



Diagnostic accuracy of artificial intelligence for detecting periodontal bone loss on dental radiographs: a systematic review

Precisión diagnóstica de la inteligencia artificial para detectar pérdida ósea periodontal en radiografías dentales: revisión sistemática

Arlin Castro Diaz¹  

¹ Universidad Iberoamericana Santo Domingo. Santo Domingo, República Dominicana.

Submitted: 01-03-2025 | Revised: 20-04-2025 | Accepted: 11-07-2025 | Published: 14-08-2025

Corresponding Author: Arlin Castro Diaz 

ABSTRACT

Introduction: Periodontitis is staged in part by radiographic alveolar bone loss, whose manual measurement is subjective and time-consuming. Artificial intelligence has been proposed to automate this measurement, but reported accuracy varies widely. **Methods:** Following PRISMA 2020, PubMed, OpenAlex, and Crossref were searched for studies published from 2018 to 2025 that evaluated artificial intelligence models for detecting, staging, or measuring periodontal bone loss on panoramic or periapical radiographs against a reference standard. Risk of bias was assessed with QUADAS-2. **Results:** Thirty-eight studies were included, most using panoramic radiographs and convolutional neural networks. Reported sensitivity ranged from 0% to 100%, specificity from 66.2% to 98%, and the area under the curve from 0.70 to 0.951. Staging and measurement studies reported accuracy from 58% to 96.8%. **Conclusions:** Artificial intelligence detects periodontal bone loss with high but heterogeneous accuracy. The evidence is strongest for detection on panoramic radiographs and weakest for severity staging and clinical measurement, and external validation remains uncommon.

Keywords: artificial intelligence; alveolar bone loss; periodontitis; radiography panoramic; diagnosis computer-assisted; systematic review.

RESUMEN

Introducción: La periodontitis se estadia en parte por la pérdida ósea alveolar radiográfica, cuya medición manual es subjetiva y laboriosa. Se ha propuesto la inteligencia artificial para automatizar esta medición, pero la precisión reportada es muy variable. **Métodos:** Siguiendo PRISMA 2020, se buscó en PubMed, OpenAlex y Crossref estudios publicados entre 2018 y 2025 que evaluaran modelos de inteligencia artificial para detectar, estadiar o medir la pérdida ósea periodontal en radiografías panorámicas o periapicales frente a un estándar de referencia. El riesgo de sesgo se evaluó con QUADAS-2. **Resultados:** Se incluyeron treinta y ocho estudios, la mayoría con radiografías panorámicas y redes neuronales convolucionales. La sensibilidad reportada osciló entre 0% y 100%, la especificidad entre 66,2% y 98%, y el área bajo la curva entre 0,70 y 0,951. Los estudios de estadificación y medición reportaron exactitud entre 58% y 96,8%. **Conclusiones:** La inteligencia artificial detecta la pérdida ósea periodontal con una precisión alta pero heterogénea. La evidencia es más sólida para la detección en radiografías panorámicas y más débil para la estadificación de la severidad y la medición clínica, y la validación externa sigue siendo poco común.

Palabras clave: inteligencia artificial; pérdida de hueso alveolar; periodontitis; radiografía panorámica; diagnóstico por computador; revisión sistemática.

INTRODUCTION

Periodontitis is staged in part by the radiographic extent of alveolar bone loss, yet the measurement that underpins this staging is the least reproducible element of the diagnosis. Two clinicians examining the same radiograph frequently disagree on the percentage of bone loss at a given site, and this disagreement propagates directly into the assigned stage and into treatment planning.⁽¹⁾ Artificial intelligence has been proposed as a way to make this measurement objective and repeatable, but the studies that evaluate it report results that span the full range from barely usable to apparently expert-level, which leaves clinicians without a clear expectation of what these tools can do.

The appeal of automating bone loss measurement is concrete. Manual measurement requires the examiner to locate the cemento-enamel junction, trace the bone crest, and express the ratio as a percentage, steps that are slow and operator-dependent even for calibrated readers. The steps are identical in every case, and a model that learns them from annotated examples can repeat them without fatigue and without the between-examiner variation that characterizes human measurement. A model that performs these steps reliably could standardize staging across clinicians, flag underdiagnosed sites, and support epidemiological screening. For this reason, the number of primary studies has grown rapidly since 2018, when the first deep learning models for radiographic periodontal assessment appeared.^(2,3)

The consequences of misclassification are clinical and economic. Underestimating bone loss can leave progressive disease untreated, whereas overestimating it can lead to unnecessary periodontal therapy or extraction. The 2017 classification was designed to improve prognostic precision, but its radiographic component relies on a measurement that few clinicians perform identically. Automation therefore promises not only speed but consistency, and the question addressed here is whether that consistency comes at an acceptable cost in accuracy.

The evaluation of these models follows the framework of diagnostic accuracy studies, in which an index test, the model, is compared against a reference standard, most often expert annotation. The central difficulty is that the reference standard for radiographic bone loss is itself imperfect, so a model can appear inaccurate because it disagrees with an imperfect reference rather than because it is wrong. This distinction, which recurs throughout the results, is essential to interpreting the reported numbers.

The staging framework that these measurements feed into is itself standardized. The 2017 classification of periodontal and peri-implant diseases defines stages I through IV partly by the percentage of radiographic bone loss at the site of greatest loss, with numeric thresholds that separate mild from moderate from severe disease. Because these thresholds are numeric, a model that measures bone loss directly can in principle assign a stage automatically, which is precisely what several of the included studies attempt.⁽⁴⁻⁶⁾ The clinical value of the models therefore depends on whether their measurements are accurate in millimeters, and not merely on whether they detect the presence of disease.

Earlier reviews have summarized parts of this literature. A review of artificial intelligence for radiographic detection of periodontal bone loss concluded that the technology is efficient and broadly accurate,⁽⁷⁾ and a review of models for diagnosing gingivitis and periodontal disease reached a similarly favorable but cautious conclusion.⁽⁸⁾ More recent reviews of artificial intelligence for periodontitis classification and diagnosis report high pooled accuracy but note the same recurring problems: heterogeneous reference standards, a predominance of single-center datasets, and few external validations.⁽⁹⁻¹¹⁾

What is missing is a synthesis restricted to the radiographic measurement of bone loss itself, which is the task with the most direct clinical consequence for staging, and one that distinguishes detection from staging from measurement. The aim of this systematic review was to synthesize how accurately artificial intelligence models detect and stage periodontal bone loss on panoramic and periapical radiographs and to characterize the architectures, reference standards, and methodological quality of the included studies.

