Clinical Oral Investigations (2026) 30:207
https://doi.org/10.1007/s00784-026-06882-z

RESEARCH

Evaluating AI diagnostic accuracy in approximal dental caries detection on bitewing radiographs

Kostis Giannakopoulos[1] · Argyro Kavadella[1] · Dimitrios Paraskevis[1,2] · Aristidis Arhakis[3] · Miltiadis A. Makrygiannakis[1,4] · Eleftherios G. Kaklamanos[1,3,5]

Received: 4 September 2025 / Accepted: 14 April 2026

Abstract

Objectives To evaluate the diagnostic accuracy of the Diagnocat™ artificial intelligence (AI) system for caries detection on bitewing radiographs compared with expert human examiners, with emphasis on differences between enamel and dentin lesions.

Materials and methods A sample of 100 digital bitewing radiographs (1540 surfaces) was retrospectively selected from the European University Cyprus dental clinic database using a systematic backward screening method. Radiographs were obtained with a standardized phosphor plate system and anonymized before analysis. Two independent experts (operative dentistry and oral radiology) established the reference standard. AI and human assessments were binarized (caries/no caries; enamel/dentin) and compared. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were calculated, and statistical significance was tested across detection categories.

Results Diagnocat™ showed high specificity (94.3%, 95% CI: 92.4%–96.0%) and NPV (96.1%, 95% CI: 94.7%–97.3%), with an overall accuracy of 91.6% (95% CI: 89.7%–93.4%). Sensitivity was moderate (73.1%, 95% CI: 65.9%–79.9%), and PPV was 64.7% (95% CI: 57.7%–71.5%). Agreement with the expert consensus was substantial (Cohen's $\kappa = 0.638$). For enamel lesions, sensitivity and specificity were 73.3% (95% CI: 62.8%–82.7%) and 92.9% (95% CI: 91.0%–94.7%) with moderate agreement with the consensus (Cohen's $\kappa = 0.492$) and for dentin lesions they were 72.8% (95% CI: 61.8%–83.8%) and 92.8% (95% CI 90.9%–94.6%) with moderate agreement with the consensus (Cohen's $\kappa = 0.468$). NPV remained high ($\geq 98.0\%$), while PPV was low (42.0% and 39.2%), across lesion types. Detection patterns differed significantly between AI and the reference standard ($p < 0.001$).

Conclusions Diagnocat™ demonstrated good diagnostic performance in ruling out caries. However, its overall lower sensitivity emphasizes the need for clinician oversight, especially in detecting early-stage disease.

Clinical relevance This study offers an independent validation of Diagnocat™ using bitewing radiographs. It demonstrates lesion-depth–specific insights, showing that while AI is highly reliable for excluding disease, its predictive value remains limited.

Keywords Artificial intelligence · Bitewing radiographs · Caries detection · Diagnostic accuracy · Computer-aided diagnosis

✉ Kostis Giannakopoulos
k.giannakopoulos@external.euc.ac.cy

[1] School of Dentistry, European University Cyprus, 6 Diogenous str, Nicosia 2404, Cyprus

[2] Medical School, National and Kapodistrian University of Athens, 75 Mikras Asias str, Athens 11527, Greece

[3] School of Dentistry, Aristotle University of Thessaloniki, Aristotle University of Thessaloniki Campus, Thessaloniki 54124, Greece

[4] School of Dentistry, National and Kapodistrian University of Athens, 2 Thivon str, Athens 11527, Greece

[5] Hamdan Bin Mohammed College of Dental Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences, P.O. Box: 505055, Dubai Healthcare City, Dubai, United Arab Emirates

Published online: 29 April 2026

Introduction

Artificial intelligence (AI) is increasingly transforming dentistry by enhancing diagnostic accuracy, supporting treatment planning, and improving radiographic interpretation [1, 2]. Modern AI systems, particularly deep learning models, can analyze dental data (e.g. radiographs, photographs, chart notes) rapidly and consistently, assisting clinicians in decision-making. Several studies in medical applications, where AI algorithms have begun to match or even surpass human experts in certain diagnostic tasks [3–5].

Dental caries remains a prevalent chronic disease worldwide, necessitating early detection for effective management and prevention. Traditional methods rely primarily on visual examination and radiographic interpretation, with bitewing radiographs considered the gold standard for proximal caries detection. The interpretation of radiographs has traditionally depended on the clinician's expertise, training, and experience. However, the interpretation of these radiographs is subject to inter- and intra-observer variability, potentially leading to inconsistent diagnoses [6, 7].

Standardized assessment systems, such as the International Caries Detection and Assessment System (ICDAS), have improved the consistency of caries diagnosis [8]. Ekstrand et al. demonstrated strong correlation between ICDAS radiographic criteria and histological findings, validating ICDAS as a standardized assessment tool [9].

In dental radiology, AI applications now span bitewing, periapical, panoramic and cone-beam computed tomography (CBCT) images for detecting different pathologies. Deep learning algorithms, particularly CNNs, have shown remarkable potential and several studies show promising results. Casalegno et al. developed a deep learning model that holds promise for increasing the speed and accuracy of caries detection [10]. Similarly, Schwendicke et al. conducted a systematic review showing that AI systems achieved accuracy of 0.82–0.89 in caries detection [11] and several recent studies show better detection accuracy by neural networks than by dentists [12]. Lee et al. evaluated a CNN that achieved up to 89% accuracy, 92% sensitivity and 94% specificity in detecting dental caries in premolars and molars [13] while Devito et al. reported significantly better performance of a CNN than human observers [14]. Two recent meta-analyses reported pooled sensitivity and specificity of approximately 0.94/0.91 and 0.87/0.89 respectively, for AI systems detecting proximal caries on bitewing radiographs [15, 16]. These models often exceed human dentists in sensitivity for early lesions, while specificity is comparable or slightly lower [17].

Diagnocat™ (Diagnocat Co. Ltd., San Francisco CA, USA) is a commercially available, cloud-based AI software that uses CNNs to automatically detect and label dental pathologies, including primary and secondary caries, on radiographic images. Previous studies evaluating Diagnocat™ have reported moderate to substantial agreement with expert examiners and diagnostic performance broadly comparable to that of trained clinicians. Other notable systems include Denti.AI, and deep learning frameworks such as ResNet, AssistDent and YOLO, which have shown high precision for dental findings in research settings. Overall, comparative studies suggest that AI can significantly assist dentists in caries diagnosis but should not yet be used independently [18–24]. However, these studies have largely been conducted using internally curated or non-independent datasets.

