

Deep learning for dental implant brand classification: a systematic review of imaging modalities, methodological variability, and model performance systematic review

Ozden Melis Durmaz Yilmaz^{ID} · Ikbal Leblebicioglu Kurtulus

Received: 14 May 2026 / Accepted: 11 September 2026

© The Author(s) 2026

Abstract

Background Accurate identification of dental implant brands from radiographic images is clinically important, particularly when previous treatment records are unavailable. Deep learning (DL) models have shown promising performance for automated implant classification, yet it remains unclear how reliably this performance translates across different imaging modalities, datasets, and validation strategies.

Objective To systematically evaluate the diagnostic performance and methodological characteristics of deep learning models for dental implant brand classification using panoramic, periapical, and mixed radiographic datasets.

Methods This systematic review was reported in accordance with the PRISMA 2020 guidelines. PubMed, Scopus, Web of Science, and IEEE Xplore databases were searched for studies published between 2020 and 2026. Studies employing CNN-based, YOLO-based, transformer-based, AutoML, or hybrid deep learning models for implant brand classification were included. Extracted data included imaging modality, dataset characteristics, model architecture, validation strategy, and diagnostic performance metrics. Risk of bias was assessed using the QUADAS-2 tool. Because the performance metrics reported across the included studies were inconsistent and not statistically poolable, a structured narrative synthesis was performed in accordance with the Synthesis Without Meta-analysis (SWiM) guidelines rather than a meta-analysis.

Results Twenty studies met the inclusion criteria. Reported diagnostic accuracy ranged approximately from 75% to 99%. Studies using smaller datasets, limited implant diversity, and manually cropped implant regions generally reported higher performance, whereas large multicenter datasets demonstrated comparatively lower but potentially more clinically realistic results. Panoramic radiographs were the most commonly used imaging modality, although periapical and mixed datasets also demonstrated strong diagnostic capability. Earlier studies primarily used transfer learning-based CNN architectures, while more recent investigations incorporated attention mechanisms, hybrid architectures, synthetic augmentation, and automated object-detection workflows. However, true geographically independent external validation remained uncommon.

Conclusion Deep learning models may demonstrate potential for dental implant brand classification. Nevertheless, diagnostic performance is influenced by dataset composition, implant diversity, preprocessing workflow,

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



and validation strategy rather than model architecture alone. Future studies should prioritize multicenter external validation, standardized reporting, explainable AI integration, and clinically applicable automated workflows.

Keywords Dental implants · Deep learning · Artificial intelligence · Implant classification · Panoramic radiography · Systematic review

Introduction

Dental implant treatments have become one of the most widely utilized prosthetic solutions in contemporary dentistry due to their high long-term success rates in both partially and fully edentulous patients [1]. However, the global increase in implant placement has introduced a critical clinical challenge: the accurate identification of previously placed implant systems [2–4]. This issue is particularly relevant in cases where patient records are incomplete or unavailable. Furthermore, variability in implant-specific connection designs and prosthetic components complicates appropriate restorative and surgical planning [2].

Radiographic evaluation plays a central role in implant identification. Panoramic radiographs are widely used in clinical practice due to their ability to provide a comprehensive overview of the maxillofacial region in a standardized and accessible manner [2, 5, 6]. However, despite these advantages, panoramic imaging is subject to inherent limitations, including geometric distortion, superimposition of anatomical structures, and reduced spatial resolution compared to intraoral radiographs. In contrast, periapical radiographs may provide more detailed visualization of implant-specific features, such as thread design, tapering, and fixture morphology, which are essential for precise implant system identification. Nevertheless, panoramic radiographs remain one of the most commonly available and routinely acquired imaging modalities in large-scale diagnostic workflows, making them highly relevant for the development of automated diagnostic tools [7].

Recent advances in artificial intelligence (AI), particularly deep learning (DL)-based image analysis, have significantly transformed medical and dental radiology [8–10]. Among these approaches, convolutional neural networks (CNNs), YOLO-based object detection models, modern CNN and hybrid architectures, and hybrid deep learning frameworks have increasingly been applied to radiographic image analysis. These models can learn complex spatial patterns directly from raw radiographic images, enabling automated detection and classification while reducing the reliance on the handcrafted feature extraction required by conventional image-analysis pipelines [11–13].

In the context of implant identification, CNN-based models have demonstrated the ability to recognize subtle morphological characteristics of implant systems from radiographic images, thereby facilitating automatic classification of implant brands and reducing operator-dependent variability [2, 5, 14]. Previous studies have reported promising diagnostic performance, with accuracy values ranging approximately between 75% and 99% [2, 5, 14–18].

Given the increasing heterogeneity in study designs, imaging modalities, and deep learning methodologies, there is a critical need for a comprehensive and methodologically robust synthesis of the existing literature. In particular, variations in dataset composition, number of implant classes, validation strategies, and reporting standards complicate direct comparison of model performance across studies. Furthermore, the potential for selection bias—particularly in reviews that selectively include high-performing models—may lead to overestimation of diagnostic capabilities. To our knowledge, this is the first systematic review synthesizing implant classification studies across panoramic, periapical, and mixed imaging modalities while explicitly examining methodological heterogeneity and validation strategies.

Therefore, this systematic review aims to evaluate the diagnostic performance of deep learning models for dental implant brand classification across different radiographic modalities, including panoramic, periapical, and mixed datasets. In addition, this review investigates the impact of methodological factors—such as model

architecture, data augmentation strategies, and validation approaches—on reported performance outcomes, with a focus on identifying sources of variability and potential bias in the current literature.

Materials and methods

Study design and reporting guidelines

This systematic review was reported in accordance with the PRISMA 2020 statement for reporting systematic reviews [19]. Additionally, the synthesis of findings without meta-analysis was reported in accordance with the Synthesis Without Meta-analysis (SWiM) reporting guideline [20].

The review was guided by the following question: in patients undergoing radiographic evaluation, what is the diagnostic performance of deep learning models for dental implant brand or system classification on panoramic, periapical, or mixed radiographs, and which methodological factors influence the reported performance?

The review protocol was not prospectively registered; however, to enhance transparency, the protocol was deposited in the Open Science Framework (OSF) after completion of the review (registration DOI: <https://doi.org/10.17605/OSF.IO/984AP>). The eligibility criteria, search strategy, and synthesis methods were established before study selection, and no post hoc methodological modifications were introduced. Nevertheless, the eligibility criteria, search strategy, and synthesis methods were established before completion of study selection and data synthesis, and no post hoc methodological modifications were introduced during the review process.

Eligibility criteria

The eligibility criteria were defined based on study design, imaging modality, methodological characteristics, and outcomes of interest.

Inclusion criteria

Original research articles employing deep learning architectures, including CNN-, transformer-, YOLO-, AutoML-, or hybrid models, for dental implant brand or system classification. These architecture families represent distinct design paradigms: standard CNNs perform whole-image classification, YOLO-based models jointly localize and classify the implant, and transformer-based and AutoML frameworks differ in how features are learned. Object-detection architectures were included explicitly, as they are increasingly used in recent implant studies.

Studies using panoramic, periapical, or mixed radiographic datasets were included to better reflect real-world clinical imaging conditions and reduce modality-related selection bias.

Reporting at least one diagnostic performance metric (e.g., accuracy, AUC, F1-score, sensitivity, specificity).

Published in peer-reviewed journals between January 2020 and June 2026.

Articles written in English. The restriction to English-language articles was applied for feasibility; the resulting potential for language bias is acknowledged as a limitation.

Exclusion criteria

Studies focusing exclusively on implant detection or segmentation without explicit implant brand classification outcomes were excluded.

Non-deep learning approaches (e.g., classical machine learning or rule-based systems).

Studies lacking sufficient methodological transparency (e.g., missing validation details).

Studies based exclusively on synthetic or simulated datasets without real clinical radiographic images were excluded.

Reviews, conference abstracts, editorials, or non-peer-reviewed literature.

When duplicate publications or substantially overlapping datasets addressed the same research question, only the most comprehensive version was included. When overlapping datasets were used to address distinct objectives, such as robustness testing or human–AI comparison, this overlap was recorded and considered during interpretation.

No performance-based thresholds (e.g., accuracy cut-offs) were applied as inclusion or exclusion criteria in order to avoid selection bias.

Search strategy

A comprehensive literature search was conducted across the following electronic databases: PubMed (MEDLINE), Scopus, Web of Science, IEEE Xplore. The search strategy combined terms related to dental implants, artificial intelligence, deep learning, and radiographic imaging modalities. The complete electronic search strategies for all databases are provided in Supplementary Table S1.

