South American Publishing

SAP Artificial Intelligence in Dentistry. 2026;2:107
doi: 10.62486/aid2026107

REVIEW

Diagnostic accuracy of deep learning for detecting periapical lesions on radiographs and cone-beam computed tomography: a systematic review

Precisión diagnóstica del aprendizaje profundo para detectar lesiones periapicales en radiografías y tomografía computarizada de haz cónico: una revisión sistemática

Sandra Ivette López Romo[1]
Olga Peña Silva[1]
Luis Fabián Enríquez Jácome[1]

[1] Universidad Técnica de Ambato, Facultad de Ciencias de la Salud, Carrera de Enfermería. Ambato, Ecuador.

Submitted: 28-01-2026 | Revised: 20-03-2026 | Accepted: 25-05-2026 | Published: 27-05-2026

Corresponding Author: Luis Fabián Enríquez Jácome

ABSTRACT

Introduction: Deep learning models for detecting periapical lesions report near-perfect accuracy under internal validation but much lower sensitivity in clinical conditions, and the magnitude of this gap is unresolved. Methods: A systematic review was conducted following PRISMA 2020. Scopus, PubMed and Crossref were searched up to 2026 for diagnostic accuracy studies of deep learning for periapical lesion detection on periapical radiographs, panoramic radiographs and cone-beam computed tomography against an expert reference standard. Screening was assisted by artificial intelligence and verified by the authors. Results: Twenty-eight primary studies were included. Reported accuracy ranged from 0.84 to 0.96 under internal validation but fell to sensitivities between 33.3% and 93.9% in external or clinical evaluations, with the lowest figures on panoramic radiographs using a commercial tool. Model performance on cone-beam computed tomography was generally higher than on two-dimensional radiographs. Conclusions: Deep learning detects periapical lesions with high accuracy under controlled conditions, but its reported performance is inflated relative to real-world use, particularly on panoramic radiographs. External validation on clinically representative datasets is required before these tools can be recommended for routine diagnosis.

Keywords: artificial intelligence; deep learning; periapical periodontitis; radiography; cone-beam computed tomography; systematic review.

RESUMEN

Introducción: Los modelos de aprendizaje profundo para detectar lesiones periapicales alcanzan una precisión casi perfecta en la validación interna, pero una sensibilidad mucho menor en condiciones clínicas, y la magnitud de esa brecha sigue sin resolverse. Métodos: Se realizó una revisión sistemática según PRISMA 2020. Se buscó en Scopus, PubMed y Crossref hasta 2026 estudios de precisión diagnóstica del aprendizaje profundo para detectar lesiones periapicales en radiografías periapicales, panorámicas y tomografía computarizada de haz cónico frente a un patrón de referencia experto. El cribado fue asistido por inteligencia artificial y verificado por los autores. Resultados: Se incluyeron veintiocho estudios primarios. La precisión notificada osciló entre 0,84 y 0,96 en la validación interna, pero descendió a sensibilidades de entre el 33,3% y el 93,9% en evaluaciones externas o clínicas, con las cifras más bajas en radiografías panorámicas con una herramienta comercial. El rendimiento en tomografía computarizada de haz cónico fue en general superior al de las radiografías bidimensionales. Conclusiones: El aprendizaje profundo detecta lesiones periapicales con alta precisión en condiciones controladas, pero su rendimiento notificado está inflado respecto al uso real, especialmente en radiografías panorámicas. Se requiere validación externa en conjuntos de datos clínicamente representativos antes de recomendar estas herramientas para el diagnóstico habitual.

Palabras clave: inteligencia artificial; aprendizaje profundo; periodontitis apical; radiografía; tomografía computarizada de haz cónico; revisión sistemática.

© 2026; The Authors. This Open access article is distributed under CC BY 4.0 licence.

INTRODUCTION

The published evidence on deep learning for periapical lesion detection is split against itself. On one side, studies that train and test a model on a single curated dataset report accuracy that approaches perfection: Endres et al. reported a detection accuracy of 0.93 in 2020, and Mousavi et al. reported an accuracy of 0.96 with an area under the curve of 1.00 in 2026.(1,2) On the other side, studies that move the same tools into conditions that resemble practice report far weaker performance: Kazimierczak et al. measured a sensitivity of 33.3% for a commercial artificial intelligence system reading panoramic radiographs, and Hadzic et al. measured a sensitivity of 86.7% for a validated model on a clinically representative cone-beam computed tomography set.(3,4) The distance between those two sets of numbers is the unresolved question that this review addresses.

The contradiction matters because periapical lesions are common, consequential and easy to miss. Apical periodontitis is among the most prevalent findings in dental radiography, and a lesion that is overlooked can progress to pain, tooth loss or the failure of endodontic treatment.(5) Two-dimensional radiographs understate the true extent of disease because of anatomical superimposition, and even experienced examiners disagree on the presence and size of a lesion, which is precisely why a reliable automated detector would be valuable. A model that reads radiographs as accurately in a clinic as it does on the dataset that trained it would change endodontic diagnosis; a model that does not is a source of false reassurance, because a clinician who trusts the output may stop looking for what the model missed.

The imaging itself is part of the difficulty, and the three modalities that dominate the literature present different challenges to an automated reader. Periapical radiographs are widely available and show the apex in detail, but they flatten three-dimensional anatomy into two dimensions and can hide a lesion behind superimposed structures. Panoramic radiographs cover the whole dentition in a single exposure and are therefore the natural place to screen for incidental lesions, but their spatial resolution and distortion make small apical radiolucencies hard to resolve. Cone-beam computed tomography removes the superimposition and shows the full three-dimensional extent of a lesion, which is why it is treated as the most informative modality despite its higher radiation dose and cost. A deep learning model inherits all of these properties from the image it is given, so the modality of the input is as important a determinant of performance as the architecture of the model, and any synthesis that ignores modality risks attributing to the model what is in fact a property of the image.

