QUBAHAN JOURNAL OF MEDICAL SCIENCES
VOL. 3, November 2026
https://doi.org/10.48161/qjms.v3a90

Classification of Oral Cancer Through Artificial Intelligence-Based Histopathological Image Analysis: A Comprehensive Review of Diagnostic Methods and Clinical Challenges

Dilovan Asaad Zebari [1*], Nechirvan Asaad Zebari[2]

[1] Faculty of Computing and Information Technology, Sohar 311, Sohar University, Oman;
[2] Department of Technical Engineering, University of Mohaghegh Ardabili, Iran.

* **Corresponding author:** dilo.asaad@mail.com.

ABSTRACT: Background: Artificial intelligence (AI) is increasingly applied to oral squamous cell carcinoma (OSCC) detection and oral epithelial dysplasia (OED) grading in digital histopathology. **Objective:** To characterize diagnostic tasks, model families, reported performance, and evidence relevant to clinical generalizability. **Methods:** We developed a structured, task-stratified evidence map from 39 topic-specific bibliographic sources, supplemented by two methodological guidance references and three cross-modal methodological examples. Sources were classified by evidence role, and published performance estimates were interpreted according to endpoint and validation setting. Comprehensive multi-database retrieval records, independent eligibility screening, and completed study-level risk-of-bias assessments were not available; the synthesis is therefore provisional. **Results:** The topical inventory comprised 27 candidate classification or grading studies, three adjacent original investigations, five secondary reviews, two dataset sources, and two clinical or pathology background publications. Several classifiers reported high internal accuracy, but differences in case mix, diagnostic labels, prediction units, and independent testing precluded a valid cross-study ranking. No meta-analysis was undertaken because comparable patient-level diagnostic contingency data were not consistently available. **Conclusion:** The principal barrier to clinical translation is not model discrimination alone but reproducible patient-independent evaluation, dependable reference diagnoses, external validation, and demonstrated benefit in clinical workflows. The review defines evidence requirements for more reliable future investigations.

Keywords: oral squamous cell carcinoma, oral epithelial dysplasia, histopathology, deep learning, diagnostic accuracy, external validation.

I. INTRODUCTION

Oral cancer remains a significant global health challenge because of its aggressive progression, associated morbidity, and potentially unfavorable outcomes when diagnosis is delayed. Oral squamous cell carcinoma (OSCC), the predominant histological type of oral malignancy, is primarily diagnosed through microscopic examination of tissue biopsies. Histopathological assessment enables pathologists to identify characteristic malignant features, including nuclear atypia, abnormal cellular differentiation, increased mitotic activity, and invasion of the underlying connective tissue [1]. In contrast, oral epithelial dysplasia (OED) is evaluated according to architectural disturbances and cytological abnormalities that indicate varying degrees of malignant potential. Although histopathological examination remains essential for diagnosis, interpretation may be influenced by tissue heterogeneity, differences in pathological expertise, and interobserver variability. These challenges have encouraged the development of artificial intelligence (AI)-assisted image classification systems to improve diagnostic consistency and support pathological decision-making [1–4].

Recent advances in machine learning and deep learning have substantially expanded the capabilities of automated oral histopathological image analysis. Conventional approaches use handcrafted morphological and textural features, whereas convolutional neural networks (CNNs) and transfer learning architectures automatically extract hierarchical image representations. More recently, researchers have investigated

Vision Transformers, hybrid CNN–Transformer architectures, weakly supervised whole-slide learning, and multimodal models to capture complex cellular characteristics and broader tissue relationships [5–21]. Several studies have reported high classification accuracy under controlled experimental conditions, demonstrating the potential of these technologies for distinguishing normal, dysplastic, and malignant tissue. However, classification performance varies considerably according to dataset composition, diagnostic categories, image magnification, preprocessing procedures, and evaluation strategies. Importantly, models developed to distinguish normal tissue from invasive OSCC cannot automatically be considered suitable for dysplasia grading, histological differentiation assessment, or prediction of future malignant transformation [5].

Despite promising technological developments, important challenges remain regarding the generalizability and clinical reliability of AI-based classification models. Many investigations rely on relatively small or repeatedly used histopathological datasets, while differences in training–testing separation, pathological annotations, and institutional imaging conditions complicate direct comparisons. Although some models achieve near-perfect accuracy during internal evaluation, their performance may deteriorate when tested on independent datasets. Furthermore, high predictive accuracy does not necessarily establish clinical usefulness, particularly when external validation, diagnostic uncertainty, explainability, and patient-level assessment are insufficiently addressed. Existing reviews have also highlighted substantial heterogeneity in study design, diagnostic modalities, performance reporting, and methodological quality [22–24]. Consequently, a more integrated evaluation is required to distinguish experimental classification success from evidence supporting reliable clinical implementation.

This review makes three contributions within the limits of its available source material. First, it separates 39 topic-specific sources into candidate diagnostic studies and distinct supporting evidence roles, avoiding the conflation of primary classifier results with datasets or secondary reviews. Second, it interprets performance by clinical endpoint and validation context, emphasizing patient independence, reference-label validity, and external transport rather than ranking isolated accuracy percentages. Third, it formulates a practical, validation-oriented framework for explainability, calibrated uncertainty, and pathologist-assisted evaluation. The present work is a structured preliminary evidence synthesis; it should not be described as a completed exhaustive systematic review until independently reproducible search, screening, extraction, and quality-assessment procedures have been documented [25, 26].

II. METHODS

This structured review addresses AI classification of digitized human oral histopathology, including binary OSCC detection, OED grading, and multiclass differentiation of normal, dysplastic, and malignant epithelium. The primary study period is January 2023 through November 2026. Candidate studies were required to address human oral histological material and evaluate a classifier or grading approach with a relevant diagnostic endpoint. Investigations restricted to clinical photographs, radiographic imaging, other cancer sites, prognosis without contemporaneous diagnostic classification, or image acquisition alone were not treated as primary evidence of diagnostic accuracy. Dataset reports, clinical background publications, and secondary reviews were retained only for provenance and context. These criteria specify the intended eligibility screen for a future finalized systematic review rather than retrospective claims that independent full-text screening has been completed. Table 1 distinguishes defined scope and completed source categorization from uncompleted review procedures.

Table 1. Study Eligibility Criteria and Review Status

Element	Criterion or procedure	Status
Period	January 2023– November 2026	Defined
Population	Digitized human oral histopathology	Candidate records mapped
Index test	AI classification of OSCC or OED	27 candidate primary records
Endpoints	Diagnosis or histopathological grading	Tasks stratified
Not primary evidence	Reviews, datasets, prognosis-only and acquisition-only studies	Contextualized separately
Screening	Independent dual screening and exclusion log	Not documented
Risk of bias	QUADAS-3 informed judgments	Planned; not scored
Pooling	Comparable diagnostic 2×2 data	Not performed

The starting material was an existing bibliography of 39 topic-specific sources, rather than database exports from a comprehensive literature search. Two additional references address diagnostic-test-quality and AI imaging reporting methods. Three additional publications on feature fusion in other medical imaging modalities provide methodological context, bringing the reference list to 44 entries. These five supporting references are excluded from the 39-source topical evidence inventory. An illustrative search concept for future reproducible retrieval is ("oral squamous cell carcinoma" OR "oral epithelial dysplasia" OR "oral cancer" OR OSCC) AND (histopatholog* OR "digital pathology" OR "whole slide" OR "whole-slide") AND ("deep learning" OR "machine learning" OR "artificial intelligence" OR transformer OR classification). The specific databases, platform syntax, search execution dates, complete result exports, deduplication decisions, and independent screening logs have not been recorded. Accordingly, Figure 1 depicts provisional source accounting and must not be interpreted as a completed selection flow with verified exclusion totals.

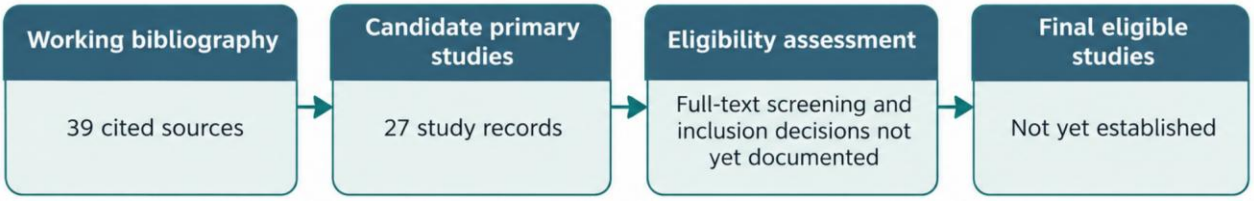


FIGURE 1. Working Bibliography and Provisional Study Selection

The 39 topic-specific sources were provisionally categorized by evidence role and clinical objective. To convert this inventory into a reproducible systematic review, at least two assessors would need to independently screen titles, abstracts, and full texts; document disagreements and exact full-text exclusion reasons; and preserve the complete audit trail. A standardized extraction sheet should then capture the study setting, diagnostic categories, independent numbers of patients/slides/patches, pathology reference standard and adjudication, source dataset and duplication checks, train–validation–test partitioning unit, model-selection procedure, test-site independence, thresholds, confusion-matrix counts, sensitivity, specificity, receiver-operating-characteristic area, calibration, and confidence intervals. Missing items should be designated “not reported” or “not verified,” never inferred from image totals. A justified estimate-level appraisal using the QUADAS-3 framework [25], complemented by AI imaging reporting criteria [26], remains to be completed. No independent risk-of-bias scores or pooled effect estimates are claimed.

https://doi.org/10.48161/qjms.v3a90

III. RESULTS: COMPOSITION AND CHARACTERISTICS OF THE EVIDENCE BASE

The preliminary evidence inventory contains 39 distinct topic-specific references covering AI-based oral histopathology, relevant diagnostic applications, source datasets, and clinical context. A provisional classification yields 27 candidate original diagnosis or grading studies, three adjacent original investigations, five secondary reviews, two dataset publications, and two clinical or pathology background sources. The two reporting/quality guidance references and three cross-modal methodological examples are not counted among these 39 topical sources. These figures describe bibliographic roles, not a completed eligibility screen: the 27 candidate studies are neither a confirmed final review sample nor necessarily 27 independent patient cohorts.

