Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Artificial Intelligence in Prosthodontic, Restorative and Implant Dentistry: A Reporting-Quality Scoping Review

Deepjyoti Roy*[1], Banani Das[2], Anindita Das[3]

[1*]Department of Computer Science and Engineering, Assam down town University, Guwahati, Assam, India, deepjyoti2roy@gmail.com
[2,3]Department of Computer Science and Engineering, Assam down town University, Guwahati, Assam, India

ABSTRACT

To map the reporting completeness and risk of bias of primary artificial intelligence (AI) studies in prosthodontics, restorative dentistry and implant dentistry, and to identify recurring methodological deficiencies. A scoping review was conducted following the Arksey and O'Malley framework with Levac and Joanna Briggs Institute refinements, reported in accordance with PRISMA-ScR. Primary studies of AI models for prosthodontic, restorative or implant tasks were charted for dataset provenance, external validation, code and data availability, reference-standard definition, sample-size justification and handling of class imbalance. Each study was routed to a design-appropriate instrument: reporting completeness was scored against a condensed twelve-domain adaptation of the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) for imaging and generative studies and against TRIPOD+AI for prediction models, while risk of bias was judged with QUADAS-2 or PROBAST+AI as appropriate.

Nineteen studies (2016–2024) were charted: six restorative, eight prosthodontic, five implant. Fifteen used convolutional neural networks, three generative adversarial networks and one a knowledge-based ontology. All nineteen used private datasets and none provided a code- or data-availability statement. Eighteen of nineteen (95%) were single-centre, reported no external validation, offered no sample-size justification and declared no adherence to any reporting guideline; 17/19 (89%) did not address class imbalance; and 15/19 (79%) defined the reference standard by expert annotation without reported inter-examiner reliability. Under QUADAS-2, 5/13 studies were at high and 7/13 at unclear overall risk of bias; under PROBAST+AI, 2/3 were at high risk, driven by the analysis domain.

Keywords: artificial intelligence; deep learning; prosthodontics; restorative dentistry; dental implants; risk of bias; reporting standards.

INTRODUCTION

Artificial intelligence (AI), and in particular deep learning, has moved quickly from a research curiosity to a component of everyday digital dentistry. Prosthodontics, restorative dentistry and implant dentistry are especially exposed to this shift because their workflows are already digital and image-rich: intraoral and desktop scanning, cone-beam computed tomography, panoramic and periapical radiography, and computer-aided design and manufacturing all generate the structured data on which learning algorithms depend [1]. The last five years have consequently seen a rapid accumulation of primary studies and of secondary reviews describing what AI can do in these fields, from caries and restoration detection to automated crown design, finish-line determination, implant-system recognition and peri-implant assessment [2–11].

Whether this activity translates into safe clinical benefit depends less on headline accuracy than on methodological soundness and transparent reporting. In medical AI more broadly, repeated meta-research has shown that impressive performance figures often rest on fragile foundations. Systematic appraisals have documented an over-reliance on small, single-institution datasets, an almost routine absence of external validation, undeclared data leakage, optimistic framing, and a near-total lack of shared code or data [12–16]. Such deficiencies inflate apparent performance and are a principal reason why so few models reach the clinic. There is no obvious reason to expect dentistry to be exempt.

To counter these problems the methodological community has produced a family of reporting guidelines and risk-of-bias instruments. For the medical-imaging studies that dominate dental AI, the Checklist for

Artificial Intelligence in Medical Imaging (CLAIM) provides item-level reporting guidance, updated in 2024 [17,18]. For prediction models, the TRIPOD+AI reporting statement and the companion PROBAST+AI risk-of-bias tool extend the established TRIPOD and PROBAST instruments to machine learning [19–21], while QUADAS-2 remains the validated instrument for judging bias in diagnostic-accuracy studies [22]. These are complemented by STARD for diagnostic-accuracy reporting [23], by SPIRIT-AI and CONSORT-AI for AI trial protocols and reports [24,25], and by a dentistry-specific author checklist [26]. Together they make it possible to ask not merely whether an AI model performed well, but whether the study was reported completely enough to be believed and reproduced. Critically, as Figure 1 shows, most of these instruments post-date the bulk of the primary dental-AI literature, so the studies now shaping clinical expectations were designed and published without them.

