AMERICAN JOURNAL OF BIOMEDICINE AND PHARMACY
https://biojournals.us/index.php/AJBP
ISSN: 2997-7177

Article

Evaluation of the Role of Artificial Intelligence in Treatment Planning and Orthodontics: Narrative Review

Lara Talib Sharif[1], Lava Talib Shireef[2]
1. Kirkuk Health Department, Iraqi Ministry of Health, Iraq
2. Renewable Energy Research Center-Kirkuk, Northern Technical University, Iraq

Citation: Sharif, L. T., Shireef, L. T. Evaluation of the Role of Artificial Intelligence in Treatment Planning and Orthodontics: Narrative Review. American Journal of Biomedicine and Pharmacy 2026, 3(9), 91-105.

Received: 18th Jul 2026
Revised: 05th Aug 2026
Accepted: 25th Aug 2026
Published: 08th Sept 2026

Abstract: AI is incorporated in almost all stages of the digital orthodontic workflow, from the automated analysis of diagnostic records, the simulation of treatment outcomes, and remote monitoring of tooth movement. In a very short duration, there has been an explosion of models that are published, which has outpaced critical appraisal of their clinical reliability. The aim of this study is to assess the current role of AI in orthodontic treatment planning through the synthesis of reported diagnostic and predictive performance across the main clinical domains. Furthermore, we will examine the influence of task complexity and study design on performance, and finally identify the methodological, ethical, and regulatory barriers that continue to limit routine clinical integration. A narrative review was conducted utilising PubMed/MEDLINE, Scopus, and Web of Science available for articles published from January 2009 to June 2026. Priority was given to systematic reviews, scoping reviews, meta-analyses, and primary diagnostic –accuracy or predictive-modelling studies published in English. The primary authors provided performance metrics, which were extracted without pooling and are presented accordingly. Seven domain were identified cephalometric landmark identification, malocclusion classification, extraction decision support, skeletal maturation assessment, orthognathic outcome prediction, clear aligner planning, remote monitoring. The reported performance follows a consistent gradient determined by the structure of the task and not by its clinical relevance. The tooth segmentation task achieved 98% performance, the two-dimensional landmark detection task achieved 98.3% performance, the binary crossbite classification task achieved 98.6% performance, and the extraction decision task achieved an area under the curve of 91.2%. Multiclass tasks fall to 71–91%. Fine-grained spatial tasks are lowest, with successful detection rates of 67.5% within 2 mm for posteroanterior landmarks and a pooled three-dimensional landmarking error of 2.44 mm. performance also declines when models are across multiple centres: cervical vertebral maturation staging accuracy fell from 98% in a single-centre cohort to 71.1–78.3% in a six-centre cohort involving 3,600 images. AI is delivers its most robust and reproducible benefits in reduced analysis time and reduce inter-observer variability rather than in unambiguous diagnostic superiority over experienced clinicians. given limited external validation, dataset homogeneity, inconsistent reporting and unresolved ethical and medico-legal question, AI is best positioned as a decision-support adjunct requiring clinician oversight at every stage. priorities for the field are multicentre externally validated datasets and prospective designs with patient-relevant endpoints, routine use of explainable AI and adherence to established reporting frameworks.

Keywords: Orthodontics, Treatment Planning, Artificial Intelligence, Machine Learning, Deep Learning, Cephalometrics, Clear Aligners

Introduction

Orthodontic treatment planning is a multifactorial clinical process in which a diagnosis and the an individualised treatment plan are arrived by integrating the chief complaint of the patient, clinical examination, cephalometric analysis, dental cast or intraoral scan measurements, and radiographic assessment. All of these inputs need to be interpreted and interpretation introduces variability. Staging of skeletal maturation, cephalometric tracing and the decision to extract or preserve teeth all depended substantially on the operator's training, experience and clinical philosophy, i.e. with the consequence that different clinicians presented with identical records may arrive at materially different treatment plans. this variability is well recognized within the specialty and has long motivated interest in more objective, reproducible methods of analysis [1, 2]. The diagnostic substrate of orthodontics has been transformed over the past two decades. Digital radiography, cone-beam computed tomography, intraoral scanning, and three-dimensional facial imaging have replaced or supplemented their analogue predecessors, generating large volumes of structured and machine-readable clinical data [3]. The almost ubiquitous accumulating of digital records has coincided with significant advances in computing methods that can learn directly from such data, and the coming together of the two has created the ideal condition for AI to enter orthodontic practice. Artificial intelligence is an umbrella term describing computational systems that perform tasks ordinarily requiring human intelligence. It encompasses machine learning (ML) in which algorithms infer patterns from data without being explicitly programmed with rules; artificial neural networks (ANNs) which process information through layers of interconnected computational units; and deep learning (DL), a subset of ML employing multilayered networks that learn hierarchal representations. convolutional neural networks (CNNs), a deep learning architecture optimised for image analysis, have become the dominant approach for image-based tasks in orthodontics [2, 4]. The clinical literature represented in Figureure 1 often confuses the relationship between those terms. The application of these methods becoming increasingly popular to orthodontics use. A scoping review by Monill González et al. mapped the early literature and concluded that AI was already being applied to diagnosis, landmark detection, and treatment decision-making, but the evidence base was heterogeneous and immature [1]. A more recent scoping review by Gracea et al. identified 71 eligible studies and classified them into three domains – diagnostics (n=29), anatomical landmark identification (n=20) and treatment planning (n=22), with Landmark identification the most extensively investigated [5]. Related Syntheses have documented parallel activity in skeletal age determination, malocclusion classification, orthognathic surgical planning, and clear aligner therapy [3, 4, 6]. Despite this volume of work, several question relevant to the practising orthodontist remain incompletely answered. Reported accuracies vary widely between studies, and it is not always clear whether that variation reflects genuine differences in model capability, differences in difficulty of task being performed, or differences in study design and reference standard. Comparisons with human examiners'are inconsistently reported. external validation is frequently absent, so the generalizability of reported performance to a different clinic, scanner, or patient population is often unknown. Ethical and medico legal frameworks governing AI-assisted clinical decisions remain unsettled [7]. Existing reviews have catalogued applications extensively, but fewer have examined systematically why performance differs so markedly between domains. The aim of this narrative review is therefore to evaluate the current role of AI in orthodontic treatment planning by synthesizing the available evidence across principal clinical applications. The specific objectives include both summarizing reported diagnostic and predictive performance in each domain, to examine how performance relates to task structure, dataset composition, and study design etc. to compare AI performance with that of human examiners where such comparisons have been made, and to identify the methodological, ethical, and the regulatory barriers that currently constrain clinical integration that is together with the research priorities most likely to address them.

