Joko Kusnoto FKG

Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

Artikel 1

Document Details

Submission ID	trn:oid:::3618:138250528
Submission Date	May 9, 2026, 9:09 AM GMT+7
Download Date	May 9, 2026, 9:16 AM GMT+7
File Name	Accuracy of Orthodontic Malocclusion Detection - COMPLETE.pdf
File Size	27.5 MB

18 Pages

7,771 Words

44,539 Characters

9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 15 words)

Exclusions

- 1 Excluded Source

Match Groups

18 Not Cited or Quoted 9%
Matches with neither in-text citation nor quotation marks

1 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

8%	Internet sources
3%	Publications
3%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

18 Not Cited or Quoted 9%
Matches with neither in-text citation nor quotation marks

1 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

8%	Internet sources
3%	Publications
3%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	**www.e-hir.org**	**3%**
2	Internet	**repository.unair.ac.id**	**2%**
3	Student papers	**University of Florida on 2023-04-12**	**<1%**
4	Internet	**academic.oup.com**	**<1%**
5	Student papers	**Universitas Trisakti on 2025-09-11**	**<1%**
6	Publication	**Syeda Leena Mumtaz, Abdulrahim Shamayleh, Hussam Alshraideh, Adnane Guell...**	**<1%**
7	Internet	**assets.ctfassets.net**	**<1%**
8	Internet	**www.journals.aiac.org.au**	**<1%**
9	Publication	**U Venkatesh, Periyasamy Aravind Gandhi. "Prediction of COVID-19 Outbreaks Usi...**	**<1%**
10	Internet	**www.mdpi.com**	**<1%**

11 Internet	
alzres.biomedcentral.com	<1%
12 Internet	
pdfs.semanticscholar.org	<1%
13 Internet	
www.researchsquare.com	<1%
14 Internet	
www.scirp.org	<1%

www.e-hir.org

pISSN: 2093-3681
eISSN: 2093-369X

HIR

Healthcare Informatics Research

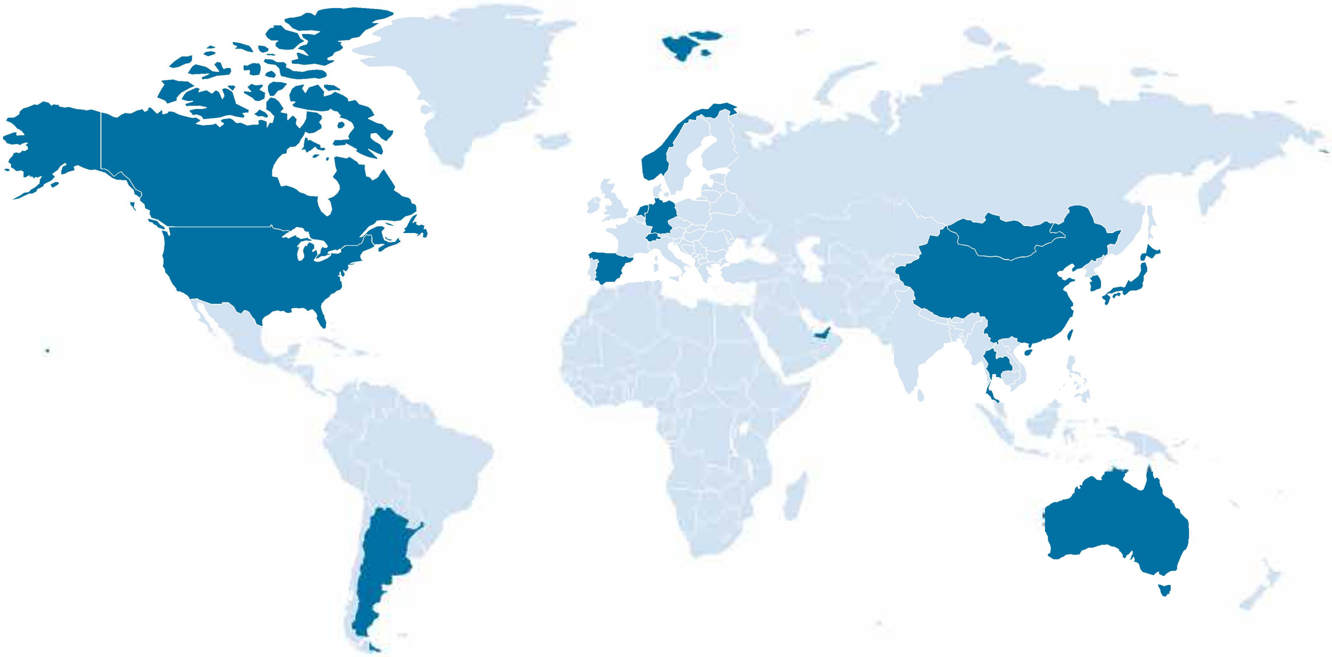
Vol.32 • No.2 • April 2026

Special Issue: Data Science for Healthcare

HIR Healthcare Informatics Research Vol. 32 • No. 2 • April 2026 pages 115-200

The Korean Society of
Medical Informatics

Editorial board



Editor

Hyejung Chang	*Kyung Hee University School of Management, Korea*
David Westfall Bates	*Harvard University School of Medicine, USA*

Editorial Member

Elizabeth Borycki	*University of Victoria School of Health Information Science, Canada*
Young Moon Chae	*Mongolian National University of Medical Sciences, Mongolia*
Jeeyae Choi	*University of North Carolina Wilmington School of Nursing, USA*
Jinwook Choi	*Seoul National University College of Medicine, Korea*
Mona Choi	*Yonsei University College of Nursing, Korea*

HIR Healthcare Informatics Research

Ju Han Kim	*Seoul National University College of Medicine, Korea*
Jung A Kim	*Hanyang University School of Nursing, Korea*
Kwang Gi Kim	*Gachon University College of Medicine, Korea*
Minkyung Kim	*RAK hospital, United Arab Emirates*
Michio Kimura	*Hamamatsu University School of Medicine, Japan*
Hyoun-Joong Kong	*Seoul National University Hospital, Korea*
Kyehwa Lee	*Asan Medical Center, Korea*
Daniel Luna	*Hospital Italiano de Buenos Aires, Argentina*
Shuqian Luo	*Capital University of Medical Sciences Medical Imaging Laboratory, China*
Fernando Martin-Sanchez	*Institute of Health Carlos III, Spain*
Rae-Woong Park	*Ajou University College of Medicine, Korea*
Charles Safran	*Harvard University School of Medicine, USA*
Nawanan Theera-Ampornpunt	*Mahidol University Ramathibodi Hospital, Thailand*
Alfred Winter	*University of Leipzig Institute of Medical Informatics, Statistics and Epidemiology, Germany*
Joo Heung Yoon	*University of Pittsburgh Medical Center, USA*

Statistics Editor

Hyeoun-Ae Park	*Seoul National University College of Nursing, Korea*

Ethics Editor

Eun-Young Kim	*Inje University College of Medicine, Korea*
In Ho Kwon	*Dong-A University College of Medicine, Korea*

