Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FROM DIAGNOSTIC AUTOMATION TO TRUSTWORTHY CLINICAL INTELLIGENCE: A SYSTEMATIC SCOPING REVIEW AND EVIDENCE-TO-DEPLOYMENT ROADMAP FOR ARTIFICIAL INTELLIGENCE IN DIGITAL DENTISTRY

[1] GUMMA PARVATHI DEVI, [2] DR. PRATHIPATI RATNA KUMAR,

Research scholar, Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India.

Assistant Professor, Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India

Corresponding email : parvathi3.gumma@gmail.com

ABSTRACT

Digital dentistry increasingly reports high-performing computational models, yet a central knowledge gap remains: existing reviews often catalogue technologies or deployment concerns without a common framework that distinguishes analytical performance from transportability, clinical validity, clinical utility, and lifecycle safety. Consequently, it remains unclear what level of real-world use is actually justified by a reported benchmark result. This systematic scoping review critically appraised literature available through July 2026; 147 records were screened and 56 studies were included for evidence synthesis. Rather than pooling incomparable accuracy values, the review evaluated data provenance, partitioning and leakage, reference standards, external validation, calibration and uncertainty, fairness, human factors, cybersecurity, prospective impact, and lifecycle monitoring across imaging, natural-language processing, multimodal and foundation models, robotics, AR/VR, tele-operation, synthetic data, and digital twins. The synthesis identifies an accuracy-to-deployment gap: evidence is comparatively mature for selected two-dimensional imaging and segmentation tasks, whereas multimodal assistants, generative systems, robotics, tele-operation, and digital twins remain limited by external, prospective, human-factor, or lifecycle evidence. The novelty of this review is an integrated evidence-to-deployment taxonomy, an M0-M5 clinical utility ladder, minimum reporting requirements, and staged evidence gates. The new knowledge created is that evidence maturity, not nominal accuracy alone, determines the defensible level of clinical use, and that the validation burden increases as autonomy, multimodality, generative behaviour, and post-deployment change increase. These outputs provide a practical basis for designing, reviewing, and governing clinically credible digital-dentistry research.

Keywords—*Artificial Intelligence; Digital Dentistry; Clinical Translation; Foundation Models; Multimodal Learning; Explainable AI; Calibration; External Validation; Robotics; Lifecycle Governance.*

HIGHLIGHTS

- Introduces a five-layer evidence-to-deployment framework linking technical performance to clinical utility and lifecycle safety.
- Separates analytical validity, clinical validity, workflow impact, economic value, and post-market governance.
- Critically evaluates leakage, calibration, uncertainty, external validation, explainability, fairness, human factors, and cybersecurity.
- Adds a maturity map for conventional AI, foundation models, multimodal assistants, robotics, AR/VR, and digital twins.
- Provides a minimum reporting checklist and staged research agenda suitable for clinically oriented dental-AI studies.

1. INTRODUCTION

Modern digital dentistry generates heterogenous data at nearly all points of care: intraoral and panoramic X-rays, CBCT, 3D scanning of surfaces, photos, periodontal readings, operative notes, prescription records, lab records, and telemetry from devices. Artificial intelligence can turn all this data into detection, segmentation, risk stratification, treatment planning, simulation, and workflow aids. But the key scientific challenge is no longer whether the model gets a high score on some carefully curated dataset. Instead, the

important question becomes whether the model maintains accuracy, calibration, fairness, interpretability, security, and utility in its interactions with new patients, new devices, new institutions, new users, and new clinical environments. [21]–[50]

Previous literature in the field is comprised mostly of retrospective models' development and general narrative review articles. In such cases, algorithms that have highly variable levels of evidence maturity tend to be put together at the same level. An algorithm that was tested using a hold-out sample set from a single hospital, an algorithm tested using a phantom, and an algorithm that was tested using a prospective clinical trial are not equivalent forms of evidence.

A critical synthesis is needed now because digital dentistry has expanded from relatively bounded image-classification tasks to foundation models, multimodal assistants, robotics, tele-operation, and continuously updated systems, while the maturity of supporting evidence remains highly uneven [55]-[61]. Without a common appraisal lens, an internally validated single-site result can be interpreted as if it carried the same evidential weight as external validation, prospective workflow testing, or lifecycle monitoring. That mismatch can distort research priorities, inflate clinical expectations, and encourage premature claims of readiness. The present study is therefore justified by the need to determine not merely whether a technology performs well, but what evidence supports a particular level of use, what evidence is still missing, and which next validation step would most meaningfully reduce uncertainty.

The current review therefore takes a critical systematic scoping approach. The review keeps in mind the need for comprehensive scope for an evolving multidisciplinary subject while analyzing evidence through four translational questions: What clinical problem does the research address? How representative and controlled are the data? How trustworthy is the validation? What more needs to be done for application in practice? The review structure follows an innovative five-tier taxonomy based on: data, intelligence, assurance, clinical, and governance.

The key aims are to (i) map significant AI and new digital technologies within dentistry; (ii) compare datasets, models, evaluation criteria, and validation techniques; (iii) differentiate between proof-of-concept results and clinical application; (iv) highlight methodological and publication flaws that result in non-reproducibility or exaggerated claims; and (v) outline a tangible agenda for dental AI research and governance. Key aspects of this systematic review are summarized in Table 1, and an evidence-to-deployment framework is shown in Figure 1.

Table 1. Distinctive Contributions of the Review

No.	Distinct contribution of this review
1	Introduces a five-layer evidence-to-deployment taxonomy rather than a technology-only catalogue.
2	Separates analytical performance from clinical validity, utility, and lifecycle safety.
3	Applies a deployment-readiness lens across NLP, imaging, robotics, AR/VR, synthetic data, and multimodal systems.
4	Provides minimum reporting requirements and a staged research agenda for clinically credible dental AI.
5	Identifies an accuracy-to-deployment gap and converts it into explicit evidence-maturity levels and next-step validation gates.

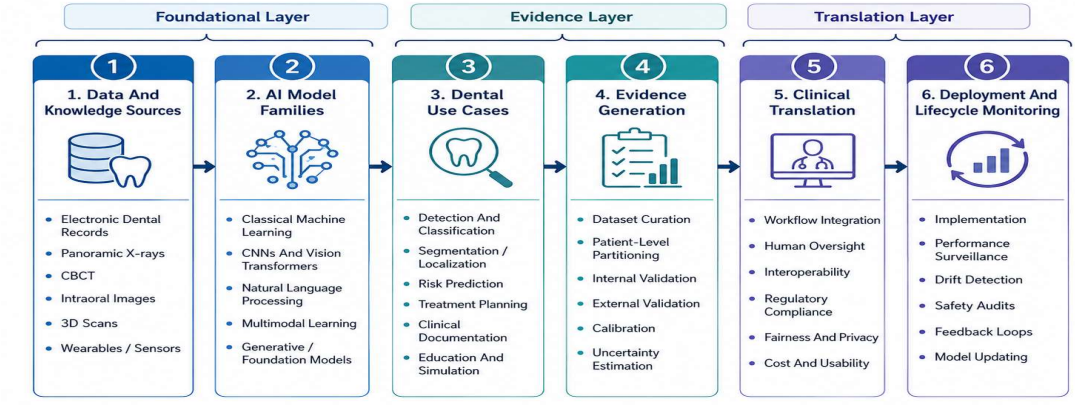

Figure 1. Evidence-to-deployment taxonomy proposed for artificial intelligence in dentistry.

2. REVIEW DESIGN AND REPORTING METHOD

This review adopts a critical systematic scoping design because the evidence spans heterogeneous diagnostic, prognostic, educational, robotic, and workflow studies that cannot be reduced responsibly to one pooled effect. Reporting follows PRISMA 2020 and PRISMA-ScR through explicit eligibility criteria, reproducible search concepts, sequential screening, systematic abstraction, and structured synthesis. Interpretation is conducted at both the clinical-task and study levels, whereas validation is judged across the complete data-to-decision pathway. This distinction prevents strong technical performance from being interpreted as clinical validity when the dataset, comparator, partitioning strategy, workflow, or lifecycle evidence remains inadequate.