Materials and Methods

This narrative review was conducted on the basis of a literature search in the PubMed/MEDLINE, Scopus, and Web of Science databases for articles published between January 2009 and June 2026. Search terms were combined using Boolean operators and comprised: “artificial intelligence”, “machine learning”, “deep learning”, “convolutional neural network”, “orthodontics”, “treatment planning”, “cephalometric”, “landmark”, “malocclusion”, “extraction”, “cervical vertebral maturation”, “orthognathic surgery”, “clear aligner”, and “remote monitoring”. Reference lists of all included reviews were hand-searched for further sources. systematic reviews, scoping reviews, meta-analyses, and primary diagnostic accuracy or predictive modelling studies published in English were prioritised. case reports, conference abstracts, editorials, and opinion articles were excluded. Where multiple publications reported the same dataset, the most complete version was retained.

Because this is a narrative rather than a systematic review, no formal risk-of-bias appraisal or meta-analytic pooling was undertaken. Findings are synthesized thematically across the domains in which AI has been applied. All quantitative values are reported as stated by primary authors, with the sample size and study design provided alongside so that readers may judge comparability. Where a Figureure originates in a single study within a review, this is stated explicitly. Manufacturer sources were only consulted to establish the existence and regulatory status of named commercial products and are identified as such.

Basic Concepts of Artificial Intelligence Related to Orthodontics

Machine Learning refers to algorithms that learn patterns from data without explicit rule-based programming. Orthodontics more commonly reports methods including decision trees, random forests, support vector machines, gradient-boosted trees and ensemble or stacking classifiers that combine several base learners. These approaches typically operate on structured, tabular inputs such as cephalometric measurements or model analysis values. Deep learning is a subset of machine learning based on multilayered artificial neural networks. convolutional neural networks the dominant architecture for image-based tasks, use convolutional filters to extract hierarchical spatial features from radiographs, photographs, or 3D scans, beginning with simple edges in early layers and ultimately culminating in complex anatomical conFigureurations in deeper layers. Architectures frequently reported in the orthodontic literature include ResNet, DenseNet, VGG, EfficientNet, Xception and Mask R-CNN [6, 8, 9]. A recurring criticism deep learning in clinical context is opacity: a network may produce an accurate output without any interpretable account of which image features drove it. Explainable AI (XAI) techniques address this by generating visual heatmaps indicating the image regions most influential to a prediction. Their application in orthodontics remains uncommon but is methodologically important, since it is the mechanism by which a clinician can determine whether a model has learned relevant anatomy or an incidental artefact of the training set [9].

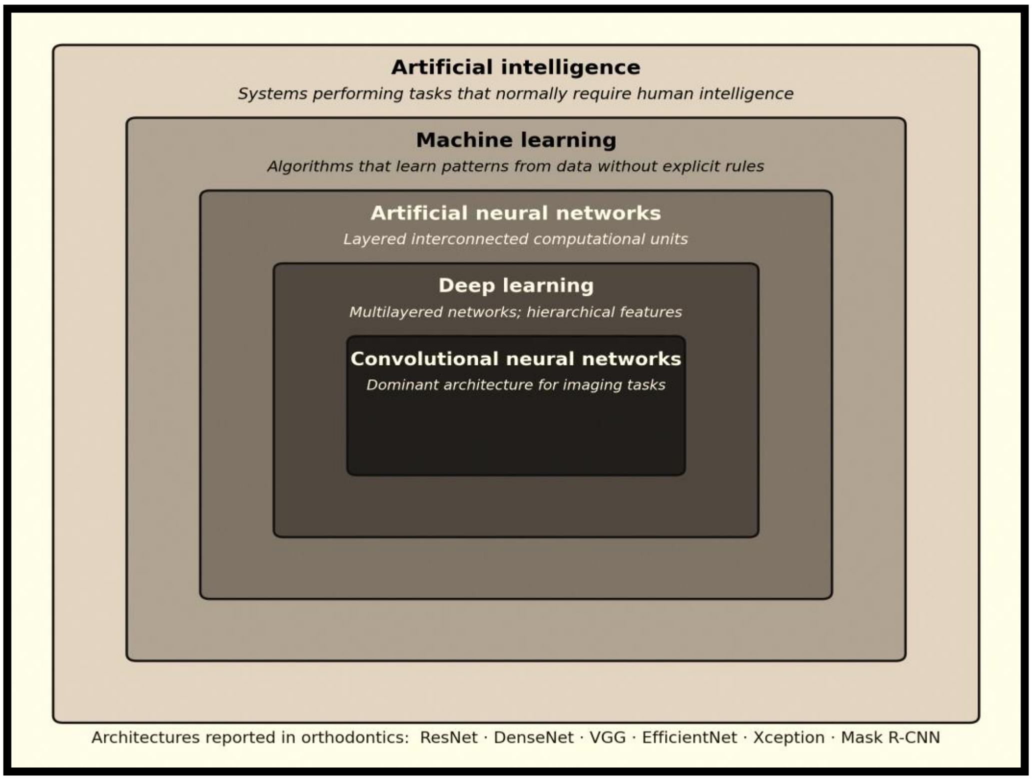

Figure 1. Hierarchy of Artificial intelligence and machine learning techniques in orthodontic literature.

Uses of AI in treatment planning of orthodontic

AI applications have focused on every stage of the orthodontic process including records, diagnosis, planning, outcome simulation and monitoring as shown in Figure 2. The next paragraphs will consider each domain in turn.

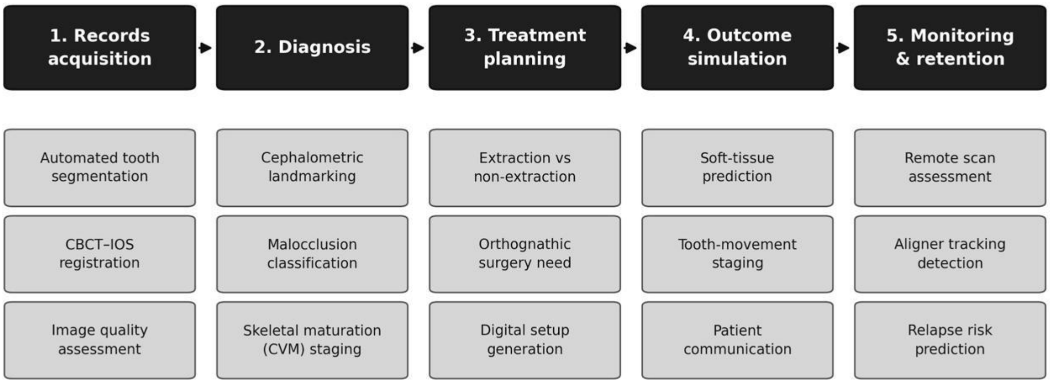

Figure 2. Utilization of artificial intelligence in treatment planning of orthodontic

Although the importance of artificial intelligence but Clinician oversight is still required at every stage of the planning treatment.

