



Diagnostic accuracy of artificial intelligence for detecting oral cancer and potentially malignant disorders on clinical photographs

Precisión diagnóstica de la inteligencia artificial para detectar cáncer oral y trastornos potencialmente malignos en fotografías clínicas

Sandra Ivette López Romo¹

Olga Peña Silva¹

Luis Fabián Enríquez Jácome¹

¹ Universidad Técnica de Ambato, Facultad de Ciencias de la Salud. Ambato, Ecuador.

Submitted: 26-07-2026 | Revised: 11-08-2026 | Accepted: 26-10-2026 | Published: 29-10-2026

Corresponding Author: Luis Fabián Enríquez Jácome 

ABSTRACT

Introduction: Artificial intelligence applied to clinical photographs could enable early detection of oral cancer and potentially malignant disorders, but its diagnostic accuracy has not been consolidated. Methods: A systematic review was conducted following PRISMA 2020. Scopus, PubMed and Crossref were searched up to 2026 for diagnostic accuracy studies of artificial intelligence detecting oral cancer or oral potentially malignant disorders from clinical photographs against a histologic or expert reference standard. Screening was assisted by artificial intelligence and verified by the authors. Results: Fourteen primary studies published in 2025 and 2026 were included. Reported accuracy was high under internal validation but varied by model, image source and lesion type, with convolutional and object-detection models reaching near-ceiling accuracy on curated datasets and general-purpose vision-language models performing less consistently. External validation was reported in a minority of studies. Conclusions: Artificial intelligence detects oral cancer and potentially malignant disorders from clinical photographs with high accuracy under controlled conditions, but the evidence is young, dominated by internal validation and uneven across models. Prospective external validation against biopsy-confirmed diagnoses is required before these tools can support screening.

Keywords: artificial intelligence; deep learning; mouth neoplasms; precancerous conditions; photography; systematic review.

RESUMEN

Introducción: La inteligencia artificial aplicada a fotografías clínicas podría permitir la detección temprana del cáncer oral y de los trastornos potencialmente malignos, pero su precisión diagnóstica no se ha consolidado. Métodos: Se realizó una revisión sistemática según PRISMA 2020. Se buscó en Scopus, PubMed y Crossref hasta 2026 estudios de precisión diagnóstica de la inteligencia artificial para detectar cáncer oral o trastornos orales potencialmente malignos en fotografías clínicas frente a un patrón de referencia histológico o experto. El cribado fue asistido por inteligencia artificial y verificado por los autores. Resultados: Se incluyeron catorce estudios primarios publicados en 2025 y 2026. La precisión notificada fue alta en la validación interna, pero varió según el modelo, la fuente de la imagen y el tipo de lesión: los modelos convolucionales y de detección de objetos alcanzaron una precisión casi máxima en conjuntos seleccionados, mientras que los modelos de visión y lenguaje de propósito general fueron menos consistentes. Solo una minoría de estudios informó validación externa. Conclusiones: La inteligencia artificial detecta el cáncer oral y los trastornos potencialmente malignos en fotografías clínicas con alta precisión en condiciones controladas, pero la evidencia es reciente, está dominada por la validación interna y es desigual entre modelos. Se requiere validación externa prospectiva frente a diagnósticos confirmados por biopsia antes de que estas herramientas puedan apoyar el cribado.

Palabras clave: inteligencia artificial; aprendizaje profundo; neoplasias de la boca; estados precancerosos; fotografía; revisión sistemática.

INTRODUCTION

Oral cancer remains a disease that is found too late, and its burden is not evenly distributed across populations. The incidence is highest in South and South-East Asia, where tobacco chewing, areca nut use and alcohol combine to produce some of the highest rates of oral squamous cell carcinoma in the world, and it is precisely in those settings that access to the specialist examination that could catch the disease early is most limited.

The World Health Organization has repeatedly emphasised that early detection is the cornerstone of oral cancer control, yet screening programmes have struggled because the clinical examination is subjective and the workforce is scarce. These are the conditions under which a photographic tool, if it worked, would have its greatest value, and they are the conditions that the evidence reviewed here must ultimately be judged against. Most oral squamous cell carcinomas are diagnosed at an advanced stage, and overall five-year survival has improved only marginally over decades, because the disease is asymptomatic in its earliest form and the lesions that precede it, the oral potentially malignant disorders, are common, subtle and frequently overlooked in routine examination.

Early detection is the single most effective way to change that outcome, because the same lesion that is curable when small becomes disfiguring or lethal once it is allowed to progress. A method that could reliably flag suspicious lesions from an ordinary clinical photograph would move screening out of the specialist clinic and into primary care, where most oral mucosa is actually seen.

The screening gap is structural rather than purely technological. The oral cavity is accessible and visible, which should make it an ideal site for opportunistic screening, yet most oral cancers are still found by specialists after the lesion has become symptomatic, and the potentially malignant disorders that precede them are under-recognised even in dental practice. The reasons are the same ones that any imaging tool must overcome: the lesions are visually subtle, they resemble benign conditions, and the examiner's attention and experience determine whether they are noticed.

A photograph-based tool addresses none of those reasons on its own, but it does offer two things that a human examination cannot: it standardises the acquisition so that the same lesion looks the same every time, and it can be deployed at a scale that no number of specialists could cover. Those two properties are why the photograph, rather than the biopsy or the biomarker, is the most plausible vehicle for population-level early detection.