Eligible articles included research that evaluated AI-based or computationally intelligent dental applications and provided enough details regarding data, methodology, evaluation, or implementation for a critical appraisal of the study. Opinion papers, non-dental applications, articles without technical or clinical evaluations, abstracts without full text and repeated experiments were excluded from the analysis. The process of screening resulted in 147 references, with 82 unique references being selected for the full-text analysis and 56 included studies in total. Eligibility criteria are summarised in Table 2, and the process of screening is illustrated in Figure 2.

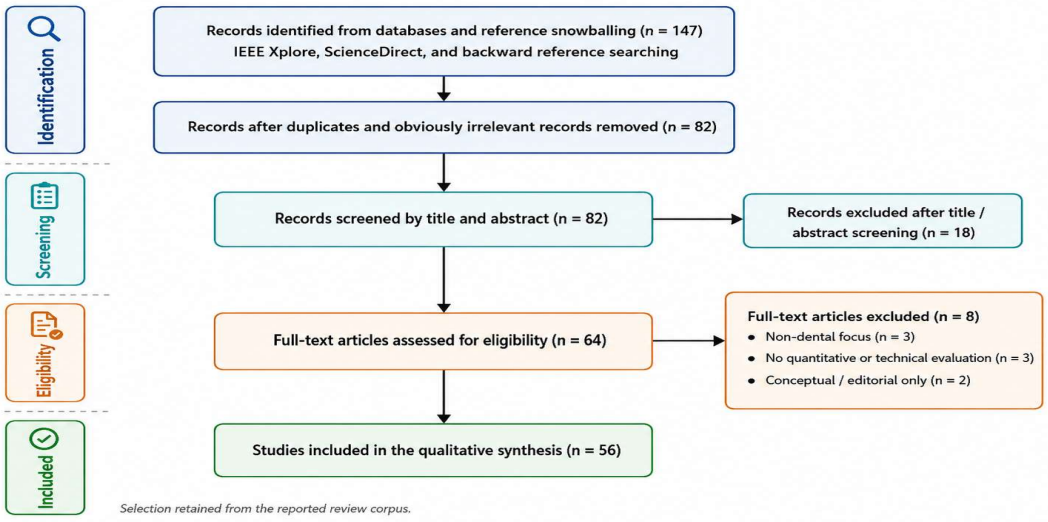

Figure 2. PRISMA-guided selection process retained from the reported review corpus.

Table 2. Eligibility Criteria and Operational Definitions

Criterion	Operational definition used for critical synthesis
Population/context	Dental patients, dental images or records, dental professionals, students, simulators, phantoms, or dental devices.
Intervention/technology	AI/ML, deep learning, transformers, NLP, robotics, AR/VR, synthetic data, tele-operation, or related intelligent systems.
Outcomes	Diagnostic/segmentation performance, calibration, efficiency, usability, safety, learning outcomes, robotic precision, or deployment evidence.
Minimum technical reporting	Data source, sample size or modality, model/technology description, and quantitative or structured qualitative evaluation.
Exclusions	Non-dental work, purely conceptual commentary, inaccessible full text, duplicate experiment, or no evaluable method/outcome.

2.1 Data Extraction and Critical Appraisal

The framework of analysis for each paper includes the clinical task, data modality, source of samples, data partition unit, modelling family, control group, measures of evaluation, internal or external validation design, approach to explainability, environment of implementation, and limitations of the research. One of the most critical differences for assessment is whether the unit of splitting coincides with the clinical unit. Image-level randomisation may lead to leakage of information in case the same patients' images are present in both train and test datasets.

Six different areas of potential risk of bias are evaluated, including participant and data selection, reference standard and annotation quality, model/index design, prevention of leakage and overfitting, statistical analysis, and generalisability to practical workflows. The articles are evaluated as low, medium, or high-risk studies rather than being given an arbitrary score. Studies of high risk include those where there is a small or convenience sample, no information about the class balance, test data used in model selection, only accuracy reported in imbalanced problems, or no external validation in the face of claims of clinical equivalence.

Table 3. Risk-of-Bias and Applicability Appraisal Framework

Appraisal domain	Lower concern	Higher concern
Data selection	Consecutive or clearly sampled data; multicenter diversity	Convenience sample; unclear exclusions; narrow demographics
Reference standard	Multiple trained annotators; adjudication; reliability reported	Single unblinded annotator; unclear diagnostic criteria
Partitioning	Patient-level, temporal, or external split	Image-level random split; duplicate or correlated samples
Model development	Prespecified pipeline; nested tuning; ablation	Test-set tuning; incomplete preprocessing description
Statistics	Confidence intervals, calibration, sensitivity analysis	Single point estimates; accuracy alone; no uncertainty
Applicability	Workflow-relevant prospective or silent deployment	Laboratory-only evaluation with broad clinical claims

3. LANDSCAPE OF DENTAL AI DATA AND TASKS

There are fundamental limits imposed by the data quality of Dental AI applications. Panoramic and bitewing radiographs are prevalent due to their widespread availability, standardisation, and compatibility with well-developed algorithms based on computer vision in two dimensions. The CBCT allows for localisation in three dimensions and surgical planning, but at the cost of greater annotation effort, variability in scanners, reconstruction techniques, and computational power. Intraoral photographs are used to diagnose caries, plaque, gingival conditions, and lesions but are affected by illumination, distance from the camera, presence of saliva, occlusion, and photographer skills. Surface scans and dental casts give detailed geometry for tooth segmentation, annotation, orthodontic analysis, and prosthetics design.

The discipline needs to shift focus away from the size of the dataset as being the indicator of its quality to dataset utility. An extremely large dataset that is homogeneous may not be as useful as a smaller dataset from several centres that contains sampling, reference standards, subgroup population and testing outside of the training set. All research publications must disclose the number of patients involved, number of images/records, data acquisition machines used, period studied, exclusions, prevalence and class imbalance.

Critical interpretation: the evidence suggests that dataset volume should not be treated as a surrogate for validity. A smaller, multi-centre, clinically representative dataset with patient-level separation and transparent reference standards can provide stronger evidence than a much larger homogeneous dataset. The practical implication is that future studies should report the provenance and diversity of evidence before emphasizing scale.

4. NATURAL-LANGUAGE PROCESSING AND CLINICAL RECORDS

NLP technology is able to convert dentist's notes into structured elements, code, relationships, and time-stamped events. Rule-based NLP technology provided traceability but was not portable across organisations. Convolutional neural networks, BiLSTM-CRF, and transformer-based approaches enhanced context representation while being less dependent on manually created dictionaries. The use cases of such NLP applications include diagnosis and procedure extraction, drug surveillance, coding, registries, and clinical summarisation. [1]–[20]

F1 scores greater than 0.90 for entity extraction have been achieved and are very promising, but their results are heavily dependent on the characteristics of the annotation guidelines and the documentation style in a specific location. A language model created for one organisation, using its forms, will not be able to perform if there are differences in terms of abbreviations, spelling, terminology of specialties, or languages. The NLP community in dentistry must take into account the use of macro and class-based metrics, negation handling, abbreviation handling, error classification, and cross-site evaluations. There is much room for progress with LLMs in

summarisation and QA, but the problems with hallucinations, prompt sensitivity, provenance, and privacy are also important.

