dentistry journal

MDPI

Review

Artificial Intelligence in Endodontics: A Narrative Review and Critical Synthesis of Diagnostic, Predictive, and Operational Applications

Ashwin Narayanan [1,2,*], Dhanavel Chakravarthy [2], Vinaytosh Mishra [3], Bassem Mohamed [1] and Tina Susan Titus [1]

[1] Department of Restorative Sciences, College of Dentistry, Gulf Medical University, Ajman P.O. Box 4184, United Arab Emirates
[2] Department of Conservative Dentistry and Endodontics, Indira Gandhi Institute of Dental Sciences (IGIDS), Sri Balaji Vidyapeeth University, Pondicherry 607402, India
[3] Thumbay College of Management and AI in Healthcare, Gulf Medical University, Ajman P.O. Box 4184, United Arab Emirates
* Correspondence: dr.ashwin@gmu.ac.ae; Tel.: +971-52-526-4686

Abstract

Background/Objectives: Artificial intelligence (AI) research in endodontics has grown quickly, but the strength of the evidence differs between applications. This narrative review with structured evidence mapping examined diagnostic imaging, canal morphology, outcome prediction, large language models (LLMs), robotics and augmented reality, and operational uses such as case-difficulty assessment, triage and scheduling. **Methods:** Two anchor reviews were combined with a documented PubMed search to 30 September 2026, and studies were classified by design, sample, reference standard, validation, and evidence of clinical utility. **Results:** Periapical lesion detection has the largest evidence base: one meta-analysis reported pooled sensitivity of 0.925 and specificity of 0.852, although most included studies were at risk of bias and none was prospective. Fracture detection relies mainly on ex vivo data and is sensitive to image resolution. Outcome models showed modest to moderate discrimination in single-centre data. LLM studies reported 40% to 100% correct, appropriate, or concordant responses across different benchmarks and scoring rules, and outputs varied on repeated querying. Difficulty-assessment and scheduling tools have appeared since 2020 but have not been evaluated prospectively in clinical workflow. **Conclusions:** Technical performance has not yet been shown to translate into clinical benefit, and AI in endodontics should currently be used only as clinician-supervised decision support.

Keywords: artificial intelligence; machine learning; deep learning; large language model; endodontics; external validation; case difficulty; triage; clinical decision support

Check for updates

Academic Editors: Adrián Lozano and José Luis Sanz

Received: 23 September 2026
Revised: 14 October 2026
Accepted: 20 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The term "artificial intelligence" was introduced in the 1955 Dartmouth proposal, which suggested that aspects of human intelligence could be described precisely enough for a machine to simulate them [1]. Dentistry adopted it late, because machine learning needs digital, labelled data, which became routine only with digital radiography and cone-beam computed tomography (CBCT) [2,3]. Early endodontic applications were narrow, such as an artificial neural network (ANN) for locating the minor apical foramen in cadaver teeth [4]. Convolutional neural networks (CNNs) later shifted attention to image interpretation, first on periapical and panoramic radiographs and then on CBCT [5]. Since 2023, large language

models (LLMs) and robotic systems for endodontic microsurgery have opened further lines of work [3].

Several reviews already cover parts of this field. Narrative overviews have described applications and future directions [3,6], data preparation and ethics [7], and image-based diagnosis [8]. Systematic reviews have addressed single questions: deep learning for periapical radiolucencies [9], AI in endodontic diagnosis and treatment evaluation [10], AI in endodontic education [11], and case-difficulty assessment tools [12]. The present review differs in three respects. First, it grades the maturity of each domain against stated criteria (data realism, validation level, and evidence of clinical utility) rather than by headline performance. Second, it keeps technical performance separate from clinical usefulness and reports each metric by type, instead of placing sensitivity, accuracy, Dice coefficients, and agreement statistics on one scale. Third, it covers operational applications (difficulty assessment, triage, and scheduling) and the 2025–2026 LLM studies that retested models over time; neither is covered in the earlier reviews.

The aim was to map the evidence across diagnostic, predictive, generative, and operational uses of AI in endodontics, to state what each body of evidence can and cannot support, and to identify validation priorities. The authors are developing a clinician-supervised triage and case-allocation system for a postgraduate endodontic clinic. That interest led us to examine operational applications in detail; to limit its influence, the same appraisal criteria were applied to every domain.

2. Materials and Methods

2.1. Review Design

This is a narrative review with structured evidence mapping, prepared with reference to the SANRA criteria for narrative reviews [13]. It is not a systematic review: no protocol was registered, and no pooling was attempted because studies differed in imaging modality, reference standard, unit of analysis, and outcome metric.

2.2. Information Sources and Search

The evidence base was assembled in two stages. First, two anchor reviews [3,6] and the primary studies they cite were used to capture work published up to early 2024. Second, PubMed/MEDLINE was searched from inception to 30 September 2026, with particular attention to records from January 2024 onwards. A core string combined AI terms (e.g., artificial intelligence, machine learning, deep learning, neural network, large language model, ChatGPT) with endodontic terms (endodont*, root canal, periapical, apical periodontitis, pulp*). Domain strings were added for periapical lesions, fractures and resorption, canal morphology, prognosis, LLMs, robotics and augmented reality, education, and operational tasks (case difficulty, complexity, triage, referral, allocation, scheduling, and workflow). Full strings and record counts are given in Appendix A (Table A1). Reference lists of included systematic reviews were checked, and citations of the anchor reviews were tracked forward.

2.3. Eligibility and Selection

We included peer-reviewed original studies (clinical, retrospective, in vitro, ex vivo, or simulation-based) and systematic reviews that applied AI, machine learning, or LLMs to an endodontic diagnostic, prognostic, treatment, educational, or operational task and reported quantitative performance. Full-length peer-reviewed conference papers were accepted for crack detection, because much of that work appears in engineering proceedings. Editorials, conference abstracts, preprints, non-English reports, and studies in which the endodontic task was incidental were excluded. Titles and abstracts were screened against these criteria,

and full texts were checked for records that appeared to meet them or where eligibility was unclear. Because the purpose was mapping rather than pooling, studies were chosen purposively within each domain: systematic reviews where available, primary studies with the largest datasets or strongest validation, and studies that expose specific limitations. The core string returned 1045 records (764 from 2024 onwards); the operational string returned 105 records, from which seven primary studies on difficulty assessment, triage, or scheduling were identified. A PRISMA flow diagram is not included because we selectively chose the most relevant papers for each category rather than filtering a large search database.