Editorial Assistant

Soyeong Park	*KOSMI, Korea*

Byung-Chul Chang	*CHA Bundang Medical Center, Korea* 🇰🇷
Polun Chang	*National Yang-Ming University Institute of Biomedical Informatics, Taiwan* 🇹🇼
W. Ed Hammond	*Duke University Center for Health Informatics, USA* 🇺🇸
Sun Huh	*Hallym University College of Medicine, Korea* 🇰🇷
Myeng-Ki Kim	*Seoul National University College of Dentistry, Korea* 🇰🇷
Suk Wha Kim	*CHA Bundang Medical Center, Korea* 🇰🇷
Kwok Chan Lun	*National University of Singapore Department of Information Systems, Singapore* 🇸🇬
Ronald C. Merrell	*Virginia Commonwealth University School of Medicine, USA* 🇺🇸
Anne Moen	*University of Oslo Department of Nursing Science, Norway* 🇳🇴
Chun-Por Wong	*Ruttonjee & Tang Shiu Kin Hospital Integrated Medical Service, Hong Kong* 🇭🇰

Editorial Office
608, 6th floor, Twin City Namsan Office Building, 366 Hangang-daero, Yongsan-gu, Seoul, 04323, Korea
Tel: +82-2-733-7637, +82-2-734-7637 E-mail: hir@kosmi.org

pISSN 2093-3681
eISSN 2093-369X

Vol. 32 • No. 2 • April 2026

HIR
Healthcare Informatics Research

Editorial

Original Articles

Case Report

Communication

www.e-hir.org

Original Article

Healthc Inform Res. 2026 April;32(2):166-178.
https://doi.org/10.4258/hir.2026.32.2.166
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Accuracy of Orthodontic Malocclusion Detection Using Multiple AI Models: A Comparative Study

Hillda Herawati[1], **Joko Kusnoto**[2], **Indrayadi Gunardi**[3], **Anggit Wirasto**[4], **Tri Erri Astoeti**[5]
[1]Doctoral Program in Dental Sciences, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia
[2]Department of Orthodontics, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia
[3]Department of Oral Medicine, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia
[4]Informatics Study Program, Faculty of Science and Technology, Universitas Harapan Bangsa, Purwokerto, Indonesia
[5]Department of Dental Public Health and Preventive Dentistry, Faculty of Dentistry, Universitas Trisakti, Jakarta, Indonesia

Objectives: This study aimed to evaluate and compare the accuracy of multiple artificial intelligence (AI) models (ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot) in detecting orthodontic malocclusion features in standardized multiview intraoral photographs. The reference standard was assessment by an orthodontist. **Methods:** A cross-sectional observational study was conducted using five standardized intraoral photographs (frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal) obtained from 50 children aged 9–12 years. The following eight malocclusion parameters were assessed: anterior crowding, diastema, overjet, overbite, molar relationship, canine relationship, crossbite, and dental arch symmetry. Diagnostic accuracy and agreement between each AI model and the orthodontist were evaluated using Cohen's kappa (κ) and the area under the receiver operating characteristic curve (AUC). **Results:** Agreement between the AI models and the orthodontist ranged from poor to moderate across all orthodontic domains, with Cohen's κ values ranging from -0.15 to 0.63. Visually prominent alignment features, including anterior crowding and diastema, demonstrated comparatively higher agreement (κ, 0.00–0.63) and discriminatory performance, with AUC values ranging from 0.56 to 0.85. In contrast, parameters requiring precise spatial interpretation, such as sagittal relationships, overbite, crossbite, and arch morphology, showed consistently low agreement (κ, -0.15 to 0.38) and poor to near-random classification performance, with AUC values predominantly ranging from 0.41 to 0.70 and, in some cases, approaching 0.50. **Conclusions:** Current multimodal AI models demonstrate limited, parameter-dependent accuracy in detecting orthodontic malocclusions from intraoral photographs. These findings emphasize the limitations of general-purpose AI systems for orthodontic decision support and highlight the need for task-specific models trained on clinically annotated datasets.

Keywords: ROC Curve, Artificial Intelligence, Diagnosis, Malocclusion, Orthodontics

---

Submitted: January 14, 2026
Revised: March 27, 2026
Accepted: April 4, 2026

Corresponding Author
Joko Kusnoto
Department of Orthodontics, Faculty of Dentistry, Universitas Trisakti, Jl. Kyai Tapa no. 260, Jakarta 11440, Indonesia. Tel: +62-21-5672731, E-mail: joko.k@trisakti.ac.id (https://orcid.org/0009-0009-8955-4459)

I. Introduction

Malocclusion is a prevalent oral health problem among children and adolescents, with reported prevalence rates as high as 30%-40% in some populations [1]. Early detection is essential to prevent functional impairment and reduce the need for more complex orthodontic interventions later in [1,2]. Traditional orthodontic screening methods rely largely on direct clinical examinations by trained specialists, which may limit access to needed dental care, particularly in community settings such as schools [2].

Recent advances in artificial intelligence (AI) have shown promise for improving diagnostic accuracy in orthodontics, including through the application of machine learning and deep learning to various imaging modalities. For example, convolutional neural networks (CNNs) have been used effectively in cephalometric analysis and panoramic image interpretation, yielding accurate automated orthodontic diagnoses [3]. However, the use of intraoral photographs remains relatively underexplored, despite their value as non-invasive, low-cost clinical records suitable for early screening and teleorthodontics [4,5]. The diagnostic potential of photographs has nevertheless been recognized, and previous studies have reported strong diagnostic performance for dental conditions assessed from these images [6,7].

In recent years, large language models (LLMs), including ChatGPT and similar AI frameworks, have emerged with the ability to interpret both textual and visual inputs. Studies evaluating LLM responses in orthodontics have reported moderate to high consistency but substantial variability in accuracy relative to assessments by orthodontic specialists [8]. Although these models can generate fluent and well-structured responses, they may also provide incomplete or misleading information, underscoring the need for caution-clinical interpretation [9-11]. These observations support the view that LLMs may serve as adjunctive tools in orthodontic practice rather than replacements for expert judgment.

Comparative studies have identified discrepancies between LLM-generated responses and assessments by orthodontic experts, particularly in malocclusion classification and treatment decision-making [9,12]. Although these models may have value for preliminary educational purposes, existing evidence suggests that they cannot replace the nuanced assessments of experienced clinicians [1,3,11]. This limitation highlights the need for further research to rigorously validate AI-generated outputs against established clinical standards.

To evaluate the accuracy of malocclusion classification from clinical images by AI models, including advanced versions such as ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot, standardized inputs are required for comparison with expert evaluations. At present, research on the performance of multimodal LLMs in analyzing standardized multiview intraoral photographs against comprehensive orthodontic diagnostic criteria remains limited [7,9]. Further studies are therefore needed in thisarea.

II. Methods

1. Study Design and Population

This cross-sectional observational study was designed to evaluate the diagnostic performance of multimodal LLMs in detecting orthodontic malocclusion features from standardized intraoral photographs. The study focused exclusively on image-based diagnostic interpretation and did not incorporate clinical examination, radiographic analysis, or treatment planning. The study used retrospective clinical photographs obtained from 50 children aged 9–12 years in Cimahi, West Java, Indonesia. No clinical intervention, treatment modification, or follow-up assessment was performed as part of the study.

2. Sample Size

This observational cross-sectional study aimed to evaluate the diagnostic accuracy of multimodal LLMs in detecting orthodontic malocclusion from intraoral images. The study sample consisted of 50 pediatric participants, each contributing five intraoral images obtained from different views, including frontal, right lateral, left lateral, maxillary occlusal, and mandibular occlusal views. The total sample size was 250 images, which was considered sufficient to cover key aspects of orthodontic malocclusion, including alignment, sagittal, vertical, transverse, and arch morphology.