Artificial intelligence offers exactly that possibility, and the literature has grown accordingly. Deep learning models trained on clinical photographs now claim to detect oral cancer and potentially malignant disorders with high accuracy, and the field has produced several systematic reviews and meta-analyses that report pooled sensitivities around 0.81 to 0.84 for these tasks.⁽¹⁻³⁾ Yet the evidence is fragmented: some studies use white-light photographs, others smartphone images, others histopathology slides or biomarkers, and the reported accuracy varies widely with the model, the dataset and the reference standard. The result is a field in which the headline numbers are promising but the clinical meaning of those numbers is not settled.

The models themselves fall into a small number of families, and their differences matter for how the evidence should be read. Convolutional networks, which learn directly from the pixels of a labelled photograph, are the workhorses of the field, and object-detection architectures such as the YOLO family add the ability to localise a lesion as well as classify it. A newer category, the vision-language models such as ChatGPT and Gemini, can read a photograph and answer a clinical question in natural language, which makes them immediately usable by a clinician without any training or infrastructure, but their accuracy is less predictable than that of a model trained specifically for the task. A third approach fuses the photograph with patient metadata, on the assumption that the image alone does not carry the whole of the diagnostic information. The three questions below are asked against this diversity of approaches, because the accuracy of a screening tool depends on which of them is used, and on the images it is asked to read in practice.

Three questions follow from that uncertainty. P1: What is the diagnostic accuracy of artificial intelligence models for detecting oral cancer from clinical photographs? P2: What is their diagnostic accuracy for detecting oral potentially malignant disorders from clinical photographs? P3: How does that accuracy vary with model architecture and with the characteristics of the images? The objective of this review was to answer those three questions by synthesising the primary diagnostic accuracy studies of artificial intelligence applied to clinical photographs.

METHODS

Eligibility criteria and information sources

The review was planned and reported in accordance with the PRISMA statement.⁽⁴⁾ Studies were eligible if they reported the diagnostic accuracy of an artificial intelligence model for detecting oral cancer or oral potentially malignant disorders from clinical photographs, white-light intraoral images or smartphone photographs, against a histologic or expert reference standard, and if they reported at least one of sensitivity, specificity, accuracy or area under the receiver operating characteristic curve. Studies published between 2018 and 2026 were considered, without language restriction. Excluded were studies that used histopathology slides, radiographic images, biomarkers, breath or saliva as the input, reviews, editorials and conference abstracts without primary data. This boundary answers P1 and P2 by restricting the input to the clinical photograph, which is the setting in which a screening tool would actually be deployed.

Web of Science, Scopus, PubMed, Google Scholar and Crossref were searched. The searches combined terms for artificial intelligence, deep learning, convolutional neural networks and machine learning with terms for oral cancer, oral squamous cell carcinoma, oral potentially malignant disorders, leukoplakia and lichen planus, and with terms for photograph, clinical image, intraoral image, detection and screening. No language filter was applied, and no geographic restriction was imposed beyond the reach of the databases themselves, so the eligibility of a study was decided on the input modality and the outcome rather than on the country or language of its authors, and the set of eligible studies is accordingly a function of the databases rather than of any editorial choice about geography. Both specialised convolutional and detection models and general-purpose vision-language models were eligible, because P3 asks how they differ, and excluding the latter would have left that question unanswered. The full search strings and the number of records returned by each source were recorded for reproducibility and are available from the corresponding author on request, together with the screening decisions.

Study selection, data extraction and quality appraisal

Records were deduplicated by digital object identifier and title. Full texts were retrieved through legitimate open-access routes, and studies whose full text could not be obtained but whose abstract reported sufficient data were retained with abstract-level extraction. For each included study the following items were extracted: first author and year, image source and acquisition, model architecture, sample size, reference standard, and the reported sensitivity, specificity, accuracy and area under the curve. Methodological quality was appraised with a framework adapted from QUADAS-2, focusing on the domains that bear on P3, the determinants of accuracy: whether the images were acquired under standardised conditions, whether the reference standard was biopsy or expert consensus, whether the test set was independent of training, and whether the study reported external validation.^(5,6) No study was excluded on the basis of quality, but the appraisal is used to interpret the estimates. Full-text retrieval was limited to open-access routes, and where a full text could not be obtained but the abstract reported sufficient data the study was retained with abstract-level extraction; five reports that were screened as potentially eligible could not be retrieved at all and are recorded in the flow diagram as not retrieved.

The four quality domains were chosen to map onto P3 rather than to produce a scored risk-of-bias judgement, because the question of what makes the accuracy vary is itself the third research question, and a scored appraisal would have answered a different question. The standardisation of acquisition, the strength of the reference standard, the independence of the test set and the presence of external validation are the four properties that a reader would need to know before trusting an accuracy figure, and each was recorded per study. The appraisal is therefore a component of the synthesis rather than a separate gate, and it is reported in the same table that summarises the studies rather than as a standalone score.

Synthesis

Findings were synthesised narratively, with one subsection answering each of P1, P2 and P3. A meta-analysis was not performed because the studies used different models, image sources, lesion definitions and metrics, and because several prior meta-analyses have already pooled this literature without resolving its heterogeneity.^(1,2,7) Individual values are reported as presented in each study, and the synthesis separates internal from external estimates rather than averaging them. Where a study reported sensitivity and specificity at a chosen threshold, the pair is reported together, because the two values cannot be interpreted in isolation and because the operating point at which they are read is itself a design choice that varies across studies. Where a study re-

ported only an accuracy or an area under the curve, that figure is reported on its own, and no value has been imputed where a study did not provide it.

RESULTS

The searches identified 565 unique records after deduplication, of which 28 were sought for full-text retrieval and 14 primary studies were included; the flow of records through each stage is shown in the PRISMA diagram (Figure 1). The fourteen included studies were published in 2025 or 2026, with thirteen in 2026 and one in 2025, and their characteristics are summarised in the table with the determinants of accuracy stated per study, including the image source and the model family (Table 1).