2.4. Data Extraction, Maturity Grading, and Metrics

For each study, we extracted design and data source, sample size, reference standard, validation strategy, reported metrics, and main limitations (Table 1). Individual risk-of-bias assessment was not performed; where systematic reviews applied QUADAS-2, their judgements are reported. Evidence maturity was graded on five levels (Figure 1): (1) simulated, in vitro or ex vivo data; (2) retrospective clinical data with internal validation (cross-validation or a hold-out split from the same source); (3) external or multicentre validation on data from an independent site or device; (4) clinician-assistance (reader) studies testing whether AI changes clinicians' performance; and (5) prospective clinical evaluation or studies of patient, workflow, or educational outcomes. Levels 1–3 describe technical performance and levels 4–5 clinical utility. Uncontrolled clinical reports (single-patient case reports without a comparator) were not assigned a maturity level; they were recorded separately as clinical feasibility evidence, because they show that a technique can be performed in a patient but do not test whether it improves clinical performance, workflow, or outcomes. A domain was considered more mature only when several studies reached higher levels, not on the strength of a single high value.

Table 1. Representative studies by domain: design, sample, validation, reported results, and principal limitation.

Study	Design and Data	Sample	Validation	Reported Results (Metric Type)	Main Limitation
Periapical lesion detection					
Sadr 2023 [9]	SR and meta-analysis (QUADAS-2, GRADE)	18 studies (6 pooled)	—	Pooled sensitivity 0.925, specificity 0.852 (detection)	12/18 studies at risk of bias; no prospective studies
Ekert 2019 [14]	Retrospective; panoramic radiographs	2001 tooth segments	Internal	AUC 0.85; sensitivity 0.65; specificity 0.87	Reference = majority of 6 examiners; PPV 0.49
Setzer 2020 [15]	Retrospective; limited-FOV CBCT	20 volumes, 61 roots	Internal (5-fold CV)	Detection accuracy 0.93; specificity 0.88; lesion Dice 0.52	Very small dataset
Kirnbauer 2022 [16]	Retrospective; CBCT	144 scans	Internal (4-fold CV)	Sensitivity 0.97; specificity 0.88 (detection)	Single centre
Chau 2025 [17]	Retrospective; CBCT (SAM-based)	185 scans, all with lesions	Internal	Diagnostic accuracy 98.9%; voxel sensitivity 72.4%, specificity 99.9%; Dice 0.70	Lesion-only dataset; voxel specificity inflated by background

Table 1. *Cont.*

Study	Design and Data	Sample	Validation	Reported Results (Metric Type)	Main Limitation
Fracture and crack detection					
Shah 2018 [18]	Simulated cracks; hr-CBCT	6 molars (36 simulated scans)	Internal (3-fold CV)	AUC 0.97	Synthetic cracks only
Vicory 2021 [19]	Ex vivo teeth; hr-CBCT, micro-CT	31 cracked, 14 control	Statistical separation test	Ex vivo set $p = 0.74$; synthetic set $p = 0.026$	Not a diagnostic metric; affected by smoothing filters
Sahu 2022 [20]	Ex vivo teeth; micro-CT, hr-CBCT	45 teeth (31 cracked)	Internal (hold-out)	AUC 0.88 micro-CT vs. 0.56 hr-CBCT; dentists' accuracy 0.55–0.70 vs. 0.38–0.41	Ex vivo; small sample
Fukuda 2020 [21]	Retrospective; panoramic radiographs	300 images, 330 fractured teeth	Internal (5-fold CV)	Recall 0.75; precision 0.93; F 0.83	Clearly visible fractures only; no specificity
Resorption and working length					
Mohammad-Rahimi 2024 [22]	Retrospective; periapical radiographs	190 radiographs, 470 ROIs	Internal (10-fold CV, hold-out)	Accuracy 85.6% (DINO); recall 77.4%, F1 82.9% (MoCo v2)	Single source; small test set
Saghiri 2012 [4]	Cadaver model; ANN	50 teeth	—	Correct location 96% (ANN) vs. 76% (endodontists)	Cadaver setting
Root-canal morphology					
Lin 2021 [23]	Ex vivo teeth; CBCT with micro-CT labels	30 teeth (5 test)	Internal (hold-out)	Dice 96.2% tooth, 86.8% pulp cavity	Very small test set
Sherwood 2021 [24]	Retrospective; CBCT	135 volumes (35 test)	Internal (hold-out)	Dice 0.660–0.768; sensitivity 0.720–0.786	Single centre
Duman 2024 [25]	Retrospective; CBCT axial sections	922 sections, 153 patients	Internal (hold-out)	Sensitivity 0.92; precision 0.83; AUC 0.84	Section-level analysis
Yang 2024 [26]	Retrospective; panoramic, CBCT reference	805 patients, 4 institutes	External (4 datasets)	AUC 0.797–0.945	Lowest on low-resolution data
Zhu 2026 [27]	Retrospective multicentre CBCT; reader study	1 internal, 3 external centres	External	AUC 94.4% internal, 85.2–90.0% external; aided dentists more accurate	Uncommon condition
Outcome prediction					
Lee 2023 [28]	Retrospective; periapical radiographs	598 premolars	Internal (5-fold CV)	Accuracy 67.0%; AUC 0.638	Radiograph-only input
Qu 2022 [29]	Retrospective; clinical and radiographic variables	234 teeth, 178 patients	Internal (hold-out, 5-fold CV)	Accuracy 0.80; AUC 0.88 (GBM), 0.83 (RF)	Small; single centre
Large language models and education					
Karaca 2026 [30]	Simulated scenarios; 6 LLMs	50 scenarios	—	Appropriateness 40–76% (two-specialist rating)	Simulated cases
Kumar 2026 [31]	Anonymised cases; 3 LLMs	40 cases	—	Agreement 92.5–100% (κ 0.85–1.00)	Single-expert reference
Sismanoglu 2026 [32]	Guideline yes/no items; 2 LLMs, weekly × 3	200 items	Repeated testing	Accuracy 92.8% vs. 84.8%; κ 0.95 vs. 0.38	Closed-format items