From both methodological and clinical perspectives, external validation using independent image samples is essential to assess the robustness, reproducibility, and real-world applicability of AI-based diagnostic systems prior to widespread clinical implementation. Performance may vary when algorithms are applied to datasets acquired under different clinical conditions or evaluated against independent expert reference standards.

Therefore, the present study aims to externally validate the diagnostic performance of the Diagnocat™ AI software for detecting dental caries on bitewing radiographs, using an independent image dataset, in comparison with human observers. By assessing sensitivity and specificity relative to expert consensus reference standard, this study will help determine if Diagnocat™ can reliably replicate expert radiographic caries diagnosis. The null hypothesis is that there is no difference in caries detection performance between experienced clinicians and the Diagnocat™ AI system.

Materials and methods

This diagnostic accuracy study adhered to ethical guidelines for diagnostic research and the European University Cyprus Institutional Committee on Ethics and Bioethics approved the protocol (EUC ETHICS COMMITTEE 2025-1). The study was designed and reported in accordance with the Standards for Reporting Diagnostic Accuracy Studies (STARD 2015) guidelines.

Sample size calculation, selection criteria, and technical aspects

A sample of 100 digital bitewing radiographs was selected from the imaging database of the European University Cyprus dental clinic, comprising 1540 approximal surfaces and included posterior teeth with various types of dental restorations and differing levels of caries lesions, ensuring a clinically diverse dataset.

Radiographs were selected using a systematic backward screening method: a fixed date was defined, and all bitewing radiographs taken before this date were retrieved in reverse chronological order. Each image was assessed sequentially according to the predefined inclusion and exclusion criteria. Radiographs that did not meet the criteria were excluded, and screening continued backwards until 100 eligible radiographs had been identified. This approach ensured a transparent and reproducible selection process, minimized selection bias, and ensured that all radiographs taken before the index date had an equal opportunity for inclusion, contingent solely on meeting the eligibility criteria.

Inclusion criteria were: (i) radiographs displayed posterior teeth (premolars and/or molars) with clearly visible approximal and occlusal surfaces; (ii) demonstrated acceptable diagnostic quality, including appropriate exposure, contrast, and no motion artefacts; (iii) showed non-overlapped interproximal contacts; (iv) contained teeth with no restorations or with restorations that still allowed clear assessment of the surfaces for potential lesions; (v) represented standard bitewing projections with correct geometric characteristics.

Exclusion criteria were: (i) radiographs exhibited technical errors such as cone-cut, motion blur, or cropping that impaired diagnostic interpretation; (ii) presented overlapping interproximal contacts that obscured evaluation of interproximal caries; (iii) included teeth with extensive metallic restorations, or other structures that obstructed assessment of the surfaces of interest; (iv) reflected non-standard projections or incorrect angulation that distorted tooth anatomy and prevented reliable assessment; (v) were radiographs of pediatric patients depicting deciduous or mixed dentition.

Power analysis was not performed because of the absence of prior data on the expected differences between the two diagnostic methods. Nevertheless, the inclusion of 1,540 examined surfaces was considered sufficient to allow detection of potential differences in the diagnostic accuracy of the AI method at the surface level. All bitewing radiographs were acquired using the Gendex GX-770 x-ray machine, operating at 70kVp, 7 mA (Gendex Corp, Milwaukee, Wisconsin, USA), using VistaScan phosphor imaging plates size 2 and the VistaScan Nano Easy scanner (Dürr Dental SE, Bietigheim-Bissingen, Germany). Images were exported in JPEG format and anonymized before analysis.

Reference standard

A reference standard dataset for caries presence was established by consensus between two experienced university faculty members: an Associate Professor of Operative Dentistry (KG) and an Assistant Professor of Oral and Maxillofacial Radiology (AK), with over 25 years of clinical experience. All radiographs were viewed under ambient light, using a Dell Latitude 3420 Laptop (Dell Inc, Texas, United States), on a 14-inch display screen with a 1360 × 768 resolution, allowing the display of 256 shades of grey (8-bit system). Viewers were allowed to perform image manipulation processes, such as magnification and/or contrast / brightness adjustment, as they would do in the clinical setting to enhance their diagnostic accuracy. First, each observer independently examined the bitewing radiographs to identify proximal caries lesions. Occlusal primary and secondary (adjacent to restorations) radiolucencies were not included in the analysis. The identified tooth surfaces, caries presence or absence, and ICDAS radiographic scores per surface [8], as well as the total number of lesions per radiograph, were recorded on a spreadsheet. For the statistical analysis, ICDAS scores RA1 and RA2 (lesions confined to enamel), were grouped as “enamel lesions” and scores RA3–RC6 (lesions extending into dentin), were grouped as “dentin lesions”, following conventions used in radiographic ICDAS-based validation studies [25, 26]. Before the study, both observers underwent a calibration session, reviewing sample radiographs and aligning on ICDAS diagnostic criteria to ensure consistency. The assessments were conducted blinded to each assessor’s judgment and patient clinical information, providing an unbiased diagnostic evaluation.

Following the independent evaluations, any disagreements were resolved during a joint consensus session. All radiographs with discrepant assessments were re-evaluated by the observers until agreement was reached, thereby ensuring that no indeterminate classifications remained. The resulting consensus diagnosis served as the reference standard. The agreement between the two expert clinicians (inter-observer agreement) was assessed. Inter-observer agreement statistics reported in this study refer to the independent pre-consensus evaluations and were included to document examiner calibration prior to consensus formation.

Test method

Following the development of the reference standard, all radiographs were analyzed with Diagnocat™ using the same laptop computer as during the viewing process. This software was selected because it represents a commercially available, clinically oriented system, allowing assessment of AI performance under realistic conditions relevant to routine clinical workflows. It has been extensively investigated in similar studies for estimating the diagnostic accuracy of both 2D and 3D images [21, 27–30].

