REVIEW ARTICLE

https://doi.org/10.47102/annals-acadmedsg.2026200

Efficacy of AI in the diagnosis of oral squamous cell carcinoma from clinical images: A systematic review and meta-analysis

Yu Shin Chan*[1] *MRCS*, Keerthana Thangaraja*[2] *MBBS*, Owen Tsung Wen Ho[1] *MRCS*, Brigitte Tok[2] *MBBS*, Clarissa Cheong[1] *MRCS*, Joy Au[1] *MRCS*, Serena Lo[3] *PhD*, Ming Yann Lim[1] *FAMS*, Hao Li[1] *FAMS*, Ernest Weizhong Fu[1] *FAMS*, Yijin Jereme Gan[1] *FAMS*

ABSTRACT

Introduction: Oral squamous cell carcinoma (OSCC) is a common malignancy that is frequently diagnosed at advanced stages, partly because of limitations in conventional diagnostic approaches. This systematic review aimed to evaluate the diagnostic performance of artificial intelligence (AI) models in detecting OSCC from clinical images.

Methods: PubMed, ScienceDirect, the Cochrane Library, and Embase were searched from January 2000 to September 2024. Studies assessing the diagnostic accuracy of AI models for OSCC using clinical images were included. Studies that did not involve clinical images, were not specific to OSCC, or were not published in English were excluded. In accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines, study selection was per-formed independently by 4 reviewers, with disagreements resolved through consensus. Subsequently, diagnostic accuracy data were extracted, risk of bias was assessed using Quality Assessment of Diagnostic Accuracy Studies using AI and thematic analysis was performed by 2 reviewers. Where appropriate, a random effects meta-analysis was conducted.

Results: Nine studies comprising 47,987 oral clinical images were included, with 6 studies in the meta-analysis. AI models demonstrated a pooled sensitivity of 0.94 (95% confidence interval [CI] 0.87–0.97), specificity of 0.95 (95% CI 0.86–0.98), and an area under the curve of 0.975 (95% bootstrap CI 0.911–0.986). AI systems generally achieved higher diagnostic accuracy than general practitioners and performance comparable to specialists under controlled study conditions.

Conclusion: AI models show strong diagnostic perfor-mance for OSCC detection from clinical images. However, current evidence is largely derived from retrospective and heterogeneous datasets without external validation. Prospective and blinded studies are required to deter-mine their role as decision-support tools in real-world settings.

Keywords: artificial intelligence, cancer, oral squamous cell carcinoma, otorhinolaryngology, public health

CLINICAL IMPACT

What is New

- AI demonstrates high diagnostic performance for OSCC detection in clinical images, with a pooled sensitivity of 0.94, specificity of 0.95, and strong discriminatory performance across 6 studies included in the meta-analysis.
- This review identifies the current evidence gap between promising diagnostic accuracy and clinical readiness, highlighting limited external validation, heterogeneous retrospective datasets, and concerns regarding independent data partitioning.

Clinical Implications

- AI could serve as an adjunctive screening and decision-support tool, particularly in primary care and resource-limited settings where access to specialist assessment is limited.
- Clinical implementation requires prospective, externally validated studies before routine use, with evaluation of real-world diagnostic performance, clinician adoption, and potential effects on referral and biopsy rates.

INTRODUCTION

Oral squamous cell carcinoma (OSCC) is the most prevalent and life-threatening malignancy of the oral cavity, representing approximately 90% of all oral malignancies.[1] Lip and oral cavity cancers accounted for an estimated 452,205 new cases and 194,108 deaths worldwide in 2024. These estimates encompass cancers of the lip and oral cavity and should not be interpreted as OSCC-specific figures.[2] Despite signifi-cant advancements in treatment modalities—such as surgery, radiotherapy, and chemotherapy—the 5-year survival rate for OSCC remains relatively low,

The Annals is an open access journal, allowing non-commercial use under CC BY-NC-SA 4.0.

* Joint first authors
[1] Department of Otorhinolaryngology, Tan Tock Seng Hospital, Singapore
[2] Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[3] Clinical Research and Innovation Office, Tan Tock Seng Hospital, Singapore
Correspondence: Dr Yu Shin Chan, Department of Otorhinolaryngology, Head and Neck Surgery, Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore 308433. Email: chanyushin@gmail.com
Accepted: 28 October 2026
Published Online First: 10 November 2026

particularly when the disease is diagnosed at advanced stages.[3] Early detection is therefore crucial, as patients diagnosed in the initial stages have significantly better prognoses and quality-of-life outcomes than those diagnosed later. However, current diagnostic approaches are often limited by subjective interpretation, inter-observer variability, and limited access to specialist care, especially in low-resource settings.[4]

The diagnosis of OSCC typically begins with clinical suspicion followed by histological diagnosis via tissue sampling. Clinically, it usually presents as a non-healing ulcer, an exophytic mass, or a mixed red and white patch (erythroleukoplakia), frequently affecting the lateral tongue and the floor of the mouth.[5] Imaging plays a crucial role in staging, with contrast-enhanced computed tomography (CT) assessing tumour extent and bone invasion. Magnetic resonance imaging (MRI) offers superior soft tissue resolution, especially for perineural and marrow involvement.[6] 18F-fluorodeoxyglucose positron emission tomography (PET)/CT is useful for detecting nodal and distant metastases.[7] Histologically, OSCC is characterised by invasive nests or cords of atypical squamous cells with varying degrees of keratinisation, and is graded by degree of differentiation.[8]

