Dental Research Journal

Review Article

Evaluating the diagnostic accuracy of neural network models in detecting oral potentially malignant disorders and oral cancer using mobile photographs: An umbrella review

P. D. Madan Kumar[1], Rajeshwari Selvam[2], Sasidharan Sivakumar[3], Lavanya C[4], Ranganathan Kannan[4]

Departments of [1]Public Health Dentistry, [2]ICMR Project Research Scientist and [4]Oral and Maxillofacial Pathology, Ragas Dental College and Hospital, Chennai, Tamil Nadu, [3]Discovery Research Division, Indian Council of Medical Research, ICMR Headquarters, New Delhi, India

ABSTRACT

Background: Oral cancer (OC) and oral potentially malignant disorders (OPMDs) remain major global public health challenges, particularly in low-and middle-income countries. Although early detection substantially improves prognosis, limited healthcare infrastructure restricts timely diagnosis. Artificial intelligence (AI) enabled, mobile phone-based diagnostic systems offer a promising, accessible solution, and multiple systematic reviews have demonstrated their potential. However, uncertainty persists regarding the comparative performance of AI models across diverse real-world settings.

Aim: An umbrella review was aimed at evaluating the comparative performance of different AI models in detecting OC and OPMD.

Materials and Methods: This research identified six systematic reviews from databases such as Medline (via PubMed), Web of Science, Scopus, and EMBASE through October 2024 which were checked at the title, abstract, and full-text levels. The risk of bias (ROB) was then assessed using the Joanna Briggs Institute's ROB assessment tool.

Results: Across included reviews, pooled sensitivity and specificity for AI-based detection ranged from 88% to 92%, with reported diagnostic odds ratios ranging from 114 to 2549, indicating strong discriminatory performance. Deep learning architectures such as EfficientNet and ResNet consistently demonstrated high diagnostic accuracy, while hybrid approaches (e.g., MLSO + SVM) showed promising performance in selected analyses. However, substantial heterogeneity was observed across studies (I^2 often >85%), reflecting variability in populations, image acquisition protocols, and model architectures.

Conclusion: Deep learning models like EfficientNet and ResNet are favored in clinical diagnostics for their exceptional performance and adaptability. Hybrid approaches, such as MLSO + SVM, also show great potential by combining the strengths of traditional and modern methods effectively.

Key Words: Artificial intelligence, deep learning, early diagnosis, intraoral photography, mobile health, oral cancer

Received: 01-Apr-2025
Revised: 03-Jan-2026
Accepted: 20-Mar-2026
Published: 15-May-2026

Address for correspondence:
Dr. P. D. Madan Kumar, Department of Public Health Dentistry, Ragas Dental College and Hospital, Chennai - 600 119, Tamil Nadu, India.
E-mail: madankumar21@yahoo.co.in

Access this article online

Website: www.drj.ir
www.drjjournal.net
www.ncbi.nlm.nih.gov/pmc/journals/1480
DOI: 10.4103/drj.drj_172_25

How to cite this article: Kumar PD, Selvam R, Sivakumar S, Lavanya C, Kannan R. Evaluating the diagnostic accuracy of neural network models in detecting oral potentially malignant disorders and oral cancer using mobile photographs: An umbrella review. Dent Res J 2026;23:18.

INTRODUCTION

Oral cancer (OC) and oral potentially malignant disorders (OPMDs) pose significant global health challenges due to high incidence rates, delayed diagnosis, and associated mortality.[1] Early detection of these conditions is vital, as timely intervention greatly improves survival rates and reduces the need for invasive treatments associated with advanced disease stages.[2,3] Access to specialized diagnostics for OC remains limited, especially in low-resource settings.[4] This disparity highlights the need for accessible, cost-effective, and accurate screening solutions that can reach underserved populations and support early intervention efforts. Addressing this challenge requires a leap from traditional methods to innovative, technology-driven solutions that can bridge the accessibility gap.

The rapid development of artificial intelligence (AI) and deep learning has new opportunities to address these diagnostic challenges. In particular, neural network models, with their ability to process large datasets and detect subtle patterns, have shown promise in identifying early signs of malignancy in various medical imaging applications.[5] Studies are being carried out to investigate usage of smartphone images to identify OPMD and OC.[6,7] Such applications could transform screening practices, making early detection accessible to individuals outside of traditional healthcare settings[8,9]. This approach is particularly relevant for remote and underserved areas where mobile phones are widely available, but access to specialized diagnostic care is limited.[10]

Mobile phone-based neural networks, with their capacity to analyze vast datasets and recognize intricate patterns, present an innovative pathway to identify early signs of OPMDs and OCs, aiming to assist healthcare providers and even laypersons in identifying suspicious lesions.[11] However, questions regarding the diagnostic accuracy of these models in real-world situations remain critical. Unlike controlled clinical environments, photographs taken with mobile phones are subject to variations in lighting, angles in which they are taken, image quality, and device specifications, all of which could impact the model's reliability.[12] In addition, the performance of neural network models can vary depending on their architecture, the diversity of the training data, and the presence of confounding factors, such as benign oral conditions that resemble OPMD.[13]

Many published literature has evaluated the diagnostic accuracy of neural network models in detecting OPMD and OC using mobile phone photographs. Systematic reviews that focused on the diagnostic performance of these models using images from various modalities such as computed tomography, histopathological slide images, spectra images, autofluorescence images, and clinical intra-oral photographs are present.[11,14]

Despite the growing interest in AI for healthcare, there remains significant uncertainty regarding the specific AI models that deliver optimal performance across diverse clinical and operational settings. Several systematic reviews have highlighted AI in diagnosing OC and OPMDs that fall short of identifying the most effective models tailored to specific contexts, such as resource-limited environments. This lack of clarity hinders the translation of AI-based solutions from research to real-world applications, particularly in regions where early detection could make the most significant impact.