The version used was fully locked and trained before this evaluation, with no additional training on our sample. The

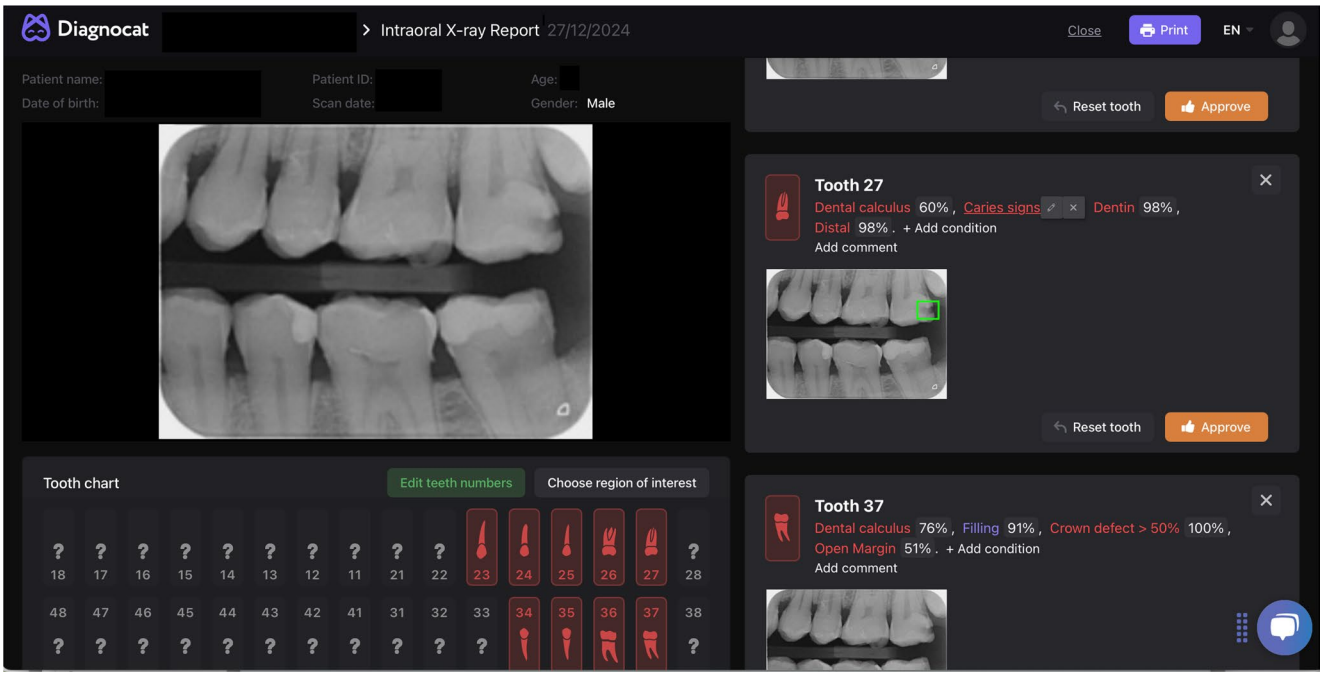

Fig. 1 Diagnocat™ output for a bitewing radiograph showing automatic detection of caries lesions. This visualization reflects how the AI software highlights potential caries-related findings. Green-colored overlay highlights the suspected area of decay, with associated probability scores displayed. All other detected pathologies by the software were disregarded for the purposes of this study

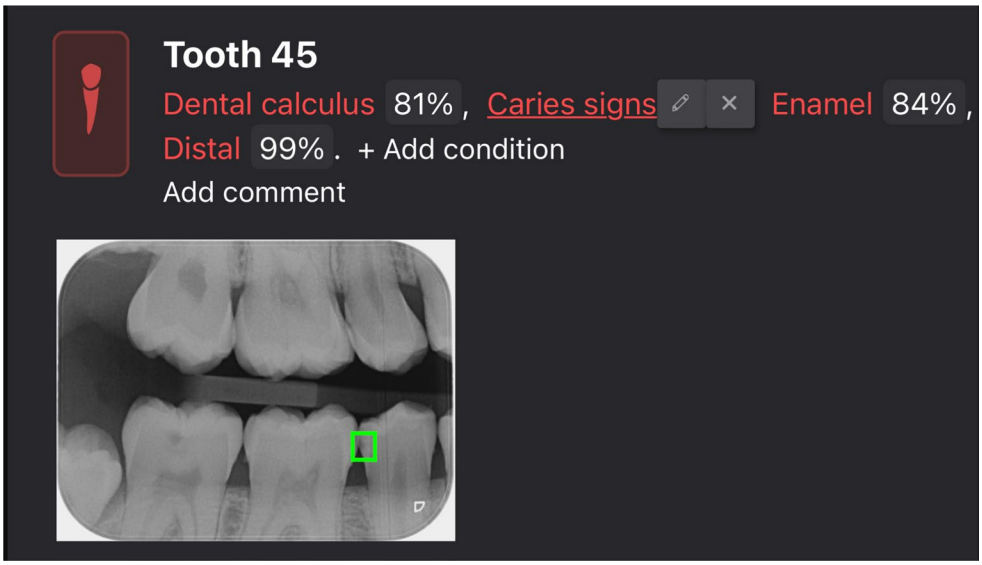

Fig. 2 Example of Diagnocat™ classification of detected caries on a bitewing radiograph. Lesions are categorized by the software into "enamel", "dentin", or "secondary caries," with visual color-coded overlays and accompanied by probability scores. This classification is clinically relevant, as differentiation between the depth of enamel and dentin lesions informs treatment planning and risk assessment

software detects various dental pathologies on each tooth and assigns a probability score to each detected finding. For the purposes of the present study, only caries lesions with software-assigned probability scores ≥ 50% were considered positive. The study evaluated the software using its default diagnostic output, and probability scores were not exported separately for threshold optimization analyses. All other detected pathologies were disregarded (Fig. 1). For each radiograph, the software-generated output included colored overlays indicating suspected caries areas and classified lesions into "enamel" and "dentin", caries (Fig. 2). The outcome of the AI assessment for each tooth surface was binarized as "caries" vs. "no caries" and as "enamel lesions" and "dentin lesions" to allow for direct comparison with the human reference standard. The lesions reported by the software were registered in a spreadsheet.

Table 1 Positive and negative diagnoses regarding the identification of caries lesions on each surface by the two expert clinicians

Any caries lesion	Assessor 2		
Assessor 1	Positive	Negative	**Subtotal**
Positive	154	38	192
Negative	24	1324	1348
Subtotal	178	1362	**1540**

Cohen's kappa = 0.80959

Table 2 Positive and negative diagnoses regarding the identification of caries lesions on each surface by the AI software compared to the reference standard dataset

Any caries lesion	AI software assessment		
Reference standard dataset	Positive	Negative	**Subtotal**
Positive	141	52	193
Negative	77	1270	1347
Subtotal	218	1322	**1540**

Statistics

The agreement between the two expert clinicians was assessed using Cohen's Kappa. All paired comparisons of proportions were conducted using a cluster-robust McNemar-type approach with patient as the clustering unit.