3. Ethical Approval and Consent to Participate

Written informed consent was obtained from all participants and their legal guardians before study initiation. Ethical approval for the study protocol was granted by the Health Research Ethics Committee of the Faculty of Dentistry, Universitas Trisakti (No. 1006/S3/KEPK/FKG/9/2025).

4. Image Collection

Five images were acquired for each participant, comprising frontal, left lateral, right lateral, maxillary occlusal, and mandibular occlusal views (Figure 1). All images were captured using an iPhone 15 (Apple Inc., Cupertino, CA, USA) at a resolution of 6048 × 4032 pixels (24 megapixels). All images were obtained by a single investigator who was calibrated for the study, following standardized intraoral photography protocols to ensure consistent angulation, lighting, and field of view.

5. Image Preparation and Pre-processing

All images were anonymized before analysis by removing patient-identifying information. The images were standardized

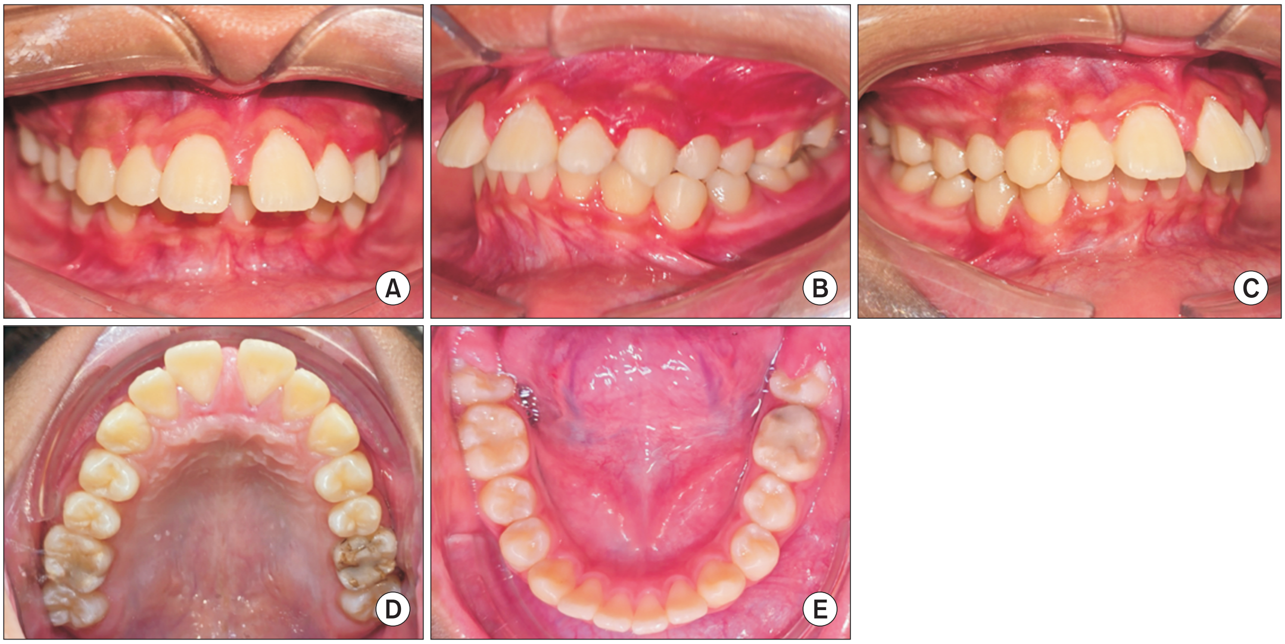


Figure 1. Sample clinical images obtained from orthodontic patients: (A) frontal view, (B) left lateral view, (C) right lateral view, (D) maxillary occlusal view, and (E) mandibular occlusal view.

without digital enhancement, filtering, contrast adjustment, or color correction. However, the images were cropped to remove extraneous objects, such as hands and mirrors, that were visible during image acquisition. Images were analyzed at their original resolution to preserve clinically relevant visual features, such as tooth arrangement, occlusal relationships, and arch morphology.

6. Orthodontic Diagnostic Framework

Orthodontic assessment in this study followed a structured diagnostic framework based on established clinical standards, including Proffit's definitionsof malocclusion, Angle's classification of sagittal relationships, and the Dental Health Component of the Index of Orthodontic Treatment Need (IOTN) [13-15].

For analytical consistency and effective image-based evaluation, orthodontic parameters were organized into five diagnostic domains reflecting different levels of spatial complexity:

1) Alignment domain: This domain included assessment of anterior crowding and diastema in both the maxillary and mandibular arches. These features are commonly recognized indicators in orthodontic screening [16].
2) Sagittal relationship domain: This domain comprised assessment of overjet, molar relationship, canine relationship, and overall Angle classification, focusing on anteroposterior dental relationships that are central to orthodontic diagnosis [17].
3) Vertical relationship domain: This domain emphasized overbite assessment, categorized as normal, deep bite, or open bite, representing clinically important vertical discrepancies [17].
4) Transverse relationship domain: This domain assessed the presence and type of crossbite (anterior or posterior, unilateral or bilateral), capturing transverse discrepancies that are often subtle on two-dimensional imaging [18].
5) Arch morphology domain: This domain evaluated dental arch symmetry using occlusal views, highlighting structural characteristics shaped by underlying skeletal patterns [19].

All parameters were evaluated systematically within this unified framework by both the orthodontist and the AI models, thereby ensuring a structured assessment approach. For statistical analysis, outcomes were dichotomized as normal or abnormal to enable direct comparison of diagnostic agreement and discrimination [20]. The detailed mapping of multilevel orthodontic parameters to binary categories is presented in Table 1.

7. Generative AI Multimodal Large Language Models

In this study, four publicly available multimodal LLMs with image-text processing capabilities were evaluated: ChatGPT 5.2 Pro (OpenAI), Claude 4.5 Sonnet (Anthropic), Gemini 3 Fast (Google), and Microsoft Copilot (Microsoft) [12,21]. These models were accessed through their official web-based

Table 1. Mapping of multi-class orthodontic parameters into binary categories for statistical analysis

Domain	Parameter	Original category	Binary category
Alignment	Maxillary crowding	None	Normal
		Mild, Moderate, Severe	Abnormal
	Mandibular crowding	None	Normal
		Mild, Moderate, Severe	Abnormal
	Diastema	None	Normal
		Presence (maxilla or mandible)	Abnormal
Sagittal	Overjet	Normal	Normal
		Increased, Edge-to-edge, Reverse	Abnormal
	Molar relationship (right & left)	Class I	Normal
		Class II, Class III	Abnormal
	Canine relationship (right & left)	Class I	Normal
		Class II, Class III	Abnormal
	Angle classification	Class I	Normal
		Class II, Class III	Abnormal
Vertical	Overbite	Normal	Normal
		Deep bite, Open bite	Abnormal
Transverse	Crossbite (right & left)	None	Normal
		Anterior, Posterior, Anterior + Posterior	Abnormal
Arch morphology	Arch symmetry	Symmetrical (maxilla/mandible)	Normal
		Asymmetrical	Abnormal

