

Evaluating Diagnostic Accuracy in Jaw Pathologies on Orthopantomograms: A Comparative Study between Oral Radiologists and AI-Driven Chatgpt Analysis

Dr. Yashika Kewalramani, Arjun Singh Parihar

Oral Physician and Dentomaxillofacial Radiologist Government College of Dentistry, Indore, Choithram International School

Abstract- Artificial Intelligence (AI) and its incorporation into dental imaging, particularly in the interpretation of radiographs known as Orthopantomograms, has led to many promising advancements. However, its clinical utility and diagnostic consistency remain subjects of investigation when compared to the judgment of trained oral radiologists. This study evaluates the diagnostic precision and variability between experienced oral radiologists and a widely accessible AI model “ChatGPT”, in analyzing different confirmed Jaw Pathologies through Orthopantomograms. By using systematic assessment methods, the study aims to ensure a balanced and objective examination of the potential incorporation of AI in oral radiodiagnosis.

Index Terms- : Radiography, Panoramic, Artificial Intelligence, Oral lesions.

I. INTRODUCTION

Artificial intelligence (AI) is transforming healthcare by improving diagnostic accuracy, aiding clinical decisions, and streamlining workflows in fields like radiology and dentistry^{1,2}. Orthopantomograms (OPGs) are essential for detecting jaw pathologies but can be difficult to interpret due to overlapping anatomy and subtle features³. Advances in large language models (LLMs), such as OpenAI’s ChatGPT-4o, have shown strong diagnostic performance in imaging, sometimes comparable to radiologists^{4,5}. In oral radiology, ChatGPT-4o has correctly identified up to 78% of complex cases, performing better than earlier models^{6,7}. Studies suggest it can serve as a useful adjunct, providing consistent reasoning and second- opinion support^{4,8}. This study evaluates ChatGPT’s diagnostic accuracy compared with oral radiologists in identifying jaw lesions on OPGs, using histopathology as the gold standard.

Materials and Methods Study Design

This retrospective, cross-sectional diagnostic accuracy study compared the interpretation of orthopantomograms (OPGs) of jaw lesions by oral and maxillofacial radiologists with ChatGPT. Histopathology served as the gold standard for all diagnoses. The study followed methods from previous AI-based diagnostic research in radiology and dentistry⁹⁻¹².

Sample Selection

Fifty anonymized OPGs with histopathologically confirmed jaw lesions were collected from multiple CBCT diagnostic

centers. All radiographs were de-identified to maintain patient confidentiality.

Inclusion Criteria

- Patients of any age or sex with histopathologically confirmed jaw lesions (odontogenic, non-odontogenic cysts, benign and malignant tumors, fibro-osseous lesions).
- High-quality digital panoramic radiographs with clear resolution.
- Complete clinical and histopathological records.
- Exclusion Criteria
- Poor-quality radiographs (low resolution, distortion, or exposure errors).
- Missing or incomplete clinical or histopathological documentation.
- Recurrent lesions or prior surgical/radiographic interventions altering imaging features.

Radiologist Evaluation

Three experienced oral and maxillofacial radiologists independently reviewed all 50 OPGs and provided the most likely diagnosis for each case¹⁰.

ChatGPT Evaluation

Standardized text descriptions of each OPG (covering lesion site, borders, internal structure, and relation to nearby anatomy) were input into ChatGPT (GPT-4 model). Diagnostic terms were avoided, and the model generated the most likely diagnosis for each case. This design mirrors

approaches used in prior studies assessing LLMs with structured clinical and radiological case inputs.^{11,12}

Reference Standard

Histopathological confirmation was used as the definitive diagnostic reference for evaluating the accuracy of both radiologist and ChatGPT predictions¹¹.

