors have no conflicts of interest to declare.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Regulatory Statement

This review was conducted in accordance with the PRISMA 2020 guidelines for systematic reviews and meta-analyses, as recommended by the EQUATOR Network.

The protocol for this systematic review was prospectively registered in the PROSPERO database under registration number CRD420251083609.

Disclosure

All data extraction forms, analytic code, and datasets used in this review are available upon request from the corresponding author.

Ethics statement

Ethical approval and informed consent were obtained for all primary studies included in this review where reported. If such information was not reported in the original study, it was noted accordingly.

List of abbreviations

OSCC – Oral Squamous Cell Carcinoma

OPMDs – Oral Potentially Malignant Disorders

CNN – Convolutional Neural Network

ANN – Artificial Neural Network

WSI – Whole Slide Image

AUC – Area Under the Curve

AUROC – Area Under the Receiver Operating Characteristic Curve

miRNA – microRNA

BDT – Boosted Decision Tree

QUADAS-AI – Quality Assessment of Diagnostic Accuracy Studies – Artificial Intelligence extension

GRADE – Grading of Recommendations, Assessment, Development and Evaluation

OED – Oral Epithelial Dysplasia

REFERENCES

1. Ó SF, Lima JRM, Oliveira SMD, Cruz RAN, Medeiros CGL. Oral cancer and public health. Braz J Implant Health Sci. 2024;4(8):1-7.
2. Liyanage V, Tao M, Park JS, Wang KN, Azimi S. Malignant and premalignant oral lesions classification and diagnosis with deep neural networks. J Dent. 2023;137:104657. https://doi.org/10.1016/j.jdent.2023.104657. PMid:37574105.
3. Bashir RMS, Shephard AJ, Mahmood H, Azarmehr N, Raza SEA, Khurram SA, et al. A digital score of peri-epithelial lymphocytic activity predicts malignant transformation in oral epithelial dysplasia. J Pathol. 2023;260(4):431-42. https://doi.org/10.1002/path.6094. PMid:37294162.
4. Montoro JRMC, Hicz HA, Souza L, Livingstone D, Melo DH, Tiveron RC, et al. Prognostic factors in squamous cell carcinoma of the oral cavity. Braz J Otorhinolaryngol. 2008;74(6):861-6. https://doi.org/10.1016/S1808-8694(15)30146-4. PMid:19582342.
5. 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(4):1123-30. https://doi.org/10.1111/odi.13825. PMid:33636041.
6. Pruthi N, Singh R, Kumar A, Sharma P, Gupta M, Yadav S. Applying machine learning for enhanced microRNA analysis: a companion risk tool for oral squamous cell carcinoma in standard care incisional biopsy. Biomolecules. 2024;14(4):458. https://doi.org/10.3390/biom14040458. PMid:38672474.
7. Talwar V, Singh P, Mukhia N, Vinod PK, Rathore S, Singh A, et al. AI-assisted screening of oral potentially malignant disorders using smartphone-based photographic images. Cancers. 2023;15(16):4120. https://doi.org/10.3390/cancers15164120. PMid:37627148.
8. 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. 2022;17(8):e0273508. https://doi.org/10.1371/journal.pone.0273508. PMid:36001628.
9. Chaudhary N, Rai A, Rao AM, Faizan MI, Augustine J, Chaurasia A, et al. High-resolution AI image dataset for diagnosing oral submucous fibrosis and squamous cell carcinoma. Sci Data. 2024;11(1):1050. https://doi.org/10.1038/s41597-024-03836-6. PMid:39333529.
10. Shephard AJ, Mahmood H, Raza SEA, Araújo ALD, Santos Silva AR, Lopes MA, et al. Development and validation of an artificial intelligence-based pipeline for predicting oral epithelial dysplasia malignant transformation. Commun Med. 2025;5(1):186. https://doi.org/10.1038/s43856-025-00873-z. PMid:40394272.
11. Ahmad M, Irfan MA, Sadique U, Haq IU, Jan A, Khattak MI, et al. Multi-method analysis of histopathological images for early diagnosis of oral squamous cell carcinoma using deep learning and hybrid techniques. Cancers (Basel). 2023;15(21):5247. https://doi.org/10.3390/cancers15215247. PMid:37958422.
12. Wuttisarnwattana P, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P, Vicharueang S, et al. Precise identification of oral cancer lesions using artificial intelligence. Stud Health Technol Inform. 2024;316:1096-7. https://doi.org/10.3233/SHTI240601. PMid:39176572.
13. Fati SM, Senan EM, Javed Y. Early diagnosis of oral squamous cell carcinoma based on histopathological images using deep and hybrid learning approaches. Diagnostics. 2022;12(8):1899. https://doi.org/10.3390/diagnostics12081899. PMid:36010249.
14. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25(1):30-6. https://doi.org/10.1038/s41591-018-0307-0. PMid:30617336.
15. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the QUADAS-AI extension. Nat Med. 2021;27(2):185-7.
16. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan: a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210. https://doi.org/10.1186/s13643-016-0384-4. PMid:27919275.
17. Alabi RO, Elmusrati M, Sawazaki Calone I, Kowalski LP, Haglund C, Coletta RD, et al. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. Int J Med Inform. 2020;136:104068. https://doi.org/10.1016/j.ijmedinf.2019.104068. PMid:31923822.