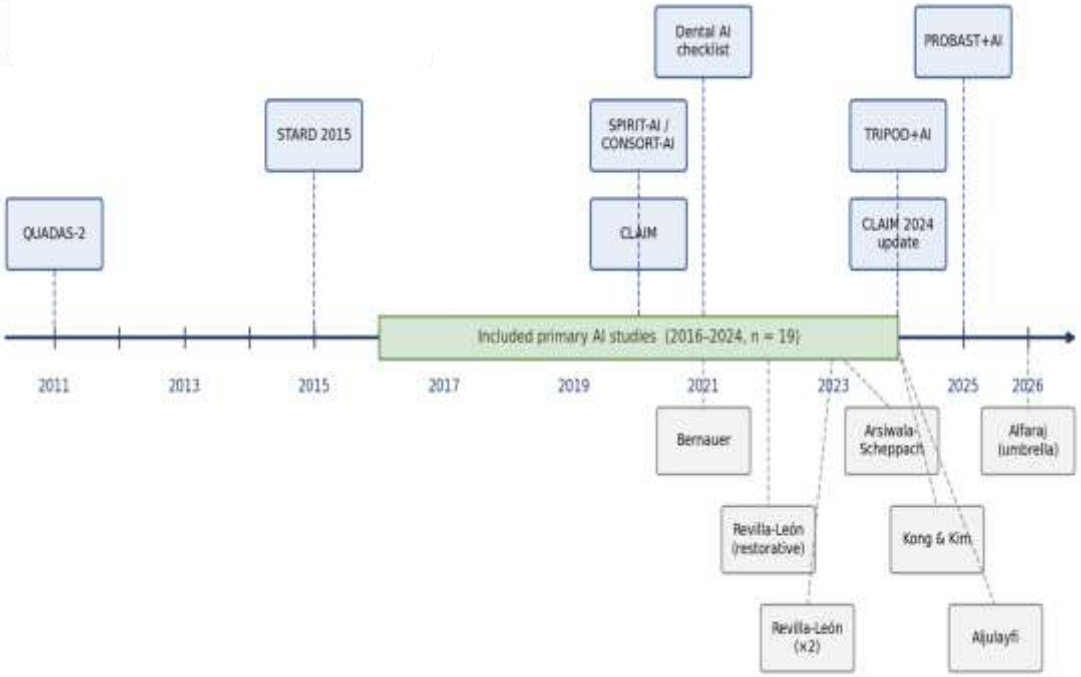

Figure 1. Chronology of reporting guidelines and risk-of-bias instruments (above the axis) relative to the secondary reviews of AI in prosthodontics, restorative and implant dentistry (below the axis) and to the publication window of the primary studies charted in this review. Most AI-specific appraisal instruments became available only after the majority of the primary evidence had been published.

Existing reviews in these fields have not yet asked that question systematically. The available syntheses are organised around applications and pooled performance: Bernauer and colleagues surveyed AI in prosthodontics and found the evidence sparse and immature [2]; Revilla-León and co-workers mapped applications in restorative dentistry, tooth-supported and removable prosthodontics, and implant dentistry [3–5]; Kong and Kim scoped AI for crown design [6]; and Arsiwala-Scheppach and colleagues scoped machine learning across all of dentistry to 2021 [7]. More recently, an umbrella review pooled eleven systematic reviews and concluded that, despite accuracy frequently exceeding 80–90%, methodological quality remains low [8], echoing an earlier comprehensive review in which fewer than one in five appraised studies were at low risk of bias [9]. Narrative reviews of CAD/CAM and removable prosthodontics reach similar conclusions [10,11]. Yet these works either evaluate applications rather than reporting, or—in the case of the umbrella review—appraise reviews rather than the primary studies beneath them, and the one dentistry-wide methodological scoping review predates the uptake of CLAIM and the arrival of TRIPOD+AI and PROBAST+AI [7,8].

A focused, primary-study, item-level appraisal of reporting completeness and risk of bias—using instruments matched to each study design and spanning prosthodontics, restorative and implant dentistry together—is therefore missing. The objectives of this scoping review were: (i) to map the characteristics of primary AI studies in these three domains; (ii) to assess their reporting completeness against CLAIM and TRIPOD+AI; (iii) to characterise their risk of bias using QUADAS-2 and PROBAST+AI; and (iv) to derive practical recommendations for authors, reviewers and journals.

Methodology

The review was designed a priori as a scoping review following the framework of Arksey and O'Malley [27], with the refinements proposed by Levac and colleagues [28] and the Joanna Briggs Institute guidance [29]. Reporting follows the PRISMA extension for Scoping Reviews (PRISMA-ScR) [30]; a completed PRISMA-ScR checklist accompanies this submission. As the review used only published data, ethical

approval was not required.

Eligibility criteria

Eligibility was defined using the Population–Concept–Context (PCC) mnemonic. Studies were eligible if they developed, trained or evaluated an AI model—machine learning, including deep learning, generative and classical or knowledge-based approaches—for a clinical task in (a) prosthodontics, including fixed and removable prosthodontics, crown or fixed-partial-denture design, shade determination, finish-line detection and occlusal design; (b) restorative or operative dentistry, including caries and restoration detection, segmentation or assessment; or (c) implant dentistry, including implant-system identification, treatment planning, peri-implant assessment and implant-outcome prediction. Only peer-reviewed primary studies reporting quantitative performance were eligible. Reviews, editorials, conference abstracts without full data, preprints, and studies confined to orthodontics, endodontics or oral and maxillofacial surgery were excluded. No date limit was applied; publications in languages other than English were excluded at full-text stage.

Information sources and search

PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore and Embase were searched from database inception to the census date, supplemented by hand-searching the reference lists of the prior reviews [2–11]. The search combined AI concepts (“artificial intelligence”, “machine learning”, “deep learning”, “neural network”, “convolutional”, “generative adversarial”, “transformer”) with domain concepts (“prosthodontic*”, “crown”, “fixed partial denture”, “removable partial denture”, “restorative”, “caries”, “dental implant”, “peri-implant*”, “shade”) using Boolean operators, with database-specific adaptation of controlled vocabulary. The full strategy for each database is deposited with the registered protocol to permit exact replication.

Selection and data charting

Records were de-duplicated in reference-management software and screened independently and in duplicate by two reviewers, first on title and abstract and then on full text. Inter-reviewer agreement was quantified using Cohen’s kappa, and disagreements were resolved by discussion with recourse to a third reviewer. A charting form was developed a priori, piloted on five studies and refined iteratively in line with the reflexive approach recommended by Levac and colleagues [28]. Charted variables comprised: authors and year; country; domain and clinical task; AI model family and specific architecture; data modality; dataset size and train/validation/test partition; number of contributing centres; data provenance; external or independent validation; availability of code and of data; the reference standard and who established it, including any reported inter-examiner reliability; whether a sample-size justification was reported; whether class imbalance was addressed; the primary performance metrics; and any declared adherence to a reporting guideline.

Critical appraisal and instrument routing

Because no single instrument fits every AI study design, each study was routed to the most appropriate contemporary tool (Figure 2). Reporting completeness of imaging, computer-vision and generative studies was assessed against CLAIM [17,18], and that of prediction-model studies against TRIPOD+AI [19]. Risk of bias was judged with QUADAS-2 across its four domains—patient selection, index test, reference standard, and flow and timing [22]—for diagnostic-accuracy studies, and with PROBAST+AI across its participants, predictors, outcome and analysis domains [20] for prediction models. Generative and design studies, for which a diagnostic-accuracy bias framework is not applicable, were appraised for reporting only and discussed narratively.