Search strategies were developed using combinations of controlled vocabulary and free-text terms related to dental implants, implant identification, radiographic imaging, artificial intelligence, deep learning, convolutional neural networks, YOLO, object detection, transformers, and transfer learning. The final literature search was conducted in June 2026. In addition to database searches, the reference lists of included studies were manually screened to identify additional relevant articles. Duplicate records were removed using reference management software, followed by manual verification. The review was intentionally restricted to peer-reviewed journal articles to ensure a consistent minimum standard of methodological reporting and reproducibility across the heterogeneous AI literature; grey literature (conference abstracts, preprints, and theses) was not searched. The potential for publication bias arising from this restriction is addressed in the limitations.

Study selection

Two independent reviewers (O.M.D.Y. and I.L.K) screened titles and abstracts of all retrieved records. Full-text articles were then assessed for eligibility based on predefined criteria. The study selection process is illustrated using a PRISMA flow diagram.

Data extraction

Data were extracted independently by two reviewers using a standardized extraction form. The following variables were recorded:

- Study characteristics (authors, year, country),
- Dataset properties (sample size, number of implant systems, data source),
- Model characteristics (architecture type, transfer learning, auxiliary learning strategies),
- Preprocessing techniques and data augmentation methods,
- Validation strategies (k-fold cross-validation, hold-out validation, external validation),
- Performance metrics (accuracy, AUC, F1-score, sensitivity, specificity).

Titles, abstracts, and full texts were independently screened in duplicate by two reviewers (O.M.D.Y. and I.L.K) against the predefined eligibility criteria. Disagreements were resolved through discussion until consensus was reached; in the few cases where consensus could not be reached, a third reviewer (S.Y) was consulted to adjudicate.

Risk of bias assessment

Risk of bias was assessed using the QUADAS-2 tool across four domains [21]: Patient Selection, Index Test, Reference Standard, Flow and Timing. Each domain was rated as “low,” “high,” or “unclear” risk of bias. To improve transparency, domain-level judgments were supported by study-specific justifications derived from the original articles. In addition, the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) was used as a conceptual reporting framework to inform the interpretation of study reporting quality in AI-relevant domains such as external validation, transparency, and reproducibility [22]; it was not applied as a formal risk-of-bias assessment tool. As QUADAS-2 was originally developed for primary diagnostic accuracy studies, its judgements in this review should be interpreted as an adaptation to the AI context.

Data synthesis and analysis

Due to substantial heterogeneity in study designs, dataset characteristics, number of implant classes, validation strategies, and reported performance metrics, a quantitative meta-analysis was not considered appropriate. Instead, a structured narrative synthesis was performed in accordance with SWiM recommendations. Subgroup analyses were conducted based on imaging modality (panoramic, periapical, and mixed), model architecture (CNN-based, YOLO-based, and transformer-based), validation strategy (internal vs. external), and dataset characteristics (sample size and number of implant classes). Where available, additional methodological factors such as data augmentation and explainable AI techniques were also qualitatively evaluated.

Consideration of bias and evidence limitations

Given the nature of the included studies, additional sources of bias were considered qualitatively, including:

- Selection bias related to dataset composition,
- Publication bias favoring high-performing models,
- Center bias due to single-center retrospective datasets,
- Validation bias arising from reliance on internal validation methods.

These factors were taken into account during interpretation of the findings.

Result

Study selection

The study selection process is illustrated in Fig. 1. The search was subsequently updated and extended through June 2026 to ensure inclusion of the most recent evidence. The final, combined search of four electronic databases (PubMed, Scopus, Web of Science, and IEEE Xplore), restricted to English-language, peer-reviewed articles, identified a total of 715 records. After removal of 146 duplicate records, 569 records were screened on the basis of title and abstract. Of these, 512 records were excluded for not meeting the predefined eligibility criteria, leaving 57 full-text articles that were assessed for eligibility. Following full-text evaluation, 37 articles were excluded; the reasons for exclusion, with corresponding counts, are reported in Supplementary Table S2. Ultimately, 20 studies met all predefined inclusion criteria and were included in the qualitative synthesis, five of which were recently published studies (2025–2026) identified during the updated search. The most frequent reasons for full-text exclusion were the absence of an implant brand or system classification outcome, the use of non-deep-learning methodologies, and insufficient methodological or performance data for extraction.

Identification**Records identified (n = 715)**

PubMed 69 · Scopus 575 · WoS 70 · IEEE 1

Duplicate records removed

(n = 146)

Screening**Records screened (n = 569)****Records excluded**

title / abstract (n = 512)

Full-text articles assessed (n = 57)**Reports excluded (n = 37)**

No brand classification: 15

Not deep learning: 9

Insufficient data: 9

Overlapping dataset: 4

Included**Studies included in review (n = 20)**

Qualitative synthesis (SWiM)

Databases: PubMed, Scopus, Web of Science, IEEE Xplore

Filters: 2020–2026 · English · peer-reviewed articles

Full-text exclusion reasons detailed in Supplementary Table S2

Fig. 1 PRISMA 2020 flow diagram illustrating the study selection process

Studies were not excluded on the basis of reported diagnostic performance in order to minimize selection bias and ensure comprehensive representation of the available evidence. In addition, studies using overlapping datasets were retained when they addressed distinct objectives, such as robustness analysis, human–AI collaboration, synthetic augmentation, or automated detection workflows.

Characteristics of included studies

The characteristics of the included studies are summarized in Table 1. The included studies demonstrated substantial methodological heterogeneity in terms of imaging modality, dataset size, implant diversity, model architecture, preprocessing strategies, and validation methodologies.

Panoramic radiographs were the most frequently used imaging modality and were utilized in several studies, including Sukegawa et al. [15], Sukegawa et al. [16], Lubbad et al. [5], Tiryaki et al. [17], and Arijj et al. [18]. Periapical radiographs were employed in studies such as Kim et al. [23], Kim et al. [24], Kong et al. [25], and

Table 1 Characteristics and methodological features of included studies on deep learning–based dental implant brand classification

Study	Country	Imaging Modality	Dataset Characteristics	Implant Systems (<i>n</i>)	Main AI Approach	ROI Strategy	Validation Strategy	External Validation	Explainable AI	Key Methodological Notes
Kim et al. [23], 2020	Korea	Periapical	Single-center retrospective dataset	4	Transfer learning CNN	Manual cropped ROI	Hold-out split (6:2:2)	No	CAM	Periapical implant classification with explainability integration
Lee & Jeong [30], 2020	Korea	Panoramic + Periapical	Single-center retrospective dataset	3	Transfer learning CNN	Manual cropped ROI	Hold-out validation	No	No	Human expert comparison and modality-specific analysis
Lee et al. [27], 2020	Korea	Panoramic + Periapical	Multicenter retrospective dataset	6	Automated deep learning framework (AutoDL)	Manual cropped ROI	Hold-out validation	No	No	Large multi-center dataset with clinician comparison
Sukegawa et al. [15], 2021	Japan	Panoramic	Single-center retrospective dataset	12	Multi-task CNN framework	Manual cropped ROI	4-fold cross-validation	No	Grad-CAM	Simultaneous implant and treatment-stage classification
Sukegawa et al. [16], 2022	Japan	Panoramic	Single-center retrospective dataset	13	CNN + attention mechanism	Manual cropped ROI	4-fold cross-validation	No	Grad-CAM + attention heatmaps	Attention-based learning with statistical comparison and effect-size analysis
Kim et al. [24], 2022	Korea	Periapical	Single-center retrospective dataset	3	YOLO-based detection/classification	Bounding-box annotation	Hold-out validation	No	No	Combined implant detection and classification workflow
Lee et al. [39], 2021	Korea	Panoramic	Multicenter retrospective dataset	6	Automated deep learning framework (AutoDL)	Manual cropped ROI	Hold-out validation with independent test cohort	Partial	Activation maps	Human–AI collaborative implant classification study
Kong et al. [25], 2023	Korea	Periapical	Single-center retrospective dataset	4	Cloud-based AutoML framework	Manual cropped ROI	Hold-out validation	No	No	Low-code AI workflow with limited algorithmic transparency

Table 1 (continued)