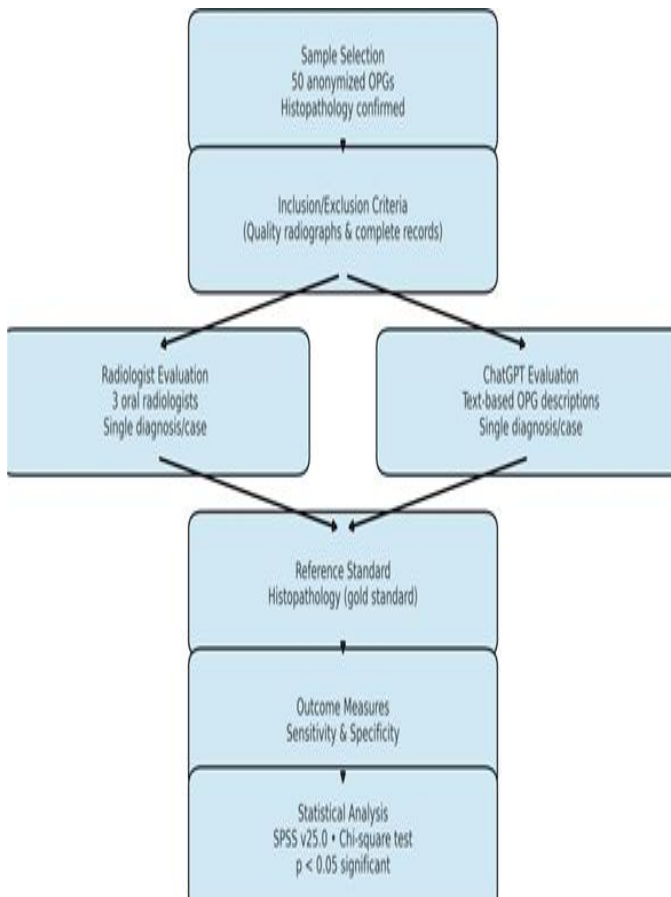
Outcome Measures

The main outcome measures were sensitivity and specificity for both radiologists and ChatGPT.

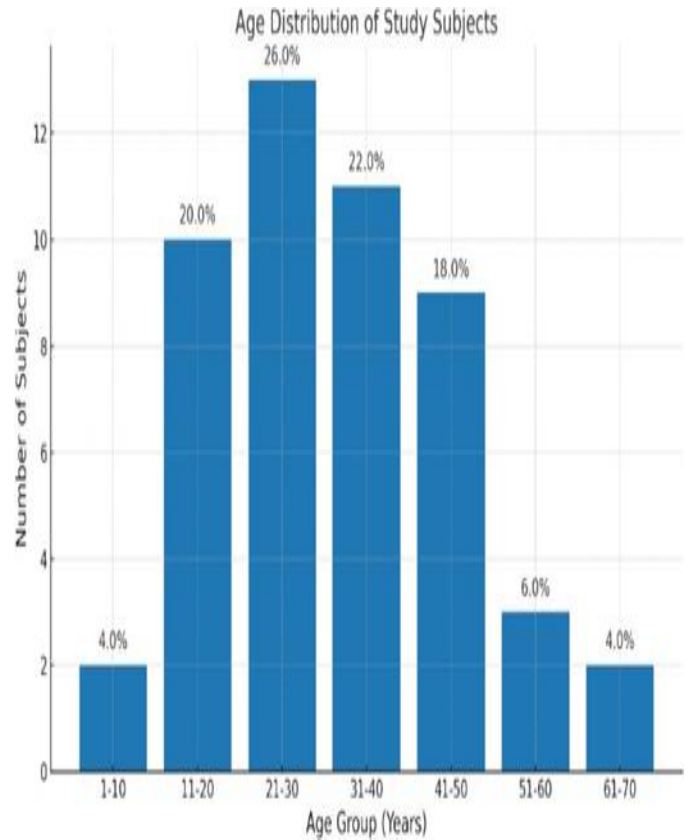
Statistical Analysis

Data were analyzed using SPSS version 25.0 (IBM Corp., Armonk, NY, USA). Results were expressed as frequencies and percentages. Diagnostic accuracy between radiologists and ChatGPT was compared with the Chi-square test^{12,14}. A p-value <0.05 was considered statistically significant.

Graphical Schematic



II. RESULTS



The average age of subjects was 32.6 years (range: 9–64), with most cases in the 21–30 (26%) and 31–40 (22%) age groups, showing higher prevalence in young and middle-aged adults. Both males and females were equally affected (50% each), indicating no gender bias.