Figure 2 visualizes the distribution of source categories, with candidate original classification and grading studies constituting the largest part of the collected literature. This emphasis reflects the range of work on detecting oral squamous cell carcinoma (OSCC), distinguishing normal and malignant tissue, and assessing oral epithelial dysplasia (OED). Nevertheless, secondary reviews, source datasets, and adjacent investigations provide different types of information and cannot contribute additional independent estimates of diagnostic accuracy.

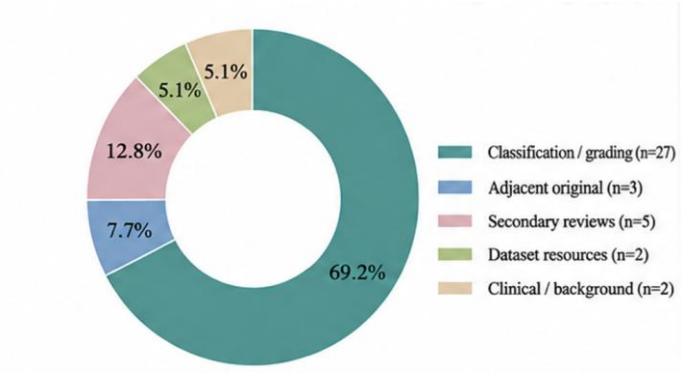


FIGURE 2. Composition of the Preliminary Evidence Inventory

Figure 3 and Table 2 make the evidence-role distinctions explicit. Of the 39 topical references, 27 (69.2%) are candidate original diagnostic or grading reports, and 12 (30.8%) provide adjacent or supporting material. This source-level categorization helps prevent double counting of review papers and dataset records as independent diagnostic evaluations. It does not, however, resolve possible overlap between primary articles that reuse a public repository or related patient material; study-level provenance and near-duplicate screening remain necessary.

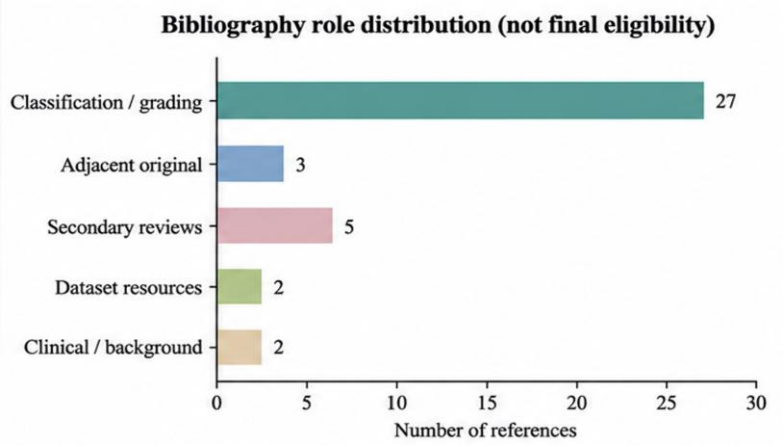


FIGURE 3. Distribution of Bibliographic Sources by Evidence Role

Table 2. Classification of Sources by Evidence Role

Evidence role	Reference numbers	Count	Use in review
Candidate diagnosis / grading	[1–3, 5–21, 27–33]	27	Candidate primary records; full-text eligibility pending
Adjacent original studies	[34- 36]	3	Prognosis, progression, or image acquisition context
Secondary reviews	[22-24, 37, 38]	5	Context and reference discovery; not primary trial data
Dataset publications	[39, 40]	2	Dataset provenance; not classifier evaluations
Clinical / pathology context	[41, 4]	2	Clinical background and observer variation

Table 3 summarizes the 27 candidate primary publications by diagnostic target, model family, selected performance results, and reported evaluation context. It is a descriptive extraction matrix, not an independently verified quality appraisal. Binary cancer detection, multigrade dysplasia assessment, normal/OED/OSCC classification, and whole-slide decisions have different reference standards, class distributions, and potential error costs. Consequently, accuracy figures are reported only within their individual study settings and are not treated as a common-scale basis for comparison.

Table 3. Characteristics and Validation of Candidate Classification Studies

Study	Year	Diagnostic task	Model	Accuracy	Validation
[1]	2023	OSCC vs normal	Custom CNN	97.82%	Internal cross-validation
[2]	2023	OSCC vs normal	ResNet50 transfer	96.60%	Internal
[3]	2023	OSCC diagnosis	ML / deep networks	NR	Not extracted
[5]	2024	OSCC vs normal	Optimized hybrid	98.75%	Internal
[6]	2025	OSCC vs normal	EfficientNet-B3	99.84%	Internal
[7]	2025	Multiclass OSCC	Xception / XAI	NR	Grad-CAM; macro-AUC 0.929
[27]	2026	OSCC vs normal	CNN / ViT models	NR	External; domain shift
[28]	2026	OSCC vs normal	Domain-generalization	NR	External domain shift
[8]	2026	Four-class lesions	CNN–ViT fusion	98.45%	AUC 0.99; multiclass
[9]	2026	OSCC vs normal	EfficientNet-B0	97.60%	Internal benchmark
[29]	2024	OED grading	E-MOD-plus	81.30%	External accuracy 86.5%

Study	Year	Diagnostic task	Model	Accuracy	Validation
[30]	2023	OED grading	Architecture comparison	NR	WSI / annotated regions
[31]	2025	OED grading	Vision Transformer	NR	Grading study
[10]	2025	OSCC vs normal	Texture ML	NR	Feature-based
[11]	2025	Mouth cancer classification	Morphology + U-Net	NR	Segmentation-aided
[12]	2026	OSCC vs normal	Gated CNN–ViT fusion	89.95%	696 images; internal
[13]	2026	Normal/OED/OSCC	EfficientNet-B0	97.48%	Multiclass validation
[14]	2026	Normal / OED grades	DenseNet-121	95.47%	2,080 images; F1 0.94
[15]	2026	OSCC vs normal	Random Forest / CNN	98.80%	CNN; internal
[16]	2026	OSCC vs normal	Stain-robust ML/MIL	87.35%	test set 79.34%
[17]	2026	Weakly supervised OSCC	UNI-2 + MIL	NR	External AUC; audit needed
[18]	2026	OSCC vs normal	5-CNN comparison	NR	Internal comparison
[19]	2026	OSF-associated OSCC	Whole-slide MIL	NR	Internal AUC 0.986
[20]	2026	OSCC vs normal	Attention hybrid ViT	98.97%	External check needed
[32]	2025	Oral cancer detection	XAI classifier	NR	XAI emphasis
[33]	2026	OED grading	CNN	NR	Validation to audit
[21]	2026	Normal/OPMD/OSCC	Multimodal LLM	66.70%	Prompt-based; distinct task

Among binary classification investigations, Das et al. reported 97.82% accuracy with a customized CNN [1], Panigrahi et al. reported 96.60% with ResNet50 transfer learning [2], Ragab and Asar reported 98.75% with an optimized hybrid model [5], and Abd El-Aziz et al. reported 99.84% using EfficientNet-B3 [6]. These are promising within-study outcomes, but differences in test sets, augmentation, optimization, and splitting strategies preclude a valid ranking based on accuracy alone.

The broader candidate set illustrates this limitation. Talo and Ozer reported 89.95% for gated convolutional–transformer fusion [12], whereas Jagathesh and Narayanan reported 98.97% with an attention-based hybrid transformer [20]. Mohapatra and Dash reported 98.80% for a CNN approach [15], and Ren et al. reported 87.35% and 79.34% across different test sets [16]. Such variation supports closer inspection of data provenance, prediction units, reference labels, and external testing, rather than interpreting higher numerical accuracy as inherent architectural superiority.