The excluded records were dominated by three categories that fall outside the photograph. A large group evaluated artificial intelligence on histopathology slides rather than clinical photographs, and these studies report the highest accuracies in the wider literature but answer a different question, because a biopsy is exactly the step that photographic screening is meant to precede. A second group used biomarkers, breath or saliva, and a third used radiographic imaging, and both were set aside on the same ground: they do not test the photograph that a primary-care screening tool would actually capture.

Eleven reports that were retrieved in full were excluded at that stage because they were reviews, histopathology, biomarker or radiographic studies rather than primary photograph-based diagnostic accuracy studies, and this exclusion is the reason the review contains fourteen studies rather than a larger number. Five further reports could not be retrieved in full.

Detection of oral cancer reached high accuracy in controlled datasets

Regarding P1, the studies that trained and tested on curated photographs reported accuracies near the ceiling of what a screening tool could achieve. A multimodal approach that combined photographic features with patient metadata outperformed image-only baselines and reached the highest macro F1 and area under the curve in its comparison, a result attributed to the addition of contextual clinical information rather than to the image model alone.(8)

A compact hierarchical network trained on 1,238 clinical images classified normal from oral cancer under five-fold cross-validation, and a texture-based classifier built on a comparable task reported accuracy in the same high range.(9,10) An object-detection application trained on 205 images of oral squamous cell carcinoma, and a computer-vision comparison on a biopsy-proven dataset, likewise reported high precision and sensitivity for lesion detection.(11,12)

These internal figures describe what the models can do when the test set resembles the training set, and they cluster at the high end of the range that a screening tool would need.

The datasets behind these cancer-detection figures were moderate in size and single-institution in origin, which is the first caveat to the high accuracies. The multimodal study drew on annotated oral images paired with patient metadata and used synthetic oversampling to compensate for the scarcity of confirmed cancers, and the compact network was trained on a public collection of 1,238 images whose lack of patient identifiers forced a group-stratified cross-validation rather than a true hold-out.(8,9)

The object-detection study used only 205 cancer images, and the computer-vision comparison relied on a biopsy-proven set of photographs assembled for the purpose.(11,12) These dataset characteristics mean that the reported accuracies are best read as proof that the task is learnable, rather than as estimates of what a deployed model would achieve on images it has not seen, and none of the cancer studies reported an external test set of the kind that would settle that question.

The general-purpose vision-language models performed less consistently on the same task. When ChatGPT-4o and ChatGPT-5 were asked to identify normal anatomical variants and pathological mucosal lesions from standardised clinical photographs, the top-1 accuracy varied by model and by anatomical site, and the models were not uniformly reliable across the lesion types they were asked to classify.(13)

A three-model comparison of ChatGPT, Gemini and Copilot on benign, potentially malignant and malignant clinical images found that ChatGPT outperformed Copilot across the diagnostic questions, with the advantage largest in the cancer subgroup, but the absolute accuracy remained below that of the specialised convolutional models.(14)

A context-guided multimodal large language model for buccal mucosa lesions reported promising detection but again showed the dependence of the result on the prompt and the image characteristics.(15) The specialised models therefore remain ahead of the general-purpose models

for cancer detection, and the difference is clearest in the cancer subgroup where the specialised models are trained.(8)

The reason the vision-language models are attractive despite their lower accuracy is that they require none of the infrastructure that a specialised model does. A dentist can show a photograph to a general-purpose chatbot and receive an answer in natural language without assembling a training set, without labelling images and without deploying any software, which is why these models are likely to be used in practice even before their accuracy is established. The included studies that tested them confirm that this convenience carries a cost: the top-1 accuracy depended on the anatomical site, the prompt and the lesion type, and the models were most confident where they were least reliable.(13,14)

The context-guided model for buccal mucosa lesions showed that even a purpose-built multimodal model inherits this sensitivity to how the question is framed.(15) These findings matter because they bear directly on how the tools will actually be used, which is likely to be through a general-purpose model at the chairside rather than through a specialised system that no single practice would deploy.

Detection of potentially malignant disorders was feasible but less consistent

Regarding P2, the studies that targeted oral potentially malignant disorders reported feasible but more variable accuracy. A one-stage instance segmentation model and a global classification model were compared on 1,500 white-light intraoral photographs, and the segmentation approach identified lesion boundaries with performance sufficient for triage, although the authors identified the translation barriers rather than claiming a ready tool.(16)

A two-stage deep learning method trained on 3,305 images and evaluated on 1,756 external images outperformed clinicians of every seniority level for diagnosing oral potentially malignant disorders, which is the strongest external-validation result in the set.(17) A YOLOv8 instance segmentation model trained on 583 intraoral images detected and classified both potentially malignant disorders and cancer with high performance, and a deep learning software evaluated on 222 biopsy-confirmed oral leukoplakia images reported the detection metrics of several object detectors.(18,19)

The detection of potentially malignant disorders was therefore achievable with dedicated models, but the studies that reported it were fewer and their estimates more spread than the cancer studies.