Diagnostic accuracy metrics were calculated separately for three binary outcome definitions: any caries, enamel lesions only, and lesions extending to dentin. We considered a match between two different assessments if tooth number, tooth surface and lesion depth were identical. From these values we calculated the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), for the AI software in detecting caries. The 95% confidence intervals were estimated using a patient-level cluster bootstrap. The agreement between the reference standard dataset and AI software assessment was evaluated using Cohen's Kappa. All paired comparisons of proportions were conducted using a cluster-robust McNemar-type approach with patient as the clustering unit. All statistical analyses were performed using IBM SPSS v.29.0. The significance level for all hypothesis testing procedures was set at $\alpha = 0.05$ ($p \leq 0.05$).

Results

The two experts independently evaluated all radiographs prior to consensus formation. Positive and negative diagnoses regarding the identification of caries lesions on each surface are presented in Table 1. Cohen's Kappa was 0.8096 for the detection of caries, indicating an almost perfect level of agreement between the assessors. Moreover, using the cluster-robust McNemar-type analysis, the mean paired difference was 0.0091 (SE: 0.006), which was not statistically significant (two-sided hypothesis; $p = 0.16$).

Table 3 Diagnostic performance of the AI software for (i) any caries (enamel + dentin combined) and (ii) depth-specific caries detection (enamel and dentin analyzed separately), compared with the reference standard dataset. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) with 95% confidence intervals are shown

Comparisons	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)
Identification of caries lesions by AI software compared to the reference standard dataset (Overall caries lesions)	73.1% (95% CI: 65.9%–79.9%)	94.3% (95% CI 92.4%–96.0%)	64.7% (95% CI 57.7%–71.5%)	96.1% (95% CI 94.7%–97.3%)
Identification of enamel caries lesions by AI software compared to the reference standard dataset (Subset analysis)	73.3% (95% CI 62.8%–82.7%)	92.9% (95% CI 91.0%–94.7%)	42.0% (95% CI 33.1%–51.2%)	98.0% (95% CI 97.1%–98.8%),
Identification of dentin caries lesions by AI software compared to the reference standard dataset (Subset analysis)	72.8% (95% CI 61.8%–83.8%)	92.8% (95% CI 90.9%–94.6%)	39.2% (95% CI 31.6%–46.6%)	98.2% (95% CI 97.2%–99.0%)

Compared to the reference standard dataset, the positive and negative diagnoses regarding identifying caries lesions on each surface by the AI software are presented in Table 2. Cohen's Kappa coefficient was 0.638 for identifying any caries, indicating substantial agreement. Accounting for clustering of surfaces within patients using the cluster-robust McNemar-type analysis, the mean paired difference was 0.016 (SE: 0.009). This difference was not statistically significant under a two-sided hypothesis ($p = 0.088$).

The sensitivity and specificity were 73.1% (95% CI 65.9%–79.9%) and 94.3% (95% CI 92.4%–96.0%), respectively, while the PPV and NPV were 64.7% (95% CI: 57.7%–71.5%) and 96.1% (95% CI: 94.7%–97.3%), respectively (Table 3). Overall agreement (accuracy) was 91.6% (95% CI: 89.7%–93.4%). Accordingly, the null hypothesis of no difference in lesion counts between the AI and the reference standard was rejected.

The positive and negative diagnoses regarding identifying enamel caries lesions on each surface by the AI software, compared to the reference standard dataset, are presented in Table 4. Cohen's Kappa coefficient was 0.492 for identifying enamel caries lesions, indicating a moderate level of agreement. Using the cluster-robust McNemar-type analysis, the mean paired difference was 0.0487 (SE: 0.0096). This difference was statistically significant (two-sided hypothesis; $p < 0.001$). However, the proportions of carious and non-carious surfaces differed significantly

Table 4 Confusion matrices showing agreement between AI system assessment and the reference standard for enamel and dentin caries lesions

Enamel caries lesions			
Reference standard dataset	AI Positive	AI Negative	**Subtotal**
Positive	74	27	101
Negative	102	1337	1439
Subtotal	176	1364	**1540**
Dentin caries lesions			
Reference standard dataset	AI Positive	AI Negative	**Subtotal**
Positive	67	25	92
Negative	104	1344	1448
Subtotal	171	1369	**1540**

($p < 0.001$). The sensitivity and specificity were 73.3% (95% CI 62.8%–82.7%) and 92.9% (95% CI 91.0%–94.7%), respectively, while the PPV and NPV were 42.0% (95% CI 33.1%–51.2%) and 98.0% (95% CI 97.1%–98.8%), respectively (Table 3). Overall agreement was 91.6% (95% CI 89.7%–93.4%).

The positive and negative diagnoses regarding identifying dentin caries lesions on each surface by the AI software, compared to the reference standard dataset, are presented in Table 4. Cohen's Kappa coefficient was 0.468 for identifying dentin caries lesions, indicating a moderate level of agreement. Using the cluster-robust McNemar-type analysis, the mean paired difference was 0.0513 (SE: 0.0098). This difference was statistically significant (two-sided hypothesis; $p < 0.001$). The sensitivity and specificity were 72.8% (95% CI 61.8%–83.8%) and 92.8% (95% CI 90.9%–94.6%), respectively, while the PPV and NPV were 39.2% (95% CI 31.6%–46.6%) and 98.2%, respectively (Table 3). Overall agreement was 91.6% (95% CI 89.7%–93.4%).

Discussion

This diagnostic accuracy study compared Diagnocat™ with expert consensus on 1,540 approximal tooth surfaces from 100 bitewing radiographs, using surface-level labels (enamel vs. dentin) to align with the AI's output. Overall, the AI software achieved substantial agreement with the reference standard for "any caries" (Cohen's κ = 0.638) and high specificity (~ 94.3%) alongside moderate sensitivity (~ 73.1%), producing a high NPV (~ 96.1%) but a lower PPV (~ 64.7%). Specificity was high (~ 94.3%), meaning the AI correctly recognized healthy surfaces most of the time, with relatively few false positives. The high NPV, implies that when the AI labels a surface as caries-free, it is very likely to be truly healthy; however, this must be interpreted alongside the low PPV, which indicates that a substantial proportion of AI-positive findings may represent false positives. Importantly, the proportions of positive/negative classifications differed significantly between AI and the reference standard (cluster-robust McNemar-type analysis $p = 0.04$), and the per-radiograph lesion counts also differed, indicating a systematic tendency of the AI to over-flag some features as carious compared with expert adjudication. By contrast, the two expert examiners showed very high inter-examiner agreement (Cohen's κ = 0.809), with no significant differences in positive/negative proportions or per-image lesion counts, supporting the robustness of the consensus reference standard [7]. Although formal intra-observer repeatability was not reassessed, prior calibration and structured consensus procedures were implemented to minimize intra-rater variability [31].