To permit a compact cross-study comparison, the CLAIM and TRIPOD+AI items were condensed a priori into twelve reporting domains judged critical to reproducibility and to the interpretation of performance: data source; ground-truth or reference standard; de-identification; data partitions; safeguards against leakage or overfitting; sample-size rationale; model architecture; definition of evaluation metrics; external or independent testing; failure or limitation analysis; code availability; and data availability. Each domain was scored as reported, partially reported or not reported. This condensation is an adaptation rather than a substitute for the full instruments, and is stated explicitly here so that the scoring is reproducible; the mapping of individual CLAIM and TRIPOD+AI items to the twelve domains is provided with the registered protocol. Two reviewers appraised every study independently. Consistent with scoping-review methodology, appraisal served to characterise the evidence base rather than to exclude studies [30].

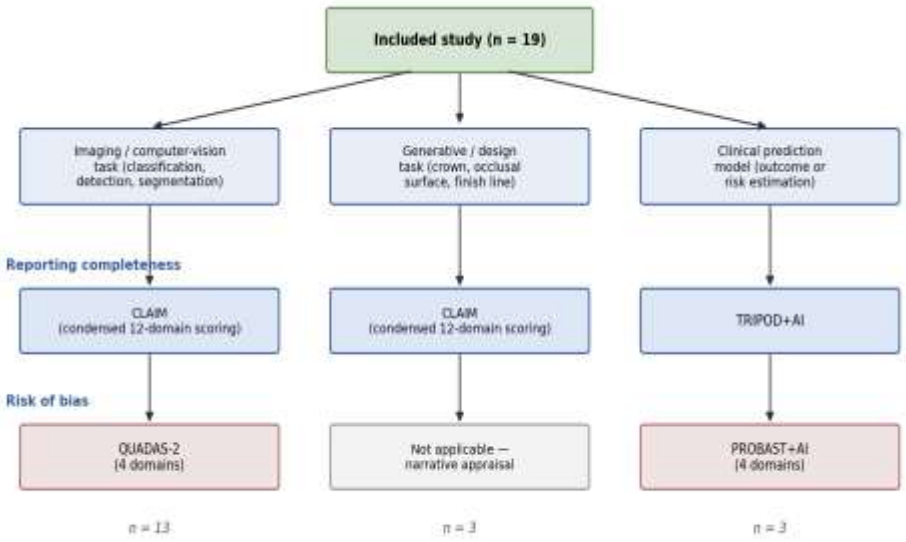

Figure 2. Routing of included studies to design-appropriate appraisal instruments. Imaging and computer-vision studies and generative design studies were scored for reporting completeness against a condensed twelve-domain adaptation of CLAIM; prediction models were scored against TRIPOD+AI. Risk of bias was judged with QUADAS-2 for diagnostic-accuracy studies and PROBAST+AI for prediction models; a diagnostic-accuracy bias framework is not applicable to generative design tasks.

Synthesis of results

Findings were summarised numerically, as counts and proportions of studies meeting each reporting domain and receiving each risk-of-bias judgement, and narratively, grouped by domain and clinical task. Given the heterogeneity of tasks, models, reference standards and outcome metrics, no meta-analysis was undertaken and no pooled estimate of accuracy is reported.

Results

Nineteen primary studies met the eligibility criteria and were charted (Figure 3). Their characteristics are presented in Table 1. Six addressed restorative dentistry, eight prosthodontics and five implant dentistry, and they were published between 2016 and 2024. Geographically the evidence base is narrow: seven studies originated in the Republic of Korea, four in Japan, three in China, two in Türkiye, and one each in Hong Kong, Germany and the Netherlands. Convolutional neural networks predominated (15/19), with generative adversarial networks used in three crown-design studies and a knowledge-based ontology with case-based reasoning in one removable-partial-denture study [44]. Radiographic images were the commonest modality (11/19: periapical, bitewing or panoramic), followed by three-dimensional scan meshes (4/19), intraoral photographs (2/19), die-scan images and tabular clinical data (1/19 each). Dataset size varied by more than three orders of magnitude, from 24 clinical cases [41] to 156,965 radiographs [46]. Reported performance was uniformly high—areas under the curve of 0.85–0.998, F1 scores up to 0.99, mean average precision up to 0.97, and generative surface deviations below 0.161 mm—a pattern that must be read against the appraisal below.