METHODS

Protocol and registration

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 statement.⁽¹²⁾

Eligibility criteria

Studies were eligible when they were original investigations that developed or evaluated an artificial intelligence, machine learning, or deep learning model for detecting, staging, segmenting, or measuring periodontal or alveolar bone loss on panoramic or periapical radiographs and reported at least one quantitative performance measure against a reference standard. Accepted measures were sensitivity, specificity, area under the receiver operating characteristic curve, Dice or intersection over union, accuracy, F1-score, intraclass correlation coefficient, or mean measurement error. Studies whose primary input was cone-beam computed tomography or an intraoral photograph were excluded, as were reviews, editorials, and studies reporting no quantitative metric. The publication window was 2018 to 2025. The task definitions were deliberately broad. Detection comprised studies that identified the presence of bone loss, staging comprised studies that assigned a severity class, measurement comprised studies that estimated bone level or the percentage of bone loss, and segmentation comprised studies that delineated the bone or landmark pixels, with studies combining tasks assigned to all relevant categories.

Information sources and search

PubMed, OpenAlex, and Crossref were searched for records published between 2018 and 2025. Queries combined terms for artificial intelligence, periodontal or alveolar bone loss, and dental radiography, with several synonymous strings run in each source to maximize recall. The PubMed query set comprised three strings: ("artificial intelligence" OR "deep learning" OR "machine learning") AND ("periodontal bone loss" OR "alveolar bone loss" OR periodontitis) AND (radiograph OR panoramic OR periapical) AND (detection OR staging OR diagnosis OR classification); "deep learning" AND ("bone loss" OR periodontitis) AND ("panoramic radiograph" OR periapical); and "convolutional neural network" AND (periodontal OR "alveolar bone") AND (radiograph OR radiography). Equivalent free-text queries were adapted to the OpenAlex and Crossref syntax, and each source returned up to 50 records per query. Records were deduplicated by digital object identifier and by title before screening.

Study selection

Screening proceeded in two stages. Title and abstract screening excluded clearly irrelevant records, and uncertain records were carried forward rather than discarded. The title and abstract screen applied the eligibility criteria conservatively, so that any record whose relevance could not be determined was retained for full-text assessment. Full text was available for a subset of reports; for the remainder, eligibility and data extraction relied on the abstract.

Data extraction

For each included study the following items were extracted: first author and year, country, model architecture, imaging projection, sample size, reference standard, task, and reported diagnostic performance. Values were recorded as reported, without recalculation, and when a study reported separate values for different lesion severities or imaging projections, the stratum-specific values were retained. When a study reported staging accuracy for the 2017 classification, the stage-specific values were recorded, and when a study reported agreement statistics, the intraclass correlation coefficient or the kappa was retained alongside the accuracy. No values were imputed, and studies for which the performance could not be extracted were excluded at the eligibility stage rather than included with missing data.

Risk of bias assessment

Risk of bias and applicability were assessed with QUADAS-2 across patient selection, the index test, the reference standard, and flow and timing.⁽¹³⁾ Because most studies reported no prospective registration and no independent external validation, risk of bias was summarized narratively by domain.

Synthesis

A meta-analysis was not performed because the studies reported heterogeneous tasks, metrics, and reference standards. Detection studies reported sensitivity and specificity, staging studies reported classification accuracy, and measurement studies reported agreement coefficients and error in millimeters, and these cannot be combined into a single estimate without strong and unjustified assumptions. Findings were synthesized narratively, grouped by imaging projection and

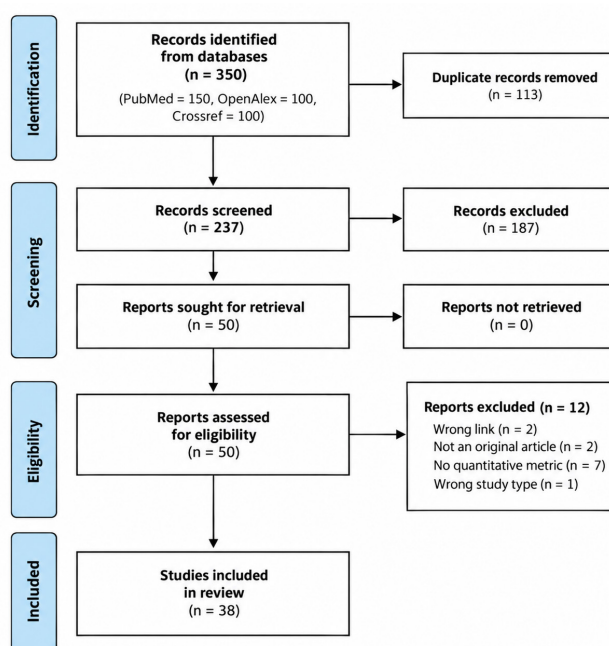
by task, and heterogeneity was described rather than statistically tested.

RESULTS

Study selection

The searches retrieved 350 records, of which 113 duplicates were removed and 237 unique records were screened (Figure 1). Title and abstract screening excluded 187 records, leaving 50 reports assessed for eligibility. Of these, 12 were excluded: 7 because no extractable quantitative metric was reported, 2 because the imaging input was not a two-dimensional dental radiograph, 2 because the outcome was not periodontal bone loss, and 1 because the report was not an original study. Thirty-eight studies were included(38).

Figure 1. PRISMA 2020 flow diagram of study selection.



Characteristics of included studies

The included studies spanned 2018 to 2025, with the largest number published in 2025 (11 studies) and 2024 (10 studies), and the earliest report in 2018 (Table 1). Panoramic radiographs were the most common input, used in 23 of 38 studies, and periapical radiographs in the remaining 15. In terms of task, 20 studies performed detection of bone loss, 9 performed severity staging, 7 performed quantitative measurement of bone level, and 2 performed pixel-level segmentation. The detection studies used bounding boxes or classification labels as their output, and the segmentation and measurement studies used pixel masks or landmark coordinates, so the format of the model output, and therefore the metric used to evaluate it, differed across the studies.

Convolutional neural networks dominated the architectures. Object detectors based on the YOLO family and region-based detectors such as Faster R-CNN and Mask R-CNN were the most frequent.(14-20) Segmentation networks based on U-Net were also common,(21-23) and several studies applied classification networks such as VGG-16, Inception, and ResNet.(24-27) A minority used transformer architectures(28,29) or a multimodal large language model.(30)

The reference standard was expert annotation in most studies, whether a single periodontist, a calibrated panel, or a consensus of specialists. Clinical periodontal assessment served as the reference in a subset,(22,31,32) and a few studies measured agreement against manual annotation.(33,34) Sample sizes ranged from fewer than 100 images(30,35) to more than 20,000 radiographs,(28,29) and only a minority reported external validation on an independent cohort.(18,36)

Table 1. Characteristics of the 38 included studies evaluating artificial intelligence for periodontal bone loss on dental radiographs.