5. AI FOR TWO- AND THREE-DIMENSIONAL DENTAL IMAGING

Imaging is by far the most advanced field of dentistry regarding AI. Convolutions, U-Nets, object detectors, and vision transformers have been employed to perform tasks like tooth numbering, caries/lesion detection, periodontal bone loss evaluation, anatomical segmentation, age prediction, implant planning, and generation of radiographs. In terms of internal validity, the literature provides some excellent results in terms of accuracies and Dice coefficients in the 90% range in multiple studies. However, such results need to be considered together with prevalence, confidence intervals, reference standard, and validation type. [21]–[24], [27], [29], [30], [34], [37]–[42], [44], [45], [47]–[50]

For detection and classification, sensitivity, specificity, precision, recall, F1 score, area under the receiver-operating-characteristic curve, and precision-recall area must be chosen based on clinical implications and balance among classes. For segmentation, Dice and intersection-over-union scores need to be supplemented with boundary or surface distances in cases where millimeter-scale anatomical accuracy is needed. Calibration is necessary anytime probabilities impact referral or treatment decisions. Decision-curve analysis helps in assessing net clinical value at different threshold levels.

Transformers can capture global dependencies and distant contexts, while CNNs preserve efficient local inductive biases. Hybrid models might be suitable, but superiority claims should be substantiated through controlled experiments using same datasets, augmentation, training, and evaluation strategies. Evaluation from outside is crucial to be fulfilled. [21], [27], [40], [47], [48], [50]

Critical interpretation: the apparent performance advantage of a model family is scientifically meaningful only when it survives a controlled comparison using the same patients, preprocessing, training budget, reference standard, and test protocol. Cross-paper leaderboard comparisons can therefore create false certainty. For clinical imaging, the more consequential question is whether performance remains calibrated and stable across sites, scanners, prevalence shifts, and clinically difficult cases.

6. ROBOTICS, AUGMENTED REALITY, VIRTUAL REALITY, AND TELE-OPERATION

Robotics aims to enhance accuracy in positioning, drilling, implantation, endodontic instrumentation, and minimally invasive surgery through kinematic control, imaging guidance, force sensing, and visual feedback. Advanced prototypes have been developed that include elaborate compensation techniques for movement of patients, tool bending, and changes in resistance. Nonetheless, much of the research is still done using models and laboratory-based testing. Implementation in the clinical setting will call for a fail-safe system, emergency stopping procedures, controlled autonomy, collision detection, sterile operation, cybersecurity, and verification of controls. [25], [35], [36], [41], [43], [46]

Anatomical visualisation, planning, navigation, and skill development can be done using virtual and augmented reality. While the benefits to education are likely given the repeatability and performance tracking possible using immersive systems, it is not the case that enhanced confidence and simulator scores lead to patient performance. Studies need to include comparisons between learning retention, procedural accuracy, error rate, cognitive load, and clinical performance using traditional methods. [26], [32]

Latency, packet loss, network reliability, security issues, and liability are all problems associated with tele-operative dentistry. Force feedback and video synchronisation have to be tested on poor networks rather than perfect laboratory networks. The best model at this stage is assisted and remotely guided intervention, not autonomous. [43]

7. MULTIMODAL LEARNING, SYNTHETIC DATA, AND DIGITAL TWINS

The multimodal models may integrate imaging, textual information, demographic data, periodontal parameters, sensor measurements, and medical history in order to make context-aware predictions. The concept is intuitively understandable for clinical applications, since a diagnosis is not based on one isolated data modality. However, the multimodal learning approach is subject to such issues as the absence of some data modality, asynchronous acquisition, data modality exploitation, and the domination of one of the available data modalities. [55]–[61]

Synthetic images and generative models can be used to reduce annotation needs, aid simulation, and improve rare class representation. Synthetic

images cannot be presumed to diversify simply due to their high image realism. Utility is determined by how well it performs downstream, leakage of privacy information, review by experts, and discovery of anatomical anomalies. The training and test set cannot be the same as the generative model's training dataset. [22], [24], [28], [33], [39]

The dental digital twin should not be thought of as a three-dimensional model but rather a continuously updated computational construct specific to the patient. Validating a dental digital twin involves longitudinal data integration, causal mechanisms, uncertainty propagation, and trajectory prediction. To date, there is only theoretical knowledge available on the matter, and thus digital twins should be introduced as a research topic.

Critical interpretation: emerging systems do not merely add new model classes; they add new failure modes. Multimodal and generative systems introduce missing-modality, provenance, hallucination, and privacy risks, while digital twins introduce longitudinal and causal-validity requirements. Their evidential burden should therefore be higher than that of a static retrospective classifier, even when headline accuracy appears similar.

8. EXPLAINABILITY, UNCERTAINTY, FAIRNESS, PRIVACY, AND SECURITY

A robust dental AI needs more than just a visualisation heat map. Saliency and attention visualisations could give us clues about the plausibility of the anatomical structures on which the model is focusing; however, these visualisations are often unstable and do not provide proof of causality. The need for explainability should depend on the user and the specific task at hand, with different techniques used for radiologists, clinicians, and patients. [53], [54], [56], [59], [60]

The need for uncertainty estimation arises in cases where image quality is poor, pathology lies beyond the data distribution learned by the model, and probability is the basis for medical intervention. Calibration curves, ECE, Brier score, confidence intervals, and methods for abstaining/referring are to be reported. The safe model must be permitted to report its uncertainty and seek expert consultation. [53], [54], [56]

Subgroup analysis of performance by age, gender, demographic subpopulation as is ethically and legally possible, scanner type, institution, and disease severity is needed for assessing fairness. Privacy techniques like federated learning will limit central data collection, but won't preclude the need for governance, membership inference, model inversion, poisoning, and communication security concerns. Threat modeling and secure aggregation are needed along with privacy claims. [53], [54], [59]

9. CLINICAL TRANSLATION AND REGULATORY READINESS

Translation of clinical applications takes place by means of analytical validity, clinical validity, clinical utility, and lifecycle safety. Analytical validity is concerned with accuracy of measurement or prediction under certain technical conditions. Clinical validity poses a question regarding the generalisation of the performance of the test to the target population and situation. Clinical utility queries if the use is beneficial over the current approach. Lifecycle safety requires post-deployment evaluation as data, devices, prevalence, users, and workflow changes. [53]–[56], [60]

The regulatory requirements for AI-powered medical devices increasingly stress total product lifecycle data, pre-set change control, data handling, human factors, transparency, cybersecurity, and real-life performance monitoring. Thus, dental AI providers must specify the purpose of use precisely, determine user characteristics and effects of failure, implement version control, describe training data source, pre-establish update process, and track drift and subgroup performance. Retrospective accuracy study is an essential technical procedure; however, it cannot replace prospective clinical assessment. [53], [54], [59], [60] Table 4 represents the maturity level and major translational flaws of the considered technology fields, and Figure 3 shows the evidence maturity of them qualitatively.

Critical interpretation: deployment readiness is a property of the complete sociotechnical system rather than the trained model alone. A technically accurate model may still be unsuitable for practice if clinicians cannot recognize uncertainty, override errors, trace provenance, or detect performance drift. This finding shifts evaluation from a one-time benchmark to an evidence chain that continues through human interaction and post-deployment monitoring.