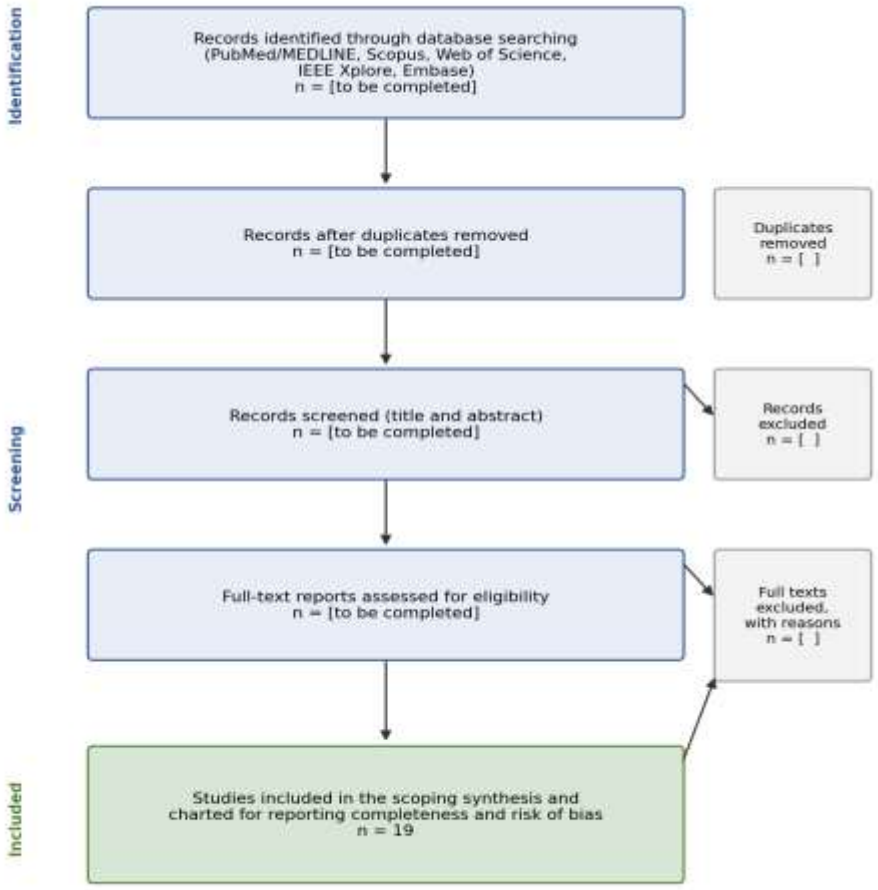

Figure 3. PRISMA-ScR flow of study identification, screening and inclusion. Screening counts are to be finalised on execution of the registered search; the included synthesis comprises the 19 studies charted here.

Table 1. Characteristics of the 19 charted artificial-intelligence studies. CNN, convolutional neural network

Study	Domain	Clinical task	Model / architecture	Modality	Dataset (n)	Centres	Key metric(s)
Lee 2018 [31]	Restorative	Caries detection	CNN (Inception-v3)	Periapical	3,000	1	AUC 0.85–0.92
Lee 2021 [32]	Restorative	Early caries detection	CNN (U-Net)	Bitewing	304 + 50	1	F1 0.64
Bayraktar 2022 [33]	Restorative	Interproximal caries	CNN (YOLO)	Bitewing	1,000	1	Sensitivity / precision
Rohrer 2022 [34]	Restorative	Restoration segmentation	CNN (U-Net)	Panoramic	1,781	1	F1 up to 0.95

Çelik 2022 [35]	Restorative	Restoration detection	CNN (10 detectors)	Panoramic	123 (684 objects)	1	mAP up to 0.97
Vinayahalingam 2021 [36]	Restorative	Automated charting	Mask R-CNN (ResNet-50)	Panoramic	2,000	1	F1 up to 0.99
Tian 2022 [37]	Prosthodontic	Crown occlusal design	Two-stage GAN	Scan / depth mesh	Not stated	1	RMS/SD < 0.161 mm
Ding 2023 [38]	Prosthodontic	Crown design	3D-DCGAN	3D casts	600 + 12	1	3D similarity; FE stress
Cho 2023 [39]	Prosthodontic	Crown design vs CAD	Commercial GAN software	Intraoral scan	30	1	Time; fit; occlusion
Choi 2024 [40]	Prosthodontic	Finish-line detection	Hybrid deep learning	Scan mesh	182	1	Distance deviation
Yamaguchi 2019 [41]	Prosthodontic	Crown debonding prediction	CNN	Die-scan images	24 cases (8,640 images)	1	Accuracy 98.5%; AUC 0.998
Takahashi 2021 [42]	Prosthodontic	Edentulous arch classification	CNN (ResNet-152)	Intraoral photographs	1,184	1	AUC 0.98 (F1 as low as 0.4)
Takahashi 2021 [43]	Prosthodontic	Prosthesis/ restoration detection	CNN (YOLOv3)	Intraoral photographs	1,904	1	AP / F per class
Chen 2016 [44]	Prosthodontic	RPD design support	Ontology+ case-based reasoning	Tabular	104	1	AUC-ROC 0.96
Sukegawa 2020 [45]	Implant	Implant-system classification	CNN (VGG16/19)	Panoramic	8,859 (11 systems)	1	Accuracy / F1 (fine-tuned VGG)

Park 2023 [46]	Implant	Implant-system classification	CNN (ResNet-50)	Panoramic + periapical	156,965	15	Accuracy 88.5%; F1 0.84
Lee & Jeong 2020 [47]	Implant	Implant-system identification	CNN (Inception-v3)	Panoramic + periapical	10,770	1	AUC 0.97
Cha 2021 [48]	Implant	Peri-implant bone-loss measurement	Region-based CNN	Periapical	708	1	AP / AR; OKS ≈ clinician
Zhang 2023 [49]	Implant	Implant-failure prediction	CNN+ clinical variables	Periapical+ panoramic	89 failures (2:1)	1	Accuracy / AUC (TRIPOD)

GAN, generative adversarial network; RPD, removable partial denture; AUC, area under the receiver-operating-characteristic curve; mAP, mean average precision; AP, average precision; AR, average recall; OKS, object keypoint similarity; FE, finite element; RMS, root mean square; SD, standard deviation.

Reporting completeness

Reporting completeness across the twelve condensed domains is presented in Table 2. Description of the data source, the model architecture and the evaluation metrics was essentially universal (19/19 for each), and data partitions were specified in 18/19. Beyond these basics, reporting thinned sharply. De-identification was stated in only 1/19, explicit safeguards against data leakage in 1/19 fully and 1/19 partially, and a rationale for sample size in 1/19. No study reported a public code repository and none shared its dataset (0/19 for both). Only the multi-centre implant study of Park and colleagues used data approximating external conditions, drawn from fifteen institutions, and even there a dedicated held-out external test set was not clearly delineated, so this domain was scored as partial [46]. A failure or limitation analysis was present in 9/19. Among the three prediction-model studies appraised against TRIPOD+AI, only Zhang and colleagues declared adherence to a reporting guideline, reported a formal sample-size calculation and adopted a deliberate class-balancing design [49]; the debonding-prediction study of Yamaguchi and colleagues [41] and the ontology-based decision-support model of Chen and colleagues [44] did none of these.