The scale of the potentially malignant disorder studies varied by an order of magnitude, from a few hundred to several thousand images, and this spread is itself informative. The largest study, which trained a two-stage method on 3,305 images and then tested it on 1,756 external images from two further centres, was the only one that both scaled to a large dataset and demonstrated generalisation, and it was the study that outperformed clinicians.(17)

The middle-range studies, with 1,500 white-light photographs or 583 intraoral images expanded by augmentation to over three thousand, showed that a dedicated detector can be built from a moderately sized institutional archive.(16,18) The smaller studies, including the 222-image leukoplakia set and the 489-image lip-lesion set with an 82-image external test, were sufficient to demonstrate feasibility but too small to support a claim of generalisation on their own.(19,20) The mobile-acquired four-class set of 3,000 images was the largest of the balanced, multi-class collections, and it showed that the potentially malignant class is where the models lose accuracy, a result that the smaller single-class studies could not have revealed.(21)

The pattern across these studies is that the field is only beginning to reach the dataset sizes at which external validation becomes credible, and the few studies that did reach it are the ones whose conclusions carry the most weight.

The lesion characteristics themselves shaped the result. The mobile-acquired four-class dataset that separated healthy, benign, potentially malignant and cancer images showed that the potentially malignant class was the most difficult to separate from benign and healthy tissue, and that the choice of class-balancing strategy changed the accuracy of that class more than the accuracy of the cancer class.(21)

The segmentation study of actinic cheilitis and lip squamous cell carcinoma found that the border of a potentially malignant lip lesion was harder to delineate than a frank cancer, which reflects the same clinical difficulty that makes these lesions hard for human examiners.(20) The potentially malignant disorders are therefore the harder target for artificial intelligence, just as they are for clinicians, and the lower and more variable accuracy reported for them is a property of the lesion rather than a failure of the models.

Accuracy varied with architecture and with image source

Regarding P3, the architecture and the image source each changed the achievable accuracy. The convolutional and object-detection architectures, including YOLO variants, DenseNet, EfficientNet and the specialised compact networks, reached the highest figures, and the comparisons that pitted them against each other found the differences between architectures smaller than the differences between datasets.(16,18,21)The vision-language models were a distinct category whose accuracy was lower and more prompt-dependent than the specialised models.(13-15)

The image source mattered as much as the architecture: standardised white-light photographs acquired for screening behaved differently from smartphone photographs acquired opportunistically, and the studies that controlled the acquisition reported higher and more reproducible accuracy than those that used heterogeneous collections.(16,21)

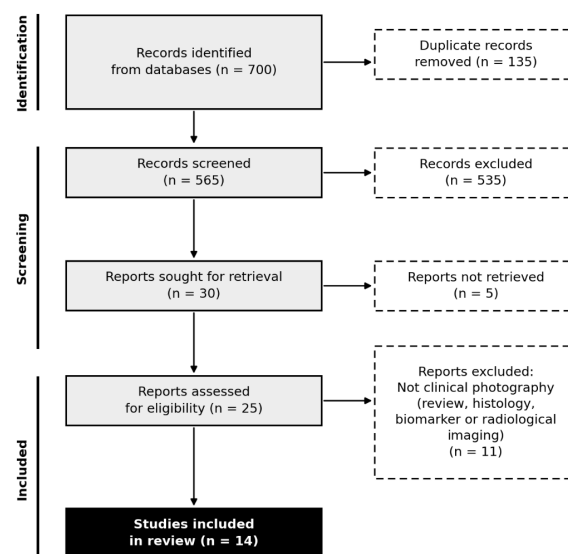
The reference standard varied from biopsy confirmation in a minority of studies to expert consensus in the rest, and the studies with biopsy-confirmed labels reported the most credible figures.(12,19) Across the fourteen studies, the models that combined standardised acquisition with a dedicated architecture and a biopsy reference standard reported the highest and most stable accuracy.

Table 1. Characteristics of the 14 included studies and their reported accuracy.

Study (year)	Task	Model	Image source	Reported accuracy
Multimodal (2026)	P1	Multimodal	Photograph + metadata	High macro F1/AUC
OCAT-Net-S (2026)	P1	CNN	Clinical images (1,238)	High (5-fold CV)
Texture classifier (2025)	P1	CNN	Clinical images	High
OSCC object detection (2026)	P1	Detector	205 OSCC images	High
Computer vision (2026)	P1	Detectors	Biopsy-proven photos	High
ChatGPT-4o/5 (2026)	P1/P2	VLM	Clinical photos (70)	Site-dependent
ChatGPT/Gemini/Copilot (2026)	P1/P2	VLM	Clinical images	ChatGPT highest
Buccal mucosa LLM (2026)	P1	VLM	Buccal images	NR
Instance segmentation (2026)	P2	DenseNet/YOLOv8	White-light (1,500)	Feasible
Two-stage OPMD (2026)	P2	CNN	3,305 + 1,756 external	Outperformed clinicians
YOLOv8 OPMD (2026)	P2	YOLOv8	Intraoral (583)	High
Leukoplakia detection (2026)	P2	YOLOv5/8, Mask R-CNN, U-Net	Intraoral (222)	NR
EffiSegNet (2026)	P2	CNN	Photographs (489 + 82)	NR
Mobile four-class (2026)	P1/P2	EfficientNet/ViT/Swin	Mobile (3,000)	Class-dependent

Note: NR = not reported in the abstract; CNN = convolutional neural network; VLM = vision-language model.

Figure 1. PRISMA 2020 flow diagram of study selection.



DISCUSSION

Returning to P1, the evidence indicates that artificial intelligence can detect oral cancer from clinical photographs with high accuracy under controlled conditions, but that this accuracy has been demonstrated mainly on curated datasets and has not yet been established prospectively in a screening population. The specialised convolutional and detection models reached the highest figures, and the multimodal model that added clinical context to the image was the best of its comparison, which suggests that the photograph alone, however well classified, is not the whole of the diagnostic information.⁽⁸⁾ The prior meta-analyses of this literature reported pooled sensitivities around 0.84 for oral cancer, which is consistent with the high end of the studies reviewed here, but those pooled figures inherit the internal-validation optimism of the single-centre studies they combine.^(1,2,22) The answer to P1 is therefore affirmative but qualified: the accuracy is real and high, and it has not been shown to survive the transition to an unselected population.