A. Cephalometric Assessment and Landmark Recognition

The most studied application in orthodontics of AI is automated cephalometric landmarking, reflecting both the tedium of manual tracing and substantial inter examiner variability it carries [5]. Jiao et al. applied a Mask Region-based CNN to 400 lateral cephalograms, using 320 for training and 80, with 19 manually identified landmarks per image. of the 1,520 landmarks in the test set, 1,494 were detected, giving a detection rate of 98.29%, with mean linear deviation of 2.19 mm from manually marked points. Notably, the per-landmark error ranged from 0.56 mm to 9.51 mm [8]. This

Distribution is clinically important a favourable mean can conceal individual landmarks whose localisation is unreliable, and landmarks differ considerably in how much error a derived measurement will tolerate. For posteroanterior cephalograms, Lee et al. developed a deep learning model trained and validated on 1,032 and compared it against two expert examiners who independently identified 19 landmarks on 82 test images. The model achieved a mean radial error of 1.87 ± 1.53 mm, with a successful detection rates of 34.7%, 67.5% and 91.5% within 1.0, 2.0 and 4.0 mm, respectively, and performance was comparable to that of the human examiners [10]. That fewer than 35% of landmarks fell within 1 mm illustrates the limits of current spatial precision.

At the level of evidence synthesis, Schwendicke et al. screened 321 records and included 19 studies published between 2017 and 2020, all employing CNNs, predominantly on two-dimensional lateral radiographs (n = 15), predominantly using publicly available datasets (n = 12), and testing a mean of 30 landmarks [11]. The heavy reliance on public benchmark collections is a limitation of the domain since performance on standardised research datasets may not transfer to heterogeneous clinical records. The authors Serafin et al. screened 252 papers published between 2020 and 2022, including 15 for qualitative synthesis and 11 for meta-analysis, with high statistical heterogeneity, three-dimentional cephalometric based on cone beam computed tomography reported a pooled mean landmarking error of 2.44 mm, and noted accuracy improved with more recent publication year [12].

B. Classification of malocclusions and support for diagnosis

CNNs applied directly to intraoral and facial photographs can classify malocclusion without radiographic input, offering a potentially radiation-free adjunct for screening and triage.

six CNN architectures were utilized to classify orthodontic photographs by Noeldeke et al. trained and compared DenseNet, EfficientNet, MobileNet, ResNet18, ResNet50, and Xception Photo. The dataset consisted of 676 photographs from 311 orthodontic patients. Among all models, Xception achieved the greatest test accuracy at 98.57% for binary presence of crossbite classification. The model conFigureuration with the best accuracy switched to DenseNet and overall accuracy dropped to 91.43% [13] when the task was extended to three classes, distinguishing lateral from frontal crossbite. Both the decline in accuracy the shift in optimal architecture provide important insights: rankings established on from a binary task do not necessarily transfer to the multiclass version of the same clinical problem.

The VGG-11 model was trained by Koch et al. on 5,266 clinical images which were classified by orthodontic experts into Angle Class I (2,322; 44%), class II (1,880; 36%) and class III (1,064; 20%). Validation and test sets consisted of 474 and 512 images, respectively. The study is distinguished by its application of three explainable-AI methods were applied to find whether classifications made by the model relied on clinically relevant anatomical regions rather than incidental image characteristics [9]. Orthodontic AI literature has not adopted this type of interpretability analysis, deserves wider adoption.

C. Support for decision on extraction or non-extraction

The decision to extract is one of the most important and contentious decisions in orthodontics, and has historically depended on clinician experience and philosophy rather than any single objective algorithm. This could indeed present a case to standardise at least the models they deploy. It does throw up a structural issue, though. All models in this domain are trained against expert decisions. This makes their output a reproduction of existing practice that may be inconsistent. Expert models won’t learn to identify the optimum treatment independently.

Suhail et al. trained several ML models on data from 287 patients, each evaluated independently by five orthodontists. Ensemble methods outperformed more complex architectures including neural networks, and a random forest ensemble achieved performance within the inter-expert disagreement range [14]. The salient finding is that the ceiling is a model that matches the level of agreement between experts has extracted the available signal from its training labels, and additional data alone will not surpass it. Köktürk et al. selected 1,000 patients and trained seven ML models on them. The participants included 500 extraction and 500 non-extraction cases with 36 variables spanning demographic information, cephalometric measurements and model measurements. A stacking classifier combining the gradient-boosted trees, support vector machines, and random forests achieved the highest performance, with an area under the curve of 91.2%.the most influential predictive variables were maxillary and mandibular arch-length discrepancy, Wits appraisal, and ANS-Me length [15]. The clinical coherence of these predicators supports the face validity of the model, because an experienced clinician would weigh all three variables.

D. Diagnosis of skeletal maturity

Due to timing of the pubertal growth spurt informs decisions on interceptive and orthopaedic treatment, automated cervical vertebral maturation (CVM) staging from lateral cephalograms has attracted considerable attention. Mohammed et al. applied a CNN-based multiclass model to approximately 1,200 cephalometric radiographs and 1,200 orthopantomograms, reporting accuracy for CVM staging of up to 98% in males, with second molar calcification proving more accurate in females [16]. Abdulqader et al. conducted a study in which various qualitative and quantitative CVM staging methods were compared using a dataset of 3,600 lateral cephalometric images drawn from six different medical centres. This dataset is one of the few genuinely multicentre datasets in the orthodontics AI literature, divided into training, validation, and test sets in a ratio of 8:1:1. The quantitative approach categorised images into six stages based on measurements of 13 cervical vertebral landmarks. while, the qualitative method assessed the morphology of the vertebrae. The overall staging accuracies of qualitative and quantitative methods were only 71.1% and 78.3% respectively, although the model achieved a landmark successful-detection rate of 97.1% within 1.0 mm [17]. This comparison of the two studies is one of the most illuminating available in this literature. Near perfect landmark localization coexisting with 71–78% staging accuracy demonstrates that the major challenge in CVM assessment lies in the reproducibility of the staging criteria themselves rather than in the measurement of the vertebrae. automating a poorly reproducible classification system yields a fast and consistent version of a poorly reproducible classification. It is also relevant that the lower Figureures stem from the larger, multicentre, more heterogeneous dataset. Which is expected the direction of change when a model is evaluated beyond a single institution. A comprehensive review by Farzanegan et al. conclude that a compatible conclusion that once again emphasizes the accuracy's sensitive dependence on the size of the data, method of labelling and model design. Expert supervision remains unavoidable [18].