Histopathology most often revealed odontogenic keratocyst (22%) and ameloblastoma (20%). Radiologists commonly diagnosed odontogenic keratocyst and dentigerous cyst (20% each), while ChatGPT most frequently diagnosed dentigerous cyst (36%) and ameloblastoma (26%).

Radiologists had fewer false positives, whereas ChatGPT showed more varied errors.

Accuracy rates were 36%, 50%, and 34% for the three radiologists, compared to 14% for ChatGPT. Overall, radiologists were significantly more accurate than ChatGPT, with no major differences among themselves.

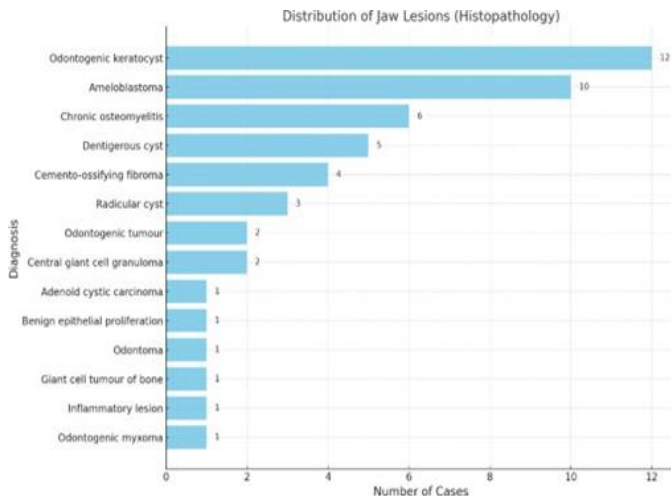
III. DIAGNOSTIC ACCURACY

Diagnosis	Diagnostic accuracy							
	Observer 1		Observer 2		Observer 3		Chat GPT	
	Value	95%CI	Value	95%CI	Value	95%CI	Value	95%CI
Adenoid cystic carcinoma	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%
Odontogenic tumour	98.00 %	89.35% to 99.95%	94.00 %	83.45% to 98.75%	94.00 %	83.45% to 98.75%	96.00 %	86.29% to 99.51%
Ameloblastoma	78.00 %	64.04% to 88.47%	82.00 %	68.56% to 91.42%	80.00 %	66.28% to 89.97%	66.00 %	51.23% to 78.79%
Benign epithelial proliferation	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%
Ossifying fibroma	92.00 %	80.77% to 97.78%	92.00 %	80.77% to 97.78%	92.00 %	80.77% to 97.78%	92.00 %	80.77% to 97.78%
Central giant cell granuloma	94.00 %	83.45% to 98.75%	92.00 %	80.77% to 97.78%	88.00 %	75.69% to 95.47%	96.00 %	86.29% to 99.51%
Chronic osteomyelitis	90.00 %	78.19% to 96.67%	94.00 %	83.45% to 98.75%	92.00 %	80.77% to 97.78%	88.00 %	75.69% to 95.47%
Odontoma	94.00 %	83.45% to 98.75%	96.00 %	86.29% to 99.51%	94.00 %	83.45% to 98.75%	88.00 %	75.69% to 95.47%
Dentigerous cyst	86.00 %	73.26% to 94.18%	92.00 %	80.77% to 97.78%	88.00 %	75.69% to 95.47%	62.00 %	47.17% to 75.35%
Giant cell tumour of bone	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%

Inflammatory lesion	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%
Odontogenic keratocyst	80.00 %	66.28% to 89.97%	88.00 %	75.69% to 95.47%	80.00 %	66.28% to 89.97%	70.00 %	55.39% to 82.14%
Odontogenic myxoma	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%	98.00 %	89.35% to 99.95%
Radicular cyst	84.00 %	70.89% to 92.83%	98.00 %	89.35% to 99.95%	94.00 %	83.45% to 98.75%	92.00 %	80.77% to 97.78%

Radiologists showed highest accuracy for odontogenic tumors (94–98%), odontomas (94–96%), and central giant cell granulomas (88–94%), with lower accuracy for ameloblastoma (78–82%) and dentigerous cysts (86–92%). ChatGPT performed well in select categories (e.g., central giant cell granuloma 96%, adenoid cystic carcinoma 98%) but underperformed in key lesions such as ameloblastoma (66%), dentigerous cyst (62%), and odontogenic keratocyst (70%). While radiologists demonstrated consistent superiority across most lesions, ChatGPT’s apparent accuracy in rare entities reflected negative case dominance rather than true diagnostic ability, limiting its clinical utility.