Study	Country	Imaging Modality	Dataset Characteristics	Implant Systems (<i>n</i>)	Main AI Approach	ROI Strategy	Validation Strategy	External Validation	Explainable AI	Key Methodological Notes
Park et al. [28], 2023	Korea	Pan- oramic + Peri- apical	Large multicenter heterogeneous dataset	27	Automated deep learning framework (AutoDL)	Manual cropped ROI	Hold-out validation	No	Grad-CAM	Large-scale multicenter study with modality comparison
Lee et al. [29], 2024	Korea	Pan- oramic + Peri- apical	Multicenter dataset with independent distorted test set	27 (training), 9 (testing)	Transfer learning CNN	Manual cropped ROI	Hold-out validation + independent distorted test cohort	Partial	Conceptually discussed	Robustness analysis under clinically distorted imaging conditions
Lubbad et al. [5], 2024	Türkiye	Panoramic	Single-center retrospective dataset	6	CNN/transformer benchmarking framework	Manual cropped ROI	Hold-out validation	No	No	Comparative evaluation of 28 deep learning architectures
Ariji et al. [18], 2025	Japan	Panoramic	Single-center dataset with external testing	8	Two-step detection/classification pipeline	Automatic detection pipeline	Hold-out validation + external testing	Partial	No	Clinically oriented automated implant detection and classification workflow
Lin et al. [26], 2025	Taiwan	Periapical	Single-center retrospective dataset	NR	YOLO-based object detection framework	Automatic detection pipeline	Hold-out validation	No	No	Modern end-to-end implant detection workflow using YOLOv10
Sukegawa et al. [38], 2025	Japan	Panoramic	Single-center retrospective dataset	10	CNN with synthetic radiographic augmentation	Manual cropped ROI	10-fold cross-validation	No	t-SNE feature visualization	STL-based synthetic panoramic image generation
Tiryaki et al. [17], 2025	Türkiye	Panoramic	Single-center retrospective dataset	5	Transfer learning CNN with ensemble fusion	Manual cropped ROI	5-fold cross-validation	No	No	Combined implant classification and angle estimation framework

Table 1 (continued)

Study	Country	Imaging Modality	Dataset Characteristics	Implant Systems (<i>n</i>)	Main AI Approach	ROI Strategy	Validation Strategy	External Validation	Explainable AI	Key Methodological Notes
Yuce et al. [31], 2025	Türkiye	Panoramic	Single-center retrospective dataset (275 patients; 1075 cropped; 5375 after augmentation)	4	Transfer learning CNN (GoogLeNet, ResNet-18, VGG16, ShuffleNet)	Manual cropped ROI+CLAHE	Hold-out validation (80:20)	No	Grad-CAM	First statistical comparison of four CNNs across four brands; first reported ShuffleNet use for implant classification
Bena-katti et al. [34], 2025	India	Pan-oramic + Peri-apical	Single-center dataset (1300 collected; 1138 retained; 1744 after augmentation)	5	DEtection TRansformer (DETR)	Automatic detection pipeline	Hold-out validation (70:15:15)	No	No	First trans-former-based detector (DETR) for implant identification; author-reported train-validation generalization gap
Lashaki et al. [32], 2025	Iran (public dataset)	Periapical	Public Robo-flow dataset (511 images; 5110 after augmentation)	3	CNN with hybrid classifiers (CSVM/CDT/CRF)+pre-trained VGG16/ResNet50/Xception	Implant-specific masking + CLAHE	5-fold cross-validation	No	No	Pairwise binary classification design; preprocessing/masking emphasis on a public, imbalanced dataset
Pachaiappan et al. [35], 2026	India	Pan-oramic + Peri-apical	Single-center dataset (4677 images; 8189 implants; ×3 augmentation)	7	YOLOv10-based object detection	Automatic detection pipeline (Roboflow)	Hold-out validation (80:10:10)	No	No	System-level (not subtype) detection and classification; marked class imbalance across brands

Table 1 (continued)

Study	Country	Imaging Modality	Dataset Characteristics	Implant Systems (<i>n</i>)	Main AI Approach	ROI Strategy	Validation Strategy	External Validation	Explainable AI	Key Methodological Notes
Ahmed et al. [33], 2026	Saudi Arabia	Periapical	Two-source dataset (6092 images; 5851 after filtering to five brands)	5	YOLOv12x object detection (YOLOv7–v12 and DF-DETR benchmarked)	Automatic detection pipeline (Roboflow)	Hold-out validation (80:10:10)	No	No	Cross-version YOLO benchmarking; human–AI comparison with Cohen’s kappa; STARD/CLAIM-adherent reporting

Lin et al. [26], whereas mixed panoramic–periapical datasets were used in multicenter investigations by Lee et al. [27], Park et al. [28], and Lee et al. [29].

Dataset size varied considerably across studies, ranging from 355 implant fixtures in a YOLOv3-based periapical study by Kim et al. [24] to 156,965 radiographic images in the large multicenter study by Park et al. [28]. Similarly, the number of implant classes varied substantially, ranging from 3 implant systems in smaller classification studies to 27 implant systems in large multicenter datasets.

Most earlier studies employed transfer learning–based convolutional neural network architectures, particularly VGG- and ResNet-based models [23, 30]. More recent studies incorporated advanced architectures and learning strategies, including attention mechanisms, transformer-based models, hybrid CNN–transformer frameworks, AutoML systems, and YOLO-based object detection pipelines [5, 16, 18, 25, 26].

Validation methodologies also differed substantially among studies. Most studies relied primarily on internal validation approaches, including hold-out validation and cross-validation, whereas only a limited number incorporated independent multicenter datasets or external testing cohorts.

The five most recent studies (2025–2026) reflected the same overall modality distribution. Yuce et al. [31] used panoramic radiographs; Lashaki et al. [32] and Ahmed et al. [33] used periapical radiographs; and Benakatti et al. [34] and Pachaiappan et al. [35] used mixed panoramic–periapical datasets. These studies also broadened the geographic representation of the evidence base beyond the predominantly Korean and Japanese cohorts of the original sample, adding investigations from Türkiye, India, Iran, and Saudi Arabia. In terms of architecture, the most recent studies were weighted toward the most recent object-detection frameworks: Benakatti et al. [34] employed a DETection TRansformer (DETR), Pachaiappan et al. [35] used YOLOv10, and Ahmed et al. [33] used YOLOv12x, whereas Yuce et al. [31] compared four transfer-learning CNNs and Lashaki et al. [32] combined CNN feature extraction with classical machine-learning classifiers and pre-trained networks. One study (Lashaki et al. [32]) relied on a publicly available radiographic dataset rather than an institutional archive, and two (Pachaiappan et al. [35]; Ahmed et al. [33]) were published in 2026, reflecting the continued and rapid expansion of this field.

Risk of bias assessment

Risk-of-bias assessment using the QUADAS-2 framework demonstrated that most included studies showed relatively low risk in the reference standard domain, as implant labels were generally derived from institutional records or clinician-confirmed implant systems. Similarly, the index test domain was frequently rated as low risk in studies that clearly described model architecture, validation procedures, and performance evaluation methods.

However, several studies using proprietary or AutoML-based systems demonstrated unclear risk due to limited algorithmic transparency and insufficient reporting of model optimization procedures [25, 27].

The most common methodological concern across the included studies was related to patient selection and dataset composition. Most investigations were based on retrospective single-center datasets with varying degrees of class imbalance and limited implant diversity [17, 28]. Several studies included substantially unequal distributions of implant systems, potentially favoring overrepresented implant classes and artificially inflating classification performance. In addition, many studies relied on manually cropped implant regions of interest prior to classification, thereby reducing workflow realism and limiting direct clinical applicability.

Another major limitation identified across the literature was the limited use of external validation. Although some studies incorporated multicenter datasets or independent testing cohorts, true geographically independent external validation remained uncommon [18, 28, 29]. In several multicenter studies, training and testing datasets were derived from overlapping institutional sources, potentially limiting generalizability despite large sample sizes. Furthermore, robustness analyses under clinically challenging imaging conditions were performed in only a limited number of studies.

The present review additionally highlights the need for a more nuanced interpretation of validation strategies in implant classification studies. Although many investigations reported internal testing procedures, the methodological rigor of validation approaches varied substantially across the literature. Building on established distinctions between internal, temporal, and geographically external validation, we propose—as an interpretive framework for the purposes of this review—a progressive validation hierarchy [36, 37]. Within this framework, validation strategies range from single-center internal hold-out validation (Level 1) and multicenter internal validation using shared datasets (Level 2), to independent robustness testing under distorted imaging conditions (Level 3), and finally true geographically independent external validation cohorts (Level 4). Importantly, most currently available studies remained within Levels 1–2, whereas only a limited number incorporated partially independent testing datasets or robustness-oriented evaluation frameworks [18, 29]. True geographically external validation across different institutions, imaging devices, and patient populations remained exceptionally uncommon. This limitation may contribute to overestimation of model generalizability and highlights the need for more rigorous real-world validation protocols in future implant AI research.

Overall, the available evidence suggests that although deep learning models demonstrate promising diagnostic performance for implant brand classification, the methodological quality and generalizability of the current literature remain constrained by retrospective study designs, ROI-dependent workflows, class imbalance, and insufficient external validation. The detailed QUADAS-2 assessment and major methodological concerns are summarized in Table 2.