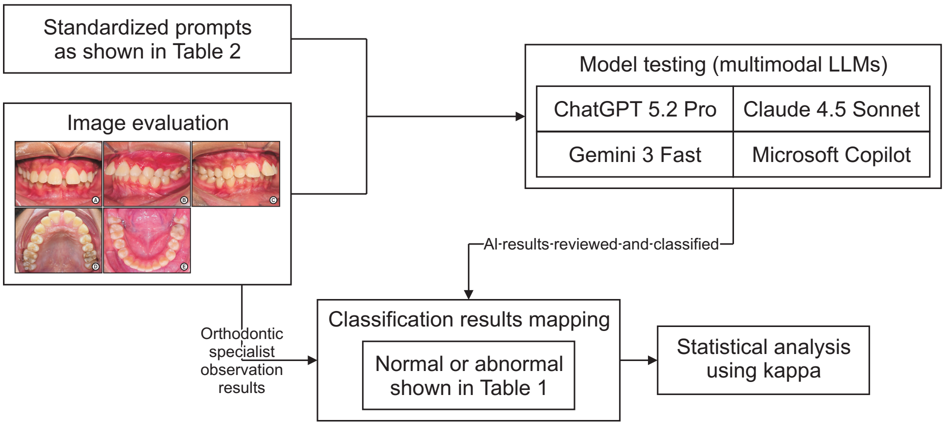


Figure 2. Schematic overview of multimodal large language model (LLM) evaluation, result mapping, and agreement analysis.

interfaces using their default system configurations, without application programming interfaces, external tools, or model fine-tuning [12]. This approach was intended to reflect how these multimodal LLMs would operate in real-world clinical use.

8. Model Testing Workflow and Prompting Strategy

For model evaluation, a structured workflow was used, as illustrated in Figure 2 and detailed in Table 2. Analysis and testing of the AI models were conducted in stages according to the orthodontic diagnostic framework. Each intraoral image was evaluated independently by each multimodal model in a blinded testing procedure. This approach ensured that each model's rating depended solely on information contained in the corresponding intraoral image. Parameter-specific prompts were used for each orthodontic category with identical syntax and structure, with the aim of approximating the logical reasoning process used by an orthodontist.

Table 2. Standardized prompts used for multimodal LLM evaluation by orthodontic domain

Orthodontic domain	Image view	Diagnostic focus	Standardized prompt
Alignment	Frontal intraoral	Anterior crowding	“Analyze this frontal intraoral photograph. Focus on anterior crowding. Identify whether crowding is present in the maxillary and/or mandibular anterior teeth, specify the affected teeth, classify severity as mild, moderate, or severe, and provide a diagnostic conclusion.”
Alignment	Frontal intraoral	Diastema	“Analyze this frontal intraoral photograph. Focus on diastema. Determine whether diastema is present, identify its location (maxillary and/or mandibular), and provide a diagnostic conclusion.”
Sagittal relationship	Lateral intraoral (right/left)	Molar relationship (M1)	“Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary first molar and mandibular first molar. Classify the molar relationship as Class I, II, or III. Report findings separately for the right and left sides.”
Sagittal relationship	Lateral intraoral (right/left)	Canine relationship (C)	“Analyze this lateral intraoral photograph. Focus on the relationship between the maxillary canine and mandibular canine. Classify the canine relationship as Class I, II, or III. Report findings separately for the right and left sides.”
Sagittal relationship	Lateral intraoral	Overjet	“Analyze this lateral intraoral photograph. Focus on overjet. Classify the overjet as normal, increased, edge-to-edge, or reverse, and provide a diagnostic conclusion.”
Vertical relationship	Frontal intraoral	Overbite	“Analyze this frontal intraoral photograph. Focus on overbite. Classify the vertical overlap as normal, deep bite, or open bite, and provide a diagnostic conclusion.”
Transverse relationship	Frontal intraoral	Crossbite	“Analyze this frontal intraoral photograph. Focus on transverse relationships. Identify the presence of crossbite, specify whether it is unilateral or bilateral, and provide a diagnostic conclusion.”
Arch morphology	Maxillary occlusal	Arch symmetry	“Analyze this maxillary occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion.”
Arch morphology	Mandibular occlusal	Arch symmetry	“Analyze this mandibular occlusal photograph. Assess the arch is symmetrical or asymmetrical, and provide a diagnostic conclusion.”

LLM: large language model.

All prompts were written in Indonesian and were kept identical throughout the study. Outputs from each model were classified into predefined orthodontic categories and then reduced to two groups: normal and abnormal.

4

9. Statistical Analysis

Cohen’s kappa (κ) was used to assess agreement between each AI model and the single orthodontist’s assessment. This method is widely used in dental research to quantify inter-rater reliability, particularly in studies examining diagnostic discrepancies[7]. By providing a comparative framework, Cohen’s κ helps indicate how closely AI model outputs align with expert clinical assessments [22]. Model performance was evaluated by dichotomizing diagnostic outputs into normal and abnormal categories and calculating the area under the receiver operating characteristic curve (AUC) to quantify discriminative ability. AUC values were interpreted according to established receiver operating characteristic criteria as follows: >0.90, excellent; 0.80–0.89, good; 0.70–0.79, fair; 0.60–0.69, poor; and <0.60, fail [23].

III. Results

A total of 50 children were included in this study. Participants were 9–12 years old (Table 3). For each participant, five standardized intraoral photographs were obtained, comprising frontal, left lateral, right lateral, maxillary occlusal, and mandibular occlusal views. In total, 250 intraoral images were analyzed and used as input for the AI models.

Across all orthodontic domains, the dataset showed substantial clinical variability while representing a typical range of malocclusions (Table 4). In the alignment domain, both arches showed a predominance of anterior crowding, most of which was mild to moderate in severity. Diastemas were observed more frequently in the mandible than in the maxilla. In the sagittal domain, normal overjet and Angle Class I relationships predominated; however, a high proportion of participants also exhibited Class II malocclusion and increased overjet. In the vertical domain, normal overbite was most common, with deep bite representing the primary anomaly and open bite occurring rarely. In the transverse

Table 3. Demographic characteristics of the subjects

Variable	Value
Sex	
Male	20 (40)
Female	30 (60)
Age (yr)	10.56 ± 0.73

Values are presented as number (%) or mean ± standard deviation.

Table 4. Integrated distribution of orthodontic diagnostic characteristics across all domains

Parameter	n (%)
Alignment domain	
Maxillary crowding	
None	6 (12)
Mild	23 (46)
Moderate	13 (26)
Severe	8 (16)
Mandibular crowding	
None	6 (12)
Mild	16 (32)
Moderate	21 (42)
Severe	7 (14)
Diastema	
Maxilla	4 (8)
Mandible	12 (24)
Sagittal relationship domain	
Overjet	
Normal	31 (62)
Increased	15 (30)
Edge-to-edge	2 (4)
Reverse	2 (4)
Molar relationship (right)	
Class I	31 (62)
Class II	18 (36)
Class III	1 (2)
Molar relationship (left)	
Class I	30 (60)
Class II	18 (36)
Class III	2 (4)
Canine relationship (right)	
Class I	28 (56)
Class II	22 (44)
Canine relationship (left)	
Class I	30 (60)
Class II	19 (38)
Class III	1 (2)
Angle classification	
Class I	27 (54)
Class II	21 (42)
Class III	2 (4)
Vertical relationship domain	
Overbite	
Normal	31 (62)
Open bite	1 (2)
Deep bite	18 (36)
Transverse relationship domain	
Crossbite (left)	
None	41 (82)
Anterior	6 (12)
Posterior	2 (4)
Anterior + posterior	1 (2)
Crossbite (right)	
None	41 (82)
Anterior	6 (12)
Posterior	3 (6)
Arch morphology domain	
Arch symmetry	
Maxilla (symmetrical)	39 (78)
Mandible (symmetrical)	40 (80)

Continued on the next column.