Multiclass classification and dysplasia grading pose additional challenges because morphological boundaries between classes may be less distinct. Mahanty et al. reported 98.45% accuracy for four-class classification using a hybrid CNN–Vision Transformer [8], and Bharti et al. reported 97.48% validation accuracy for normal, dysplastic, and malignant tissue classification [13]. A separate Bharti investigation reported 95.47% accuracy and macro-F1 of 0.94 for OED grading [14]. Peng et al. reported 81.30% internal and 86.5% external accuracy for dysplasia grading [29]. These estimates need interpretation in the context of grade definitions, annotation agreement, class prevalence, and the independence of evaluation data. Hassanein et al. reported 66.70% accuracy for a multimodal language model distinguishing normal mucosa, potentially malignant disorders, and OSCC [21]. This provides an instructive contrast to specialized supervised classifiers, although prompt-based evaluation and supervised image classification cannot be compared as equivalent experiments. Weakly supervised and whole-slide approaches by Yan et al. [17] and Offurum et al. [19] similarly extend the literature toward slide-oriented evaluation but address different tasks and test units.

Figure 4 presents the year distribution recorded in the 39-source bibliography. The observed concentration in recent publication years describes the assembled reference collection rather than a definitive temporal trend in all published literature. The earlier Rahman dataset publication [39] is retained as a source-provenance reference and does not belong to the primary 2023–2026 study window.

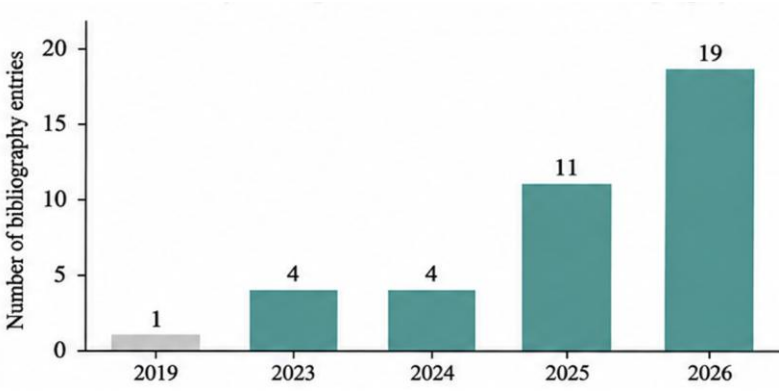


FIGURE 4. Publication-Year Distribution of Bibliographic Sources

Together, Figures 2–4 and Tables 2–3 show that original classification studies dominate the preliminary evidence inventory, while the clinical endpoints and evaluation designs remain heterogeneous. Some studies report high internal accuracy; however, the degree of patient-level independence, quality of reference diagnoses, external-site transport, and confidence-interval reporting are less consistently established in the current source audit. The results should therefore be interpreted as an evidence map that guides formal eligibility screening and task-stratified synthesis, not as the findings of an exhaustive final review.

IV. HISTOPATHOLOGY DATASETS AND DIAGNOSTIC TASKS

Dataset provenance and diagnostic label definitions are major determinants of whether published artificial intelligence (AI) results can be compared meaningfully. The Rahman oral histopathology repository contains 1,224 images from 230 patients, including 290 normal and 934 oral squamous cell carcinoma (OSCC) images acquired at 100× and 400× magnification [39]. These magnification subsets originate from the same repository and should not be treated as independent external validation cohorts. By comparison, the multi-OSCC resource describes 1,325 patients and multiscale tumor-core and tumor-edge images; its prognostic information does not itself constitute an independent diagnostic classifier evaluation [40].

The diagnostic targets also differ. Binary normal-versus-OSCC recognition, oral epithelial dysplasia (OED) grading, discrimination among normal tissue, OED, and OSCC, and recognition of OSCC associated with oral submucous fibrosis (OSF) address different clinical questions. For example, the whole-slide OED study by Peng et al. [29], the four-category grading benchmark by Bharti et al. [14], and the three-class lesion benchmark by Bharti et al. [13] should not be compared directly with binary OSCC image classifiers. Table 4 summarizes how the units of analysis, diagnostic labels, and acquisition settings affect interpretability. Figure 5 provides a visual schematic of normal mucosa and invasive OSCC for anatomical orientation; it does not represent the imaging data used in the reviewed experiments.

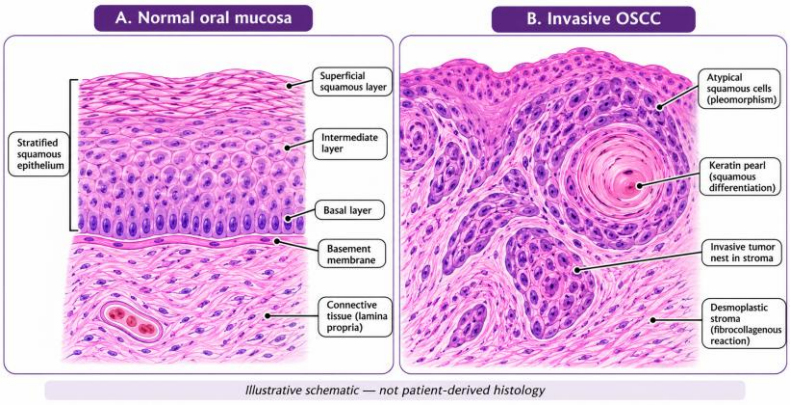


FIGURE 5. Schematic Comparison of Normal Oral Mucosa and Invasive OSCC

Figure 5 depicts an educational comparison of stratified normal oral epithelium and invasive squamous carcinoma, including epithelial disorganization, keratinization, and tumor nests within the underlying stroma. The illustration is synthetic and is not a patient-derived micrograph, study sample, or validation result; clinical interpretation requires examination of actual tissue sections by a qualified pathologist.

Table 4. Histopathology Dataset Characteristics and Comparability

Data source / design	Reported characteristics	Comparability implication
Rahman repository [39]	1,224 images; 230 patients; normal 290, OSCC 934	Binary image collection; repeated use requires checking overlap
Rahman 100× / 400× subsets	Magnification subsets from one parent source	Not independent institutions or external cohorts
Multi-OSCC resource [40]	1,325 patients; tumor core/edge at ×200, ×400, ×1000	Multiscale resource; prognostic endpoints require separation
Peng et al. [29]	56 oral leukoplakia whole slides; external tissue microarrays reported	OED grading; shifts in unit and acquisition source
Bharti et al. [14]	2,080 H&E images; four OED grading categories	Different class definitions from binary OSCC
Bharti et al. [13]	2,363 images; normal / OED / OSCC	Multiclass task; evaluate class imbalance
Rachel et al. [27]	10,817 patches from 114 patients; two external sets	Patient clustering, label adjudication and domain shift

Preprocessing is similarly relevant to the validity of reported performance. Quality control, stain normalization, image tiling, and augmentation can improve model development, but patient- or slide-level separation must occur before creating training-only patches and augmented images. Otherwise, highly similar observations may be distributed across development and test sets. Figure 6 summarizes this recommended workflow; it is a methodological illustration rather than a procedure independently confirmed for every reviewed study.

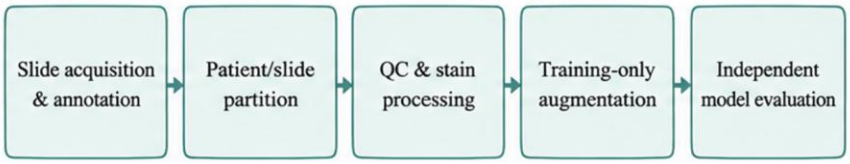


FIGURE 6. Recommended Histopathological Data Processing and Evaluation Workflow

V. ARTIFICIAL INTELLIGENCE CLASSIFICATION ARCHITECTURES

The candidate studies encompass several computational approaches, as summarized in Figure 7. Conventional machine learning combines handcrafted texture or morphological descriptors with classifiers, providing a useful baseline for oral histopathology [10, 15]. Convolutional neural networks (CNNs) and transfer learning have been investigated widely for binary OSCC recognition, including customized CNNs and pretrained ResNet- or EfficientNet-based systems [1–3, 5, 6, 9, 18]. Attention-based and Vision Transformer methods model broader image relationships, while hybrid architectures integrate convolutional features and transformer representations [8, 12, 20, 31]. Whole-slide and weakly supervised approaches aggregate information from multiple patches when detailed region-level annotations are limited [16, 17, 19].

Related methodological work outside oral histopathology provides useful examples of feature integration but does not provide oral-cancer diagnostic evidence. Zebari et al. [42] investigated fusing deep representations for brain-tumor classification using magnetic resonance images, and Salih and Abdulazeez [43] combined complementary deep features in another brain-tumor classification study. Zebari et al. [44] examined multi-fractal feature extraction and classifier fusion for mammographic breast-cancer assessment. These studies motivate consideration of complementary representations and decision fusion, but differences in tissue type, imaging modality, reference standards, and evaluation data preclude transferring their performance estimates to oral squamous cell carcinoma histopathology. They are treated as cross-modal methodological context, not as additional primary studies in the 39-source evidence inventory.