Recent developments in artificial intelligence (AI), particularly in machine learning and deep learning, have shown considerable promise in medical image analysis and pattern recognition, offering the potential to augment diagnostic accuracy and efficiency in OSCC detection.[9] AI models trained on histopathological images, clinical images, and radiographic data have demonstrated high sensitivity and specificity in distinguishing malignant from benign lesions.[10] Given the growing body of literature exploring the role of AI in OSCC detection, a systematic evaluation of the diagnostic accuracy of these models is both timely and necessary. Although AI has been used for various stages of OSCC investigation, this systematic review and meta-analysis aimed to critically appraise and synthesise existing evidence on the efficacy of AI-based models in diagnosing OSCC from clinical images.

METHODS

This review is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Guided by the Population, Intervention, Comparison, Outcome and Study Design (PICOS) elements for the inclusion criteria, the primary research question was: "How efficacious is AI in the diagnosis of oral squamous cell carcinoma from clinical images?" The review protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO; CRD420251084762) in 2025, after the literature searches were conducted in September–October 2024. The registration was therefore retrospective in relation to the search process. No subsequent amendments were made to the registered protocol.

Search strategy and selection criteria

PubMed, ScienceDirect, the Cochrane Library, and Embase were searched to retrieve articles that evaluated the diagnostic accuracy of AI models in the diagnosis of OSCC based on clinical images.

The search combined the Boolean operators AND and OR with the search terms "deep learning", "artificial intelligence", "artificial learning", "machine learning", "machine intelligence", "neural networks", "probabilistic networks", "knowledge representation", "statistical learning", "Bayesian learning", "medical decision", "decision tool", "support tool" AND "oral cavity carcinoma", "oral cavity cancer", "oropharyngeal cancer", "oropharyngeal carcinoma", "oral cavity squamous cell carcinoma".

The searches were confined to articles published from January 2000 to September 2024. All articles were published in English or included an English translation. The search was carried out between September and October 2024. The PICOS search strategy is summarised in Supplementary Table S1.

The research team consisted of 4 reviewers who independently reviewed the titles and abstracts identified by each database to determine the final list of included articles. Disagreements were resolved through consensus discussion among the reviewers. Supplementary Table S2 summarises the studies included in the systematic review and indicates those excluded from the meta-analysis (with reasons). Although the terms "oropharyngeal cancer" and "oropharyngeal carcinoma" were included in the search strategy to increase sensitivity, only studies evaluating lesions arising within the oral cavity were eligible for inclusion.

Risk of bias and applicability were assessed independently by 2 reviewers using the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI) framework, an extension of Quality Assessment of Diagnostic Accuracy Studies-2 developed specifically for AI-based diagnostic accuracy studies. As shown in Supplementary Table S3, each included study was evaluated across 4 core domains: patient selection, index test, reference standard, and flow and timing—with each rated for both risk of bias and applicability concerns. In addition, 3 AI-specific domains were assessed: independence of data partitioning (whether training and test datasets were drawn from separate, non-overlapping sources); model transparency (whether sufficient architectural and methodological detail was reported to enable replication); and integrity of the training/test split

(whether the test set remained fully independent from model development). Each domain was rated as low, high, or unclear risk of bias, based on prespecified signalling questions. Disagreements were resolved by consensus.

Only studies with histologically confirmed datasets were included in the meta-analysis. Supplementary Table S4 summarises the pooled models included in the meta-analysis. Supplementary Tables S5–S10 present the 2×2 data from each of the included models.

Data extraction

A standardised Excel spreadsheet with the following headings was used to extract data from all included articles.

1. Study characteristics
    a. Author
    b. Year of publication
    c. Country
    d. Sample size
    e. Study design
2. Methods of diagnosing OSCC
3. Data that the machine learning model is trained on
4. AI model characteristics
5. Diagnostic performance metrics
    a. Accuracy
    b. Precision
    c. Sensitivity
    d. Specificity
    e. Positive predictive value
    f. Negative predictive value
    g. Area under the curve (AUC) value
    h. F1

Statistical analysis

Statistical analyses were performed using R in RStudio (version 2022.12.0+353; RStudio Team, Vienna, Austria), primarily with the mada package. Individual-study sensitivity and specificity were calculated from the extracted 2×2 contingency tables, with 95% confidence intervals (CIs) estimated using the Clopper–Pearson exact binomial method. Pooled sensitivity and specificity with 95% CIs were estimated using a bivariate random-effects model (Reitsma model), fitted by restricted maximum likelihood. A summary receiver operating characteristic (SROC) curve was constructed using the Rutter–Gatsonis parameterisation, with the AUC calculated using the mada package. The AUC 95% CI was estimated using 2000 parametric bootstrap iterations.

Between-study heterogeneity was evaluated using the model's variance components, correlation, and Zhou–Dendukuri I^2 statistic. Following the benchmarks established by Higgins et al.,[11] I^2 values of 25%, 50%, and 75% were tentatively considered to represent low, moderate, and high inconsistency, respectively. These ranges were interpreted cautiously because they were not specifically developed for the Zhou–Dendukuri estimator and may be unreliable with only 6 studies.