E. Planning of orthognathic surgery and prediction of outcomes

In case of borderline patients, the choice between orthodontic camouflage and combined orthodontic-surgical treatment is uncertain, AI has been applied both to the decision itself and the prediction of postoperative facial soft-tissue morphology. de Oliveira et al. analysed 920 lateral radiographs of patients treated with conventional orthodontics or with orthognathic surgery, comprising 558 Class II and 362 Class III patients. They found that a combined machine learning technique accurately was able to predict the requirement of orthognathic surgery. It performed better performance in Class III than Class II patients [19]. The asymmetry is clinically plausibile since a Class III skeletal discrepancy is less amenable to dentoalveolar camouflage. As a result, the decision boundary is sharper. To predict soft-tissue outcome, Knoops et al. developed a three-dimensional morphable model trained and validated on 4,261 faces of healthy volunteers and patients of orthognathic surgery. The framework classified patients requiring surgery with 95.5% sensitivity and 95.2% specificity, and simulated surgical outcomes with a mean accuracy of 1.1 ± 0.3 mm, comparable to conventional computer-assisted planning methods [20]. The Figureures most commonly quoted in the literature originated from single study only. The scoping review of AI for post orthognathic facial soft tissue by Almarhoumi involved prediction screening 1,579 records and only included seven studies that were published between 2009 and 2023. Moreover, all these studies were retrospective in design with sample sizes ranging from five to 137 patients [21]. The review performed no meta-analysis; sensitivity and specificity were not reported at all in four of the seven included studies. The chin (jaw) is a highly predictable region. Similarly, prediction of the lower face is reasonable. However, the midface and nasal base have a poor predictive value. Overall, mean prediction errors reported in the included studies ranged from approximately 0.16 mm to 3.08 mm. The variation in errors depended on the facial region and method but were generally more accurate for the chin and lower face compared with the midface and nasal base. The surgical pathway is being expanded by the application of AI, expanding its usage according to broader syntheses. Though there are repeated limitations in each of these studies. A scoping review published in 2026 mapped the literature from 2017 to May 2025 and included 62 studies covering cephalometric analysis, virtual surgical planning, outcome prediction, and intraoperative support. The review also highlighted persistent ethical, legal and validation gaps [22]. According to 60 articles analysed by Wang et al., the literature lacks sufficient data regarding overfitting and limited interpretability associated with deep-learning methods [23].

F. Integrating clear aligner therapy and digital workflow

AI and machine learning technology are incorporated at several stages of clear aligners planning. These include automated tooth segmentation from intraoral scans, registration of digital models and generation of the digital setup, and prediction of tooth movement. Ruiz et al. conducted a scoping review screening 708 studies and included 41. Out of 41 these 16 addressed tooth segmentation while 4 registration of digital models. Additionally, 13 digital setups and 8 remote monitoring were included in this scoping review. AI could segment teeth with accuracy of 98%, and AI effectively merged cone beam computed tomography and intraoral scan data. The review assessed 13 aligner software programs available on the market for their level of automation. Finding that only one was reported to show complete automation of the orthodontic digital workflow. Also, none of the 13 marketed programs has been evaluated in any of the 41 included studies [24]. Preda et al. reached a compatible conclusion from the treatment-planning side. They have identifying seven commercial AI tools and observing mismatch between academically validated tools and commercially available systems, which generally lack published validation [6]. The distinction becomes relevant when one interprets the substantial literature that compares predicted and achieved tooth movement in aligner therapy. The research studies evaluate the output of a commercial integrated planning pipeline rather than an isolated algorithmic component, so their results should be understood as describing predictability of aligner treatment rather than performance of AI as such. Kravitz et al assessed a sample of 37 patients and 401 anterior teeth treated with Anterior Invisalign and reported an anterior tooth movement mean accuracy of 41% [25], a Figureure often inappropriately generalised to whole arch accuracy. A retrospective evaluation by Alswajy et al. involving 206 patients from three private practices was carried out with an overall predicted accuracy of 76.85%. The lowest reliability was related to arch-length parameters upper 38.79% and lower 30.02% [26]. Rossini et al. synthesising 11 studies, identified extrusion to be the least controllable movement (30% accuracy) and upper molar distalisation as most predictable (88%) [27], with subsequent systematic reviews had compatible conclusions [28, 29].

In extraction cases, systematic discrepancies occurred between planned and achieved movement. According to Dai et al., the expected anchorage control of the first molar and retraction of the central incisor was not fully achieved in 30 patients treated with maxillary first premolar extraction [30]. Furthermore, in 33 adults treated with four first premolar extractions, it was established that the first molars achieved greater mesial tipping and buccal inclination. The first molars also exhibited less constriction, mesial-lingual rotation, mesial displacement, and intrusion than planned [31]. Ren et al. studying 31 extraction patients reported about the undesired movements as the first molar-mesial movement of 2.2 mm, mesial tipping of 5.4° and intrusion of 0.45 mm, Canine distal tipping of 11.0°, lingual tipping of 4.4° and distal rotation of 4.9°, and Incisor-lingual tipping of 10.6° with extrusion of 1.5 mm [32]. To assess the clinical effectiveness of the new versatile aligner system (Invisalign Lite), Huanca Ghislanzoni et al, prospectively evaluated 21 subjects whose treatment was conducted over a period of 28 weeks. The trial reported significant differences between the predicted and achieved values for torque correction of the maxillary central incisors and for rotational correction of most teeth [33]. Patient-facing outcome visualisation is a related but different application. Mourgues et al. evaluated an AI-based smile simulation tool in 51 subjects, and study objective was set to verify whether simulated outcomes confirmed to accepted orthodontic aesthetic criteria and whether the software modified the anatomy of the teeth [34]. These tools are meant for communication rather than treatment planning, and findings relating to these tools shouldn't be extrapolated to clinical planning software.