These model categories should be evaluated against their intended diagnostic endpoint, prediction unit, data requirements, and validation design rather than ordered by architectural sophistication. Within-study benchmarks have shown that established CNNs can remain competitive with newer architectures in certain multiclass or dysplasia-grading settings [13, 14]. Moreover, histological grade prediction [34], malignant transformation prediction [35], and super-resolution imaging [36] provide related methodological evidence but are not interchangeable with contemporaneous OSCC diagnosis. The taxonomy in Figure 7 therefore organizes model families conceptually without implying a numerical ranking of diagnostic accuracy.

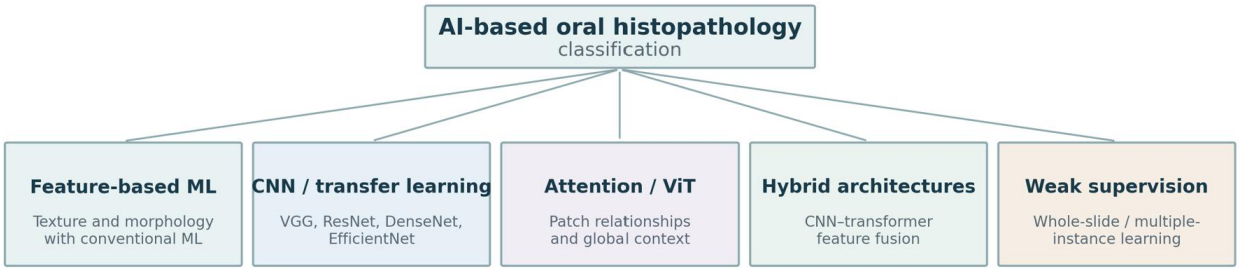


FIGURE 7. Taxonomy of AI Models for Oral Histopathology Classification

VI. COMPARATIVE ANALYSIS OF DIAGNOSTIC PERFORMANCE

Because studies differ in diagnostic endpoint, class composition, model training, and prediction unit, a single overall ranking of classification accuracy would be misleading. The following synthesis separates binary OSCC diagnosis, dysplasia grading, and multiclass lesion classification, then examines external performance and study-specific methodological conditions. Numeric outcomes originate from the working extraction and require cross-checking against original full texts, tables, and supplements before publication. When denominators and corresponding confusion matrices are unavailable, a pooled diagnostic-accuracy estimate cannot be reliably reconstructed.

BINARY OSCC CLASSIFICATION

Within studies designed to distinguish OSCC from normal oral epithelium, several CNN or transfer-learning approaches reported high internal accuracy. Das et al. reported 97.82% using a customized CNN [1], Panigrahi et al. 96.60% using ResNet50 transfer learning [2], Ragab and Asar 98.75% using an optimized hybrid pipeline [5], Abd El-Aziz et al. 99.84% for a learning-rate-tuned EfficientNet-B3 configuration [6], and Kumar et al. 97.60% with EfficientNet-B0 in a within-study architecture comparison [9]. These estimates support technical feasibility in the respective cohorts, but the distinction between internal validation and independent clinical testing remains crucial.

Other reported results demonstrate a wider performance range. Talo and Ozer reported 89.95% accuracy for gated CNN–Transformer feature fusion on a 696-image design [12], whereas Ren et al. reported 87.35% and 79.34% across two separate test datasets [16]. Such differences cannot be attributed to architecture alone: disease spectrum, patient independence, staining and scanner provenance, magnification, image versus slide evaluation, selection of operating threshold, and reference-label procedures may all vary. In particular, studies reusing the same public repository do not establish independent clinical replication merely by reporting new models.

DYSPLASIA GRADING AND MULTICLASS CLASSIFICATION

The evaluation problem becomes more difficult when models must distinguish intermediate epithelial changes. Peng et al. used 56 oral leukoplakia whole slides for OED grading and reported 81.3% internal accuracy, 86.5% external tissue-microarray accuracy, and an external area under the curve (AUC) of 0.669 [29]. The discrepancy between thresholded accuracy and discrimination reinforces the need to inspect class balance, grading definitions, prediction units, confidence intervals, and threshold selection rather than rely on a single metric. Bharti et al. reported 95.47% accuracy and macro-F1 of 0.94 using DenseNet-121 for four-class OED grading [14], while a separate normal/OED/OSCC benchmark reported 97.48% validation accuracy with EfficientNet-B0 [13].

Mahanty et al. reported 98.45% in a four-class histopathology task using combined convolutional and transformer representations [8]; a multimodal language-model assessment by Hassanein et al. reported 66.7% for its specified oral histopathology categories [21]. The tasks, training regimes, and input modalities are not equivalent; the latter result therefore should not be interpreted as a universal measure of multimodal-model capability. The clinically meaningful question is whether the output labels correspond to the intended decision (invasive cancer detection, dysplasia grade, or broader lesion triage) and whether the reference diagnoses have been independently adjudicated.

Table 5. Reported Diagnostic Performance Across Selected Studies

Study	Diagnostic endpoint	Reported result	Principal comparability limitation
[1]	Binary OSCC	97.82% accuracy	Internal evaluation
[2]	Binary OSCC	96.60% accuracy	Study-specific internal partition
[5]	Binary OSCC	98.75% accuracy	Optimization and split design
[6]	Binary OSCC	99.84% accuracy	Tuned training configuration
[9]	Binary OSCC	97.60% accuracy	Within-study benchmark
[12]	Binary OSCC	89.95% accuracy	696-image setting
[16]	Binary OSCC	87.35%; 79.34%	Separate test datasets
[29]	OED grading	81.3% internal; 86.5% external; AUC 0.669	Whole slides versus microarray setting
[13]	Normal/OED/OSCC	97.48% validation accuracy	Three-class task

[14]	Four-class OED	95.47%; macro-F1 0.94	Different label definitions
[8]	Four-class pathology	98.45% accuracy	Class composition and evaluation differ
[21]	Oral pathology three-class	66.70% accuracy	Multimodal prompt-based assessment

As summarized in Table 5, the reported values span markedly different settings and denominators. Figure 8 visualizes these published point estimates while separating binary OSCC diagnosis from multiclass and dysplasia-related outcomes. The lengths of the bars should not be used to rank architectures: apparent performance differences may reflect fundamentally different clinical endpoints rather than algorithmic superiority.

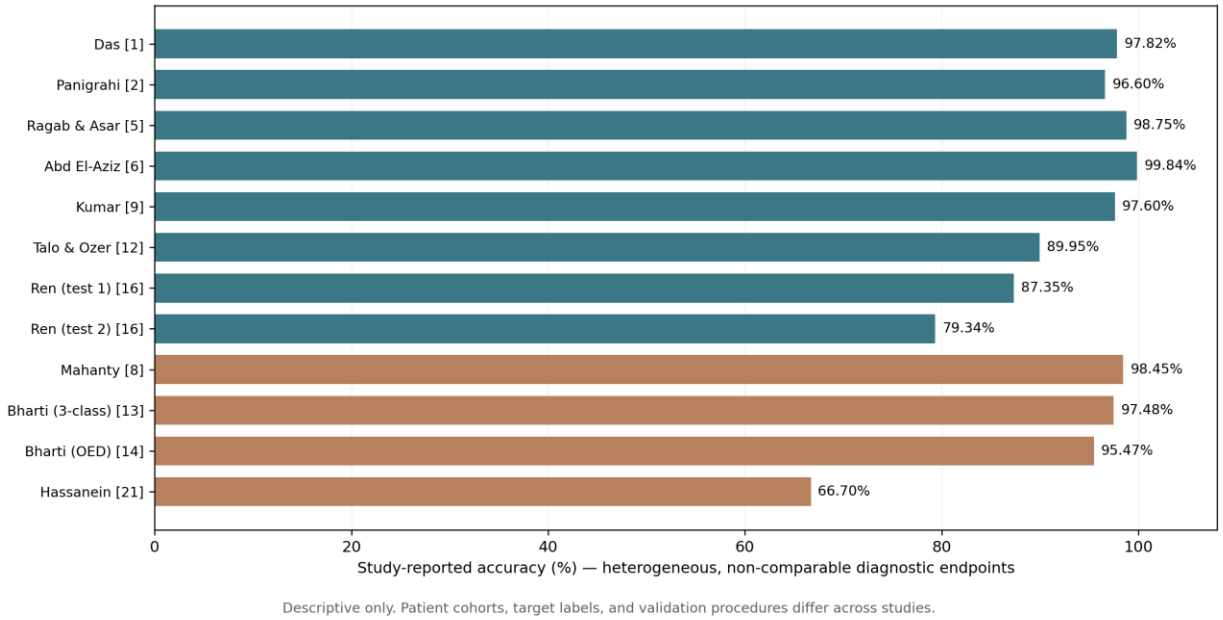


FIGURE 8. Study-Reported Accuracies Across Heterogeneous Diagnostic Tasks

EXTERNAL VALIDATION AND MODEL GENERALIZABILITY

Internal evaluation cannot independently establish that an AI classifier will work on new patients or slides acquired elsewhere. Rachel et al. evaluated multiple CNN and transformer models using 10,817 histopathology patches from 114 patients and reported deterioration on two external datasets, together with problems in diagnostic-label agreement [27]. Hua and Zhou examined cross-source generalization [28], and Ren et al. reported materially different accuracy values between testing datasets [16]. These investigations suggest that staining differences, field of view, institutional practice, selection effects, and reference-label quality can affect performance even when images are all described as oral histopathology.