To bridge this knowledge gap, we performed an umbrella review with the primary objective of systematically evaluating the comparative performance of various AI models in detecting OC and OPMD. By evaluating sophisticated deep learning frameworks such as EfficientNet and ResNet, in combination with hybrid models like MLSO + SVM, our objective is to pinpoint algorithms that deliver exceptional diagnostic precision while maintaining versatility across diverse clinical contexts.

The findings of this study will have broad implications for the integration of AI into public health strategies. By identifying the most efficient models, our study aims to facilitate the implementation of AI-driven diagnostic solutions, ensuring their dependability and suitability for application in resource-limited environments. This is particularly crucial for rural India, where access to conventional healthcare facilities is limited, and innovative, cost-effective diagnostic solutions are urgently needed.

MATERIALS AND METHODS

Review question

This umbrella review was conducted in alignment with PRISMA guidelines to address the following research question: "Which AI models demonstrate the highest diagnostic accuracy and adaptability for detecting OC and OPMD in resource-constrained settings?"

Inclusion and exclusion criteria

The criteria for eligibility were established based on the following PICOS framework:

- (P) Population: Patients undergoing evaluation for OPMD and OC
- (I) Intervention: Implementation of digital tools which uses Neural Networks for the detection of OPMD and OC (including smartphone or AI-integrated smartphone applications)
- (C) Comparison: Application of traditional diagnostic techniques is considered the gold standard (such as visual assessment)
- (O) Outcomes: Diagnostic accuracy (sensitivity, specificity), accuracy, and diagnostic odds ratio (DOR)
- (S) Study design: Systematic reviews of observational studies.

Research studies adhering to the PICO framework's "Subjects, Intervention, Control, Outcome" criteria were deemed eligible for inclusion. Systematic reviews of *ex vivo* cell studies, case reports/series, and clinical trials were specifically excluded.

Type of the study

This umbrella review encompassed systematic reviews of observational studies that analyzed the diagnostic accuracy of various neural networks in identifying OPMD and OC in digital photographs, in contrast to conventional clinical assessments.

Search strategy for article identification

Two independent researchers (L C. and R.S.) performed an electronic search for systematic reviews across several databases, including Medline (via PubMed), Web of Science, Scopus, EMBASE, and Google Scholar, covering publications up to October 2024. Detailed search strategies for each database are presented in Table 1. The search was restricted to English-language publications, with no further limitations imposed.

Screening of articles

To compile relevant articles, duplicate entries were systematically identified and removed based on title, author, and publication year after consolidating results from various databases. Two independent reviewers (L.C. and R.S.) conducted an initial screening of titles and abstracts, excluding studies that did not meet the predefined PICO framework criteria through a consensus-based approach. Any disagreements regarding article selection were resolved via discussion or, if necessary, by consulting a third and fourth reviewer (M.K. and R.K.).