18. Alhazmi FH, Alhazmi Y, Makrami A, Masmali A, Salawi N, Masmali K, et al. Application of artificial intelligence and machine learning for prediction of oral cancer risk. J Oral Pathol Med. 2021;50(5):444-50. https://doi.org/10.1111/jop.13157. PMid:33394536.
19. Kaur S, Mupparapu M. Artificial intelligence in oral cancer diagnosis and prognosis: current trends and future directions. Oral Oncol Rep. 2023 In press.
20. Marque L, Souza FN, Silva LM, Silva BMR, Oliveira FB, Santos LCC. Artificial intelligence in oral cancer diagnosis and prognosis: current perspectives. J Oral Pathol Med. 2024;53(2):123-35. https://doi.org/10.1111/jop.13467.
21. El-Naggar AK. The 2017 WHO classification of head and neck tumours. Int J Head Neck Surg. 2017;48(6):30-7.
22. Dimitriou N, Arandjelović O, Caie PD. Deep learning for whole-slide image analysis: an overview. Front Med. 2019;6:264. https://doi.org/10.3389/fmed.2019.00264. PMid:31824952.
23. Sengupta S. Challenges in AI-based diagnosis of oral cancer: data, validation, and deployment. Diagnostics. 2022;12(8):1903.
24. Bello IO, Soini Y, Salo T. Prognostic evaluation of oral tongue cancer: means, markers and perspectives (II). Oral Oncol. 2010;46(9):636-43. https://doi.org/10.1016/j.oraloncology.2010.06.008. PMid:20637679.
25. Shavlokhova V, Sandhu S, Flechtenmacher C, Koveshazi I, Neumeier F, Padrón-Laso V, et al. MobileNet-based detection of oral squamous cell carcinoma using ex vivo confocal images. J Clin Med. 2021;10(22):5326. https://doi.org/10.3390/jcm10225326.
26. Wuttisarnwattana P, Limprasert W, Suebnukarn S, Jinaporntham S, Jantana P, Vicharueang S, et al. DeepLab v3+ for segmentation of oral potentially malignant and malignant lesions in photographic images. J Imaging. 2024;10(1):45.
27. Mahmood H. Use of artificial intelligence for the diagnosis of oral cancer: a systematic review. Oral Oncol. 2020;106:104741.
28. Pruthi P. A machine learning model based on miRNA expression for early detection of oral squamous cell carcinoma. J Oral Pathol Med. 2024;53(1):65-72.
29. Rahman MM. A review of artificial intelligence applications in oral cancer diagnosis and prognosis. Front Oncol. 2023;13:1057421.
30. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI extension. Nat Med. 2021;27(2):185-7.
31. Coblis. Color Blindness Simulator [Internet]. 2025 [cited 2025 Oct 13]. Available from: https://www.color-blindness.com/coblis-color-blindness-simulator/
32. Bandeira CM, Almeida AA, Carta CFL, Almeida JD, Kaminagakura E. Como melhorar o diagnóstico precoce do câncer bucal? Braz Dent Sci. 2017;20(4):25-31. https://doi.org/10.14295/bds.2017.v20i4.1439.
33. Mohamed Y, Mohammed SA, Sulaiman AM. Oral ID® as an adjunctive tool for surgical margin assessment in patients with oral squamous cell carcinoma: a comparative study: Oral ID® as an Adjunctive Tool for Surgical Margin Assessment. Braz Dent Sci. 2022;24(4, Suppl):e2868. https://doi.org/10.4322/bds.2021.e2868.

Maria Clara Fernandes Ribeiro Dantas
(Corresponding address)
Universidade Federal de Juiz de Fora, Faculdade de Odontologia, Departamento de Odontopediatria e Ortodontia, Juiz de Fora, MG, Brazil.
Email: clarafdantas@gmail.com

Editor-in-chief: Sergio Eduardo de Paiva Gonçalves
Editor: Renata Falchete do Prado

Date submitted: 2025 Oct 13
Accept submission: 2026 Feb 04