Imaging modality analysis

Substantial variability in diagnostic performance was observed across imaging modalities.

Panoramic radiograph-based studies generally reported high classification performance under controlled conditions. Several CNN-based studies reported accuracy values exceeding 95%, particularly in datasets involving manually cropped implant regions and limited implant diversity [15–17]. However, studies involving larger and more heterogeneous multicenter datasets demonstrated comparatively lower but potentially more clinically realistic performance values. For example, Park et al. [28], which included 27 implant systems across large multicenter datasets, reported an overall accuracy of approximately 88.5%. Consistent with this pattern, Yuce et al. [31] reported 98.9% accuracy with a VGG16 model trained on manually cropped, contrast-limited adaptive histogram equalization (CLAHE)-enhanced panoramic images of four implant brands, again under controlled single-center conditions with limited implant diversity.

Periapical radiograph-based studies also demonstrated strong diagnostic performance. Accuracy values generally ranged from approximately 85% to 98% in studies using periapical imaging [23–25]. These findings may reflect the higher spatial resolution and improved visualization of implant thread morphology, tapering, and

Table 2 QUADAS-2 risk-of-bias assessment and major methodological concerns

Study	Patient Selection	Index Test	Reference Standard	Flow & Timing	Major Methodological Concerns
Kim et al. [23], 2020	Low	Low	Low	Unclear	Small single-center dataset and manually cropped ROI workflow
Lee & Jeong [30], 2020	Low	Low	Low	Unclear	Limited implant diversity and lack of external validation
Lee et al. [27], 2020	Low	Low	Low	Unclear	Possible class imbalance and limited methodological transparency of proprietary AutoDL framework
Sukegawa et al. [15], 2021	Low	Low	Low	Low	Manual ROI dependency and absence of external validation
Sukegawa et al. [16], 2022	Unclear	Low	Low	Unclear	Very limited implant diversity and operator-dependent bounding-box annotation
Kim et al. [24], 2022	Low	Unclear	Low	Unclear	Combined detection-classification workflow with limited implant diversity
Lee et al. [39], 2021	Low	Low	Low	Low	Human–AI collaborative workflow and limited external generalizability
Kong et al. [25], 2023	Low	Unclear	Low	Unclear	Limited algorithmic transparency and lack of explainable AI
Park et al. [28], 2023	Low	Low	Low	Unclear	Large multicenter dataset but no true geographically external validation; substantial class imbalance
Lee et al. [29], 2024	Low	Low	Low	Unclear	Overlapping training dataset with prior studies and limited distorted-image test cohort
Lubbad et al. [5], 2024	Low	Low	Low	Unclear	Small single-center dataset and lack of external validation
Ariji et al. [18], 2025	Low	Low	Low	Low	Limited implant diversity despite automated workflow and partial external validation
Lin et al. [26], 2025	Unclear	Low	Low	Unclear	Insufficient reporting of external validation and dataset composition
Sukegawa et al. [38], 2025	Low	Low	Low	Unclear	Synthetic image augmentation may limit direct real-world generalizability
Tiryaki et al. [17], 2025	Low	Low	Low	Unclear	Class imbalance and manually cropped ROI dependency
Yuce et al. [31], 2025	Low	Low	Low	Unclear	Manually cropped ROI; possible augmentation-related data leakage (augmentation prior to train/test split); four brands only; no external validation
Benakatti et al. [34], 2025	Unclear	Low	Low	Unclear	Small imbalanced dataset with author-reported overfitting and limited validation-set generalization; no external validation
Lashaki et al. [32], 2025	Unclear	Low	Unclear	Unclear	Public dataset with metadata-derived labels and severe class imbalance; pairwise binary rather than multiclass classification; heavy augmentation; no external validation
Pachaiappan et al. [35], 2026	Low	Low	Low	Unclear	Single-institution dataset with marked class imbalance reducing minority-brand recall; system-level classification only; no external validation
Ahmed et al. [33], 2026	Unclear	Low	Low	Low	Merged two-source dataset with substantial class imbalance (OSS-TEM overrepresented) and exclusion of low-frequency brands; no external validation despite STARD/CLAIM-adherent reporting

fixture geometry in periapical radiographs. More recent periapical studies extended this range further: Lashaki et al. [32] reported up to 98.7% accuracy using a preprocessed, mask-focused VGG16 pipeline, and Ahmed et al. [33] reported a mean average precision at IoU 0.50 (mAP@50) of 0.989 using a YOLOv12x object-detection model.

Studies incorporating mixed panoramic–periapical datasets demonstrated robust and clinically relevant performance levels [27–29]. Interestingly, Park et al. [28] reported no statistically significant performance difference between panoramic and periapical modalities, suggesting that both imaging modalities may provide comparable diagnostic information under certain conditions. Two recent mixed-modality studies published in 2025–2026 similarly reported strong performance using object-detection architectures: Pachaiappan et al. [35] achieved a detection mAP of 98.3% with YOLOv10 across seven implant systems, while Benakatti et al. [34] reported

a precision of 0.83 and recall of 0.89 ($F1 = 0.82$) using a DETR framework, although the authors explicitly acknowledged a train–validation generalization gap consistent with overfitting.

Model architecture analysis

Convolutional neural network–based architectures represented the most commonly used deep learning frameworks across the included studies. Earlier studies primarily employed transfer learning–based CNN architectures such as VGG16, VGG19, ResNet, GoogLeNet, and DenseNet [23, 30].

Several studies incorporated more advanced learning strategies. Sukegawa et al. [15] introduced a multi-task learning framework for simultaneous implant brand and treatment-stage classification, while Sukegawa et al. [16] integrated an Attention Branch Network into ResNet architectures and demonstrated that attention mechanisms improved the performance of shallower networks such as ResNet18.

More recent studies explored transformer-based and hybrid architectures. Lubbad et al. [5] evaluated 28 different CNN and modern CNN and hybrid architectures and reported that ConvNeXt-small achieved the highest overall classification performance, with an accuracy of 94.2%.

Object detection–based workflows were increasingly adopted in recent studies. Kim et al. [24], Lin et al. [26], and Arijj et al. [18] employed YOLO-based detection pipelines capable of simultaneous implant localization and classification. Among these, Arijj et al. [18] proposed a clinically oriented two-step framework integrating YOLOv7-based implant detection with EfficientNet-based implant type classification. This trend toward object detection was further reinforced by the studies added during the updated search. Pachaiappan et al. [35] and Ahmed et al. [33] implemented YOLOv10 and YOLOv12x pipelines, respectively, for simultaneous implant localization and classification, and Ahmed et al. [33] additionally benchmarked successive YOLO generations (v7–v12) against a transformer-based DF-DETR baseline, reporting incremental gains in precision, recall, and mean average precision across versions. Benakatti et al. [34] introduced the first transformer-based detector (DETR) applied to implant identification, extending the transformer paradigm from classification (e.g., ConvNeXt and vision-transformer benchmarking) to end-to-end detection.

Beyond detection, two further studies addressed architectural breadth and computational efficiency. Yuce et al. [31] provided the first formal statistical comparison of four CNN architectures (GoogLeNet, ResNet-18, VGG16, and ShuffleNet) for implant classification using Z-tests with Bonferroni correction, reporting that VGG16 significantly outperformed the other networks; this study also represented the first reported application of the light-weight ShuffleNet architecture to implant identification, suggesting potential for deployment on resource-limited devices. Lashaki et al. [32] combined CNN-based feature extraction with classical machine-learning classifiers (support vector machine, decision tree, and random forest) and pre-trained networks (VGG16, ResNet50, and Xception), reporting that hybrid CNN–random-forest and preprocessed VGG16 pipelines achieved the highest performance among the configurations evaluated.

In addition, Kong et al. [25] utilized a cloud-based AutoML framework, demonstrating that automated neural architecture search systems may achieve competitive performance while simplifying model development workflows.

Sukegawa et al. [38] introduced a synthetic radiographic image generation framework using STL-based implant models and demonstrated that biologically realistic synthetic augmentation improved classification performance and feature-space organization.

Validation strategy analysis

A critical methodological finding across the included studies was the limited use of true external validation. Most studies relied primarily on internal validation methods, including hold-out validation and cross-validation approaches [15, 17, 23].

Although several studies incorporated multicenter datasets or independent test cohorts [18, 28, 29], geographically independent external validation datasets remained uncommon. Consequently, the generalizability of many proposed models remains uncertain.

Several studies also relied heavily on manually cropped implant regions of interest prior to classification, which may limit direct clinical applicability in fully automated workflows. None of the five studies added during the updated search incorporated geographically independent external validation; all relied on internal hold-out or cross-validation within single or merged institutional datasets. Notably, Ahmed et al. [33] explicitly identified the absence of an external validation cohort as a key limitation and reported that multisite evaluation was planned, underscoring that the lack of external validation remains a recognized and unresolved gap even in the most recent literature.

