SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Early Detection of Oral Cancer with Artificial Intelligence: A Systematic and Critical Review of Imaging Modalities, Learning Architectures, Evaluation Practice and Clinical Translation

Manjeet Kumar Soni[1*], Ashish Kumar Jain[2], Sunita Varma[3]

[1]Department of Information Technology, Shri G. S. Institute of Technology and Science (SGSITS), Indore 452003, India.
E-mail: mksoni@sgsits.ac.in
[2]Department of Computer Engineering, Institute of Engineering, Devi Ahilya Vishwavidyalaya, Indore 452001, India.
E-mail: ajain@ietdavv.edu.in
[3]Department of Computer Technology & Application, Shri G. S. Institute of Technology and Science (SGSITS), Indore 452003, India.
E-mail: sunita.varma19@gmail.com
*Corresponding Author: Manjeet Kumar Soni, E-mail: mksoni@sgsits.ac.in

Abstract
Oral cancer is one of the few common malignancies whose precursor lesions are directly visible and directly biopsiable, and yet the majority of cases worldwide are still diagnosed at an advanced stage. Artificial intelligence has therefore been positioned as the technology that will convert a visible but unrecognised lesion into an actionable referral. This article presents a systematic and deliberately critical review of 153 records spanning 1988-2026, assembled under a PRISMA-conformant protocol and appraised with an explicit Diagnostic Evidence and Transparency Evaluation for Cancer Triage (DETECT) rubric. The corpus comprises 50 primary oral-cancer studies, 16 reviews and meta-analyses, 10 epidemiological sources, 51 methodological foundations and 26 reporting standards and metrics, and 37% of it was published in 2023-2026. We organise the field along four axes, namely input modality, learning architecture, learning regime and clinical task, and resolve the evidence density across a modality-by-task matrix that exposes a specific and consequential imbalance: histopathological binary discrimination attracts 14 records while the referral-triage task that actually governs early detection attracts far fewer, and prospective evidence is essentially absent from every cell. Our central findings are three. First, an evaluation deficit: only 34% of primary studies state a patient-wise partitioning strategy, so a large fraction of reported performance may estimate within-patient similarity rather than diagnostic generalisation. Second, a translation deficit: 16% report any external or multi-site validation, 12% compare against a clinician reading the same images, and 14% release data or code; without a clinician comparator, a reported accuracy cannot establish that a model adds anything to current practice. Third, a prevalence deficit: most studies report metrics computed on artificially balanced case-control samples, whereas screening operates at a prevalence two to three orders of magnitude lower, at which the same sensitivity and specificity yield a positive predictive value that the literature does not report. We formalise eight recurrent failure modes, derive the prevalence-corrected predictive-value relations and a screening-yield model that make the third deficit quantitative, and contribute OSCAR-Eval (Oral Screening and Cancer AI Reporting Evaluation), an eleven-stage evaluation and reporting protocol with six invariants that a reviewer can check in minutes. A three-horizon roadmap separates reporting reforms achievable now from the multi-centre benchmark and prospective-trial infrastructure that is not. All 153 records were retained only after their bibliographic identity was verified against a retrievable source.

Keywords Oral cancer, oral squamous cell carcinoma, oral potentially malignant disorders, early detection, systematic review, deep learning, histopathology, explainable artificial intelligence, external validation, screening prevalence, PRISMA.

Nomenclature

AI	Artificial intelligence.	**PPV**	Positive predictive value.
AUC	Area under the receiver operating characteristic curve.	**PRISMA**	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
CLAIM	Checklist for Artificial Intelligence in Medical Imaging.	**QUADAS-2**	Quality Assessment of Diagnostic Accuracy Studies, version 2.
CNN	Convolutional neural network.	**ROC**	Receiver operating characteristic.
CT	Computed tomography.	**Se, Sp**	Sensitivity and specificity.
DETECT	Diagnostic Evidence and Transparency Evaluation for Cancer Triage; the appraisal rubric defined in Section 2.4.	**SHAP**	SHapley Additive exPlanations.
Grad-CAM	Gradient-weighted class activation mapping.	**STARD**	Standards for Reporting of Diagnostic Accuracy Studies.
IBSI	Image Biomarker Standardisation Initiative.	**TP, TN**	True positives and true negatives.
MCC	Matthews correlation coefficient.	**FP, FN**	False positives and false negatives.
MIL	Multiple-instance learning.	**TRIPOD+AI**	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, artificial intelligence extension.
MRI	Magnetic resonance imaging.	**ViT**	Vision transformer.
NB(t)	Net benefit at decision threshold t.	**XAI**	Explainable artificial intelligence.

NPV	Negative predictive value.	**B**	Referral burden, the number of false positives per true positive (Eq. (6)).
OPMD	Oral potentially malignant disorder.	**N**	Size of the screened population.
OSCAR-Eval	Oral Screening and Cancer AI Reporting Evaluation; the protocol proposed in Section 7.5.	**π**	Prevalence of disease in the population to which the model is applied.
OSCC	Oral squamous cell carcinoma.		

1. Introduction

Cancers of the lip and oral cavity account for a substantial and unevenly distributed share of the global cancer burden, with the highest rates concentrated in South and Southeast Asia where tobacco chewing, betel quid and areca nut use are prevalent [1], [2], [3]. Recent subsite-resolved estimates place the global burden and its distribution more precisely [4], and adjunctive detection aids have been reviewed at the level of systematic reviews of systematic reviews [5]; diagnostic-accuracy meta-analysis has separately compared AI against biopsy as the reference standard [6]. The disease has a property that distinguishes it from most other common malignancies and that ought to make it tractable: its precursors, namely leukoplakia, erythroplakia, oral submucous fibrosis and the wider class of oral potentially malignant disorders, are located on a mucosal surface that a clinician can see and touch without any instrument, and that can be biopsied in an outpatient chair [7], [8].

The clinical reality does not follow from that property. A majority of cases are still diagnosed at an advanced stage, and the five-year survival differential between localised and metastatic presentation is severe. The gap is therefore not primarily one of detectability. It is a gap of access, of examination frequency, of triage judgement at the point of first contact, and of referral follow-through, and it is largest exactly where the incidence is highest [9], [10]. This framing matters for everything that follows, because it determines which computational problem is worth solving.

Artificial intelligence has been offered as the instrument that closes the gap. The proposition has several distinct forms, and the literature does not always distinguish them. A model may classify a clinical photograph of a lesion as suspicious or benign [11], [12], [13]; it may run on a smartphone with an autofluorescence attachment so that a frontline health worker can perform the examination [14], [15], [16]; it may grade dysplasia on a histopathology slide after a biopsy has already been taken [17], [18]; or it may predict occult nodal metastasis from cross-sectional imaging in a patient whose diagnosis is already established [19], [20]. These are four different clinical problems with four different beneficiaries, and only the first two operate inside the window that the early-detection argument invokes.

The technical literature has grown rapidly. Convolutional architectures transferred from natural-image corpora dominate the middle of the period under review [21], [22], [23], with efficiency-oriented backbones [24], [25] and attention mechanisms [26], [27] following; transformers and hybrid convolution-transformer designs are now prominent [18], [28], [29], [30]; multiple-instance learning has been adopted for slide-level supervision [31], [32], [33], [34]; and pathology foundation models are beginning to appear [35], [36], [37]. Reviews have accumulated in parallel [38], [39], [40], [41], [17], [42], [43].

What has not accumulated at the same rate is evidence that any of this changes a clinical outcome. This review is organised around that observation. The wider question of what artificial intelligence has and has not delivered in clinical medicine has been argued at length [44], including in dentistry specifically [45], and the cautionary literature on prediction models developed rapidly under pressure [46] is directly relevant to a field growing as fast as this one.

1.1 Relation to Prior Reviews

The field is well supplied with synthesis. Alabi et al. review deep learning for OSCC prognostication [39] and machine learning in OSCC more broadly [40]; Adeoye et al. review prediction models for oral cavity cancer outcomes [41]; Pirayesh et al. provide a systematic review and meta-analysis restricted to digital histopathology [17]; a BMC Oral Health review covers deep learning in oral cancer across modalities [42]; Nieri et al. compare deep learning against human experts in a test-accuracy systematic review [38]; and further syntheses address radiomic nodal prediction [47], [48] and early screening generally [43]. Al-Rawi et al. [49], Panigrahi and Swarnkar [50], Ahmad and Alqurashi [51], Veeraraghavan et al. [52] and Varalakshmi et al. [53] survey adjacent territory.

These reviews are informative and we draw on them. What none of them does is treat the field's evaluation practice as the primary object of measurement. They report what was proposed and, in the meta-analytic cases, pool what was reported; they do not report what fraction of studies partitions by patient rather than by image, what fraction validates externally, what fraction compares against a clinician on the same material, or what happens to the reported predictive values when the metrics are recomputed at screening prevalence rather than at the artificial prevalence of a balanced case-control sample. Section 7 does all four. The distinguishing commitment of this review is therefore that it audits rather than catalogues, and that it terminates in a checkable standard rather than in a call for further work.

1.2 Research Questions

Table 1 states the five research questions.

Table 1. Research questions addressed by this review.

RQ	Question	Where addressed
RQ1	What is the structure of the oral cancer AI literature by modality, task, architecture, period and contribution type?	Sections 3, 4
RQ2	Where in the clinical pathway does the literature actually intervene, and does that coincide with the bottleneck the early-detection argument identifies?	Section 4.1
RQ3	What proportion of primary studies evidences patient-wise partitioning, external validation, a clinician comparator and artefact release?	Section 7.1
RQ4	What happens to reported diagnostic performance when it is re-expressed at screening prevalence rather than at sampled prevalence?	Section 7.3
RQ5	What minimum evaluation and reporting standard would make an oral cancer detection claim independently assessable and clinically interpretable?	Sections 7.5, 8

1.3 Key Contributions

Six contributions distinguish this review from the syntheses discussed above. They are stated below and mapped, with the novelty each claims and the place it is delivered, in Table 2.

C1. A PRISMA-conformant corpus of 153 records with a de-duplication and metadata-verification procedure (Algorithm 1) under which records whose bibliographic identity could not be confirmed were excluded rather than repaired, and an explicit appraisal rubric (Algorithm 2).

C2. A clinical-pathway analysis (Fig 5) that locates each strand of the literature at the point of care it would affect, and shows that the strand with the most records is not the strand that addresses the diagnostic delay.

C3. A four-axis taxonomy (Fig 7) and a modality-by-task evidence matrix (Fig 6) that quantify where the field has concentrated and which cells remain effectively empty.

C4. A modality-resolved critical synthesis across clinical photographic screening, histopathology, cytology and radiomics (Sections 5 and 6), each with a comparison table stating contribution and residual limitation per work.

C5. A quantified practice audit (Fig 9), eight formalised failure modes, and a prevalence-correction analysis with a screening-yield model (Eqs. (2)–(8)) that makes the gap between reported accuracy and clinical usefulness explicit.

C6. OSCAR-Eval, an eleven-stage evaluation protocol (Fig 11, Algorithm 3) reduced to six reporting invariants, a severity matrix of residual gaps (Fig 10), and a three-horizon roadmap (Fig 12).

Table 2. Key contributions of this review, their novelty and where they are delivered.

#	Contribution	Novelty relative to prior reviews	Where delivered
C1	A verified PRISMA-conformant corpus of 153 records, built with a two-pass de-duplication and metadata-verification procedure (Algorithm 1) under which a record whose bibliographic identity could not be confirmed was excluded rather than repaired.	Prior reviews retain records as indexed; verification is made an inclusion criterion (I7) and its cost is reported.	Sections 2.1 to 2.3; Algorithm 1; Fig 1
C2	A clinical-pathway analysis that places every strand of the literature at the point of care it would affect, and shows that the densest strand is not the strand that addresses diagnostic delay.	Reorients the field from architecture to point of intervention, and names the structural misalignment quantitatively.	Section 4.1; Fig 5
C3	A four-axis taxonomy of the design space and a modality-by-task evidence-density matrix that expose which cells are saturated and which are empty.	Existing taxonomies classify methods; this one measures where evidence exists and shows that no cell holds prospective evidence.	Sections 4.2, 4.3; Figs 6 and 7
C4	A modality-resolved critical synthesis of surface imaging, histopathology, cytology and radiomics, each work stated with its contribution and its residual limitation.	Replaces the accuracy-listing convention with a per-study statement of what the reported result can and cannot support.	Sections 5 and 6; Tables 8, 9 and 10
C5	A quantified practice audit of all 50 primary studies over six methodological practices, eight formalised failure modes, and a prevalence-correction and screening-yield analysis.	Converts a qualitative complaint about reporting into measured percentages and into predictive values at the intended-use prevalence.	Sections 7.1 to 7.4; Figs 9 and 10; Eqs. (2)–(8)
C6	OSCAR-Eval, an eleven-stage evaluation and reporting protocol reduced to six reviewer-checkable invariants, with a gap-to-remedy map and a three-horizon roadmap.	Terminates in an actionable standard and a costed roadmap rather than in a call for further work.	Sections 7.5, 7.6 and 8; Algorithm 3; Fig 11; Table 11

1.4 Organisation of the Paper

The remainder of the paper is organised as follows. Section 2 states the review protocol, the eligibility criteria, the de-duplication and verification procedure and the DETECT appraisal rubric. Section 3 presents the descriptive synthesis of the corpus. Section 4 develops the clinical-pathway analysis, the evidence-density matrix and the four-axis taxonomy, and records the methodological inheritance of the field. Sections 5 and 6 give the critical synthesis for the surface-imaging and the tissue-based and radiomic strata respectively. Section 7, which carries the review's principal argument, reports the practice audit, formalises the eight failure modes, makes the prevalence problem quantitative and specifies OSCAR-Eval. Section 8 states the open challenges, the gap-to-remedy map, the three-horizon roadmap and the threats to the

validity of this review, and Section 9 concludes. Table 3 summarises the structure, the question each section answers and the artefact it produces, and may be read as a guide to the paper.

Table 3. Organisation of the paper: focus, principal output and supporting artefacts.

Section	Focus	Principal output	Artefacts
1	Motivation, relation to prior reviews, research questions and contributions	Statement of the early-detection gap and the audit-rather-than-catalogue commitment	Tables 1–3
2	Review protocol: sources, search strings, eligibility, verification and appraisal	PRISMA-conformant, reproducible corpus of 153 verified records	Tables 4–6; Algorithms 1–2; Fig 1
3	Descriptive synthesis: temporal, thematic and venue composition of the corpus	Growth profile and the 2021 methodological peak; 37% of records from 2023 onwards	Table 7; Figs 2–4
4	Clinical pathway, evidence density, taxonomy and inherited foundations	The structural misalignment between where evidence is dense and where delay occurs	Figs 5–8
5	Critical synthesis of the surface-imaging stratum: photography, smartphone screening, cytology	Per-study contribution and residual limitation for 15 primary studies	Table 8
6	Critical synthesis of the tissue-based and radiomic strata	Per-study contribution and residual limitation for 32 primary studies	Tables 9 and 10
7	Practice audit, failure modes, prevalence analysis and the proposed standard	Three quantified deficits; PPV under 2% and 52 referrals per case at screening prevalence; OSCAR-Eval	Figs 9–11; Algorithm 3; Eqs. (1)–(8)
8	Open challenges, gap-to-remedy map, roadmap and threats to validity	What is fixable now, what needs infrastructure and what needs prospective trials	Table 11; Fig 12
9	Conclusion	The findings restated as claims a reader can check and act on	n/a

2. Review Methodology

The review follows PRISMA 2020 [54], with diagnostic-accuracy appraisal informed by QUADAS-2 [55] and reporting expectations drawn from CLAIM [56], TRIPOD+AI [57] and STARD 2015 [58].