The rapid growth of the literature has not closed the gap. Successive systematic reviews have concluded that deep learning performs well on periapical lesion detection, but several of them pooled studies that used different reference standards, different imaging modalities and different metrics, and none resolved the conflict between internal and external estimates of performance.(6-8)The reviews that did measure pooled sensitivity and specificity reported values that sit between the optimistic and pessimistic extremes, which is what happens when a meta-analysis averages studies that were not measuring the same thing. The commercial tools that clinicians actually use have been evaluated in only a small fraction of the published work, and their real-world behaviour is the part of the evidence that is least well characterised.

The stakes of getting this right are concrete. A missed periapical lesion delays endodontic treatment and can convert a restorable tooth into a non-restorable one, whereas a false positive sends a patient toward an unnecessary intervention or an unneeded cone-beam computed tomography scan. The commercial artificial intelligence platforms that automate this reading are already marketed to dental practices, which means the question is no longer whether deep learning will be used for periapical detection but whether its users understand what the published accuracy figures actually promise. A clinician who reads a headline accuracy of 0.96 and assumes the tool performs at that level in their own practice is acting on a claim the literature has not substantiated, and the discrepancy between the headline and the reality is precisely what this review quantifies.

The objective of this systematic review was to synthesize the diagnostic accuracy of deep learning models for detecting periapical lesions on periapical radiographs and cone-beam computed tomography against expert reference standards, with particular attention to the gap between internal and external validation, and to describe how that gap behaves across imaging modalities.

METHODS

The review was planned and reported in accordance with the PRISMA 2020 statement.(9) Studies were eligible if they reported the diagnostic accuracy of a deep learning or artificial inte-

lligence model for detecting periapical lesions on periapical radiographs, panoramic radiographs or cone-beam computed tomography, against an expert or histologic reference standard, and if they reported at least one of sensitivity, specificity, accuracy or area under the receiver operating characteristic curve. A literature search was conducted using Scopus, PubMed, and Crossref to identify studies published between 2018 and 2026, without restrictions on language. Excluded were studies of caries, periodontal bone loss, implant or root fracture detection without a periapical lesion outcome, purely technical papers without a diagnostic accuracy estimate, reviews, editorials and conference abstracts without primary data. This eligibility boundary was drawn specifically to keep the synthesis centred on the question of lesion detection rather than on adjacent tasks, and the decision is justified by the review question rather than by convention.

The searches combined terms for deep learning, convolutional neural networks and artificial intelligence with terms for periapical lesions, apical periodontitis and periapical radiolucency, and with terms for detection, diagnosis and classification. In Scopus the query used field codes, for example TITLE-ABS-KEY("deep learning" OR "convolutional neural network" OR "artificial intelligence") AND TITLE-ABS-KEY("periapical lesion" OR "apical periodontitis" OR "periapical radiolucency") AND TITLE-ABS-KEY("detection" OR "diagnosis" OR "classification").

The two Scopus queries, together with PubMed and Crossref, returned the set of records that entered screening, and a record of each query, its source and the number of records returned was retained for reproducibility. The two Scopus queries returned 370 records, PubMed returned 200 and Crossref returned 100, and after deduplication the three sources yielded the 488 unique records that entered screening. The full search strings and the number of records returned by each are recorded for reproducibility and are available from the corresponding author on request, together with the screening decisions. The searches were run in 2026, and the last search was performed on the date of analysis. No language filter was applied, and no geographic restriction was imposed beyond the reach of the databases themselves.

Records were deduplicated by digital object identifier and title. Title and abstract screening was assisted by an artificial intelligence model and each decision was verified by the authors, with uncertain records carried forward to full-text screening. Full texts were retrieved through legitimate open-access routes, and studies whose full text could not be obtained but whose abstract reported sufficient data were retained with abstract-level extraction. For each included study the following items were extracted: first author and year, imaging modality, model architecture, sample size, reference standard, and the reported sensitivity, specificity, accuracy and area under the curve. Methodological quality was appraised with a framework adapted from QUADAS-2 and STARD, focusing on four domains that bear directly on the internal-external gap: whether the test set was independent of training, whether the reference standard was expert consensus or histology, whether the study was prospective or retrospective, and whether the sample reflected routine clinical rather than enriched images.(10,11) No study was excluded on the basis of quality, but the quality appraisal is used to interpret the estimates rather than to merge them. Four of the included studies could be extracted only at abstract level because a full text was not obtainable through open access, and their quantitative detail is therefore incomplete; these studies are retained because their abstracts reported diagnostic accuracy figures, but their absence of full methods limits the weight they can carry in the synthesis.

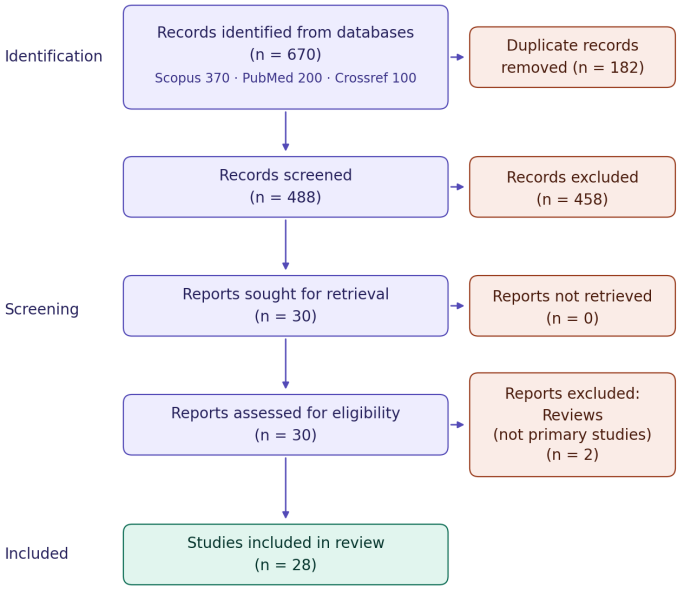
Findings were synthesised narratively by imaging modality and by validation setting. A meta-analysis was not performed because the studies used different architectures, metrics, reference standards and units of analysis, and because pooling internal and external estimates would reproduce rather than resolve the gap that motivated the review. Instead, the synthesis separates internal validation estimates from external and clinical evaluations, and reports individual values as presented in each study. This decision is the analytic counterpart of the review question: the comparison that matters is not between one model and another, but between what a model achieves on its own data and what it achieves on data it has not seen.