Table 4. Technology Domains, Evidence Strength, and Translational Weaknesses

Technology domain	Representative tasks	Typical evidence strength	Main translational weakness

NLP / EDR	Entity extraction, coding, registry creation, summarization	Moderate for local text extraction	Cross-site language variation; hallucination and provenance for LLMs
2D imaging	Tooth localization, caries, bone loss, classification	Moderate to relatively strong internal evidence	Limited external and prospective validation; class imbalance
3D CBCT	Segmentation, canal/nerve mapping, implant planning	Moderate technical evidence	Scanner heterogeneity; annotation burden; boundary accuracy
Robotics	Guided drilling, implant/endodontic assistance	Early-stage to moderate prototype evidence	Laboratory evaluation; safety certification; human factors
AR/VR	Training, planning, intraoperative overlays	Moderate educational/technical evidence	Transfer to clinical outcomes; latency and registration error
Multimodal AI	Joint image-text-risk prediction	Early-stage	Missing-modality robustness; calibration; data alignment
Digital twins	Patient-specific simulation and prediction	Conceptual/early-stage	Longitudinal validation; causal fidelity; uncertainty

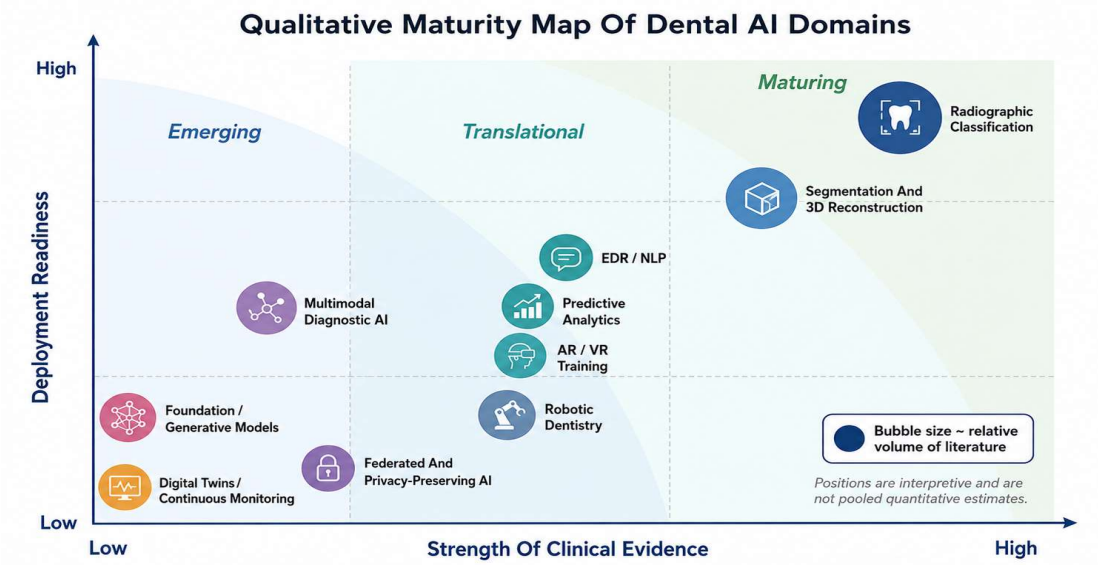

Figure 3. Qualitative maturity map based on the pattern of evidence in the reviewed corpus; positions are interpretive rather than pooled estimates.

Figure 3. Qualitative Maturity Map Based On The Pattern Of Evidence In The Reviewed Corpus; Positions Are Interpretive Rather Than Pooled Estimates.

10. COMPARATIVE EVALUATION REQUIREMENTS

Standardised reporting is one of the key reasons why robust research results do not replicate. The minimum set of requirements should be selected based on the type of task and its clinical application. Precision alone is not enough for an unbalanced diagnosis problem; Dice alone may overlook clinically relevant border errors; and a significant statistical difference may have no clinical relevance at all. The following table gives a useful guideline for reporting of future dental AI research. The minimum criteria for each type of study are listed in Table 5, and Table 6 separates propaganda from science.

Table 5. Minimum Evaluation Requirements by Study Type

Study type	Essential metrics	Essential validation	Additional clinical evidence
Binary/multiclass diagnosis	Sensitivity, specificity, precision, recall, macro-F1, ROC-AUC, PR-AUC, confidence intervals	Patient-level split; external or temporal test; calibration	Decision-curve analysis; reader study; error severity
Detection/localization	AP/mAP, sensitivity by lesion size, false positives per image	External device/site test; adjudicated reference standard	Impact on reading time and missed lesions
Segmentation	Dice, IoU, Hausdorff/surface distance, class-wise results	Patient/site separation; scanner variability	Anatomical acceptability and planning impact
NLP	Entity/relation macro-F1, exact match, error categories	Cross-site and cross-template evaluation	Documentation time, correction burden, provenance
Robotics	Position/orientation error, force error, failure rate, latency	Phantom + cadaver + staged clinical evaluation	Human factors, fail-safe behavior, complications
AR/VR education	Objective performance, retention, workload, error rate	Randomized or controlled comparator	Transfer to supervised clinical performance
Generative/synthetic data	Fidelity, diversity, privacy risk, downstream utility	Independent expert and model-based evaluation	Anatomical validity and failure taxonomy

Table 6. Replacement Of Weak Claims With Defensible Scientific Interpretations

Common claim	Why it is weak	More defensible formulation
"The model is clinically ready because accuracy is 97%."	Accuracy may be inflated by prevalence, leakage, and internal testing.	"The model achieved 97% internal accuracy; external calibration and prospective utility remain to be established."
"The AI outperformed dentists."	Reader expertise, case mix, confidence intervals, and workflow context may be unclear.	"In this retrospective reader study, the model exceeded the prespecified comparator under the tested conditions."
"Heat maps make the system explainable."	Heat maps may be unstable and do not establish causal or clinically useful explanations.	"Localization maps were assessed for plausibility; explanation fidelity and user benefit require further study."
"Federated learning guarantees privacy."	Model updates can still leak information or be attacked.	"Federated learning reduces raw-data centralization and requires secure aggregation, threat modeling, and governance."
"Synthetic data solved class imbalance."	Synthetic samples can reproduce artifacts or reduce realism.	"Synthetic augmentation improved downstream performance in the tested setting; privacy and anatomical validity were separately assessed."

11. EVIDENCE GAPS THAT SHOULD DRIVE THE NEXT GENERATION OF STUDIES

- Multi-center studies in prospective setting that will measure performance both before and after the algorithm being integrated into real-world clinical workflows.
- Benchmarking datasets that are either publicly available or governed that have individual patient partitions, rich metadata, and independent test splits.
- Comparisons using identical settings for preprocessing, optimisation, and data splits, as

opposed to comparison using numbers from another publication.

- Calibration, confidence intervals, subgroup analysis, failures, and out-of-distribution analysis routinely reported.
- Human-factor studies analyzing automation bias, overrides, cognitive workload, explanations, and attribution of responsibility.
- Clinical-economics based on observed time, additional tests, complications, referrals, and maintenance costs rather than hypothetical costs.
- Lifecycle monitoring plan in regard to data drift, model drift, device modifications, software changes, incidents, and re-training.
- Multimodal dental language models and domain models evaluated in terms of hallucinations, provenance, privacy, and domain shift.
- Robotics and tele-operating systems tested under degraded sensory input, communication, registration, and haptics.

12. A STAGED RESEARCH AGENDA FOR TRUSTWORTHY DIGITAL DENTISTRY

Realistic timeline should not forecast adoption by year alone. Adoption is contingent upon evidence gates being met. Stage 1 requires data governance, problem definition, sampling, and replicable technical baseline. Stage 2 involves robust analytical validation using patient-level partitioning, testing, calibration, and failures analysis. Stage 3 requires assessment of man-machine interaction in silent deployment and controlled reader or workflow studies. Stage 4 involves clinical utility testing, which includes patient safety, timing, referrals, and resource utilisation. Stage 5 involves lifecycle management, including updates, cybersecurity, and periodic review. Table 7 indicates the evidence gate and minimal deliverable for each stage while Figure 4 shows the associated development and lifecycle pipeline.

The same process is involved for imaging, NLP, robotics, and multimodal applications; the level of evidence required for safety assurance will vary. Autonomous or interactive systems need more validation than decision-support systems. Outputs that carry high risks must incorporate uncertainty-aware referrals and confirmation from the doctor. This process is not about removing the clinician from the loop but ensuring a loop where the responsibility is clearly defined and the limitations of the system are clearly apparent.