Table 4. Continued

domain, crossbite was uncommon, and most participants had no crossbite; when present, crossbite was usually unilateral. Symmetry predominated in both the maxillary and mandibular arches. Overall, the dataset represented a clinically variable yet typical sample suitable for comparative evaluation of multimodel AI system performance.

Following the descriptive analysis, receiver operating characteristic (ROC) analysis was performed to evaluate the discriminatory performance of the multimodal models in classifying normal versus abnormal orthodontic findings. Figure 3 shows marked domain-specific variation in model performance across the ROC curves. Parameters related to alignment demonstrated the greatest discriminatory ability, with ROC curves positioned farther from the reference diagonal, whereas sagittal relationship parameters showed only moderate discriminatory ability. In contrast, the verti-

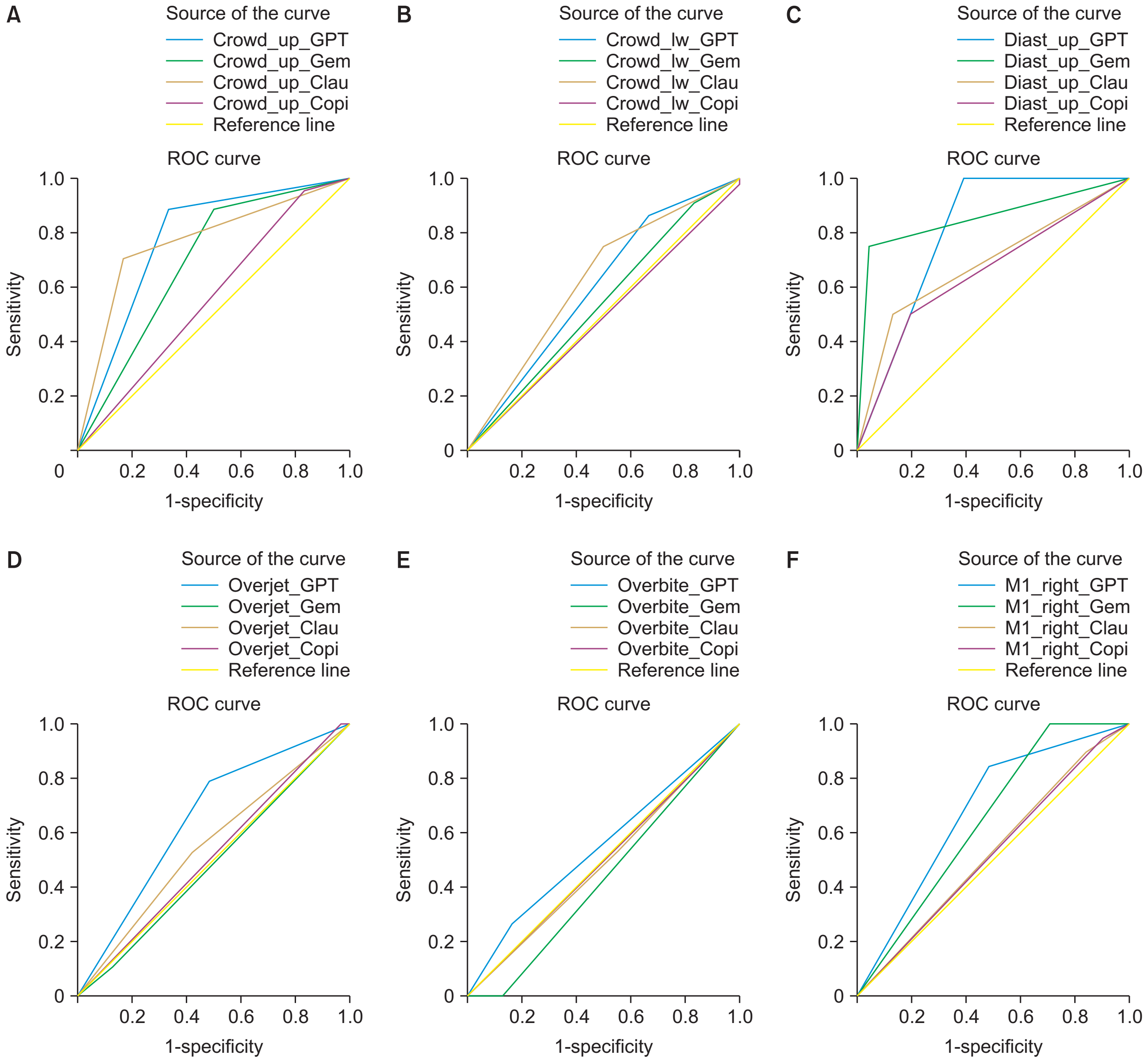


Figure 3. Integrated receiver operating characteristic (ROC) curves of multimodal AI models for orthodontic parameters (see Supplementary Figure S1 for a high-resolution version). Diagonal segments are produced by ties. Panels (A)–(N) represent ROC curves for individual orthodontic parameters evaluated from standardized multiview intraoral photographs; (A) maxillary crowding, (B) mandibular crowding, (C) maxillary diastema, (D) overjet, (E) overbite, (F) right first molar relationship, (G) left first molar relationship, (H) right canine relationship, (I) left canine relationship, (J) Angle classification, (K) right crossbite, (L) left crossbite, (M) maxillary arch symmetry, and (N) mandibular arch symmetry. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot against the orthodontist reference standard.