First author, year	Country	Architecture	Imaging	Sample	Reference standard	Task
Al Husaini 2025	NR	InceptionV3/V4, ResNet-50	panoramic	NR	NR	staging
AlGhaihab 2025	NR	Denti.AI (CNN)	periapical	39 radiographs (316 surfaces)	consensus of 3 specialists	detection
Camlet 2025	NR	ChatGPT 4.5, o1, o3, o4-mini-high	panoramic	10 panoramic (252-271 sites)	three examiners	measurement
Do 2025	Vietnam	YOLOv8	panoramic	500 panoramic	NR	staging
Hameedaldeen 2025	NR	Overjet AI	periapical	103 patients	periodontist clinical + radiographic gold standard	detection
Ibraheem 2025	NR	Second Opinion v1.1 (CAdE)	periapical	300 periapical (1030 teeth)	two oral radiologists	detection
Li 2025	China	HC-Net+	panoramic	10,881 OPGs (pretrain); 382 clinical + 760 test	periodontal specialists	detection
Li 2025	NR	object detection CNN	panoramic	558 panoramic (7359 teeth)	NR	staging
Piroonsan 2025	NR	CNN (VGG19 best)	periapical	1369 radiographs (556 patients)	NR	detection
Rezallah 2025	NR	MobileNet-v2 + YOLOv8	panoramic	817 panoramic	oral radiology specialist	detection
Widyaningrum 2025	NR	Mask R-CNN + DenseNet169	panoramic	600 panoramic (+100 external)	NR	staging
Amasya 2024	NR	DiagnoCat (Mask R-CNN + Cascade R-CNN)	panoramic	~100 radiographs	3-clinician consensus	detection
Ameli 2024	Canada	U-Net + YOLOv9	periapical	1000 train; 1582 test	clinical assessments	measurement
Chen 2024	NR	deep CNN	periapical	336 periapical (390 teeth)	three periodontists	staging
Dai 2024	NR	AlexNet/VGG16/ResNet18 + RF/SVM/NB/LR/KNN	periapical	NR	NR	staging
Jundaeng 2024	NR	YOLOv8	panoramic	2000 panoramic	NR	segmentation
Kong 2024	South Korea	SRGAN + U-Net	panoramic	NR	NR	segmentation
Kurt-Bayrakdar 2024	NR	U-Net	panoramic	1121 panoramic	expert segmentation	detection
Mardini 2024	Chile	CNN + statistical inference	panoramic	2010 images (40 test)	radiologists	measurement
Xue 2024	NR	YOLOv8 + Mask R-CNN + TransUNet	panoramic	320 patients (8462 teeth)	three periodontists	staging
Yu 2024	China	SegFormer	panoramic	NR	clinicians	measurement
Chen 2023	NR	YOLOv5 + VGG16 + U-Net ensemble	periapical	8000 periapical (27,964 teeth)	clinicians	detection
Dujic 2023	NR	ViT, BEiT, DeiT	periapical	21,819 periapical	calibrated dentists	detection
Hoss 2023	NR	5 CNNs	periapical	21,819 periapical	trained dentists	detection
Kong 2023	NR	PDCNN (two-stage CNN)	panoramic	NR	periodontitis annotations	detection
Ryu 2023	NR	Faster R-CNN	panoramic	4083 panoramic	examiner consensus (ICC 0.94)	detection
Vollmer 2023	NR	Keypoint R-CNN (ResNet-50-FPN)	panoramic	1414 + 341 external	ground truth labels	detection
Alotaibi 2022	Saudi Arabia	VGG-16	periapical	1724 periapical (1610 patients)	NR	detection
Jiang 2022	China	UNet + YOLOv4	panoramic	640 panoramic	3 periodontists	staging
Lee 2022	USA	DL segmentation (bone/tooth/CEJ)	periapical	NR	independent examiners	measurement

First author, year	Country	Architecture	Imaging	Sample	Reference standard	Task
Tsoromokos 2022	NR	CNN	periapical	1546 approximal sites (54 participants)	manual annotation	measurement
Danks 2021	UK	Hourglass network	periapical	340 radiographs (63 patients)	clinicians	measurement
Li 2021	China	Deetal-Perio (Mask R-CNN)	panoramic	NR	NR	staging
Bayrakdar 2020	Turkey	GoogleNet Inception v3	panoramic	2276 panoramic	clinical status	detection
Thanathornwong 2020	NR	Faster R-CNN	panoramic	100 panoramic	expert annotations	detection
Kim 2019	South Korea	DeNTNet	panoramic	12,179 panoramic	experienced clinicians	detection
Krois 2019	Germany	deep CNN (7-layer)	panoramic	2001 segments	three examiners	detection
Lee 2018	NR	deep CNN	periapical	1740 radiographs	clinical diagnosis	detection

Note: NR = not reported; CNN = convolutional neural network; OPG = orthopantomogram; CEJ = cementoenamel junction. Source: Compiled from data extracted from the included studies.

The studies also differed in the anatomical landmarks they measured. Most models localize the cementoenamel junction and the alveolar bone crest and express bone loss as the ratio between the two, which mirrors the clinical definition of radiographic bone loss.(21,34,37) A few studies measured bone loss at specific sites, such as approximal surfaces,(33) or combined the radiographic measurement with a clinical stage assigned according to the 2017 classification.(31,32) The choice of landmark and of unit of analysis, whether site, tooth, or patient, differs across studies and complicates the comparison of their reported errors.

A further source of variation is the landmark on which the measurement is anchored. The cementoenamel junction is difficult to localize on panoramic radiographs, and some studies anchor the measurement on the tooth apex or on adjacent structures instead, which changes the resulting bone level percentage. The studies that reported the lowest agreement with the reference standard cited this difficulty explicitly,(33,38) and the studies that reported the highest agreement used a controlled periapical projection in which the landmarks are more visible.(22,34)

The studies also originated predominantly from Asia and Europe, with China, South Korea, and Turkey contributing several reports each, and a smaller number from North America, South America, and the Middle East.(3,22,25,26,32,39) The geographic concentration is relevant because panoramic machines, exposure protocols, and patient demographics differ across regions, and a model validated in one setting may not transfer to another. Sample sizes also grew over time: early studies used a few hundred radiographs,(16,38) whereas recent studies used several thousand, and two large periapical studies shared a dataset of more than 21,000 radiographs.(28,29)

Risk of bias

Across QUADAS-2 domains, the reference standard was the principal source of concern. When bone loss is labelled by expert readers, the model is evaluated against human judgement rather than against the true bone level, and inter-reader agreement for bone loss is only moderate, so the target itself contains error. A minority of studies used clinical assessment or a calibrated consensus as a stronger reference,(31,32) and none used a criterion independent of human readers. Patient and image selection raised further concerns because cohorts were assembled retrospectively from single institutions in most studies, and few reports described consecutive or random sampling. Flow and timing were generally acceptable, but blinding was reported inconsistently, and only a minority held out an independent test set.(18,32,36) The applicability of the findings to routine practice is also uncertain. Most models were trained on images from a single institution and a single acquisition device, and performance measured internally may not transfer to a different panoramic machine or to a different population. The two studies that reported an external validation cohort illustrate this. One found the external performance lower than the internal estimate,(18) and the other reported specificity and accuracy that were retained on an independent set.(36) These mixed results suggest that external validation, where performed, tends to reduce the apparent accuracy, which is the expected direction for any model evaluated outside its training distribution. Taken together, the risk-of-bias appraisal indicates that the reported accuracy should be interpreted as an upper bound.