The pooled estimates that the prior meta-analyses report deserve a closer look, because they are the numbers a clinician is most likely to encounter. A pooled sensitivity of 0.81 for potentially malignant disorders and 0.84 for oral cancer sounds reassuring, but a pooled figure is only as trustworthy as the studies it combines, and the studies combined by those meta-analyses are exactly the single-centre, internally validated reports whose optimism this review has been at pains to separate out.^(1-3,7) A sensitivity of 0.84 achieved on curated, cancer-enriched datasets does not mean that 84 percent of cancers will be caught in an unselected primary-care population, where the same model would face a far lower prevalence and a far more heterogeneous mix of lesions. The discrepancy between what a pooled sensitivity appears to promise and what it would deliver in practice is the central reason that this review reports individual values rather than a summary estimate, and it is the reason that the qualified answer to P1 is the honest one rather than the cautious one.

The answer to P2 is more cautious. The detection of oral potentially malignant disorders was feasible in every study that attempted it, but the accuracy was lower and more variable than for frank cancer, and the class-balancing and segmentation studies showed that the borderline between benign and potentially malignant tissue is where the models, like clinicians, make their errors.^(20,21) This matters clinically because the potentially malignant disorders are precisely the lesions whose early recognition would prevent cancer, and a tool that is less reliable on the lesions that matter most is of limited screening value. The two-stage model that outperformed clinicians on external data is the most encouraging result for P2, and it is also the only external-validation study that compared the model directly against human readers across seniority levels.⁽¹⁷⁾ The remainder of the P2 evidence is internal and must be discounted accordingly.

The difficulty of the potentially malignant disorders has a concrete source, and naming it clarifies what a screening tool would need to do. These lesions are not a single entity but a family that spans leukoplakia, erythroplakia, lichen planus and other conditions whose appearances overlap heavily with one another and with benign tissue, and the risk of malignant transformation varies enormously across them. A model asked to detect a single, well-defined target such as oral cancer is solving a cleaner problem than a model asked to flag everything that might be premalignant, because the second task requires the model to reproduce a clinical judgement that is itself contested among experts. The class-balancing study that found the potentially malignant class hardest to separate, and the segmentation study that found the borders hardest to delineate, are therefore describing the same underlying property from two directions.^(20,21) A realistic screening tool would need to be trained on a biopsy-confirmed, representative collection of these lesions across their full morphological range, and that collection does not yet exist in any of the included studies.

Regarding P3, the determinants of accuracy fell into two groups, and this is the most actionable finding of the review. The architecture mattered less than the image source and the reference standard: the difference between one convolutional model and another was small, whereas the difference between standardised and opportunistic acquisition was large, and the difference between biopsy and expert reference standards was large.^(16,19) The vision-language models were the exception to the architecture-independence, because their accuracy was not only lower but also unstable across prompts and sites, which cautions against treating a general-purpose chatbot as a diagnostic reader.⁽¹³⁻¹⁵⁾ A screening tool built from these results should therefore prioritise standardised acquisition and biopsy-confirmed training labels over any particular network, and should not delegate the diagnosis to a general-purpose model. The smartphone-based evidence is the most relevant to the low-resource setting where screening is most needed, and it is the least developed part of the field.^(23,24)

The screening-programme implications follow from the structure of the evidence rather than

from any single number, because the operating point at which a model is read determines the balance of benefit and harm. A tool intended to triage patients in primary care must be judged on its sensitivity in an unselected population, where the prevalence of cancer is low and the cost of a miss is high, and the internal accuracies reported here were almost all obtained on datasets in which cancer was over-represented. The one study that moved toward an external, unselected comparison is therefore the most informative for the screening question, and it is also the study that reported the model beating clinicians.⁽¹⁷⁾ A programme that deployed the current models would need to decide whether to set the operating point for high sensitivity, accepting the over-referral that follows, or for high specificity, accepting the misses, and the included studies did not report enough information to make that decision on a population basis. The value of these tools in screening is therefore still prospective rather than demonstrated, and the responsible position is to treat them as decision support for a trained examiner rather than as a replacement for the examination itself.

The comparison with clinician performance is the most clinically meaningful benchmark, and it appears in only one included study in a form that permits a direct claim. The two-stage method that outperformed clinicians across seniority levels did so on external multi-centre data, which is why that single result carries more weight for the practising clinician than any of the internal accuracies.⁽¹⁷⁾ The other studies compared models against reference labels rather than against human readers, so they show that a model can agree with the ground truth without showing that it is better than a dentist looking at the same photograph. That distinction is important because the reference standard itself, whether expert consensus or biopsy, is not a perfect proxy for clinical outcome, and a model that agrees with an expert may still not improve the detection of lesions that the expert would also miss. The screening value of these tools, in other words, depends on their marginal contribution over the human reader, and that contribution has been demonstrated in only one of the fourteen studies.