G. Remote health monitoring and teleorthodontics

The integration of AI and remote monitoring is the most recent extension of these techniques to clinical practice, and is one of the few domains in which a marketed system has been evaluated independently. According to Polizzi et al., the objective of the scoping review was to examine existing artificial intelligence (AI) applications in teleorthodontics, examining the use of AI-assisted platforms for remote assessment of the treatment progress, appointment triage, as well as patient engagement [35]. A multicentre retrospective study of a commercial AI remote monitoring system was subsequently reported by McCray et al., analysing 3,323 assessments from 623 patients treated at multiple sites in the United States. For a binary decision on the seating of an aligner, the system achieved a sensitivity of 93.2%, a specificity of 86.2%, a positive predictive value of 89.2%, and a negative predictive value of 94.4% [36].

In this instance, the high negative predictive value is the clinically meaningful Figureure, indicating that when the system reports adequate tracking, that is a reliable report; which is the property required of a triage tool that aims to reduce unnecessary appointments. The lower specificity indicates a tendency toward false positive of unseating. In practice, this results in an additional review rather than missed pathology. This is an acceptable error profile for this application. However, it limits the extent to which chairside time can be reduced.

Table 1. Summary of utilization AI in treatment planning of orthodontic

Field	Reference	Sample	Method	Remarks
Landmarking, 2D lateral	[8]	19 landmarks ,400 cephalograms	R-CNN	Detection at 98.29%, average error of 2.19 mm (range 0.56–9.51)
Landmarking, 2D PA	[10]	82 test at 1,032 images	Deep learning	MRE 1.87 ± 1.53 mm; SDR 34.7, 67.5,91.5% at 1,2 ,4 mm.
Landmarking, 3D CBCT	[12]	11 studies (meta-analysis)	Various DL	high heterogeneity, Pooled error 2.44 mm
Malocclusion classification	[13]	311 patients, 676 photographs	Six CNN architectures	91.43% three-class (DenseNet), 98.57% binary (Xception).
Malocclusion classification	[9]	5,266 images	Visual Geometry Group-11 with three XAI techniques	Categorization of Angles class with Interpretability Verification
Extraction decision	[14]	287 patients	Ensemble / random forest	Performance is within the range of inter-expert agreement
Extraction decision	[15]	36 variables at 1000 patients	Stacking classifier	Key predictors arch-length discrepancy, Wits, ANS-Me. AUC 91.2 %.
Skeletal maturation	[16]	Approximately 1200 cephalograms with 1200 OPG radiographs	CNN multiclass	CVM staging in males have accuracy rates near 98%
Skeletal maturation	[17]	3600 images, six centres	DL qualitative and quantitative	At 1 mm SDR 97.1%, staging 71.1% / 78.3%

Orthognathic decision	[19]	920 radiographs includes:558 II and 362 III.	Combined ML	Prediction of surgical need is accurate and better in Class III.
Orthognathic outcome	[20]	4261 faces	3D morphable model	Accuracy (1.1± 0.3)mm, Sensitivity 95.5% and specificity 95.2%.
Clear aligner workflow	[24]	13 software programs from 41 studies.	Scoping review	Segmentation 98%, 1 of 13 programs fully automated
Aligner predictability	[26]	206 patients, three practices	Predicted and achieved	Overall 76.85%; arch length 38.79% ,30.02%.
Remote monitoring	[36]	623 patients, 3,323 assessments.	Commercial AI platform	NPV94.4%, Sensitivity 93.2% and specificity 86.2%.

Comparison of clinician and diagnostic performance

An analysis of performance across domains, organized according to task structure rather than clinical domain, reveals a consistent pattern (Figure. 3). The models achieve 91-99% accuracy when the ground truth is clearly defined in binary classification and segmentation tasks. For multiclass problems, the performance drops to between 71-91% and that on fine-grained spatial tasks that require millimetre level localisation, the performance is the lowest, where successful detection rate within 2 mm is only 67.5% for posteroanterior landmarks and the pooled three-dimensional landmarking error remains at 2.44 mm.

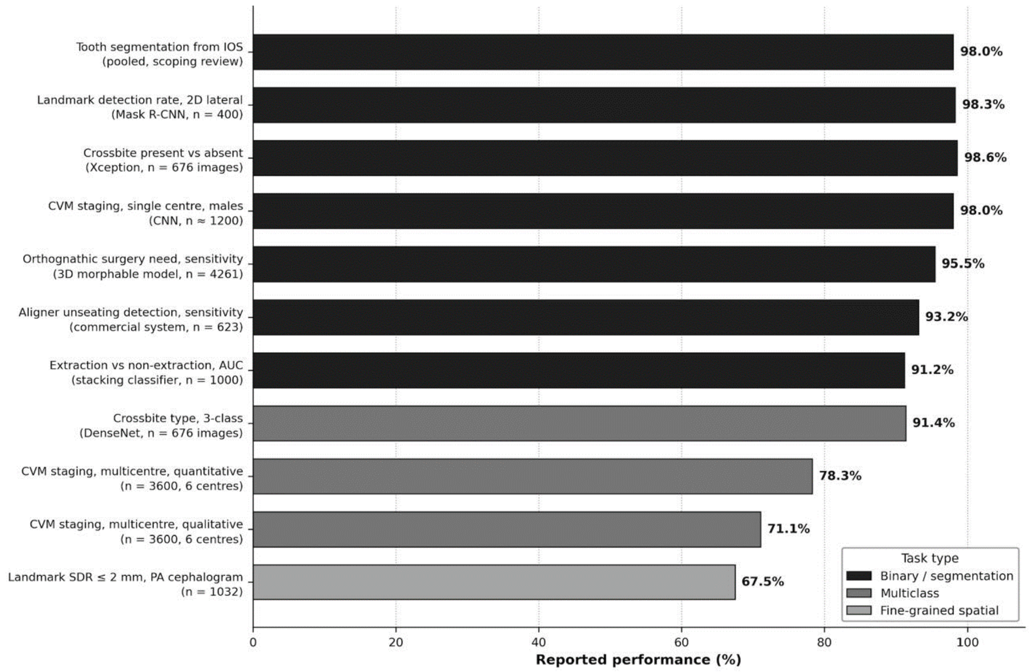

Figure 3. Reported Performance of AI models in Orthodontic Tasks.

The gradient is a property to the structure of the tasks rather than clinical importance. Moreover, it's practically related to AI tasks: the tasks that AI performs best are frequently not the tasks that dictate treatment outcome. It is straightforward to diagnose the binary cross bite detection accurately. Individualised prediction of soft-tissue response to surgery is clinically decisive but the least reliable one.