Performance analysis

Reported diagnostic performance and validation characteristics of the included studies are summarized in Table 3. Overall accuracy values ranged approximately between 75% and 99%, reflecting heterogeneity in dataset composition, implant diversity, imaging modality, validation methodology, and model architecture.

Studies involving relatively few implant systems and manually cropped implant regions frequently reported extremely high diagnostic performance, often exceeding 97% accuracy [15, 17]. In contrast, studies incorporating larger multicenter datasets and a greater number of implant systems generally reported lower but potentially more clinically realistic performance estimates [28].

Several studies additionally reported complementary performance metrics such as AUC and F1-score. Sukegawa et al. [16] reported an AUC value of 0.9993 using an attention-based ResNet18 model, whereas Lubbad et al. [5] demonstrated F1-scores exceeding 94% using ConvNeXt architectures. Park et al. [28] reported lower but clinically relevant F1-scores in highly heterogeneous multicenter datasets.

Class imbalance and implant diversity appeared to substantially influence model performance. Studies involving fewer implant classes generally demonstrated higher classification accuracy, whereas performance tended to decrease as implant diversity and dataset heterogeneity increased.

The most recent studies (2025–2026) reinforced the wide performance range and the heterogeneity of reporting metrics observed across the literature. Reported figures included 98.9% accuracy (Yuce et al. [31].), 98.7% accuracy with a ROC AUC of 0.988 (Lashaki et al. [32].), and 95.0% test accuracy (Ahmed et al. [33].) for classification-oriented studies, alongside detection-oriented metrics such as a mAP of 98.3% (Pachaiappan et al. [35].) and a mAP@50 of 0.989 with a mAP@50–95 of 0.900 (Ahmed et al. [33].). Importantly, this divergence illustrates a central obstacle to quantitative synthesis: classification studies typically reported accuracy and AUC, whereas object-detection studies reported mean average precision, and at least one study (Benakatti et al. [34].) reported only precision, recall, and F1-score without an accuracy value. These metrics are not directly comparable or statistically poolable, which—rather than methodological heterogeneity alone—was the principal reason a meta-analysis was not undertaken. Methodological design further complicated direct comparison; for example, Lashaki et al. [32]. evaluated performance using a pairwise binary classification scheme (e.g., Bicon vs. Bego), which is structurally simpler than multiclass classification and may yield higher apparent accuracy than studies discriminating among all implant systems simultaneously.

Advanced methodological findings

Several studies investigated advanced methodological strategies beyond conventional classification frameworks.

Explainable artificial intelligence approaches were incorporated in multiple studies. Sukegawa et al. [15] and Sukegawa et al. [16] utilized Grad-CAM–based visualization methods to identify implant regions contributing to classification decisions, whereas Lee et al. [39] incorporated activation mapping within a human–AI collaborative diagnostic workflow.

Table 3 Diagnostic performance and validation characteristics of included studies

Study	Best Model / Framework	Accuracy (%)	AUC	External or Independent Testing	Major Performance-Influencing Characteristics
Kim et al. [23], 2020	ResNet-50	98.0	NR	No	Small periapical dataset with manually cropped ROIs
Lee & Jeong [30], 2020	GoogLeNet Inception-v3	NR	0.979	No	Modality-specific performance differences and limited implant diversity
Lee et al. [27], 2020	Automated DCNN framework	NR	0.954	No	Multicenter mixed-modality dataset with clinician comparison
Sukegawa et al. [15], 2021	ResNet152 multi-task CNN	99.1	0.9999	No	Multi-task learning with manually cropped implant ROIs
Sukegawa et al. [16], 2022	ResNet18 with Attention Branch Network (ABN)	97.2	0.9993	No	Attention-guided shallow CNN architecture
Kim et al. [24], 2022	YOLOv3-based detection/classification framework	96.7	NR	No	Combined implant detection and classification workflow
Lee et al. [39], 2021	Automated DL algorithm (Neuro-T v2.0.1)	80.6	NR	Partial	Human–AI collaborative implant classification study
Kong et al. [25], 2023	Google Cloud AutoML Vision	98.1	NR	No	Cloud-based AutoML framework with limited algorithmic transparency
Park et al. [28], 2023	Neuro-T v3.0.1	88.5	NR	No	Large multicenter heterogeneous dataset with substantial implant diversity
Lee et al. [29], 2024	ResNet-50	95.1	NR	Partial	Robustness testing under distorted imaging conditions
Lubbad et al. [5], 2024	ConvNeXt-small	94.2	NR	No	Comparative benchmarking of CNN and transformer architectures
Ariji et al. [18], 2025	YOLOv7+EfficientNet-B0	98.9	1.000	Partial	Fully automated two-step detection and classification workflow
Lin et al. [26], 2025	Improved IB-YOLOv10	NR	NR	No	End-to-end object detection workflow using modern YOLO architecture
Sukegawa et al. [38], 2025	ResNet50 with synthetic augmentation	91.5	0.996	No	STL-based synthetic radiographic augmentation
Tiryaki et al. [17], 2025	Fusion CNN model	98.9	NR	No	Ensemble learning with implant angle estimation
Yuce et al. [31], 2025	VGG16	98.9	NR	No	Manually cropped ROIs with CLAHE; small augmented single-center dataset; ShuffleNet competitive at lower computational cost
Benakatti et al. [34], 2025	DEtection TRansformer (DETR)	NR	0.96 (PR-AUC)	No	Transformer detector (precision 0.83 / recall 0.89 / F1 0.82); author-acknowledged overfitting and limited generalization
Lashaki et al. [32], 2025	VGG16 (preprocessed)	98.7	0.988	No	Pairwise binary classification; preprocessing- and masking-driven gains on a public, imbalanced dataset
Pachaiappan et al. [35], 2026	YOLOv10m	NR	NR	No	Detection mAP 98.3% (precision 0.93 / recall 0.86 / F1 0.89); class imbalance lowered recall for minority brands (e.g., Nobel 72.7%)
Ahmed et al. [33], 2026	YOLOv12x	95.0	NR	No	Detection mAP@50 0.989 / mAP@50–95 0.900; OSSTEM overrepresented; outperformed clinicians across experience levels

NR not reported, AUC area under the curve

Human–AI interaction was specifically evaluated by Lee et al. [39], who demonstrated that AI-assisted interpretation significantly improved implant identification performance among dental professionals.

Lee et al. [29] evaluated model robustness under clinically challenging imaging conditions, including distorted angulations, overexposure, cutoff implant visualization, and foreign-body interference. Despite these distortions, the proposed deep learning system maintained relatively high classification performance.

Tiryaki et al. [17] extended deep learning applications beyond implant classification by incorporating regression-based implant angle estimation, thereby demonstrating potential utility for prosthetic treatment planning and angled abutment selection.

Finally, Arijji et al. [18] proposed a fully automated two-step workflow integrating implant detection and implant-type classification directly from panoramic radiographs, thereby improving the clinical applicability of deep learning-based implant identification systems.

Human–AI comparison was further examined by Ahmed et al. [33], who compared a YOLOv12x model against 52 clinicians spanning six experience levels across 5,200 image evaluations. The model significantly outperformed every clinician subgroup (one-way ANOVA with Tukey HSD, $p < 0.001$), and no significant differences emerged among clinicians by experience or specialization; inter-rater agreement was quantified using the Cohen kappa statistic. These findings are consistent with the earlier human–AI study by Lee et al. [39] and reinforce the potential role of AI as a decision-support tool when clinical records are unavailable.

Reporting practices also showed signs of maturation in the most recent studies. Ahmed et al. [33] reported their study in accordance with the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guideline and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM), representing one of the few included studies to adopt an AI-specific reporting framework. In addition, preprocessing strategy emerged as an explicit performance factor: Lashaki et al. [32] demonstrated that implant-specific masking combined with CLAHE-based contrast enhancement substantially improved classification performance relative to unprocessed inputs, and Yuce et al. [31] similarly applied CLAHE prior to classification, supporting the broader observation that preprocessing choices materially influence reported outcomes.

Discussion

This systematic review evaluated the diagnostic performance and methodological characteristics of deep learning (DL) models for dental implant brand classification across panoramic, periapical, and mixed radiographic datasets. The present review incorporated studies representing diverse imaging modalities, dataset complexities, and performance ranges to provide a balanced overview of the current literature.

Overall, the included studies demonstrated that deep learning models can achieve high diagnostic performance for implant brand classification, with reported accuracy values ranging from approximately 75% to 99%. However, one of the most important findings of the present review is that reported performance metrics were influenced by methodological characteristics rather than model architecture alone. In particular, dataset composition, number of implant classes, validation strategy, image preprocessing workflow, and degree of dataset heterogeneity substantially affected reported outcomes.