Comparison of correct and incorrect diagnosis (taking histopathology as gold standard) between dental surgeon’s and Chat GPT.

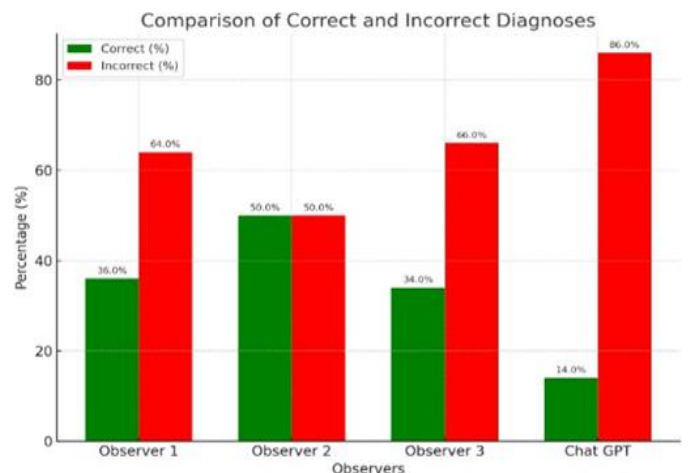


Radiologists showed highest accuracy for odontogenic tumors (94–98%), odontomas (94–96%), and central giant cell granulomas (88–94%), with lower accuracy for

ameloblastoma (78–82%) and dentigerous cysts (86–92%). ChatGPT performed well in select categories (e.g., central giant cell granuloma 96%, adenoid cystic carcinoma 98%) but underperformed in key lesions such as ameloblastoma (66%), dentigerous cyst (62%), and odontogenic keratocyst (70%). While

radiologists demonstrated consistent superiority across most lesions, ChatGPT’s apparent accuracy in rare entities reflected negative case dominance rather than true diagnostic ability, limiting its clinical utility.

Comparison of correct and incorrect diagnosis (taking histopathology as gold standard) between dental surgeon’s and Chat GPT.



Diagnosis	Observer 1		Observer 2		Observer 3		Chat GPT		Cochran Q	p-value
	Number of subjects	Percentage	Number of subjects	Percentage	Number of subjects	Percentage	Number of subjects	Percentage		
Correct	18	36.0	25	50.0	17	34.0	7	14.0	19.970	<.001
Incorrect	32	64.0	25	50.0	33	66.0	43	86.0		
Total	50	100.0	50	100.0	50	100.0	50	100.0		

Pair-wise comparison of accurate diagnosis by dental surgeon's and Chat GPT.

Pairwise	p-value
Observer 1 vs. observer 2	.118
Observer 1 vs. observer 3	1.000
Observer 1 vs. Chat GPT	.007*
Observer 2 vs. observer 3	.096
Observer 2 vs. Chat GPT	<.001*
Observer 3 vs. Chat GPT	.013*

IV. EVALUATION

This study compared ChatGPT's diagnostic accuracy in interpreting orthopantomograms of histopathologically confirmed jaw lesions with that of oral and maxillofacial radiologists. Radiologists showed consistently higher accuracy (78–98%), while ChatGPT performed lower for key lesions such as dentigerous cysts (62%) and ameloblastoma (66%). Although ChatGPT provided logical explanations, it often overpredicted common lesions and missed rare ones, showing that LLMs, while promising, are not yet reliable for image-based diagnosis^{1,2}.

Radiologists' Superior Diagnostic Performance

Radiologists outperformed ChatGPT in almost all categories, especially ameloblastoma (78–82%) and dentigerous cysts (86–92%), due to their ability to recognize subtle radiographic details. Similar findings were reported by Mitsuyama et al.³ and Qu et al.⁴, underscoring the irreplaceable role of radiologists. Tassoker et al.⁵ and Pradhan⁶ also identified radiologists as the diagnostic gold standard in oral imaging.