The sensitivity and specificity values reported below are those given in the source studies, and no recalculation was attempted when a study reported only one of the two. Because the operating point at which sensitivity and specificity are read can be chosen after the fact, and because several studies reported the two values at different thresholds, the figures should be read as indications of the range of achievable performance rather than as fixed properties of a model. This is a further reason the synthesis avoids a pooled estimate and reports individual values instead, and it is consistent with the appraisal framework that treats the choice of threshold as part of the study design rather than as a neutral reporting detail.

RESULTS

The searches identified 670 records, of which 182 were duplicates; 488 unique records were screened by title and abstract, and 28 studies were included after full-text assessment (Figure 1). Two further reports that were retrieved in full were excluded at that stage because they were reviews rather than primary studies, and the remainder of the exclusions occurred at title and abstract screening.

Figure 1. PRISMA 2020 flow diagram of study selection.



The result that speaks directly to the gap is the contrast between internal and external estimates of accuracy. Under internal validation, on test sets drawn from the same pool as training, the reported accuracy clustered between 0.84 and 0.96. Ba-Hattab et al. achieved a classifier accuracy of 84% with a sensitivity of 81% and a specificity of 86% on panoramic radiographs, Mousavi et al. reached an accuracy of 0.96 with an area under the curve of 1.00 on periapical radiographs, and Erkal et al. reported an accuracy of 90.1% with a sensitivity of 92.8% for a DeepLabV3 segmentation model on periapical radiographs.(2,8,12)

Setzer et al. and Endres et al. reported detection accuracies of 0.93 in the two earliest studies of the set, establishing the ceiling that later internal validations reproduced.(1,7) These figures describe what the models can do when the test set resembles the training set.

The external and clinical evaluations describe something different, and it is here that performance deteriorates. Kazimierczak et al. measured a sensitivity of 33.3% for the commercial Diagnocat platform reading panoramic radiographs, with an F1 score of 32.7%, and a sensitivity of 77.8% for the same platform reading cone-beam computed tomography, against a three-clinician consensus.(3)

Hadzic et al., evaluating a previously validated convolutional neural network on a clinically representative cone-beam computed tomography set of 195 scans, measured an overall sensitivity of 86.7% and a specificity of 84.3%, with sensitivity rising to 90.4% only after excluding the smallest lesions.(4)

Allihaibi et al. measured a sensitivity of 93.9% at tooth level and 86.2% at root level for Diagnocat on cone-beam computed tomography, but a specificity of only 65.2%, which indicates a trade towards over-calling rather than missing.(13)Albochi et al. compared artificial intelligence with periapical radiography and cone-beam computed tomography using cone-beam computed tomography as the reference standard and reported a similar dependence of accuracy on the modality.(14)Issa et al. evaluated Diagnocat on periapical radiographs and found that its sensitivity fell short of the expert readers against whom it was compared.(15)The same pattern of high sensitivity and lower specificity recurred across the external studies, and it is the specificity side, not the sensitivity side, that distinguishes clinical from internal estimates.

The randomised evidence points the same way. Pul et al. randomised thirty dentists to read fifty panoramic radiographs with and without artificial intelligence assistance and found that assistance raised overall accuracy from 91.6% to 93.3%, a modest but significant gain, with the benefit concentrated among less experienced readers.(16)The size of that effect is small relative to the gap between internal and external validation, and it suggests that the value of these tools in practice lies in assisting the reader rather than in replacing the reader.

A sensitivity-specificity trade-off ran through the external studies, and its direction is informative. Where a model was tuned to miss few lesions, it over-called, and where it was tuned to avoid over-calling, it missed. Allihaibi et al. reported the clearest instance of the first pattern, with a tooth-level sensitivity of 93.9% purchased at the cost of a specificity of 65.2%, meaning that the platform flagged many healthy roots as diseased.(13)

Ba-Hattab et al., by contrast, reported a more balanced pair of 81% sensitivity and 86% specificity on panoramic radiographs, and Hadzic et al. reported 86.7% sensitivity with 84.3% specificity on cone-beam computed tomography.(4,8) The commercial platform evaluated by Kazimierczak et al. sat at the conservative end on panoramic radiographs, with a sensitivity of only 33.3% but a high specificity, which describes a tool that alarms in few cases but likewise detects few lesions.(3)

None of these operating points is intrinsically correct; the point is that the choice of threshold, and therefore the balance between missing and over-calling, is a clinical decision that the studies made in different ways and that a reader must recover from the reported pair rather than from a single accuracy figure.

The prevalence of lesions in the test sets also shaped the reported values, and few studies stated it explicitly. When a dataset is enriched with lesions, the positive predictive value is inflated and the negative predictive value is deflated relative to what a screening population would produce, which is why the same sensitivity and specificity translate into different clinical meaning in a specialist endodontic clinic and in a general screening setting.

The studies that drew their images from routine clinical workload rather than from curated lesion collections are therefore the more informative for the screening use case, and they are the studies whose sensitivity figures were lowest. This prevalence effect, which the diagnostic accuracy literature treats under the heading of spectrum bias, is a further reason that the internal validation figures should not be transferred directly to a clinical setting without adjustment.