Publication bias was explored using an effective-sample-size-weighted Deeks-type funnel plot asymmetry regression, with the log diagnostic odds ratio regressed against the inverse square root of the effective sample size. A continuity correction of 0.5 was applied to all 4 cells of the 2×2 contingency table for studies containing zero cells. Sensitivity analyses were performed using alternative continuity corrections of 0.1 and 1.0 to assess the robustness of the regression results.

RESULTS

The search from the 4 databases retrieved a total of 627 articles. After removing 114 duplicates, 513 articles remained. After reviewing the titles and abstracts, 103 articles were retrieved in full text, and 1 additional article was included from the citations. After reviewing these, 9 diagnostic articles focusing on clinical images were included in the systematic review, and 6 in the meta-analysis. The results of the search are summarised in the PRISMA flow diagram (Supplementary Fig. S1).

Characteristics of included studies

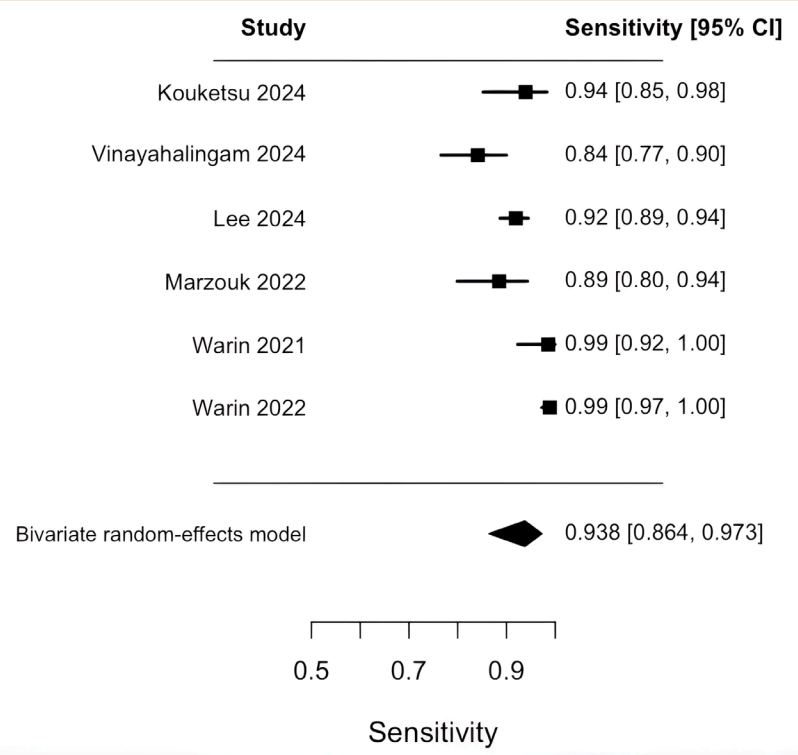
Nine eligible studies were included in the systematic review, with a total of 47,987 clinical images, of which at least 1784 were of OSCC. Six studies were included in the meta-analysis, comprising a total of 8825 clinical images. Most of the studies were from Asia, and most used convolutional neural networks (CNNs). Supplementary Table S2 summarises the characteristics of the included studies in the systematic review.

Diagnostic test accuracy

The pooled sensitivity of AI models was 0.938 (95% CI 0.864–0.973), while the pooled specificity was 0.9543 (95% CI 0.859–0.986), indicating high diagnostic accuracy with moderate consistency across studies. The pooled sensitivity and specificity are shown in the forest plots in Figs. 1 and 2. The SROC curve yielded an AUC of 0.975 (95% bootstrap CI 0.911–0.986) (Fig. 3), indicating high discriminatory performance under controlled study conditions.

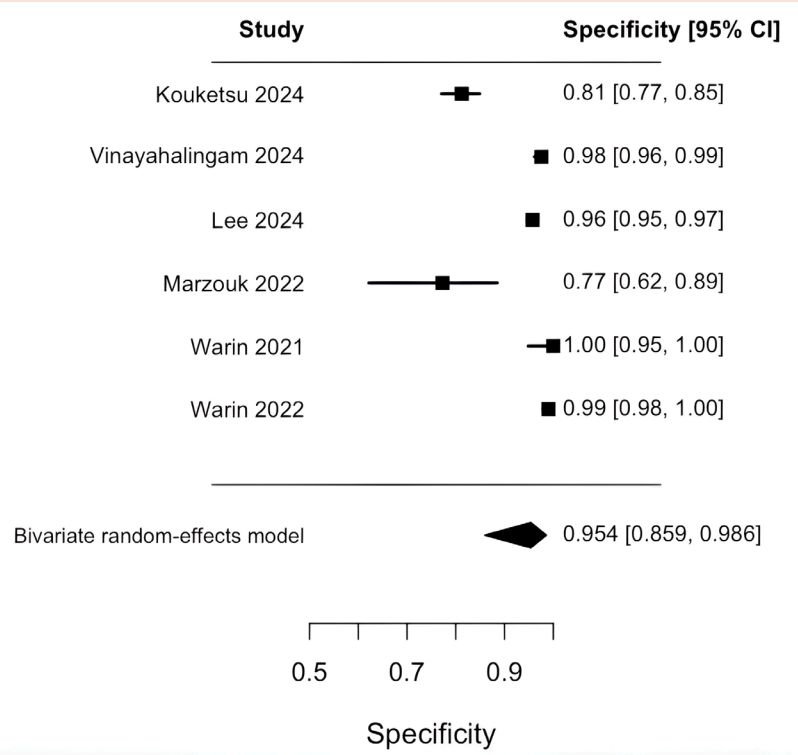
Between-study heterogeneity was observed, with a Zhou–Dendukuri I^2 of 56.7%. The estimated between-

Fig. 1. Forest plot of sensitivity.