Table 1: Search strategy for the study

Database	Search string	Numbers
Medline (PubMed)	((“artificial intelligence”[MeSH Terms] OR “deep learning”[MeSH Terms] OR “machine learning”[MeSH Terms] OR “neural networks”[MeSH Terms]) AND (“mouth neoplasms”[MeSH Terms] OR “oral cancer”[MeSH Terms] OR “oral squamous cell carcinoma”[MeSH Terms] OR “Head and neck cancer” [MeSH Terms] OR “Precancerous conditions” [MeSH Terms] OR “oral neoplasms” [MeSH Terms] OR “head and neck carcinoma” [MeSH Terms]) AND (“diagnosis, computer assisted”[MeSH Terms]))	125
Web of Science	((((ALL=(“precancerous lesion”)) OR ALL=(“precancerous lesions”)) OR ALL=(“precancerous condition”) OR ALL=(“oral premalignant lesion”) OR ALL=(opmd) OR ALL=(OPMDs) OR ALL=(“oral precancer”) OR ALL=(leukoplakia) OR ALL=(erythroplakia) OR ALL=(“oral submucous fibrosis”))) AND (((ALL=(“mobile photography”)) OR ALL=(“mobile applications”) OR ALL=(“artificial intelligence”) OR ALL=(“mobile application”))) AND TS=(systematic)	9
Scopus	((TITLE-ABS-KEY (“precancerous lesions”) OR TITLE-ABS-KEY (“precancerous lesion”) OR TITLE-ABS-KEY (“precancerous conditions”) OR TITLE-ABS-KEY (precancer) OR TITLE-ABS-KEY (“oral premalignant lesion”) OR TITLE-ABS-KEY (“oral premalignant condition”) OR TITLE-ABS-KEY (opmd) OR TITLE-ABS-KEY (opmds))) AND ((TITLE-ABS-KEY (“mobile photography”) OR TITLE-ABS-KEY (“mobile photograph”) OR TITLE-ABS-KEY (“mobile photographs”) OR TITLE-ABS-KEY (“mobile applications”) OR TITLE-ABS-KEY (mobile AND apps) OR TITLE-ABS-KEY (“mobile application”) OR TITLE-ABS-KEY (“ai application”)))	20
Embase	(“oral potentially malignant disorder”/exp OR “oral lesion, precancerous” OR “oral lesion, premalignant” OR “oral potentially malignant disease” OR “oral potentially malignant disorder” OR “oral pre-malignant lesion” OR “oral precancerous condition” OR “oral precancerous lesion” OR “oral premalignant lesion” OR “potentially malignant oral disease” OR “potentially malignant oral disorder” OR “precancerous oral lesion” OR “precarcinoma, mouth” OR “premalignant oral lesion” OR “mouth cancer”/exp OR “cancer, mouth” OR “intraoral cancer” OR “mouth cancer” OR “mouth mucosa cancer” OR “oral cancer” OR “oral cavity cancer”) AND (“mobile application”/exp OR “mobile app” OR “mobile application” OR “mobile applications” OR “mobile apps” OR “portable software app” OR “portable software application” OR “portable software applications” OR “portable software apps” OR “tablet application”) AND (“systematic review”/exp OR “review, systematic” OR “systematic review”)	7
Google scholar	(“Artificial Intelligence” OR “Deep Learning” OR “Machine Learning” OR “Neural Networks”) AND (“Oral Cancer” OR “Oral Squamous Cell Carcinoma” OR “Mouth Neoplasms”) AND (“Systematic review” OR “Systematic review and meta-analysis”)	156

Following this preliminary filtering, each reviewer undertook a rigorous evaluation of the full texts of the remaining articles, with any residual discrepancies resolved through further dialogue [Figure 1].

Data extraction

For studies that satisfied the inclusion criteria, data extraction was performed by two reviewers (L.C. and R.S.), adhering to the Joanna Briggs Institute (JBI) guidelines for umbrella reviews.[15] The extraction focused on several parameters: (a) author and year of publication, (b) type of study, (c) research setting, (d) range of years covered, (e) number of studies reviewed, (f) dataset size, (g) sources searched, (h) Neural network models used, (i) Comparison (j) tools used for appraisal, (k) outcomes measured, (l) results and findings, (m) significance of the results, and (n) heterogeneity of the studies. These details have been organized and presented in Tables 2 and 3.

Risk of bias

The risk of bias (ROB) for each study was independently evaluated by two reviewers (L.C and R.S) using the JBI systematic review critical appraisal