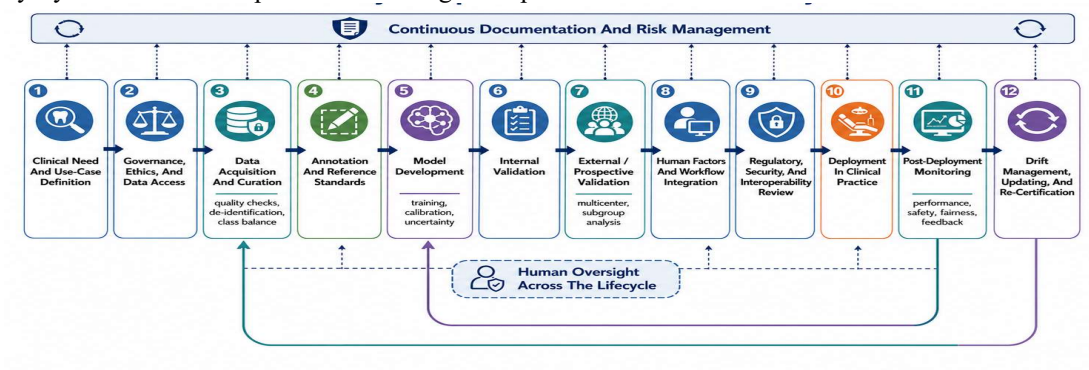

Figure 4. Development and lifecycle pipeline recommended for clinically responsible dental AI.

Table 7. Staged Research Agenda and Evidence Gates

Stage	Evidence gate	Minimum deliverable
1. Problem and data readiness	Clear intended use; representative data; reliable labels	Protocol, data dictionary, governance plan, baseline comparator
2. Analytical validity	Robust internal and external performance	Locked model, calibration, confidence intervals, error and subgroup analysis
3. Human factors	Safe and understandable clinician interaction	Usability study, override policy, explanation evaluation, training material
4. Clinical utility	Improved decision/process/outcome versus current practice	Prospective trial or rigorous workflow study, safety and economic outcomes

5. Lifecycle assurance	Performance remains acceptable after deployment	Drift dashboard, incident process, cybersecurity controls, controlled updates

13. FOUNDATION MODELS, LARGE MULTIMODAL MODELS, AND GENERATIVE DENTAL AI

The current state of dental AI is characterized by reusable pretrained models as opposed to task-specific networks. The vision foundation models learn dental features from large datasets of radiographic, CBCT volumetric, intraoral photographic, and 3D imaging data; the vision-language systems link the images with corresponding reports, questions, or treatments; and large language models help with documentation, education, information search, and conversational tasks. The major strength of such models is transfer, in which one pertained representation is reused for segmentation, detection, classification, landmarking, report generation, and cross-modal retrieval with reduced amount of task-specific labeling. The major danger of such models is that the scale hides the origin of the data, failure modes, and the distinction between linguistic plausibility and clinical accuracy.

Evaluation of dental foundation models should thus move beyond simple accuracy metrics. Tests that are required include zero-shot and few-shot transfer learning, generalisation to unseen sites, robustness with respect to hardware and acquisition modality, label efficiency, calibration, abstention, rare condition behaviour, and sensitivity to missing or conflicting modality inputs. In the case of generative models, factuality and citation provenance should be assessed independently from fluency. Evaluation of image generation work should include anatomical accuracy, use in downstream tasks, memorisation and privacy risk, and clinically impactful hallucinations.

However, there are distinct human factors that arise with the use of large multimodal models. An easy-to-understand interpretation may actually foster undue trust, even when the underlying prediction is incorrect. In such a case, the interface used within clinical practice must provide source evidence, uncertainty, modality availability, and limitations, along with an ability to differentiate between retrieved facts and generated information; and rapid correction by the clinician must be possible. Guidance on large multimodal models in healthcare practice issued by the WHO stresses the importance of governance, transparency, accountability, and protection of human rights, while modern medical-AI regulation tends more towards life cycle than one-time validation [59], [60].

Table 8. Evaluation Requirements for Foundation and Multimodal Models

Model class	Dental opportunity	Minimum evaluation	Primary safety concern
Task-specific CNN/Transformer	Caries, bone loss, lesion detection, tooth numbering	Patient-level external test; class-wise metrics; calibration	Leakage and device/site domain shift
Vision foundation model	Reusable representations across radiographs, CBCT and scans	Linear probing; few-shot transfer; multi-site benchmark; OOD testing	Hidden pretraining overlap and uncertain generalization
Vision-language model	Image–text retrieval, reporting, question answering	Grounded factuality; retrieval provenance; expert error taxonomy	Hallucination and unsupported recommendations
Large language model	Documentation, education, triage support	Scenario-based clinical evaluation; citation verification; red-team testing	Confident misinformation and automation bias
Generative image model	Augmentation, reconstruction, modality translation	Anatomical validity; privacy audit; downstream clinical utility	Fabricated pathology or erased abnormality
Multimodal clinical assistant	Joint reasoning over images, notes, history and sensors	Missing-modality tests; calibration; prospective workflow study	Cross-modal contradiction and responsibility ambiguity

14. EVIDENCE-MATURITY MATRIX AND CLINICAL UTILITY LADDER

One flaw that occurs repeatedly in reviews of AI in dentistry is the ranking of fundamentally different pieces of evidence at the same level. A retrospective internal test, an external validation, a prospective silent deployment, and a randomised workflow evaluation respond to distinct research questions. For this reason, the evidence-maturity matrix presented below can be regarded as the key instrument of interpretation. Progression is cumulative; later levels do not invalidate the former and high-performing algorithms cannot compensate for poor-quality data or workflow.

Table 9. Evidence-Maturity and Clinical Utility Ladder

Level	Core question	Representative evidence	Decision permitted
M0: Concept feasibility	Can the computational idea operate on a small curated sample?	Proof of concept; exploratory dataset	Research continuation only
M1: Analytical validity	Does the system measure the intended target under controlled conditions?	Locked internal test; leakage control; reference-standard audit	Technical comparison
M2: Transportability	Does performance persist across sites, devices and populations?	External or temporal validation; subgroup analysis; calibration	Limited clinical pilot
M3: Clinical validity	Does the output correspond to clinically meaningful states or outcomes?	Prospective or well-designed retrospective clinical study	Supervised decision support
M4: Clinical utility	Does use improve decisions, workflow, safety or patient outcomes?	Silent deployment; reader study; controlled impact study	Operational deployment with oversight
M5: Lifecycle assurance	Does performance remain safe after deployment and change?	Monitoring, drift detection, incident handling, update controls	Sustained regulated use

The reviewed literature is the best in M1 for two-dimensional image analysis and selected segmentation tasks. Certain systems reach M2 via external databases, but calibration and subgroup analysis are still unreliable. Robotics, teleoperation, digital twin models, and multimodal assistance frequently reside in M0–M2 due to the dominance of laboratory feasibility studies. The key research opportunity thus lies not in a minor accuracy improvement based on a well-known retrospective split, but in a move from M1 to M3–M5 using external, prospective, human-factors, economic, and lifecycle evidence. Table 9 presents the six-level evidence maturity and clinical utility ladder.

15. QUANTITATIVE SYNTHESIS WITHOUT MISLEADING META-ANALYSIS

Formal meta-analysis does not apply in cases where different tasks, disease definitions, sampling methods, gold standard criteria, and outcomes are applied. However, a systematic review should still involve some numerical measures. Some recommended descriptive measures include publication year, dental specialty, method, dataset size, number of sites, patient-level versus image-level splitting, external validation, prospective evaluation, model family, calibration, explanation method, and code or data availability. It is best to apply medians and interquartile ranges rather than pooled means if there is skewness. Accuracy values must never be pooled if they relate to different classes and outcomes. Table 10 shows numerical variables that could be combined in order to avoid misrepresentation of pooled numbers.