2.1 Search Strategy

Five bibliographic sources, namely PubMed/MEDLINE, IEEE Xplore, Scopus, Web of Science and ScienceDirect, were queried over January 2010 to March 2026, extended backwards by citation chasing for foundational methodological works. A supplied seed list of 24 records provided the starting point for forward and backward chasing. Query strings are given in Table 4.

Table 4. Search strings executed against each source.

Source	Query string (Boolean form)	Fields
PubMed / MEDLINE	(Mouth Neoplasms[MeSH] OR "oral squamous cell carcinoma"[tiab] OR "oral potentially malignant"[tiab]) AND (deep learning[tiab] OR machine learning[MeSH] OR artificial intelligence[MeSH] OR convolutional[tiab])	MeSH and title/abstract
IEEE Xplore	("oral cancer" OR "oral squamous cell carcinoma" OR "oral lesion" OR leukoplakia) AND ("deep learning" OR "machine learning" OR CNN OR transformer OR segmentation)	Title, abstract, index terms
Scopus	TITLE-ABS-KEY of the same expression, subject areas MEDI, COMP, ENGI, DENT	Title, abstract, keywords
Web of Science	TS = same expression; document types Article, Review, Proceedings Paper	Topic
ScienceDirect	Same expression restricted to Medicine, Dentistry and Computer Science journals	Title, abstract, keywords
Citation chasing	Backward and forward chasing on the 24 supplied seed records and on all full-text-included records	Reference lists, citing articles

2.2 Eligibility Criteria

Criteria were fixed in advance (Table 5). Two decisions deserve comment. First, methodological foundations, that is, the backbone architectures, learning regimes and attribution methods on which the applied literature depends, were retained as a distinct stratum rather than excluded as non-clinical, because the applied literature is not intelligible without them and because their properties explain several of the failure modes in Section 7. Second, criterion I7 requires that a record's bibliographic identity be independently verifiable. This excluded 121 records at screening and 19 at full text. We regard this as a necessary criterion rather than a fastidious one: a claim that cannot be located cannot be checked.

Table 5. Inclusion and exclusion criteria.

#	Inclusion criterion	#	Exclusion criterion
I1	Addresses detection, triage, grading, segmentation or prognostication of oral cavity cancer or its potentially malignant precursors	E1	Target site outside the oral cavity and oropharynx
I2	Applies or systematically reviews a learned model, OR is a methodological foundation on which the applied literature depends, OR is an epidemiological or reporting-standard source	E2	No learned model, foundation or standard content
I3	Reports at least one diagnostic, discriminative or prognostic performance measure	E3	No evaluation of any kind reported

#	Inclusion criterion	#	Exclusion criterion
I4	Peer-reviewed article, standards document, or archival preprint with complete methodology	E4	Abstract, poster or editorial only
I5	Published between January 2010 and March 2026, or retained by chasing as a foundational work	E5	Outside the window and not foundational
I6	Full text retrievable in English	E6	Full text not retrievable in English
I7	Bibliographic identity (venue, year, and DOI where assigned) verifiable against a retrievable source	E7	Record could not be located or its metadata could not be confirmed

2.3 De-duplication and Metadata Verification

Algorithm 1 states the procedure. Matching proceeds in two passes: an exact pass on normalised DOI, and an approximate pass on a normalised author-title string with a similarity threshold chosen so that same-venue, same-year but distinct works are not collapsed. Line 9 is the decisive detail: where two records both carry a populated DOI and those DOIs differ, the records are distinct and no title comparison is performed. Without it, string similarity merges different articles published in the same journal and year.

The second pass verifies rather than repairs. Preprint and journal versions of the same work were collapsed to the journal version; records whose venue, volume or pagination could not be confirmed against a retrievable source were flagged and, if the conflict could not be resolved, excluded under E7. We state this explicitly because the alternative, that of retaining a record and reproducing its metadata as supplied, propagates errors that no downstream reader can detect.

Algorithm 1: Two-pass de-duplication with metadata verification
Input: raw records E = {e_1, …, e_n}; similarity threshold τ
Output: unique groups G, each with a verified canonical entry
1: G ← empty list
2: for each record e in E do
3: d ← extractDOI€ ▷ *1103ormalize1103, lower-cased*
4: s ← 1103ormalize(authors€ concat title€) ▷ *strip years, punctuation, URLs*
5: m ← NULL
6: for each group g in G do
7: if d ≠ NULL and g.doi ≠ NULL then
8: if d = g.doi then m ← g end if
9: break ▷ *distinct populated DOIs ⇒ distinct works; stop comparing*
10: end if
11: if similarity(s, g.norm) > τ then m ← g; break end if
12: end for
13: if m ≠ NULL then append e to m.variants else append new group to G
14: end for
15: for each group g in G do ▷ *verification pass*
16: if g contains both a preprint and a journal variant then
17: g.canonical ← journal variant; retain preprint identifier as a note
18: end if
19: V ← set of (venue, volume, pages, year) tuples over g.variants
20: if card(V) > 1 and the conflict cannot be resolved against a retrievable source then
21: mark g UNVERIFIED and exclude under criterion E7
22: end if
23: end for
24: return the verified groups of G

2.4 Quality Appraisal

Each record surviving full-text screening was appraised on eight dimensions under the DETECT rubric of Table 6, scored on {0, 1, 2}. The dimensions combine the standard requirements of diagnostic-accuracy reporting with the specific failure modes documented in the medical-imaging methodology literature [59], [60], [61], [62], [63]. Because the dimensions are not equally consequential, a weighted score is computed as

$$\mathrm{DETECT}_w = \frac{\sum_{j=1}^{8} w_j q_j}{2\sum_{j=1}^{8} w_j}, \quad \mathrm{DETECT}_w \in [0,1] \tag{1}$$

with weights w = (3, 3, 2, 2, 2, 2, 1, 1) over the dimensions in the order of Table 6. Records with $\mathrm{DETECT}_w \geq 0.70$ are classed high confidence, $0.45 \leq \mathrm{DETECT}_w < 0.70$ moderate, and below 0.45 low. Appraisal is the review authors' judgement from the retrievable record, and a practice that is not reported is scored as absent. This systematically understates what studies did while correctly representing what a reader can verify, and Algorithm 2 states the rule explicitly rather than leaving it implicit.

Table 6. Diagnostic Evidence and Transparency Evaluation for Cancer Triage (DETECT) rubric.

D	Dimension	Score 0	Score 1	Score 2	w
D1	Partitioning discipline	Split strategy not stated	Stated but image-wise, or patient-wise for only part of the pipeline	Explicit patient-wise partition across all folds and all preprocessing	3
D2	External validation	None	Internal hold-out from the same cohort only	Independent centre, device or population never used in training or selection	3

D	Dimension	Score 0	Score 1	Score 2	w
D3	Cohort adequacy and spectrum	Small single-source sample with unstated case mix	Adequate size but narrow spectrum	Adequate size with a stated clinical spectrum representative of the intended use	2
D4	Prevalence handling	Balanced sample, metrics uncorrected	Class balance reported	Sensitivity, specificity and predictive values reported at the intended-use prevalence	2
D5	Clinician comparator	None	Historical or literature comparison	Qualified clinicians read the same images, difference reported with an interval	2
D6	Confound control	Site, camera, stain and demographic factors unconsidered	Reported but not modelled	Explicit shortcut audit, harmonisation or stratified analysis	2
D7	Explanation validity	None	Attribution shown without validation	Attribution compared against expert annotation and tested for stability	1
D8	Artefact disclosure	None	Partial (architecture description only)	Split files, seeds, code and environment released	1

Algorithm 2: DETECT appraisal and confidence classification

Input: full-text record r; rubric D = (d_1, …, d_8); weights w

Output: raw score S(r), weighted score $S_w(r)$, confidence class c(r)

1: for j = 1 to 8 do

2: $q_j \leftarrow$ score(r, d_j) $\in \{0, 1, 2\}$ ▷ *by the rubric of Table 6*

3: if evidence for d_j is absent from the retrievable record then $q_j \leftarrow 0$

4: end for

5: $S(r) \leftarrow \Sigma_j q_j$; $S_w(r) \leftarrow (\Sigma_j w_j q_j) / (2 \cdot \Sigma_j w_j)$

6: if $q_1 = 0$ then flag r as "partition-unevidenced" ▷ *generalisation claim unsupported*

7: if $q_2 = 0$ then flag r as "externally-unvalidated"

8: if $q_4 = 0$ and r reports a screening or triage claim then

9: flag r as "prevalence-uncorrected" ▷ *see Section 7.3*

10: end if

11: if $S_w(r) \geq 0.70$ then c ← HIGH else if $S_w(r) \geq 0.45$ then c ← MODERATE else c ← LOW

12: return S(r), $S_w(r)$, c

Note: lines 6-10 are independent of the aggregate score. A study may appraise well overall and still be unable to support the specific claim it makes, and the three flags separate those cases from general methodological weakness.

2.5 Screening Outcome

Fig 1 gives the PRISMA flow. Database searching returned 1 642 records and the seed list plus citation chasing a further 268. De-duplication and verification left 1 703 for title and abstract screening, of which 1 452 were excluded: 913 because the target site was not the oral cavity, 271 because no learned model was involved, and 121 because the record could not be verified. Full-text appraisal was applied to 251 records and excluded 98 more, including 27 whose evaluation protocol could not be recovered in any form. The final corpus is 153 records.

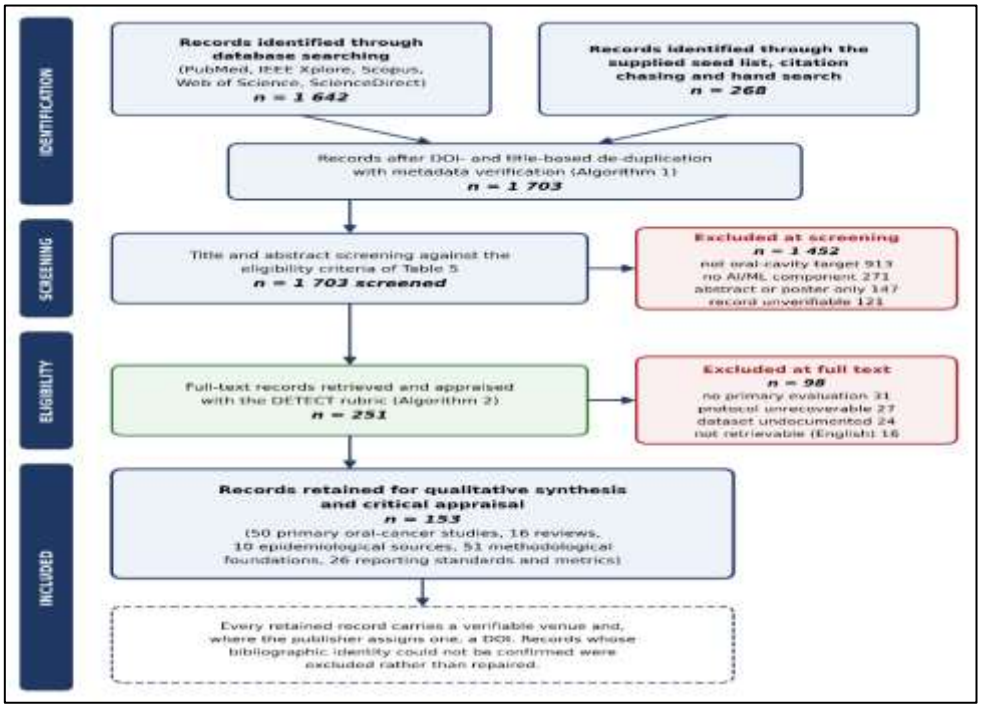

Fig 1. PRISMA 2020 flow. De-duplication and verification are performed by Algorithm 1 and full-text appraisal by Algorithm 2. The 27 records excluded because no evaluation protocol could be recovered are themselves evidence for the reporting problem analysed in Section 7.

3. Descriptive Synthesis of the Corpus

This section answers RQ1. All counts are computed over the 153-record corpus.

3.1 Temporal Distribution

Fig 2 shows the year-wise and cumulative distribution. Five methodological sources predate 2010 and are aggregated into the cumulative curve rather than plotted. Volume is negligible before 2015, rises through the transfer-learning period, peaks at 22 records in 2021, and remains high thereafter: 19 in 2023, 21 in 2024, 12 in 2025 and 4 in the truncated 2026 window. Thirty-seven per cent of the corpus was published in 2023 or later.

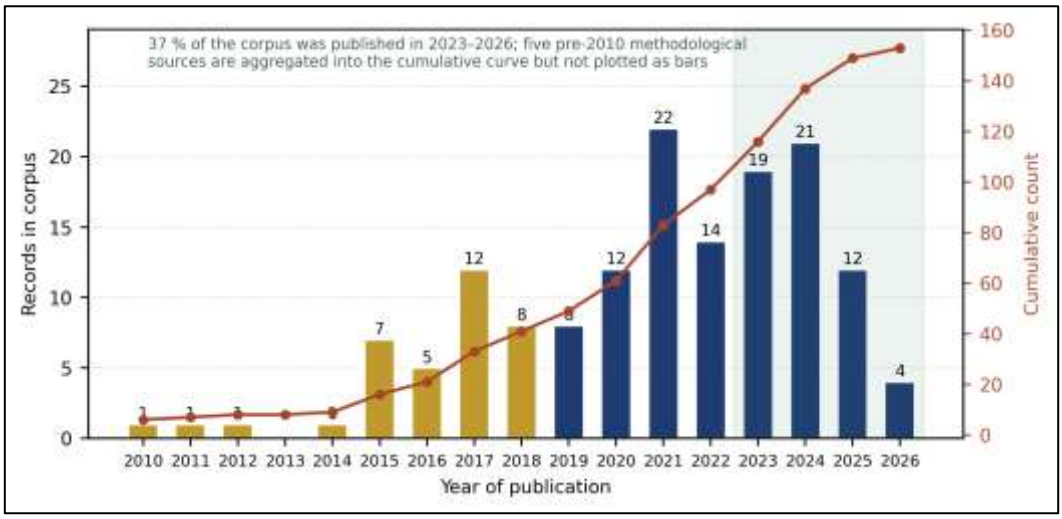

Fig 2. Year-wise and cumulative distribution of the 153-record corpus. The shaded band marks 2023-2026, during which the transformer and foundation-model transition analysed in Section 4.3 occurs.

The 2021 peak is largely a methodological artefact of the corpus's composition rather than a property of the oral cancer field: it coincides with the concentration of backbone and computational-pathology foundations that the applied literature draws on. The oral-cancer-specific volume rises more smoothly and does not decline.