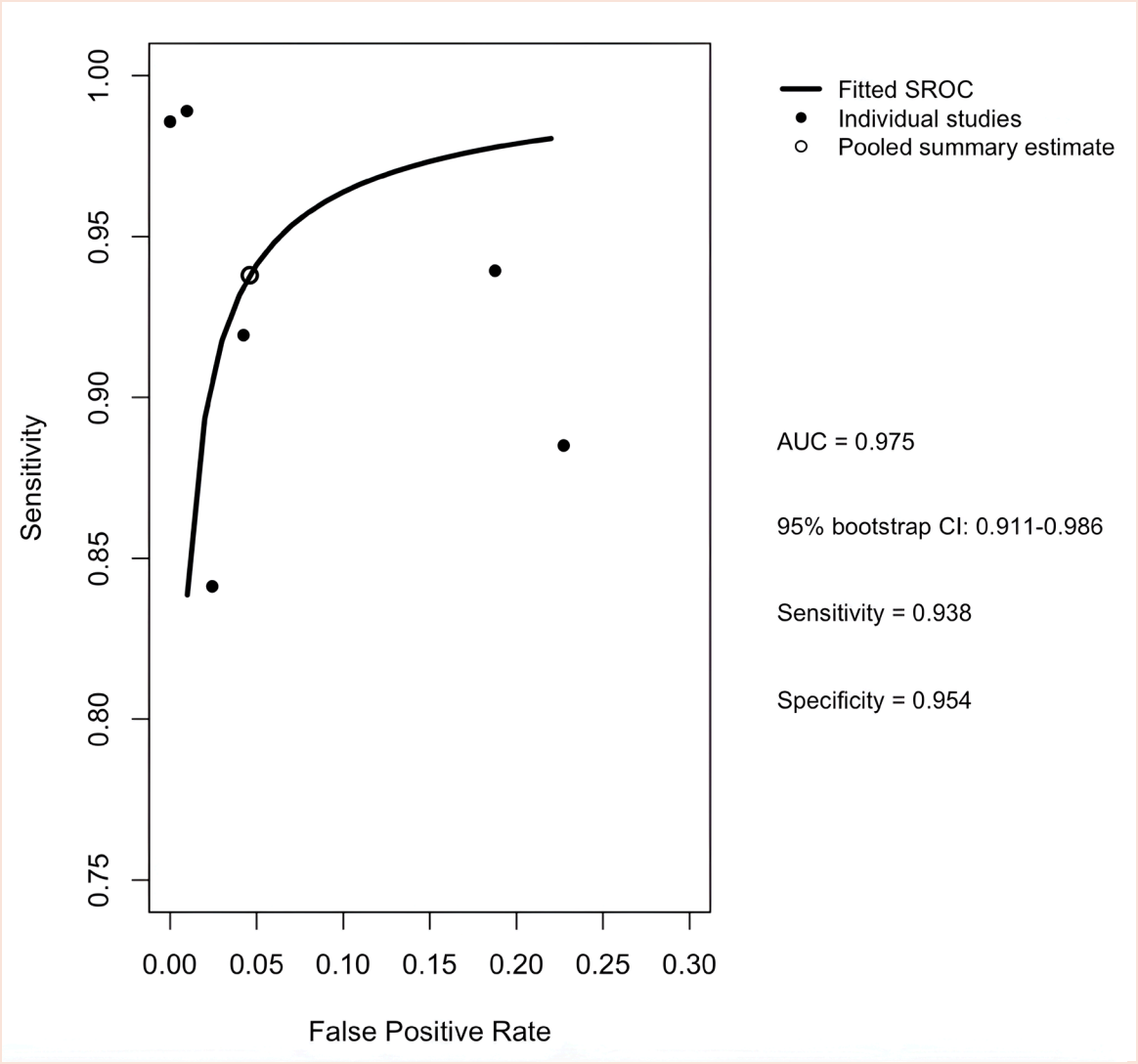
CI: confidence interval

Fig. 2. Forest plot of specificity.



CI: confidence interval

Fig. 3. Summary receiver operating characteristic curve for AI-based diagnosis of oral squamous cell carcinoma.



AI: artificial intelligence; AUC: area under the curve; CI: confidence interval; OSCC: Oral squamous cell carcinoma; SROC: summary receiver operating characteristic curve

The thick solid line represents the fitted SROC curve, the filled circles represent individual studies, and the open circle represents the pooled summary estimate. The AUC was 0.975 (95% bootstrap CI 0.911–0.986).

study variance was greater for logit false-positive rate (2.093) than for logit sensitivity (0.973), with a negative correlation between these parameters (r=-0.588). The differing variance components and negative correlation suggest heterogeneity in the relationship between sensitivity and false-positive rate, indicating that heterogeneity is primarily driven by variations in overall diagnostic accuracy across studies rather than standard threshold effects.