Table 6 distinguishes studies with specific external or multiple-test-set evidence from purely internal experiments. Figure 9 contrasts within-publication observations for Peng and Ren without confusing external accuracy with external AUC. Peng's higher external accuracy does not negate the AUC of 0.669; accuracy depends on classification thresholds and category proportions, whereas AUC evaluates ranking discrimination across thresholds. Similarly, Ren's two test-set values must not be averaged or interpreted as paired performance without verifying sample construction and uncertainty.

Table 6. External Validation and Test-Set Sensitivity Across Selected Studies

Study	Task / setting	Held-out or external finding	Critical interpretation
[29]	OED grading; external tissue microarray	86.5% external accuracy; AUC 0.669	Accuracy and AUC need complementary interpretation
[16]	OSCC classification; two test datasets	87.35% and 79.34% accuracy	Independent denominators and case mix require audit
[27]	CNN/ViT external testing	Performance deterioration; exact contrast not extracted	Domain shift and expert reassessment of labels
[28]	Generalization across pathology sources	Failure on distinct patient cohort described	Stain invariance does not ensure patient transport
[17]	Weakly supervised slide representations	External AUC described	Model and patient/slide definitions require checking

Explainability addresses an additional but different dimension of validity. Grad-CAM and other spatial displays have been applied in multiclass or hybrid classifiers [7, 8, 17, 32]. Such visualizations indicate model sensitivity to image regions; without pathologist-defined spatial references, robustness tests, and faithfulness experiments, they cannot establish that a classifier has based its prediction on pathognomonic features. A reliable clinical assessment also requires review of false-negative cancer cases, uncertainty calibration, and the consequences of erroneous dysplasia grading.

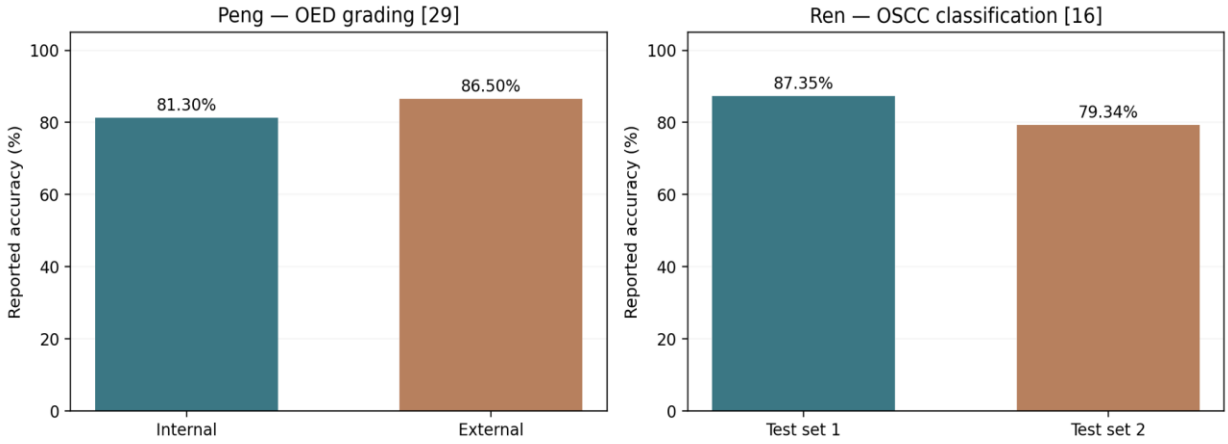


FIGURE 9. Within-Study Accuracy Estimates Across Evaluation Sets

COMPARATIVE METHODOLOGICAL ASSESSMENT

Early convolutional and transfer-learning evidence: Das et al. and Panigrahi et al. illustrated automated OSCC discrimination using custom CNN and pretrained ResNet50 implementations [1, 2]. Ananthakrishnan et al. also compared machine learning and deep architectures [3]. These investigations establish methodological feasibility within their datasets; key unresolved questions include independent

patient sampling, slide-versus-patch partitioning, blinded pathology reference standards, and the breadth of clinical disease represented.

Optimization and network comparisons: Ragab and Asar used an optimized hybrid classifier [5], Abd El-Aziz et al. studied learning-rate tuning in EfficientNet-B3 [6], and Kumar et al. evaluated EfficientNet and alternative CNN families [9]. These provide useful within-study comparisons, but hyperparameter search, test reuse, and the choice of reported best checkpoint must be considered before claiming a diagnostic advance. Prefer matched test sets, confidence intervals and prespecified comparison plans over selecting the highest accuracy reported across unrelated publications.

Intermediate lesions and grading: Peng et al. examined OED in oral leukoplakia [29]; Araújo et al. compared grading approaches [30]; Hadilou et al. investigated transformer representations [31]; and Belaldavar et al. considered CNN-based OED grading [33]. These studies are particularly relevant to clinical uncertainty because dysplasia grading is an ordinal, multi-category judgment affected by pathologist agreement. Independent consensus grading, class-specific sensitivity, severe-dysplasia misclassification, and ordinal error are more informative than pooled image accuracy alone.

Multiclass classification and visual explanation: Štifanić et al. integrated Grad-CAM into a multiclass OSCC framework [7]. Mahanty et al. investigated combined CNN–Transformer learning and multiple explanations [8]. Two distinct Bharti studies evaluated normal/OED/OSCC discrimination and multi-grade OED classification, respectively [13, 14]. Their inclusion clarifies the difference between broad lesion classification and grading within premalignant histology; neither should be treated as an interchangeable binary detection experiment. Reported interpretability should be tested against independently annotated diagnostic regions. Handcrafted features and preprocessing pipelines. Bahaa et al. evaluated texture descriptors [10], Devi et al. used morphological reconstruction and U-Net-related processing [11], Mohapatra and Dash compared conventional and deep learning methods [15], and Balel et al. compared five CNN families [18]. These approaches offer evidence about incremental technical choices. However, preprocessing parameters, augmentation, feature selection, and model tuning must be learned within the training data to prevent contamination of held-out evaluation.

Hybrid attention and whole-slide methods: Talo and Ozer explored gated CNN–Transformer fusion [12], while Jagathesh and Narayanan studied an attention-aided transformer approach [20]. Ren et al. examined robust classification across test sources [16]; Yan et al. investigated weakly supervised slide analysis [17]; and Offurum et al. studied OSCC associated with oral submucous fibrosis at the whole-slide level [19]. Aggregated slide predictions may better match some clinical workflows, but the unit of analysis and lesion definition differ from patch-based detection. Any claim that these models outperform conventional CNNs requires common patients, tasks, and evaluation rules.

Rachel et al. and Hua and Zhou provide evidence relevant to independent testing and domain generalization [27, 28]. Shao et al. addressed tumor grading linked with prognosis [34], Liu-Swetz et al. assessed malignant transformation [35], and Liu et al. studied image super-resolution [36]. These adjacent investigations inform clinical applications and image quality, but should not be entered as independent binary OSCC diagnostic-accuracy estimates. Pushparathi et al. explored explainability [32], while the multimodal language model examined by Hassanein et al. used a distinct prompting and classification design [21]. These differences reinforce the need to specify what clinical decision each algorithm is intended to support.

The central pattern is a mismatch between the precision of reported point estimates and the incomplete information needed to judge clinical transportability. A minimum extractable descriptor set should include the number of independent patients, images and slides per patient, histological reference standard and agreement, disease spectrum, site and scanner, train–test split unit, threshold selection, sensitivity and specificity by class, confidence intervals, and external performance. When a field is unavailable, it should be recorded as not reported, not inferred from image counts or percent accuracy.

Table 7. Methodological Quality Assessment Requirements

Dimension	Evidence required	Potential weakness if absent	Current status
Patient selection	Representative patients; exclusions; disease spectrum	Spectrum and sampling bias	Not assessed by full-text audit
Index test	Locked model; threshold; blinded evaluation; split design	Test leakage or optimism	Not formally assessed
Reference diagnosis	Expert pathology; label adjudication	Label misclassification	Selected studies describe review
Analysis and flow	Evaluation denominator; missing data; intervals	Selective reporting or unit mismatch	Not formally assessed
AI-specific safeguards	Patient/slide independence; outside institution	Near-duplicate contamination, domain shift	External examples discussed [27, 28]
Clinical applicability	Output matches patient-level clinical decision	Patch-to-patient extrapolation	Task-level critique completed

Table 7 clarifies the difference between identifying common threats to validity and completing formal appraisal. Table 8 operationalizes the four QUADAS-3 domains for oral histopathology AI; it identifies what each study must report before study-specific judgments can be justified [25]. CLAIM 2024 can further guide assessment of AI-imaging reporting completeness [26]. Inferences about the percentage of studies at high risk of bias cannot be made until these assessments have been undertaken and documented.