If sufficiently homogeneous subgroups can be formed, then random effects meta-analysis can be considered for one defined purpose and measure, where threshold definitions, patient populations, gold standards, and study designs are similar. Synthesis of diagnostic accuracies should emphasise the use of sensitivity/specificity pairs, hierarchical models, and threshold handling, rather than the pooling of accuracy. The heterogeneity issue must be dealt with both clinically and methodologically, rather than just statistically. Publication bias assessment is not reliable when the number of studies is small or benchmark reuse results in non-independent measures.

Table 10. Quantitative Synthesis Variables and Reporting Priorities

Synthesis variable	Why it matters	Preferred reporting
Dataset origin and centre count	Indicates spectrum and site diversity	Number of patients, images, centres, countries and devices
Partitioning unit	Detects patient or acquisition leakage	Patient-level, study-level or tooth-level split with rationale
Reference standard	Determines label credibility	Number and expertise of annotators; consensus; agreement
External validation	Tests transportability	Completely independent institution/time/device
Calibration and uncertainty	Supports thresholded decisions and abstention	ECE/Brier/calibration curve plus confidence intervals
Clinical comparator	Defines real-world relevance	General dentist, specialist, consensus panel or standard workflow

Prospective impact	Tests utility rather than discrimination	Time, decisions, referrals, complications, patient outcomes
Reproducibility	Enables verification	Code, model card, data statement, split manifest, reporting checklist

16. MINIMUM REPORTING STANDARD FOR DENTAL ARTIFICIAL INTELLIGENCE STUDIES

In order to enhance reproducibility and reviewability, the following reporting standard is suggested. This standard combines study design, data source, modeling, statistical analysis, clinical setting, and planning across the life cycle. It can be used by journals as an author checklist, and by reviewers to spot unsupported claims. Table 11 includes the suggested minimum reporting standard.

Table 11. Minimum Reporting Standard for Dental Artificial Intelligence Studies

Domain	Mandatory items
Clinical question	Intended user; intended use; target population; care setting; decision supported; clinical consequence of error
Dataset	Source, dates, inclusion/exclusion, centres, devices, demographics, prevalence, missingness, patient linkage, de-identification
Reference standard	Annotation protocol, assessor expertise, blinding, consensus, intra/inter-rater agreement, ambiguous-case handling
Partitioning	Patient-level separation, temporal separation where appropriate, external test provenance, duplicate and near-duplicate checks
Model	Architecture, initialization, preprocessing, augmentation, loss, optimizer, schedule, seeds, stopping rule, software and hardware
Comparators	Clinically relevant baseline and fair model baselines trained on identical data and tuning budgets
Statistics	Prespecified primary endpoint, confidence intervals, class-wise results, calibration, paired testing, sample-size justification
Robustness	Site/device shift, image-quality degradation, missing modality, OOD cases, subgroup analysis, failure examples
Explainability	Method, stability, sanity checks, clinician evaluation, explicit warning that explanation does not prove causality
Human factors	Interface, training, override, automation bias, time burden, disagreement resolution, final accountability
Deployment	Integration, latency, cybersecurity, privacy, monitoring, drift, incident response, retraining and change control
Transparency	Funding, conflicts, data/code availability, protocol, model card, limitations and negative findings

17. CRITICAL DISCUSSION AND SYNTHESIS

The central pattern across the reviewed literature is not a shortage of technically capable models but a mismatch between technical performance and the strength of evidence needed for clinical use. The literature is comparatively strong at demonstrating whether an algorithm can discriminate, detect, segment, or generate under controlled conditions, but substantially weaker at demonstrating whether the same behaviour persists across institutions, devices, patient groups, workflow pressures, and software updates. This distinction changes the interpretation of published results: a high metric is evidence of task performance under stated conditions, not a general certificate of clinical readiness.

17.1 The Accuracy-Utility Paradox

Two-dimensional imaging and selected segmentation tasks occupy the most mature part of the evidence landscape, yet even here external validation, calibration, subgroup analysis, and prospective impact are inconsistent. The review therefore challenges the common assumption that incremental gains in accuracy represent equivalent gains in clinical value. Once analytical performance is already high, the greater scientific return may come from testing transportability, uncertainty, decision thresholds, reader interaction, and patient-relevant outcomes rather than optimizing another fraction of a percentage point.

17.2 Why Cross-Study Performance Rankings Can Mislead

Reported metrics are conditioned by prevalence, case difficulty, annotation rules, split unit, scanner or device, preprocessing, and comparator design. Consequently, two studies reporting similar accuracy may provide very different levels of evidence, while a model with a slightly lower metric may be scientifically stronger if evaluated externally with confidence intervals and calibration. The reader should therefore interpret numerical performance together with evidence quality. The defensible comparison is not simply model versus model, but evidence package versus evidence package.

17.3 Emerging Technologies Increase the Evidence Burden

Foundation models, multimodal assistants, generative systems, robotics, tele-operation, and digital twins broaden the potential value of digital dentistry but also move risk beyond ordinary classification error. Hallucinated recommendations, hidden pretraining overlap, cross-modal contradiction, control latency, degraded sensing, cybersecurity, causal-model error, and update-related drift can all affect safety. The review therefore infers a risk-proportionate principle: the more autonomous, interactive, generative, multimodal, or continuously changing the system, the stronger the required evidence for uncertainty handling, human oversight, robustness, and lifecycle governance.

17.4 Practical and Scientific Takeaway

The main takeaway for researchers, reviewers, and clinicians is to replace the question 'Which model has the highest reported accuracy?' with 'What level of use is justified by the available evidence, and what is the next missing evidence gate?' The proposed maturity ladder and reporting standard make that question operational. They also provide a way to value negative findings, transportability studies, calibration work, workflow trials, and post-market monitoring as scientific contributions rather than as secondary implementation details.

18. DIFFERENCE FROM PRIOR WORK AND ACHIEVEMENT OF THIS STUDY

Compared with application-focused systematic reviews such as [55], deployment-concern reviews such as [56], dental education and foundation-model discussions such as [57], and multimodal medical-imaging implementation guidance such as [58], this study contributes a dentistry-specific framework that places heterogeneous technologies on one evidence continuum without treating their reported metrics as directly interchangeable. Its difference from prior work is therefore conceptual as well as methodological: it integrates technology mapping, risk-of-bias reasoning, maturity classification, clinical utility, minimum reporting, and lifecycle governance into a single decision-oriented synthesis. The achievement of the study is the conversion of a diffuse literature into four actionable outputs: (i) an evidence-to-deployment taxonomy, (ii) an M0-M5 evidence-maturity and clinical-utility ladder, (iii) task- and system-specific minimum evaluation/reporting requirements, and (iv) staged evidence gates that specify what should be demonstrated before stronger claims of deployment are made. This combination gives the reader a practical method for distinguishing promising proof-of-concept work from evidence that can support supervised clinical use or sustained deployment.

19. LIMITATIONS AND SELF-CRITIQUE

This review also has limitations that constrain its conclusions. First, the included evidence is highly heterogeneous in clinical task, disease definition, data source, reference standard, and outcome metric; for that reason, pooled accuracy would create a false impression of comparability and was deliberately avoided. Second, the maturity map and evidence-level assignments are interpretive syntheses rather than pooled quantitative estimates, so incomplete reporting in primary studies can affect how confidently a study is placed on the proposed ladder. Third, rapidly evolving foundation-model, multimodal, robotics, and regulatory literature means that the evidence base can change quickly after the July 2026 search boundary. Fourth, because prospective workflow, economic, and lifecycle studies remain sparse, the review can identify missing evidence and plausible deployment requirements but cannot establish that any proposed framework itself improves patient outcomes. These limitations do not weaken the need for the framework; instead, they define where it should be tested. Future work should prospectively apply the maturity ladder and reporting standard to new dental-AI studies, assess inter-reviewer agreement in maturity classification, and evaluate whether use of the framework improves reproducibility, claim calibration, and study design.