Studies using relatively small datasets with manually cropped implant regions and limited implant diversity frequently reported extremely high performance values exceeding 97% [15, 17]. In contrast, studies involving larger multicenter datasets with greater implant diversity demonstrated comparatively lower but likely more clinically realistic performance estimates. For example, Park et al. [28], which incorporated 27 implant systems and more than 156,000 radiographic images from multiple institutions, reported an overall accuracy of approximately 88.5%. These findings suggest that increasing classification complexity and real-world heterogeneity represent major challenges for implant classification systems. This observation is particularly important because many earlier implant AI studies primarily emphasized peak accuracy values without adequately considering dataset complexity or generalizability. The present review therefore expands upon previous reviews, such as Dashti et al. [11], by systematically evaluating methodological variability, imaging modality differences, validation strategies, and sources of potential bias rather than focusing solely on overall classification performance.

One of the major findings of this review is that high classification performance was achievable across both panoramic and periapical radiographic modalities. Periapical radiographs theoretically provide superior spatial resolution and improved visualization of implant-specific features such as thread geometry, tapering, and fixture

morphology, which may facilitate implant identification. Several periapical-based studies demonstrated strong diagnostic performance, including Kim et al. [23], Kong et al. [25], and Lin et al. [26]. However, panoramic radiographs also consistently demonstrated high performance levels across multiple studies [15, 17]. Importantly, Park et al. [28] reported no statistically significant difference between panoramic and periapical imaging modalities in multicenter datasets, suggesting that deep learning models may successfully extract diagnostically relevant implant features from both imaging sources under appropriate training conditions.

Nevertheless, studies incorporating mixed and heterogeneous datasets generally reported lower but potentially more clinically representative performance values. This was particularly evident in multicenter studies and robustness analyses, where imaging variability, projection distortion, exposure differences, and implant diversity more closely reflected real-world clinical scenarios. Lee et al. [29], for example, specifically evaluated model robustness under distorted imaging conditions, including overexposure, non-perpendicular angulation, cutoff implant visualization, and foreign-body interference. Despite these challenges, the proposed ResNet50-based model maintained relatively high diagnostic performance, suggesting that modern deep learning systems may exhibit meaningful robustness beyond idealized datasets.

Another important finding of this review concerns the evolution of model architectures over time. Earlier studies primarily relied on transfer learning-based CNN architectures such as VGG16, VGG19, and ResNet models [23]. These approaches generally demonstrated excellent performance, particularly when implant regions were manually cropped prior to classification. However, recent studies increasingly incorporated more sophisticated architectures, including attention-based learning frameworks, transformer-based models, AutoML systems, and object-detection pipelines.

Attention mechanisms appeared particularly promising in implant classification tasks. Sukegawa et al. [16] demonstrated that integration of an Attention Branch Network improved the performance of shallower ResNet architectures and enabled visualization of diagnostically relevant implant regions through Grad-CAM analysis. Similarly, transformer-based and hybrid architectures showed encouraging performance in more recent studies. Lubbad et al. [5], who evaluated 28 different CNN and modern CNN and hybrid architectures, reported that ConvNeXt-small achieved the highest overall classification performance. Their findings suggest that modern convolutional architectures may outperform both traditional CNNs and pure vision transformer models in implant classification tasks.

Recent studies also demonstrated a clear transition from manual ROI-dependent classification systems toward more automated and clinically applicable workflows. Earlier studies typically relied on manually cropped implant regions prior to classification, limiting direct clinical integration. In contrast, Lin et al. [26], Kim et al. [24], and Arijj et al. [18] incorporated YOLO-based object detection frameworks capable of simultaneous implant localization and classification. Among these, Arijj et al. [18] proposed a clinically oriented two-step workflow integrating YOLOv7-based implant detection with EfficientNet-based implant type classification directly from panoramic radiographs. Unlike earlier classification-only pipelines, this approach more closely reflects real-world clinical workflows and may represent an important step toward fully automated implant identification systems. This shift toward automated detection has accelerated in recent studies, particularly those published in 2025–2026, which were dominated by object-detection architectures such as DETR, YOLOv10, and YOLOv12x —indicating that end-to-end detection, rather than manually cropped classification, is increasingly the default design for new implant-identification systems. Ahmed et al. [33] further demonstrated, through cross-version benchmarking of YOLOv7–v12 against a transformer-based DF-DETR baseline, that successive architectural generations yielded measurable gains in localization and accuracy, while Benakatti et al. [34] extended the transformer paradigm to detection itself. Collectively, these studies suggest that the field is converging on clinically deployable, fully automated pipelines, although none yet pairs this automation with geographically independent external validation.

Beyond conventional classification tasks, several studies explored additional clinically relevant applications of artificial intelligence. Tiryaki et al. [17], for example, extended deep learning applications beyond implant brand classification by incorporating regression-based implant angle estimation. The authors reported a root mean square error of approximately 2.9°, suggesting potential utility for prosthetic planning and angled abutment

selection. Similarly, Lee et al. [39] demonstrated that AI-assisted interpretation significantly improved implant identification performance among dental professionals, supporting the potential role of human–AI collaboration in clinical decision support systems.

Recent studies also highlighted a key methodological barrier to evidence synthesis in this field: the inconsistency of reported performance metrics. Classification-oriented studies reported accuracy and AUC, object-detection studies reported mean average precision at varying IoU thresholds, and some studies reported only precision, recall, and F1-score [34]. Because these metrics capture different aspects of model behaviour and are computed under different task definitions, they cannot be meaningfully pooled. This observation directly supports the decision to perform a structured narrative synthesis rather than a meta-analysis, and it argues for the adoption of standardized, fully reported metric sets in future implant-AI studies.

One of the most innovative methodological developments identified in this review was the use of synthetic radiographic augmentation. Sukegawa et al. [38] introduced a novel framework using STL-based implant models to generate synthetic panoramic radiographic images. Importantly, the inclusion of biologically realistic background-processed synthetic images improved classification performance and feature-space organization compared with conventional augmentation alone. These findings suggest that anatomically realistic synthetic augmentation may improve generalizability and partially address limitations associated with small datasets and class imbalance.

Despite these promising developments, several important methodological limitations remain prevalent throughout the literature. The most significant limitation identified in the present review was the limited use of true external validation. Most studies relied primarily on internal hold-out or cross-validation approaches, which may overestimate model performance and fail to adequately assess generalizability. Although several studies incorporated multicenter datasets or independent testing cohorts [18, 28, 29], geographically independent external validation remained uncommon. Consequently, the real-world clinical performance of many proposed systems remains uncertain.

Another important limitation was the frequent reliance on manually cropped implant regions of interest. Although this preprocessing strategy simplifies classification tasks and often improves performance, it reduces clinical realism because implant localization must first be performed manually. This issue partially explains why earlier CNN-based studies often reported near-perfect classification accuracy. More recent object-detection frameworks may therefore provide more clinically meaningful performance estimates, even when overall accuracy values are comparatively lower [18, 24, 26].

Another methodological consideration identified in the present review was the partial overlap of datasets across several studies. For example, Lee et al. [29] utilized a training dataset derived largely from the same multicenter AI-Hub database previously reported by Park et al. [28], although an independent distorted-image testing cohort was additionally incorporated for robustness evaluation. Similarly, the sequential studies published by Sukegawa and colleagues likely involved overlapping institutional radiographic archives and implant pools while focusing on different methodological objectives, including multi-task learning, attention mechanisms, and synthetic augmentation strategies. Although such overlap may reflect progressive methodological refinement rather than direct duplication, it may also reduce the effective independence of the available evidence and should therefore be considered when interpreting aggregated findings across studies.

Dataset imbalance also emerged as a recurrent methodological concern. Several studies included substantially unequal distributions of implant systems, potentially biasing classification performance toward overrepresented implant brands. Tiriyaki et al. [17], for example, demonstrated lower performance in implant systems with smaller sample sizes, whereas Park et al. [28] similarly reported reduced performance for underrepresented implant classes. These findings indicate that dataset balance and implant diversity are critical determinants of model robustness. The same pattern was also observed in more recent investigations. Pachaiappan et al. [35] reported markedly reduced recall for the underrepresented Nobel class (72.7%) relative to the overrepresented Osstem class, and Ahmed et al. [33] similarly noted that performance reflected an OSSTEM-dominated class distribution, with two low-frequency brands excluded entirely from analysis. Benakatti et al. [34] attributed an explicit

train–validation generalization gap in part to limited and unevenly distributed training data. Together with the earlier findings of Tiryaki et al. [17]. and Park et al. [28]., these results confirm that class imbalance is a persistent and consequential determinant of model performance across architectures and imaging modalities.