Table 8. Study-Level Diagnostic Quality Appraisal Framework

QUADAS-3 domain	Information to examine	Relevant failure mode
Participants	Recruitment, case spectrum, exclusions, institutional setting	Convenience sampling; repeated patches from few patients
Index Test	Frozen model, threshold, blinding, test isolation	Test-set optimization; unacknowledged model selection
Target Condition	Histological labels, expert confirmation, disagreement	Ambiguous dysplasia grade; uncertain reference labels
Analysis	Independent denominator, missing observations, intervals, analysis unit	Patient leakage; no uncertainty; selective metric reporting

The provisional synthesis supports three evidence-based cautions. First, differences in diagnostic targets and prediction units make headline accuracies unsuitable for a cross-paper leaderboard. Second, robust clinical transport requires verified patient-independent partitioning, reliable pathology reference standards, and truly external testing. Third, explainability and confidence reporting must be evaluated separately from classifier discrimination. The tables in this section specify appraisal requirements; they do not imply that independent, study-level bias ratings or a quantitative meta-analysis have been completed.

VII. EXPLAINABLE ARTIFICIAL INTELLIGENCE AND CLINICAL READINESS

Explainable artificial intelligence (XAI) is increasingly presented as a means of making computational histopathology more transparent. In oral squamous cell carcinoma (OSCC) and oral epithelial dysplasia (OED) classification, gradient-based activation mapping, SHAP feature attribution, and transformer attention visualization can indicate which regions or encoded features influenced a model output. These tools, however, answer a narrower question than diagnostic correctness: a salient region may be correlated with the predicted label without containing the histological evidence a pathologist would use.

Visualization alone cannot establish that a classifier recognizes malignant epithelial invasion, dysplastic architecture, or clinically meaningful cellular atypia [7, 8, 17, 32].



FIGURE 10. Framework for Evaluating AI Explanations in Oral Histopathology

Grad-CAM and Grad-CAM++ yield spatially coarse class-related heatmaps, whose apparent agreement with disease regions can be influenced by layer selection and image resolution. SHAP estimates depend on how features are represented and which background or perturbation assumptions are used, while attention weights do not automatically provide faithful explanations. For an interpretable diagnostic claim, authors should specify the explanation method, define the target class, evaluate localization against independently adjudicated annotations, and test whether the explanation remains stable under benign changes in stain, rotation, or scanner conditions. Reader studies are additionally required to determine whether a displayed explanation improves pathologist accuracy, confidence calibration, or reporting time, rather than simply appearing plausible.

The distinction between visual explanation and demonstrated clinical utility is summarized in Table 9. This table does not score the reviewed studies; instead, it specifies the minimum evidence needed to substantiate an interpretation or deployment claim. Figure 10 places attribution maps between the model decision and two independent evaluation stages: pathological validity and practical clinical benefit. Its key message is that a persuasive heatmap is not, by itself, a validated diagnostic explanation.

Table 9. Evidence Requirements for Explainability and Clinical Evaluation

Method	Available output	Evidence needed for clinical interpretation
Grad-CAM / Grad-CAM++	Class-related activation map	Localization agreement with expert-adjudicated tissue regions; stability tests
SHAP	Attribution to represented features	Feature definitions, reference/background choice, sensitivity of attribution
Transformer attention	Attention across image patches	Faithfulness evaluation; compare against pathological localization
Calibration and uncertainty	Probability confidence or abstention	Calibration curve, error analysis, prespecified referral threshold
Pathologist-assisted decision	Diagnostic prediction with explanation	Blinded or prospective reader evaluation; patient-level benefit and workload

VIII. FRAMEWORK FOR RELIABLE AI-BASED CLASSIFICATION

The evidence synthesized in this review supports a validation-oriented framework that treats data representativeness, reference-label quality, and external transport as prerequisites for trustworthy AI classification. The pathway begins with independent patient and slide sampling, transparent reporting of diagnostic class definitions, and adjudicated histopathological labels. Before augmentation or patch extraction, all images from the same patient, and where appropriate the same biopsy or whole slide, should

be assigned to one data partition. Otherwise, near-duplicate morphology may be shared across training and test sets, creating optimistic performance estimates even when the nominal image counts are large.

Subsequent stages comprise quality-controlled image acquisition, train-only stain and magnification normalization, prespecified model development, and evaluation at clinically meaningful units. Model comparison should use commonly held-out specimens, fixed decision thresholds, and class-specific outcomes rather than selecting the maximum accuracy across unrelated experiments. Sensitivity, specificity, discrimination, calibration and uncertainty should be reported alongside confidence intervals and subgroup results when possible. A locked model should then be tested at independent institutions and across scanner/stain distributions, with label disagreements and performance deterioration documented.

Table 9 and the methodological concerns described in Sections 4–7 motivate the evidence gates depicted in Figure 11. The final stage is a pathologist-supervised reader or prospective workflow study that measures not only classifier performance but also whether AI assistance changes diagnostic decisions, workload, and error patterns. This framework is a structured synthesis of existing limitations rather than a newly trained or externally validated clinical classifier; its utility must be tested in future empirical investigations.

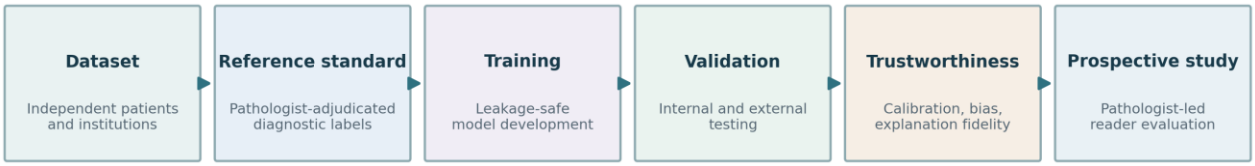


FIGURE 11. Evidence-Based Framework for Clinical Validation of Oral Histopathology AI

IX. RESEARCH GAPS AND FUTURE PRIORITIES

Three linked evidence gaps currently restrict interpretation of AI performance in oral histopathology. First, the studies differ in their target conditions: normal-versus-OSCC discrimination, OED grading, combined normal/OED/OSCC classification, and malignant-transformation prediction are not interchangeable endpoints. Second, the apparent quantity of training data may reflect repeated images, related patches, or reuse of one public repository rather than independent patients or institutions. Third, external testing, reference-label adjudication, probability calibration, and patient-level diagnostic performance are incompletely established for a substantial portion of the candidate evidence base. These limitations prevent defensible universal rankings based on maximum reported accuracy alone.

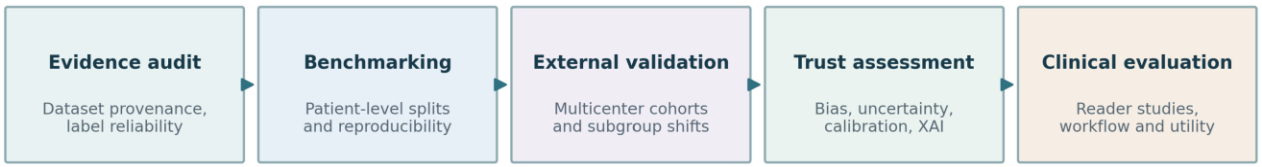


FIGURE 12. Research Priorities for Clinical Translation of Histopathology AI

Table 10 converts these concerns into testable design priorities. Its central recommendation is to prespecify the clinical decision, specimen unit, independent reference standard, and operating threshold before training the model. Prospective multicenter cohorts should report institution-specific sensitivity and specificity with uncertainty intervals; OED grading studies should document class-specific disagreement and adjudication; and explanation studies should compare localization to expert tissue annotations. The sequence in Figure 12 prioritizes data provenance and

reproducible benchmarks before calibration, clinical reader testing, and workflow integration. This ordering is intended to reduce the risk of advancing computationally impressive but poorly transportable systems.

Table 10. Research Gaps and Measurable Evaluation Priorities

gap	Recommended research design	Measurable endpoint
Repeated public image sources	Independent institution and patient-level test cohort	External sensitivity/specificity, 95% confidence intervals
Related patches across splits	Partition patients before patching/augmentation	Zero patient overlap; documented duplicate audit
Imbalanced diagnostic classes	Prespecified class-stratified analysis	Macro-F1, class-specific recall and confusion matrix
Reference-label disagreement	Blinded multi-pathologist adjudication	Agreement statistics and adjudication record
Stain and scanner domain shift	Center-stratified robustness testing	Center-wise discrimination and calibration change
Unvalidated attribution maps	Expert region masks and perturbation tests	Spatial agreement and explanation faithfulness
Unknown clinical impact	Prospective or blinded reader study	Incremental diagnosis quality, workload and decision utility

X. DISCUSSION

The assembled oral histopathology literature demonstrates technical promise across conventional machine learning, CNN transfer learning, hybrid architectures, and slide-level approaches. Nevertheless, a reported accuracy percentage is interpretable only in relation to its target diagnosis, patient spectrum, pathology labels, threshold selection, and unit of prediction. Patch-level recognition of malignant tissue does not itself establish performance for whole-biopsy grading, diagnostic resolution of borderline dysplasia, or improvement in pathologist decisions. Binary OSCC detection, OED grading, multiclass lesion discrimination, and prediction of later malignant transformation therefore require distinct evidentiary interpretations. The comparatively wide variation across selected test settings is better viewed as a signal of design heterogeneity than as proof that one model family is superior.