20. CONCLUSION

This critical scoping review shows that the principal scientific challenge in digital dentistry has shifted from demonstrating that computational systems can achieve high benchmark performance to demonstrating that their outputs remain valid, useful, safe, and governable across real clinical conditions. The most consistent gap is an accuracy-

to-deployment gap: technical performance is commonly reported, whereas transportability, calibration, subgroup reliability, human factors, prospective utility, economic impact, and lifecycle assurance are much less consistently established. Therefore, the level of clinical claim should be determined by evidence maturity rather than by headline accuracy alone.

The scientific contribution of this work is the integration of previously fragmented technical and translational evidence into a common decision framework. Specifically, the study contributes a five-layer evidence-to-deployment taxonomy, an M0-M5 evidence-maturity and clinical-utility ladder, a minimum reporting standard, and staged evidence gates for progression from problem definition to lifecycle assurance. These constructs add significantly to already published knowledge because they do not merely summarize what technologies have been studied; they explain how different forms of evidence should be weighted, why seemingly strong results may still be insufficient for deployment, and what next evidence is required to justify a stronger level of use. The resulting new knowledge is a risk-proportionate principle for digital dentistry: as systems become more autonomous, multimodal, generative, interactive, or continuously updated, the evidential burden must expand from accuracy toward uncertainty management, human oversight, robustness, provenance, cybersecurity, and post-deployment monitoring. This provides researchers, reviewers, clinicians, and regulators with a concrete basis for designing and judging future work on clinically trustworthy digital dentistry.

FUNDING

This review received no external funding.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

The study synthesizes previously published literature and did not directly involve human participants, identifiable personal data, or animals; institutional ethical approval and informed consent were therefore not applicable.

DATA AVAILABILITY

All evidence analyzed in this review is available in the cited publications. The article reports the search scope, eligibility criteria, extraction fields, appraisal domains, and evidence-synthesis framework. No new patient-level or identifiable data were generated or analyzed.

ACKNOWLEDGMENT

The authors acknowledge the academic support of the Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad.

REFERENCES

[1] Pethani, F., & Dunn, A. G. (2023). Natural language processing for clinical notes in dentistry: A systematic review. Journal of Biomedical Informatics.

[2] Shastry, K. A., & Shastry, A. (2023). An integrated deep learning and natural language processing approach for continuous remote monitoring in digital health. Decision Analytics Journal.

[3] Rahimi-Ardabili, H., Dras, M., & Coiera, E. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. International Journal of Medical Informatics

[4] McMaster, C., Chan, J., & Liew, D. F. L. (2023). Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. Journal of Biomedical Informatics.

[5] Hossain, E., Rana, R., Higgins, N., Soar, J., & Barua, P. D. (2023). Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. Computers in Biology and Medicine.

[6] Landolsi, M. Y., Hlaoua, L., & Ben Romdhane, L. (2023). Information extraction from electronic medical documents: State of the art and future research directions. Knowledge and Information Systems.

[7] Singleton, J., Li, C., Akpunonu, P. D., & Abner, E. L. (2023). Using natural language processing to identify opioid use disorder in electronic health record data. International Journal of Medical Informatics.

[8] Wyles, C. C., Fu, S., Odum, S. L., Rowe, T., & Habet, N. A. (2023). External validation of natural language processing algorithms to extract common data elements in THA operative notes. The Journal of Arthroplasty

[9] Raza, S., & Schwartz, B. (2023). Constructing a disease database and using natural language processing to capture and standardize free-text clinical information. Scientific Reports.

[10] Tavabi, N., Pruneski, J., Golchin, S., & Singh, M. (2024). Building large-scale registries

from unstructured clinical notes using a low-resource natural language processing pipeline. Artificial Intelligence in Medicine.

[11] Amosa, T. I., Izhar, L. I. B., Sebastian, P., & Ismail, I. B. (2023). Clinical errors from acronym use in electronic health records: A review of NLP-based disambiguation techniques. IEEE Xplore.

[12] Liu, F., Liu, M., Li, M., Xin, Y., Gao, D., & Wu, J. (2023). Automatic knowledge extraction from Chinese electronic medical records and rheumatoid arthritis knowledge graph construction. PubMed Central.

[13] Houssein, E. H., Mohamed, R. E., & Ali, A. A. (2023). Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. Scientific Reports.

[14] Li, C., Zhang, Y., Weng, Y., Wang, B., & Li, Z. (2023). Natural language processing applications for computer-aided diagnosis in oncology. MDPI Journals.

[15] Wang, C., Yao, C., Chen, P., Shi, J., Gu, Z., & Zhou, Z. (2023). Artificial intelligence algorithm with ICD coding technology guided by embedded electronic medical record system in medical record information management. ScienceDirect.

[16] Raza, S., Dolatabadi, E., Ondrusek, N., & Rosella, L. (2023). Discovering social determinants of health from case reports using natural language processing: Algorithmic development and validation. Springer.

[17] Kumar, A., & Gond, A. (2023). Natural language processing: Healthcare achieving benefits via NLP. ScienceOpen.

[18] Tavabi, N., Singh, M., Pruneski, J., & Kiapour, A. M. (2023). Systematic evaluation of common natural language processing techniques to codify clinical notes. PLOS ONE.

[19] Poniszewska-Marańda, A., & Vynogradnyk, E. (2023). Medical data transformations in healthcare systems with the use of natural language processing algorithms. MDPI Journals.

[20] Cai, J., Chen, S., Guo, S., Wang, S., Li, L., & Liu, X. (2023). RegEMR: A natural language processing system to automatically identify premature ovarian decline from Chinese electronic medical records. Springer.

[21] Lee, Y., et al. (2025). Oral-anatomical knowledge-informed semi-supervised learning for 3D dental CBCT segmentation and lesion detection. IEEE Transactions on Automation Science and Engineering, 22, 11205–11218. https://doi.org/10.1109/TASE.2025.3530936

[22] Sunilkumar, A. P., Parida, B. K., & You, W. (2024). Recent advances in dental panoramic X-ray synthesis and its clinical applications. IEEE Access, 12, 141032–141051. https://doi.org/10.1109/ACCESS.2024.3422650

[23] Bussaban, L., Boonchieng, E., Panyarak, W., Charuakkra, A., & Chulamanee, P. (2024). Age group classification from dental panoramic radiographs using deep learning techniques. IEEE Access, 12, 139962–139973. https://doi.org/10.1109/ACCESS.2024.3466953

[24] Sunilkumar, A. P., Parida, B. K., Moon, S.-Y., & You, W. (2025). Panoramic X-ray synthesis from dental CBCT using patient-tailored simulated geometry with uniform projection sampling. IEEE Access, 13, 100752–100763. https://doi.org/10.1109/ACCESS.2025.3573744

[25] Xia, Z., et al. (2024). Robotics application in dentistry: A review. IEEE Transactions on Medical Robotics and Bionics, 6(3), 851–867. https://doi.org/10.1109/TMRB.2024.3408321

[26] Bhat, S. H., Hareesha, K. S., Kamath, A. T., Kudva, A., Vineetha, R., & Nair, A. (2024). A framework to enhance the experience of CBCT data in real-time using immersive virtual reality: Impacting dental pre-surgical planning. IEEE Access, 12, 45442–45455. https://doi.org/10.1109/ACCESS.2024.3375770

[27] Wu, Z., Huangjian, & Zhengsong. (2025). A feature selection and enhanced reuse framework for detection and classification of dental diseases in panoramic dental X-rays images. IEEE Access, 13, 70741–70751. https://doi.org/10.1109/ACCESS.2025.3561362

[28] Zhao, M., et al. (2024). Enlarge the error prediction dataset in 3-D printing: An unsupervised dental crown mesh generator. IEEE Transactions on Computational Social Systems, 11(6), 7929–7940. https://doi.org/10.1109/TCSS.2024.3417388

[29] Li, K.-C., et al. (2024). Detection of tooth position by YOLOv4 and various dental problems based on CNN with bitewing radiograph. IEEE Access, 12, 11822–11835.

https://doi.org/10.1109/ACCESS.2023.3348788

[30] Suresh Kumar M, Vijaya Bhaskar Sadu, Mayura Shelke, Vivekanandhan V (2026). Enhanced optimization of the diagnosis of liver disease with an intelligent adaptive neuro-fuzzy inference system, *Biomedical Signal Processing and Control*, vol(117), 109494, https://doi.org/10.1016/j.bspc.2026.109494.