Risk of bias assessment

Risk of bias assessment using the QUADAS-AI framework is presented in Supplementary Table S3. The index test and reference standard domains were rated as low risk across all 9 studies. Patient selection was rated high risk in 2 studies—Marzouk et al. (2022) and Birur et al. (2022)—because of unclear case ascertainment methods, and unclear in 4 others. Among the AI-specific domains, data partitioning independence and training/test split integrity were the most frequent sources of concern, rated high risk in 2 studies and unclear in 5, reflecting limited reporting of dataset separation practices in the included literature. Model transparency was rated low risk in 4 studies and unclear in the remainder.

Publication bias

Potential publication bias was explored using a Deeks-type funnel plot asymmetry analysis (Supplementary Fig. S2). The primary analysis, using a continuity

correction of 0.5, showed no statistically significant funnel plot asymmetry (P=0.603). Sensitivity analyses using alternative continuity corrections of 0.1 and 1.0 yielded similarly non-significant results (P=0.780 and P=0.516, respectively), indicating that the findings were robust to the choice of continuity correction. However, given the small number of studies included in the meta-analysis (n=6), the analysis had limited statistical power, and publication bias cannot be excluded.

Performance of AI diagnostic models

The studies reviewed used a variety of algorithms whose accuracy varied with algorithm architecture, data preprocessing, and training strategies.

Nine studies were included in the systematic review, of which 6 provided eligible data for the quantitative meta-analysis. The meta-analysis comprised Kouketsu et al. (2024),[12] Vinayahalingam et al. (2024), Lee et al. (2024), Marzouk et al. (2022), Warin et al. (2021)[13] and Warin et al. (2022). These studies contributed paired diagnostic accuracy data from which sensitivity and specificity were estimated.

The remaining 3 studies were included in the qualitative synthesis but not in the meta-analysis because the classification was not histologically determined. Their findings are discussed separately and were not used to calculate the pooled diagnostic accuracy estimates or the SROC curve.

Across the included literature, reported diagnostic performance varied according to model architecture, preprocessing methods, dataset characteristics, and validation procedures. The following discussion distinguishes studies that contribute to the quantitative synthesis from those that provide additional qualitative evidence.

The observed between-study heterogeneity warrants consideration when interpreting these findings. The greater variance in logit false-positive rate compared with logit sensitivity suggests that false-positive rates varied more substantially across studies. The negative correlation between logit sensitivity and logit false-positive rate likely reflects differences in overall model accuracy, driven by study-specific characteristics such as AI model architecture, dataset composition, and validation procedures. These findings highlight the importance of standardised validation protocols and assessment of diagnostic performance across clinically relevant thresholds.

The included studies show that preprocessing and data augmentation were associated with improvements in diagnostic performance, although the statistical significance of these improvements was not reported. Various forms of augmentation, such as random oversampling or weight balancing, help address class imbalance and improve the efficiency of model training. Zhang et al. (2024) reported that integrating such preprocessing steps increased the model's accuracy from 95.11% to 98.34% and precision from 96.14% to 98.54%.[14] However, no P value or CI was recorded. Lee et al. (2024) introduced an innovative augmentation approach termed mosaic-based soft labelling, which achieved 94.4% accuracy, outperforming conventional over-sampling (91.1%), and weight balancing (92.8%) methods.[15] Enhancements such as normalisation and contrast adjustment help reduce inconsistencies caused by lighting and image quality, supporting the development of more generalisable models.[16] By introducing variations, these techniques improve algorithms by training them on a dataset more representative of real-world clinical circumstances. Although beneficial, these preprocessing techniques can be resource-intensive, requiring computational power and technical expertise, and may introduce data bias and limit scalability.[17]

Warin et al. (2022) conducted a comparative evaluation of several CNN models for OSCC classification, revealing that DenseNet-169 and ResNet-101 outperformed other CNNs and even oral and maxillofacial specialists in diagnostic accuracy.[18] However, because of their large architectures, these models require high-performance computing, limiting their suitability for mobile applications. In contrast, lightweight models such as SqueezeNet showed lower performance (AUC 0.71–0.88) but offered a more feasible solution for mobile deployment.

Zhang et al. (2024) explored gated recurrent units (GRUs), a type of recurrent neural network (RNN), which are less commonly used for image classification than CNNs. Traditional RNNs process sequential data by feeding the output of 1 time step as input to the next, creating a "loop".[19] However, they struggle with long-term dependencies because the gradient can vanish or explode as it propagates through many time steps. GRUs address this issue via gating mechanisms that selectively filter information, thereby improving training efficiency.[20,21] Zhang et al.'s proposed gated recurrent unit network/Northern Goshawk optimisation algorithm (GRU/NGO-LC), which incorporates Lévy flight and chaos theory, significantly outperformed CNN models such as ResNet-101 and R-CNN. It achieved accuracy, precision, and sensitivity of 98.34%, 98.54%, and 98.60%, respectively, compared with ResNet-101's 91.95%, 93.02%, and 86.71%.