Table 2. Reporting completeness against a condensed twelve-domain adaptation of CLAIM (imaging and generative studies) and TRIPOD+AI († prediction-model studies).

Study	DS	GT	DI	DP	LK	SS	AR	EM	EV	FA	CD	DA
Lee 2018 [31]	Y	Y	—	Y	—	—	Y	Y	—	—	—	—
Lee 2021 [32]	Y	Y	—	Y	—	—	Y	Y	—	—	—	—
Bayraktar 2022 [33]	Y	Y	—	Y	—	—	Y	Y	—	—	—	—
Rohrer 2022 [34]	Y	Y	—	Y	—	—	Y	Y	—	Y	—	—
Çelik 2022 [35]	Y	Y	—	Y	—	—	Y	Y	—	Y	—	—

Vinayahalingam 2021 [36]	Y	Y	—	Y	—	—	Y	Y	—	Y	—	—
Tian 2022 [37]	Y	Y	—	—	—	—	Y	Y	—	—	—	—
Ding 2023 [38]	Y	Y	—	Y	—	—	Y	Y	—	Y	—	—
Cho 2023 [39]	Y	Y	—	Y	—	—	—	Y	—	Y	—	—
Choi 2024 [40]	Y	Y	—	Y	—	—	Y	Y	—	—	—	—
Takahashi 2021a [42]	Y	Y	—	Y	—	—	Y	Y	—	Y	—	—
Takahashi 2021b [43]	Y	Y	—	Y	—	—	Y	Y	—	—	—	—
Sukegawa 2020 [45]	Y	Y	—	Y	—	—	Y	Y	—	—	—	—
Park 2023 [46]	Y	Y	—	Y	—	—	Y	Y	P	Y	—	—
Lee & Jeong 2020 [47]	Y	Y	—	Y	—	—	Y	Y	—	—	—	—
Cha 2021 [48]	Y	Y	—	Y	—	—	Y	Y	—	Y	—	—
Yamaguchi 2019 [41] †	Y	Y	—	Y	P	—	Y	Y	—	Y	—	—
Chen 2016 [44] †	Y	Y	—	Y	—	—	Y	Y	—	—	—	—
Zhang 2023 [49] †	Y	Y	Y	Y	Y	Y	Y	Y	—	Y	—	—

Y, reported; —, not reported; P, partially reported. DS, data source; GT, ground truth or reference standard; DI, de-identification; DP, data partitions; LK, leakage or overfitting safeguards; SS, sample-size rationale; AR, architecture specified; EM, evaluation metrics defined; EV, external or independent test; FA, failure or limitation analysis; CD, code available; DA, data available.

Risk of bias

Risk-of-bias judgements are presented in Table 3 and summarised in Figure 4. Under QUADAS-2, only 1/13 diagnostic-accuracy studies was at low overall risk of bias, 7/13 were unclear and 5/13 were high. Patient selection was the weakest domain, with 9/13 unclear and 2/13 high, reflecting retrospective single-centre convenience sampling and, in two studies, the exclusion of low-quality images in ways that limit applicability [35,36]. The reference standard was unclear in 6/13 and high in 1/13, typically because expert annotation was used without histological confirmation or any reported measure of inter-examiner agreement. Flow and timing was at low risk in all 13 studies, since a single internal reference standard was applied uniformly. The multi-centre implant study was the only study at low risk across all four domains [46]. Under PROBAST+AI, the analysis domain drove the risk: 2/3 studies were at high risk overall, owing to very small event counts, augmentation-inflated image totals and absent external validation in the debonding model (24 cases) [41] and the decision-support model (104 cases) [44], while the implant-failure model was more robust but still limited by its single-centre design [49].

Table 3. Risk-of-bias judgements. For QUADAS-2, D1–D4 denote patient selection, index test, reference standard, and flow and timing. For PROBAST+AI, D1–D4 denote participants, predictors, outcome and analysis. Generative and design studies [37–39] were appraised for reporting completeness only (Table 2) and are discussed narratively.

Study	Instrument	D1	D2	D3	D4	Overall
Lee 2018 [31]	QUADAS-2	Unclear	Unclear	High	Low	High
Lee 2021 [32]	QUADAS-2	Unclear	Low	Unclear	Low	High
Bayraktar2022 [33]	QUADAS-2	Unclear	Unclear	Unclear	Low	Unclear
Rohrer2022 [34]	QUADAS-2	Low	Low	Unclear	Low	Unclear
Çelik 2022 [35]	QUADAS-2	High	Unclear	Unclear	Low	High
Vinayahalingam	QUADAS-2	High	Low	Low	Low	High

2021 [36]						
Choi 2024 [40]	QUADAS-2	Unclear	Unclear	Unclear	Low	Unclear
Takahashi 2021a [42]	QUADAS-2	Unclear	High	Low	Low	High
Takahashi 2021b [43]	QUADAS-2	Unclear	Unclear	Low	Low	Unclear
Sukegawa 2020 [45]	QUADAS-2	Unclear	Unclear	Low	Low	Unclear
Park 2023 [46]	QUADAS-2	Low	Low	Low	Low	Low
Lee & Jeong 2020 [47]	QUADAS-2	Unclear	Unclear	Low	Low	Unclear
Cha 2021 [48]	QUADAS-2	Unclear	Unclear	Unclear	Low	Unclear
Yamaguchi 2019 [41]	PROBAST+AI	Unclear	Low	Low	High	High
Chen 2016 [44]	PROBAST+AI	Unclear	Unclear	Low	High	High
Zhang 2023 [49]	PROBAST+AI	Low	Low	Low	Unclear	Unclear