Dataset composition is the second determinant. Where same task performed for A multiple centre and single centre, the result revealed that multicentre Figureure is consistently lower for clinical evaluation. Obviously in CVM staging accuracy was 98% in a single-centre cohort but fell to a multicentre range of 71.1% to 78.3% in a six centre cohort for 3,600 images [16, 17]. The upper bound rather than the expected clinical value should be interpreted as single-centre performance.

In many studies, performance against human examiners was assessed, and it was found that this was statistically roughly comparable to, and in narrowly constrained tasks in some ways better than, that of experts [9, 10, 13]. Nonetheless, in their broader scoping and systematic reviews, it was concluded that AI has not yet surpassed experienced clinicians across the full breadth of orthodontic diagnostic and planning tasks, and the supervision of human remains essential at all stages of treatment [1, 5, 6]. The most robustly demonstrated benefits include reductions in analysis time and in inter- and intra-examiner variability, rather than unambiguous superiority in diagnostic accuracy.

Table 2. Comparison between human examiner and AI performance.

Domain and Reference	Human examiner performance	AI performance	Remarks
PA cephalometric landmarking [10]	Two specialist examiners, 19 landmarks	MRE (1.87 ± 1.53) mm	Comparable
Extraction decision [14]	5 orthodontists, independent assessment	Ensemble Random forest	In the range of inter-expert agreement
Orthognathic surgical need [20]	Conventional computer-assisted planning	Sensitivity 95.5%and specificity 95.2%	Similar precision (1.1 mm)
Angle classification [9]	Expert-assigned reference labels	VGG-11 with XAI verification	Classification based on relevant anatomy
CVM staging [17]	Rater-dependent conventional staging	71.1–78.3% (six centres)	Limited by reference standard
Analysis time [17]	Manual expert assessment	Automated batch processing	Time saving is most important benefit

Restrictions, Ethical and leagal consideration.

The clinical translation is still limited by various methodological and ethical constraints. datasets homogeneity and bias in algorithms. Most published models are developed and validated on single-centre datasets, often from homogenous populations. This brings up valid concerns regarding generalisability and systematic underperformance in groups of patients underrepresented in training data. Craniofacial morphology differs greatly across populations. As such, a model that was trained predominantly on one population may contain measurable bias when applied to another.

Absence of the external validation, sample sizes in many studies remain modest and there is often no validation on data from an independent centre, scanner or population [5]. The repeated drop in performance when we test these models in multiple centres shows that it is not a theoretical concern.

Inconsistent reporting. Challenges arise in the comparison and pooling of studies due to the heterogeneous reporting of metrics, reference standards, and study design. This is reflected in the high statistical heterogeneity of the meta-analysis of three-dimensional landmarking [12]. Typically, structures reporting frameworks are available, including the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [37] and dental-specific checklists for authors, reviewers and readers [38], but adherence remains inconsistent.

Limitations of reference standards. In various fields, the performance limit on models is set not by what an algorithm can do, but rather by how reproducible the human judgement used as ground truth is. This applies to CVM staging, where the classification system itself is rater-dependent [17, 18], and to extraction decision-making, where models are trained against expert opinions that themselves disagree [14].

Interpretability. Clinical trust and regulatory approval are limited due to the opacity of deep learning models. There currently exist explainable AI methodologies that have been utilised quite positively [9], but are not yet commonplace.

Uncertainty in ethics and medico legal. Mörch et al. analysed AI ethics in dentistry via a scoping review. It was determined that issues were clustered in concerns of prudence, equity, privacy, professional responsibility, and democratic participation, and noted that ethical issues are inconsistent relative to the pace of technical development [7]

There are many questions remain unanswered, including how informed consent will be obtained if treatment decisions are made with the assistance of AI, who governs patient imaging data that is used to train the model, and Who will bear the professional and legal responsibility if the treatment decision that was guided by AI results in a suboptimal outcome. Regulatory clearance of software as a medical device demonstrates compliance with safety and intended-use requirements but does not establish comparative clinical effectiveness [39].

Table 3. Strategies to mitigate limitations in orthodontic Artificial intelligence research

Limitation	Evidence	Mitigation strategy
Single-centre, homogeneous datasets	The accuracy level drops from nearly 98% to between 71 and 78 % when tested across six centres [16, 17]	Multicentre, demographically diverse training and test sets
Absent external validation	Frequently unreported across domains [5]	External test sets assert important clinical readiness as the requirement.
Reliance on public benchmark data	12 of 19 landmarking studies used public datasets [11]	Evaluation on heterogeneous routine clinical records
Fluctuating report.	meta-analysis with high heterogeneity [12]	Abide by CLAIM [37] and dental AI checklists [38]
Non-reproducible reference standards	Decisions regarding CVM staging and extraction rely on the rater. [14, 17, 18]	Reform of classification criteria, agreement on the truth.
Model opacity	Interpretability rarely assessed	Routine explainable AI, as applied by Koch et al. [9]
Surrogate endpoints	Accuracy reported	Planning time, treatment duration, refinements, occlusal indices
Commercial systems unvalidated	None of 13 aligner programs was evaluated in 41 studies [24]	Independent evaluation of marketed software
Unresolved ethics and liability	As addressed in literature [7], the consistency of the data is Inconsistently.	Schemes designed to guarantee consent transfer, data governance, accountability

Future Directions

The following is a summary of priorities identified in the literature:

- Demographically diverse datasets from multicenters help in establishing generalizability and detecting algorithmic bias. Data-sharing initiatives could significantly speed up this process.
- The standard of evidence which establishes clinical readiness should progressively shift from retrospective single-centre accuracy studies to prospective external validation studies.

- Diagnostic accuracy is to be supplemented with patient-relevant endpoints. A tool is considered beneficial to patients based on planning time, the number of aligner refinements, overall duration of the treatment, occlusal outcome indices and patient-reported experience.
- It is important to ensure routine explainability to facilitate clinician trust as well as regulatory evaluation of the model.
- Using established frameworks like CLAIM for standardized reporting would enhance comparability and enable meaningful meta-analysis.
- Modifying the reference standard can yield more benefits than refining algorithms in domains where human ground truth is the limiting factor, particularly for CVM staging and extraction decision-making.
- Independent evaluation of commercial systems remains a significant outstanding need, given that the software routinely used in the clinic is mostly different from the models reported in the literature. [6, 24]

Conclusion

Artificial intelligence has established a wide-ranging evidence base for nearly every domain of orthodontic treatment planning. The results show strong performance for binary classification and segmentation tasks, a moderate performance for multiclass problems, and the weakest performance on fine-grained spatial and individualised predictive tasks that most directly influence treatment outcome. Performance declines consistently when models are evaluated across multiple centres, indicating that single-centre results should be interpreted as an upper bound. The most definitely demonstrated benefits are improved time efficiency and reduced inter-observer variability rather than clear diagnostic superiority over expert clinicians. Routine uptake continues to be constrained by limited external validation, homogeneous datasets, inconsistent reporting, reference standards that lack reproducibility, and unresolved ethical and medico-legal issues.