The GRU/NGO-LC also surpassed the standard GRU and GRU/NGO (91.27% and 90.54% accuracy) and outperformed 5 other modern optimisation methods.[11] However, whether these differences are statistically significant cannot be determined, as neither CIs nor confusion matrix were provided.

The study suggests that the GRU/NGO-LC outperforms both the standard GRU method and other CNN models for OSCC classification. Despite its performance, the study did not assess the algorithm's computational efficiency or memory burden.

Although CNNs were the most common algorithms in these studies, some studies adopted hybrid models incorporating multiple architectures together with preprocessing and augmentation strategies. Hybrid models included in this review demonstrated higher diagnostic accuracy than conventional deep learning models, but were heterogeneous and difficult to interpret. Marzouk et al. (2022) introduced AIDTL-OCCM, a hybrid CNN incorporating contrast enhancement, DenseNet-169 feature extraction, and chimp optimisation algorithm. It achieved 90.08% accuracy, 88.6% recall, and 89.1% precision, outperforming 5 other highly optimised CNNs. The next best performing model, Inception-V4, had 85.64% accuracy, 85.7% recall, and 87.8% precision.[22] This suggests that AIDTL-OCCM has potential for the classification of OSCC in clinics. However, the study did not detail the model's processing speed or computational demands compared with other conventional models, limiting clinical applicability. Vinayahalingam et al. (2024) evaluated a hybrid combining Mask R-CNN and a Swin Transformer for multiclass lesion classification, including OSCC, leukoplakia, and oral lichen planus. This model leveraged spatial feature extraction from CNNs and transformer-based global context analysis, achieving 95.5% accuracy (CI 93.8–97.2) and 84.3% sensitivity (CI 80.4–88.2).[23] This hybrid architecture enhances pattern detection and improves the integration of information from different regions of an image. Although promising, it lacked comparison with other CNN/transformer-based models.

DISCUSSION

OSCC remains the most common and deadly malignancy of the oral cavity, with early detection playing a pivotal role in improving survival outcomes. Conventional diagnosis relies on clinical examination, histopathology, and imaging modalities such as CT, MRI, and PET/CT for staging; however, these approaches are constrained by subjectivity, cost, and limited accessibility in resource-constrained settings.

Recent advances in AI, particularly deep learning, have demonstrated growing potential to support OSCC diagnosis. This systematic review and meta-analysis demonstrates, as a proof of concept, that AI models trained on clinical images can achieve high diagnostic performance, with a pooled sensitivity of 0.94, specificity of 0.95, and AUC of 0.975. These findings suggest strong discriminatory performance across the included retrospective datasets. AI offers several theoretical advantages over traditional approaches, including objective pattern recognition, rapid image analysis, and the possibility of deployment in settings where specialist expertise or advanced imaging is unavailable. However, current evidence primarily reflects feasibility rather than readiness for routine clinical implementation.

AI versus healthcare professionals

Across the included studies, AI models demonstrated variable diagnostic performance relative to general practitioners and specialist clinicians. Some models achieved higher accuracy than clinician comparators under controlled study conditions, whereas others performed similarly or less well. These comparisons should be interpreted cautiously because of differences in study populations, model architectures, datasets, and validation methods. In Warin et al. (2022), large CNN architectures such as DenseNet-169 and ResNet-101 achieved accuracies of 98.98% (CI 98.13–99.51) and 93.27% (CI 91.51–94.75), res-pectively, compared with 89.59% (CI 87.29–91.25) for spe-cialists.[15] Although these results suggest that AI systems can achieve expert-level performance in retro-spective image-based evaluations, such large models require substantial computational resources, limiting their practicality outside tertiary care environments.

Smaller, lightweight models such as SqueezeNet and Swin-S demonstrated lower diagnostic accuracy—84.59% (CI 82.2–86.8) and 79.18% (CI 76.5–81.69%), respectively—but offer greater feasibility for mobile and point-of-care applications. SqueezeNet achieved an accuracy of 84.59%, marginally exceeding the 83.27% (CI 80.78–85.55)[18] reported for GPs, whereas Swin-S achieved a slightly lower accuracy. However, Birur et al. reported that MobileNet and VGG19 achieved higher sensitivities (82% and 87%) and accuracies (79% and 87%) than GPs, who demonstrated a sensitivity of 60% and an accuracy of 70%.[24] These findings suggest that diagnostic performance varies across lightweight architectures, and that computational efficiency should be evaluated alongside accuracy. Overall, these findings support the feasibility of AI as an assistive screening tool in primary care.

Potential clinical applications

AI holds promise as a decision-support tool that may reduce diagnostic error and assist clinicians, particularly in resource-limited settings.[22,23] Lightweight models such as MobileNet enable near real-time lesion assessment from smartphone images, while larger models may complement diagnostic workflows in specialist centres.[24] AI-based tools could also lower barriers by facilitating preliminary self-screening and prompting earlier clinical consultation. AI models with high diagnostic sensitivity may help

identify suspicious lesions that warrant further clinical assessment, reducing the likelihood of missed OSCC. Models with high specificity may help reduce false-positive assessments and unnecessary referrals or biopsies. The practical effects of these approaches require evaluation in prospective clinical studies.