A second theme is the distinction between technical replication and clinical independence. Public histology datasets are valuable research resources, but repeated experiments on related images can overstate the breadth of independent evidence unless patient overlap, magnification subsets, and near-duplicate patches are audited. Studies using held-out specimens commonly offer stronger within-study model comparisons than cross-paper headline scores. Reports of evaluation at additional institutions or datasets [16, 27, 28] are especially informative, although external labels, disease spectrum, scanners, and case-selection mechanisms must also be inspected. A genuine transportability claim should include prespecified thresholds, site-specific sensitivity and specificity with uncertainty intervals, calibration, and transparent error analysis.

Explainability forms a complementary, not substitutive, evidence dimension. Grad-CAM, SHAP, and transformer attention may illustrate model behavior, but convincing clinical interpretation additionally requires pathology-grounded localization, perturbation-based faithfulness testing, and assessment of whether explanations improve reader decisions. The proposed clinical validation framework in Figure 11 organizes the necessary steps from specimen independence and adjudicated labels to locked external testing and prospective pathologist-assisted evaluation. It is a conceptual pathway, not an experimentally demonstrated model or evidence that AI assistance improves patient outcomes.

These conclusions must be qualified by the review design. The current bibliography was not obtained through a saved comprehensive multi-database export, and independent screening, complete full-text

extraction, and estimate-level risk-of-bias scoring remain outstanding. Selected numerical estimates and bibliographic details also require publisher-level verification. As a result, this manuscript is best characterized as a structured, provisional evidence review, not a completed diagnostic-test-accuracy systematic review or meta-analysis. Section 11 details the resulting limitations and the outstanding tasks for a defensible submission.

XI. LIMITATIONS OF THE REVIEW

This review is limited first by substantive heterogeneity. The candidate publications differ in histopathological source populations, magnification, stain preparation, labels, model architectures, test-set selection, and evaluation units. Because comparable patient-level confusion-matrix data are incompletely available, neither a pooled sensitivity–specificity estimate nor a reliable architecture ranking could be justified. The descriptive accuracy charts consequently summarize published point estimates, not effect sizes measured on a common validation cohort.

Second, the available diagnostic evidence is vulnerable to dataset reuse, related image patches, inadequate reporting of patient-level partitioning, and uneven external evaluation. Variation in the histopathology reference standard and disagreements in OED grading may further affect apparent accuracy. The present inventory does not establish that every candidate study has these weaknesses; it identifies risks that require independent article-level examination rather than unverified generalized judgments.

Third, the review process itself is incomplete. The 39 topical sources were assembled as a working bibliography, not through an archived comprehensive database search with duplicate removal and independently documented eligibility decisions. Full-text verification, independent study selection, and completed estimate-level QUADAS-3 judgments are not yet available. The 27 candidate studies must therefore not be presented as a finalized sample, and the completeness of literature coverage remains unknown.

Finally, emphasis on recent publications may omit foundational studies, and inconsistent primary reporting restricts extraction of denominators, confidence intervals, calibration, and clinically meaningful outcomes. Future work should document database-specific search dates and strings, full screening decisions, traceable extraction, verified primary metrics, and transparent quality appraisal. Until those steps are complete, this review provides a useful preliminary evidence map rather than definitive publication-ready systematic-review conclusions.

XII. CONCLUSION

AI-assisted oral histopathological classification now encompasses handcrafted-feature models, CNN transfer learning, Vision Transformers, hybrid systems, and weakly supervised whole-slide analysis. The collected publications show encouraging performance in several controlled settings, but diagnostic endpoints, histological labels, prediction units, and test conditions differ too much to allow a valid universal ranking. The most consequential research priority is therefore to establish whether a model remains dependable across independent patients, pathology services, and clinical decision contexts, with sensitivity, specificity, calibration, uncertainty, and reader benefit measured transparently. The proposed framework links specimen-level data independence, expert-adjudicated reference diagnoses, external model assessment, explanation faithfulness, and prospective clinical evaluation. Because the present evidence inventory and appraisal remain provisional, the review should be viewed as a structured foundation for a future fully reproducible synthesis rather than conclusive evidence of clinical effectiveness.

XIII. REFERENCES

[1] Das, M., Dash, R., & Mishra, S. K. (2023). Automatic detection of oral squamous cell carcinoma from histopathological images of oral mucosa using deep convolutional neural network. International Journal of Environmental Research and Public Health, 20(3), 2131. https://doi.org/10.3390/ijerph20032131

[2] Panigrahi, S., Nanda, B. S., Bhuyan, R., Kumar, K., Ghosh, S., & Swarnkar, T. (2023). Classifying histopathological images of oral squamous cell carcinoma using deep transfer learning. Heliyon, 9(3), e13444. https://doi.org/10.1016/j.heliyon.2023.e13444

[3] Ananthakrishnan, B., Shaik, A., Kumar, S., Narendran, S. O., Mattu, K., & Kavitha, M. S. (2023). Automated detection and classification of oral squamous cell carcinoma using deep neural networks. Diagnostics, 13(5), 918. https://doi.org/10.3390/diagnostics13050918

[4] Ng, G. T. E., Phang, S. C., Yu, K. S., Tiwari, L., Khurram, S. A., Sloan, P., & Kujan, O. (2025). Understanding interobserver variability of pathologists to improve oral epithelial dysplasia grading. Oral Diseases, 31(3), 838–845. https://doi.org/10.1111/odi.15078

[5] Ragab, M., & Asar, T. O. (2024). Deep transfer learning with improved crayfish optimization algorithm for oral squamous cell carcinoma recognition using histopathological images. Scientific Reports, 14, 25348. https://doi.org/10.1038/s41598-024-75330-3

[6] Abd El-Aziz, A. A., Mahmood, M. A., & Abd El-Ghany, S. (2025). Enhancing early detection of oral squamous cell carcinoma: A deep learning approach with LRT-enhanced EfficientNet-B3 for accurate and efficient histopathological diagnosis. Diagnostics, 15(13), 1678. https://doi.org/10.3390/diagnostics15131678

[7] Štifanić, J., Štifanić, D., Anđelić, N., & Car, Z. (2025). Explainable AI for oral cancer diagnosis: Multiclass classification of histopathology images and Grad-CAM visualization. Biology, 14(8), 909. https://doi.org/10.3390/biology14080909

[8] Mahanty, C., Shieh, C.-S., Horng, M.-F., Nallamalla, A., Patro, S. G. K., Khan, S., & Engida, T. (2026). Explainable vision transformer framework for multi-class classification and prognostic interpretation of oral cancer in histopathology images. Discover Oncology, 17, Article 728. https://doi.org/10.1007/s12672-026-04871-y

[9] Kumar, J., Kumar, M., & Jindal, M. K. (2026). Efficient deep learning models for oral squamous cell carcinoma classification in histopathological images. Scientific Reports, 16, Article 21734. https://doi.org/10.1038/s41598-026-44424-5

[10] Bahaa, S., Ramadan, O., Darwish, Z. E., El-Fiky, E., & Saleh, M. M. (2025). Texture-based image feature analysis for the classification of oral squamous cell carcinoma using a machine learning approach. BMC Oral Health, 26, 105. https://doi.org/10.1186/s12903-025-07482-1

[11] Devi, M. S., Priya, S., Desai, U., Acharya, B., & Moreira, F. (2025). Enhanced histopathologic image analysis for mouth cancer classification using morphological reconstruction and UNet. Journal of Imaging Informatics in Medicine. https://doi.org/10.1007/s10278-025-01717-x

[12] Talo, F., & Ozer, A. B. (2026). Classification of oral squamous cell carcinoma from histopathological images using a hybrid deep learning model. Diagnostics, 16, 2885. https://doi.org/10.3390/diagnostics16182885

[13] Bharti, A., Kashyap, N., Kamboj, M., Joshi, K., Hooda, S., Mishra, D., Singh, H., & Pal, D. (2026). Comprehensive benchmarking of deep learning architectures for multiclass histopathological classification of oral epithelial lesions. Journal of Oral Biology and Craniofacial Research, 16, 101478. https://doi.org/10.1016/j.jobcr.2026.101478

[14] Bharti, A., Kashyap, N., Kamboj, M., Joshi, K., Hooda, S., Mishra, D., Singh, H., & Pal, D. (2026). Deep learning-based automated detection and grading of oral epithelial dysplasia: A comparative histopathological image analysis. Journal of Dentistry, 173, 106771. https://doi.org/10.1016/j.jdent.2026.106771

[15] Mohapatra, A., & Dash, R. (2026). A unified framework for oral squamous cell carcinoma detection using machine learning and deep learning techniques. Discover Computing, 29, 169. https://doi.org/10.1007/s10791-026-10054-8

[16] Ren, Y., Li, R., & Chen, D. (2026). A histopathology image-based computer-aided classification study for oral squamous cell carcinoma. Frontiers in Oncology, 16, 1806332. https://doi.org/10.3389/fonc.2026.1806332