The overestimation of disease presence clinically, poses a serious risk of overtreatment, as this may result in unnecessary interventions for lesions that do not require treatment [32]. It was reported that AI usage for caries detection increased the treatment intensity, as significantly more enamel caries lesions were detected and managed non-/micro-invasively or invasively [32]. Consequently, false-positive diagnoses can compromise the cost-effectiveness of caries detection and treatment, as unnecessary treatment of initially healthy teeth can increase the likelihood of more extensive interventions over time [33]. Dentists must critically interpret the findings of the algorithm while using their inherent higher specificity, and combine them with the clinical examination before initiating any irreversible intervention [12, 33]. It is important to note that this AI software provides a continuous probability estimate for the presence of caries. For the purposes of binary classification, a 50% probability threshold was selected, corresponding to the default operational setting of the Diagnocat™ platform, in the absence of an externally validated, clinically endorsed alternative cutoff. While higher thresholds (e.g., 80–90%) would be expected to improve specificity, such thresholds would need to be clinically validated and calibrated against patient-level outcomes. The present study therefore focuses on diagnostic performance benchmarking rather than optimization of clinical decision thresholds.

When stratified by lesion depth, performance patterns were broadly similar for enamel and dentin lesions. The enamel–dentin junction was selected as the analytical boundary because it represents a biologically and radiographically meaningful threshold commonly used in ICDAS-based validation studies. In both categories, specificity and NPV remained high, whereas sensitivity was moderate and PPV very low. These findings suggest that the AI software is more reliable in excluding disease than in confidently confirming lesion presence, regardless of lesion depth.

This consideration is more critical in low-prevalence populations, such as the one in the present study, where PPV can be low and false-positive results can give rise to

overtreatment. Therefore, high specificity is required to ensure accurate identification of sound surfaces [33].

Our aggregate performance (sensitivity ~ 73%, specificity ~ 94%, κ ~ 0.638) aligns with independent Diagnocat™ validation on intraoral radiographs, where per-surface sensitivity ~ 0.51–0.76 and specificity ~ 0.88–0.97 were reported [21]. In the broader AI-for-caries literature on bitewings, meta-analyses report high pooled sensitivity and specificity (e.g., 0.94/0.91 and 0.87/0.89) with model- and dataset-dependent variability [15, 16]. Our results fall within the lower-middle range for sensitivity but upper range for specificity, which likely reflects a stringent reference standard (calibrated experts using ICDAS constructs [8, 9]) and real-world images rather than curated datasets. The high NPV we observed is consistent with meta-analytic conclusions that AI is generally reliable at excluding disease on bitewings; however, variability in PPV across studies highlights the need for cautious interpretation of positive findings [15, 16]. Another recent review and meta-analysis by Luke et al. similarly concluded that many AI models demonstrate high sensitivity and specificity in caries detection, sometimes exceeding 90% on both metrics, but with substantial variability across dasasets and lesion definitions [34]. Diagnocat™ results are mid-to-high for specificity and mid-range for sensitivity under these constraints.

Comparative AI-vs-human studies reinforce these patterns. In a randomized controlled trial on intraoral radiographs, an AI slightly exceeded dentists in both sensitivity and specificity (~ 88% and 91% vs. 84% and 88%), with accuracy 89% vs. 86% [35]. Conversely, AssistDent on bitewings increased enamel-only sensitivity for dentists but reduced specificity, illustrating a sensitivity–specificity trade-off when AI prompts are used for early lesions [24]. Our findings demonstrate that Diagnocat™ exhibited high specificity and negative predictive value, but relatively lower sensitivity and positive predictive value. This indicates that the AI software is more reliable at correctly identifying sound surfaces than detecting all caries lesions, particularly subtle demineralizations. Consequently, while the software can provide useful guidance, it may miss early lesions and does not replace clinical judgement.

Beyond adult intraoral radiography, other recent work further contextualizes our findings. A pediatric pilot study using CNNs for approximal caries on periapical radiographs showed promising discrimination in children aged 5–12 years, supporting AI's utility across age groups [36]. In 3D imaging, AI-assisted CBCT caries detection has also been explored. A decision-support system evaluated on 500 CBCT volumes enhanced reader confidence and diagnostic performance [31]. More broadly, the Diagnocat™ platform has been iteratively developed and clinically studied in CBCT workflows since its early deployments; a 2021 investigation described a multi-module deep learning system (including caries and periapical modules) that improved dentists' diagnostic performance in clinical settings [37]. A recent technical report evaluated Diagnocat™'s radiological report function on CBCT and suggested satisfactory interpretive performance, while emphasizing the need for further validation [38]. These studies, although on CBCT rather than bitewings, show an ongoing refinement trajectory for Diagnocat™ across modalities; however, its algorithmic architecture and training datasets are proprietary and not publicly disclosed, limiting external reproducibility. This lack of transparency regarding Diagnocat™'s algorithmic structure and training datasets reflects a broader challenge in commercial AI systems, where insufficient technical disclosure reduces explainability and limits scientific validation.

Strengths of our study include the use of real-world bitewings, ICDAS-informed calibration [8, 9], surface level analysis, and blinded human vs. AI readings—factors that likely improved internal validity. Limitations include a relatively modest sample size, which may constrain generalizability; however, at the surface level, the 1540 examined surfaces were considered sufficient. Furthermore, the random selection of radiographs from a dental clinic's patient pool enhances the representativeness of the sample for the general population. Moreover, variations in radiographic equipment, imaging protocols, and patient positioning across clinical environments may affect reproducibility, meaning that results obtained in our dataset cannot be assumed to generalize without caution. All bitewing radiographs were exported in JPEG format and reviewed on 14-inch, 1360 × 768 resolution, 8-bit grayscale monitors under ambient light conditions. While this setup may reduce the visibility of subtle radiolucencies compared to higher-resolution, diagnostic-grade monitors, it reflects the imaging system available at our university and simulates real clinical conditions, as these are the same computers used by students in dental operatories. Also, identical image formats and viewing conditions were applied for both human observers and the AI software, minimizing the risk of systematic bias. Future research with larger, multi-center datasets, ideally encompassing diverse populations and imaging systems, will be essential to confirm the robustness and generalizability of our findings. Another limitation is the absence of a true biologic gold standard such as histological validation of caries. We defined ground truth by expert radiographic consensus, which is a practical and common approach in imaging studies but not infallible. Thus, the AI's sensitivity and specificity are contingent on radiographic detection only, not actual lesion activity or cavitation. Along these lines, we did not incorporate clinical data (such as visual inspection findings, patient caries risk assessment, etc.) into the reference standard. In practice, dentists combine radiographic findings