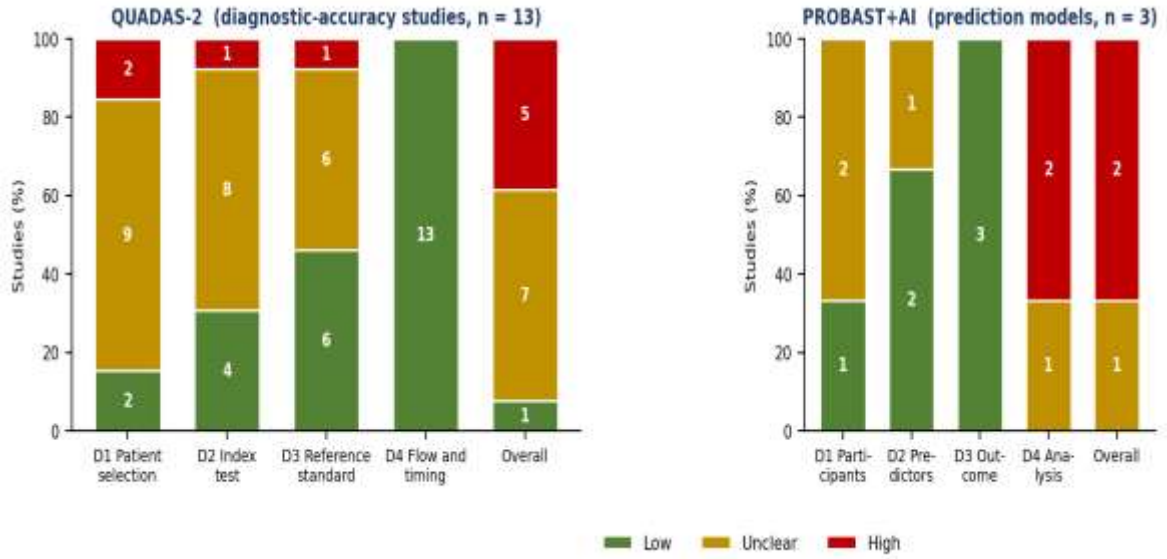

Figure 4. Distribution of risk-of-bias judgements by domain. Numbers within the bars are counts of studies. Patient selection and reference standard are the weakest QUADAS-2 domains; the analysis domain dominates risk under PROBAST+AI.

Recurring deficiencies across the evidence base

Aggregated across all nineteen studies, the deficiencies were strikingly consistent (Figure 5). Every study used a private, non-public dataset and none provided a code- or data-availability statement (19/19 for each). Eighteen of nineteen (95%) drew on a single centre, reported no external or independent validation, offered no sample-size justification, declared no adherence to any reporting guideline and gave no de-identification statement. Class imbalance was left unaddressed in 17/19 (89%), and in 15/19 (79%) the reference standard was an expert annotation for which no inter-examiner reliability was reported. These patterns held across all three clinical domains and were largely independent of the specific task, the imaging modality or the size of the dataset.

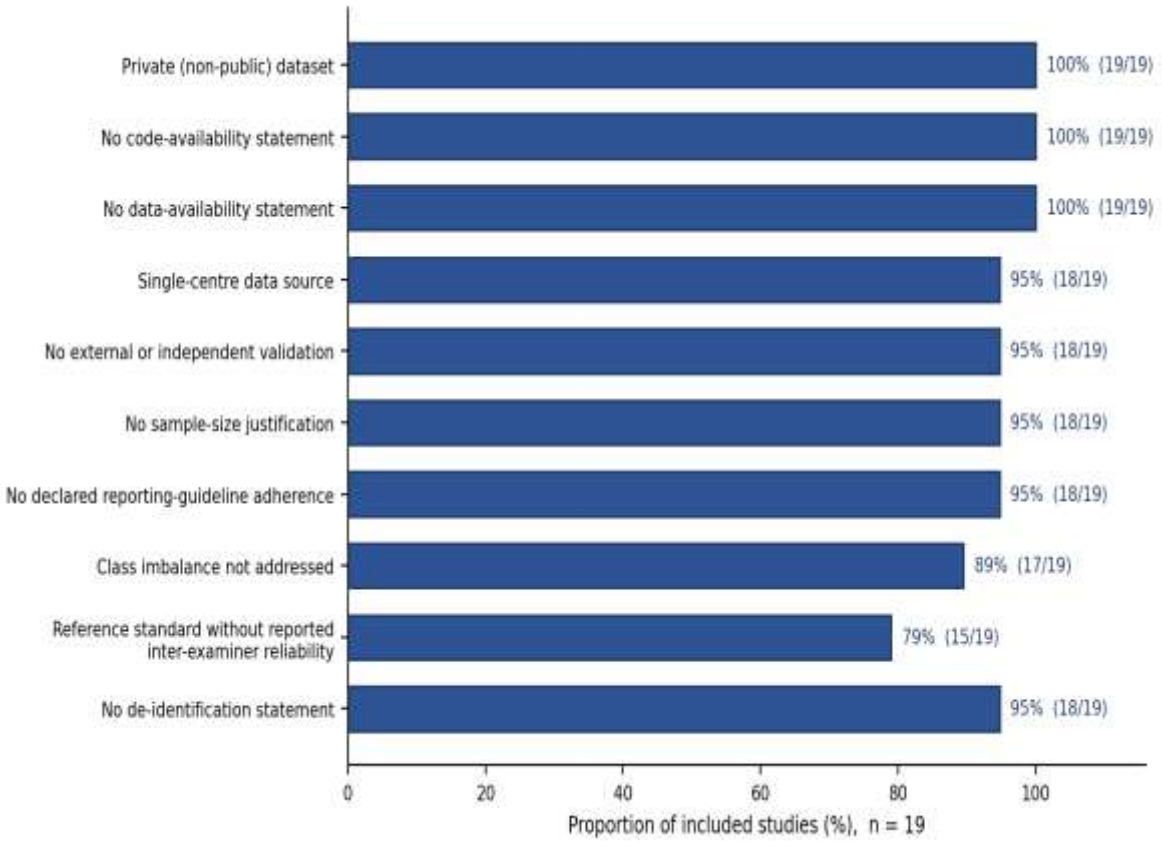

Figure 5. Prevalence of key reporting and methodological deficiencies across the 19 charted studies. Percentages are of all included studies; counts are given in parentheses.