The combination of expert reference standards, retrospective convenience sampling, and internal test sets inflates apparent performance relative to what would be observed in routine practice, and only a handful of studies took the steps, such as external validation or a clinical reference, that would temper this inflation.

Diagnostic accuracy

Table 2. Reported diagnostic performance of artificial intelligence models for periodontal bone loss.

First author, year	Reported performance
Al Husaini 2025	accuracy 96.8% (ResNet-50)
AlGhaihab 2025	sensitivity 76%/65%; specificity 86%/90%; accuracy 81%/76%
Camlet 2025	tooth-count kappa 0.52-0.69; deemed insufficient for clinical use
Do 2025	precision 0.95-0.97; recall 0.94-0.96 (bone level/CEJ)
Hameedaldeem 2025	sensitivity 82-100%; specificity 89-96%
Ibraheem 2025	sensitivity 91.1%; specificity 93.1%
Li 2025	AUC 94.2% vs 85.6% specialists; accuracy >92.4%
Li 2025	AUC 0.79; accuracy 0.72; mAP 0.88
Piroonsan 2025	sensitivity 0.82; specificity 0.67; AUC 0.70-0.77; accuracy 0.75
Rezallah 2025	sensitivity 0.7-1.0; accuracy 0.88
Widyaningrum 2025	specificity 0.88; accuracy 0.80
Amasya 2024	accuracy 0.980 / 0.993; kappa 0.956-0.974
Ameli 2024	accuracy 95.62%; ICC >0.94
Chen 2024	accuracy 0.728; 0.97 (stage III); PCC 0.828
Dai 2024	accuracy 0.872-0.968
Jundaeng 2024	sensitivity 90-100%; specificity 96-98%; accuracy 97-98%
Kong 2024	Dice 0.91; IoU 84.9%
Kurt-Bayrakdar 2024	sensitivity 1.0; AUC 0.951; accuracy 0.994
Mardini 2024	sensitivity 0.23 (light-moderate); 0 (severe)
Xue 2024	accuracy 89.45%; ICC 0.806
Yu 2024	sensitivity 0.742; IoU 0.789; F1 0.867
Chen 2023	accuracy 0.97 (bone loss); 0.90 overall
Dujic 2023	AUC 0.899-0.918; accuracy 83.4-85.2%
Hoss 2023	sensitivity 88.8-90.7%; specificity 66.2-71.2%; AUC 0.884-0.913
Kong 2023	accuracy 0.762
Ryu 2023	AUC 0.91
Vollmer 2023	AP 0.694-0.758; recall 0.611
Alotaibi 2022	accuracy 0.73 (binary); 0.59 (severity)
Jiang 2022	accuracy 0.77
Lee 2022	Dice 0.91; AUC 0.89-0.90; accuracy 0.85
Tsoromokos 2022	ICC 0.601; mean %ABL 23.1 vs 27.8
Danks 2021	accuracy 0.58; mean PBL error 10.69%
Li 2021	accuracy 0.896 / 0.824
Bayrakdar 2020	sensitivity 0.94; specificity 0.88; accuracy 0.91
Thanathornwong 2020	sensitivity 0.84; specificity 0.88; F-measure 0.81
Kim 2019	F1 0.75 vs clinicians 0.69
Krois 2019	sensitivity 0.81; specificity 0.81; accuracy 0.81
Lee 2018	accuracy 81.0% / 76.7% (periodontally compromised teeth)

Note: AUC = area under the receiver operating characteristic curve; F1 = F1-score; ICC = intraclass correlation coefficient; ABL = alveolar bone level. Source: Compiled from data extracted from the included studies.

Reported performance varied widely across tasks and studies (Table 2). Among detection studies, sensitivity ranged from 0% for severe bone loss in one study that found the model unable to detect advanced lesions(39)to 100% for total alveolar bone loss in a U-Net study.(21)Thirteen studies reported sensitivity and ten reported specificity. Specificity ranged from 66.2%(29)to 98%,(19) and most detection studies that reported both measures obtained sensitivity and specificity at or above 80%.(2,16,26,27,40)Eight studies reported the area under the curve, which is a threshold-independent measure and therefore better suited to comparing models that used different operating points.

The area under the receiver operating characteristic curve was reported in eight studies and ranged from 0.70 for a classification study of intrabony defects(27)to 0.951 for a U-Net detection study.(21)Intermediate values were reported for detection on periapical radiographs, with area under the curve between 0.884 and 0.918 in a large multi-model study,(28,29) and for detection on panoramic radiographs, with a value of 0.91.(17)One multi-center study reported an area under the curve of 0.942 for the model, which exceeded the 0.856 achieved by the periodontal specialists who served as the reference.(32)

Among the detection studies, the reference standard conditioned the reported accuracy. Studies that used a clinical or consensus reference reported sensitivity and specificity above 0.85,(31,40,41)whereas studies that used a single expert reader or a manual annotation reported more moderate agreement.(2,35) One study that directly compared the model with the clinicians who defined the reference reported that the model exceeded the clinicians on the area under the curve, a result that is possible only when the reference is imperfect.(32)This pattern suggests that the ceiling of reported accuracy is set by the reference standard, and that comparisons between studies with different references should be interpreted with caution.

Staging and measurement studies reported lower and more variable accuracy than detection studies. Staging accuracy ranged from 0.58 for a severity classification in periapical radiographs(38)to 0.968 in a study that combined deep features with a random forest classifier,(42)with intermediate values near 0.77 to 0.89,(4-6)and a deep CNN reported accuracy of 0.728 overall but 0.97 for stage III lesions.(37) Measurement studies reported intraclass correlation coefficients from 0.601(33)to above 0.94,(22)and mean bone-level errors around 5% to 11% of the bone length.(6,38) Segmentation studies reported Dice coefficients from 0.91(23)and intersection over union up to 0.849,(23)with one study reporting a Dice of 0.91 and staging area under the curve near 0.90.(34)

The staging studies illustrate the spread more concretely. A Mask R-CNN pipeline reported accuracy of 0.896 in one center and 0.824 in another,(5)an ensemble combining YOLOv8, Mask R-CNN, and TransUNet reported accuracy of 89.45% with an intraclass correlation of 0.806 against three periodontists,(6)and a U-Net plus YOLOv4 pipeline reported accuracy of 0.77.(4)The variability among these staging studies, all of which target the same clinical classification, is larger than the variability among detection studies, which is consistent with the greater difficulty of the task.

The segmentation studies, which are a prerequisite for quantitative measurement, reported pixel-level agreement that was generally high. A U-Net enhanced with a super-resolution network reported a Dice coefficient of 0.91 and intersection over union of 0.849,(23)and a YOLOv8 model reported sensitivity of 90% to 100% and specificity of 96% to 98% for tooth, cemento-enamel junction, and bone level segmentation.(19)These segmentation results support the feasibility of automatic landmark localization, even though the downstream measurement accuracy remains more variable.