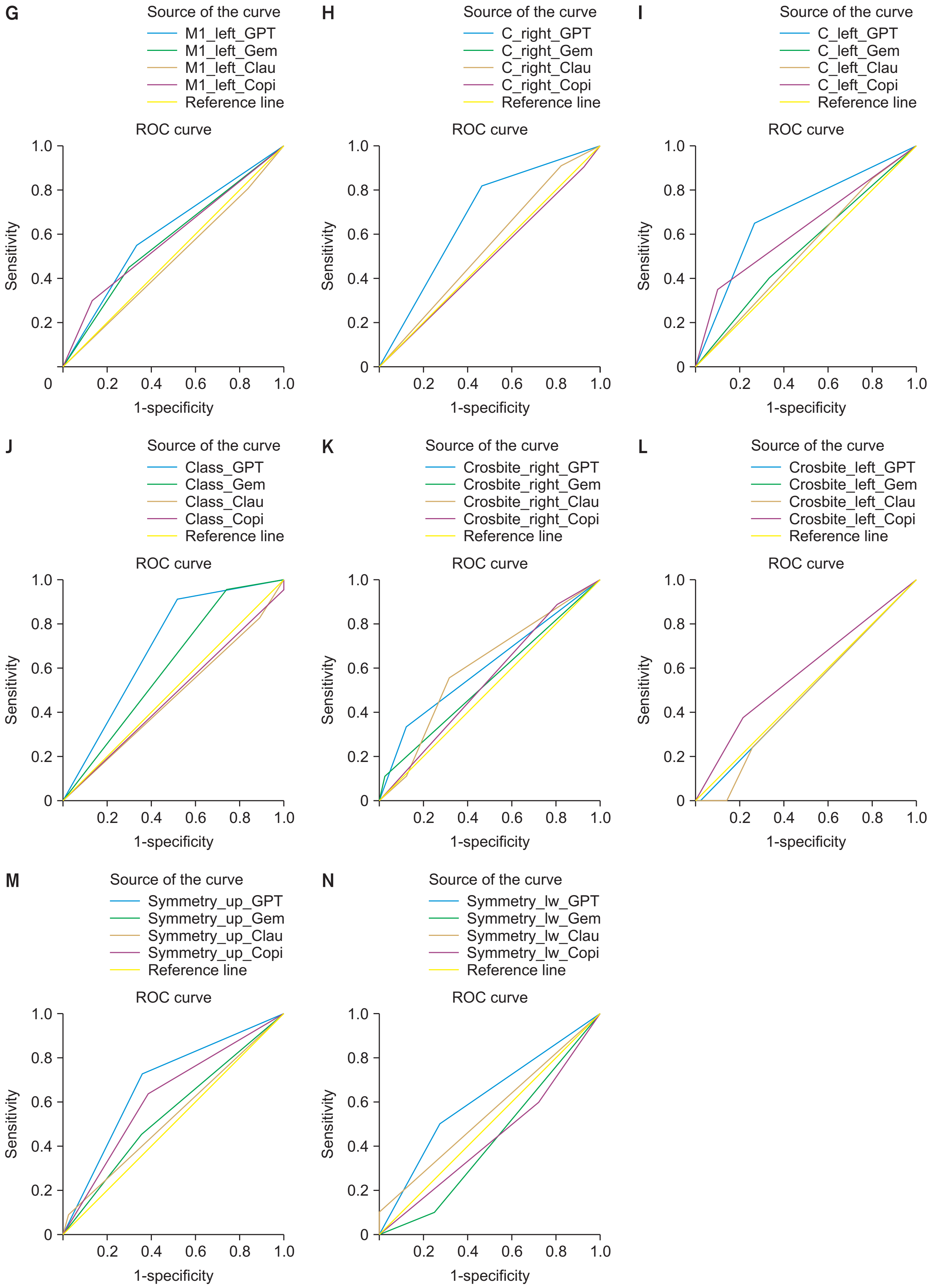


Figure 3. Continued

cal, transverse, and arch morphology domains showed near-random classification performance, with their ROC curves clustering close to the reference diagonal. These findings indicate that multimodal models perform better for visually prominent orthodontic features than for complex spatial relationships.

As shown in Table 5, the discriminative performance of the multimodal AI models varied substantially across orthodontic tasks. The alignment domain showed the highest discriminative performance, with Gemini achieving the highest AUC value of 0.85, followed by ChatGPT, with AUC values approaching 0.80, indicating relatively strong discrimination for visually observable features such as crowding and diastema. Claude showed similar but slightly lower values, whereas Copilot demonstrated the lowest discriminative performance.

Discriminative performance declined in the sagittal relationship domain, in which the highest AUC values, up to 0.70, were observed for ChatGPT, representing fair discrimination, whereas the other models performed less well. In contrast, all models showed poor discriminative performance in the vertical and transverse relationship domains, with AUC values close to the random-classification threshold of 0.50. Discriminative performance was lowest in the arch morphology domain, in which AUC values did not reach the threshold for fair discrimination, reflecting the continued limitations of multimodal models in processing complex dental geometric features accurately.

Analysis of Cohen's κ statistics for agreement between the multimodal models and the orthodontist showed poor to moderate agreement across all domains (Table 6). In the alignment domain, agreement was generally poor, although Gemini achieved moderate agreement, with κ values up to 0.63. ChatGPT and Claude showed lower agreement, whereas Copilot showed only minimal agreement. Agreement in the sagittal relationship domain was also poor overall, although ChatGPT performed relatively better, with κ values up to 0.38; Gemini and Claude performed less well, and Copilot showed minimal agreement. In the vertical, transverse, and arch morphology domains, agreement was consistently poor, with κ values near zero or below zero across models. Overall, the combined ROC-AUC and κ findings suggest that, although multimodal AI models can moderately discriminate visually salient alignment features, agreement with orthodontist evaluation remains limited, particularly for parameters requiring precise spatial interpretation.

Figure 4 presents a radar-chart comparison of the discrimination abilities of the four AI models. ROC analysis indicates notable differences in discriminative performance among the models, with ChatGPT showing the highest overall performance, followed by Gemini, whereas Claude and Microsoft Copilot showed more limited performance, as reflected by ROC curves that lay closer to the reference line.

According to the radar chart, ChatGPT showed the best overall balance across sensitivity, specificity, positive predictive value, negative predictive value, and accuracy. Gemini showed comparatively strong sensitivity and positive and negative predictive values, although its overall balance across performance indices was somewhat lower. Claude and Copilot showed the weakest overall performance, particularly

Table 5. Integrated ROC-AUC performance of multimodal AI models across orthodontic domains

Orthodontic domain	Parameters included	AUC				Overall interpretation
		ChatGPT 5.2 Pro	Gemini 3 Fast	Claude 4.5 Sonnet	Microsoft Copilot	
Alignment	Crowding (maxillary, mandibular), Diastema (maxillary)	0.60–0.80	0.69–0.85	0.68–0.77	0.56–0.65	Poor–Good
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.65–0.70	0.53–0.65	0.47–0.55	0.48–0.52	Poor
Vertical relationship	Overbite	0.55	0.44	0.49	0.50	Fail
Transverse relationships	Crossbite (R/L)	0.49–0.61	0.50–0.54	0.48–0.60	0.54–0.58	Fail–Poor
Arch morphology	Symmetry (maxillary, mandibular)	0.41–0.68	0.43–0.61	0.47–0.60	0.44–0.63	Fail–Poor

ROC: receiver operating characteristic, AUC: area under the curve, AI: artificial intelligence, R: right, L: left.

AUC values were interpreted based on established ROC guidelines: ≥0.90 (excellent), 0.80–0.89 (good), 0.70–0.79 (fair), 0.60–0.69 (poor), and <0.60 (fail).

Table 6. Summary of agreement (Cohen's kappa) between AI models and orthodontist assessment by orthodontic domains

Orthodontic domain	Parameters included	Cohen's kappa				Overall agreement
		ChatGPT 5.2 Pro	Gemini 3 Fast	Claude 4.5 Sonnet	Microsoft Copilot	
Alignment	Crowding (maxillary, mandibular), Diastema (maxillary)	0.11–0.33	0.15–0.63	0.16–0.25	0.00–0.17	Poor–Moderate
Sagittal relationships	Overjet, M1 (R/L), Canine (R/L), Angle classification	0.20–0.38	0.08–0.21	-0.03–0.07	-0.04–0.23	Poor–Moderate
Vertical relationship	Overbite	0.13	-0.15	0.05	-0.01	Poor
Transverse relationships	Crossbite (R/L)	-0.02–0.13	0.00–0.05	0.09–0.11	0.03–0.07	Poor
Arch morphology	Symmetry (maxillary, mandibular)	0.07–0.27	-0.15–0.15	0.07–0.15	-0.07–0.19	Poor

AI: artificial intelligence, R: right, L: left.