Table 1. *Cont.*

Study	Design and Data	Sample	Validation	Reported Results (Metric Type)	Main Limitation
Large language models and education					
Büker 2025 [33]	True/false questions; 5 chatbots, 3×/day × 4 days	60 questions	Repeated testing	Accuracy 74.0–82.5%	Closed-format; variation over days
Doumani 2025 [10]	Systematic review (imaging and LLM)	10 studies	—	Accuracy 75–99%	No meta-analysis; heterogeneous
Ardila 2025 [11]	SR and meta-analysis; student-based	5 studies	—	Hedges g 1.48 favouring AI vs. students	Not AI-assisted learning; $I^2 \approx 84\%$
Case difficulty, triage, and scheduling					
Mallishery 2020 [34]	AAE forms; ANN	500 cases	Not stated in abstract	Sensitivity 94.96%	Structured form input; single centre
Qu 2023 [35]	Retrospective CBCT; regression and XGBoost	341 teeth, 261 patients	Internal (5-fold CV)	R^2 0.797 (XGBoost)	Microsurgery only
Karkehabadi 2024 [36]	Retrospective periapical radiographs; CNN	1386 radiographs, 2 sites	Internal (10-fold CV) and hold-out vs. 7 examiners	Accuracy 87.6% (VGG16)	Poor reliability of human reference
Catmabacak 2026 [37]	Standardised AAE cases; 2 LLMs, 30 repeats	30 cases (5400 classifications)	Repeated testing	Least accurate and stable for moderate difficulty	Simulated cases
Gupta 2026 [38]	Simulated cases; structured decision support	60 cases	—	Agreement 83–97% with experts	Exploratory; simulated
Titu 2026 [39]	Radiographs; vision–language models	3600 radiographs	Internal	>90% reproduction of practitioner triage	Mainly text-generation metrics
Ahlat 2026 [40]	Historical demand data; ML + optimisation	7 procedure types	Simulation	Estimated capacity 113 vs. 36 patients/day	Not implemented prospectively
Robotics and augmented reality					
Farronato 2023 [41]	In vitro; 3D-printed jaws	90 access cavities	—	Entry-point deviation 0.51–0.63 mm; angular 8.5°	In vitro
Qiao 2026 [42]; Wan 2025 [43]	Case reports; autonomous robot	1 patient each	—	Uneventful 6-month follow-up	No comparator

Values are as reported in each source and are not comparable across rows because metrics, units of analysis, and reference standards differ. AAE, American Association of Endodontists; ANN, artificial neural network; AUC, area under the ROC curve; CV, cross-validation; FOV, field of view; GBM, gradient boosting machine; hr-CBCT, high-resolution CBCT; LLM, large language model; PPV, positive predictive value; RF, random forest; ROI, region of interest; SAM, Segment Anything Model; SR, systematic review.

Metrics are reported as named in the source. Accuracy is the proportion of correct classifications; sensitivity (recall), specificity, and precision refer to detection unless stated otherwise; the Dice similarity coefficient measures voxel or pixel overlap in segmentation; the area under the receiver operating characteristic curve (AUC) summarises discrimination; appropriateness is an expert judgement that a response was clinically acceptable; and

agreement is concordance with a reference rater (percentage or κ). These quantities are not interchangeable and are not compared directly.

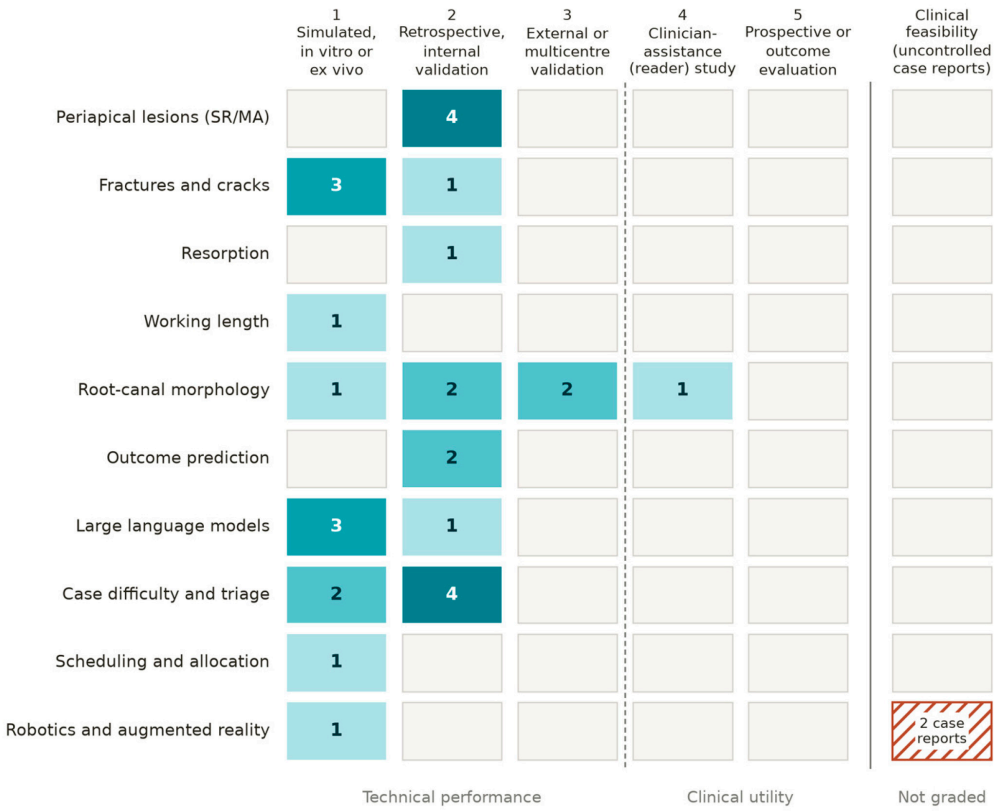


Figure 1. Evidence map of AI applications in endodontics by maturity level. Filled cells show that at least one study cited in this review reached that level; numbers give the count of such studies. Levels 1–3 reflect technical performance and levels 4–5 reflect clinical utility. The separate, ungraded column on the right records uncontrolled clinical feasibility reports (single-patient case reports without a comparator), which do not meet the criteria for level 5; the hatched cell shows the two robotic case reports [42,43]. SR/MA, systematic review with meta-analysis available. The map summarises the studies reviewed here and is not a complete census of the literature.