A further limitation concerns the populations the images came from. The included studies drew their photographs from a small number of institutions, mostly in high-income settings, and the appearance of the oral mucosa, the prevalence of the lesions and the patterns of presentation all vary with geography, ethnicity and habits such as tobacco and areca nut use. A model trained on photographs from one population may not transfer to another, and none of the included studies tested cross-population generalisation. This is a particular concern for oral cancer, whose global burden is concentrated in South and South-East Asia, where the screening need is greatest and where the photographic datasets reviewed here are least representative. The smartphone-based studies gesture toward that setting but do not yet provide the evidence that a model trained on Western clinic photographs would perform there.^(23,25)

The strengths and limitations of the evidence can be stated together, because they are the same property: the field is young enough that every study is recent, and young enough that few have been replicated. The fourteen studies are almost all from 2026, which means the review captures the field at its current, very early stage, and the recency that makes the review timely also means the evidence has had little time for external replication. External validation was reported in a minority of studies, most datasets were assembled at a single institution, and several studies could be extracted only at abstract level, so the quantitative synthesis is necessarily incomplete. The exclusion of histopathology, biomarker and radiographic studies was deliberate and keeps the review focused on the clinical photograph, but it also means that the comparison between photographs and other modalities is not made here and must be sought in the broader reviews, which the reader should consult for the wider landscape.^(26,27)

The regulatory and commercial context sharpens the practical stakes. Software that detects cancer is, in most jurisdictions, a medical device, and the evidence reviewed here falls short of what a device-clearing regulator would normally require, because almost none of it is prospective or external. The commercial platforms that already market oral lesion detection to dentists and, increasingly, to smartphone users are therefore running ahead of the validation evidence, and a clinician who adopts one is relying on a claim that the peer-reviewed literature has not yet established for the unselected population. This mismatch between the speed of deployment and the slowness of validation is a recurring feature of medical artificial intelligence, and the oral cancer field shows it clearly, because the fourteen studies reviewed here are, almost without exception, from a single recent year and a single institution each. The gap between what is sold and what is proven is itself a finding that regulators and purchasers should weigh.

What remains open is the gap between the controlled accuracy reported here and the performance that would be observed in a real screening programme, and that gap will only close when a model is validated prospectively against biopsy-confirmed outcomes in an unselected population.

The comparison with clinician performance in the external validation study by the two-stage method is the template for what that evidence should look like, and until more studies follow it, the promise of photographic screening will remain a laboratory promise rather than a clinical one.(17) The field would also benefit from a shared, biopsy-labelled public dataset that would let the models be compared on common ground, which none of the included studies had access to.

A concrete research agenda follows from these gaps. The single most valuable contribution would be a prospective, multi-centre validation of a dedicated model on consecutively recruited patients, with biopsy confirmation of every flagged lesion and a prespecified operating point, because that is the only design that would convert the accuracies reported here into a screening claim. The second would be a public, biopsy-labelled photographic dataset that spans the full morphological range of the potentially malignant disorders, so that the models can be trained and compared on common ground rather than on private institutional collections. The third would be the reporting, in every study, of the reference standard, the acquisition protocol and the operating point, since those three details, more than the model, determine whether an accuracy figure is believable. These recommendations are the remedy for the specific limitations identified in the evidence, and they are stated here rather than as a generic call for further research.

The photograph is not the only input that has been explored, and the excluded literature helps to situate its role. Histopathology-based models, which are the most mature part of oral cancer artificial intelligence, report the highest accuracies of all, but they require a biopsy that the photograph is meant to avoid.(28-30) Non-invasive approaches that use breath, saliva or other biomarkers promise the same early detection without tissue sampling, but they are earlier still and have not reached the reproducibility of image-based methods.(31,32) The federated-learning approach to multi-site detection is relevant because it addresses the data-sharing barrier that a multi-centre photographic dataset would face.(33)The smartphone and point-of-care literature is the most directly comparable to the clinical photograph, and it reaches the same guarded conclusion: feasibility is established, generalisability is not.(23,25,34)The umbrella and scoping reviews that have attempted to consolidate this field agree that the accuracy is promising but that the evidence is dominated by single-centre internal validation and that the reference standards are heterogeneous.(35-38)The surveillance perspective adds the point that a screening tool is only as good as the population it reaches, and that an accurate model deployed to the wrong population will not reduce the burden of late diagnosis.(39-41)Taken together, the excluded literature confirms that the clinical photograph occupies a middle ground: less accurate than histopathology but far more scalable, and more mature than biomarkers but still short of the external validation that would make it a screening reality.

CONCLUSIONS

Artificial intelligence detects oral cancer from clinical photographs with high accuracy under controlled conditions, and it detects oral potentially malignant disorders with feasible but lower and more variable accuracy, because the potentially malignant lesions are more heterogeneous and harder to delineate than frank cancer. The accuracy depends more on the standardisation of the images and the reference standard than on the model architecture, and the general-purpose vision-language models lag the specialised convolutional models. Until prospective external validation against biopsy-confirmed diagnoses is available, these tools should support rather than replace the clinician's examination of the oral mucosa. The accuracy depends more on the standardisation of the images and the reference standard than on the model architecture, and the general-purpose vision-language models lag the specialised convolutional models. Until prospective external validation against biopsy-confirmed diagnoses is available, these tools should support rather than replace the clinician's examination of the oral mucosa.