with clinical context; our study isolated the radiographic task to evaluate the AI, but this means some "lesions" identified (by both AI and experts) might have been misdiagnosed in comparison to a real clinical decision-making scenario. Another limitation is that we evaluated only one AI software (Diagnocat™) and a specific version of its algorithm. AI software can improve over time with new training data and algorithm updates; our results are a snapshot of the technology's capability at the time of the study. It's possible that newer versions of Diagnocat™ or other AI platforms have achieved higher sensitivity without sacrificing specificity, or vice versa. We did not compare Diagnocat™ to other available caries detection AIs in this study, so we cannot extrapolate these findings to all systems on the market. Finally, our study focused purely on diagnostic performance; we did not assess workflow integration or real-time use of the AI in clinic. It remains to be seen how dentists interact with such software in practice – for example, whether too many false positive alerts might cause alert fatigue, or how much time an AI actually saves during interpretation.

Future research should include prospective clinical trials evaluating AI-assisted caries detection in routine practice, including its impact on diagnostic accuracy, treatment decisions, and potential overtreatment. In cases of equivocal findings on 2D radiographs, adjunctive 3D imaging (CBCT) has been suggested as a possible validation approach [20]; however, its use must remain strictly justified and in accordance with ALARA/ALADA principles [39], and should not be employed routinely for caries diagnosis [40–42]. Such investigations will help refine the responsible integration of AI tools alongside clinician judgement.

Conclusions

Diagnocat™ demonstrated high specificity and substantial agreement with expert consensus on bitewing radiographs, with moderate sensitivity overall and relatively worse performance for dentin than enamel lesions; however, these differences were not statistically significant. The system's high NPV indicates reliable exclusion of disease on AI-negative surfaces, whereas low PPV, significant cluster-robust McNemar type differences, and low sensitivity underscore the need for clinician confirmation of AI-positive findings. Based on both our results and the broader evidence, we advocate a conservative, adjunctive role for AI in caries detection: it can assist clinicians by providing a second opinion and improving the consistency of radiographic interpretation, but it cannot replace the need for expert clinical judgment. Used wisely, AI has the potential to streamline diagnostics, standardize caries assessments, and possibly improve early detection, which could benefit patient outcomes. Yet, the dentist should remain in control, verifying AI-identified lesions and integrating clinical context before making treatment decisions, thereby harnessing the strengths of AI while safeguarding against its limitations. Continued improvements in AI algorithms, larger validation studies (such as incorporating different populations and radiographic techniques), and real-world trials will further clarify how these tools can be optimally integrated into dental practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00784-026-06882-z.

Author contributions CRediT authorship contribution statement. K.G.: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing-original draft, Project administration, Supervision.A.K.: Conceptualization, Investigation, Methodology, Writing – review and editingD.P.: Formal analysis, Data curation, Writing – review and editing, Validation.A.A.: Methodology, Visualization, Writing – review and editing.M.A.M.: Methodology, Visualization, Writing – review and editing.E.G.K.: Formal analysis, Data curation, Writing – review and editing, Supervision.

Funding Open access funding provided by the Cyprus Libraries Consortium (CLC). No external funding was received for the present study. Diagnocat™ subscription was paid by the European University Cyprus.

Data availability Data can be provided upon request.

Declarations

Competing interests The authors declare no competing interests.

References

1. Mallineni SK, Sethi M, Punugoti D, Kotha SB, Alkhayal Z, Mubaraki S, Almotawah FN, Kotha SL, Sajja R, Nettam V, Thakare AA, Sakhamuri S (2024) Artificial Intelligence in Dentistry: A Descriptive Review. Bioeng (Basel) 11(12):1267. https://doi.org/10.3390/bioengineering11121267PMID: 39768085; PMCID: PMC11673909
2. Chen YW, Stanley K, Att W (2020) Artificial intelligence in dentistry: current applications and future perspectives. Quintessence Int 51(3):248–257. https://doi.org/10.3290/j.qi.a43952. Erratum