The segmentation studies are also informative in their own right. They show that networks can reliably separate tooth, bone, and the cemento-enamel junction, which are the anatomical components required for any quantitative analysis. The highest agreement was reported for the bone and the cemento-enamel junction, with sensitivity and specificity at or above 96%, and the lowest for the tooth itself, which is the least ambiguous structure to a human observer.(19)This pattern, in which the network is most precise on the structures that are hardest for humans to localize consistently, is what makes the technology attractive despite its current limits.

The strongest detection results came from models applied to panoramic radiographs with large training sets. A U-Net trained on 1,121 panoramic radiographs reported sensitivity of 1.0 and an area under the curve of 0.951,(21)and a commercial system evaluated against a three-clinician consensus reported accuracy of 0.98 for binary and 0.99 for multiclass detection.(41)A DeNTNet model trained on 12,179 panoramic radiographs reported an F1-score of 0.75, which exceeded the 0.69 achieved by the clinicians who annotated the data.(3)These results indicate that the ceiling for detection, at least on curated internal test sets, is close to the ceiling of the reference standard itself.

The contrast between detection and staging is clearest within studies that performed both. A pipeline that segmented bone, tooth, and the cemento-enamel junction reported a Dice coefficient of 0.91 for segmentation but a case diagnosis accuracy of 0.85,(34)and another reported

high segmentation agreement but staging accuracy of 0.77.(4)The pattern across these studies is consistent: the model localizes the relevant anatomy well, but the final severity assignment, which depends on a ratio and a threshold, is where accuracy is lost.

A commercial software evaluated against a consensus of three clinicians reached accuracy of 0.98 for binary and 0.99 for multiclass bone loss detection,(41)whereas a multimodal large language model evaluated on panoramic images was judged insufficient for clinical use on the basis of its bone level estimates.(30)The wide spread between these endpoints reflects differences in the model, the reference standard, and the difficulty of the task rather than a single underlying capability.

Detection performance also differed between panoramic and periapical radiographs. On panoramic radiographs, a U-Net model reported sensitivity of 1.0 and an area under the curve of 0.951 for total alveolar bone loss,(21)a Faster R-CNN model reported sensitivity of 0.84 and specificity of 0.88,(16)and a transfer learning model based on GoogleNet Inception reported sensitivity of 0.94 and specificity of 0.88.(26)On periapical radiographs, an ensemble of a detector, a classifier, and a segmentation network reported accuracy of 0.97 for bone level detection,(15)and a pipeline combining MobileNet-v2 with YOLOv8 reported accuracy of 0.88 with sensitivity of 0.7 to 1.0.(43)An early deep learning study of periodontally compromised teeth on periapical radiographs reported accuracy of 81.0% for premolars and 76.7% for molars,(44)and a two-stage CNN reported accuracy of 0.762 for radiographic bone loss classification.(45)

The measurement studies reported agreement against manual or clinical annotation, and here the results diverged. A segmentation network that delineated the bone, the tooth, and the cemento-enamel junction reported a Dice coefficient of 0.91 and a staging area under the curve between 0.89 and 0.90,(34)and a cascading SegFormer model reported intersection over union of 0.789, an F1-score of 0.867, and a correlation with clinician measurements above 0.85.(46)A U-Net plus YOLOv9 pipeline reported accuracy of 95.62% and intraclass correlation above 0.94 against clinical assessment.(22)By contrast, an hourglass network reported a mean bone level error of 10.69% and a severity staging accuracy of only 0.58,(38)and a CNN evaluated against manual annotation reached an intraclass correlation of 0.601.(33)Pixel-level agreement is therefore high, whereas agreement with the clinical ground truth is more variable. The measurement studies that reported the highest agreement used a clinical or consensus reference and a controlled projection, whereas those that used a single manual annotation reported lower agreement, which again points to the reference standard as a determinant of the measured accuracy.

The field also evolved over the review period. The earliest studies used comparatively small datasets and reported moderate accuracy.(2,3,44)Studies published in 2024 and 2025 applied more recent architectures to larger datasets and reported higher accuracy, including several above 90%.(21,22,41)The reporting of measurement error and of external validation also increased in the more recent reports,(18,36) although external validation remains uncommon.

DISCUSSION

This review found that artificial intelligence detects periodontal bone loss on dental radiographs with high but heterogeneous accuracy, and that the reported performance depends more on the task and the reference standard than on any particular architecture. Detection of bone loss is the most mature task, with sensitivity and specificity frequently above 80%, whereas severity staging and quantitative measurement remain weaker and more variable, and the single study that used an external validation cohort reported lower performance than internal estimates.

These findings sit between the optimistic and the cautious positions in the existing literature. The earlier reviews concluded that artificial intelligence is efficient and broadly accurate for radiographic bone loss detection(7)and for diagnosing gingivitis and periodontal disease,(8)whereas more recent reviews emphasized that pooled accuracy is high but the certainty of the evidence is limited by risk of bias and the scarcity of external validation.(9,10) A review of digital technologies for periodontal screening likewise reported promising accuracy alongside limited clinical translation.(11)A review specific to marginal bone loss around fixed prostheses also reported high accuracy but cautioned that the evidence is limited by small samples and heterogeneous reference standards.(47)The present review reconciles these views: the detection literature supports optimism, but the staging and measurement literature, which is the part that actually drives clinical decisions, supports caution.

The earlier reviews provide useful anchors for these conclusions. Tariq and colleagues, reviewing artificial intelligence for radiographic detection of periodontal bone loss, reported that the models were efficient and accurate but noted the heterogeneity of the underlying studies.

(7)Revilla-León and colleagues, reviewing models for gingivitis and periodontal disease, reported high accuracy with a caveat about maturity and clinical readiness.(8)The present review adds to these syntheses a clearer separation of the tasks, and this separation is what explains the mixed conclusions of the earlier reviews.

The quantitative anchor points reinforce this reading. Detection studies on panoramic radiographs report sensitivity and specificity above 0.85 in most cases,(16,21,26)whereas the best measurement studies report bone level errors of 5% to 10%,(6,22) and the weakest staging studies report accuracy near chance for the most severe categories.(39)The spread is therefore not random noise; it follows a gradient from detection, through staging, to measurement, and this gradient is the most reliable message of the review.

The heterogeneity has identifiable sources. First, the task differs: detection of the presence of bone loss is easier than staging its severity or measuring its extent, and the metrics differ accordingly. Second, the reference standard varied from a single reader to a clinical assessment to a consensus panel, and the one study that compared the model against clinical staging found that the model outperformed the clinicians on the area under the curve, which suggests that the reference, not the model, may be the limiting factor in some studies.(32)Third, panoramic and periapical radiographs present different challenges, because panoramic images contain more anatomy and distortion whereas periapical images capture the bone crest in more detail. Fourth, sample sizes ranged from fewer than 100 to more than 20,000 images, and models trained and tested within a single institution tend to overstate performance. Fifth, one study evaluated a general-purpose multimodal language model rather than a dedicated vision model and found it insufficient, which shows that task-specific training matters.(30)Sixth, the unit of analysis differed, with some studies reporting patient-level accuracy and others tooth-level or site-level, which further complicates direct comparison.