Two methodological features of the newly included studies further illustrate how study design shapes reported performance. First, Lashaki et al. [32] evaluated their models using a pairwise binary classification scheme, which simplifies the discrimination task relative to simultaneous multiclass classification and may inflate apparent accuracy; this represents a concrete example of how task formulation— independent of architecture— affects headline performance figures. Second, the same study relied on a publicly available radiographic dataset, in contrast to the institutional archives used by most other included studies. While public datasets improve reproducibility and benchmarking, dataset-specific labelling conventions and image characteristics may limit the transferability of reported performance to routine clinical radiographs. These considerations reinforce the central observation of this review that dataset composition, preprocessing, and task design influence reported performance at least as strongly as model architecture.

The present review additionally identified important sources of potential bias within the existing literature. Publication bias likely favors studies reporting exceptionally high performance values, whereas studies with lower or moderate accuracy may remain underreported. Selection bias may also occur when investigators selectively exclude challenging implant systems or low-quality radiographs. To reduce this bias, the present review intentionally avoided performance-based inclusion thresholds and incorporated studies with broader methodological variability and wider performance ranges.

Reporting bias also remains problematic. Many studies failed to report comprehensive diagnostic metrics such as F1-score, sensitivity, specificity, confidence intervals, calibration performance, or class-wise error analysis. Most studies primarily emphasized discrimination metrics such as accuracy and AUC, whereas calibration analysis and predictive uncertainty estimation were rarely reported. Similarly, explainable artificial intelligence techniques remained underutilized across the literature. Although several studies incorporated Grad-CAM or feature-space visualization approaches [15, 16], most studies still functioned largely as “black-box” systems. Limited interpretability may hinder clinician trust and slow clinical adoption of AI-assisted implant identification systems. Encouragingly, the most recent studies showed early signs of improved reporting discipline. Ahmed et al. [33] reported their work in accordance with both STARD and the CLAIM framework—one of the few included studies to apply an AI-specific reporting standard addressing external validation, transparency, and reproducibility. Wider adoption of such frameworks, alongside routine reporting of calibration, confidence intervals, and class-wise performance, would substantially strengthen the comparability and clinical credibility of future implant-AI research.

From a clinical perspective, automated implant brand classification systems may provide substantial practical benefits. Such systems could assist clinicians when implant records are unavailable, support prosthetic maintenance and component compatibility assessment, facilitate revision procedures, and improve treatment planning efficiency. In particular, object-detection–based workflows and AI-assisted decision support systems may eventually enable real-time chairside implant identification. However, widespread clinical implementation will require robust external validation, standardized evaluation protocols, improved interpretability, and prospective clinical testing.

Several limitations of the present systematic review should also be acknowledged. First, substantial methodological heterogeneity precluded quantitative meta-analysis. Second, inconsistencies in reporting standards limited direct comparison across studies. Third, many included studies were retrospective and single-center in design, potentially limiting external generalizability. Fourth, although the review protocol was not prospectively registered, it was retrospectively deposited in a public repository (OSF; <https://doi.org/10.17605/OSF.IO/984AP>) after completion of the review to enhance transparency. Finally, despite attempts to minimize selection bias, the possibility of publication bias cannot be completely excluded.

Future research should prioritize large-scale multicenter external validation, inclusion of broader implant system diversity, standardized AI reporting guidelines, and prospective clinical validation studies. In addition,

integration of explainable AI methods, calibration analysis, federated learning approaches, and fully automated end-to-end workflows may further improve the reliability and clinical applicability of deep learning systems for dental implant identification.

Conclusion

This systematic review indicates that, in answer to the guiding question, deep learning models may achieve high diagnostic performance for dental implant brand classification on panoramic, periapical, and mixed radiographs, but that this performance is reported almost exclusively under internally validated, single-center conditions. Although reported accuracy was generally high, it varied substantially across studies and appeared to be associated more with dataset composition, implant diversity, preprocessing workflow, and validation strategy than with model architecture alone. Because these observations derive from a narrative synthesis of heterogeneous, mostly retrospective studies with limited external validation, they should be regarded as hypothesis-generating rather than as definitive evidence of clinical readiness.

The findings further suggest that implant AI research appears to have progressively evolved from conventional CNN-based classification systems toward more clinically applicable frameworks incorporating attention mechanisms, synthetic augmentation, and automated detection-classification workflows. Nevertheless, the current literature remains limited by important methodological concerns, including substantial dataset heterogeneity, class imbalance, inconsistent reporting standards, and a persistent lack of geographically independent external validation.

Future research should prioritize multicenter prospective validation, broader implant system diversity, standardized reporting frameworks, and development of fully automated and interpretable AI workflows. Greater emphasis on explainable artificial intelligence, calibration analysis, and real-world clinical evaluation will be essential for reliable clinical integration of deep learning systems for dental implant identification.

Abbreviations

DL	Deep learning
AI	Artificial intelligence
CNN	Convolutional neural networks
YOLO	You only look once
AUC	Area under the curve
Grad-CAM	Gradient-weighted class activation mapping
SWiM	Synthesis without meta-analysis
ABN	Attention branch network
DETR	Detection TRansformer
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
STARD	The Standards for Reporting of Diagnostic Accuracy
CLAHE	Contrast limited adaptive histogram equalization
QUADAS	Quality assessment of diagnostic accuracy studies
mAP	Mean Average Precision

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1186/s12903-026-09926-8>.

Acknowledgements The authors thank Assoc.Prof. Serkan Yilmaz, (Nigde Ömer Halisdemir University, Faculty of Dentistry, Department of Oral and Maxillofacial Radiology), for serving as an independent adjudicator and resolving the small number of disagreements that arose during study selection. Dr. Yilmaz had no other role in the study and has approved being

named in this acknowledgement.

Authors' contributions O.M.D.Y. Conceptualization, methodology, data curation, investigation, resource, validation formal analysis and writing - original draft, writing - review & editing. I.L.K. Methodology, data curation, validation, investigation, writing - original draft, writing - review & editing.

Funding The authors did not receive any funding from any organizations.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. French D, Ofec R, Levin L. Long term clinical performance of 10 871 dental implants with up to 22 years of follow-up: A cohort study in 4247 patients. *Clin Implant Dent Relat Res*. 2021;23(3):289–97. <https://doi.org/10.1111/cid.12994>.
2. Kurtulus IL, Lubbad M, Yilmaz OMD, Kilic K, Karaboga D, Basturk A, Akay B, Nalbantoglu U, Yilmaz S, Ayata M. A robust deep learning model for the classification of dental implant brands. *J Stomatol Oral Maxillofac Surg*. 2024;125(5):101818. <https://doi.org/10.1016/j.jormas.2024.101818>.
3. Sharma D, Jhingta P, Bathla SC. Challenges in dental implant identification and need of universal dental implant identification, numbering, and nomenclature system. *Int J Prosthodont Restor Dent*. 2020;10(4):146–50. <https://doi.org/10.5005/jp-journals-10019-1291>.
4. Takahashi T, Nozaki K, Gonda T, Mameno T, Wada M, Ikebe K. Identification of dental implants using deep learning—pilot study. *Int J Implant Dent*. 2020;6(1):53. <https://doi.org/10.1186/s40729-020-00250-6>.
5. Lubbad MA, Kurtulus IL, Karaboga D, Kilic K, Basturk A, Akay B, Nalbantoglu OU, Yilmaz OMD, Ayata M, Yilmaz S. A comparative analysis of deep learning-based approaches for classifying dental implants decision support system. *J Imaging Inf Med*. 2024;37(5):2559–80. <https://doi.org/10.1007/s10278-024-01086-x>.
6. Majid IA, ur Rahaman SM, Kumar H, Sowbhagya MB, Alikutty FK. Radiographic prescription trends in dental implant site. *J Dent Implant*. 2014;4(2):140–3. <https://doi.org/10.4103/0974-6781.140874>.
7. Sahrman P, Kühl S, Dagassan-Berndt D, Bornstein MM, Zitzmann NU. Radiographic assessment of the peri-implant site. *Periodontol 2000*. 2024;95(1):70–86. <https://doi.org/10.1111/prd.12577>.
8. Katsumata A. Deep learning and artificial intelligence in dental diagnostic imaging. *Jpn Dent Sci Rev*. 2023;59:329–33. <https://doi.org/10.1016/j.jdsr.2023.09.004>.
9. Yilmaz S, Tasyurek M, Amuk M, Celik M, Canger EM. Developing deep learning methods for classification of teeth in dental panoramic radiography. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2024;138(1):118–27. <https://doi.org/10.1016/j.oooo.2023.02.021>.
10. Mohammad-Rahimi H, Rokhshad R, Bencharit S, Krois J, Schwendicke F. Deep learning: a primer for dentists and dental researchers. *J Dent*. 2023;130:104430. <https://doi.org/10.1016/j.jdent.2023.104430>.
11. Dashti M, Londono J, Ghasemi S, Zare N, Samman M, Ashi H, Amirzade-Iranaq MH, Khosraviani F, Sabeti M, Khurshid Z. Comparative analysis of deep learning algorithms for dental caries detection and prediction from radiographic images: A comprehensive umbrella review. *PeerJ Comput Sci*. 2024;10:e2371. <https://doi.org/10.7717/peerj-cs.2371>.