Discussion

This scoping review appraised, at the level of individual reporting domains and across prosthodontics, restorative and implant dentistry together, how completely primary AI studies are reported and how far they are at risk of bias. The central finding is a marked disconnect between performance and provenance: studies report near-ceiling accuracy while omitting the very information—external validation, dataset access, code, sample-size rationale and reference-standard reliability—that would allow those claims to be trusted, reproduced or generalised. In this respect dental AI closely mirrors the wider medical-AI literature, in which comparable appraisals have repeatedly found high apparent accuracy resting on non-reproducible, internally validated, single-centre designs [12–16].

The result is consistent with, and extends, the existing dental evidence. Prior reviews concluded that methodological quality in this field is low and that few studies reach a low risk of bias [2,8,9], but they reached that conclusion largely as an adjunct to describing applications and performance, or—in the case of the umbrella review—by appraising reviews rather than the primary studies beneath them [8]. By routing each primary study to a design-appropriate, AI-specific instrument, the present review converts that general impression into a specific, domain-level diagnosis and updates the dentistry-wide methodological scoping review that predated these instruments [7]. It also demonstrates that the deficiencies are not confined to any one sub-discipline: caries detection, crown design and implant identification share almost the same reporting profile, which points to a field-wide norm rather than to isolated shortcomings.

The chronology in Figure 1 helps explain, though it does not excuse, this pattern. Fifteen of the nineteen studies were published before TRIPOD+AI and the CLAIM update became available, and all of them before PROBAST+AI. Authors were therefore working without instruments that now define the expected standard. The corollary is that the field should improve rapidly if these instruments are adopted, and that reviews such as this one need periodic repetition to detect whether that improvement is occurring.

These findings matter clinically. A model that has never been tested outside the institution and scanner on which it was built cannot be assumed to perform there, and a reference standard defined by a single unblinded annotator may encode that annotator's errors as ground truth. The problem is compounded in prosthodontics by the frequent use of undisclosed commercial CAD algorithms as comparators or components, because neither their training data nor their decision logic can be scrutinised. Without external validation and transparent reporting, the high accuracies now populating the literature offer weak assurance of real-world benefit and may foster premature confidence among clinicians and

manufacturers alike.

Several concrete steps would improve the situation:

Authors should adopt CLAIM or TRIPOD+AI when designing and writing AI studies, and should appraise their own risk of bias with QUADAS-2 or PROBAST+AI before submission.
The reference standard should be defined explicitly, established by more than one calibrated examiner, and accompanied by a reported measure of inter-examiner reliability.
Sample size should be justified prospectively, and the handling of class imbalance stated; augmentation should never be presented as an increase in independent sample size.
Partitioning should be performed at patient rather than image level, and safeguards against data leakage described explicitly.
External validation on data from a different centre, scanner or time period should become an expectation rather than an exception.
Code and de-identified data should be shared wherever ethically and legally possible, and shared public benchmarks for prosthodontic and implant tasks should be developed to enable independent replication.
Journals in the field, including this one, could accelerate the shift by requiring a completed AI reporting checklist at submission, as several imaging and prediction-model journals now do.

Strengths and limitations

The principal strength of this review is its design-matched, domain-level appraisal across three clinical fields, which yields a more actionable picture than application-focused syntheses. Several limitations should be acknowledged. As a scoping review it maps and characterises the evidence without pooling it, so no quantitative estimate of diagnostic accuracy is offered. The twelve-domain condensation of CLAIM and TRIPOD+AI, while stated a priori and reproducible, is an adaptation and is not interchangeable with item-by-item scoring against the full instruments; absolute completeness rates would differ under full scoring, although the relative pattern is unlikely to change. Some appraisal items could be scored only from the accessible portions of full texts; where a detail was absent from the sources available it was recorded as not reported, which is epistemically distinct from confirming that the authors did not perform the step, and the ratings are therefore provisional pending confirmation against the complete published records. Excluding non-English publications may have narrowed the geographical range further. Finally, the field is evolving quickly, and the search should be refreshed immediately before publication so that recent studies—which may report to a higher standard—are captured.

Conclusions

Within prosthodontics, restorative and implant dentistry, AI studies report consistently high accuracy but fall well short of contemporary reporting and risk-of-bias standards. External validation is almost never performed, code and data are not shared, sample sizes are not justified, and reference standards are often incompletely characterised. Until these gaps are closed through standardised reporting, external validation and open-science practices, current performance claims should be regarded as promising but unproven, and clinical adoption should proceed with corresponding caution.

Declarations

Ethical approval: Not required; this review used only previously published data.
Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.
Conflicts of interest: The authors declare no conflict of interest.
Data availability: The charting form, the CLAIM/TRIPOD+AI domain mapping and the completed PRISMA-ScR checklist are available with the registered protocol and from the corresponding author on reasonable request.