The difference between detection and staging accuracy has a mechanistic explanation. Detecting the presence of bone loss requires the model to recognize a local radiographic pattern, a radiolucency at the bone crest, which convolutional networks learn readily. Staging and measurement require the model to localize the cemento-enamel junction and the bone crest precisely and to compute a ratio, which compounds errors at both landmarks. The studies that report the largest measurement errors attribute them to landmark localization rather than to lesion recognition.(33,38) Detection accuracy is therefore consistently high while staging and measurement accuracy remain modest.

The clinical consequence of this heterogeneity is that a pooled figure is misleading. A clinician who reads that artificial intelligence detects bone loss with an accuracy of 0.9 cannot know whether that figure applies to the detection of bone loss on a panoramic radiograph, which is a task the models do well, or to the staging of early bone loss on a periapical radiograph, which they do less well. A clinician who needs to distinguish stage II from stage III periodontitis, a decision that changes the treatment plan, cannot rely on a tool whose staging accuracy is 0.58, even when the same tool detects bone loss with high sensitivity. The findings therefore support using these tools for detection and as a second reader, while keeping human judgement for staging and measurement until dedicated validation studies are available.

The regulatory and commercial landscape adds a further consideration. A number of the evaluated tools are commercially available or embedded in imaging software, and one commercial tool reached accuracy above 0.98 in a consensus-based evaluation,(41)whereas another reported more modest sensitivity for bone loss.(35)The difference between these products shows that performance cannot be assumed from the vendor or from the underlying architecture, and that each tool requires its own validation evidence. Regulators and purchasers should therefore demand task-specific, externally validated accuracy rather than a general claim of artificial intelligence capability.

The choice of projection also matters for deployment. Panoramic radiographs cover the whole dentition in a single acquisition and are therefore the natural target for screening, but they suffer from magnification and overlap that degrade landmark precision. Periapical radiographs offer sharper bone crest detail but cover fewer teeth and require more acquisitions. The detection advantage of panoramic radiographs and the measurement advantage of periapical radiographs observed in this review are consistent with these known physical properties, and a pragmatic workflow might combine a panoramic screening model with a periapical confirmation model.

The findings also bear on the workflow in which these tools would be used. Panoramic radiographs are already acquired routinely for many dental patients, and a model that runs on this image without additional acquisition could flag suspected bone loss during a visit that was scheduled for another reason. This opportunistic screening use is the most plausible near-term application, because it adds no cost or radiation, and the detection accuracy reported here is sufficient

to make the flag useful. The staging and measurement tasks, by contrast, would be used to plan treatment, and their lower accuracy argues for keeping those steps under clinician control until validation improves.

At the population level, periodontitis is common, frequently undiagnosed, and treatable when identified early, which makes a low-cost detection tool a candidate for opportunistic screening. The detection accuracy reported here, with sensitivity and specificity above 0.85 on panoramic radiographs in the strongest studies, is comparable to that of other screening instruments already in clinical use.⁽¹⁾ A model embedded in routine panoramic imaging could therefore identify patients who warrant periodontal assessment, although the false-positive rate would need to be acceptable, and the diagnostic yield would still depend on a clinician confirming the finding.

The strengths of this review are the systematic search across three bibliographic databases, the explicit 2018 to 2025 window, the separation of detection from staging and measurement, and the recording of values as reported without recalculation. The limitations mirror those of the primary studies. Full text was available for only a subset of reports, so some data were extracted from abstracts. Screening and extraction were performed by a single author with artificial intelligence assistance rather than two independent reviewers, which raises the risk of selection error and warrants a dual-reviewer design in future updates. A meta-analysis was not possible given the heterogeneous tasks and metrics, and publication bias may inflate the observed performance, because studies with unfavorable results are less likely to be published. Finally, the review did not appraise the certainty of the evidence with a formal framework such as GRADE, and this appraisal would be a useful addition to future updates once the reporting of diagnostic accuracy becomes more uniform. The review's own process also has limits. The searches were restricted to three bibliographic databases and to records retrievable in those sources, and studies indexed only elsewhere may have been missed. The reliance on abstracts for a subset of studies means that some extraction fields, such as the reference standard, may be incomplete for those reports.

For clinical practice, the results support cautious adoption of detection models as a second reader, particularly on panoramic radiographs, while reserving staging and measurement decisions for the clinician. A detection model that flags sites of suspected bone loss can direct the examiner's attention and reduce missed lesions, whereas a staging model that assigns a severity class carries a risk of misclassification that directly affects the treatment plan. For research, the most informative next studies would validate models prospectively across multiple centers, use clinical or histologic reference standards, report measurement error in millimeters rather than classification accuracy alone, and register their protocols. The broader adoption of artificial intelligence in dental radiology depends on evidence of this kind rather than on further internal accuracy claims.^(48,49)

The findings also carry implications for how this literature should be reported and reviewed. Studies should state the reference standard, the unit of analysis, and the landmark definitions explicitly, and should report both detection accuracy and measurement error when both are relevant. Reviews of this literature should separate detection from staging and measurement, because pooling them obscures the very difference on which the clinical value of the tools depends. Until reporting converges on these conventions, clinicians and guideline developers will need to judge each model on the specific task and population for which it was validated, rather than on a single summary figure.

The research agenda that follows from these findings has three parts. The first is prospective, multi-center validation of detection and staging models on heterogeneous acquisition devices, so that the external validity of the current estimates can be established. The second is a shift from classification accuracy to measurement error as the primary endpoint, because the clinical decision depends on the bone level in millimeters, not on a class label. The third is reporting that separates detection, staging, and measurement and that states the reference standard, the landmarks, and the unit of analysis explicitly. Progress on these three points would convert a literature of promising but heterogeneous demonstrations into an evidence base on which clinical adoption can be justified.

CONCLUSIONS

This systematic review synthesized the diagnostic accuracy of artificial intelligence for periodontal bone loss on dental radiographs and found high but heterogeneous performance. Detection of bone loss on panoramic radiographs is the most mature application, with sensitivity and specificity frequently above 80%, whereas severity staging and quantitative measurement are less accurate and more variable. The reference standard was expert annotation in most studies, and risk of bias was moderate to high, chiefly because of unblinded expert references, retrospective convenience sampling, and scarce external validation. Prospective, externally validated studies with clinical or histologic reference standards are required before these models can be recommended for staging and measurement in routine practice. The gradient from detection through staging to measurement is the most reliable message of this evidence, and it should guide both clinical expectations and the design of future validation studies. Detection is ready for cautious clinical support, staging is approaching readiness, and measurement remains a research task until its error is reduced to the level that staging decisions require. These three gradations, more than any single accuracy figure, describe where the field stands and what remains to be done.