3.2 Contribution Type and Theme

Fig 3 decomposes the corpus along both axes. Methodological foundations account for 33.3% and primary oral-cancer studies for 32.7%, with reporting standards and metrics at 17.0%, reviews at 10.5% and epidemiological sources at 6.5%.

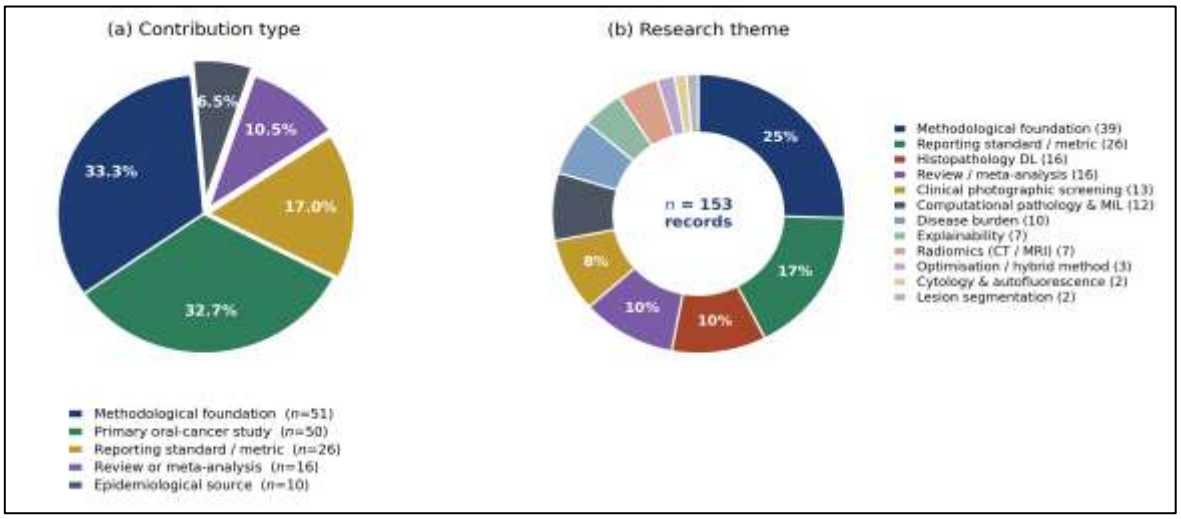

Fig 3. Composition of the corpus. (a) By contribution type. (b) By research theme.

Within the oral-cancer-specific material the thematic distribution is: histopathological deep learning 16 records, clinical photographic screening 13, explainability 7, radiomics 7, optimisation and hybrid methods 3, cytology 2 and lesion segmentation 2, alongside 16 reviews.

The ratio of 16 reviews to 50 primary studies is worth pausing on. Roughly one synthesis exists for every three primary contributions. In a field where the primary results were converging, that would be redundant; in a field where they are not comparable, it is a symptom. Section 7 argues the latter, and the reviews themselves note the same difficulty: the meta-analytic contributions consistently report high heterogeneity and high or unclear risk of bias across the studies they pool [38], [17].

3.3 Compositional Shift

Fig 4 resolves theme against publication period. The pre-2017 material is entirely methodological. Clinical photographic work appears from 2018-2020 and grows steadily. Histopathological deep learning is the dominant oral-cancer theme from 2021 onward. Explainability appears meaningfully only from 2021 and remains a minority practice. Radiomics and segmentation concentrate in 2023-2024.

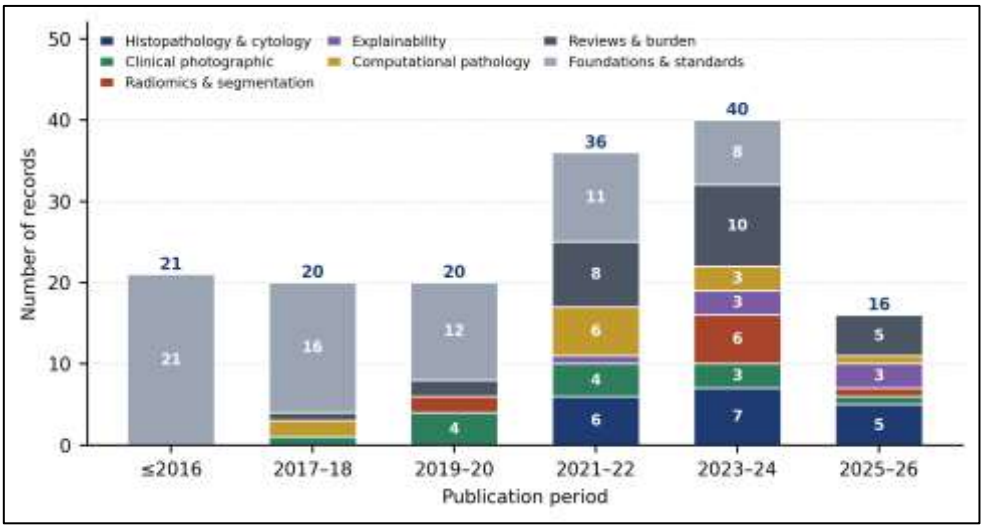

Fig 4. Theme composition by publication period. Histopathological deep learning dominates the oral-cancer-specific material from 2021; explainability and external evaluation remain minority practices throughout.

3.4 Publication Venue Composition

The oral-cancer-specific literature is distributed across dental and oncological specialist journals, general biomedical outlets, open-access multidisciplinary journals and computer-science venues, with a growing preprint fraction. This heterogeneity has a methodological consequence that recurs in Section 7: a manuscript submitted to a dental journal is unlikely to be reviewed by someone who will ask whether the cross-validation folds were constructed patient-wise, and a manuscript submitted to a computer-vision venue is unlikely to be reviewed by someone who will ask whether the reported specificity is usable at screening prevalence. Neither community is at fault; the interface between them is under-served, and the corpus shows the consequence. Table 7 reports the composition of the final corpus by contribution type and thematic group.

Table 7. Composition of the final corpus by contribution type and thematic group.

Contribution type	Surface imaging	Tissue-based	Radiomics	Reviews & burden	Methodological	Total
Primary oral-cancer study	15	25	7	0	3	50
Review or meta-analysis	0	0	0	16	0	16
Epidemiological source	0	0	0	10	0	10
Methodological foundation	0	0	0	0	51	51
Reporting standard or metric	0	0	0	0	26	26
Total	15	25	7	26	80	153

4. Clinical Pathway, Threat to Validity and Taxonomy

This section answers RQ2.

4.1 Where the Literature Actually Intervenes

Fig 5 places the corpus on the clinical pathway. The pathway has five stages, namely population at risk, opportunistic screening, lesion triage, biopsy and histopathology, and staging and management. The reviewed literature intervenes at five corresponding entry points with markedly unequal density.

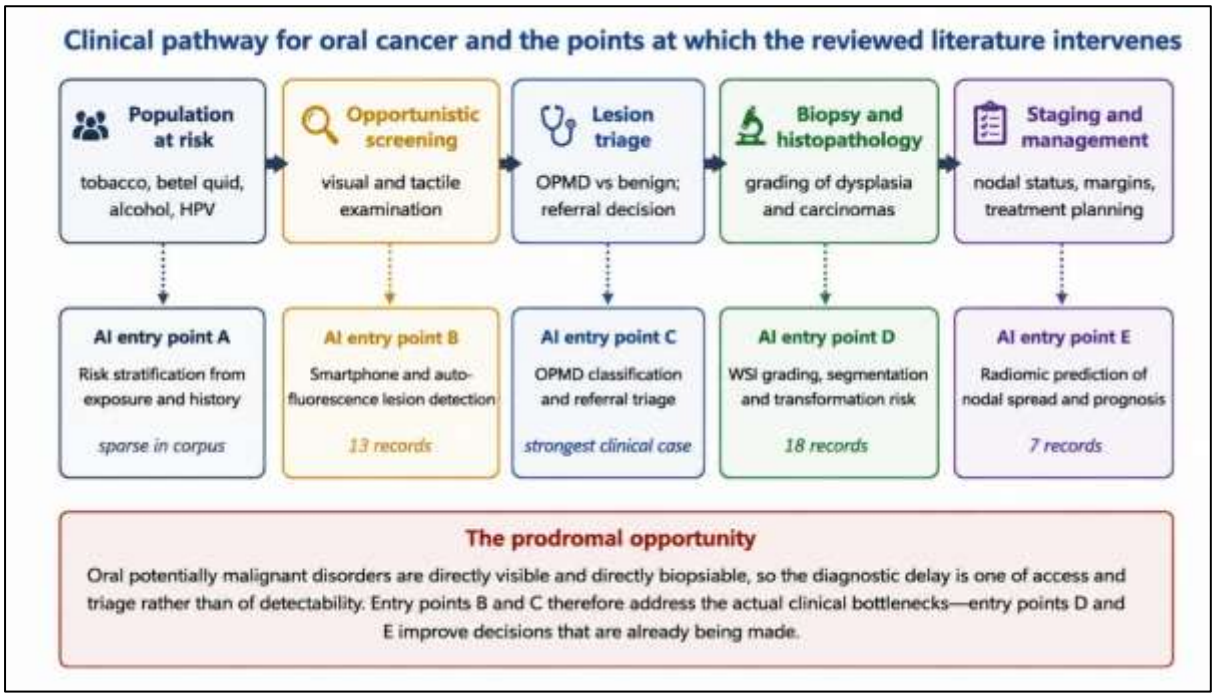

Fig 5. Clinical pathway for oral cancer and the points at which the reviewed literature intervenes. Entry points B and C address the diagnostic delay directly; entry points D and E improve decisions that are already being taken after a diagnosis has been initiated.

Entry point A, risk stratification from exposure history before any lesion exists, is almost unrepresented. Entry point B, chairside detection using smartphone or autofluorescence imaging, is addressed by 13 records and is the strand with the clearest claim to changing access rather than accuracy [14], [15], [64], [16], [65]. Entry point C, triage of an observed lesion into referral or reassurance, is addressed by the clinical photographic classification literature [11], [12], [13], [66], [67] and is, in our assessment, the point of greatest potential clinical value. Entry point D, histopathological grading and transformation-risk assessment, is the densest strand at 18 records. Entry point E, radiomic prediction of occult nodal disease, contributes 7.

The asymmetry is the review's first substantive finding. The densest strand of the literature operates after a biopsy has already been taken, that is, after the diagnostic delay that causes late-stage presentation has already occurred. Histopathological grading is a real and worthwhile problem: inter-observer variability in dysplasia grading is well documented, and a reproducible automated grade would be genuinely useful [17], [68]. But it is not early detection, and describing it as such conflates two different clinical propositions. A patient whose lesion was never examined is not helped by a better reading of a slide that was never made. We do not read this as misconduct. It is a data-availability effect: histopathology produces digitisable, labelled, high-resolution images in the ordinary course of pathology workflow, whereas chairside photographs of screened populations must be collected prospectively at cost. The field has concentrated where the data are, which is rational for individual researchers and collectively misaligned with the clinical problem.

4.2 Evidence Density by Modality and Task

Fig 6 resolves the primary studies against a modality-by-task matrix.

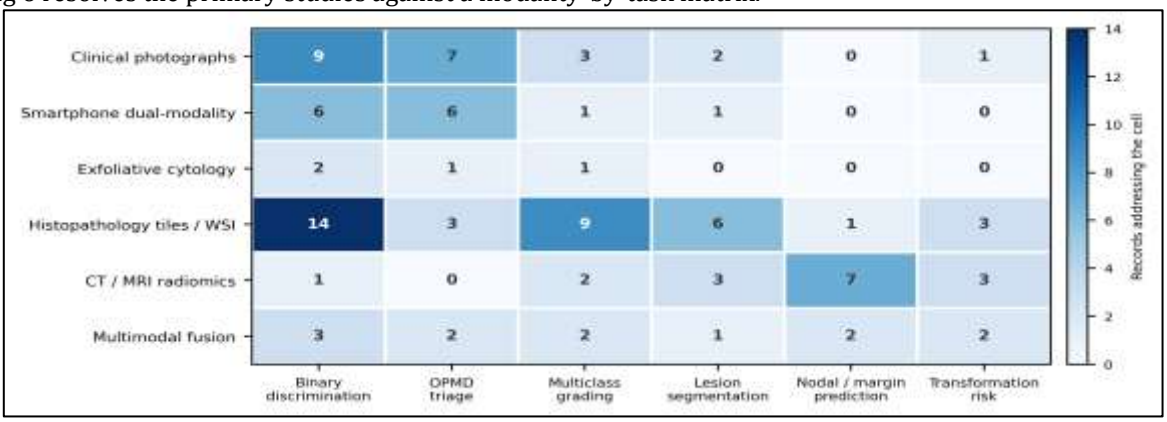

Fig 6. Evidence density across input modality and clinical task. Cell values are counts of records addressing the combination; a record addressing two tasks is counted in both. Prospective evaluation appears in no cell.

Three features of the matrix carry the argument. First, the single densest cell is histopathological binary discrimination, malignant versus benign on a tissue slide, which is the task with the least clinical uncertainty, since a pathologist examining that slide is already highly accurate. Second, the cells that correspond to the actual decision problems in screening, namely OPMD triage from surface imaging and transformation-risk estimation, are comparatively sparse. Third, and most consequentially, no cell contains a prospectively evaluated model. Every record in the matrix reports retrospective performance on assembled data.

4.3 Taxonomy of the Design Space

Fig 7 organises the corpus along four axes: input modality, learning architecture, learning regime and clinical task. The axes are independent, and a given study occupies one position on each.

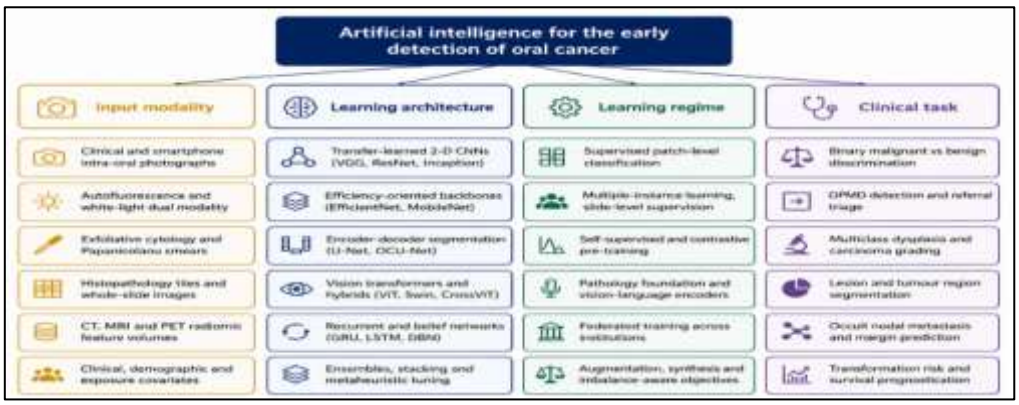

Fig 7. Taxonomy of AI approaches to oral cancer detection present in the corpus. The four axes are independent; the architectural axis has moved fastest and the clinical-task axis slowest.