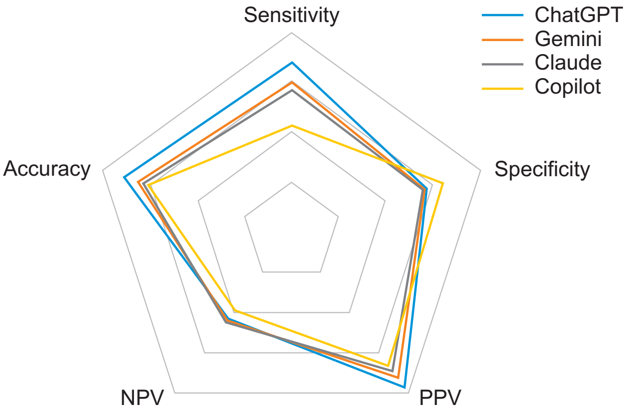


Figure 4. Comparative diagnostic performance of AI models based on classification metrics and receiver operating characteristic (ROC) curve. Each panel compares the discriminatory performance of ChatGPT 5.2 Pro, Gemini 3 Fast, Claude 4.5 Sonnet, and Microsoft Copilot. AI: artificial intelligence, NPV: negative predictive value, PPV: positive predictive value.

with respect to sensitivity and accuracy, with somewhat reduced specificity as well.

Taken together, these findings indicate that multimodal AI models show domain-specific performance, with higher discrimination for visually salient alignment features and lower discrimination and agreement for parameters requiring accurate spatial interpretation. These results emphasize the domain-specific limitations of general-purpose AI systems in the analysis of dental images.

IV. Discussion

This study provides an exploratory analysis of the clinical applicability of a multiview intraoral photograph dataset and the diagnostic performance of multimodal large language models in assessing orthodontic malocclusion. The dataset showed substantial clinical variability across alignment, sagittal, vertical, transverse, and arch morphology domains, all of which are important for evaluating AI tools [1]. Notably, the sample included both male and female participants across the target age range and was supplemented with multi-angle images, enabling assessment of both simple and complex orthodontic features [6].

Across these domains, a clear pattern emerged. Parameters related to anterior alignment, such as crowding and diastema, showed higher agreement (κ, 0.00–0.63) and AUC values (0.56–0.85). The relatively stronger performance observed for alignment parameters may be attributable to their visually salient, inherently two-dimensional characteristics, which can be inferred directly from intraoral photographs without requiring complex spatial interpretation. Features such as crowding and diastema appear as explicit variations in tooth spacing and position, making them well suited to pattern recognition. In addition, alignment assessment relies on relatively stable and easily identifiable anatomical landmarks, which may reduce variability across evaluators and analytical approaches. These features are also less susceptible to perspective distortion caused by variation in camera positioning than sagittal relationships and arch morphology, both of which depend on more complex interarch spatial relationships and three-dimensional interpretation.

In contrast, recent studies have found that geometrically precise parameters, such as molar and canine relationships and dental arch morphology, show suboptimal performance, often reflected by lower AUC values, because they depend on fine-grained spatial relationships and three-dimensional interpretation that are difficult to infer from two-dimensional images [24]. The implications of these findings support the hypothesis that general multimodal learning models and LLMs may have value in preliminary orthodontic analysis, butthat their performance varies markedly across areas of orthodontic assessment [1]. This interpretation is further supported by recent large-scale benchmark evaluations showing that strong performance on broad benchmarks does not necessarily translate into effective reasoning in scientific and clinical domains [25].

Our results suggest that the evaluative performance of large language models is stronger for anterior than for posterior relationships. These findings are consistent with previous studies examining AI capabilities in orthodontics. For example, Hack et al. [4] reported that AI models perform better in identifying conspicuously visible features, such as anterior tooth position, than in assessing complex occlusal relationships that require robust spatial reference frameworks. Similarly, Stetzel et al. [15] found that deep learning models perform well in predicting aesthetically oriented components of the IOTN, illustrating the strengths of AI in visually oriented analyses. However, these findings contrast with those of the present study with respect to sagittal and transverse parameter assessment, while still remaining broadly consistent with prior reports that AI models are challenged when evaluating parameters that depend on strong spatial references [24].

This study also demonstrated variation in agreement and accuracy across the evaluated large language models (Tables 4 and 5). These intermodel differences may reflect differences in underlying architectures and training paradigms. For example, Transformer-based models, including Vision Transformers, process images holistically and may therefore detect general irregularities in dental alignment efficiently, although they remain limited in making subtle distinctions among densely interdependent orthodontic categories such as overjet and molarintercuspation [26]. These observations are consistent with previous reviews showing that CNNs trained on task-specific datasets outperform general AI systems, including LLMs, on tasks requiring high anatomical specificity [7]. The observed differences in outputs among the evaluated large language models may therefore reflect variation in visual abstraction, depth of reasoning, and internal representational capacity than true conceptual understanding [27]. Overall, the aggregated findings likely reflect the representational constraints imposed by nonspecialized training data [28]. Notably, cross-model analyses in scientific AI research have shown high error correlations among leading large language models, suggesting shared inductive biases rather than independent failure modes [25].

From a healthcare informatics perspective, evaluating several LLMs together may be viewed as a form of collective model behavior. Although each model has distinct strengths and limitations, broad patterns remain evident, including stronger performance in detecting alignment anomalies and much weaker performance in identifying complexspatial characteristics [1,29]. These findings highlight the importance of learning from multiple models collectively while also recognizing the limitations of AI and the continuing need for human expertise [28]. The observed convergence across models further suggests that aggregating general-purpose LLMs may offer limited benefit for inherently spatial clinical tasksreinforcing the need for task-specific AI systems [25]. Future studies should therefore focus on developing task-specific orthodontic AI systems rather than continuing to test general-purpose LLMs alone. An important next step will be the development of appropriately labeled, standardized multiview intraoral photograph datasets curated by orthodonticspecialists [24]. Such datasets could support supervised learning models capable of finer-scale spatial reasoning and clinically interpretable outputs [15], although this will require collaboration between orthodontists and information technology specialists.

This study has several limitations, including variability in image quality, use of a single orthodontist as the reference standard, and a relatively small sample size, all of which may limit generalizability. Reliance on a single evaluator may introduce subjectivity; therefore, the observed Cohen's κ values may reflect both AI performance and variability in expert judgment. Future studies should include multiple orthodontists and assess inter-rater reliability to establish a more robust reference standard. In addition, some models generated unsolicited treatment suggestions rather than strictly diagnostic outputs, highlighting limited output controllability. Furthermore, none of the models explicitly expressed diagnostic uncertainty or provided clinical disclaimers, raising important ethical considerations regarding the use of multimodal AI in healthcare settings. Moreover, all models were evaluated using a single standardized prompting framework. Therefore, the reported results reflect performance under this specific prompting configuration only. Alternative prompting strategies were not evaluated, and different prompt designs may lead to variation in performance; however, ensemble prompting may be considered as an alternative. This limitation should be considered when interpreting the findings. In conclusion, current multimodal

AI models demonstrate limited, parameter-dependent accuracy in detecting orthodontic malocclusions from intraoral photographs. These findings emphasize the limitations of general-purpose AI systems for orthodontic decision support and highlight the need for task-specific models trained on clinically annotated datasets.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

ORCID

Hillda Herawati (https://orcid.org/0009-0001-4222-9159)
Joko Kusnoto (https://orcid.org/0009-0009-8955-4459)
Indrayadi Gunardi (https://orcid.org/0000-0002-5525-5559)
Anggit Wirasto (https://orcid.org/0000-0002-1707-5979)
Tri Erri Astoeti (https://orcid.org/0000-0002-9341-504X)

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.4258/hir.2026.32.2.166.