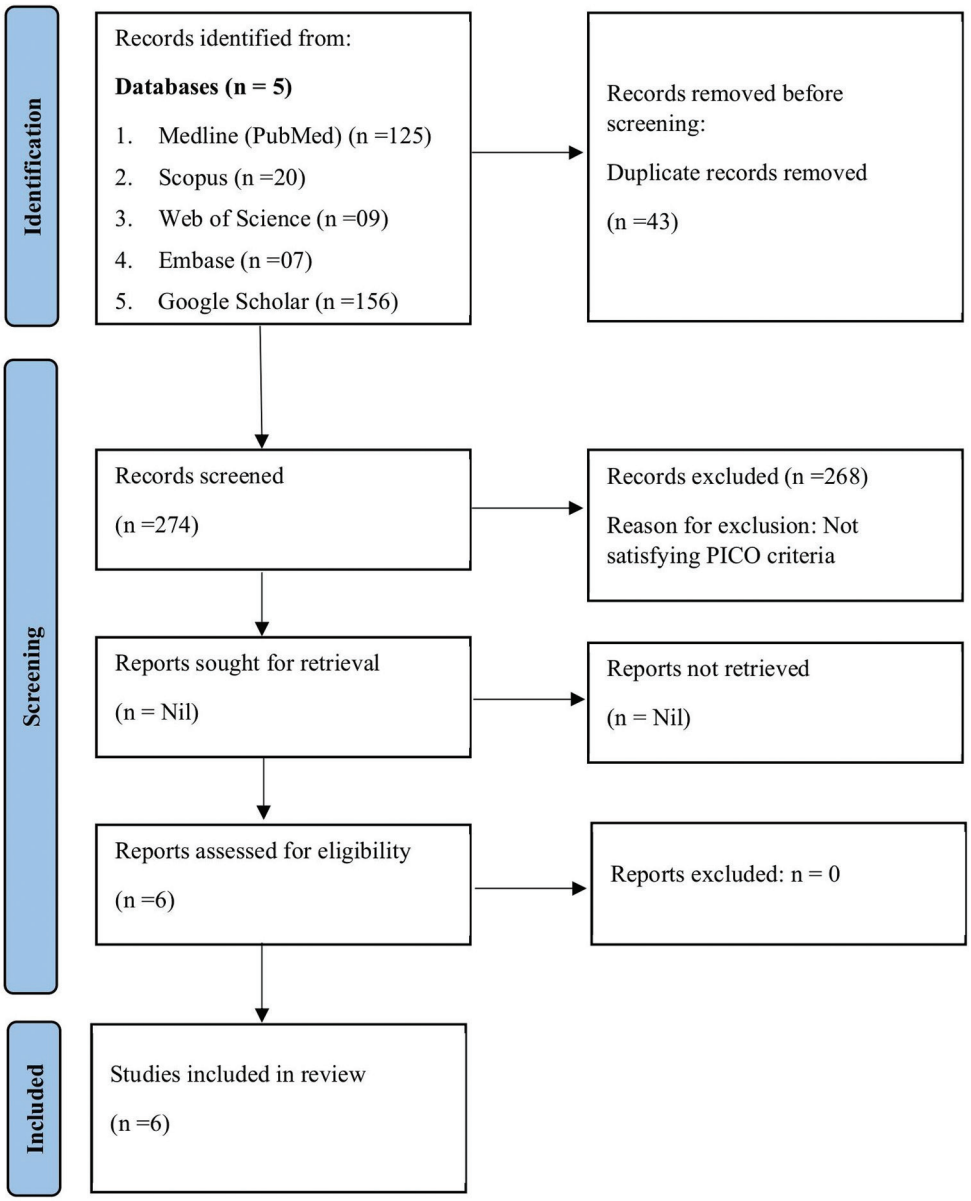

Figure 1: PRISMA flow diagram illustrating the study selection process for the umbrella review. The figure depicts the identification, screening, eligibility assessment, and inclusion of systematic reviews evaluating artificial intelligence-based detection of oral potentially malignant disorders and oral cancer using mobile photographs. The process followed PRISMA guidelines, including removal of duplicates, title and abstract screening, full-text assessment, and final inclusion of articles.

Table 2: General characteristics of included systematic reviews

Author, year	Study type	Study setting	Range of years	Number of studies	Dataset size	Sources searched	Models used	Comparison	Appraisal instrument
Rokhshad *et al.*, 2024[16]	Systematic Review and Meta - Analysis	Iran	2019 to June 2023	36 studies for qualitative and 15 studies for quantitative synthesis	46,339	PubMed, Web of Science, Scopus, Embase, IEEE, ArXiv, medRxiv, and Google Scholar	CNN, NN, SVM	Visual examination	QUADAS-2
Ferro *et al.*, 2022[17]	Systematic Review and Meta - Analysis	UK	Until February, 2022	35 studies were suitable for qualitative synthesis, and 31 for quantitative analysis	44,605	EMBASE, PubMed, Cochrane CENTRAL, and dblp	Classical (Linear discriminant analysis, Quadratic discriminant analysis, Decision tree, RVM, SVM) and Modern (NN based on DenseNet121, ResNet101, EfficientNet, VGG19, HRNet-W18, EfficientNet-b4)	Optical imaging, thermal imaging, and VOC analysis	QUADAS-2
Kavyashree *et al.*, 2024[18]	Systematic review	India	2014–2023	73	Not mentioned	IEEE Explore, Elsevier, Springer, John Wiley, Inderscience, Hindawi, Taylor Francis	SVM, logistic regression, kNN, decision tree classifier, LDA, QDA, AdaBoost Classifiers, MLP and GMM classifiers, ELM, CART algorithm, Random forest algorithm, SDC, Hybrid ABC-PSO Classifier, BLDA Classifier, RLS, MLP with back propagation, and DNN	Not mentioned	Not assessed
Beristain-Colorado *et al.*, 2024[19]	Systematic review	Mexico	Till July 2022	9	1581	PubMed, ClinicalTrials, Scopus, Google Scholar, Web of Science	ResNet-101 Faster R-CNN, EfficientNet-B0, CNN architecture: YOLOv5 EfficientNet-B4 Benign, CNN with DenseNet-121 and Faster R-CNN, automated DL algorithm using cascaded CNNs	Histopathological examination	QUADAS-2
Warin and Suebnukarn, 2024[20]	Systematic review	Thailand	2000–2023	23	2280 Clinical images	Medline, Scopus, Google Scholar	Faster R-CNN, YOLOv4, RetinaNet, CenterNet2		QUADAS-2
de Chauveron *et al.*, 2024[21]	Systematic review	France	2017–2023	11	12,996	Pubmed, Cochrane, Scopus, Embase, Dentistry and oral science source, dblp, Arxiv, Google Scholar, and OpenGrey (gray literature)	ResNet50 VGG-Skip, EffiscientNetb4, VGG19, NASNet, MLSO + SVM, CNN + ISSA, Probabilistic neural Network and CNN	Histopathological examination	HAS grading

LDA: Linear discriminant analysis; QDA: Quadratic discriminant analysis; SDC: Softmax discriminant classifier; RLS: Regularized least squares; DNN: Deep neural network; CNN: Convolutional neural network; NN: Neural network; SVM: Support vector machine; RVM: Relevance vector machine; DL: Deep learning; MLSO: Multi-Level Swarm Optimization; MLP: Multilayer Perceptron; GMM: Gaussian Mixture Model; ELM: Extreme Learning Machine; CART: Classification and Regression Tree; SDC: Softmax Discriminant Classifier; VOC: Volatile Organic Compounds, HAS: Haute Autorité de Santé