3. Diagnostic Imaging

3.1. Periapical Lesions

A systematic review of deep learning for periapical radiolucencies included 18 studies and pooled six; sensitivity was 0.925 (95% CI 0.862–0.960) and specificity 0.852 (95% CI 0.810–0.885), but 12 studies were at risk of bias, and none was prospective [9]. Ekert et al. analysed 2001 tooth segments on panoramic radiographs and reported an AUC of 0.85, sensitivity of 0.65, and specificity of 0.87 [14]; the reference standard was the majority decision of six examiners, so the model could agree with clinicians but not be shown to exceed them. On CBCT, Setzer et al. reported a lesion-detection accuracy of 0.93 and specificity of 0.88 from only 20 limited field-of-view volumes (61 roots), with a lesion Dice coefficient of 0.52 [15]. Kirnbauer et al. reported sensitivity of 0.97 and specificity of 0.88 on 144 scans using four-fold cross-validation [16]. CBCT-SAM, which adapts the Segment Anything foundation model, reached a diagnostic accuracy of 98.9% on 185 scans; its segmentation sensitivity was 72.4% and specificity 99.9% at voxel level, with a Dice coefficient of 0.70 [17]. Detection-level and voxel-level figures are not comparable: voxel specificity is inflated by the large number of background voxels, and every CBCT-SAM scan contained a lesion. All of these CBCT studies used internal validation only (level 2).

3.2. Fractures and Cracks

Crack detection is at an earlier stage and depends heavily on image quality. Shah et al. simulated longitudinal cracks in high-resolution CBCT (hr-CBCT) scans of six extracted molars and obtained a mean AUC of 0.97 in three-fold cross-validation [18]. Vicory et al. imaged 31 cracked and 14 control extracted teeth; cracked and control teeth were not significantly separated in this ex vivo set ($p = 0.74$), whereas separation was significant in a smaller, previously published set of synthetic hr-CBCT cracks ($p = 0.026$) [19]. The authors attributed the difference partly to smoothing filters in the scanner's endodontic mode. Sahu et al. tested a 3D U-Net on 45 ex vivo teeth: the AUC was 0.88 (accuracy 0.82) on micro-CT but 0.56 (accuracy 0.56) on hr-CBCT, while three dentists reached accuracies of 0.55–0.70 on micro-CT and 0.38–0.41 on hr-CBCT [20]. A previous review reported the 0.88 value as accuracy [3]; in the primary report, it is an AUC. On clinical panoramic radiographs, Fukuda et al. reported recall of 0.75 and precision of 0.93 for vertical root fractures, but only fractures with clearly visible lines were included, and specificity could not be estimated [21].

3.3. Resorption

Resorption has received little attention. Using 190 periapical radiographs (470 regions of interest) with CBCT-supported ground truth, Mohammad-Rahimi et al. found that the self-supervised DINO model reached a mean cross-validated accuracy of 85.6% in separating external cervical resorption, caries, and healthy teeth; MoCo v2 had the highest hold-out recall (77.4%) and F1 score (82.9%) [22]. Data came from a single source.

4. Root-Canal Morphology

External validation and reader studies have begun to appear in this domain. Lin et al. trained a U-Net on 30 extracted teeth with micro-CT-derived labels; on a five-tooth test set, Dice was 96.2% for the tooth and 86.8% for the pulp cavity [23]. For C-shaped canals in 135 CBCT volumes, Sherwood et al. reported Dice coefficients of 0.768, 0.736, and 0.660 for three U-Net variants [24]. Duman et al. detected second mesiobuccal canals on 922 axial CBCT sections from 153 patients with sensitivity of 0.92, precision of 0.83, and an AUC of 0.84 [25]. Two multicentre studies show what happens away from the development site. For C-shaped canals on panoramic radiographs of 805 patients from four institutions, AUC across four external datasets ranged from 0.797 to 0.945 and was lowest on low-resolution images [26]. For palatal radicular grooves on CBCT, diagnostic AUC was 94.4% internally and 85.2–90.0% at three external centres; in the same study, model assistance improved dentists' accuracy and shortened interpretation time [27]. These are the clearest level-3 and level-4 examples in the reviewed literature.

5. Treatment Outcome Prediction

Predicting healing is harder than detecting disease. Lee et al. trained PRESSAN-17, a CNN with self-attention, on preoperative radiographs of 598 premolars with three-year outcomes; mean five-fold accuracy was 67.0% against 63.4% for ResNet-18, and the mean AUC was 0.638 [28]. For endodontic microsurgery (234 teeth, 178 patients), Qu et al. reported accuracy of 0.80 for both gradient boosting (AUC 0.88) and random forest (AUC 0.83) models built on eight clinical and radiographic predictors [29]. Both were single-centre and internally validated; neither has external validation or reported calibration, and operator- and restoration-related factors known to affect outcome were absent or limited. These models cannot yet guide individual treatment decisions.

6. Large Language Models

LLM studies differ from imaging studies in almost every design feature. Reported proportions of correct, appropriate, or concordant responses ranged from 40% to 100% (Table 1), but these figures measure different things. The lower bound comes from Karaca et al., in which two specialists rated six LLMs' responses to 50 simulated scenarios; appropriateness ranged from 40% (ChatGPT-3.5) to 76% (Claude) and was higher for straightforward than complex cases [30]. The upper bound comes from Kumar et al., in which ChatGPT-5 agreed with one experienced endodontist on all 40 anonymised cases ($\kappa = 1.00$), with Gemini at 97.5% and Perplexity at 92.5% [31]. With a single-rater reference, this is agreement rather than accuracy.

Three factors account for much of this spread. The first is benchmark construction: yes/no items derived from position statements [32], true/false questions graded by difficulty [33], simulated vignettes [30], and anonymised cases [31] test different abilities, and closed items are easier to score than open diagnostic reasoning. The second is the reference standard and scoring rule. The third is output variability. Sismanoglu et al. posed 200 guideline items weekly for three weeks; ChatGPT-5 was more accurate than Gemini-2.5 Flash (92.8% vs. 84.8%; OR 2.37, $p = 0.004$) and far more consistent (κ 0.95 vs. 0.38) [32]. Büker et al. asked five chatbots 60 questions three times daily over four days; overall accuracy ranged from 74.0% to 82.5%, and some models changed significantly across days [33]. In 5400 repeated case-difficulty classifications, instability was greatest for moderate-difficulty cases [37].

Most studies did not report prompt wording, system instructions, sampling temperature, model version or date of access, and commercial models change without notice. Fixed and published prompts, logged versions and dates, repeated sampling, and benchmarks built from real cases with independent multi-expert reference standards should therefore be minimum requirements. LLM evidence is currently at levels 1–2, and no prospective study has shown that LLM support improves endodontic decisions or outcomes.