References

1. Monill-Gonzalez A, Rovira-Calatayud L, d'Oliveira NG, Ustrell-Torrent JM. Artificial intelligence in orthodontics: where are we now? A scoping review. Orthod Craniofac Res 2021;24 Suppl 2:6-15. https://doi.org/10.1111/ocr.12517
2. Albalawi F, Abalkhail KA. Trends and application of artificial intelligence technology in orthodontic diagnosis and treatment planning: a review. Appl Sci 2022;12(22):11864. https://doi.org/10.3390/app122211864
3. Subramanian AK, Chen Y, Almalki A, Sivamurthy G, Kafle D. Cephalometric analysis in orthodontics using artificial intelligence: a comprehensive review. Biomed Res Int 2022;2022:1880113. https://doi.org/10.1155/2022/1880113
4. Hack M, Dragulin B, Hack L, ElSaafin M, Dumitrescu I, Stan D, et al. Comparative study on the results of orthodontic diagnostics by using algorithms generated by artificial intelligence and simple algorithms. Med Pharm Rep 2024;97(2):215-21. https://doi.org/10.15386/mpr-2702
5. Liu J, Zhang C, Shan Z. Application of artificial intelligence in orthodontics: current state and future perspectives. Healthcare (Basel) 2023;11(20):2760. https://doi.org/10.3390/healthcare11202760
6. Ryu J, Kim YH, Kim TW, Jung SK. Evaluation of artificial intelligence model for crowding categorization and extraction diagnosis using intraoral photographs. Sci Rep 2023;13(1):5177. https://doi.org/10.1038/s41598-023-32514-7
7. Zhang R, Zhang L, Zhang D, Wang Y, Huang Y, Wang D, et al. Development and evaluation of a deep learning model for occlusion classification in intraoral photographs. PeerJ 2025;13:e20140. https://doi.org/10.7717/peerj.20140
8. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(1):33. https://doi.org/10.1007/s10916-023-01925-4
9. Zheng J, Ding X, Pu JJ, Chung SM, Ai QY, Hung KF, et al. Unlocking the potentials of large language models in orthodontics: a scoping review. Bioengineering (Basel) 2024;11(11):1145. https://doi.org/10.3390/bioengineering11111145
10. Wu X, Huang Y, He Q. A large language model improves clinicians' diagnostic performance in complex critical illness cases. Crit Care 2025;29(1):230. https://doi.org/10.1186/s13054-025-05468-7
11. Hatia A, Doldo T, Parrini S, Chisci E, Cipriani L, Montagna L, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. J Clin Med 2024;13(3):735. https://doi.org/10.3390/jcm13030735
12. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023;25:e51580. https://doi.org/10.2196/51580
13. Jasim ES, Kadhom ZM, Al-Groosh D. Comparative evaluation of orthodontic treatment needs index and the dental aesthetic index to assess the need for orthodontic treatment from the participants' perspective: a cross-sectional study. J Orthod Sci 2025;14:30. https://doi.org/10.4103/jos.jos_24_25
14. Roy P, Roy P, Koley S, Sheet S. Bolton's ratio variations in Angle's Class I, Class II and Class III malocclusions: an observational study. J Clin Exp Dent 2025;17(3):e280-5.

https://doi.org/10.4317/jced.62591

15. Stetzel L, Foucher F, Jang SJ, Wu TH, Fields H, Schumacher F, et al. Artificial intelligence for predicting the aesthetic component of the index of orthodontic treatment need. Bioengineering (Basel) 2024;11(9):861. https://doi.org/10.3390/bioengineering11090861
16. Brook PH, Shaw WC. The development of an index of orthodontic treatment priority. Eur J Orthod 1989; 11(3):309-20. https://doi.org/10.1093/oxfordjournals.ejo.a035999
17. Nguyen SM, Nguyen MK, Saag M, Jagomagi T. The need for orthodontic treatment among vietnamese school children and young adults. Int J Dent 2014;2014:132301. https://doi.org/10.1155/2014/132301
18. Kozanecka A, Sarul M, Kawala B, Antoszewska-Smith J. Objectification of orthodontic treatment needs: does the classification of malocclusions or a history of orthodontic treatment matter? Adv Clin Exp Med 2016; 25(6):1303-12. https://doi.org/10.17219/acem/62828
19. Braun S, Hnat WP, Fender DE, Legan HL. The form of the human dental arch. Angle Orthod 1998;68(1):29-36. https://doi.org/10.1043/0003-3219(1998)068<0029:TFOTHD>2.3.CO;2
20. Huynh N, Zhang J, Pliska B, Amin R, Narang I, Chadha N, et al. Prevalence of altered craniofacial morphology in children with OSA. J Sleep Res 2025;34(5):e70060. https://doi.org/10.1111/jsr.70060
21. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA 2023; 330(9):866-9. https://doi.org/10.1001/jama.2023.14217
22. Masood Y, Masood M, Zainul NN, Araby NB, Hussain SF, Newton T. Impact of malocclusion on oral health related quality of life in young people. Health Qual Life Outcomes 2013;11(1):25. https://doi.org/10.1186/1477-7525-11-25
23. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol 2022;75(1):25-36. https://doi.org/10.4097/kja.21209
24. Vishwanathaiah S, Fageeh HN, Khanagar SB, Maganur PC. Artificial intelligence its uses and application in pediatric dentistry: a review. Biomedicines 2023;11(3):788. https://doi.org/10.3390/biomedicines11030788
25. Song Z, Lu J, Du Y, Yu B, Pruyn TM, Huang Y, et al. Evaluating large language models in scientific discovery [Internet]. Ithaca (NY): arXiv.org; 2025 [cited 2026 Mar 1]. Available from: https://arxiv.org/abs/2512.15567.
26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale [Internet]. Ithaca (NY): arXiv.org; 2020 [cited 2026 Mar 1]. Available from: https://arxiv.org/abs/2010.11929v1.
27. Azizi S, Hatampoor S, Tahamtan S. Applications of artificial intelligence in diagnosis and treatment planning of orthodontics: a narrative review. Saudi Dent J 2025; 37(7-9):70. https://doi.org/10.1007/s44445-025-00077-0
28. Ahmed AE, Aljohani WF, Abu Rukbah LK, Rajhi SA, Najmi NK, Zughlul MK, et al. The role of artificial intelligence in stroke imaging in emergency settings: a systematic review. Cureus 2025;17(10):e93941. https://doi.org/10.7759/cureus.93941
29. Jiang L, Chai Y, Li M, Liu M, Fok R, Dziri N, et al. Artificial hivemind: The open-ended homogeneity of language models (and beyond) [Internet]. Ithaca (NY): arXiv.org; 2025 [cited 2026 Mar 1]. Available from: https://arxiv.org/abs/2510.22954.