Limitations

This review has several important limitations. Only 6 studies were eligible for meta-analysis, with 3 additional studies included qualitatively, limiting statistical power and generalisability. Substantial heterogeneity in datasets, preprocessing techniques, and model architectures hindered direct comparison across studies. Most models were CNN-based, introducing potential architectural bias, and many lacked external validation. Diagnostic overlap between early OSCC and oral potentially malignant disorders (OPMDs) further complicates model performance and remains a challenge for both AI systems and clinicians. Publication bias could not be excluded, as the small number of studies limited the power to detect funnel plot asymmetry.

Although the pooled AUC of 0.975 (95% bootstrap CI 0.911–0.986) indicates high discriminatory performance, these findings should be interpreted in the context of the predominantly retrospective and controlled study settings. Curated image datasets may not adequately represent the variability encountered in routine clinical practice, including differences in image quality, lighting conditions, lesion presentation, and disease prevalence. Furthermore, models evaluated using clearly distinguishable malignant and nonmalignant lesions may demonstrate inflated diagnostic performance compared with more challenging real-world presentations, particularly early OSCC and oral potentially malignant disorders. The lower bound of the bootstrap CI highlights uncertainty in the pooled AUC estimate but should not be interpreted as a prediction of real-world performance. Prospective external validation using representative clinical populations is therefore neces-sary to establish the generalisability and clinical utility of these models.

Future research directions

Future studies could stratify by oral cavity subsite (e.g. tongue versus buccal mucosa versus floor of mouth). Future studies could also benchmark AI systems against clinicians with clearly defined levels of training and experience. Such analyses may identify anatomical locations or clinical scenarios where AI provides the greatest incremental benefit.

Future studies should also focus on multiclass models capable of differentiating OSCC from benign lesions and diagnostically complex OPMDs. In Vinayahalingam's 2024 study, OSCC was classified with high accuracy using their hybrid deep learning model, whereas the detection of oral lichen planus and leukoplakia was only moderately effective.[23]

Further expansion of lesion segmentation would be particularly useful for hard-to-visualise oropharyngeal/nasopharyngeal regions to aid clinicians in surgical planning, biopsy targeting, or radiotherapy planning.

Other imaging modalities could also benefit from further study. AI is increasingly applied in advanced imaging such as CT, MRI, and histopathology. In CT and MRI, AI can support tumour segmentation, staging, and prediction of treatment response, aiding preoperative planning and personalised therapy. In histopathology, AI models facilitate automated classification, grading, and prognosis, providing reproducible and quantitative insights that assist pathologists.

Multimodal integration of clinical, radiographic, histological, and molecular data may improve staging, personalised treatment, and prognostication by enhancing diagnostic accuracy, reducing interobserver variability, and supporting comprehensive patient management. Transformer-based architectures are particularly suited for such complex data integration.

Finally, in real-world deployment, many high-performing models require substantial computational resources, depend on network availability, and may face language barriers. They should be tested in clinical environments, with future trials focusing on practical outcomes such as diagnostic speed, clinician adoption, and reduction in unnecessary biopsies.

Regulation, reliability, and outlook

Despite promising accuracy in controlled settings, widespread clinical adoption of AI remains constrained by limited external validation, dataset bias, and regulatory challenges. These findings should be interpreted as proof of concept rather than evidence of clinical readiness. Additionally, issues of transparency, accountability, data privacy, and medicolegal responsibility must be addressed. Explainable AI techniques may improve clinician trust by providing interpretable outputs.[28] Ultimately, AI is unlikely to replace clinicians in the near future. Instead, it can be positioned as an adjunct that augments clinical judgment. Although AI excels at processing large datasets and identifying subtle patterns, it lacks the contextual reasoning, empathy, and ethical judgment intrinsic to clinical care. At this point, its optimal role may lie in supporting—not supplanting—the human elements of medicine. However, it is prudent to note that current performance estimates reflect early-stage implementations and are likely to be surpassed as model architectures mature, datasets diversify, and validation extends into real-world clinical environments.

CONCLUSION

As a proof of concept, AI models demonstrate strong potential for OSCC detection from clinical images and can achieve diagnostic performance comparable to clinicians in retrospective settings. AI has the potential to improve early detection and streamline diagnostics. However, translation into routine practice requires rigorous prospective validation, explainability, and assessment within real-world clinical workflows. AI can be a complementary decision-support tool that enhances, rather than replaces, clinical expertise in OSCC diagnosis.

Supplementary materials

Table S1. PICOS search strategy.

Table S2. Summary of characteristics of included studies in the systematic review.

Table S3. Risk of bias and applicability assessment of included studies using the QUADAS-AI framework.

Table S4. Summary of pooled models included in meta-analysis.

Table S5. Kouketsu et al. 2024—2×2 contingency table.

Table S6. Vinayahalingam et al. 2024—2×2 contingency table.

Table S7. Lee et al. 2024—2×2 contingency table.

Table S8. Marzouk et al. 2022—2×2 contingency table.

Table S9. Warin et al. 2021—2×2 contingency table.

Table S10. Warin et al. 2022—2×2 contingency table.

Fig. S1. PRISMA flow diagram.

Fig. S2. Deeks' funnel plot.

Acknowledgements

The authors sincerely thank all the researchers and clinicians whose work and insights contributed to this systematic review and meta-analysis. They also appreciate the support of their institution and colleagues, whose valuable guidance and encouragement have been instrumental throughout this study.