7. Robotics and Augmented Reality

In an in vitro study on 3D-printed jaws, Farronato et al. used markerless augmented reality to guide 90 access cavities; mean deviation from the plan was 0.51 mm at the entry point and 0.77 mm apically in anterior teeth and premolars, with 8.5° angular deviation, and 0.63 mm at the entry point in molars [41]. Autonomous robotic systems integrating CBCT and intraoral scan data have been reported for microsurgery on a maxillary first molar combined with sinus floor elevation [42] and on the mesiobuccal root of a maxillary second molar [43], each with uneventful six-month follow-up. These are single case reports that show feasibility in expert centres but provide no comparison with freehand or static-guided surgery. They are therefore mapped as clinical feasibility evidence, outside the graded levels, and not as level 5 (Figure 1).

8. Education and Clinic Operations

In a systematic review of five student-based studies, pooled estimates favoured AI over students on diagnostic tasks (Hedges g = 1.48, 95% CI 0.60–2.36; $I^2 \approx 84\%$), with an accuracy advantage of about 20 percentage points in the two analysable LLM studies [11]. Because AI was compared with students, not students with and without AI, this does not show that AI improves learning.

Operational applications form a small but growing literature. Mallishery et al. trained an ANN on 500 completed American Association of Endodontists (AAE) Case Difficulty Assessment forms and reported sensitivity of 94.96% against calibrated endodontists [34]. Qu et al. predicted microsurgical difficulty from CBCT measurements of 341 teeth, with

XGBoost performing best ($R^2 = 0.797$) [35]. Karkehabadi et al. classified difficulty from 1386 periapical radiographs from two sites; VGG16 reached 87.6% accuracy, and all models exceeded seven human examiners, whose inter-examiner reliability was poor [36]. A structured decision-support system showed 83–97% agreement with experts on 60 simulated cases [38]. For triage, fine-tuned vision–language models reproduced more than 90% of practitioners' triage assessments on 3600 radiographs, although evaluation relied mainly on text-generation metrics [39]. For scheduling, demand forecasting combined with mixed-integer linear programming increased estimated daily capacity in an academic endodontic clinic from 36 to 113 patients in simulation, with balanced workloads across students, residents and specialists [40].

These studies show technical feasibility, but they inherit the weaknesses of their reference standards: a scoping review found that difficulty tools, including the AAE form, show variable inter- and intra-rater reliability and uneven validation [12]. To the best of our knowledge, no study has yet evaluated an integrated pathway (screening, difficulty assessment, and allocation) in a working clinic, with faculty verification of each AI recommendation and outcomes such as allocation agreement, time to allocation, caseload balance among trainees, and patient safety. In postgraduate clinics, allocation must also match case difficulty to trainee competence and ensure breadth of exposure, which referral tools designed for general practice do not address. This is a research opportunity, not an established application.

9. Evidence Overview

Table 1 lists the design, sample, validation, and principal limitation of each study discussed, and Figure 1 maps the maturity levels reached in each domain.

10. Discussion

10.1. Technical Performance Versus Clinical Utility

Almost all of the evidence reviewed sits at levels 1–2. A high cross-validated sensitivity is a technical result; it does not show that clinicians using the tool diagnose more accurately, that treatment decisions change, or that patients benefit. Only one study in this review measured clinicians' performance with and without AI [27], and none measured patient outcomes. Reporting and evaluation frameworks such as TRIPOD + AI for prediction models [44] and DECIDE-AI for early clinical evaluation of decision support [45], together with dentistry-specific checklists [46], set out what the next generation of studies should report.

10.2. External Validation and Domain Shift

Performance obtained with one scanner, protocol, population, and annotation team may not hold elsewhere. Domain shift can arise from scanner type, voxel size, field of view, exposure and reconstruction or smoothing filters; from differences in case mix and disease prevalence; from changes in annotation practice; and, for LLMs, from changes to the model itself. In the reviewed studies, crack detection fell from an AUC of 0.88 on micro-CT to 0.56 on hr-CBCT [20], smoothing filters removed crack signal [19], C-shaped canal classification was weakest on low-resolution external data [26], and groove detection lost several AUC points at external centres [27]. Domain adaptation can help; Fourier-based adaptation, for example, improved CBCT tooth segmentation after a change in scanner parameters [47]. It does not replace external testing. Spectrum effects add to the problem: CBCT-SAM was developed only on scans containing lesions [17], and the panoramic fracture study selected clearly visible fractures [21]. Both choices are likely to overstate performance in unselected patients.

10.3. Reference Standards

Reference standards range from stereomicroscopy [4] and micro-CT [23] to examiner consensus [14], a single expert [31], and follow-up outcomes [28]. When the reference is clinician consensus, a model can match clinicians but cannot be shown to outperform them; when it is a single clinician, the study measures agreement. Difficulty and triage models face an added problem, because the scoring systems used as ground truth are themselves only moderately reliable [12,36].

10.4. Implications

AI output in endodontics should currently be treated as a prompt for clinician review. For periapical lesions, the evidence justifies further evaluation as a reading aid, not autonomous diagnosis. Routine reliance on AI for fracture detection, outcome prediction, or LLM advice is not supported. Any operational tool should leave the final allocation decision with faculty and log disagreements for audit. Data governance, consent for secondary use of images, explainability, automation bias, and accountability apply to every domain [7,8]. Research priorities are multicentre external validation with full reporting of devices and protocols, reader and prospective studies of clinical utility, calibrated outcome models using clinical and radiographic predictors, standardised LLM evaluation, and shared, governed datasets.

11. Limitations

The review's narrative design and purposive selection may have introduced selection bias, and the documented search used PubMed only. Duplicate independent screening, formal risk-of-bias assessment of individual studies, and protocol registration were not undertaken. Only English-language reports were included. The maturity levels are the authors' framework rather than a validated instrument, and LLM findings may date quickly. The authors' research interest in operational AI has been disclosed.

The evidence itself is limited by single-centre retrospective designs, small samples (in several studies, fewer than 50 teeth or CBCT volumes), heterogeneous reference standards and metrics, the absence of standardised public datasets, and inconsistent reporting of confidence intervals, disease prevalence, and calibration. Publication bias towards positive results is likely, and prospective and outcome studies are almost entirely absent.