REFERENCES

1. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99(7):769-774.
2. Bernauer SA, Zitzmann NU, Joda T. The use and performance of artificial intelligence in prosthodontics: a systematic review. Sensors (Basel). 2021;21(19):6628.
3. Revilla-León M, Gómez-Polo M, Vyas S, Barmak AB, Özcan M, Att W, et al. Artificial intelligence applications in restorative dentistry: a systematic review. J Prosthet Dent. 2022;128(5):867-875.
4. Revilla-León M, Gómez-Polo M, Vyas S, Barmak AB, Gallucci GO, Att W, et al. Artificial intelligence models for tooth-supported fixed and removable prosthodontics: a systematic review. J Prosthet Dent. 2023;129(2):276-292.
5. Revilla-León M, Gómez-Polo M, Vyas S, Barmak BA, Gallucci GO, Att W, et al. Artificial intelligence

applications in implant dentistry: a systematic review. J Prosthet Dent. 2023;129(2):293-300.
6. Kong HJ, Kim YL. Application of artificial intelligence in dental crown prosthesis: a scoping review. BMC Oral Health. 2024;24(1):937.
7. Arsiwala-Scheppach LT, Chaurasia A, Müller A, Krois J, Schwendicke F. Machine learning in dentistry: a scoping review. J Clin Med. 2023;12(3):937.
8. Alfaraj A, Limones Á, Ahmad S, Aljubairah F, Albalawi S, Albesher M, et al. Harnessing artificial intelligence in prosthodontics and implant dentistry: an umbrella review of systematic evidence. J Prosthodont. 2026;35(2):127-142.
9. Aljulayfi IS, Almatrafi AH, Althubaitiy RO, Alharbi FA, Alqahtani MS, Gufran K. The potential of artificial intelligence in prosthodontics: a comprehensive review. Med Sci Monit. 2024;30:e944310.
10. Yeslam HE, Freifrau von Maltzahn N, Nassar HM. Revolutionizing CAD/CAM-based restorative dental processes and materials with artificial intelligence: a concise narrative review. PeerJ. 2024;12:e17793.
11. Ali IE, Tanikawa C, Chikai M, Ino S, Sumita Y, Wakabayashi N. Applications and performance of artificial intelligence models in removable prosthodontics: a literature review. J Prosthodont Res. 2024;68(3):358-367.
12. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.
13. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3(3):199-217.
14. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e297.
15. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021;4(1):65.
16. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. Diagn Progn Res. 2022;6(1):13.
17. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2):e200029.
18. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6(4):e240300.
19. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
20. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.
21. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008.
22. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536.
23. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.
24. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-1374.
25. Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, et al. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent. 2021;107:103610.
26. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol. 2005;8(1):19-32.
27. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci. 2010;5:69.
28. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc. 2015;13(3):141-146.
29. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169(7):467-473.

30. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent. 2018;77:106-111.
31. Lee S, Oh SI, Jo J, Kang S, Shin Y, Park JW. Deep learning for early dental caries detection in bitewing radiographs. Sci Rep. 2021;11(1):16807.
32. Bayraktar Y, Ayan E. Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs. Clin Oral Investig. 2022;26(1):623-632.
33. Rohrer C, Krois J, Patel J, Meyer-Lueckel H, Rodrigues JA, Schwendicke F. Segmentation of dental restorations on panoramic radiographs using deep learning. Diagnostics (Basel). 2022;12(6):1316.
34. Çelik B, Çelik ME. Automated detection of dental restorations using deep learning on panoramic radiographs. Dentomaxillofac Radiol. 2022;51(8):20220244.
35. Vinayahalingam S, Goey RS, Kempers S, Schoep J, Cherici T, Moin DA, et al. Automated chart filing on panoramic radiographs using deep learning. J Dent. 2021;115:103864.
36. Tian S, Wang M, Dai N, Ma H, Li L, Fiorenza L, et al. DCPR-GAN: dental crown prosthesis restoration using two-stage generative adversarial networks. IEEE J Biomed Health Inform. 2022;26(1):151-160.
37. Ding H, Cui Z, Maghami E, Chen Y, Matinlinna JP, Pow EHN, et al. Morphology and mechanical performance of dental crown designed by 3D-DCGAN. Dent Mater. 2023;39(3):320-332.
38. Cho JH, Yi Y, Choi J, Ahn J, Yoon HI, Yilmaz B. Time efficiency, occlusal morphology, and internal fit of anatomic contour crowns designed by dental software powered by generative adversarial network: a comparative study. J Dent. 2023;138:104739.
39. Choi J, Ahn J, Park JM. Deep learning-based automated detection of the dental crown finish line: an accuracy study. J Prosthet Dent. 2024;132(6):1286.e1-1286.e9.
40. Yamaguchi S, Lee C, Karaer O, Ban S, Mine A, Imazato S. Predicting the debonding of CAD/CAM composite resin crowns with artificial intelligence. J Dent Res. 2019;98(11):1234-1238.
41. Takahashi T, Nozaki K, Gonda T, Ikebe K. A system for designing removable partial dentures using artificial intelligence. Part 1. Classification of partially edentulous arches using a convolutional neural network. J Prosthodont Res. 2021;65(1):115-118.
42. Takahashi T, Nozaki K, Gonda T, Mameno T, Ikebe K. Deep learning-based detection of dental prostheses and restorations. Sci Rep. 2021;11(1):1960.
43. Chen Q, Wu J, Li S, Lyu P, Wang Y, Li M. An ontology-driven, case-based clinical decision support model for removable partial denture design. Sci Rep. 2016;6:27855.
44. Sukegawa S, Yoshii K, Hara T, Yamashita K, Nakano K, Yamamoto N, et al. Deep neural networks for dental implant system classification. Biomolecules. 2020;10(7):984.
45. Park W, Huh JK, Lee JH. Automated deep learning for classification of dental implant radiographs using a large multi-center dataset. Sci Rep. 2023;13(1):4862.
46. Lee JH, Jeong SN. Efficacy of deep convolutional neural network algorithm for the identification and classification of dental implant systems, using panoramic and periapical radiographs: a pilot study. Medicine (Baltimore). 2020;99(26):e20787.
47. Cha JY, Yoon HI, Yeo IS, Huh KH, Han JS. Peri-implant bone loss measurement using a region-based convolutional neural network on dental periapical radiographs. J Clin Med. 2021;10(5):1009.
48. Zhang C, Fan L, Zhang S, Zhao J, Gu Y. Deep learning-based dental implant failure prediction from periapical and panoramic films. Quant Imaging Med Surg. 2023;13(2):935-945.