The architectures and the models they produced were many. Object-detection models of the YOLO family were the most common: Hirata et al. developed a YOLOv8 framework, Saber et al. compared YOLOv8, YOLOv11 and YOLOv12 for periapical index scoring and reported precisions of 88.5% and 89.1% for the two newer versions, Oztekin et al. added ordinal severity grading to a YOLO framework for panoramic radiographs, and Shahmaleki et al. applied YOLOv7 to paediatric panoramic radiographs.(17-20)

U-Net and its variants dominated the cone-beam computed tomography and segmentation studies: Setzer et al. used a U-Net for lesion segmentation in cone-beam computed tomography, Kirnbauer et al. used a two-step U-Net for osteolytic lesion detection, Öksüzoğlu et al. used nnU-Net v2 for three-dimensional segmentation, and Chau et al. proposed CBCT-SAM for lesion detection.(7,21-23)Other architectures included ResNet and ConvNeXt for classification,(24-26)a self-designed grading network that outperformed ResNet variants on cone-beam computed tomography periapical index classification,(27)and pretrained transformers for segmentation and lesion detection.(28)

Boztuna et al. segmented periapical lesions on panoramic radiographs with a deep learning model, Szabó et al. detected lesions on panoramic radiographs, and Hamdan et al. evaluated a no-code model for detecting periapical radiolucencies.(5,29,30)Vision-language models entered the evidence only recently: Akyüz et al. and Hassanein et al. each compared a general-purpose model against expert endodontists, with high agreement on some categories but variable per-lesion accuracy.(31,32)

Within individual studies, the between-model differences were consistently smaller than the between-setting differences, which is itself informative. Saber et al. found the three YOLO versions nearly equivalent, with precisions of 88.5% and 89.1% for the two newest, and Erkal et al. found DeepLabV3 and HRNet separated by only a few percentage points on the same dataset.(12,18) Zhao et al. reported that their self-designed PAINet outperformed the ResNet baselines, and Liu et al. reported that the integrated YoCNET improved on a single classification model, but in each case the gains from the architecture were modest relative to the gains from moving to a stronger reference standard or a more representative test set.(25,27) The contrast between these small architectural differences and the large internal-external differences supports the conclusion that the architecture is not where the field's uncertainty lies, and that the next meaningful improvement in reported accuracy will come from better evaluation rather than from a better network.

The scale of the studies and the unit of analysis differed widely, and both affect how the numbers should be read. The smallest studies used a few dozen images and the largest used more than a thousand, with sample sizes ranging from 20 cone-beam computed tomography volumes in the earliest study by Setzer et al. to 1,305 periapical radiographs in the classification studies by Liu and colleagues.(7,24,25)

Some studies reported accuracy at the level of the tooth, others at the level of the root or the lesion, and a few at the level of the whole image, which means that a sensitivity of 90% in one study is not the same quantity as a sensitivity of 90% in another. The studies that reported lesion-level metrics on small, augmented datasets, such as the ensemble study by Mousavi et al., produced the most optimistic figures, whereas the studies that reported tooth- or root-level metrics on large, clinically representative datasets produced the more modest ones.(2,4) This scaling pattern is consistent with the internal-external gap and reinforces that the gap is partly a consequence of how the studies were constructed rather than of any single model's weakness.

The temporal spread of the evidence tracks the evolution of the field. The two earliest studies, published in 2020, established U-Net and convolutional classification as the reference approaches, and the years that followed added progressively more capable architectures, from ResNet and VGG16 through the YOLO object detectors to the transformers and vision-language models of 2026.(1,7,28,31)

Ten of the twenty-eight studies appeared in 2026 alone, which reflects how quickly the literature is growing and how recently the field reached its current volume. This recency is a double-edged property: it means the review captures a fast-moving field at a specific point, but it also means that the newest models, whose performance may differ from the models reported here, are the least well validated, because external validation studies lag behind the release of new architectures. The conclusion that internal estimates outpace external ones is therefore unlikely to be overturned by newer models, since every new model enters the same pattern of internal enthusiasm followed by slower, more sober external testing.

Across imaging modalities, performance on cone-beam computed tomography exceeded performance on two-dimensional radiographs in the studies that compared them directly. The commercial platform in the study by Kazimierczak et al. more than doubled its sensitivity when moving from panoramic to cone-beam computed tomography, and the cone-beam computed tomography models developed by Öksüzoğlu et al. and Chau et al. reported segmentation and detection accuracies at the upper end of the range.(3,22,23)

The eleven studies that used cone-beam computed tomography, either alone or in comparison, reported higher and more stable estimates than the seven studies restricted to periapical radiographs or the six restricted to panoramic radiographs, which is consistent with the additional three-dimensional information that cone-beam computed tomography provides. The cone-beam computed tomography studies reported the most technically advanced pipelines, and their results show the ceiling that three-dimensional imaging permits. Setzer et al. achieved a voxel-level segmentation that allowed lesion detection accuracy of 0.93 with a specificity of 0.88 in the earliest U-Net implementation, and Kirnbauer et al. extended the same two-step tooth-localisation and segmentation approach to a larger routine clinical dataset.(7,21)

Öksüzoğlu et al. reported fully automated three-dimensional segmentation with nnU-Net v2, and Chau et al. proposed CBCT-SAM specifically for lesion identification, both achieving detection metrics near the top of the range reported in this review.(22,23) The grading and classification studies on cone-beam computed tomography, including the PAINet grading system of Zhao et al. and the VGG16 transfer-learning classification of Akalin et al., addressed the additional question of severity rather than presence, and their accuracies were likewise higher than the two-dimensional detection studies.(26,27) The consistency of this pattern across different architectures and reference standards is what makes the modality conclusion robust rather than the product of a single favourable study.