Table 3: Inference from included studies

Author, Year	Appraisal result	Outcome assessed	Key results	Significance	Heterogeneity
Rokhshad *et al.*, 2024[16]	Index test domain: High risk 13%, low risk 43%, Patient selection: Low risk 53%, Reference standard: Low risk 10%, high risk 23%, unclear 67%	Sensitivity, Specificity, DOR	AI accuracy: 74%–100%. Clinicians: 61%–98%. Pooled DOR: Malignant (155), Cancerous (114)	Significant	Sensitivity: I^2=85.75%, Specificity: I^2=97.40%
Ferro *et al.*, 2022[17]	Domain 1: High risk 11%, low risk 26%, unclear risk 63%, Domain 2: Low risk 43%, unclear risk 54%, Domain 3: Low risk 71%, unclear risk 29%, Domain 4: Low risk 69%, unclear risk 31%	Sensitivity, FPR, AUC	Classical: Sensitivity 90.4%, FPR 15.1%, AUC 91.5%. Modern: Sensitivity 88.3%, FPR 13.9%, AUC 93.2%	Insignificant	Sensitivity: I^2=62%, Specificity: I^2=84%
Kavyashree *et al.*, 2024[18]	Not assessed	Accuracy	Accuracy: SVM (100%), Logistic Regression (100%), CNN (99.3%), Random Forest (90%), MLP (94.1%)	-	-
Beristain-Colorado *et al.*, 2024[19]	High risk: Patient Selection (100%). Applicability concerns: 33%	Sensitivity, Specificity, AUC	NN accuracy >85%. Useful in remote clinical settings	-	-
Warin and Suebnukarn, 2024[20]	Low risk: Applicability (74.1%), Reference Standard (40.7%), Patient Selection (59.3%)	Sensitivity, Specificity, AUC, DOR	The pooled sensitivity: 92%, specificity: 92%, DOR: 2549.08. Precision: 76.7%–98.0%, Recall: 81.0%–92.0%, F1 score: 79.3%–89.0%, and AUC: 0.74–0.91, Accuracy: 76.0%–98.58%	-	-
de Chauveron *et al.*, 2024[21]	Intermediate evidence (Grade 4C)	Sensitivity, Specificity, Accuracy	MLSO + SVM performed best. VGG with attention performed poorly despite 2155 images	-	-

DOR: Diagnostic odds ratio; AUC: Area under the curve; CNN: Convolutional neural network; SVM: Support vector machine; VGG: Visual Geometry Group; MLP: Multilayer Perceptron; FPR: False Positive Rate; MLSO: Multi-Level Swarm Optimization

tool [Figure 2].[15] Discrepancies in their assessments were resolved through consultation with a third reviewer (M.K.) to achieve consensus. The ROB findings were visualized using the RoBvis tool.

RESULTS

Study selection and its characteristics

The study selection process was carried out independently by two authors (L.C and R.S). This involved an initial screening of titles and abstracts, followed by a detailed evaluation of full texts according to set inclusion and exclusion criteria. In the end, data extraction was conducted on six articles [Figure 1]. According to the JBI methodological guidelines, the data were systematically organized and presented in Tables 2 and 3. Table 3 offers a consolidated overview of the outcome measures derived from the included studies. The research predominantly investigated the diagnostic accuracy of various AI models, focusing on metrics such as sensitivity and specificity.

Main outcome of the study

The results of this Umbrella review demonstrate that the six systematic reviews and meta-analyses evaluated provide compelling evidence on the effectiveness of AI in clinical diagnostics across diverse settings and methodologies. The studies span geographical regions including Iran, the UK, India, Mexico, Thailand, and France, covering a wide range of years from 2000 to 2023. They collectively employed various AI models, traditional machine learning techniques, and robust appraisal frameworks to compare diagnostic performance against conventional methods.

Although several reviews reported pooled diagnostic metrics, a quantitative meta–meta-analysis was not performed due to substantial methodological heterogeneity, including differences in study designs, outcome definitions, AI architectures, validation strategies, and pooling methods. Overlap of primary studies and inconsistent reporting of pooled estimates further precluded reliable secondary quantitative synthesis. Accordingly, findings were narratively synthesized in line with JBI guidance, emphasizing comparative trends rather than pooled effect estimates.

Rokhshad *et al.*[16] reported high AI diagnostic accuracy (74%–100%), exceeding clinician performance (61%–98%), with pooled DORs of 155 for malignant and 114 for cancerous lesions. However,