REFERENCES

1. Kassebaum NJ, Bernabé E, Dahiya M, Bhandari B, Murray CJL, Marcenes W. Global Burden of Severe Periodontitis in 1990–2010. *J Dent Res*. 2014;93(11):1045–1053. doi: 10.1177/0022034514552491
2. Krois J, Ekert T, Meinhold L, Golla T, Kharbot B, Wittemeier A, et al. Deep Learning for the Radiographic Detection of Periodontal Bone Loss. *Sci Rep*. 2019;9(1):8495. doi: 10.1038/s41598-019-44839-3
3. Kim J, Lee HS, Song IS, Jung KH. DeNTNet: Deep Neural Transfer Network for the detection of periodontal bone loss using panoramic dental radiographs. *Sci Rep*. 2019;9(1):17615. doi: 10.1038/s41598-019-53758-2
4. Jiang L, Chen D, Cao Z, Wu F, Zhu H, Zhu F. A two-stage deep learning architecture for radiographic staging of periodontal bone loss. *BMC Oral Health*. 2022;22(1):106. doi: 10.1186/s12903-022-02119-z
5. Li H, Zhou J, Zhou Y, Chen Q, She Y, Gao F, et al. An Interpretable Computer-Aided Diagnosis Method for Periodontitis From Panoramic Radiographs. *Front Physiol*. 2021;12:655556. doi: 10.3389/fphys.2021.655556
6. Xue T, Chen L, Sun Q. Deep learning method to automatically diagnose periodontal bone loss and periodontitis stage in dental panoramic radiograph. *Journal of Dentistry*. 2024;150:105373. doi: 10.1016/j.jdent.2024.105373
7. Tariq A, Nakhi FB, Salah F, Eltayeb G, Abdulla GJ, Najim N, et al. Efficiency and accuracy of artificial intelligence in the radiographic detection of periodontal bone loss: A systematic review. *Imaging Sci Dent*. 2023;53(3):193. doi: 10.5624/isd.20230092
8. Revilla-León M, Gómez-Polo M, Barmak AB, Inam W, Kan JYK, Kois JC, et al. Artificial intelligence models for diagnosing gingivitis and periodontal disease: A systematic review. *The Journal of Prosthetic Dentistry*. 2023;130(6):816–824. doi: 10.1016/j.prosdent.2022.01.026
9. Zhang J, Deng S, Zou T, Jin Z, Jiang S. Artificial intelligence models for periodontitis classification: A systematic review. *Journal of Dentistry*. 2025;156:105690. doi: 10.1016/j.jdent.2025.105690
10. Deepa DS. Diagnostic Accuracy of Artificial Intelligence- Based Models in Periodontitis: A Systematic Review Meta-Analysis. *SEEJPH*. 2025:896–910. doi: 10.70135/seejph.vi.4739
11. Farina R, Simonelli A, Trombelli L, Ettmayer JB, Schmid JL, Ramseier CA. Emerging Applications of Digital Technologies for Periodontal Screening, Diagnosis and Prognosis in the Dental Setting. *J Clin Periodontology*. 2025;52(S29):211–245. doi: 10.1111/jcpe.14156
12. 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
13. 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
14. Do KQ, Thai TT, Lam VQ, Nguyen TT. Development and validation of artificial intelligence models for automated periodontitis staging and grading using panoramic radiographs. *BMC Oral Health*. 2025;25(1):1623. doi: 10.1186/s12903-025-07025-8
15. Chen CC, Wu YF, Aung LM, Lin JCY, Ngo ST, Su JN, et al. Automatic recognition of teeth and periodontal bone loss measurement in digital radiographs using deep-learning artificial intelligence. *Journal of Dental Sciences*. 2023;18(3):1301–1309. doi: 10.1016/j.jds.2023.03.020
16. Thanathornwong B, Suebnukarn S. Automatic detection of periodontal compromised teeth in digital panoramic radiographs using faster regional convolutional neural networks. *Imaging Sci Dent*. 2020;50(2):169. doi: 10.5624/isd.2020.50.2.169
17. Ryu J, Lee DM, Jung YH, Kwon O, Park S, Hwang J, et al. Automated Detection of Periodontal Bone Loss Using Deep Learning and Panoramic Radiographs: A Convolutional Neural Network Approach. *Applied Sciences*. 2023;13(9):5261. doi: 10.3390/app13095261
18. Vollmer A, Vollmer M, Lang G, Straub A, Kübler A, Gubik S, et al. Automated Assessment of Radiographic Bone Loss in the Posterior Maxilla Utilizing a Multi-Object Detection Artificial Intelligence Algorithm. *Applied Sciences*. 2023;13(3):1858. doi: 10.3390/app13031858
19. Jundaeng J, Chamchong R, Nithikathkul C. Advanced AI-assisted panoramic radiograph analysis for periodontal prognostication and alveolar bone loss detection. *Front Dent Med*. 2025;5:1509361. doi: 10.3389/fdmed.2024.1509361
20. Li X, Chen K, Zhao D, He Y, Li Y, Li Z, et al. Deep Learning for Staging Periodontitis Using Panoramic Radiographs. *Oral Diseases*. 2025;31(7):2273–2281. doi: 10.1111/odi.15269
21. Kurt-Bayrakdar S, Bayrakdar İŞ, Yavuz MB, Sali N, Çelik Ö, Köse O, et al. Detection of periodontal bone loss patterns and furcation defects from panoramic radiographs using deep learning algorithm: a retrospective study. *BMC Oral Health*. 2024;24(1):155. doi: 10.1186/s12903-024-03896-5
22. Ameli N, Gibson MP, Kornerup I, Lagraverre M, Gierl M, Lai H. Automating bone loss measurement on periapical radiographs for predicting the periodontitis stage and grade. *Front Dent Med*. 2024;5:1479380. doi: 10.3389/fdmed.2024.1479380
23. Kong V, Lee EY, Kim KA, Shon HS. Integrating Super-Resolution with Deep Learning for Enhanced Periodontal Bone Loss Segmentation in Panoramic Radiographs. *Bioengineering*. 2024;11(11):1130. doi: 10.3390/bioengineering11111130
24. Al Husaini MAS, Habaebi MH, Yadav S. Automated Classification Radiograph of Periodontal Bone Loss Using Deep Learning. *Biomed Eng Comput Biol*. 2025;16:11795972251405305. doi: 10.1177/11795972251405305
25. Alotaibi G, Awawdeh M, Farook FF, Aljohani M, Aldhafiri RM, Aldhoayan M. Artificial intelligence (AI) diagnostic tools: utilizing a convolutional neural network (CNN) to assess periodontal bone level radiographically—a retrospective study. *BMC Oral Health*. 2022;22(1):399. doi: 10.1186/s12903-022-02436-3