12. Conclusions

Most AI evidence in endodontics remains at the stage of technical development with internal validation. Periapical lesion detection has the largest evidence base, and morphology segmentation has the first external validation and reader studies; crack detection, outcome prediction, LLM advice, and operational tools are at an earlier stage. Reported performance should be read in light of metric type, dataset, and validation level, not as evidence of clinical benefit. For now, AI in endodontics is best used as decision support under clinician supervision. External validation, prospective evaluation, and outcome-based studies, including faculty-verified evaluation of triage and case allocation in teaching clinics, are the priorities.

Author Contributions: Conceptualization, A.N.; methodology, A.N.; investigation, A.N.; resources, A.N., D.C., V.M., B.M. and T.S.T.; writing—original draft preparation, A.N.; writing—review and editing, A.N., V.M. and T.S.T.; supervision, D.C., V.M. and B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analysed in this study. The search strings and record counts are provided in Appendix A; all studies discussed are available in the cited publications.

Acknowledgments: AI-assisted tools were used to support language editing, formatting, and manuscript organisation. During the preparation of this manuscript, the authors used Claude Sonnet 5.5 (Anthropic, accessed via the claude.ai web interface) to design and generate the graphical abstract. For the graphical abstract, no text-to-image generative model was used, and the evidence content was taken from the authors' own abstract and manuscript. The authors reviewed, edited, and verified all content and take full responsibility for the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The authors are developing an AI-supported triage and case-allocation system for a postgraduate endodontic clinic, as disclosed in the Introduction.

Appendix A

Table A1 lists the PubMed search strings. All searches were run on 30 September 2026 with publication dates from inception to 30 September 2026; counts for January 2024 onwards are shown separately. AI denotes the block ("artificial intelligence"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "neural network*"[tiab] OR "convolutional"[tiab] OR "large language model*"[tiab] OR chatbot*[tiab] OR ChatGPT[tiab]); ENDO denotes (endodont*[tiab] OR "root canal"[tiab] OR periapical[tiab] OR "apical periodontitis"[tiab] OR pulp*[tiab]).

Table A1. PubMed search strings and record counts (search date 30 September 2026).

Search	PubMed String	Records (All Years)	Records (2024–)
S1 Core	AI AND ENDO	1045	764
S2 Periapical lesions	AI AND (periapical[tiab] OR "apical periodontitis"[tiab] OR "apical lesion"[tiab])	420	287
S3 Fracture, crack, resorption	AI AND ("root fracture*"[tiab] OR crack*[tiab] OR "cervical resorption"[tiab] OR "root resorption"[tiab]) AND (tooth[tiab] OR teeth[tiab] OR dental[tiab] OR endodont*[tiab])	115	79
S4 Morphology	AI AND ("canal morphology"[tiab] OR "C-shaped"[tiab] OR mesiobuccal[tiab] OR "root canal anatomy"[tiab] OR "pulp cavity"[tiab] OR "root morphology"[tiab])	85	63
S5 Prognosis	AI AND ENDO AND (prognos*[tiab] OR outcome*[tiab] OR "treatment success"[tiab] OR predict*[tiab])	420	298
S6 LLMs	("large language model*"[tiab] OR chatbot*[tiab] OR ChatGPT[tiab] OR Gemini[tiab] OR "generative artificial intelligence"[tiab]) AND ENDO	160	158
S7 Robotics, AR	(robot*[tiab] OR "augmented reality"[tiab]) AND (endodont*[tiab] OR "root canal"[tiab] OR "apical surgery"[tiab] OR microsurgery[tiab]) AND (tooth[tiab] OR teeth[tiab] OR dental[tiab] OR endodont*[tiab])	98	81

Table A1. *Cont.*

Search	PubMed String	Records (All Years)	Records (2024–)
S8 Operational	AI AND ENDO AND ("case difficulty"[tiab] OR complexity[tiab] OR triag*[tiab] OR referral*[tiab] OR allocat*[tiab] OR allot*[tiab] OR scheduling[tiab] OR workflow[tiab] OR "case assignment"[tiab])	105	82
S9 Education	AI AND ENDO AND (student*[tiab] OR education*[tiab] OR postgraduate*[tiab] OR residen*[tiab] OR trainee*[tiab])	159	145

Searches were limited by publication date (PDAT). Record counts are those returned on the search date.