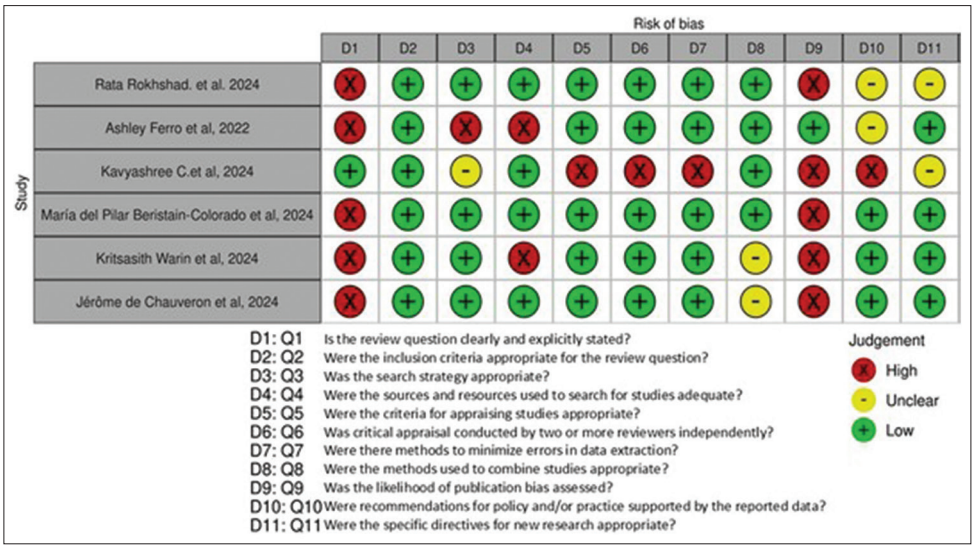

Figure 2: Risk of bias assessment of included systematic reviews using the Joanna Briggs Institute critical appraisal tool. The figure summarizes domain-wise risk of bias judgments across included reviews, including clarity of the research question, search strategy, study appraisal, data extraction, publication bias assessment, and applicability. Visualization was generated using the Robvis tool, highlighting variability in methodological quality that contributes to heterogeneity.

substantial heterogeneity in sensitivity (I^2 = 85.75%) and specificity (I^2 = 97.40%) was observed. Methodological appraisal identified high ROB related to research question clarity, search strategy, appraisal criteria, independent review, and publication bias assessment, limiting the overall reliability despite robust inclusion criteria.

Ferro *et al.*[17] compared classical and modern AI models, reporting high sensitivity (90.4%) and area under the curve (AUC) (91.5%) for classical methods, while modern neural networks achieved slightly higher AUC (93.2%) and lower false positive rates (13.9% vs. 15.1%). However, methodological limitations were noted in research question clarity, search strategy, appraisal criteria, independent review, and publication bias assessment, along with moderate heterogeneity in sensitivity (I^2 = 62%) and specificity (I^2 = 84%).

Kavyashree *et al.*[18] Reported very high accuracy for SVM and logistic regression (100%) and near-comparable performance for convolutional neural networks (CNNs) (99.3%), while random forest (90%) and Multilayer Perceptron (MLP) (94.1%) performed less well. Methodological concerns were limited, primarily involving unclear appraisal criteria (D5), search strategy (D3), publication bias assessment (D9), and future research directives (D11), with strengths in independent review and data extraction supporting overall reliability. Similarly, Beristain-Colorado *et al.*[19] demonstrated neural network accuracy exceeding 85% in remote clinical settings, with generally low ROB; limitations were confined to research question clarity (D1), appraisal criteria (D5), and one unclear domain (D9), while robust search and data extraction methods enhanced study reliability.

Warin and Suebnukarn[20] presented strong evidence for the utility of AI in diagnostics, reporting pooled sensitivity and specificity values of 92%, with an impressive DOR of 2549.08. Precision, recall, and F1 scores were consistent with these metrics, and AUC ranged from 0.74 to 0.91, underscoring the reliability of the models tested. This study has a moderate ROB. Issues exist in research question clarity, study appraisal, and publication bias assessment. In addition, the search strategy (D3) and policy recommendations (D10) are unclear. Despite these concerns, strong methodologies in data extraction, independent review, and future research directives make it relatively stronger than highly biased studies. Finally, de Chauveron *et al.*[21] identified MLSO + SVM as the top-performing model in their analysis, while VGG architectures underperformed despite substantial data inclusion. This study mirrors the bias trends of Rokhshad *et al.*[16] and Ferro *et al.*,[17] with high risks in research question clarity, search strategy, study appraisal, and publication bias. The lack of clarity in policy recommendations and research directives further weakens its reliability. Despite good inclusion criteria and data extraction methods, the study's overall ROB is high. The only criterion not applicable was the evaluation of publication bias, due

to the inclusion of predominantly qualitative evidence synthesis in some studies. Therefore, the overall ROB across the included reviews is considered to be minimal [Figure 2].

Overall, the results affirm the potential of AI, particularly deep learning and hybrid architectures, in enhancing diagnostic accuracy and reliability across clinical domains. However, the variability in study quality, significant heterogeneity, and frequent methodological limitations highlight the need for standardized approaches and more rigorous evaluation in future research.

DISCUSSION

The findings of the reviewed studies highlight the growing potential of AI in clinical diagnostics, aligning with existing literature on the topic. Previous research has consistently demonstrated the ability of AI models, particularly deep learning architectures like CNNs, to surpass traditional diagnostic methods in terms of accuracy and efficiency. For example, studies outside this review have reported diagnostic accuracies exceeding 90% for AI systems in detecting malignancies and other complex conditions, verifying the high accuracies noted by Rokhshad *et al*.[16] and Kavyashree *et al*.[18] Similarly, the comparative edge of modern neural networks, as seen in Ferro *et al*.,[17] aligns with earlier findings that architectures like ResNet and EfficientNet outperform classical approaches in sensitivity and specificity across various datasets.