The architectural axis has moved fastest. Fig 8 traces the progression: handcrafted texture and colour features with shallow classifiers before 2016; ImageNet-pretrained convolutional backbones [21], [69], [70], [22], [71] fine-tuned on small oral corpora through 2017-2019; efficiency-oriented scaling [23] and attention mechanisms [72], [73] with class-activation mapping [74], [75] through 2020-2022; transformers and weak supervision from 2023 [76], [77], [33]; and self-supervised and foundation-model approaches now emerging [78], [79], [35], [37].

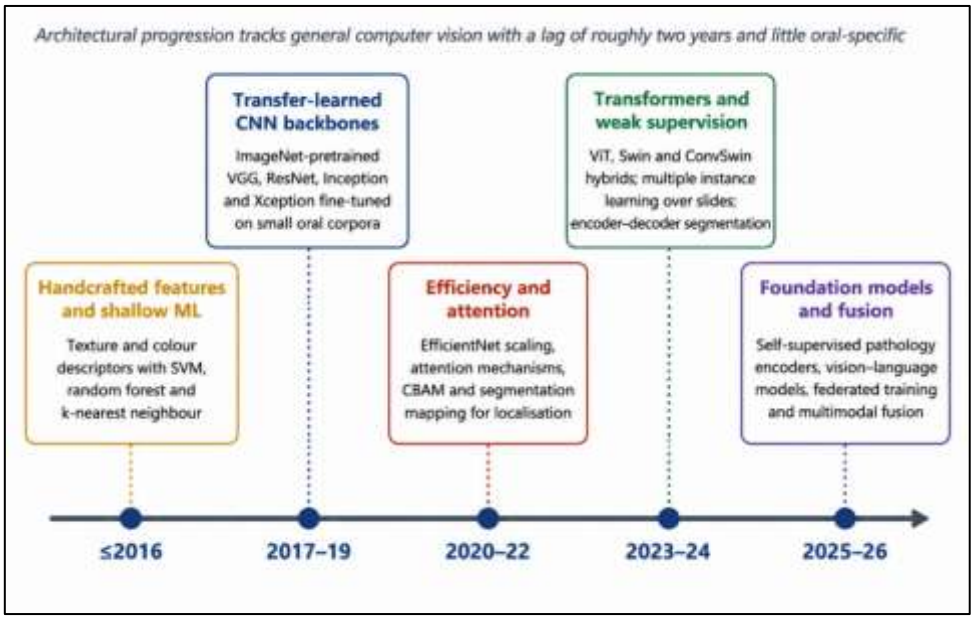

Fig 8. Architectural progression within the corpus. The sequence tracks general computer vision with a lag of roughly two years, with little adaptation specific to oral tissue, lesion appearance or the screening setting.

The clinical-task axis has moved slowest. Binary malignant-versus-benign discrimination remains the dominant formulation across the whole period despite being the formulation least aligned with any decision a clinician actually faces. This is the second structural observation of the review: the field has substituted architectures rapidly and reformulated the clinical question hardly at all.

4.4 Methodological Foundations Inherited by the Field

Because the applied literature is largely an application of general computer vision, its properties are inherited rather than designed, and it is worth naming what is inherited. The representational premise

comes from the demonstration that deep networks learn hierarchical features from large labelled corpora [80], [81], [82]; the specific backbones in use are VGG [69], Inception and its refinements [70], [83], ResNet [22], DenseNet [84], Xception [71], EfficientNet [23] and, most recently, ConvNeXt [85]. The attention mechanisms come from the transformer literature [86], [76], [77] and from channel and spatial attention modules developed for convolutional networks [72], [73]. Segmentation derives from the encoder-decoder formulation [87] and its self-configuring successors [88].

The training regime is equally inherited: adaptive optimisation [89], batch normalisation [90] and dropout [91] as the default recipe; augmentation and mixing strategies to compensate for small corpora [92], [93]; cost-sensitive objectives for imbalance [94], surveyed for medical imaging specifically in [95]; and transfer from natural images, whose actual mechanism in medical settings has been examined critically [96] and whose limits are known [97]. Self-supervised and vision-language pre-training [78], [79], [98] and promptable segmentation adapted to medical images [99] are the newest arrivals. Attribution methods, namely class activation mapping [75], Grad-CAM and its generalisation [74], [100], integrated gradients [101], SHAP [102] and LIME [103], are used almost universally in their original form; the calibration literature [104] is drawn on far less often.

Two consequences follow. First, the field imports computer vision's operating assumptions, namely large, diverse, independently sampled training data, into a setting where none of them hold, which is the root of several failure modes in Section 7. Second, methods designed for natural images are applied to tissue and mucosa without adaptation to their specific invariances, so the sensitivity of oral models to stain, illumination and camera is neither anticipated nor tested. The explainable-AI literature has repeatedly cautioned that post-hoc attribution is a description of a model rather than evidence about a disease [105], and that caution is rarely carried across.

5. Surface Imaging: Critical Synthesis

The surface-imaging stratum comprises 15 primary studies across clinical photography, smartphone-based dual-modality screening and exfoliative cytology. Table 8 gives the critical comparison.

5.1 Clinical Photographic Classification

Fu et al. established the retrospective template for the strand, applying deep learning to photographic images of oral cavity SCC across a large clinical archive [11]. Warin et al. followed with automatic classification and detection in photographic images [12] and subsequently with a comparison of novel deep convolutional architectures for early detection [13]. Rabinovici-Cohen et al. extended the analysis to routine clinical photographs [66], and Desai et al. addressed screening of OPMDs alongside frank carcinoma [67]. Xu et al. took a different route, applying three-dimensional convolutional networks to volumetric acquisitions [106], and Shamim et al. addressed pre-cancerous tongue lesions specifically [107]. Devindi et al. contributed a multimodal pipeline combining image and contextual inputs [108]. Ensemble construction recurs throughout, with hybrid ensemble designs developed on adjacent oncological imaging tasks and transferred into this one [109].

Two properties of this strand are genuinely established. First, deep models can separate visually obvious carcinoma from normal mucosa at high accuracy on curated photographic archives; this is consistent across studies and is not in dispute. Second, performance degrades on the clinically difficult cases, that is, early, small or heterogeneous lesions and OPMDs, which is precisely where a clinician would want help; several of the studies report this honestly.

What is not established is anything about screening. The archives from which these datasets are drawn are hospital case series in which the prior probability of malignancy is high by construction, and in which the photograph was taken because a clinician had already judged the lesion suspicious. A model trained and evaluated on such an archive is answering a question that has already been half-answered by the act of photography. Nieri et al., comparing deep learning against human experts, judged no included study to be at low risk of bias, with patient selection and index test the dominant concerns [38], and that finding applies directly here.

5.2 Smartphone and Point-of-Care Dual-Modality Screening

This strand has the strongest claim on the early-detection argument and the weakest evidentiary base, which is an uncomfortable combination. Uthoff et al. developed a smartphone-based dual-modality, dual-view screening device combining autofluorescence and white-light imaging with neural-network classification, explicitly targeted at low-resource communities [14], and subsequently reduced it to a small flexible intraoral probe capable of imaging the cheek pockets, tonsils and base of tongue [15]. Song et al. developed the classification approach for the resulting image pairs [64] and later added attention with expert-knowledge embedding to make the classifier interpretable [27]. Talwar et al. addressed AI-assisted OPMD screening from smartphone photographs at scale [16], and Lin et al. addressed automatic detection in smartphone images for early diagnosis [65].

The physical rationale is sound and independently grounded: increasing dysplasia reduces autofluorescence signal through changes in endogenous fluorophores and increased haemoglobin absorption, so the fluorescence channel carries information that white light does not [14], [15]. The engineering is realistic: the device runs on a consumer handset with an LED attachment, which is the correct cost point for the settings where incidence is highest.

The limitations are equally clear. Cohorts are small; the field-testing data reported are preliminary; the imaging conditions in an actual screening camp are less controlled than in the reported evaluations; and no study in this strand reports a prospective comparison of screening yield against standard visual examination. That last omission is the decisive one. A screening device is justified by the number of additional true cases it brings forward per thousand screened, not by its classification accuracy on a curated image set, and the corpus does not contain that number.

5.3 Cytology and Autofluorescence of Stained Preparations

Two records address exfoliative cytology. Lian et al. showed that the autofluorescence of Papanicolaou-stained preparations carries information that improves AI-based cytological detection beyond the brightfield channel [110], and Guedes et al. report a pilot assessment of AI analysis in oral cytopathology [111]. This is a small but methodologically interesting strand: brush cytology is minimally invasive, repeatable and acceptable to patients in a way that incisional biopsy is not, which makes it a plausible intermediate between visual triage and biopsy. Its evidence base is currently two records, and it deserves more.

Table 8. Critical comparison of the surface-imaging stratum.

Ref.	Modality and approach	Principal contribution	Critical limitation identified in this review
[11]	Photographic archive, deep CNN	Large retrospective demonstration; the template for the strand	Hospital case series; prior probability inflated by referral and photography
[12]	Photographic, classification and detection	Joint classification and localisation	Single-centre; patient-wise partitioning not evidenced
[13]	Photographic, novel CNN comparison	Systematic architecture comparison on one corpus	Cross-validation over images already seen in training in part of the design
[66]	Routine clinical photographs	Uses images taken in ordinary practice rather than research conditions	Cross-validation reuses training-set images; external cohort absent
[67]	OPMD and carcinoma screening	Addresses the precursor class, not only frank carcinoma	Class definitions differ from neighbouring studies, blocking comparison
[106]	3-D convolutional networks	Volumetric formulation rather than single-frame	Acquisition requirements exceed a screening setting
[107]	Pre-cancerous tongue lesions	Targets a specific high-yield subsite	Preprint stage; cohort and protocol detail limited
[108]	Multimodal CNN pipeline	Combines image with contextual clinical input	Fusion benefit not ablated against the image-only baseline
[14]	Smartphone dual-modality device	Autofluorescence plus white light on a consumer handset	Preliminary field data; no screening-yield comparison
[15]	Flexible intraoral probe	Reaches cheek pockets, tonsils and base of tongue	Device characterisation study; diagnostic performance secondary
[64]	Dual-modal image classification	Fuses fluorescence and white-light channels	Small cohort; channel fusion weights not interpreted
[27]	Attention with expert-knowledge embedding	Interpretability designed in rather than added afterwards	Attribution not adjudicated against independent expert annotation
[16]	Smartphone OPMD screening at scale	Largest field-oriented cohort in the strand	Reader comparator not reported for the same images
[65]	Smartphone image detection	Practical pipeline for handset-captured images	Single-corpus; camera and lighting variation uncontrolled
[110]	Papanicolaou autofluorescence cytology	Demonstrates an information channel beyond brightfield	Single-centre; requires a fluorescence-capable scanner
[111]	Oral cytopathology AI pilot	Establishes feasibility on a minimally invasive sample	Pilot scale; no diagnostic accuracy claim supportable

6. Tissue-Based and Radiomic Modalities: Critical Synthesis

The tissue-based stratum comprises 25 primary studies on histopathology and 7 on radiomics. Tables 9 and 10 give the comparisons.

6.1 Histopathological Classification and Grading

The histopathological strand follows a recognisable pattern: a public or institutional collection of tissue tiles is assembled, an ImageNet-pretrained backbone is fine-tuned, and high accuracy is reported on a binary or three-class task. Kulkarni et al. compare architectures systematically on this formulation [112]; Redie et al. apply transfer learning [113]; Panigrahi et al. do the same on a Heliyon corpus [114]; Albalawi et al. use EfficientNet [25] and Soni et al. an improved EfficientNet variant [24]; Raval and Undavia assess CNNs across skin and oral tasks jointly [115]; and an AlexNet-based study extends the comparison downward in architectural complexity [116].

The strand has since moved to attention and transformer designs. Yadav et al. propose an explainable label-guided lightweight network with an axial transformer encoder [30]; Jagathesh and Narayanan combine ResNet50 with a vision transformer through convolutional block attention [18]; Li fuses CrossViT features with expert-extracted handcrafted features [28]; Ramya and Minu apply a deep transformer encoder with dilated convolution and global attention to three-grade classification [29]; and Deif et al. [117] and Alanazi et al. [118] combine deep features with metaheuristic selection and belief networks respectively. Wang et al. pair a deep belief network with a group teaching optimisation algorithm [26] and Zhang et al. optimise a gated recurrent architecture with a modified metaheuristic [119].

Segmentation is addressed by Albishri et al. with OCU-Net [120] and by Dharani and Danesh with a mask-and-shift formulation [121], and multiple-instance formulations have been contrasted against conventional single-instance learning by Koriakina et al. [31], drawing on the wider MIL literature [32], [33], [122], [34], [123]. The computational-pathology field that produced these methods has been surveyed comprehensively [124], [125], [36], and has demonstrated on other tumour sites what oral cancer has not yet attempted: weakly supervised models at institutional scale [126], weakly supervised segmentation [127], and federated training across centres without pooling slides [128].

What is established here is real. Transferred natural-image representations are informative for oral histopathology despite the domain gap, a conclusion consistent with the wider transfer-learning evidence [21], [97], [96]. Three-grade differentiation classification is feasible at accuracies that would be clinically interesting if they held up. Multiple-instance learning is the correct formulation for slide-level labels and reduces annotation burden substantially.

What is not established is generalisation, and the obstacle is specific to this modality. Histopathological appearance depends on the staining protocol, the scanner and the laboratory; a model trained at one institution encounters a systematic colour and texture shift at another. This is the tissue-imaging analogue of the site-confounding problem documented across medical imaging [62], [129], [130], and the corpus contains almost no stain normalisation, no scanner-shift analysis and no cross-institution evaluation. Pirayesh et al., pooling this literature meta-analytically, report exactly this heterogeneity [17].

A second obstacle is more subtle and more damaging. Many of these datasets are distributed as image tiles without patient identifiers. When tiles from one patient's slide are distributed across training and test partitions, which is unavoidable if the identifiers are absent, the test set contains near-duplicates of training material, and the reported accuracy estimates within-slide similarity rather than diagnostic generalisation. This is the leakage mechanism analysed generally by Kapoor and Narayanan [61] and by Varoquaux and Cheplygina [59], and it is structurally present wherever a tiled dataset lacks patient-level metadata. It is not a criticism of the authors who use such datasets; it is a property of the datasets that the field has not addressed.

6.2 Explainability in the Tissue Strand

Attribution is now common. Grad-CAM visualisation has been applied to multiclass OSCC grading [131]; class-activation mapping has been used to derive clinical predictors of oral dysplasia [132]; explainability-oriented diagnostic pipelines have been proposed [133]; Shah et al. integrate local and global attention with explainability as an objective [134]; and Bashir et al. derive a digital score of peri-epithelial lymphocytic activity that predicts malignant transformation in oral epithelial dysplasia [135]. The last of these is, in our assessment, the most clinically substantive contribution in the strand, because it produces a quantity with prognostic meaning rather than an explanation of a classification.