12. Jeon S-J, Yun J-P, Yeom H-G, Shin W-S, Lee J-H, Jeong S-H, Seo M-S. Deep-learning for predicting C-shaped canals in mandibular second molars on panoramic radiographs. *Dentomaxillofac Radiol.* 2021;50(5):20200513. <https://doi.org/10.1259/dmfr.20200513>.
13. Lian L, Zhu T, Zhu F, Zhu H. Deep learning for caries detection and classification. *Diagnostics.* 2021;11(9):1672. <https://doi.org/10.3390/diagnostics11091672>.
14. Sukegawa S, Yoshii K, Hara T, Yamashita K, Nakano K, Yamamoto N, Nagatsuka H, Furuki Y. Deep neural networks for dental implant system classification. *Biomolecules.* 2020;10(7):984. <https://doi.org/10.3390/biom10070984>.
15. Sukegawa S, Yoshii K, Hara T, Matsuyama T, Yamashita K, Nakano K, Takabatake K, Kawai H, Nagatsuka H, Furuki Y. Multi-task deep learning model for classification of dental implant brand and treatment stage using dental panoramic radiograph images. *Biomolecules.* 2021;11(6):815. <https://doi.org/10.3390/biom11060815>.
16. Sukegawa S, Yoshii K, Hara T, Tanaka F, Yamashita K, Kagaya T, Nakano K, Takabatake K, Kawai H, Nagatsuka H. Is attention branch network effective in classifying dental implants from panoramic radiograph images by deep learning? *PLoS ONE.* 2022;17(7):e0269016. <https://doi.org/10.1371/journal.pone.0269016>.
17. Tiryaki B, Ozdogan A, Guller MT, Miloglu O, Oral EA, Ozbek IY. Dental implant brand and angle identification using deep neural networks. *J Prosthet Dent.* 2025;133(6):1528–34. <https://doi.org/10.1016/j.prosdent.2023.07.022>.
18. Arijji Y, Kusano K, Fukuda M, Wakata Y, Nozawa M, Kotaki S, Arijji E, Baba S. Two-step deep learning models for detection and identification of the manufacturers and types of dental implants on panoramic radiographs. *Odontology.* 2025;113(2):788–98. <https://doi.org/10.1007/s10266-024-00989-z>.
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372. <https://doi.org/10.1136/bmj.n71>.
20. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, Hartmann-Boyce J, Ryan R, Shepperd S, Thomas J. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ.* 2020;368. <https://doi.org/10.1136/bmj.l6890>.
21. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeftang MM, Sterne JA, Bossuyt PM. Group* Q-: QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529–36. <https://doi.org/10.7326/0003-4819-155-8-201110180-00009>.
22. Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, Kahn CE Jr, Esteva A, Karthikesalingam A, Mateen B. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med.* 2021;27(10):1663–5. <https://doi.org/10.1038/s41591-021-01517-0>.
23. Kim J-E, Nam N-E, Shim J-S, Jung Y-H, Cho B-H, Hwang JJ. Transfer learning via deep neural networks for implant fixture system classification using periapical radiographs. *J Clin Med.* 2020;9(4):1117. <https://doi.org/10.3390/jcm9041117>.
24. Kim H-S, Ha E-G, Kim YH, Jeon KJ, Lee C, Han S-S. Transfer learning in a deep convolutional neural network for implant fixture classification: A pilot study. *Imaging Sci Dent.* 2022;52(2):219. <https://doi.org/10.5624/isd.20210287>.
25. Kong HJ. Classification of dental implant systems using cloud-based deep learning algorithm: an experimental study. *J Yeungnam Med Sci.* 2023;40(Suppl):S29–36. <https://doi.org/10.12701/jyms.2023.00465>.
26. Lin Y-J, Chen S-L, Lu Y-C, Lin X-M, Mao Y-C, Chen M-Y, Yang C-S, Chen T-Y, Li K-C, Tu W-C. Deep learning-assisted diagnostic system: Implant brand detection using improved IB-YOLOv10 in periapical radiographs. *Diagnostics.* 2025;15(10):1194. <https://doi.org/10.3390/diagnostics15101194>.
27. Lee J-H, Kim Y-T, Lee J-B, Jeong S-N. A performance comparison between automated deep learning and dental professionals in classification of dental implant systems from dental imaging: a multi-center study. *Diagnostics.* 2020;10(11):910. <https://doi.org/10.3390/diagnostics10110910>.
28. Park W, Huh J-K, Lee J-H. Automated deep learning for classification of dental implant radiographs using a large multi-center dataset. *Sci Rep.* 2023;13(1):4862. <https://doi.org/10.1038/s41598-023-32118-1>.
29. Lee J-H, Kim Y-T, Lee J-B. Identification of dental implant systems from low-quality and distorted dental radiographs using AI trained on a large multi-center dataset. *Sci Rep.* 2024;14(1):12606. <https://doi.org/10.1038/s41598-024-63422-z>.
30. Lee J-H, Jeong S-N. 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.* 2020;99(26):e20787. <https://doi.org/10.1097/MD.00000000000020787>.
31. Yuce H, Acikgoz G, Al-Jumaili S, Ucan ON, Ozlu Ucan G, Dere KA. Automatic identification of dental implant brands with deep learning algorithms. *Dentomaxillofac Radiol.* 2025;54(7):588–96. <https://doi.org/10.1093/dmfr/twaf054>.
32. Lashaki RA, Raiesi Z, Razavi N, Goodarzi M, Najafzadeh H. Optimized classification of dental implants using convolutional neural networks and pre-trained models with preprocessed data. *BMC Oral Health.* 2025;25(1):535. <https://doi.org/10.1186/s12903-025-05704-0>.
33. Ahmed WM, Azhari AA, Almufti A, Alsadah ZM, Ahmed AFA, Lahi A, Fawaz KA. Development and evaluation of an AI model for dental implant type detection: A comparison of diagnostic accuracy between a deep learning model and dental professionals. *J Prosthodont.* 2026;35(1):81–93. <https://doi.org/10.1111/jopr.70064>.

34. Benakatti V, Nayakar RP, Anandhalli M. Advanced deep learning techniques for recognition of dental implants. *J Oral Biol Craniofac Res.* 2025;15(2):215–20. <https://doi.org/10.1016/j.jobcr.2025.01.016>.
35. Pachaiappan B, Shafiullah RS, Balaraman G, Bhuvaneshwar YP, Sadasivan S. Deep learning-based object detection of dental implant systems in panoramic and periapical radiographs. *J Prosthet Dent.* 2026. <https://doi.org/10.1016/j.prosdent.2026.01.029>.
36. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol.* 2016;69:245–7. <https://doi.org/10.1016/j.jclinepi.2015.04.005>.
37. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol Artif Intell.* 2022;4(3):e210064. <https://doi.org/10.1148/ryai.210064>.
38. Sukegawa S, Yoshii K, Hara T, Tanaka F, Taki Y, Inoue Y, Yamashita K, Nakai F, Nakai Y, Miyazaki R. Optimizing dental implant identification using deep learning leveraging artificial data. *Sci Rep.* 2025;15(1):3724. <https://doi.org/10.1038/s41598-025-87579-3>.
39. Lee J-H, Kim Y-T, Lee J-B, Jeong S-N. Deep learning improves implant classification by dental professionals: a multi-center evaluation of accuracy and efficiency. *J Periodontal Implant Sci.* 2021;52(3):220. <https://doi.org/10.5051/jpis.2104080204>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Ozden Melis Durmaz Yilmaz¹  · Ikbal Leblebicioglu Kurtulus²

✉ Ozden Melis Durmaz Yilmaz
ozdenmelisdurmaz47@gmail.com

Ikbal Leblebicioglu Kurtulus
ikbal_leblebicioglu@hotmail.com

¹ Department of Prosthodontics, Faculty of Dentistry, Nigde Omer Halisdemir University, Nigde 51240, Türkiye

² Faculty of Dentistry, Department of Prosthodontics, Erciyes University, Kayseri 38039, Türkiye