Implications of heterogeneity and sources of variability

Substantial heterogeneity was consistently observed across the included reviews, with several meta-analyses reporting I^2 values exceeding 85% for sensitivity and specificity. Although pooled estimates indicate high overall diagnostic performance, such heterogeneity limits generalizability and suggests that AI model accuracy is highly context dependent. This variability likely arises from differences in patient populations, image acquisition conditions (camera specifications, resolution, lighting, focus, and angulation), AI architectures and training strategies, and validation approaches, with many studies relying solely on internal validation. While formal subgroup or meta-regression analyses were not feasible, future stratification by model type, imaging conditions, and validation strategy may aid interpretation. Importantly, the observed heterogeneity underscores the need for standardized imaging protocols, transparent model reporting, and rigorous external validation to support reliable clinical implementation.

Some results differ from broader findings in the literature. While most studies reviewed here report high accuracy and low false positive rates for AI models, heterogeneity in sensitivity and specificity metrics was a recurring issue. For instance, Rokhshad *et al*.[16] reported significant variability with I^2 values exceeding 85%, suggesting inconsistencies in model performance across included datasets. Other meta-analyses in the field have also noted heterogeneity but often to a lesser degree, potentially reflecting methodological differences such as dataset preprocessing or study inclusion criteria. Furthermore, while classical machine learning models like SVM and logistic regression performed exceptionally well in the review by Kavyashree *et al*.,[18] these findings contrast with more recent studies where such methods are increasingly outperformed by ensemble and deep learning methods, especially on larger datasets.

The strengths of this analysis lie in its comprehensive approach, incorporating diverse geographic settings, a wide range of AI models, and robust appraisal tools such as Quality assessment of diagnostic accuracy studies-2. This breadth provides a holistic view of AI's applicability in clinical diagnostics. Nevertheless, certain limitations must be acknowledged. Many studies exhibited unclear or high risks, as noted by Ferro *et al*.[17] and Beristain-Colorado *et al*.[19] These issues could bias the pooled results and limit the generalizability of findings. Furthermore, the heterogeneity in metrics such as sensitivity and specificity, particularly in Rokhshad *et al*.[16] and Warin and Suebnukarn[20] underscores the need for standardized reporting and evaluation frameworks to enhance comparability across studies.

Another significant limitation is the lack of detailed subgroup analyses or stratification by factors such as population characteristics, imaging modalities, or model architectures. For instance, while modern AI models consistently outperformed classical approaches, their performance varied significantly across different diagnostic tasks and imaging techniques. Including such stratifications could provide deeper insights into the specific conditions or scenarios where AI excels or underperforms. In addition, certain studies, such as those by Kavyashree

et al.[18] and Beristain-Colorado *et al.*,[19] lacked appraisal of methodological rigor, which could affect the reliability of their findings.

Future research should address these limitations by adopting standardized methodologies and reporting guidelines to improve the consistency and reliability of results. Greater emphasis on external validation using independent datasets is also essential to assess the generalizability of AI models. Studies should also focus on exploring hybrid models that combine the strengths of classical and modern techniques, as suggested by the promising performance of hybrid architectures like MLSO + SVM reported by de Chauveron *et al.*[21]

In conclusion, while the findings affirm AI's transformative potential in clinical diagnostics, they also highlight significant gaps in methodology and reporting. Addressing these gaps will be critical in realizing the full potential of AI in delivering accurate, reliable, and scalable diagnostic solutions across diverse clinical settings.

CONCLUSION

Based on the comprehensive analysis of the studies, modern deep learning architectures consistently demonstrated superior performance compared to classical machine learning models, making them the most recommended for clinical diagnostics. Among these, models such as EfficientNet, ResNet, and hybrid configurations like MLSO + SVM showed the highest levels of accuracy, sensitivity, and specificity across various tasks and datasets.

EfficientNet and ResNet, particularly their advanced versions such as EfficientNet-B4 and ResNet-101, excelled in diverse applications, achieving accuracy rates above 90% in multiple studies. These models also exhibited robust adaptability to different imaging modalities and datasets, underscoring their versatility and reliability in clinical diagnostics. The hybrid MLSO + SVM model emerged as particularly noteworthy, combining the strengths of classical and modern techniques, delivering the best performance in its respective study with strong generalization capabilities.

In contrast, classical models like SVM and Logistic Regression, although achieving high accuracy in certain datasets (e.g., 100% in specific tasks), generally lagged behind deep learning architectures in terms of overall diagnostic power and scalability. Their limited ability to handle complex, high-dimensional data makes them less suitable for modern clinical applications compared to more advanced neural networks.

Therefore, based on current evidence, deep learning models like EfficientNet and ResNet are highly recommended for clinical diagnostics due to their superior performance and adaptability. Hybrid models, such as MLSO + SVM, also hold significant promise, particularly in scenarios where leveraging the complementary strengths of classical and modern methods is beneficial. Future efforts should focus on refining these models further and validating their performance across diverse clinical contexts to ensure consistent and reliable diagnostic outcomes.

Protocol registration

The protocol for this study can be reviewed on the International Prospective Register of Systematic Reviews (PROSPERO) database, referenced under the registration number: CRD420250656539.

Patient and public involvement

This research was conducted, reported, and disseminated without direct engagement or input from patients or the general public.

Data availability statement

Data can be obtained through a reasonable request process.