Ethics statement

Not applicable; this was a systematic review with no direct patient involvement and no identifying information.

Declaration

No funding was received for this paper. The authors declare that they have no affiliations or financial involvement with any commercial organisation with a direct financial interest in the subject or materials discussed in the manuscript.

REFERENCES

1. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. Oral Oncol 2009;45:309-16.
2. Ervik M, Laversanne M, Lam F, et al. Global Cancer Observatory: Cancer Today (2024 estimates, version 1.0) [Internet]. Lyon, France: International Agency for Research on Cancer; 2026 [cited 3 November 2026]. https://gco.iarc.who.int/today.
3. Shield KD, Ferlay J, Jemal A, et al. The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. CA Cancer J Clin 2017;67:51-64.
4. Speight PM, Khurram SA, Kujan O. Oral potentially malignant disorders: risk of progression to malignancy. Oral Surg Oral Med Oral Pathol Oral Radiol 2018;125:612-27.
5. Rabinovici-Cohen S, Fridman N, Weinbaum M, et al. From Pixels to Diagnosis: Algorithmic Analysis of Clinical Oral Photos for Early Detection of Oral Squamous Cell Carcinoma. Cancers (Basel) 2024;16:1019.
6. Pałasz P, Adamski Ł, Górska-Chrząstek M, et al. Contemporary Diagnostic Imaging of Oral Squamous Cell Carcinoma - A Review of Literature. Pol J Radiol 2017;82:193-202.
7. Subramaniam N, Poptani H, Andrew S, et al. Imaging Advances in Oral Cavity Cancer and Perspectives from a Population in Need: Consensus from the UK-India Oral Cancer Imaging Group. J Head Neck Physicians Surg 2021;9:4-12.
8. Müller S, Boy SC, Day TA, et al. Data Set for the Reporting of Oral Cavity Carcinomas: Explanations and Recommendations of the Guidelines From the International Collaboration of Cancer Reporting. Arch Pathol Lab Med 2019;143:439-46.
9. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
10. Ariji Y, Fukuda M, Nozawa M, et al. Automatic detection of cervical lymph nodes in patients with oral squamous cell carcinoma using a deep learning technique: a preliminary study. Oral Radiol 2021;37:290-6.
11. Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ 2003;327:557-60.
12. Kouketsu A, Doi C, Tanaka H, et al. Detection of oral cancer and oral potentially malignant disorders using artificial intelligence-based image analysis. Head & Neck 2024;46:2066-76.
13. Warin K, Limprasert W, Suebnukarn S, et al. Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. J Oral Pathol Med 2021;50:911-18.
14. Zhang L, Shi R, Youssefi N. Oral cancer diagnosis based on gated recurrent unit networks optimized by an improved version of Northern Goshawk optimization algorithm. Heliyon 2024; 10:e32077.
15. Lee SJ, Oh HJ, Son YD, et al. Enhancing deep learning classification performance of tongue lesions in imbalanced data: mosaic-based soft labeling with curriculum learning. BMC Oral Health 2024;24:161.
16. Seoni S, Shahini A, Meiburger KM, et al. All you need is data preparation: A systematic review of image harmonization techniques in multi-center/device studies for medical support systems. Comput Methods Programs Biomed 2024;250:108200.
17. Maharana K, Mondal S, Nemade B. A review: Data pre-processing and data augmentation techniques. Glob Transit Proc 2022;3:91-9.

18. Warin K, Limprasert W, Suebnukarn S, et al. AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer. PLoS One 2022;17:e0273508.
19. Li M, Jiang Y, Zhang Y, et al. Medical image analysis using deep learning algorithms. Front Public Health 2023;11:1273253.
20. Das S, Tariq A, Santos T, et al. Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research. In: Colliot O, Editor, Machine Learning for Brain Disorders. New York: Humana; 2023.
21. Yin S, Yin K, Chen W, et al. VisionGRU: A Linear-Complexity RNN Model for Efficient Image Analysis. arXiv preprint arXiv:2412.18178. 23 December 2024.
22. Marzouk R, Alabdulkreem E, Dhahbi S, et al. Deep Transfer Learning Driven Oral Cancer Detection and Classification Model. Comput Mater Contin 2022;73:3905-20.
23. Vinayahalingam S, van Nistelrooij N, Rothweiler R, et al. Advancements in diagnosing oral potentially malignant disorders: leveraging Vision transformers for multi-class detection. Clin Oral Investig 2024;28:364.
24. Birur NP, Song B, Sunny SP, et al. Field validation of deep learning based Point-of-Care device for early detection of oral malignant and potentially malignant disorders. Sci Rep 2022;12:14283.
25. Brown C, Nazeer R, Gibbs A, et al. Breaking bias: The role of artificial intelligence in improving clinical decision-making. Cureus 2023;15:e36415.
26. Hirosawa T, Shimizu T. A narrative review of artificial intelligence in medical diagnostics. Comput Mater Contin 2025;83:3919-44.
27. Song B, Sunny S, Li S, et al. Mobile-based oral cancer classification for point-of-care screening. J Biomed Opt 2021;26:065003.
28. Rosenbacke R, Melhus Å, McKee M, et al. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. JMIR AI 2024;3:e53207.