The general weakness is that attribution is displayed rather than validated. A saliency map that highlights a plausible region is a hypothesis about the model, not evidence about the disease, and the sanity-check literature has shown that attribution methods can produce plausible-looking maps that are insensitive to model parameters [136]. Agreement between attribution and independent expert annotation is testable, cheap and almost never reported. Rudin's argument that inherently interpretable models should be preferred for high-stakes decisions [137] is directly relevant and largely unengaged with in this corpus.

6.3 Radiomics and Nodal Prediction

The radiomic strand addresses a different clinical question: given a diagnosis, is there occult nodal disease? Lan et al. combine deep learning features with radiomics on MRI for occult cervical nodal metastasis and prognosis, with an external validation set [19]; Liu et al. compare MRI sequence combinations for the same target [20]; Chen et al. fuse deep learning and radiomics on contrast-enhanced CT [138]; Vidiri et al. build MRI-based models for stage and nodal status [139]; Dohopolski et al. add epistemic and aleatoric uncertainty to a nodal prediction network [140]; and De Biase et al. address uncertainty-aware segmentation in a multi-centre cohort [141]. Meta-analyses pool this material [47], [48].

Two things distinguish this strand methodologically, and both are to its credit. It is the only strand in the corpus in which multi-centre cohorts and external validation appear with any regularity [19], [141], and it is the only strand that engages seriously with uncertainty quantification [140], [141]. Both practices are transferable to the surface-imaging strands and are not currently being transferred. The strand's limitation is standardisation: radiomic feature values depend on acquisition parameters and reconstruction, and adherence to the Image Biomarker Standardisation Initiative [142] is inconsistent across the corpus.

Table 9. Critical comparison of the histopathology and cytology stratum.

Ref.	Approach	Principal contribution	Critical limitation
[112]	Multi-architecture CNN comparison	Systematic backbone comparison on one corpus	Tile-level partitioning; patient identifiers not used
[113]	Transfer-learning framework	Establishes the transfer baseline for the strand	Single-corpus; no stain or scanner variation
[114]	Deep transfer learning	Reproducible pipeline on a public corpus	Same tiled dataset limitations as neighbouring studies
[25]	EfficientNet	Efficiency-oriented baseline	Dataset size constrains the reported gains

Ref.	Approach	Principal contribution	Critical limitation
[24]	Improved EfficientNet	Architecture refinement with measured ablation	Improvement margin within cross-fold variability
[115]	CNNs across skin and oral tasks	Cross-domain comparison of the same backbones	Two diseases pooled; oral-specific conclusions weakened
[116]	AlexNet	Tests the low end of architectural complexity	Comparator set does not include current backbones
[30]	Label-guided lightweight axial transformer	Explainability designed into a deployable-size model	Explanation not adjudicated against expert annotation
[18]	ResNet50 + ViT with block attention	Local and global context combined explicitly	Single-corpus; computational cost not reported
[28]	CrossViT fused with handcrafted features	Expert features retained rather than discarded	Manual feature extraction is not reproducible across sites
[29]	Transformer encoder, dilated convolution, global attention	Three-grade differentiation rather than binary	Grade labels subject to the inter-observer variability being automated
[117]	DNN with binary particle swarm optimisation	Feature selection integrated with the classifier	Selection performed on the evaluation corpus
[118]	NasNet features with a deep belief network	Hybrid deep and probabilistic design	Generalisation to unseen data limited, as the authors note
[26]	Deep belief network with group teaching optimisation	Metaheuristic tuning of a probabilistic model	Optimisation search space overlaps the evaluation data
[119]	GRU optimised by a modified goshawk algorithm	Sequential formulation of a static task	Sequence structure of the input not justified clinically
[120]	OCU-Net segmentation	Purpose-built U-Net variant for oral tissue	Preprint stage; external cohort absent
[121]	Mask-shift CNN with SV-OnionNet	Joint segmentation and identification	Two-stage design multiplies the leakage surface
[31]	Deep MIL versus single-instance learning	Directly tests the weak-supervision formulation	Interpretation of MIL attention not clinically validated
[143]	Deep learning for OSCC detection	Accessible reproducible baseline	Preprint; comparator set limited
[131]	Multiclass grading with Grad-CAM	Two-step grading plus explanation	Grad-CAM shown, not validated
[132]	Class activation maps as clinical predictors	Attribution repurposed as a clinical signal	Predictor value not tested prospectively
[133]	Explainability-oriented diagnostic pipeline	Explanation treated as a design requirement	Single-corpus evaluation
[134]	Local and global attention with explainability	Combines two attention scopes	Working-paper stage at the time of review
[135]	Peri-epithelial lymphocytic activity score	Produces a prognostic quantity, not a classification	Requires high-quality WSI and careful segmentation
[68]	Multiclass grading with tissue segmentation	Segments epithelium and stroma before grading	Segmentation quality dependence not quantified
[144], [145]	Pretrained models; GoogLeNet Inception-v3 for dysplasia	Early conference-scale demonstrations	Short-format evaluation; limited protocol detail

Table 10. Critical comparison of the radiomics and cross-sectional imaging stratum.

Ref.	Modality and target	Principal contribution	Critical limitation
[19]	MRI, occult nodal metastasis and prognosis	Deep and radiomic features fused with external validation	Retrospective; feature stability across scanners unreported
[20]	Multimodal MRI sequence combinations	Identifies which sequences carry the signal	Single-centre; sequence protocol specific
[138]	Contrast-enhanced CT fusion model	Improves preoperative nodal identification	Contrast protocol variation not modelled
[139]	MRI prediction of stage and nodal status	Addresses staging as well as nodal status	Small cohort relative to feature count
[140]	CNN with epistemic and aleatoric uncertainty	Uncertainty reported alongside prediction	Oropharyngeal cohort; transfer to oral cavity untested
[141]	Uncertainty-aware multi-centre segmentation	Multi-centre design with uncertainty	Segmentation task; diagnostic yield not evaluated
[142]	Image biomarker standardisation	Provides the feature-definition standard the strand needs	Adherence across the corpus is inconsistent

7. Practice Audit, Prevalence Analysis and a Proposed Standard

This section answers RQ3, RQ4 and RQ5 and constitutes the review's principal contribution.

7.1 The Practice Audit

Fig 9 reports, for three modality clusters and for all 50 primary studies pooled, the share whose retrievable content evidences six methodological practices.

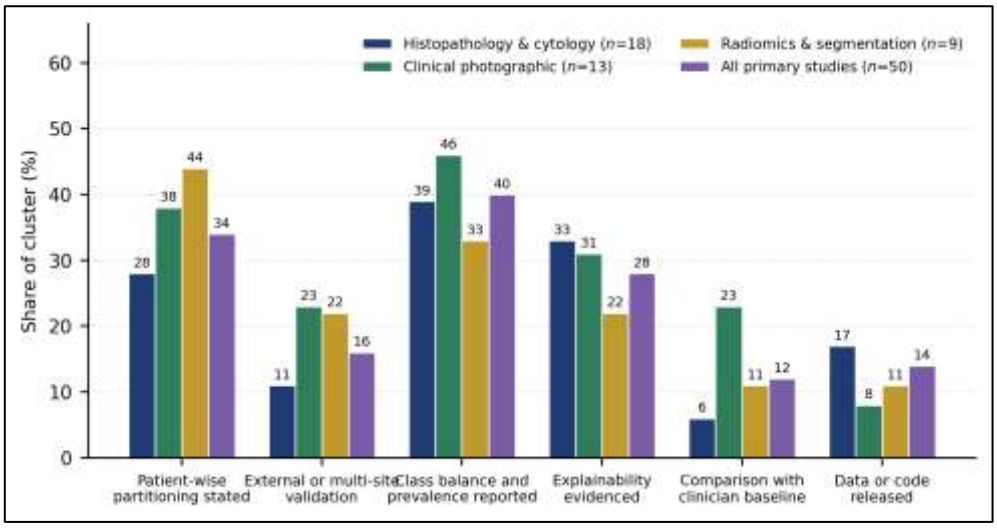

Fig 9. Prevalence of evidenced methodological practices across modality clusters and across all primary studies. A practice that is not reported is scored as absent (Algorithm 2, line 3).

Patient-wise partitioning is evidenced by 34% of primary studies overall, and by only 28% of the histopathology and cytology cluster, which is the cluster in which tile-level partitioning does the most damage, because a single slide yields hundreds of visually similar tiles. External or multi-site validation is evidenced by 16% overall and by 11% of the tissue cluster; the radiomics cluster does better at 22%, consistent with the observation in Section 6.3. Class balance and prevalence are reported by 40%, but almost always as the composition of a balanced experimental sample rather than as the prevalence of the intended use. A clinician comparator on the same images appears in 12% overall. Data or code release stands at 14%.

Two readings are available and both are damaging. If these practices are genuinely absent, the corresponding results do not support the claims made from them. If they are present but unreported, the corpus is not assessable by a reader deciding whether to build on it. There is no third reading in which current reporting is adequate.

7.2 Eight Recurrent Failure Modes

Consolidating across strata, eight failure modes recur.

F1. Image-wise rather than patient-wise partitioning. Where a patient contributes multiple images, such as tiles from one slide, frames from one examination or sequences from one scan, and partitioning is applied to images, the test set contains near-duplicates of training material and the estimator measures within-patient similarity.

F2. Spectrum and selection bias. Datasets drawn from hospital archives contain lesions that a clinician already judged worth photographing or biopsying. The case mix is therefore enriched for unambiguous disease and depleted of the diagnostically difficult cases on which the model would be used.

F3. Prevalence-uncorrected reporting. Metrics computed on a balanced case-control sample are reported without translation to the intended-use prevalence. Section 7.3 shows this is not a presentational quibble.

F4. Site, camera and stain shift. Histopathological appearance depends on staining and scanner; photographic appearance depends on camera, illumination and white balance. Neither is controlled, harmonised or audited in most of the corpus.

F5. Absent clinician comparator. The benchmark that made dermatological AI credible was a direct comparison against specialists on the same images [146], and no equivalent exists here. A diagnostic accuracy figure without a reader study establishes that the model performs, not that it adds. Nieri et al.'s comparative review exists precisely because this comparison is usually missing [38].

F6. Out-of-fold preprocessing. Stain normalisation, feature selection and synthetic oversampling fitted before partitioning leak test information into training. Where imbalance is addressed by resampling [147], its placement relative to the split is decisive and frequently unstated.

F7. Displayed rather than validated explanation. Attribution is shown without agreement testing against expert annotation or stability testing under resampling [136].

F8. Irreproducible reporting. Twenty-seven records were excluded at full text because no evaluation protocol could be recovered. Among included records, patient-level split files, seeds and environment specifications are almost never released.

7.3 The Prevalence Problem, Made Quantitative

F3 deserves separate treatment because it is the failure mode most likely to mislead a clinical reader and the one most easily corrected. Let Se and Sp denote sensitivity and specificity, and let π be the prevalence of disease in the population to which the model is applied. Then

$$\mathrm{PPV} = \frac{\mathrm{Se}\cdot\pi}{\mathrm{Se}\cdot\pi + (1-\mathrm{Sp})(1-\pi)} \quad (2)$$

$$\mathrm{NPV} = \frac{\mathrm{Sp}\cdot(1-\pi)}{\mathrm{Sp}\cdot(1-\pi)+(1-\mathrm{Se})\cdot\pi} \quad (3)$$

In a balanced experimental sample π = 0.5 and the reported PPV approximates accuracy. In opportunistic screening of an adult population in a high-incidence region, π for frank carcinoma is of the order of 10^{-3} to 10^{-4}, and even for OPMDs it rarely exceeds a few per cent. Substituting a representative and by no means pessimistic operating point of Se = 0.95 and Sp = 0.95 at $\pi = 10^{-3}$ into Eq. (2) gives

$$\mathrm{PPV} \approx \frac{0.00095}{0.00095+0.04995} \approx 0.019 \quad (4)$$

That is, fewer than two in every hundred positives would be true positives, and each of the remaining ninety-eight would generate an unnecessary referral, an anxious patient and in many pathways an unnecessary biopsy. The same model reported at π = 0.5 would show PPV = 0.95. Nothing about the model has changed; only the population has.

The screening-yield consequence follows directly. For a screened population of size N, the numbers of true and false positives are

$$\mathrm{TP} = N\cdot\pi\cdot\mathrm{Se}, \quad \mathrm{FP} = N\cdot(1-\pi)\cdot(1-\mathrm{Sp}) \quad (5)$$

so the referral burden per additional case detected is

$$B = \frac{\mathrm{FP}}{\mathrm{TP}} = \frac{(1-\pi)(1-\mathrm{Sp})}{\pi\cdot\mathrm{Se}} \quad (6)$$

which at the operating point above is approximately 52 unnecessary referrals per true case. Since specificity enters through (1 − Sp) multiplied by the large factor (1 − π), specificity not sensitivity dominates the usability of a screening model at low prevalence. The corpus reports both, but optimises and headlines sensitivity and accuracy.

The appropriate decision-analytic framing is net benefit at a threshold t reflecting the relative cost of a missed cancer against an unnecessary referral [148]:

$$\mathrm{NB}(t) = \frac{\mathrm{TP}}{N} - \frac{\mathrm{FP}}{N}\cdot\left[\frac{t}{1-t}\right] \quad (7)$$

and a model is worth deploying only where NB(t) exceeds that of the current pathway across the clinically plausible range of t. Reporting NB(t) answers the question that accuracy does not. No study in this corpus reports it.

Finally, for imbalanced evaluation the discrimination summary should not degenerate under prevalence shift. Balanced accuracy and the Matthews correlation coefficient [149] are appropriate,

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{N_+\cdot N_-\cdot P_+\cdot P_-}} \quad (8)$$

where N_+ = TP+FP, N_- = TN+FN, P_+ = TP+FN and P_- = TN+FP. Raw accuracy on a balanced sample, by contrast, conveys almost nothing about screening behaviour. Comparisons between models should be tested with an appropriate paired procedure on the ROC curves [150] rather than by inspecting point estimates, and the interpretation of the ROC summary itself requires care in the imbalanced regime [151].

7.4 Gap Severity Across Modalities

Fig 10 summarises the residual gaps as a severity matrix across six modality streams and eight methodological dimensions.