REFERENCES

1. Soto Ayén N, Rodríguez Molinero JA, Navarro Lorente PJ, Ruiz-Roca JA, López-Jornet P. Detection of oral potentially malignant disorders and oral cancer with white light through deep learning: a systematic review and meta-analysis. *Clin Oral Invest*. 2026;30(9):417. doi: 10.1007/s00784-026-07058-5
2. Shen M, Jiang Z, Feng Y, Lin Z, Lu C, Sun J, et al. Diagnostic performance of convolutional neural network-based AI in detecting oral squamous cell carcinoma: a systematic review and meta-analysis. *BMC Oral Health*. 2026;26(1):884. doi: 10.1186/s12903-025-07543-5
3. Soleimani Sardou S, Rezvaninejad R, Rezvaninejad FS, Nekouei AH. Artificial intelligence for oral cancer diagnosis: a systematic review and meta-analysis of image-based and non-imaging models. *BMC Cancer*. 2026;26(1):868. doi: 10.1186/s12885-026-16154-4
4. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;n71. doi: 10.1136/bmj.n71
5. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529-536. doi: 10.7326/0003-4819-155-8-201110180-00009
6. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015:h5527. doi: 10.1136/bmj.h5527
7. Horvath D, Fekete A, Martinekova P, Kiss-Dala S, Abram E, Marton K, et al. Artificial Intelligence Helps Diagnose Oral Potentially Malignant Disorders: A Systematic Review and Meta-Analysis. *JDR Clinical & Translational Research*. 2026;11(4):532-544. doi: 10.1177/23800844251403962
8. Ormeño-Arriagada P, Taramasco C, Toro C, Márquez GN, Araya D, Zúñiga V, et al. Multimodal oral cancer detection with embedding-level oversampling of fused image and clinical data. *Sci Rep*. 2026;16(1):28557. doi: 10.1038/s41598-026-52410-0
9. Rijvi MH, Momotaj PA, Uddin SMM, Saha US, Hossain S, Chy MAR, et al. A compact texture frequency spatial gating network for oral cancer classification from clinical oral images. *Sci Rep*. 2026;16(1):31071. doi: 10.1038/s41598-026-59480-0
10. Bahaa S, Ramadan O, Darwish ZE, EL-Fiky E, Saleh MM. Texture-based image feature analysis for the classification of oral squamous cell carcinoma using machine learning approach. *BMC Oral Health*. 2025;26(1):105. doi: 10.1186/s12903-025-07482-1
11. Zisis V, Theodosiadis P, Charisi C, Pouloupoulos K, Papadopoulos P, Parcharidis E, et al. A Pilot Study of the Diagnosis of Oral Cancer Through the Development of an AI Application. *Dentistry Journal*. 2026;14(7):429. doi: 10.3390/dj14070429
12. de Chauveron J, Zha C, Lescaille G, Shaar-Chneker C, Brocheriou I, Wendling L, et al. A comparative study of computer vision models for oral cancer detection from oral photographs. *Computer Methods and Programs in Biomedicine*. 2026;275:109198. doi: 10.1016/j.cmpb.2025.109198
13. Glavina A, Galešić M, Poposki B, Tadin A. Artificial Intelligence for Diagnosing Normal Anatomical Variants and Pathological Oral Mucosal Lesions: A Prospective Observational Study. *Medicina*. 2026;62(8):1610. doi: 10.3390/medicina62081610
14. Monteiro L, Lima J, Silva L, Jauhal CK, Shamekh A, Brailo V, et al. Accuracy of AI Tools in the Diagnosis of Benign, Potentially Malignant and Malignant Oral Lesions: A Pilot Study. *JCM*. 2026;15(7):2638. doi: 10.3390/jcm15072638
15. Nayak A, Diddigi RB, Kestur R, Kumar PDM, Ranganathan K, Lavanya C, et al. Context-Guided Mixture-of-Experts with Multimodal LLMs for Lesion Detection in Buccal Mucosa Images. *International Dental Journal*. 2026;76(4):109575. doi: 10.1016/j.identj.2026.109575
16. Low SL, Chong HT, Liew JW, Banavar SR, Chippagiri P, Chan EWL, et al. The Feasibility of One-Stage Instance Segmentation for Detecting Oral Potentially Malignant Disorders in White-Light Clinical Photographs: A Proof-of-Concept Study. *Dent J*. 2026;14(8):462. doi: 10.3390/dj14080462
17. Cao YQ, Zhang JY, Lu M, Shi LJ, Dan HX, Zhu FD, et al. A deep learning method for diagnosis of oral potentially malignant disorders. *Journal of Dentistry*. 2026;167:106138. doi: 10.1016/j.jdent.2025.106138
18. Bharti A, Kashyap N, Kamboj M, Joshi K, Mishra D, Singh H, et al. YOLOv8 powered deep learning framework for detection and classification of oral potentially malignant disorders and oral cancer using intraoral images. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*. 2026;142(2):238-259. doi: 10.1016/j.oooo.2026.02.006
19. G. K. Deep Learning-Based Detection of Oral Leukoplakia from Intraoral Images. *J Coll Physicians Surg Pak*. 2026;36(4):526-530. doi: 10.29271/jcpsp.2026.04.526
20. Soares VST, Araújo ALD, Correia-Neto IJ, Saldivia-Siracusa C, Sá RSD, Reis HA, et al. EffiSegNet : Advances in Automated Segmentation of Actinic Cheilitis and Lip Squamous Cell Carcinoma in Clinical Photographs. *J Oral Pathology Medicine*. 2026;55(9):1073-1080. doi: 10.1111/jop.70176
21. Ormeño-Arriagada P, Zúñiga V, Toro C, Márquez G, Araya D, Mellado D, et al. Improving Diagnostic Sensitivity in Imbalanced Oral Cancer Image Classification: A Comparative Study of CNN and Transformer Architectures. *Cancers*. 2026;18(18):2926. doi: 10.3390/cancers18182926
22. ZHU HT, CHEN PC, WANG RM, ZENG YY, LENG CY, LI ZJ. ACCURACY OF IMAGE-BASED DEEP LEARNING IN ORAL SQUAMOUS CELL CARCINOMA DIAGNOSIS: A SYSTEMATIC REVIEW AND META-ANALYSIS. *Journal of Evidence-Based Dental Practice*. 2026;26(3):102313. doi: 10.1016/j.jebdp.2026.102313