[31] Mayya, V., King, C., Vu, G. T., & Gurupur, V. (2024). Empirical study of feature selection methods in regression for large-scale healthcare data: A case study on estimating dental expenditures. IEEE Access, 12, 153564–153579. https://doi.org/10.1109/ACCESS.2024.3482192

[32] Samuel, S., Elvezio, C., Khan, S., Bitzer, L. Z., Moss-Salentijn, L., & Feiner, S. (2024). Visuo-haptic VR and AR guidance for dental nerve block education. IEEE Transactions on Visualization and Computer Graphics, 30(5), 2839–2848. https://doi.org/10.1109/TVCG.2024.3372125

[33] Kokomoto, K., Okawa, R., Nakano, K., & Nozaki, K. (2024). Tooth development prediction using a generative machine learning approach. IEEE Access, 12, 87645–87652. https://doi.org/10.1109/ACCESS.2024.3416748

[34] Naufal, M. F., Fatichah, C., Astuti, E. R., & Putra, R. H. (2025). DIR-YOLO: Segmentation of alveolar bone and mandibular canal in CBCT images for dental implant recommendation. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3621623

[35] Bharadwaj, S. R., Nayak, V. M., Prabhu, N., & Shenoy, P. (2025). Design and validation of an IMU-based handpiece angulation measurement system for guided dental implant surgery. IEEE Access, 13, 176713–176728. https://doi.org/10.1109/ACCESS.2025.3619409

[36] Shi, Y., Kim, S., Zou, Q., & Yi, B.-J. (2025). Development of a novel 7-DOF parallel-serial hybrid robotic system for precise removal of dental caries. IEEE/ASME Transactions on Mechatronics. https://doi.org/10.1109/TMECH.2025.3603020

[37] Dascalu, T., et al. (2025). DentAssignNet: Assignment network for dental cast labeling in the presence of dental abnormalities. IEEE Journal of Biomedical and Health Informatics, 29(7), 4981–4990. https://doi.org/10.1109/JBHI.2025.3549685

[38] R Sheeja, Vijaya Bhaskar Sadu, R Shobana, R Kala (2025). Brain multi modality image inpainting via deep learning based edge region generative adversarial network, *Technology and Health Care*, 33(3), https://doi.org/10.1177/09287329241300986.

[39] Yu, W., Guo, X., Li, W., Liu, X., Chen, H., & Yuan, Y. (2025). ToothMaker: Realistic panoramic dental radiograph generation via disentangled control. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2025.3588466

[40] Almalki, A., Almalki, A., & Latecki, L. J. (2025). Self-supervised masked deep embeddings for dental caries detection. IEEE Access, 13, 161206–161216. https://doi.org/10.1109/ACCESS.2025.3606811

[41] Zou, Q., Yi, B.-J., & Shi, Y. (2024). Mechanical design and analysis of a novel 6-DOF hybrid robot with large orientation workspace for dental application. IEEE Access, 12, 150824–150843. https://doi.org/10.1109/ACCESS.2024.3468953

[42] Zannah, R., Bashar, M., Mushfiq, R. B., Chakrabarty, A., Hossain, S., & Jung, Y. J. (2024). Semantic segmentation on panoramic dental X-ray images using U-Net architectures. IEEE Access, 12, 44598–44612. https://doi.org/10.1109/ACCESS.2024.3380027

[43] Shi, Y., Zou, Q., Yi, B.-J., & Maeng, S. J. (2025). Latency and reliability optimization of a WebRTC-based remote dental surgery system with real-time force feedback. IEEE Access, 13, 176845–176857. https://doi.org/10.1109/ACCESS.2025.3620994

[44] Lai, Y., et al. (2025). Automated dental landmarks recognition in special models via neural network. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2025.3590442

[45] VB Sadu, RS Kumar, BS Kumar, T Kavitha, HK Chapala, MK Chakravarthi (2024), Evaluating Machine Learning Models for Multimodal Probability-Based Energy

Forecasting. *Process Integration and Optimization for Sustainability*, vol(8), pp:1209-1222. https://doi.org/10.1007/s41660-024-00428-0.

[46] Cheng, H.-F., Ho, Y.-C., & Chen, C.-W. (2025). Autonomous dental surgery for root canal treatment: Compensating for robot-patient misalignment and file deflection. IEEE Transactions on Automation Science and Engineering, 22, 21412–21429. https://doi.org/10.1109/TASE.2025.3611997

[47] Vijaya Bhaskar Sadu, Kumar Abhishek, Omaia Mohammed Al-Omari, Sandhya Rani Nallola, Rajeev Kumar Sharma, Mohammad Shadab Khan (2024). Enhancement of cyber security in IoT based on ant colony optimized artificial neural adaptive Tensor flow. *Network: Computation in Neural Systems*, 36(3), https://doi.org/10.1080/0954898X.2024.2336058. https://doi.org/10.1109/TMM.2024.3428349

[48] Zhuang, S., Wei, G., Cui, Z., & Zhou, Y. (2025). Robust hybrid learning for automatic teeth segmentation and labeling on 3D dental models. IEEE Transactions on Multimedia, 27, 792–803. https://doi.org/10.1109/TMM.2023.3289760

[49] Kong, X., et al. (2024). A low-cost portable 50 mT MRI scanner for dental imaging. IEEE Transactions on Instrumentation and Measurement, 73, 1–11. https://doi.org/10.1109/TIM.2023.3329089

[50] Ahn, J. S., & Cho, Y.-R. (2024). Weighted sparse convolution and transformer feature aggregation networks for 3D dental segmentation. IEEE Access, 12, 135172–135184. https://doi.org/10.1109/ACCESS.2024.3462521

[51] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

[52] Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine, 169(7), 467–473. https://doi.org/10.7326/M18-0850

[53] U.S. Food and Drug Administration. (2025). Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations. Draft guidance for industry and Food and Drug Administration staff.

[54] International Medical Device Regulators Forum. (2025). Good machine learning practice for medical device development: Guiding principles (IMDRF/AIML WG/N88 FINAL:2025).

[55] Araidy, S., Batshon, G., & Mirochnik, R. (2025). Artificial intelligence applications in dentistry: A systematic review. Oral, 5(4), Article 90. https://doi.org/10.3390/oral5040090

[56] Lal, A., Nooruddin, A., & Umer, F. (2025). Concerns regarding deployment of AI-based applications in dentistry: A review. BDJ Open, 11, Article 27. https://doi.org/10.1038/s41405-025-00319-7

[57] Claman, D., & Sezgin, E. (2024). Artificial intelligence in dental education: Opportunities and challenges of large language models and multimodal foundation models. JMIR Medical Education, 10, e52346. https://doi.org/10.2196/52346

[58] Huang, S. C., et al. (2025). A systematic review and implementation guidelines of multimodal foundation models in medical imaging. npj Digital Medicine. PMID: 40343333.

[59] World Health Organization. (2025). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. World Health Organization. ISBN 978-92-4-008475-9.

[60] Dr. R. Deepa, Vijaya Bhaskar Sadu, Dr. A. Sivasamy (2024). Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records. *AIP Advances*, 14(3), https://doi.org/10.1063/5.0191990.

[61] Lv, H., Haq, I., Du, J., et al. (2026). A benchmark multimodal oro-dental dataset for large vision-language models. Scientific Data, 13, Article 990. https://doi.org/10.1038/s41597-026-07342-9