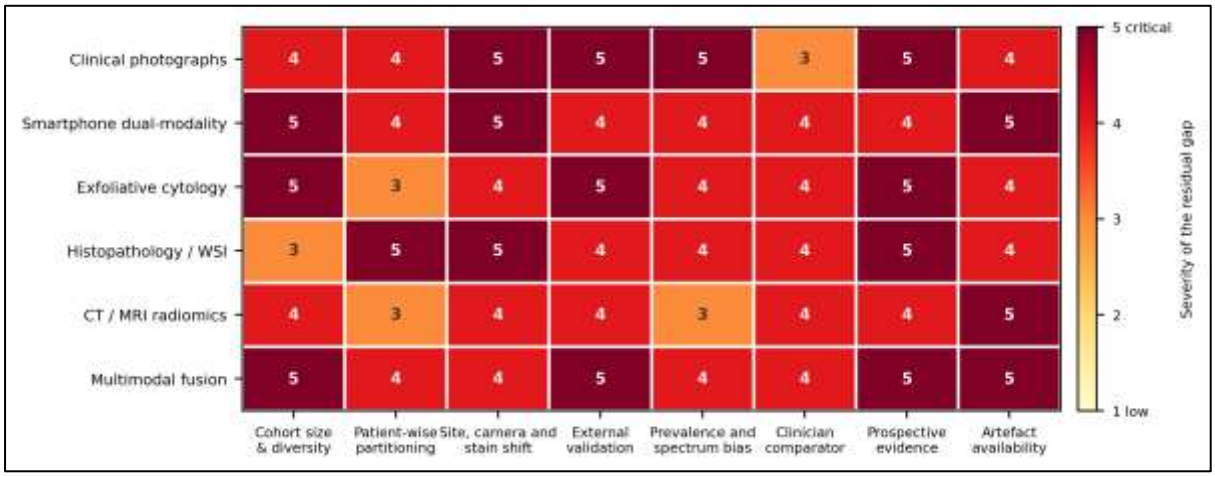

Fig 10. Severity matrix of residual gaps. Scores are the review authors' appraisal on a 1 (low) to 5 (critical) scale, derived from the practice audit of Section 7.1 and the modality analyses of Sections 5 and 6.

Prospective evidence scores 4 or 5 in every row, which is the single most uniform finding in the matrix and the clearest statement of where the field stands: after more than a decade of work, there is no prospective evidence that any of these models improves detection in practice.

7.5 OSCAR-Eval

Fig 11 and Algorithm 3 specify the proposed protocol. Every stage is achievable with existing tooling.

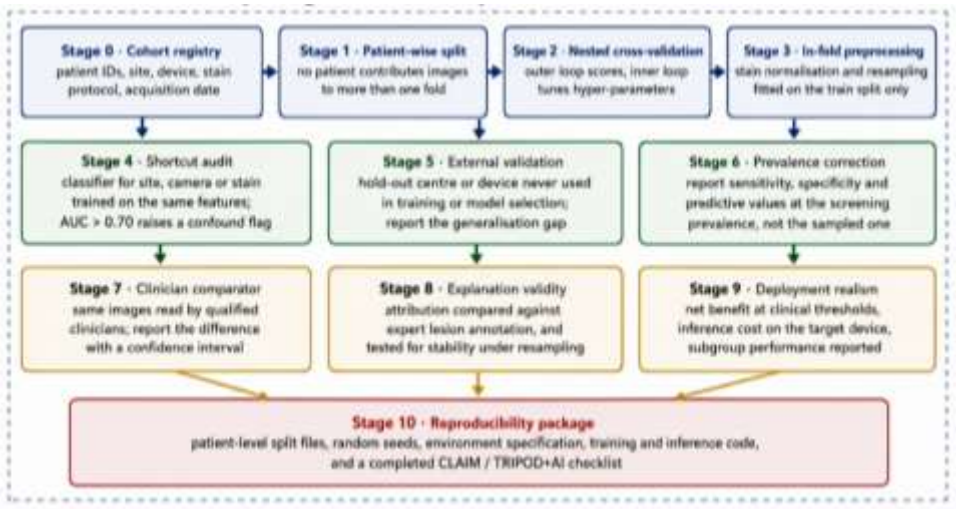

Fig 11. OSCAR-Eval, the proposed eleven-stage evaluation and reporting protocol. Stages 0-3 remove leakage; stages 4-6 establish generalisation and clinical interpretability; stages 7-9 establish incremental value, explanation validity and deployment realism; stage 10 makes the result reproducible.

Algorithm 3: OSCAR-Eval evaluation and reporting protocol

Input: dataset D with patient IDs and acquisition metadata; external cohort D_{ext}; model family F
Output: a disclosure record sufficient for independent clinical assessment

1: assert every image in D carries a patient ID and acquisition metadata; else abort ▷ *Stage 0*
2: P ← patientWiseStratifiedFolds(D, K) ▷ *Stage 1*
3: assert no patient appears in more than one fold of P
4: for each outer fold k in P do ▷ *Stage 2*
5: fit preprocessing π_k (stain normalisation, scaling) on $train_k$ only ▷ *Stage 3*
6: if resampling is used then apply it inside $train_k$ only
7: select hyper-parameters by inner cross-validation on $train_k$ only
8: $f_k \leftarrow train(F, \pi_k(train_k))$; score on $\pi_k(test_k)$
9: end for
10: g ← train auxiliary classifier for site / camera / stain on the same features ▷ *Stage 4*
11: if AUC(g) > 0.70 then flag a confound and report it
12: evaluate f on D_{ext}, never used above; report the generalisation gap ▷ *Stage 5*
13: report Se, Sp and, by Eqs. (2)–(3), PPV and NPV at the intended-use prevalence ▷ *Stage 6*
14: report referral burden B by Eq. (6) and net benefit NB(t) by Eq. (7)
15: qualified clinicians read the same images; report the difference with a CI ▷ *Stage 7*
16: compute attribution; test agreement with expert lesion annotation and stability ▷ *Stage 8*
17: report subgroup performance and inference cost on the target device ▷ *Stage 9*
18: release patient-level split files, seeds, environment, training and inference code ▷ *Stage 10*
19: return the disclosure record
20: a record missing lines 3, 13 or 18 is incomplete and its claims are not clinically assessable

Note: line 13 is the stage most often omitted and the one that most changes the reader's conclusion. It costs two arithmetic operations on numbers the study already has, and it converts an accuracy figure into a statement about what would happen in the population the model targets.

7.6 Six Reporting Invariants

The protocol above is a full standard; what a reviewer needs is a short checkable list. We propose six invariants. (i) The partition is patient-wise and the paper says so explicitly. (ii) All preprocessing, including stain normalisation and resampling, is fitted inside the training fold. (iii) Either an external cohort, centre or device is evaluated, or the paper states plainly that generalisation is untested and confines its claims accordingly. (iv) Predictive values are reported at the intended-use prevalence, with the assumed prevalence stated. (v) A clinician comparator reads the same images, or the paper states that no incremental-value claim is being made. (vi) Split files, seeds and inference code are released. A paper satisfying these six may still be wrong; it can at least be assessed. A paper failing any of them cannot be, whatever accuracy it reports.

8. Open Challenges and Research Roadmap

Table 11 maps every gap identified in this review to the failure modes that produce it, the remedy we propose and the horizon on which that remedy is realistic. Fig 12 places the same material on a three-horizon timeline.

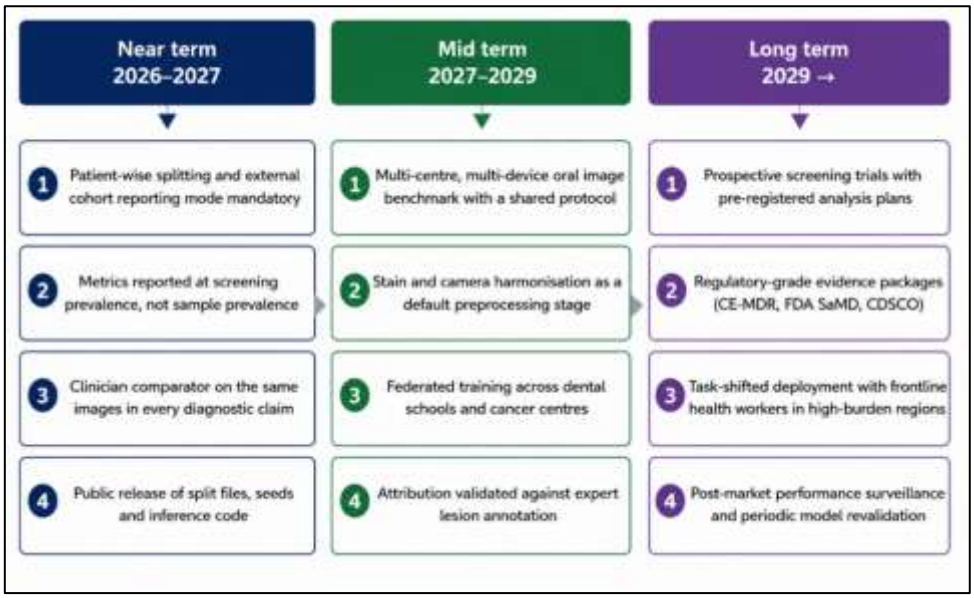

Fig 12. Three-horizon research roadmap. Near-term items are reporting and evaluation reforms achievable without new data collection; mid-term items require multi-centre benchmark and harmonisation infrastructure; long-term items require prospective trials, regulatory evidence and deployment programmes.

8.1 Benchmark Infrastructure

The single most useful investment the field could make is not architectural. It is a multi-centre, multi-device oral image benchmark with published patient-wise splits, documented acquisition conditions, and a stated case mix that reflects a screening rather than a referral population. The histopathology strand needs the equivalent: a tiled dataset distributed with patient identifiers and stain metadata, so that patient-wise partitioning becomes possible rather than merely recommended. Without these, the comparison between any two published models remains uninterpretable, because the between-dataset variance exceeds the between-method variance the reported numbers are supposed to express. The computational pathology community has built comparable resources for other tumour sites [34], [128], [36]; oral cancer has not.

8.2 Harmonisation and Shortcut Control

Stain normalisation for histopathology and colour, illumination and white-balance normalisation for photographic capture should be default preprocessing stages, and the shortcut audit of OSCAR-Eval Stage 4 should be a standard ablation. The mechanism is well established outside this field: a model presented with a nuisance variable that correlates with the label will use it, and will appear to succeed until the correlation breaks [62], [129], [130]. Oral cancer datasets are unusually exposed to this because case and control images are frequently acquired in different settings, the cases in a tertiary clinic and the controls in a dental school, so the acquisition variable is almost perfectly confounded with the label.

8.3 Reformulating the Clinical Task

Section 4 argued that the field has substituted architectures rapidly and reformulated the clinical question hardly at all. Three reformulations would materially increase clinical value. First, referral triage rather than diagnosis: the operationally useful output is a calibrated referral recommendation with a stated cost ratio, not a malignancy label. Second, malignant-transformation risk over time rather than current status: the studies that produce a prognostic quantity [135], [19] are more clinically substantive per unit of effort than those that produce a classification. Third, screening yield rather than accuracy: the endpoint that justifies a screening technology is additional cases detected per thousand screened at an acceptable referral burden, and Eqs. (5)–(7) show how to report it from quantities studies already possess.

8.4 Deployment Realism in High-Burden Settings

The incidence of oral cancer is concentrated in settings where specialist density is low and where the realistic operator is a frontline health worker with a handset, not a clinician with a workstation. This has three consequences the corpus mostly does not address. Inference must run on the device, offline, within a power and thermal budget; the model must degrade gracefully rather than confidently when image quality is poor; and the output must be a referral action, not a probability, because the operator's decision space is binary. The smartphone strand [14], [15], [16] has the right target and needs the evaluation discipline of Section 7 to make its case.

8.5 Foundation Models and Federated Training

Pathology foundation models [35], [36], [37] and self-supervised pre-training [78], [79] offer a plausible route past the small-dataset constraint that limits every strand in this corpus, because their representations are learned outside any single institution's staining and acquisition protocol. Federated training [152], [153], [128] offers a route past the data-sharing constraint that prevents multi-centre datasets from being assembled centrally. Both are promising and both are, at present, promises: no record in this corpus evaluates either on an oral cancer task with an external cohort. The honest statement is that the field has identified the right tools and has not yet tested them on its own problem.

Table 11. Mapping of identified gaps to failure modes, remedies and horizon.

Gap	Failure modes	Proposed remedy	Horizon
Image-wise partitioning of multi-image patients	F1	Patient-wise folds asserted explicitly; datasets redistributed with patient identifiers	Near
Hospital-archive case mix	F2	Stated clinical spectrum matched to intended use; screening-population sampling	Mid
Metrics reported at sampled prevalence	F3	PPV and NPV by Eqs. (2)–(3) at the intended-use prevalence, with that prevalence stated	Near
Referral burden unreported	F3	Burden B by Eq. (6) and net benefit NB(t) by Eq. (7) across clinical thresholds	Near
Stain, scanner and camera shift	F4	Harmonisation as default preprocessing; shortcut audit as a standard ablation	Near
No clinician comparator	F5	Reader study on the same images with the difference and its interval reported	Near
Preprocessing fitted outside the fold	F6	In-fold normalisation, selection and resampling, with placement stated	Near
Attribution displayed, not validated	F7	Agreement with expert lesion annotation; stability under resampling	Mid
No artefacts released	F8	Split files, seeds, environment and inference code released with the paper	Near
No shared oral benchmark	F1, F2, F4	Multi-centre, multi-device benchmark with published patient-wise splits	Mid
Foundation and federated approaches untested here	n/a	Evaluation of pathology foundation encoders and federated training on oral tasks with external cohorts	Mid
Task formulation misaligned with the decision	F5	Calibrated referral triage and transformation-risk endpoints replacing binary labels	Mid
No prospective evidence in any modality	F2, F5	Prospective screening studies with pre-registered analysis plans and yield endpoints	Long
No regulatory-grade evidence	F8	Evidence packages meeting CE-MDR, FDA SaMD and CDSCO expectations	Long
Deployment on frontline devices unproven	n/a	On-device inference with graceful degradation and referral-action outputs; post-market surveillance	Long

8.6 Threats to the Validity of This Review

Four threats are material and we state them rather than qualify them away. First, selection: the corpus was seeded from a supplied list of 24 records reflecting its compilers' interests and then extended by systematic search, so the resulting distribution is not a neutral sample of the field; the cytology strand at two records is almost certainly under-represented, and non-English clinical literature is absent by construction. Second, appraisal subjectivity: DETECT scores and the critical-limitation cells in Tables 8, 9 and 10 are the review authors' judgement from the retrievable record, and a practice present but unreported is scored as absent by design, which understates what studies did while correctly representing what a reader can verify. Third, the practice-audit percentages in Fig 9 are computed over cluster membership as assigned in this review, and a different assignment would move them by several points, though not enough to reverse the ordering that carries the argument. Fourth, the prevalence analysis in Section 7.3 uses representative rather than study-specific operating points, because most studies do not report the sensitivity-specificity pair at a stated threshold; the conclusion follows from the structure of Eqs. (2) and (6) rather than from the particular numbers, but a reader should treat the illustrative figures as illustrative.

We also record a limitation of the corpus construction itself. Criterion I7 excluded 140 records whose bibliographic identity could not be verified. Some of those were certainly legitimate works with poorly indexed metadata, and excluding them trades recall for verifiability. We consider the trade correct for a review whose thesis concerns evidentiary standards, but it is a trade and not a free choice.

9. Conclusion

This review examined 153 records on artificial intelligence for the early detection of oral cancer, spanning 1988 to early 2026, assembled under a PRISMA-conformant protocol in which every retained record's bibliographic identity was verified against a retrievable source and 140 records were excluded because it could not be.

The descriptive picture is one of steady growth and rapid architectural turnover. The critical picture resolves into three quantified deficits and one structural misalignment.

The evaluation deficit: 34% of primary studies evidence patient-wise partitioning, falling to 28% in the histopathology cluster where tile-level splitting does the most damage. A reported accuracy obtained without patient-wise partitioning may be measuring within-patient similarity, and the reader cannot tell which.