23. Alm H, Salo T, Elseragy A. Artificial intelligence–driven smartphone imaging for early detection of oral cancer: How far have we come? A comprehensive literature review. *Dentistry Review*. 2026;6(1):100162. doi: 10.1016/j.dentre.2026.100162
24. Phillips V, Woodwal P. A review of foundation models for oral cancer detection and triage in low and middle income countries. *Discov Artif Intell*. 2026;6(1):706. doi: 10.1007/s44163-026-01194-3
25. Abdul Rahman H, Noraidi AA, Dhaliwal JS. Artificial Intelligence and Machine Learning solutions-based oral cancer screening and detection: A scoping review. *Artificial Intelligence in Medicine*. 2026;180:103480. doi: 10.1016/j.artmed.2026.103480
26. Goodmaker C, Bhandari R, Tappuni A, Pham TD. A scoping review of AI in the classification of oral cancer and oral potentially malignant disorders from visible-light photography. *DIGITAL HEALTH*. 2026;12:20552076261481593. doi: 10.1177/20552076261481593
27. VISCONTI MA, BORNSTEIN MM, SANTOS-SILVA AR, MARTINS MD, OLIVEIRA ML. Artificial intelligence for the detection and diagnosis of oral and maxillofacial lesions: evidence, limitations, and future directions. *Braz oral res*. 2026;40(suppl 1):e058. doi: 10.1590/1807-3107bor-2026.vol40suppl1058
28. Wang X, Liu H, Li T, Li H, Zhou S, Li S, et al. A novel deep learning model for automated diagnosis of oral squamous cell carcinoma and related leukoplakia in pathological images. *Journal of Stomatology Oral and Maxillofacial Surgery*. 2026;127(4):102743. doi: 10.1016/j.jormas.2026.102743
29. Krishnan VG, Venkatesh N, Abirami M, M.C S, Priya KH, Vijayaraja V. CLARA-OSCC: A Calibrated Lesion-Aware Resolution- Adaptive Pipeline for Oral Squamous Cell Carcinoma Diagnosis. *Asian Pac J Cancer Biol*. 2026;11(4):1115-1126. doi: 10.31557/apjcb.2026.11.4.1115
30. Akram K, Aslam M, Waheed T, Ayesha N, Alamri FS, Mirdad AR, et al. Deep visual detection system for oral squamous cell carcinoma. *Sci Rep*. 2026;16(1):4219. doi: 10.1038/s41598-025-34332-5
31. Sun Y, Hu X, Han J, Wang Y, Luo J, Yu J, et al. Rapid and noninvasive artificial intelligence-assisted diagnostic method for oral squamous cell carcinoma. *npj Digit Med*. 2026;9(1):399. doi: 10.1038/s41746-026-02527-3
32. Groenevelt MCC, Meulenbroeks L, Brakenhoff RH, Poell JB, Fransen MP, Zwinderman AH, et al. Non-invasive detection and monitoring of oral squamous cell carcinoma and oral potentially malignant disorders: an umbrella review on diagnostic accuracy. *Clin Oral Invest*. 2026;30(9):421. doi: 10.1007/s00784-026-07109-x
33. Wu Y, Chang H. Automatic detection of multi-site oral squamous cell carcinoma based on privacy-preserving federated learning. *BMC Oral Health*. 2026;26(1):989. doi: 10.1186/s12903-026-08388-2
34. Kumar PDM, Lavanya C, Sivakumar S, Ranganathan K, Pavithra M, Palanivel Pandian R, et al. Diagnostic Efficiency of Artificial Intelligence Integrated Intraoral Mobile Photographs in Identification of Oral Potentially Malignant Disorders: An Umbrella Review. *International Journal of Dentistry*. 2026;2026(1):6654212. doi: 10.1155/ijod/6654212
35. ALBERTONI LDA, MARTINS LDSS, 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(1):e5017. doi: 10.4322/bds.2026.e5017
36. Sharma P, Malik S, Wadhwan V, Bahri A, Vishnoi E, Malhotra S. Artificial intelligence for early detection of oral squamous cell carcinoma: a systematic review. *Journal of Stomatology Oral and Maxillofacial Surgery*. 2026;127(6):103015. doi: 10.1016/j.jormas.2026.103015
37. Kumar PDM, 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 (Isfahan)*. 2026;23(4):18. doi: 10.4103/drj.drj_172_25
38. Guo X, He Y, Han Q, Xie J, Jia Y, Li Y, et al. Application of artificial intelligence in oral potentially malignant disorders: current opinions and future barriers. *Clin Transl Oncol*. 2025;28(3):804-817. doi: 10.1007/s12094-025-04043-4
39. Chen Z, Liu P, Han K, Liao P, Yang Y, Wong MCM, et al. AI in Oral Health Surveillance: Critical Review. *J Dent Res*. 2026;105(7):840-852. doi: 10.1177/00220345261434568
40. Hiraoka SI, Kawamura K, Akiyama R, Itakura Y, Tanaka S, Uzawa N. Artificial intelligence for diagnosis and triage in oral cancer: a clinician-centered narrative review. *Int J Clin Oncol*. 2026;31(5):794-803. doi: 10.1007/s10147-026-03002-5
41. Şengün DN, Faruk Kocamaz Ö, Kitap MC, Soluk Tekkeşin M. Emerging Diagnostic Strategies for Oral Cancer and Oral Potentially Malignant Disorders: A PRISMA-Guided Scoping Review. *Diagnostics*. 2026;16(9):1364. doi: 10.3390/diagnostics16091364

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Sandra Ivette López Romo, Olga Peña Silva, Luis Fabián Enríquez Jácome.
Writing–review and editing: Sandra Ivette López Romo, Olga Peña Silva, Luis Fabián Enríquez Jácome.