The methodological quality of the studies, appraised on the four domains adapted from QUADAS-2, varied in ways that map onto the internal-external gap. The test set was reported as independent of training in most studies, but the independence consisted in a random split of a single dataset rather than a genuinely separate acquisition, which is a weaker form of independence than external validation.

The reference standard was expert consensus in every study that reported it, and histologic confirmation was absent, so the ground truth itself carries measurement error for small or subtle lesions. The studies were almost entirely retrospective, and the samples were enriched for lesions rather than drawn consecutively from clinical workload, which raises the apparent prevalence and inflates the apparent accuracy. These quality characteristics are reported here rather than used

to exclude studies, but they explain why the internal estimates cluster near the ceiling while the external estimates do not, and they are the reason the synthesis treats the two kinds of estimate separately.

The characteristics of the twenty-eight included studies are summarised in the table, without repeating them here (Table 1).

Table 1. Characteristics of the 28 included studies.

Study (year)	Modality	Model	Reference standard	Reported accuracy
Setzer (2020)	CBCT	U-Net	Expert	0.93
Endres (2020)	PR	CNN	Expert	0.93
Kirnbauer (2022)	CBCT	U-Net	Expert	NR
Ba-Hattab (2023)	PAN	CNN	Expert	0.84
Issa (2023)	PR	Diagnocat	Expert	NR
Boztuna (2024)	PAN	Deep learning	Expert	NR
Liu (2024)	PR	YoCNET	Expert	NR
Zhao (2024)	CBCT	PAINet/ResNet	Expert	NR
Chen (2024)	CBCT	Transformer	Expert	NR
Kazimierczak (2024)	PAN/CBCT	Diagnocat	Expert	33.3% / 77.8% sensitivity
Hadzic (2023)	CBCT	CNN	Expert	86.7% sensitivity
Liu (2025)	PR	ConvNeXt/ResNet34	Expert	NR
Saber (2025)	PR	YOLOv8/11/12	Expert	88.5-89.1% precision
Akalin (2025)	CBCT	VGG16	Expert	NR
Pul (2025)	PAN	AI assistance	CBCT	91.6% vs 93.3%
Allihaibi (2025)	CBCT	Diagnocat	Expert	93.9% sensitivity
Szabó (2025)	PAN	Deep learning	Expert	NR
Chau (2025)	CBCT	CBCT-SAM	Expert	NR
Shahmaleki (2026)	PAN	YOLOv7	Expert	NR
Erkal (2026)	PR	DeepLabV3/HRNet	Expert	90.1%
Öksüzoğlu (2026)	CBCT	nnU-Net v2	Expert	NR
Hamdan (2026)	PR	No-code AI	Expert	NR
Oztekin (2026)	PAN	YOLO	Expert	NR
Akyüz (2026)	PR	Vision-language	Expert	NR
Hirata (2026)	PR	YOLOv8	Expert	NR
Hassanein (2026)	PR	Vision-language	Expert	NR
Mousavi (2026)	PR	Ensemble CNN	Expert	0.96
Albochi (2026)	PR/CBCT	AI	CBCT	NR

Note: PR = periapical radiograph; PAN = panoramic radiograph; CBCT = cone-beam computed tomography; NR = not reported in the abstract.

DISCUSSION

Contrary to the expectation that deep learning would have settled periapical lesion detection by now, the evidence shows that the field has not reconciled its internal and external estimates, and the discrepant authors named in the introduction represent two ends of a spectrum rather than two isolated outliers. Endres and Mousavi, who reported near-perfect accuracy, trained and tested on curated single-centre datasets in which the prevalence and appearance of lesions were controlled.(1,2)

Kazimierczak and Hadzic, who reported the lowest figures, tested commercial or previously validated models on images that reflect the mix of anatomy, image quality and disease that a clinic actually presents.(3,4) The reconciliation is therefore not that one set of authors is wrong, but that they are measuring different things, and the honest summary of the evidence is that deep learning is excellent at the task it was trained for and markedly weaker at the task a dentist needs it for.

The reason the gap has persisted is that most of the literature reports internal validation as if it were the whole story. The studies that reported the highest accuracies used training and test sets drawn from the same distribution, which is the classic source of optimism in diagnostic accuracy studies of machine learning, and the QUADAS-2 domain that captures this, the applicability of the sample to the target population, was the domain most frequently violated.(10)

The commercial tools, which are the ones most likely to reach a clinician, were evaluated in only a handful of the included studies, and their performance was systematically worse than the custom models reported elsewhere, which suggests that the gap is partly a gap between academic models and deployed products.(3,13)

The reference standard itself contributes to the residual disagreement, because expert consensus rather than histology was used in almost every study, and expert readers themselves disagree on small or subtle lesions.(33)

A further reason the estimates diverge is the absence of a shared benchmark. Each study assembled its own dataset, defined its own reference standard and chose its own operating point for reporting sensitivity and specificity, so the numbers cannot be placed on a common scale, and the field has no equivalent of the public benchmarks that have stabilised other areas of medical imaging. The studies that used the same commercial platform, Diagnocat, are the closest thing to a common reference, and it is telling that they produced the most consistent and the most modest results across different centres.(3,13,15)

When a common tool is evaluated under a common protocol, the reported performance converges toward the lower end of the range, which is itself evidence that the higher figures in the single-centre studies are inflated by their construction rather than by superior models. Establishing a public, labelled periapical dataset with an agreed reference standard would do more to resolve the question than any number of additional single-centre reports.

Nonetheless, the modality comparison offers a clear and useful conclusion. Performance on cone-beam computed tomography was consistently higher and more stable than on two-dimensional radiographs, and the difference is explained by the three-dimensional information that cone-beam computed tomography provides, which removes the superimposition that obscures lesions on periapical and panoramic radiographs.(22,23)

Whereas the gap between internal and external validation is a problem to be solved, the modality difference is a property of the imaging rather than the model, and it argues that the highest-value deployment of these tools is on cone-beam computed tomography, where the marginal benefit to the reader is largest.