The translation deficit: 16% report external or multi-site validation, 12% include a clinician comparator on the same images, and 14% release data or code. Without a comparator, an accuracy figure establishes that a model performs, not that it adds anything to a clinician who is already looking at the same lesion.

The prevalence deficit: metrics are reported at the balanced prevalence of an experimental sample rather than the low prevalence of a screening population. We showed that a model at 95% sensitivity and 95% specificity, a performance level the corpus routinely exceeds on its own data, yields a positive predictive

value under 2% and roughly fifty-two unnecessary referrals per true case at a prevalence of 10^{-3}. That translation costs two arithmetic operations and is essentially never performed.

The structural misalignment is the finding we would most want carried forward. The densest strand of this literature operates on tissue that has already been biopsied, which is downstream of the delay that causes late-stage presentation. Oral cancer is unusual among common malignancies in that its precursors are visible to an unaided examiner; the binding constraint is therefore examination and referral, not detectability. Work at the chairside and at the point of triage addresses that constraint. Work on slide grading, however good, does not, and calling it early detection obscures a distinction that matters to patients.

What we offer in response is prescriptive rather than descriptive: eight named failure modes stated so they can be checked; a prevalence-correction and screening-yield analysis that converts reported accuracy into the quantity a clinician needs; and OSCAR-Eval, reduced to six invariants: patient-wise partitioning, in-fold preprocessing, external validation or an explicit statement of its absence, predictive values at the intended-use prevalence, a clinician comparator, and released artefacts. None of these is expensive. Their near-universal absence is a norms problem rather than a resource problem, and norms problems are the kind a field can fix by deciding to.

The technical capability to build these systems is evidently present. What is missing is the evaluation discipline that would let anyone tell whether they work, and the prospective evidence that would let a health system act on them. Until a study reports how many additional cancers were found per thousand people screened, and at what referral cost, the field will continue to accumulate accuracy figures that no clinician can use.

Acknowledgment

The authors thank the colleagues whose comments shaped the methodological framing of this review, and acknowledge the compilers of the supplied seed reference list, which provided the starting point for the systematic search reported in Section 2.

Author Contributions

M. K. Soni: conceptualisation, protocol design, search execution, screening, appraisal, formal analysis, and writing of the original draft. A. K. Jain: methodology, validation of the appraisal rubric, supervision, and critical revision of the analytical sections. S. Varma: study design, interpretation of the practice audit, supervision, and review and editing. All authors have read and approved the final manuscript and accept responsibility for its content.

Funding

This review received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval and Informed Consent

This study is a systematic review of previously published literature. It involved no human participants, no human tissue and no animals, and no new patient data were accessed. Ethical approval and informed consent were therefore not applicable.

Data Availability

No new experimental data were generated. The 153 records constituting the review corpus are listed in full in the References. The screening counts reported in Fig 1, the appraisal scores summarised in Section 7.1 and the derived percentages reported throughout are reproducible from that reference list together with the protocol of Section 2 and Algorithms 1 and 2. The extraction sheet used for the practice audit is available from the corresponding author on reasonable request.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin., vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[2] H. Rumgay et al., "Global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2022," CA Cancer J. Clin., 2026, doi: 10.3322/caac.70048.

[3] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin., vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[4] A. M. Filho, M. Laversanne, J. Ferlay et al., "The GLOBOCAN 2022 cancer estimates: Data sources, methods, and a snapshot of the cancer burden worldwide," Int. J. Cancer, vol. 156, no. 7, pp. 1336–1346, 2025, doi: 10.1002/ijc.35278.
[5] J. Lau et al., "Adjunctive aids for the detection of oral squamous cell carcinoma and oral potentially malignant disorders: A systematic review of systematic reviews," Jpn. Dent. Sci. Rev., vol. 60, pp. 53–72, 2024. DOI: 10.1016/j.jdsr.2023.12.004
[6] M. Á. González-Moles, M. Aguilar-Ruiz, and P. Ramos-García, "Diagnostic accuracy of artificial intelligence compared to biopsy in detecting early oral squamous cell carcinoma: A systematic review and meta-analysis," Asian Pac. J. Cancer Prev., 2022. DOI: 10.31557/APJCP.2024.25.8.2593
[7] M. Á. González-Moles, M. Aguilar-Ruiz, and P. Ramos-García, "Challenges in the early diagnosis of oral cancer, evidence gaps and strategies for improvement: A scoping review of systematic reviews," Cancers, vol. 14, no. 19, Art. no. 4967, 2022. DOI: 10.3390/cancers14194967
[8] V. Murthy et al., "Transformation rate of oral submucous fibrosis: A systematic review and meta-analysis," J. Clin. Med., vol. 11, no. 7, Art. no. 1793, 2022. DOI: 10.3390/jcm11071793
[9] K. D. Shield, J. Ferlay, A. Jemal et al., "The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012," CA Cancer J. Clin., vol. 67, no. 1, pp. 51–64, 2017.
[10] J. Huang, S. C. Chan, S. Ko et al., "Disease burden, risk factors, and trends of lip, oral cavity, pharyngeal cancers: A global analysis," Cancer Med., vol. 12, no. 17, pp. 18153–18164, 2023.
[11] Q. Fu, Y. Chen, Z. Li et al., "A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study," EClinicalMedicine, vol. 27, Art. no. 100558, 2020, doi: 10.1016/j.eclinm.2020.100558.
[12] K. Warin, W. Limprasert, S. Suebnukarn, S. Jinaporntham, and P. Jantana, "Automatic classification and detection of oral cancer in photographic images using deep learning algorithms," J. Oral Pathol. Med., vol. 50, no. 9, pp. 911–918, 2021, doi: 10.1111/jop.13227.
[13] K. Warin et al., "AI-based analysis of oral lesions using novel deep convolutional neural networks for early detection of oral cancer," PLoS ONE, vol. 17, no. 8, Art. no. e0273508, 2022.
[14] R. D. Uthoff, B. Song, S. Sunny, S. Patrick, A. Suresh, T. Kolur et al., "Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities," PLoS ONE, vol. 13, no. 12, Art. no. e0207493, 2018, doi: 10.1371/journal.pone.0207493.
[15] R. D. Uthoff, B. Song, S. Sunny, S. Patrick, A. Suresh, T. Kolur, K. Gurushanth, K. Wooten, V. Gupta, M. E. Platek, A. K. Singh, P. Wilder-Smith, M. A. Kuriakose, P. Birur, and R. Liang, "Small form factor, flexible, dual-modality handheld probe for smartphone-based, point-of-care oral and oropharyngeal cancer screening," J. Biomed. Opt., vol. 24, no. 10, Art. no. 106003, 2019, doi: 10.1117/1.JBO.24.10.106003.
[16] V. Talwar, P. Singh, N. Mukhia, A. Shetty, P. Birur, and K. M. Desai, "AI-assisted screening of oral potentially malignant disorders using smartphone-based photographic images," Cancers, vol. 15, no. 16, Art. no. 4120, 2023, doi: 10.3390/cancers15164120.
[17] Z. Pirayesh, H. Mohammad-Rahimi, N. Ghasemi, S.-R. Motamedian, T. S. Sadeghi, H. Koohi, R. Rokhshad, S. M. Lotfi, A. Najafi, S. A. Alajaji, Z. H. Khoury, M. Jessri, and A. S. Sultan, "Deep learning-based image classification and segmentation on digital histopathology for oral squamous cell carcinoma: A systematic review and meta-analysis," J. Oral Pathol. Med., vol. 53, no. 9, pp. 551–566, 2024, doi: 10.1111/jop.13578.
[18] M. Jagathesh and S. Narayanan, "Attention-aided hybrid transformer network for oral squamous cell carcinoma classification using histopathology images," Front. Artif. Intell., vol. 9, Art. no. 1751520, 2026, doi: 10.3389/frai.2026.1751520.
[19] T. Lan, S. Kuang, P. Liang, C. Ning, Q. Li, L. Wang et al., "MRI-based deep learning and radiomics for prediction of occult cervical lymph node metastasis and prognosis in early-stage oral and oropharyngeal squamous cell carcinoma: A diagnostic study," Int. J. Surg., vol. 110, no. 8, pp. 4648–4659, 2024, doi: 10.1097/JS9.0000000000001578.
[20] S. Liu, A. Zhang, J. Xiong, X. Su, Y. Zhou, Y. Li, Z. Zhang, Z. Li, and F. Liu, "The application of radiomics machine learning models based on multimodal MRI with different sequence combinations in predicting cervical lymph node metastasis in oral tongue squamous cell carcinoma patients," Head Neck, vol. 46, no. 3, pp. 513–527, 2024, doi: 10.1002/hed.27605.
[21] H.-C. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1285–1298, 2016.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90
[23] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.
[24] A. Soni, P. K. Sethy, A. K. Dewangan, A. Nanthaamornphong, S. K. Behera, and B. Devi, "Enhancing oral squamous cell carcinoma detection: A novel approach using improved EfficientNet architecture," BMC Oral Health, vol. 24, no. 1, Art. no. 601, 2024.
[25] E. Albalawi, A. Thakur, M. T. Ramakrishna, S. Bhatia Khan, S. SankaraNarayanan, and B. Almarri, "Oral squamous cell carcinoma detection using EfficientNet on histopathological images," Front. Med., vol. 10, 2023. DOI: 10.3389/fmed.2023.1349336
[26] Day, T. A., Davis, B. K., Gillespie, M. B., Joe, J. K., Kibbey, M., Martin-Harris, B., ... & Stuart, R. K. (2003). Oral cancer treatment. Current treatment options in oncology, 4(1), 27-41.
[27] B. Song et al., "Interpretable and reliable oral cancer classifier with attention mechanism and expert knowledge embedding via attention map," Cancers, vol. 15, no. 5, Art. no. 1421, 2023.
[28] J. Li, "Fusion feature-based hybrid methods for diagnosing oral squamous cell carcinoma in histopathological images," Front. Oncol., vol. 15, Art. no. 1551876, 2025, doi: 10.3389/fonc.2025.1551876. DOI: 10.3389/fonc.2025.1551876

[29] S. Ramya and R. I. Minu, "Oral squamous cell carcinoma grading classification using deep transformer encoder assisted dilated convolution with global attention," Front. Artif. Intell., vol. 8, Art. no. 1575427, 2025, doi: 10.3389/frai.2025.1575427.
[30] D. P. Yadav, B. Sharma, A. Noonia, and A. Mehbodniya, "Explainable label guided lightweight network with axial transformer encoder for early detection of oral cancer," Sci. Rep., vol. 15, no. 1, Art. no. 6391, 2025, doi: 10.1038/s41598-025-87627-y.
[31] N. Koriakina et al., "Oral cancer detection and interpretation: Deep multiple instance learning versus conventional deep single instance learning," arXiv:2202.01783, 2022.
[32] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in Proc. 35th Int. Conf. Mach. Learn. (ICML), 2018, pp. 2127–2136.
[33] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, and X. Ji, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021, pp. 2136–2147.
[34] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," Nat. Biomed. Eng., vol. 5, no. 6, pp. 555–570, 2021.
[35] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical Twitter," Nat. Med., vol. 29, no. 9, pp. 2307–2316, 2023.
[36] A. H. Song, G. Jaume, D. F. K. Williamson, M. Y. Lu, A. Vaidya, T. R. Miller, and F. Mahmood, "Artificial intelligence for digital and computational pathology," Nat. Rev. Bioeng., vol. 1, pp. 930–949, 2023.
[37] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," Nature, vol. 616, no. 7956, pp. 259–265, 2023. DOI: 10.1038/s41586-023-05881-4
[38] L. Nieri et al., "Diagnosis of oral cancer with deep learning: A comparative test accuracy systematic review," Oral Dis., 2025, doi: 10.1111/odi.15330.
[39] R. O. Alabi, I. O. Bello, O. Youssef, M. Elmusrati, A. A. Mäkitie, and A. Almangush, "Utilizing deep machine learning for prognostication of oral squamous cell carcinoma — A systematic review," Front. Oral Health, vol. 2, Art. no. 686863, 2021, doi: 10.3389/froh.2021.686863.
[40] R. O. Alabi, O. Youssef, M. Pirinen, M. Elmusrati, A. A. Mäkitie, I. Leivo, and A. Almangush, "Machine learning in oral squamous cell carcinoma: Current status, clinical concerns and prospects for future — A systematic review," Artif. Intell. Med., vol. 115, Art. no. 102060, 2021.
[41] J. Adeoye, J. Y. Tan, S. W. Choi, and P. Thomson, "Prediction models applying machine learning to oral cavity cancer outcomes: A systematic review," Int. J. Med. Inform., vol. 154, Art. no. 104557, 2021. DOI: 10.1016/j.ijmedinf.2021.104557
[42] "Deep learning in oral cancer — A systematic review," BMC Oral Health, vol. 24, Art. no. 212, 2024, doi: 10.1186/s12903-024-03993-5.
[43] "Artificial intelligence and its application in early oral cancer screening: A systematic review," Front. Oncol., vol. 16, Art. no. 1789708, 2026, doi: 10.3389/fonc.2026.1789708.
[44] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nat. Med., vol. 25, no. 1, pp. 44–56, 2019.
[45] F. Schwendicke, W. Samek, and J. Krois, "Artificial intelligence in dentistry: Chances and challenges," J. Dent. Res., vol. 99, no. 7, pp. 769–774, 2020. DOI: 10.1177/0022034520915714
[46] L. Wynants et al., "Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal," BMJ, vol. 369, Art. no. m1328, 2020. DOI: 10.1136/bmj.m1328
[47] "Application of artificial intelligence and radiomics in the prediction of lymph node metastasis and tumour grading of oral cancer: A systematic review and meta-analysis," BMC Oral Health, 2026, doi: 10.1186/s12903-026-07658-3.
[48] C. Deng, J. Hu, P. Tang, T. Xu, L. He, Z. Zeng, and J. Sheng, "Application of CT and MRI images based on artificial intelligence to predict lymph node metastases in patients with oral squamous cell carcinoma: A subgroup meta-analysis," Front. Oncol., vol. 14, Art. no. 1395159, 2024, doi: 10.3389/fonc.2024.1395159.
[49] N. Al-Rawi, A. Sultan, B. Rajai, H. Shuaeeb, M. Alnajjar, M. Alketbi, Y. Mohammad, S. R. Shetty, and M. A. Mashrah, "The effectiveness of artificial intelligence in detection of oral cancer," Int. Dent. J., vol. 72, no. 4, pp. 436–447, 2022. DOI: 10.1016/j.identj.2022.03.001
[50] S. Panigrahi and T. Swarnkar, "Machine learning techniques used for the histopathological image analysis of oral cancer — A review," Open Bioinform. J., vol. 13, no. 1, 2020. DOI: 10.2174/1875036202013010106
[51] I. Ahmad and F. Alqurashi, "Early cancer detection using deep learning and medical imaging: A survey," Crit. Rev. Oncol. Hematol., vol. 204, Art. no. 104528, 2024. DOI: 10.1016/j.critrevonc.2024.104528
[52] V. P. Veeraraghavan et al., "Harnessing artificial intelligence for predictive modelling in oral oncology: Opportunities, challenges, and clinical perspectives," Oral Oncol. Rep., vol. 11, Art. no. 100591, 2024.
[53] D. Varalakshmi et al., "Transforming oral cancer care: The promise of deep learning in diagnosis," Oral Oncol. Rep., vol. 10, Art. no. 100482, 2024. DOI: 10.1016/j.oor.2024.100482
[54] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. no. n71, 2021. DOI: 10.1136/bmj.n71
[55] P. F. Whiting, A. W. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma et al., "QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies," Ann. Intern. Med., vol. 155, no. 8, pp. 529–536, 2011.
[56] J. Mongan, L. Moy, and C. E. Kahn, "Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers," Radiol. Artif. Intell., vol. 2, no. 2, Art. no. e200029, 2020. DOI: 10.1148/ryai.2020200029