Artificial intelligence in orthodontics, based on present evidence, is best described as a decision-support adjunct, with judgement and accountability remaining with treating clinician. The realization of its potential will depend on multicentre validation, standardized reporting, and explainable model design, and independent evaluation of systems that are deployed in clinical practice.

References

[1] A. Monill-González, L. Rovira-Calatayud, N. G. d'Oliveira, and J. M. Ustrell-Torrent, "Artificial intelligence in orthodontics: Where are we now? A scoping review," *Orthodontics & Craniofacial Research*, vol. 24, pp. 6–15, 2021, doi: 10.1111/ocr.12517.

[2] N. Kazimierczak, W. Kazimierczak, Z. Serafin, P. Nowicki, J. Nożewski, and J. Janiszewska-Olszowska, "AI in orthodontics: Revolutionizing diagnostics and treatment planning—A comprehensive review," *Journal of Clinical Medicine*, vol. 13, no. 2, Art. no. 344, 2024, doi: 10.3390/jcm13020344.

[3] S. B. Khanagar *et al*., "Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making: A systematic review," *Journal of Dental Sciences*, vol. 16, no. 1, pp. 482–492, 2021, doi: 10.1016/j.jds.2020.05.022.

[4] S. Azizi, S. Hatampoor, and S. Tahamtan, "Applications of artificial intelligence in diagnosis and treatment planning of orthodontics: A narrative review," *The Saudi Dental Journal*, vol. 37, no. 7, Art. no. 70, 2025, doi: 10.1007/s44445-025-00077-0.

[5] R. S. Gracea *et al*., "Artificial intelligence for orthodontic diagnosis and treatment planning: A scoping review," *Journal of Dentistry*, vol. 152, Art. no. 105442, 2025, doi: 10.1016/j.jdent.2024.105442.

[6] F. Preda, N. Elshazly, and R. Jacobs, “AI-enhanced orthodontic treatment planning: A scoping review on evidence-based clinical application with commercial software overview,” *Journal of Dentistry*, Art. no. 106112, 2025, doi: 10.1016/j.jdent.2025.106112.

[7] C. M. Mörch *et al.*, “Artificial intelligence and ethics in dentistry: A scoping review,” *Journal of Dental Research*, vol. 100, no. 13, pp. 1452–1460, 2021, doi: 10.1177/00220345211013808.

[8] Z. Jiao, Z. Liang, Q. Liao, S. Chen, H. Yang, G. Hong, and H. Gui, “Deep learning for automatic detection of cephalometric landmarks on lateral cephalometric radiographs using the Mask Region-based Convolutional Neural Network: A pilot study,” *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 137, no. 5, pp. 554–562, 2024, doi: 10.1016/j.oooo.2024.02.003.

[9] P. J. Koch, J. E. Cejudo Grano de Oro, M. Büttner, L. T. Arsiwala-Scheppach, J. De Geer, H. Meyer-Lueckel, and F. Schwendicke, “Deep learning for Angle classification based on intraoral photographs: An interpretability perspective,” *BMC Oral Health*, vol. 26, no. 1, Art. no. 210, 2026, doi: 10.1186/s12903-025-07550-6.

[10] H. Lee, J. M. Cho, S. Ryu, S. Ryu, E. Chang, Y. S. Jung, and J. Y. Kim, “Automatic identification of posteroanterior cephalometric landmarks using a novel deep learning algorithm: A comparative study with human experts,” *Scientific Reports*, vol. 13, no. 1, Art. no. 15506, 2023, doi: 10.1038/s41598-023-42870-z.

[11] F. Schwendicke *et al.*, “Deep learning for cephalometric landmark detection: Systematic review and meta-analysis,” *Clinical Oral Investigations*, vol. 25, no. 7, pp. 4299–4309, 2021, doi: 10.1007/s00784-021-03990-w.

[12] M. Serafin *et al.*, “Accuracy of automated 3D cephalometric landmarks by deep learning algorithms: Systematic review and meta-analysis,” *La Radiologia Medica*, vol. 128, no. 5, pp. 544–555, 2023, doi: 10.1007/s11547-023-01629-2.

[13] B. Noeldeke, S. Vassis, M. Sefidroodi, R. Pauwels, and P. Stoustrup, “Comparison of deep learning models to detect crossbites on 2D intraoral photographs,” *Head & Face Medicine*, vol. 20, no. 1, Art. no. 45, 2024, doi: 10.1186/s13005-024-00448-8.

[14] Y. Suhail, M. Upadhyay, A. Chhibber, and Kshitiz, “Machine learning for the diagnosis of orthodontic extractions: A computational analysis using ensemble learning,” *Bioengineering*, vol. 7, no. 2, Art. no. 55, 2020, doi: 10.3390/bioengineering7020055.

[15] B. Köktürk, H. Pamukçu, and Ö. Gözüaçık, “Evaluation of different machine learning algorithms for extraction decision in orthodontic treatment,” *Orthodontics & Craniofacial Research*, vol. 27, pp. 13–24, 2024, doi: 10.1111/ocr.12811.

[16] M. H. Mohammed, Z. Q. Omer, B. B. Aziz, J. F. Abdulkareem, T. M. A. Mahmood, F. A. Kareem, and D. N. Mohammad, “Convolutional neural network-based deep learning methods for skeletal growth prediction in dental patients,” *Journal of Imaging*, vol. 10, no. 11, Art. no. 278, 2024, doi: 10.3390/jimaging10110278.

[17] A. A. Abdulqader, F. Jiang, B. S. Almaqrami, F. Cheng, J. Yu, Y. Qiu, and J. Li, “Deep learning approaches for quantitative and qualitative assessment of cervical vertebral maturation staging systems,” *PLOS ONE*, vol. 20, no. 5, Art. no. e0323776, 2025, doi: 10.1371/journal.pone.0323776.