References

1. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Mag.* **2006**, *27*, 12. [CrossRef]
2. Schwendicke, F.; Samek, W.; Krois, J. Artificial Intelligence in Dentistry: Chances and Challenges. *J. Dent. Res.* **2020**, *99*, 769–774. [CrossRef] [PubMed]
3. Setzer, F.C.; Li, J.; Khan, A.A. The Use of Artificial Intelligence in Endodontics. *J. Dent. Res.* **2024**, *103*, 853–862. [CrossRef] [PubMed]
4. Saghiri, M.A.; Garcia-Godoy, F.; Gutmann, J.L.; Lotfi, M.; Asgar, K. The reliability of artificial neural network in locating minor apical foramen: A cadaver study. *J. Endod.* **2012**, *38*, 1130–1134. [CrossRef] [PubMed]
5. Heo, M.S.; Kim, J.E.; Hwang, J.J.; Han, S.S.; Kim, J.S.; Yi, W.J.; Park, I.W. Artificial Intelligence in Oral and Maxillofacial Radiology: What Is Currently Possible? *Dentomaxillofac. Radiol.* **2021**, *50*, 20200375. [CrossRef] [PubMed]
6. Aminoshariae, A.; Kulild, J.; Nagendrababu, V. Artificial Intelligence in Endodontics: Current Applications and Future Directions. *J. Endod.* **2021**, *47*, 1352–1357. [CrossRef] [PubMed]
7. Mohammad-Rahimi, H.; Sohrabniya, F.; Ourang, S.A.; Dianat, O.; Aminoshariae, A.; Nagendrababu, V.; Dummer, P.M.H.; Duncan, H.F.; Nosrat, A. Artificial intelligence in endodontics: Data preparation, clinical applications, ethical considerations, limitations, and future directions. *Int. Endod. J.* **2024**, *57*, 1566–1595. [CrossRef] [PubMed]
8. Fontenele, R.C.; Jacobs, R. Unveiling the power of artificial intelligence for image-based diagnosis and treatment in endodontics: An ally or adversary? *Int. Endod. J.* **2025**, *58*, 155–170. [CrossRef] [PubMed]
9. Sadr, S.; Mohammad-Rahimi, H.; Motamedian, S.R.; Zahedrozegar, S.; Motie, P.; Vinayahalingam, S.; Dianat, O.; Nosrat, A. Deep Learning for Detection of Periapical Radiolucent Lesions: A Systematic Review and Meta-analysis of Diagnostic Test Accuracy. *J. Endod.* **2023**, *49*, 248–261.e3. [CrossRef] [PubMed]
10. Doumani, M.; Almaqboul, F.; Alduwaysan, S.S.S.; Alzahrani, M.A.; Al Ghamdi, S.A.; Alzahrani, M.N.; Alanazi, A.T.; Al Ghamdi, N.M.; Alsalem, A.H.; Alturaif, D.; et al. Effectiveness of Artificial Intelligence in Endodontic Diagnosis and Treatment Evaluation: A Systematic Review. *Cureus* **2025**, *17*, e96091. [CrossRef] [PubMed]
11. Ardila, C.M.; Pineda-Vélez, E.; Vivares-Builes, A.M. Artificial Intelligence in Endodontic Education: A Systematic Review with Frequentist and Bayesian Meta-Analysis of Student-Based Evidence. *Dent. J.* **2025**, *13*, 489. [CrossRef] [PubMed]
12. Lin, G.S.S.; Chandler, N.P.; Duncan, H.F.; Daniel, B.K.; Aziz, A.; Friedlander, L.T. Endodontic Complexity: A Scoping Review of Case Difficulty Assessment Tools and Indices. *Int. Endod. J.* **2026**, *59*, 1358–1375. [CrossRef] [PubMed]
13. Baethge, C.; Goldbeck-Wood, S.; Mertens, S. SANRA—A Scale for the Quality Assessment of Narrative Review Articles. *Res. Integr. Peer Rev.* **2019**, *4*, 5. [CrossRef] [PubMed]
14. Ekert, T.; Krois, J.; Meinhold, L.; Elhennawy, K.; Emara, R.; Golla, T.; Schwendicke, F. Deep Learning for the Radiographic Detection of Apical Lesions. *J. Endod.* **2019**, *45*, 917–922.e5. [CrossRef] [PubMed]
15. Setzer, F.C.; Shi, K.J.; Zhang, Z.; Yan, H.; Yoon, H.; Mupparapu, M.; Li, J. Artificial Intelligence for the Computer-aided Detection of Periapical Lesions in Cone-beam Computed Tomographic Images. *J. Endod.* **2020**, *46*, 987–993. Erratum in *J. Endod.* **2026**, *52*, 156. https://doi.org/10.1016/j.joen.2025.09.022. [CrossRef] [PubMed]
16. Kirnbauer, B.; Hadzic, A.; Jakse, N.; Bischof, H.; Stern, D. Automatic Detection of Periapical Osteolytic Lesions on Cone-beam Computed Tomography Using Deep Convolutional Neuronal Networks. *J. Endod.* **2022**, *48*, 1434–1440. [CrossRef] [PubMed]
17. Chau, K.K.; Zhu, M.; AlHadidi, A.; Wang, C.; Hung, K.; Wohlgemuth, P.; Lam, W.Y.H.; Liu, W.; Yuan, Y.; Chen, H. A novel AI model for detecting periapical lesion on CBCT: CBCT-SAM. *J. Dent.* **2025**, *153*, 105526. [CrossRef] [PubMed]