What would close the gap is not a better architecture but a better evaluation. The ideal study is prospective, samples patients consecutively rather than lesions selectively, uses an independent reference standard that is itself calibrated, reports sensitivity and specificity at a prespecified threshold rather than the best achievable, and states the unit of analysis so that its numbers can be compared with others. Hadzic et al. came closest to this design among the included studies, and its sensitivity of 86.7%, lower than almost every internal validation, is the most credible single estimate in the set.(4)

Until more studies follow that model, the headline accuracies will continue to overpromise, and clinicians will continue to have to discount the published numbers when deciding whether to trust a tool with a patient's diagnosis.

The findings can be set against the earlier reviews of this literature. The meta-analyses that pooled studies reported favourable summary accuracy, but they did not separate internal from external validation and therefore inherited the optimism of the single-centre studies they combined.(6,8,34-40)

The reviews that restricted themselves to a single modality, such as panoramic radiography or cone-beam computed tomography, reached more consistent conclusions than those that pooled across modalities, because the modality itself changes the achievable performance. The present review reaches a more guarded conclusion precisely because it makes the internal-external separation, and it treats the randomised and external validation studies as the most informative rather than as outliers. This difference in analytic strategy, rather than a difference in the underlying studies, is what produces the more cautious verdict offered here.

The limitations of the evidence are integrated into these comparisons rather than separated from them. The studies were almost entirely retrospective, the samples were frequently enriched for lesions rather than drawn consecutively from clinical workload, and the unit of analysis varied between teeth, roots and images, all of which make the numeric estimates difficult to compare and, more importantly, tend to flatter the models.

The geographic concentration of the studies, most of which drew their datasets from a sin-

gle institution or country, raises a further question about how well the models generalise across populations, image acquisition equipment and practice patterns, since a model trained on one hospital's radiographs may behave differently on another's, and none of the included studies tested this directly. The two randomised and the handful of external validation studies are the most credible evidence in the set, and they are also the studies that produced the most modest numbers.

The implications follow directly. A dentist who adopts one of these tools should expect it to assist rather than replace the reader, and should place most trust in it on cone-beam computed tomography, where the evidence is strongest. A researcher who reports a new model should distinguish internal from external validation explicitly and should not present an internal test-set accuracy as a clinical claim.

The external studies show that figures above 0.90 obtained on single-centre, enriched datasets do not survive contact with clinical images, and treating them as clinical evidence would overstate the tool. The field would be better served by fewer, larger external validation studies than by further single-centre internal validations, and by a shared reference standard that would allow the estimates to be pooled rather than compared impressionistically.

The commercial context adds a layer that the academic literature only partly captures. The platforms that automate periapical lesion detection are marketed to dental practices on the strength of accuracy claims, yet the evidence reviewed here shows that the published external evaluations of at least one widely used platform are materially weaker than the internal figures that circulate in the research literature.

A dentist who purchases such a tool is therefore buying a claim that the peer-reviewed evidence has not yet established, and the discrepancy between marketing and validation is itself a finding that bears on how these products should be regulated. Diagnostic software that influences treatment decisions is a medical device in several jurisdictions, and the weakness of the external evidence reviewed here argues that post-market validation, and not only pre-market accuracy, should be a condition of continued use.

The strengths of the review should be set against these limitations. The search spanned three major databases with Scopus as the primary source, the eligibility boundary was drawn explicitly around periapical lesion detection so that the synthesis does not drift into adjacent tasks, and the analytic separation of internal from external validation is a deliberate feature of the design rather than a post hoc device.

The inclusion of the randomised trial and the commercial-platform evaluations, which are often excluded from technical reviews, is what allows the review to speak to clinical use rather than only to model development. The reporting of individual study values rather than a pooled estimate is a limitation in one sense, because it forgoes a single summary number, but it is also the property that makes the internal-external gap visible at all, and a summary estimate would have concealed the very finding the review set out to describe.

The practical workflow that the evidence supports is a second reader, not a primary one. The randomised trial by Pul et al. showed that the benefit of artificial intelligence assistance is real but small and is concentrated in less experienced readers, which means the tool adds most value when it flags a lesion that a junior or distracted reader might miss, and least value when it is allowed to replace a trained eye.(16)

On panoramic radiographs, where the commercial tools performed worst, the output should be treated as a screening prompt rather than a diagnosis, and any flagged lesion should be confirmed on cone-beam computed tomography or with a targeted periapical radiograph before treatment is planned. On cone-beam computed tomography itself, the models are accurate enough that their segmentation and detection can plausibly save time and improve consistency, but even there the specificity figures show that over-calling is a genuine risk that the reader must correct. These recommendations are the concrete translation of the internal-external gap into clinical behaviour, and they are more specific than the generic caution that the tools should be used carefully.

CONCLUSIONS

Deep learning detects periapical lesions with high accuracy under controlled internal validation, but its reported performance overstates what the same models achieve on clinically representative images, where sensitivity falls to between 33.3% and 93.9% depending on modality and product.

The gap between internal and external validation is the central finding of this review and the central caution for clinical use, and it is widest on panoramic radiographs and narrowest on

cone-beam computed tomography, where the three-dimensional information produces the most consistent and most credible estimates.

The randomised evidence shows that these tools are best deployed as a second reader that flags rather than decides, and the commercial platforms in particular should not be judged on the strength of internal accuracies alone. Until external validation on unenriched, prospective datasets with calibrated reference standards becomes the norm, these tools should be treated as assistants to the dentist rather than as independent diagnostic readers.