[17] Yan, J., Sun, L., & Wang, S. (2026). A weakly supervised deep learning model for OSCC diagnosis and lesion highlighting in histopathology images. Translational Cancer Research, 15, 531. https://doi.org/10.21037/tcr-2026-0470

[18] Balel, Y., Sağtaş, K., Teke, F., et al. (2026). Deep learning-based automated detection of oral squamous cell carcinoma in histopathological images: A comparative study of five CNN architectures. Odontology. https://doi.org/10.1007/s10266-026-01448-7

[19] Offurum, C. C., Ekanayake, R., Jayasinghe, R. D., et al. (2026). A deep learning framework to classify oral squamous cell carcinoma arising in oral submucous fibrosis using whole-slide images. Journal of Dentistry, 176, 107014. https://doi.org/10.1016/j.jdent.2026.107014

[20] Jagathesh & Narayanan. (2026). Attention-aided hybrid transformer network for oral squamous cell carcinoma classification using histopathology images. Frontiers in Artificial Intelligence. https://doi.org/10.3389/frai.2026.1751520

[21] Hassanein, F. E. A., Hassan, M. T., Javed, M., Hamouda, M. A., Amer, M., Saad, F. A., Madkour, G. G., Mohamed, M. H., & Abou-Bakr, A. (2026). Multiclass diagnostic performance and error patterns of a multimodal large language model in oral histopathology: A WHO-aligned study. Journal of Oral Biosciences, 68(5), 100802. https://doi.org/10.1016/j.job.2026.100802

[22] Pirayesh, Z., Mohammad-Rahimi, H., Ghasemi, N., Motamedian, S.-R., Sadeghi, T. S., Koohi, H., Rokhshad, R., Lotfi, S. M., Najafi, A., Alajaji, S. A., Khoury, Z. H., Jessri, M., & Sultan, A. S. (2024). Deep learning-based image classification and segmentation on digital histopathology for oral squamous cell carcinoma: A systematic review and meta-analysis. Journal of Oral Pathology & Medicine, 53(9), 551–566. https://doi.org/10.1111/jop.13578

[23] Zhu, H.-T., Chen, P.-C., Wang, R.-M., Zeng, Y.-Y., Leng, C.-Y., & Li, Z.-J. (2026). Accuracy of image-based deep learning in oral squamous cell carcinoma diagnosis: A systematic review and meta-analysis. Journal of Evidence-Based Dental Practice, 26(3), 102313. https://doi.org/10.1016/j.jebdp.2026.102313

[24] Sharma, P., Malik, S., Wadhwan, V., Bahri, A., Vishnoi, E., & Malhotra, S. (2026). Artificial intelligence for early detection of oral squamous cell carcinoma: A systematic review. Journal of Stomatology, Oral and Maxillofacial Surgery, 127(6), 103015. https://doi.org/10.1016/j.jormas.2026.103015

[25] Whiting, P. F., Tomlinson, E., Rutjes, A. W. S., et al. (2026). QUADAS-3: A revised tool for the quality assessment of diagnostic test accuracy studies. Annals of Internal Medicine. https://doi.org/10.7326/ANNALS-25-02104

[26] Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., & Kahn, C. E. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiology: Artificial Intelligence, 6(4), e240300. https://doi.org/10.1148/ryai.240300

[27] Rachel, B. R. J., Maraimalai, N., Sood, A., Singh, Y., Jain, T., Raj, B. K., Trimukhe, A., Nayyar, V., Mishra, D., Surya, V., Yadav, R., Singh, H., Sikka, K., & Ahmed, T. (2026). Generalizability of deep learning architectures for oral squamous cell carcinoma classification in histopathology: A comparative CNN–Vision Transformer study with external validation and label-quality analysis. Journal of Imaging Informatics in Medicine. Advance online publication. https://doi.org/10.1007/s10278-026-02355-7

[28] Hua, Z., & Zhou, Y. (2026). Domain-generalizable oral cancer classification: Bridging multi-center histopathological datasets. Frontiers in Medicine, 13, 1951350. https://doi.org/10.3389/fmed.2026.1951350

[29] Peng, J., Xu, Z., Dan, H., et al. (2024). Oral epithelial dysplasia detection and grading in oral leukoplakia using deep learning. BMC Oral Health, 24, 434. https://doi.org/10.1186/s12903-024-04191-z

[30] Araújo, A. L. D., Silva, V. M., Moraes, M. C., et al. (2023). The use of deep learning state-of-the-art architectures for oral epithelial dysplasia grading: A comparative appraisal. Journal of Oral Pathology & Medicine, 52, 980–987. https://doi.org/10.1111/jop.13477

[31] Hadilou, M., Mahdavi, N., Keykha, E., Ghofrani, A., Tahmasebi, E., & Arabfard, M. (2025). Artificial intelligence-based vision transformer application for grading histopathological images of oral epithelial dysplasia: A step towards AI-driven diagnosis. BMC Cancer, 25, 780. https://doi.org/10.1186/s12885-025-14193-x

[32] Pushparathi, V. P. G., Narayan, S. R. S. V., Pratheeba, R. S., & Naveen, V. (2025). Histopathological image analysis and enhanced diagnostic accuracy explainability for oral cancer detection. Cancer Investigation, 43, 726–739. https://doi.org/10.1080/07357907.2025.2559103

[33] Belaldavar, C., Angadi, P. V., Mudenagudi, U., Patil, U., & Tabib, R. A. (2026). Development and validation of a CNN model for grading oral epithelial dysplasia. Revista Española de Patología, 59, 100869. https://doi.org/10.1016/j.patol.2026.100869

[34] Shao, T., Ni, P., Wang, C., Li, J., & Lv, X. (2025). Prediction of pathological grade of oral squamous cell carcinoma and construction of prognostic model based on deep learning algorithm. Discover Oncology, 16, 976. https://doi.org/10.1007/s12672-025-02144-8

[35] Liu-Swetz, Y., et al. (2025). AI-driven prediction of progression to oral squamous cell carcinoma using a multiresolution pathology model. npj Digital Medicine. https://pubmed.ncbi.nlm.nih.gov/41233507/

[36] Liu, Z., Wang, P., Deng, N., et al. (2025). Feasibility study of single-image super-resolution scanning system based on deep learning for pathological diagnosis of oral epithelial dysplasia. Frontiers in Medicine, 12, 1550512. https://doi.org/10.3389/fmed.2025.1550512

[37] Ardila, C. M., Pineda-Vélez, E., & Díaz-Laclaustra, A. I. (2026). Artificial intelligence-assisted histopathologic diagnosis and grading of oral epithelial dysplasia: A systematic review and functional meta-synthesis. Head & Neck Pathology, 20, 74. https://doi.org/10.1007/s12105-026-01937-9

[38] Alajaji, S. A., Khoury, Z. H., Jessri, M., Sciubba, J. J., & Sultan, A. S. (2024). An update on the use of artificial intelligence in digital pathology for oral epithelial dysplasia research. Head & Neck Pathology, 18, 38. https://doi.org/10.1007/s12105-024-01643-4

[39] Rahman, T. Y. (2019). A histopathological image repository of the normal epithelium of the oral cavity and oral squamous cell carcinoma (Version 2) [Data set]. Mendeley Data. https://doi.org/10.17632/ftmp4cvtmb.2

[40] Guan, J., Guo, J., Chen, Q., Chen, J., Cai, Y., He, Y., Huang, Z., Wang, Y., & Xie, Y. (2026). A high-magnification histopathology image dataset for oral squamous cell carcinoma diagnosis and prognosis. Scientific Data, 13, 371. https://doi.org/10.1038/s41597-026-06736-z

[41] Verma, M., Kharbanda, O. P., Mehrotra, R., et al. (2025). NAMS task force report on oral cancer. Annals of the National Academy of Medical Sciences, 61(3), 272–309. https://doi.org/10.25259/ANAMS_TFR_12_2024

[42] Zebari, N. A., Mohammed, C. N., Zebari, D. A., Mohammed, M. A., Zeebaree, D. Q., Marhoon, H. A., Abdulkareem, K. H., Kadry, S., Viriyasitavat, W., Nedoma, J., & Martinek, R. (2024). A deep learning fusion model for accurate classification of brain tumors in Magnetic Resonance images. CAAI Transactions on Intelligence Technology, 9(4), 790–804. https://doi.org/10.1049/cit2.12276

[43] Salih, M. S., & Abdulazeez, A. M. (2024). A fusion-based deep approach for enhanced brain tumor classification. Journal of Soft Computing and Data Mining, 5(1), 183–193. https://doi.org/10.30880/jscdm.2024.05.01.015

[44] Zebari, D. A., Ibrahim, D. A., Zeebaree, D. Q., Mohammed, M. A., Haron, H., Zebari, N. A., Damaševičius, R., & Maskeliūnas, R. (2021). Breast cancer detection using mammogram images with improved multi-fractal dimension approach and feature fusion. Applied Sciences, 11(24), 12122. https://doi.org/10.3390/app112412122