18. Shah, H.; Hernandez, P.; Budin, F.; Chittajallu, D.; Vimort, J.B.; Walters, R.; Mol, A.; Khan, A.; Paniagua, B. Automatic quantification framework to detect cracks in teeth. *Proc. SPIE Int. Soc. Opt. Eng.* **2018**, *10578*, 105781K. [CrossRef] [PubMed]
19. Vicory, J.; Chandradevan, R.; Hernandez-Cerdan, P.; Huang, W.A.; Fox, D.; Qdais, L.A.; McCormick, M.; Mol, A.; Walters, R.; Marron, J.S.; et al. Dental microfracture detection using wavelet features and machine learning. *Proc. SPIE Int. Soc. Opt. Eng.* **2021**, *11596*, 115961R. [CrossRef] [PubMed]
20. Sahu, P.; Vicory, J.; McCormick, M.; Khan, A.; Geha, H.; Paniagua, B. Wavelet Guided 3D Deep Model to Improve Dental Microfracture Detection. In *Applications of Medical Artificial Intelligence (AMAI 2022)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13540, pp. 150–160. [CrossRef] [PubMed]
21. Fukuda, M.; Inamoto, K.; Shibata, N.; Ariji, Y.; Yanashita, Y.; Kutsuna, S.; Nakata, K.; Katsumata, A.; Fujita, H.; Ariji, E. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. *Oral Radiol.* **2020**, *36*, 337–343. [CrossRef] [PubMed]
22. Mohammad-Rahimi, H.; Dianat, O.; Abbasi, R.; Zahedrozegar, S.; Ashkan, A.; Motamedian, S.R.; Rohban, M.H.; Nosrat, A. Artificial Intelligence for Detection of External Cervical Resorption Using Label-Efficient Self-Supervised Learning Method. *J. Endod.* **2024**, *50*, 144–153.e2. [CrossRef] [PubMed]
23. Lin, X.; Fu, Y.; Ren, G.; Yang, X.; Duan, W.; Chen, Y.; Zhang, Q. Micro-Computed Tomography-Guided Artificial Intelligence for Pulp Cavity and Tooth Segmentation on Cone-beam Computed Tomography. *J. Endod.* **2021**, *47*, 1933–1941. [CrossRef] [PubMed]
24. Sherwood, A.A.; Sherwood, A.I.; Setzer, F.C.; Sheela Devi, K.; Shamili, J.V.; John, C.; Schwendicke, F. A Deep Learning Approach to Segment and Classify C-Shaped Canal Morphologies in Mandibular Second Molars Using Cone-beam Computed Tomography. *J. Endod.* **2021**, *47*, 1907–1916. [CrossRef] [PubMed]
25. Duman, Ş.B.; Çelik Özen, D.; Bayrakdar, I.Ş.; Baydar, O.; Alhaija, E.S.A.; Helvacıoğlu Yiğit, D.; Çelik, Ö.; Jagtap, R.; Pileggi, R.; Orhan, K. Second mesiobuccal canal segmentation with YOLOv5 architecture using cone beam computed tomography images. *Odontology* **2024**, *112*, 552–561. [CrossRef] [PubMed]
26. Yang, S.; Kim, K.D.; Kise, Y.; Nozawa, M.; Mori, M.; Takata, N.; Katsumata, A.; Ariji, Y.; Park, W.; Ariji, E. External Validation of the Effect of the Combined Use of Object Detection for the Classification of the C-Shaped Canal Configuration of the Mandibular Second Molar in Panoramic Radiographs: A Multicenter Study. *J. Endod.* **2024**, *50*, 627–636. [CrossRef] [PubMed]
27. Zhu, Q.; Fu, W.; Lin, Y.; Li, J.; Tang, W.; Zhang, Y.; Zhang, R.; Lu, G.; Lin, Y.; Shen, J.; et al. A Unified Deep Learning Framework for Visual Diagnosis of Palatal Radicular Grooves in CBCT Scans: A Multicenter Validation Study. *J. Endod.* **2026**, *52*, 1136–1147. [CrossRef] [PubMed]
28. Lee, J.; Seo, H.; Choi, Y.J.; Lee, C.; Kim, S.; Lee, Y.S.; Lee, S.; Kim, E. An Endodontic Forecasting Model Based on the Analysis of Preoperative Dental Radiographs: A Pilot Study on an Endodontic Predictive Deep Neural Network. *J. Endod.* **2023**, *49*, 710–719. [CrossRef] [PubMed]
29. Qu, Y.; Lin, Z.; Yang, Z.; Lin, H.; Huang, X.; Gu, L. Machine learning models for prognosis prediction in endodontic microsurgery. *J. Dent.* **2022**, *118*, 103947. [CrossRef] [PubMed]
30. Karaca, B.; Çakmak, Y.E.; Erkal, D. Clinical Relevance of Large Language Models in Endodontics: Diagnostic Appropriateness Based on 50 Simulated Case Scenarios. *Aust. Endod. J.* **2026**, *52*, 130–138. [CrossRef] [PubMed]
31. Kumar, V.; Mann, N.S.; Sharma, K. Large Language Models for Endodontic Diagnosis: A Comparative Study Against an Expert Reference Standard. *Aust. Endod. J.* **2026**, *52*, 185–191. [CrossRef] [PubMed]
32. Sismanoglu, S.; Isik, V.; Kayahan, M.B. Performance of large language models in endodontics: Accuracy, consistency, and benchmarking with consensus guidelines. *BMC Oral Health* **2026**, *26*, 1043. [CrossRef] [PubMed]
33. Büker, M.; Sümbüllü, M.; Arslan, H. Comparative Performance of Chatbots in Endodontic Clinical Decision Support: A 4-Day Accuracy and Consistency Study. *Int. Dent. J.* **2025**, *75*, 100920. [CrossRef] [PubMed]
34. Mallishery, S.; Chhatpar, P.; Banga, K.S.; Shah, T.; Gupta, P. The precision of case difficulty and referral decisions: An innovative automated approach. *Clin. Oral Investig.* **2020**, *24*, 1909–1915. [CrossRef] [PubMed]
35. Qu, Y.; Wen, Y.; Chen, M.; Guo, K.; Huang, X.; Gu, L. Predicting case difficulty in endodontic microsurgery using machine learning algorithms. *J. Dent.* **2023**, *133*, 104522. [CrossRef] [PubMed]
36. Karkehabadi, H.; Khoshbin, E.; Ghasemi, N.; Mahavi, A.; Mohammad-Rahimi, H.; Sadr, S. Deep learning for determining the difficulty of endodontic treatment: A pilot study. *BMC Oral Health* **2024**, *24*, 574. [CrossRef] [PubMed]
37. Catmabacak, E.D.; Valandova, I.C. Evaluating large language models for endodontic case difficulty assessment: Accuracy, consistency, temporal stability, and agreement. *BMC Oral Health* **2026**, *26*, 933. [CrossRef] [PubMed]
38. Gupta, R. Beta testing and clinical evaluation of an artificial intelligence-powered clinical decision support system for endodontic case assessment. *J. Conserv. Dent. Endod.* **2026**, *29*, 399–405. [CrossRef] [PubMed]
39. Titu, M.F.S.; Pavel, M.A.; Apurba, A.Z.; Ahmed, S.; Rahman, S.; Dudley, J.; Farook, T.H.; Troiano, G. Error-Tolerant Multimodal Vision-Language Models for Endodontic Triaging: A Cross-Sectional Study. *Int. J. Dent.* **2026**, *2026*, 4148741. [CrossRef] [PubMed]

40. Ahlat, M.; Simsek, C. Machine Learning and MILP for Capacity Optimization and Educational Workload in Academic Endodontics. *J. Dent. Educ.* **2026**, *online ahead of print*. [CrossRef] [PubMed]
41. Farronato, M.; Torres, A.; Pedano, M.S.; Jacobs, R. Novel method for augmented reality guided endodontics: An in vitro study. *J. Dent.* **2023**, *132*, 104476. [CrossRef] [PubMed]
42. Qiao, W.; Qin, L.; Chen, C.; Lin, Y.; Zhang, X.; Zhou, Y.; Meng, L. A Novel Robot-Assisted Access in Endodontic Microsurgery Combined with Sinus Floor Elevation for Maxillary First Molar. *J. Endod.* **2026**, *52*, 464–471. [CrossRef] [PubMed]
43. Wan, M.; Huang, L.; Li, X.; Li, S.; Wu, Q.; Gong, C.; Li, Y.; Yang, X. Endodontic microsurgery utilizing an autonomous robotic system for the maxillary second molar: A case report. *BMC Oral Health* **2025**, *25*, 1686. [CrossRef] [PubMed]
44. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ* **2024**, *385*, e078378. [CrossRef] [PubMed]
45. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933. [CrossRef] [PubMed]
46. Khurshid, Z.; Osathanon, T.; Shire, M.A.; Schwendicke, F.; Samaranayake, L. Artificial Intelligence in Dentistry: A Concise Review of Reporting Checklists and Guidelines. *Int. Dent. J.* **2026**, *76*, 109322. [CrossRef] [PubMed]
47. Sahu, P.; Fishbaugh, J.; Vicory, J.; Khan, A.; Paniagua, B. 3D Fourier Domain Adaptation for Improving CBCT Tooth Segmentation under Scanner Parameter Shift. In *Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 2023*; IEEE: New York, NY, USA, 2023; pp. 1–5. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.