in: Quintessence Int. 2020;51(5):430. doi: 10.3290/j.qi.a44465. PMID: 32020135
3. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118. https://doi.org/10.1038/nature21056. Epub 2017 Jan 25. Erratum in: Nature. 2017;546(7660):686. doi: 10.1038/nature22985. PMID: 28117445; PMCID: PMC8382232
4. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young KC, De Fauw J, Shetty S (2020) International evaluation of an AI system for breast cancer screening. Nature 577(7788):89–94. https://doi.org/10.1038/s41586-019-1799-6. Epub 2020 Jan 1. Erratum in: Nature. 2020;586(7829):E19. doi: 10.1038/s41586-020-2679-9. PMID: 31894144
5. Ehteshami Bejnordi B, Veta M, van Johannes P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, the CAMELYON16 Consortium, Hermsen M, Manson QF, Balkenhol M, Geessink O, Stathonikos N, van Dijk MC, Bult P, Beca F, Beck AH, Wang D, Khosla A, Gargeya R, Irshad H, Zhong A, Dou Q, Li Q, Chen H, Lin HJ, Heng PA, Haß C, Bruni E, Wong Q, Halici U, Öner MÜ, Cetin-Atalay R, Berseth M, Khvatkov V, Vylegzhanin A, Kraus O, Shaban M, Rajpoot N, Awan R, Sirinukunwattana K, Qaiser T, Tsang YW, Tellez D, Annuscheit J, Hufnagl P, Valkonen M, Kartasalo K, Latonen L, Ruusuvuori P, Liimatainen K, Albarqouni S, Mungal B, George A, Demirci S, Navab N, Watanabe S, Seno S, Takenaka Y, Matsuda H, Ahmady Phoulady H, Kovalev V, Kalinovsky A, Liauchuk V, Bueno G, Fernandez-Carrobles MM, Serrano I, Deniz O, Racoceanu D, Venâncio R (2017) Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 318(22):2199–2210. https://doi.org/10.1001/jama.2017.14585PMID: 29234806; PMCID: PMC5820737
6. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet (2018);392(10159):1789–1858. doi: 10.1016/S0140-6736(18)32279-7. Epub 2018 Nov 8. Erratum in: Lancet. 2019;393(10190):e44. https://doi.org/10.1016/S0140-6736(19)31047-5. PMID: 30496104; PMCID: PMC6227754
7. Pliskin JS, Shwartz M, Gröndahl HG, Boffa J (1984) Reliability of coding depth of approximal carious lesions from non-independent interpretation of serial bitewing radiographs. Community Dent Oral Epidemiol 12(6):366–370. https://doi.org/10.1111/j.1600-0528.1984.tb01473.x
8. PittsNB, Ekstrand KR, Foundation ICDAS (2013) International Caries Detection and Assessment System (ICDAS) and its International Caries Classification and Management System (ICCMS) - methods for staging of the caries process and enabling dentists to manage caries. Community Dent Oral Epidemiol 41(1):e41–52. https://doi.org/10.1111/cdoe.12025
9. Ekstrand KR, Gimenez T, Ferreira FR, Mendes FM, Braga MM (2018) The International Caries Detection and Assessment System - ICDAS: A Systematic Review. Caries Res 52(5):406–419 Epub 2018 Mar 8. PMID: 29518788
10. Casalegno F, Newton T, Daher R, Abdelaziz M, Lodi-Rizzini A, Schürmann F, Krejci I, Markram H (2019) Caries Detection with Near-Infrared Transillumination Using Deep Learning. J Dent Res 98(11):1227–1233. https://doi.org/10.1177/0022034519871884Epub 2019 Aug 26. PMID: 31449759; PMCID: PMC6761787
11. Schwendicke F, Golla T, Dreher M, Krois J (2019) Convolutional neural networks for dental image diagnostics: A scoping review. J Dent 91:103226. https://doi.org/10.1016/j.jdent.2019.103226Epub 2019 Nov 5. PMID: 31704386
12. Cantu AG, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, Elhennawy K, Schwendicke F (2020) Detecting caries lesions of different radiographic extension on bitewings using deep learning. J Dent 100:103425. https://doi.org/10.1016/j.jdent.2020.103425
13. Lee JH, Kim DH, Jeong SN, Choi SH (2018) Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent 77:106–111. https://doi.org/10.1016/j.jdent.2018.07.015Epub 2018 Jul 26. PMID: 30056118
14. Devito KL, de Souza Barbosa F, Felippe Filho WN (2008) An artificial multilayer perceptron neural network for diagnosis of proximal dental caries. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 106(6):879–884 Epub 2008 Aug 20. PMID: 18718785
15. Carvalho BKG, Nolden EL, Wenning AS, Kiss-Dala S, Agócs G, Róth I, Kerémi B, Géczi Z, Hegyi P, Kivovics M (2024) Diagnostic accuracy of artificial intelligence for approximal caries on bitewing radiographs: A systematic review and meta-analysis. J Dent 151:105388. https://doi.org/10.1016/j.jdent.2024.105388
16. Ammar N, Kühnisch J (2024) Diagnostic performance of artificial intelligence-aided caries detection on bitewing radiographs: a systematic review and meta-analysis. Jpn Dent Sci Rev 60:128–136. https://doi.org/10.1016/j.jdsr.2024.02.001Epub 2024 Feb 29. PMID: 38450159; PMCID: PMC10917640
17. Issa J, Jaber M, Rifai I, Mozdziak P, Kempisty B, Dyszkiewicz-Konwińska M (2023) Diagnostic Test Accuracy of Artificial Intelligence in Detecting Periapical Periodontitis on Two-Dimensional Radiographs: A Retrospective Study and Literature Review. Med (Kaunas) 59(4):768. https://doi.org/10.3390/medicina59040768PMID: 37109726; PMCID: PMC10142688
18. Hamdan MH, Tuzova L, Mol A, Tawil PZ, Tuzoff D, Tyndall DA (2022) The effect of a deep-learning tool on dentists' performances in detecting apical radiolucencies on periapical radiographs. Dentomaxillofac Radiol 51(7):20220122. https://doi.org/10.1259/dmfr.20220122Epub 2022 Sep 12. PMID: 35980437; PMCID: PMC9522978
19. Li S, Liu J, Zhou Z, Zhou Z, Wu X, Li Y, Wang S, Liao W, Ying S, Zhao Z (2022) Artificial intelligence for caries and periapical periodontitis detection. J Dent 122:104107. https://doi.org/10.1016/j.jdent.2022.104107Epub 2022 Mar 24. PMID: 35341892
20. Kwiatek J, Leśna M, Piskórz W, Kaczewiak J (2025) Comparison of the Diagnostic Accuracy of an AI-Based System for Dental Caries Detection and Clinical Evaluation Conducted by Dentists. J Clin Med 14(5):1566. https://doi.org/10.3390/jcm14051566
21. Szabó V, Szabó BT, Orhan K, Veres DS, Manulis D, Ezhov M, Sanders A (2024) Validation of artificial intelligence application for dental caries diagnosis on intraoral bitewing and periapical radiographs. J Dent 147:105105. https://doi.org/10.1016/j.jdent.2024.105105Epub 2024 May 29. PMID: 38821394
22. Güneç HG, Ürkmez EŞ, Danaci A, Dilmaç E, Onay HH, Cesur Aydin K (2023) Comparison of artificial intelligence *vs.* junior dentists' diagnostic performance based on caries and periapical infection detection on panoramic images. Quant Imaging Med Surg 13(11):7494–7503. https://doi.org/10.21037/qims-23-762Epub 2023 Sep 22. PMID: 37969638; PMCID: PMC10644137
23. Moidu NP, Sharma S, Chawla A, Kumar V, Logani A (2022) Deep learning for categorization of endodontic lesion based on radiographic periapical index scoring system. Clin Oral Investig 26(1):651–658. https://doi.org/10.1007/s00784-021-04043-yEpub 2021 Jul 2. PMID: 34213664
24. Devlin H, Williams T, Graham J, Ashley M (2021) The ADEPT study: a comparative study of dentists' ability to detect