[18] P. Farzanegan, M. Zarabadi, S. Najary, M. A. Tahmasbi, and M. Behnaz, “The role of artificial intelligence in orthodontics for determining skeletal age based on cervical vertebra maturation degree: A comprehensive review,” *Health Science Reports*, vol. 8, no. 11, Art. no. e71487, 2025, doi: 10.1002/hsr2.71487.

[19] P. H. J. de Oliveira *et al.*, “Artificial intelligence as a prediction tool for orthognathic surgery assessment,” *Orthodontics & Craniofacial Research*, vol. 27, no. 5, pp. 785–794, 2024, doi: 10.1111/ocr.12805.

[20] P. G. Knoops *et al.*, "A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery," *Scientific Reports*, vol. 9, no. 1, Art. no. 13597, 2019, doi: 10.1038/s41598-019-49506-1.

[21] A. A. Almarhoumi, "Accuracy of artificial intelligence in predicting facial changes post-orthognathic surgery: A comprehensive scoping review," *Journal of Clinical and Experimental Dentistry*, vol. 16, no. 5, Art. no. e624, 2024, doi: 10.4317/jced.61500.

[22] K. Janáková *et al.*, "The role of artificial intelligence in orthognathic surgery: A scoping review," *Dentistry Journal*, vol. 14, no. 5, Art. no. 286, 2026, doi: 10.3390/dj14050286.

[23] T. Wang, H. Chen, G. Song, and B. Han, "Traditional and artificial intelligent methods in predicting maxillofacial soft tissue morphology after orthognathic surgery: A narrative review," *International Journal of Dentistry*, vol. 2025, no. 1, Art. no. 6268492, 2025, doi: 10.1155/ijod/6268492.

[24] D. C. Ruiz, S. Mureșanu, X. Du, B. M. Elgarba, R. C. Fontenele, and R. Jacobs, "Unveiling the role of artificial intelligence applied to clear aligner therapy: A scoping review," *Journal of Dentistry*, vol. 154, Art. no. 105564, 2025, doi: 10.1016/j.jdent.2025.105564.

[25] N. D. Kravitz, B. Kusnoto, E. BeGole, A. Obrez, and B. Agran, "How well does Invisalign work? A prospective clinical study evaluating the efficacy of tooth movement with Invisalign," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 135, no. 1, pp. 27–35, 2009, doi: 10.1016/j.ajodo.2007.05.018.

[26] W. Alswajy, H. Baeshen, G. Al-Turki, and F. Alsulaimani, "The reliability of ClinCheck® accuracy before and after Invisalign® treatment—A multicenter retrospective study," *Applied Sciences*, vol. 13, no. 8, Art. no. 4670, 2023, doi: 10.3390/app13084670.

[27] G. Rossini, S. Parrini, T. Castroflorio, A. Deregibus, and C. L. Debernardi, "Efficacy of clear aligners in controlling orthodontic tooth movement: A systematic review," *The Angle Orthodontist*, vol. 85, no. 5, pp. 881–889, 2015, doi: 10.2319/061614-436.1.

[28] L. Galan-Lopez, J. Barcia-Gonzalez, and E. Plasencia, "A systematic review of the accuracy and efficiency of dental movements with Invisalign®," *The Korean Journal of Orthodontics*, vol. 49, no. 3, pp. 140–149, 2019, doi: 10.4041/kjod.2019.49.3.140.

[29] A. Papadimitriou, S. Mousoulea, N. Gkantidis, and D. Kloukos, "Clinical effectiveness of Invisalign® orthodontic treatment: A systematic review," *Progress in Orthodontics*, vol. 19, no. 1, Art. no. 37, 2018, doi: 10.1186/s40510-018-0235-z.

[30] F. F. Dai, T. M. Xu, and G. Shu, "Comparison of achieved and predicted tooth movement of maxillary first molars and central incisors: First premolar extraction treatment with Invisalign," *The Angle Orthodontist*, vol. 89, no. 5, pp. 679–687, 2019, doi: 10.2319/090418-646.1.

[31] F. F. Dai, T. M. Xu, and G. Shu, "Comparison of achieved and predicted crown movement in adults after 4 first premolar extraction treatment with Invisalign," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 160, no. 6, pp. 805–813, 2021, doi: 10.1016/j.ajodo.2020.06.041.

[32] L. Ren *et al.*, "The predictability of orthodontic tooth movements through clear aligner among first-premolar extraction patients: A multivariate analysis," *Progress in Orthodontics*, vol. 23, no. 1, Art. no. 52, 2022, doi: 10.1186/s40510-022-00447-y.

[33] L. H. Ghislanzoni, Z. Kalemaj, M. Manuelli, C. Magni, A. Polimeni, and A. Lucchese, "How well does Invisalign ClinCheck predict actual results: A prospective study," *Orthodontics & Craniofacial Research*, vol. 27, no. 3, pp. 465–473, 2024, doi: 10.1111/ocr.12752.

[34] T. Mourgues, M. J. González-Olmo, L. Huanca Ghislanzoni, C. Peñacoba, and M. Romero-Maroto, "Artificial intelligence in aesthetic dentistry: Is treatment with aligners clinically realistic?" *Journal of Clinical Medicine*, vol. 13, no. 20, Art. no. 6074, 2024, doi: 10.3390/jcm13206074.

[35] A. Polizzi, S. Serra, R. Leonardi, and G. Isola, "Clinical applications of artificial intelligence in teleorthodontics: A scoping review," *Medicina*, vol. 61, no. 7, Art. no. 1141, 2025, doi: 10.3390/medicina61071141.

[36] J. F. McCray *et al.*, "Accuracy of DentalMonitoring's artificial intelligence in detecting aligner tracking issues: A retrospective multi-centric study," *BMC Oral Health*, vol. 26, no. 1, Art. no. 239, 2026, doi: 10.1186/s12903-025-07580-0.

[37] J. Mongan, L. Moy, and C. E. Kahn Jr., "Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers," *Radiology: Artificial Intelligence*, vol. 2, no. 2, Art. no. e200029, 2020, doi: 10.1148/ryai.2020200029.

[38] F. Schwendicke *et al.*, "Artificial intelligence in dental research: Checklist for authors, reviewers, readers," *Journal of Dentistry*, vol. 107, Art. no. 103610, 2021.

[39] Align Technology, "Align Technology launches Align X-ray Insights, an AI computer-aided detection software, in the European Union and United Kingdom," press release, Mar. 10, 2025.