REFERENCES

1. Endres MG, Hillen F, Salloumis M, Sedaghat AR, Niehues SM, Quatela O, et al. Development of a Deep Learning Algorithm for Periapical Disease Detection in Dental Radiographs. Diagnostics. 2020;10(6):430. doi: 10.3390/diagnostics10060430
2. Mousavi SE, Jafari M, Tahmasbi M, Razavi M, Shooshtari SS. Stacked Ensemble Deep Learning Models for Accurate Detection and Size Stratification of Periapical Lesions on Intraoral Radiographs. Aust Endodontic J. 2026:aej.70113. doi: 10.1111/aej.70113
3. Kazimierczak W, Wajer R, Wajer A, Kiian V, Kloska A, Kazimierczak N, et al. Periapical Lesions in Panoramic Radiography and CBCT Imaging—Assessment of AI's Diagnostic Accuracy. JCM. 2024;13(9):2709. doi: 10.3390/jcm13092709
4. Hadzic A, Urschler M, Press JNA, Riedl R, Rugani P, Štern D, et al. Evaluating a Periapical Lesion Detection CNN on a Clinically Representative CBCT Dataset—A Validation Study. JCM. 2023;13(1):197. doi: 10.3390/jcm13010197
5. Szabó V, Orhan K, Dobó-Nagy C, Veres DS, Manulis D, Ezhov M, et al. Deep Learning-Based Periapical Lesion Detection on Panoramic Radiographs. Diagnostics. 2025;15(4):510. doi: 10.3390/diagnostics15040510
6. Basso Á, Salas F, Hernández M, Fernández A, Sierra A, Jiménez C. Machine learning and deep learning models for the diagnosis of apical periodontitis: a scoping review. Clin Oral Invest. 2024;28(11):600. doi: 10.1007/s00784-024-05989-5
7. Setzer FC, Shi KJ, Zheng Z, Yan H, Yoon H, Mupparapu M, et al. Artificial Intelligence for the Computer-aided Detection of Periapical Lesions in Cone-beam Computed Tomographic Images. Journal of Endodontics. 2020;46(7):987-993. doi: 10.1016/j.joen.2020.03.025
8. Ba-Hattab R, Barhom N, Osman S, Naceur I, Odeh A, Asad A, et al. Detection of Periapical Lesions on Panoramic Radiographs Using Deep Learning. Applied Sciences. 2023;13(3):1516. doi: 10.3390/app13031516
9. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021:n71. doi: 10.1136/bmj.n71
10. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011;155(8):529-536. doi: 10.7326/0003-4819-155-8-201110180-00009
11. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015:h5527. doi: 10.1136/bmj.h5527
12. Erkal D, Çakmak YE, Er K, Kuştarcı A. Automated periapical lesion segmentation and area-based PAI indexing: a comparative deep learning study on periapical radiographs. BMC Oral Health. 2026;26(1):1290. doi: 10.1186/s12903-026-08645-4
13. Allihaibi M, Koller G, Mannocci F. Diagnostic accuracy of an artificial intelligence-based platform in detecting periapical radiolucencies on cone-beam computed tomography scans of molars. Journal of Dentistry. 2025;160:105854. doi: 10.1016/j.jdent.2025.105854
14. Albochi M, Kamel W, Sabet N, ElBoraey M. Diagnostic accuracy of artificial intelligence in detecting periapical lesions: A comparative study with cone-beam computed tomography and conventional periapical radiography. J Clin Exp Dent. 2026;18(10):e1156-e1163. doi: 10.4317/jced.64327
15. Issa J, Jaber M, Rifai I, Mozdziak P, Kempisty B, Dyszkiewicz-Konwińska M. Diagnostic Test Accuracy of Artificial Intelligence in Detecting Periapical Periodontitis on Two-Dimensional Radiographs: A Retrospective Study and Literature Review. Medicina. 2023;59(4):768. doi: 10.3390/medicina59040768
16. Pul U, Tichy A, Pitchika V, Schwendicke F. Impact of artificial intelligence assistance on diagnosing periapical radiolucencies: A randomized controlled trial. Journal of Dentistry. 2025;160:105868. doi: 10.1016/j.jdent.2025.105868
17. Hirata N, Panyarak W, Klanliang K, Charuakkra A, Wantanajittikul K. Development of a YOLOv8-based deep learning framework for detection and classification of periapical radiolucencies in dental radiographs using the periapical index. J Assoc Med Sci. 2026;59(2):316-326. doi: 10.66285/jams.2026.063
18. Saber S, Abou El Nasr H, Torky AA, Saif N. Automated assessment of periapical health based on the radiographic periapical index using YOLOv8, YOLOv11, and YOLOv12 one-stage object detection algorithms. Sci Rep. 2025;15(1):36487. doi: 10.1038/s41598-025-21761-5
19. Oztekin F, Katar O, Avci N, Konus F. Automated Detection and Ordinal Severity Grading of Periapical Lesions on Panoramic Radiographs Using a Lesion-Preserving Tiling and Ordinal-Aware YOLO Framework. Diagnostics. 2026;16(17):2778. doi: 10.3390/diagnostics16172778
20. Shahmaleki P, Gezginci PA, Kırelli Y, Yuce F, Buyuk C. Automated detection of periapical lesions in pediatric panoramic radiographs using YOLOv7: a retrospective internal validation study. BMC Pediatr. 2026;26(1):846. doi: 10.1186/s12887-026-07354-9
21. Kirnbauer B, Hadzic A, Jakse N, Bischof H, Stern D. Automatic Detection of Periapical Osteolytic Lesions on Cone-beam Computed Tomography Using Deep Convolutional Neuronal Networks. Journal of Endodontics. 2022;48(11):1434-1440. doi: 10.1016/j.joen.2022.07.013
22. Öksüzoğlu H, Topuz Ö, Bilgir E, Çelik Ö, Bayrakdar IŞ. Fully automated 3D segmentation of periapical lesions on CBCT images using nnU-Net v2. BMC Med Imaging. 2026;26(1):240. doi: 10.1186/s12880-026-02316-0
23. Chau KK, Zhu M, AlHadidi A, Wang C, Hung K, Wohlgemuth P, et al. A novel AI model for detecting periapical lesion on CBCT: CBCT-SAM. Journal of Dentistry. 2025;153:105526. doi: 10.1016/j.jdent.2024.105526