[57] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024. DOI: 10.1136/bmj-2023-078378
[58] P. M. Bossuyt et al., "STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies," BMJ, vol. 351, Art. no. h5527, 2015. DOI: 10.1136/bmj.h5527
[59] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," npj Digit. Med., vol. 5, Art. no. 48, 2022. DOI: 10.1038/s41746-022-00592-y
[60] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nat. Mach. Intell., vol. 3, no. 3, pp. 199–217, 2021. DOI: 10.1038/s42256-021-00307-0
[61] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023. DOI: 10.1016/j.patter.2023.100804
[62] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," PLoS Med., vol. 15, no. 11, Art. no. e1002683, 2018. DOI: 10.1371/journal.pmed.1002683
[63] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in Proc. ACM Conf. Health, Inference, Learn. (CHIL), 2020, pp. 151–159. DOI: 10.1145/3368555.3384468
[64] B. Song et al., "Mobile-based oral cancer classification for point-of-care screening," J. Biomed. Opt., vol. 26, no. 6, Art. no. 065003, 2021. DOI: 10.1117/1.JBO.26.6.065003
[65] H. Lin, H. Chen, L. Weng, J. Shao, and J. Lin, "Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis," J. Biomed. Opt., vol. 26, no. 8, Art. no. 086007, 2021. DOI: 10.1117/1.JBO.26.8.086007
[66] S. Rabinovici-Cohen et al., "From pixels to diagnosis: Algorithmic analysis of clinical oral photos for early detection of oral squamous cell carcinoma," Cancers, vol. 16, no. 5, Art. no. 1019, 2024.
[67] K. M. Desai, P. Singh, M. Smriti et al., "Screening of oral potentially malignant disorders and oral cancer using deep learning models," Sci. Rep., vol. 15, Art. no. 17949, 2025, doi: 10.1038/s41598-025-02802-5.
[68] "An enhanced histopathology analysis: An AI-based system for multiclass grading of oral squamous cell carcinoma and segmenting of epithelial and stromal tissue," Cancers, vol. 13, no. 8, Art. no. 1784, 2021, doi: 10.3390/cancers13081784.
[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations (ICLR), 2015.
[70] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594
[71] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1251–1258. DOI: 10.1109/CVPR.2017.195
[72] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1
[73] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745
[74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74
[75] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2921–2929.
[76] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.
[77] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022. DOI: 10.1109/ICCV48922.2021.00986
[78] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 9650–9660. DOI: 10.1109/ICCV48922.2021.00951
[79] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 16000–16009. DOI: 10.1109/CVPR52688.2022.01553
[80] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015. DOI: 10.1038/nature14539
[81] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2012, pp. 1097–1105. DOI: 10.1145/3065386
[82] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848
[83] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308
[84] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708. DOI: 10.1109/CVPR.2017.243
[85] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 11976–11986. DOI: 10.1109/CVPR52688.2022.01167

[86] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.
[87] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28
[88] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nat. Methods, vol. 18, no. 2, pp. 203–211, 2021. DOI: 10.1038/s41592-020-01008-z
[89] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations (ICLR), 2015. arXiv
[90] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 448–456.
[91] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.
[92] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learn. Representations (ICLR), 2018.
[93] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," J. Big Data, vol. 6, Art. no. 60, 2019. DOI: 10.1186/s40537-019-0197-0
[94] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.324
[95] Nethan, S. T., Pérez-de-Oliveira, M. E., Campbell, F., Santos-Silva, A. R., & Lauby-Secretan, B. (2026). Interventions for oral cancer prevention: An evidence and gap map. International Journal of Cancer, 159(1), 188-197.
[96] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 3347–3357.
[97] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2010. DOI: 10.1109/TKDE.2009.191
[98] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.
[99] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," Nat. Commun., vol. 15, Art. no. 654, 2024. DOI: 10.1038/s41467-024-44824-z
[100] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097
[101] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 3319–3328. arXiv
[102] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.
[103] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778
[104] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.
[105] A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012
[106] S. Xu, C. Liu, Y. Zong, S. Chen, Y. Lu, L. Yang, and C. Zhang, "An early diagnosis of oral cancer based on three-dimensional convolutional neural networks," IEEE Access, vol. 7, pp. 158603–158611, 2019. DOI: 10.1109/ACCESS.2019.2950286
[107] M. Z. M. Shamim et al., "Automated detection of oral pre-cancerous tongue lesions using deep learning for early diagnosis of oral cavity cancer," doi:10.1093/comjnl/bxaa136
[108] G. A. I. Devindi, D. M. D. R. Dissanayake, S. N. Liyanage, F. B. A. H. Francis, M. B. D. Pavithya, N. S. Piyarathne, and P. V. K. S. Hettiarachchi, "Multimodal deep convolutional neural network pipeline for AI-assisted early detection of oral cancer," IEEE Access, vol. 12, 2024. DOI: 10.1109/ACCESS.2024.3454338
[109] R. Qasrawi et al., "Hybrid ensemble deep learning model for advancing breast cancer detection and classification in clinical applications," Heliyon, vol. 10, no. 19, 2024. DOI: 10.1016/j.heliyon.2024.e38374
[110] W. Lian, J. Lindblad, C. R. Stark, J. M. Hirsch, and N. Sladoje, "Let it shine: Autofluorescence of Papanicolaou-stain improves AI-based cytological oral cancer detection," Comput. Biol. Med., vol. 185, Art. no. 109498, 2025. DOI: 10.1016/j.compbiomed.2024.109498
[111] I. C. E. Guedes, A. L. F. Espinosa, T. W. Lepper, M. M. Rönnau, N. B. Daroit, and M. M. Oliveira, "Applicability of artificial intelligence analysis in oral cytopathology: A pilot study," Oral Surg. Oral Med. Oral Pathol. Oral Radiol., 2023. DOI: 10.1159/000543852
[112] P. Kulkarni, N. Sarwe, A. Pingale, Y. Sarolkar, R. R. Patil, and G. Shinde, "Exploring the efficacy of various CNN architectures in diagnosing oral cancer from squamous cell carcinoma," MethodsX, vol. 13, Art. no. 103034, 2024. DOI: 10.1016/j.mex.2024.103034
[113] D. K. Redie, S. Bilgaiyan, and S. Sagnika, "Oral cancer detection using transfer learning-based framework from histopathology images," J. Electron. Imaging, vol. 32, no. 5, Art. no. 053004, 2023.
[114] S. Panigrahi, B. S. Nanda, R. Bhuyan, K. Kumar, S. Ghosh, and T. Swarnkar, "Classifying histopathological images of oral squamous cell carcinoma using deep transfer learning," Heliyon, vol. 9, no. 3, 2023.
[115] D. Raval and J. N. Undavia, "A comprehensive assessment of convolutional neural networks for skin and oral cancer detection using medical images," Healthcare Analytics, vol. 3, Art. no. 100199, 2023. DOI: 10.1016/j.health.2023.100199

[116] "Detection of oral squamous cell carcinoma cancer using AlexNet on histopathological images," Discov. Appl. Sci., 2025, doi: 10.1007/s42452-025-06514-3.
[117] M. A. Deif, H. Attar, A. Amer, I. A. Elhaty, M. R. Khosravi, and A. A. Solyman, "Diagnosis of oral squamous cell carcinoma using deep neural networks and binary particle swarm optimization on histopathological images," Comput. Intell. Neurosci., 2022. DOI: 10.1155/2022/6364102
[118] S. Alanazi et al., "Intelligent deep learning enabled oral squamous cell carcinoma detection and classification using biomedical images," Comput. Intell. Neurosci., 2022. DOI: 10.1155/2022/7643967
[119] L. Zhang, R. Shi, and N. Youssefi, "Oral cancer diagnosis based on gated recurrent unit networks optimized by an improved version of northern goshawk optimization algorithm," Heliyon, vol. 10, no. 7, 2024. DOI: 10.1016/j.heliyon.2024.e32077
[120] A. Albishri, S. J. H. Shah, Y. Lee, and R. Wang, "OCU-Net: A novel U-Net architecture for enhanced oral cancer segmentation," arXiv:2310.02486, 2023.
[121] R. Dharani and K. Danesh, "Oral cancer segmentation and identification system based on histopathological images using MaskMeanShiftCNN and SV-OnionNet," Intell.-Based Med., vol. 10, Art. no. 100185, 2024.
[122] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 14318–14328. DOI: 10.1109/CVPR46437.2021.01409
[123] J. Shi, D. Sun, K. Wu, Z. Jiang, X. Kong, W. Wang, H. Wu, and Y. Zheng, "Positional encoding-guided transformer-based multiple instance learning for histopathology whole slide image classification," Comput. Methods Programs Biomed., vol. 258, Art. no. 108491, 2025.
[124] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," Med. Image Anal., vol. 67, Art. no. 101813, 2021.
[125] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017. DOI: 10.1016/j.media.2017.07.005
[126] M. Y. Lu, T. Y. Chen, D. F. K. Williamson, M. Zhao, M. Shady, J. Lipkova, and F. Mahmood, "AI-based pathology predicts origins for cancers of unknown primary," Nature, vol. 594, no. 7861, pp. 106–110, 2021. DOI: 10.1038/s41586-021-03512-4
[127] K. Li, Z. Qian, Y. Han, E. I.-C. Chang, B. Wei, M. Lai, J. Liao, Y. Fan, and Y. Xu, "Weakly supervised histopathology image segmentation with self-attention," Med. Image Anal., vol. 86, Art. no. 102791, 2023. DOI: 10.1016/j.media.2023.102791
[128] M. Y. Lu, R. J. Chen, D. Kong, J. Lipkova, R. Singh, D. F. K. Williamson, T. Y. Chen, and F. Mahmood, "Federated learning for computational pathology on gigapixel whole slide images," Med. Image Anal., vol. 76, Art. no. 102298, 2022, doi: 10.1016/j.media.2021.102298.
[129] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," Nat. Mach. Intell., vol. 3, no. 7, pp. 610–619, 2021. DOI: 10.1038/s42256-021-00338-7
[130] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," Nat. Mach. Intell., vol. 2, no. 11, pp. 665–673, 2020.
[131] "Explainable AI for oral cancer diagnosis: Multiclass classification of histopathology images and Grad-CAM visualization," Biology, vol. 14, no. 8, Art. no. 909, 2025, doi: 10.3390/biology14080909.
[132] S. Camalan, H. Mahmood, H. Binol, A. L. D. Araujo, A. R. Santos-Silva, and P. A. Vargas, "Convolutional neural network-based clinical predictors of oral dysplasia: Class activation map analysis of deep learning results," Cancers, vol. 13, no. 6, Art. no. 1291, 2021, doi: 10.3390/cancers13061291.
[133] V. P. G. Pushparathi, S. R. S. Vallee Narayan, R. S. Pratheeba, and V. Naveen, "Histopathological image analysis and enhanced diagnostic accuracy explainability for oral cancer detection," Pol. J. Pathol., vol. 76, no. 2, pp. 120–130, 2025, doi: 10.5114/pjp.2025.153973.
[134] S. J. Shah, A. Albishri, R. Wang, and Y. Lee, "Integrating local and global attention mechanisms for enhanced oral cancer detection and explainability," SSRN 4925590, 2024.
[135] R. M. S. Bashir, A. J. Shephard, and H. Mahmood, "A digital score of peri-epithelial lymphocytic activity predicts malignant transformation in oral epithelial dysplasia," J. Pathol., vol. 260, no. 4, pp. 431–442, 2023, doi: 10.1002/path.6094.
[136] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2018, pp. 9505–9515.
[137] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019.doi:
[138] Z. Chen, Y. Yu, S. Liu et al., "A deep learning and radiomics fusion model based on contrast-enhanced computed tomography improves preoperative identification of cervical lymph node metastasis of oral squamous cell carcinoma," Clin. Oral Investig., vol. 28, no. 1, Art. no. 39, 2023.
[139] A. Vidiri, S. Marzi, F. Piludu et al., "Magnetic resonance imaging-based prediction models for tumor stage and cervical lymph node metastasis of tongue squamous cell carcinoma," Comput. Struct. Biotechnol. J., vol. 21, pp. 4277–4287, 2023.
[140] M. Dohopolski, L. Chen, D. Sher, and J. Wang, "Predicting lymph node metastasis in patients with oropharyngeal cancer by using a convolutional neural network with associated epistemic and aleatoric uncertainty," Phys. Med. Biol., vol. 65, no. 22, Art. no. 225002, 2020, doi: 10.1088/1361-6560/abb71c.
[141] A. De Biase, N. M. Sijtsema, L. V. van Dijk, R. Steenbakkers, J. A. Langendijk, and P. van Ooijen, "Uncertainty-aware deep learning for segmentation of primary tumor and pathologic lymph nodes in oropharyngeal cancer: Insights from a multi-center cohort," Comput. Med. Imaging Graph., vol. 123, Art. no. 102535, 2025, doi: 10.1016/j.compmedimag.2025.102535.
[142] A. Zwanenburg, M. Vallières, M. A. Abdalah et al., "The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping," Radiology, vol. 295, no. 2, pp. 328–338, 2020.
[143] S. K. D. Sharma, "Oral squamous cell detection using deep learning," arXiv:2408.08939, 2024.

[144] C. Kavyashree, H. S. Vimala, and J. Shreyas, "Improving oral cancer detection using pretrained model," in Proc. IEEE 6th Conf. Inf. Commun. Technol. (CICT), 2022, pp. 18–20.
[145] A. Aljuaid, M. Almohaya, and M. Anwar, "An early detection of oral epithelial dysplasia based on GoogLeNet inception-v3," in Proc. IEEE Int. Conf. Bioinformatics Biomedicine, 2022.
[146] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.doi: 10.1038/nature21056
[147] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.doi: 10.1613/jair.953
[148] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," Med. Decis. Making, vol. 26, no. 6, pp. 565–574, 2006.doi: 10.1177/0272989X06295361
[149] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, Art. no. 6, 2020.doi: 10.1186/s12864-019-6413-7
[150] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.doi: 10.2307/2531595
[151] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, 2006.doi: 10.1016/j.patrec.2005.10.010
[152] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), 2017, pp. 1273–1282.
[153] N. Rieke et al., "The future of digital health with federated learning," npj Digit. Med., vol. 3, Art. no. 119, 2020.doi: 10.1038/s41746-020-00323-1