26. KURT S, ÇELİK Ö, BAYRAKDAR İŞ, ORHAN K, BİLGİR E, ODABAS A, et al. SUCCESS OF ARTIFICIAL INTELLIGENCE SYSTEM IN DETERMINING ALVEOLAR BONE LOSS FROM DENTAL PANORAMIC RADIOGRAPHY IMAGES. *Cumhuriyet Dental Journal*. 2020;23(4):318-324. doi: 10.7126/cumudj.777057
27. Piroonsan K, Pimolbutr K, Tansiratanawong K. Classifying Three-Wall Intraoral Defects from Intraoral Radiographs Using Deep Learning-Based Convolutional Neural Network Models. *Eur J Dent*. 2025;19(03):721-728. doi: 10.1055/s-0044-1791784
28. Dujic H, Meyer O, Hoss P, Wölfle UC, Wülk A, Meusburger T, et al. Automatized Detection of Periodontal Bone Loss on Periapical Radiographs by Vision Transformer Networks. *Diagnostics*. 2023;13(23):3562. doi: 10.3390/diagnostics13233562
29. Hoss P, Meyer O, Wölfle UC, Wülk A, Meusburger T, Meier L, et al. Detection of Periodontal Bone Loss on Periapical Radiographs—A Diagnostic Study Using Different Convolutional Neural Networks. *JCM*. 2023;12(22):7189. doi: 10.3390/jcm12227189
30. Camlet A, Kusiak A, Ossowska A, Świetlik D. Advances in Periodontal Diagnostics: Application of Multi-Modal Language Models in Visual Interpretation of Panoramic Radiographs. *Diagnostics*. 2025;15(15):1851. doi: 10.3390/diagnostics15151851
31. Hameedalddeen A, Natto ZS, Alhassani AA, Hayes C. Validate artificial intelligence for the diagnosis of periodontal disease. *BMC Oral Health*. 2025;26(1):348. doi: 10.1186/s12903-025-07496-9
32. Li Y, Cui Z, Mei L, Xie Y, Marini L, Pelekos G, et al. A novel AI-powered radiographic analysis surpasses specialists in stage II–IV periodontitis detection: a multicenter diagnostic study. *npj Digit Med*. 2025;8(1):691. doi: 10.1038/s41746-025-02077-0
33. Tsoromokos N, Parinussa S, Claessen F, Moin DA, Loos BG. Estimation of Alveolar Bone Loss in Periodontitis Using Machine Learning. *International Dental Journal*. 2022;72(5):621-627. doi: 10.1016/j.identj.2022.02.009
34. Lee C, Kabir T, Nelson J, Sheng S, Meng H, Van Dyke TE, et al. Use of the deep learning approach to measure alveolar bone level. *J Clin Periodontology*. 2022;49(3):260-269. doi: 10.1111/jcpe.13574
35. AlGhaihab A, Moretti AJ, Reside J, Tuzova L, Huang YS, Tyndall DA. Automatic Detection of Radiographic Alveolar Bone Loss in Bitewing and Periapical Intraoral Radiographs Using Deep Learning Technology: A Preliminary Evaluation. *Diagnostics*. 2025;15(5):576. doi: 10.3390/diagnostics15050576
36. Widyaningrum R, Astuti ER, Soetojo A, Faadiya AN, Nurrachman AS, Kinanggiti ND, et al. Hybrid two-stage CNN for detection and staging of periodontitis on panoramic radiographs. *Journal of Oral Biology and Craniofacial Research*. 2025;15(6):1392-1399. doi: 10.1016/j.jobocr.2025.08.019
37. Chen IH, Lin CH, Lee MK, Chen TE, Lan TH, Chang CM, et al. Convolutional-neural-network-based radiographs evaluation assisting in early diagnosis of the periodontal bone loss via periapical radiograph. *Journal of Dental Sciences*. 2024;19(1):550-559. doi: 10.1016/j.jds.2023.09.032
38. Danks RP, Bano S, Orishko A, Tan HJ, Moreno Sancho F, D'Aiuto F, et al. Automating Periodontal bone loss measurement via dental landmark localisation. *Int J CARS*. 2021;16(7):1189-1199. doi: 10.1007/s11548-021-02431-z
39. Cerda Mardini D, Cerda Mardini P, Vicuña Iturriaga DP, Ortuño Borroto DR. "Determining the efficacy of a machine learning model for measuring periodontal bone loss". *BMC Oral Health*. 2024;24(1):100. doi: 10.1186/s12903-023-03819-w
40. Ibraheem WI, Jain S, Ayoub MN, Namazi MA, Alfaqih AI, Aggarwal A, et al. Assessment of the Diagnostic Accuracy of Artificial Intelligence Software in Identifying Common Periodontal and Restorative Dental Conditions (Marginal Bone Loss, Periapical Lesion, Crown, Restoration, Dental Caries) in Intraoral Periapical Radiographs. *Diagnostics*. 2025;15(11):1432. doi: 10.3390/diagnostics15111432
41. Amasya H, Jaju PP, Ezhov M, Gusarev M, Atakan C, Sanders A, et al. Development and validation of an artificial intelligence software for periodontal bone loss in panoramic imaging. *Int J Imaging Syst Tech*. 2024;34(1):e22973. doi: 10.1002/ima.22973
42. Dai F, Liu Q, Guo Y, Xie R, Wu J, Deng T, et al. Convolutional neural networks combined with classification algorithms for the diagnosis of periodontitis. *Oral Radiol*. 2024;40(3):357-366. doi: 10.1007/s11282-024-00739-5
43. Rezallah NNF, Sherif G, Abdelkarim AZ, Afifi S. Enhancing Periodontal Bone Loss Diagnosis Through Advanced AI Techniques. *Applied Sciences*. 2025;15(12):6832. doi: 10.3390/app15126832
44. Lee JH, Kim DH, Jeong SN, Choi SH. Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. *J Periodontal Implant Sci*. 2018;48(2):114. doi: 10.5051/jpis.2018.48.2.114
45. Kong Z, Ouyang H, Cao Y, Huang T, Ahn E, Zhang M, et al. Automated periodontitis bone loss diagnosis in panoramic radiographs using a bespoke two-stage detector. *Computers in Biology and Medicine*. 2023;152:106374. doi: 10.1016/j.compbiomed.2022.106374
46. Yu H, Ye X, Hong W, Shi R, Ding Y, Liu C. A cascading learning method with SegFormer for radiographic measurement of periodontal bone loss. *BMC Oral Health*. 2024;24(1):325. doi: 10.1186/s12903-024-04079-y
47. Admin, Syeda Abeerah Tanveer, Bibi Fatima, Robia Ghafoor. Diagnostic accuracy of artificial intelligence versus manual detection in marginal bone loss around fixed semicolon. a systematic review. *J Pak Med Assoc*. 2024;74(4):S37-S42. doi: 10.47391/jpma.aku-9s-06
48. Hung K, Montalvao C, Tanaka R, Kawai T, Bornstein MM. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. *Dentomaxillofacial Radiology*. 2020;49(1):20190107. doi: 10.1259/dmfr.20190107
49. Ding H, Wu J, Zhao W, Matinlinna JP, Burrow MF, Tsoi JKH. Artificial intelligence in dentistry—A review. *Front Dent Med*. 2023;4:1085251. doi: 10.3389/fdmed.2023.1085251

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Arlin Castro Diaz.
Writing–review and editing: Arlin Castro Diaz.