Financial support and sponsorship

This study is a part of an ICMR project (Project ID IIRP-2023-1049) funded by Small extramural grants – 2023.

Conflicts of interest

The authors of this manuscript declare that they have no conflicts of interest, real or perceived, financial or non-financial in this article.

REFERENCES

1. Parak U, Lopes Carvalho A, Roitberg F, Mandrik O. Effectiveness of screening for oral cancer and oral potentially malignant disorders (OPMD): A systematic review. Prev Med Rep 2022 Dec;30:01-08.
2. González-Moles MÁ, Aguilar-Ruiz M, Ramos-García P. Challenges in the early diagnosis of oral cancer, evidence gaps and strategies for improvement: A scoping review of systematic reviews. Cancers (Basel) 2022; Oct 10;14(19):4967.
3. Warnakulasuriya S, Kujan O, Aguirre-Urizar JM, Bagan JV, González-Moles MÁ, Kerr AR, *et al.* Oral potentially malignant disorders: A consensus report from an international seminar

on nomenclature and classification, convened by the WHO collaborating centre for oral cancer. Oral Dis 2021 Nov; 27(8):1862–80.
4. Mohan P, Richardson A, Potter JD, Coope P, Paterson M. Opportunistic screening of oral potentially malignant disorders: A public health need for India. JCO Glob Oncol 2020 Nov;(6):688–96.
5. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, *et al.* Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin 2019 Mar; 69(2):127–57.
6. Talwar V, Singh P, Mukhia N, Shetty A, Birur P, Desai KM, *et al.* AI-assisted screening of oral potentially malignant disorders using smartphone-based photographic images. Cancers (Basel) 2023 Aug 16;15(16):4120.
7. Lin H, Chen H, Weng L, Shao J, Lin J. Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis. J Biomed Opt 2021 Aug 28;26(08):01-16.
8. Ilhan B, Guneri P, Wilder-Smith P. The contribution of artificial intelligence to reducing the diagnostic delay in oral cancer. Oral Oncol 2021 May;116:01-07.
9. Haron N, Zain RB, Nabillah WM, Saleh A, Kallarakkal TG, Ramanathan A, *et al.* Mobile phone imaging in low resource settings for early detection of oral cancer and concordance with clinical oral examination. Telemed J E Health 2017 Mar; 23(3):192–9.
10. Pedroso CM, Warnakulasuriya S, Santos-Silva AR. Teledentistry in the detection of oral potentially malignant disorders and oral cancer in the Latin American region: A review of literature with current possibilities. Explor Digit Health Technol 2024 Oct 22;291–301.
11. Hegde S, Ajila V, Zhu W, Zeng C. Artificial intelligence in early diagnosis and prevention of oral cancer. Asia Pac J Oncol Nurs 2022 Dec;9(12): 01-06.
12. Hunt B, Ruiz A, Pogue B. Smartphone-based imaging systems for medical applications: A critical review. J Biomed Opt 2021 Apr 15;26(04):01-22.
13. Uppal S, Kumar Shrivastava P, Khan A, Sharma A, Kumar Shrivastav A. Machine learning methods in predicting the risk of malignant transformation of oral potentially malignant disorders: A systematic review. Int J Med Inform 2024 Jun; 186:01-15.
14. Sahoo RK, Sahoo KC, Dash GC, Kumar G, Baliarsingh SK, Panda B, *et al.* Diagnostic performance of artificial intelligence in detecting oral potentially malignant disorders and oral cancer using medical diagnostic imaging: A systematic review and meta-analysis. Front Oral Health 2024 Nov 6;5:01-14.
15. Aromataris E, Fernandez R, Godfrey C, Holly C, Khalil H, Tungpunkom P. Umbrella reviews. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Ch. 10. Adelaide, South Australia, Australia: JBI; 2020.
16. Rokhshad R, Mohammad-Rahimi H, Price JB, Shoorgashti R, Abbasiparashkouh Z, Esmaeili M, *et al.* Artificial intelligence for classification and detection of oral mucosa lesions on photographs: A systematic review and meta-analysis. Clin Oral Investig 2024 Jan 13;28(1):88.
17. Ferro A, Kotecha S, Fan K. Machine learning in point-of-care automated classification of oral potentially malignant and malignant disorders: A systematic review and meta-analysis. Sci Rep 2022 Aug 13;12(1):01-15.
18. Kavyashree C, Vimala HS, Shreyas J. A systematic review of artificial intelligence techniques for oral cancer detection. Healthc Anal 2024 Jun;5:01-18.
19. Beristain-Colorado MD, Castro-Gutiérrez ME, Torres-Rosas R, Vargas-Treviño M, Moreno-Rodríguez A, Fuentes-Mascorro G, *et al.* Application of neural networks for the detection of oral cancer: A systematic review. Dent Med Probl 2024;61(1):121-8.
20. Warin K, Suebnukarn S. Deep learning in oral cancer- a systematic review. BMC Oral Health 2024 Feb 10;24(1):212.
21. de Chauveron J, Unger M, Lescaille G, Wendling L, Kurtz C, Rochefort J. Artificial intelligence for oral squamous cell carcinoma detection based on oral photographs: A comprehensive literature review. Cancer Med 2024 Jan;13(1):01-12.