24. Liu J, Jin C, Wang X, Pan K, Li Z, Yi X, et al. A comparative analysis of deep learning models for assisting in the diagnosis of periapical lesions in periapical radiographs. BMC Oral Health. 2025;25(1):801. doi: 10.1186/s12903-025-06104-0
25. Liu J, Liu X, Shao Y, Gao Y, Pan K, Jin C, et al. Periapical lesion detection in periapical radiographs using the latest convolutional neural network ConvNeXt and its integrated models. Sci Rep. 2024;14(1):25429. doi: 10.1038/s41598-024-75748-9
26. Akalin F, Özkan Y. Deep Learning-Based Hybrid Scenario for Classification of Periapical Lesions in Cone Beam Computed Tomography. Symmetry. 2025;17(9):1392. doi: 10.3390/sym17091392
27. Zhao T, Wu H, Leng D, Yao E, Gu S, Yao M, et al. An artificial intelligence grading system of apical periodontitis in cone-beam computed tomography data. Dentomaxillofac Radiol. 2024;53(7):447-458. doi: 10.1093/dmfr/twae029
28. Chen RQ, Lee Y, Yan H, Mupparapu M, Lure F, Li J, et al. Leveraging Pretrained Transformers for Efficient Segmentation and Lesion Detection in Cone-Beam Computed Tomography Scans. Journal of Endodontics. 2024;50(10):1505-1514.e1. doi: 10.1016/j.joen.2024.07.012
29. Boztuna M, Firincioglulari M, Akkaya N, Orhan K. Segmentation of periapical lesions with automatic deep learning on panoramic radiographs: an artificial intelligence study. BMC Oral Health. 2024;24(1):1332. doi: 10.1186/s12903-024-05126-4
30. Hamdan M, Yu KW, Miller M, Sidow SJ, Badr Z, Uribe SE. Evaluation of a no-code AI model for detecting periapical radiolucencies: impact of anatomical region on diagnostic performance. BMC Oral Health. 2025;26(1):52. doi: 10.1186/s12903-025-07529-3
31. Akyüz İE, Kıvırcık BE, Aslan T. Artificial Intelligence-assisted Periapical Radiographic Assessment: Lesion Detection, Endodontic Complication Analysis, and Review of Clinical Treatment Recommendations. Journal of Endodontics. 2026;52(7):1194-1205. doi: 10.1016/j.joen.2026.04.008
32. Hassanein FEA, Hussein RR, Elgarhy MR, Maher SM, Hassen A, Heidar S, et al. Artificial Intelligence Versus Human Dental Expertise in Diagnosing Periapical Pathosis on Periapical Radiographs: A Multicenter Study. Bioengineering. 2026;13(2):232. doi: 10.3390/bioengineering13020232
33. Alkhattab OR, Bokhary A. Reference Standards in Artificial Intelligence Detection of Apical Periodontitis: A Narrative Review of What Reported Accuracy Measures. Diagnostics. 2026;16(19):3225. doi: 10.3390/diagnostics16193225
34. Pul U, Schwendicke F. Artificial intelligence for detecting periapical radiolucencies: A systematic review and meta-analysis. Journal of Dentistry. 2024;147:105104. doi: 10.1016/j.jdent.2024.105104
35. Sadr S, Mohammad-Rahimi H, Motamedian SR, Zahedrozegar S, Motie P, Vinayahalingam S, et al. Deep Learning for Detection of Periapical Radiolucent Lesions: A Systematic Review and Meta-analysis of Diagnostic Test Accuracy. Journal of Endodontics. 2023;49(3):248-261.e3. doi: 10.1016/j.joen.2022.12.007
36. Tasdemir T, Tassoker M, Arpacı A, Kula A, Kocabiyik FE, İyişenyürek ME, et al. Diagnostic performance of artificial intelligence in periapical radiography: a systematic review. Odontology. 2026;114(4):1848-1864. doi: 10.1007/s10266-026-01401-8
37. Ahmad IA, Ba-Hattab R, Al-Maweri SA, Tamimi F. Deep Learning Models for Detection of Periapical Radiolucent Lesions on Panoramic Radiographs: A Systematic Review and Meta-Analysis. Int Endodontic J. 2026:iej.70207. doi: 10.1111/iej.70207
38. da Silva-Filho JE, Sousa ZDS, Caracas-de-Araújo AP, Fornagero LDS, Machado MP, de Aguiar AWO, et al. Deep learning for detecting periapical bone rarefaction in panoramic radiographs: a systematic review and critical assessment. Dentomaxillofac Radiol. 2025;54(6):405-419. doi: 10.1093/dmfr/twaf044
39. Bansal RK, Arya A, Singh B, Singla M. Comparative Efficiency of Artificial Intelligence in Detecting Periapical Lesions Using Periapical Radiographs, Orthopantomogram, and Cone-beam Computed Tomography: A Scoping Review. Indian Journal of Dental Sciences. 2025;17(1):32-37. doi: 10.4103/ijds.ijds_132_24
40. Septina F, Kiswanjaya B. Evaluating Deep Learning AI for Periapical Lesion Detection Across Panoramic, Periapical, and CBCT Radiographs: A Systematic Review. TODENTJ. 2026;20(1):e18742106391708. doi: 10.2174/0118742106391708260406104035

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

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