enamel-only proximal caries in bitewing radiographs with and without the use of AssistDent artificial intelligence software. Br Dent J 231(8):481–485. https://doi.org/10.1038/s41415-021-3526-6Epub 2021 Oct 22. PMID: 34686815; PMCID: PMC8536492
25. Machiulskiene V, Campus G, Carvalho JC, Dige I, Ekstrand KR, Jablonski-Momeni A, Maltz M, Manton DJ, Martignon S, Martinez-Mier EA, Pitts NB, Schulte AG, Splieth CH, Tenuta LMA, Ferreira Zandona A, Nyvad B (2020) Terminology of Dental Caries and Dental Caries Management: Consensus Report of a Workshop Organized by ORCA and Cariology Research Group of IADR. Caries Res 54(1):7–14. https://doi.org/10.1159/000503309Epub 2019 Oct 7. PMID: 31590168
26. Kühnisch J, Aps JK, Splieth C et al (2024) ORCA-EFCD consensus report on clinical recommendation for caries diagnosis. Paper I: caries lesion detection and depth assessment. Clin Oral Invest 28:227. https://doi.org/10.1007/s00784-024-05597-3
27. Makrygiannakis MA, Giannakopoulos K, Kavadella A, Paraskevis D, Kaklamanos EG (2025) Diagnostic accuracy of an artificial intelligence-based software in detecting supernumerary and congenitally missing teeth in panoramic radiographs. Eur J Orthod 47(4):cjaf054. https://doi.org/10.1093/ejo/cjaf054PMID: 40616472; PMCID: PMC12228091
28. Amasya H, Alkhader M, Serindere G, Futyma-Gąbka K, Aktuna Belgin C, Gusarev M, Ezhov M, Różyło-Kalinowska I, Önder M, Sanders A, Costa ALF, Castro Lopes SLP, Orhan K (2023) Evaluation of a Decision Support System Developed with Deep Learning Approach for Detecting Dental Caries with Cone-Beam Computed Tomography Imaging. Diagnostics (Basel) 13(22):3471. https://doi.org/10.3390/diagnostics13223471PMID: 37998607; PMCID: PMC10669958
29. Mema H, Gaxhja E, Alicka Y, Gugu M, Topi S, Giannoni M, Pietropaoli D, Altamura S (2025) Application of AI-Driven Software Diagnocat in Managing Diagnostic Imaging in Dentistry: A Retrospective Study. Appl Sci 15(17):9790. https://doi.org/10.3390/app15179790
30. Allihaibi M, Koller G, Mannocci F (2025) Diagnostic accuracy of an artificial intelligence-based platform in detecting periapical radiolucencies on cone-beam computed tomography scans of molars. J Dent 160:105854. https://doi.org/10.1016/j.jdent.2025.105854Epub 2025 May 31
31. Mosavat F, Ahmadi E, Amirfarhangi S et al (2023) Evaluation of diagnostic accuracy of CBCT and intraoral radiography for proximal caries detection in the presence of different dental restoration materials. BMC Oral Health 23:419. https://doi.org/10.1186/s12903-023-02954-8
32. Mertens S, Krois J, Cantu AG, Arsiwala LT, Schwendicke F (2021) Artificial intelligence for caries detection: randomized trial. J Dent 115:103849. https://doi.org/10.1016/j.jdent.2021.103849
33. Schwendicke F, Rossi JG, Göstemeyer G et al (2020) Cost-effectiveness of Artificial Intelligence for Proximal Caries Detection. J Dent Res 100(4):369–376. https://doi.org/10.1177/0022034520972335
34. Luke AM, Rezallah NNF (2025) Accuracy of artificial intelligence in caries detection: a systematic review and meta-analysis. Head Face Med 21(1):24. https://doi.org/10.1186/s13005-025-00496-8PMID: 40181403; PMCID: PMC11969992
35. Das M, Shahnawaz K, Raghavendra K, Kavitha R, Nagareddy B, Murugesan S (2024) Evaluating the Accuracy of AI-Based Software vs Human Interpretation in the Diagnosis of Dental Caries Using Intraoral Radiographs: An RCT. J Pharm Bioallied Sci 16(Suppl 1):S812–S814. https://doi.org/10.4103/jpbs.jpbs_1029_23Epub 2024 Feb 29. PMID: 38595404; PMCID: PMC11001121
36. Yavsan ZS, Orhan H, Efe E, Yavsan E (2025) Diagnosis of approximal caries in children with convolutional neural networks based detection algorithms on radiographs: A pilot study. Acta Odontol Scand 84:18–25. https://doi.org/10.2340/aos.v84.42599PMID: 39761112; PMCID: PMC11734307
37. Ezhov M, Gusarev M, Golitsyna M, Yates JM, Kushnerev E, Tamimi D, Aksoy S, Shumilov E, Sanders A, Orhan K (2021) Author Correction: Clinically applicable artificial intelligence system for dental diagnosis with CBCT. Sci Rep 11(1):22217. https://doi.org/10.1038/s41598-021-01678-5. Erratum for: Sci Rep. 2021;11(1):15006. doi: 10.1038/s41598-021-94093-9. PMID: 34754062; PMCID: PMC8578540
38. Feltraco LT, Rossetto C, Yeung AWK, Soares MQS, Oenning AC (2025) Utility of the radiological report function of an artificial intelligence system in interpreting CBCT images: a technical report. Dentomaxillofac Radiol 54(3):239–244. https://doi.org/10.1093/dmfr/twaf004
39. Jaju PP, Jaju SP (2015) Cone-beam computed tomography: Time to move from ALARA to ALADA. Imaging Sci Dent 45(4):263–265. https://doi.org/10.5624/isd.2015.45.4.263Epub 2015 Dec 17. PMID: 26730375; PMCID: PMC4697012
40. Special Committee to Revise the Joint AAE/AAOMR Position Statement on use of CBCT in Endodontics (2015) AAE and AAOMR Joint Position Statement: Use of Cone Beam Computed Tomography in Endodontics 2015 Update. Oral Surg Oral Med Oral Pathol Oral Radiol 120(4):508–512
41. Bhatt M, Coil J, Chehroudi B, Esteves A, Aleksejuniene J, MacDonald D (2021) Clinical decision-making and importance of the AAE/AAOMR position statement for CBCT examination in endodontic cases. Int Endod J 54(1):26–37. https://doi.org/10.1111/iej.13397Epub 2020 Oct 7. PMID: 32964475
42. European Commission (2012) Directorate-General for Energy, Cone beam CT for dental and maxillofacial radiology – Evidence-based guidelines. Publications Office. https://data.europa.eu/doi/ https://doi.org/10.2768/21874