ChatGPT's Diagnostic Limitations

ChatGPT achieved moderate accuracy in some categories (e.g., central giant cell granuloma 96%, adenoid cystic carcinoma 98%), but its performance dropped significantly for clinically important lesions such as ameloblastoma (66%), dentigerous cyst (62%), and odontogenic keratocyst (70%). Overprediction of common entities and false-positive diagnoses (e.g., residual cyst, squamous cell carcinoma, cherubism) were frequent. These limitations align with Schmidt et al.⁷ and Hong et al.⁸, who noted that ChatGPT often generates confidently incorrect outputs and lacks nuance in imaging-driven tasks. Unlike convolutional neural networks (CNNs), which directly process image features, ChatGPT relied only on text-based descriptions, reducing sensitivity by Durmazpinar and Ekmekci⁹. This contrast highlights the need for multimodal AI systems that integrate image-level interpretation with natural language reasoning.

Lesion-Specific Diagnostic Trends

Histopathology most often showed odontogenic keratocyst (22%) and ameloblastoma (20%). Radiologists diagnosed these accurately, while ChatGPT underperformed. Its high accuracy in rare lesions likely reflected dataset bias rather than true diagnostic ability. Similar results were seen by Sunet al.¹² and Eriksen et al.¹³, who reported LLMs overgeneralize and struggle with complex radiographs. Guimaraes et al.¹⁴ also showed that AI struggles with overlapping features, reinforcing radiologists' importance.

Comparison with Literature

Our findings align with studies showing ChatGPT is better as a support tool than a standalone system (Mago & Sharma¹⁰). Domain-specific tuning improves performance^{5,6}, but general LLMs remain limited. Khan & O'Sullivan¹¹ and Choi et al.¹⁵ found ChatGPT effective in structured case vignettes, highlighting its reasoning strengths but imaging weaknesses. Thirunavukarasu et al.¹⁶ showed ChatGPT's reasoning matches medical student level but falls short in radiology. CNN-based multimodal models, as shown by Durmazpinar & Ekmekci⁹, achieved ~90% accuracy, exceeding ChatGPT.

Clinical Implications

ChatGPT should not be used as a primary diagnostic tool in oral and maxillofacial radiology. Its utility lies in generating differential diagnoses, supporting dental education, and serving as a triage aid when paired with specialized datasets^{8,10}. Radiologists remain essential for definitive diagnosis and management planning.

Future Research

- Domain-specific AI training with curated, annotated dental imaging datasets^{5,6}.
- Multimodal AI development integrating imaging and text-based reasoning⁹.
- Explainable AI tools to enhance clinician trust.
- Large multi-center studies including rare lesions for benchmarking.

V. CONCLUSION

This study demonstrates that radiologists remain superior to ChatGPT in interpreting jaw pathologies on OPGs, achieving higher sensitivity, specificity, and diagnostic accuracy. Our results align with literature emphasizing ChatGPT's educational potential but diagnostic limitations^{3,4,7,12,13,14}. While ChatGPT is a promising adjunct, future efforts should focus on multimodal AI and domain training to achieve clinically reliable results. Until then, expert radiologist interpretation with histopathological correlation remains the diagnostic gold standard.

REFERENCES

1. Mykhalko YO, Filak YF, Dutkevych-Ivanska YV, Sabadosh MV, Rubtsova YI. From open-ended to multiple-choice: evaluating diagnostic performance and consistency of ChatGPT, Google Gemini and Claude AL. *Wiad Lek.* 2024;77(9):1852–1856. doi:10.36740/WLek/195125.
2. Panwar P, Gupta S. A review: Exploring the role of ChatGPT in the diagnosis and treatment of oral pathologies. *Oral Oncology Reports.* 2024 Jun 1;10:100225.
3. Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. *Eur Radiol.* 2025;35:1938–1947. doi:10.1007/s00330-024-11032-8.
4. Qu RW, et al. Diagnostic and management applications of ChatGPT in structured otolaryngology scenarios. *OTO Open.* 2023;7(3):e67.
5. Tassoker M. Exploring ChatGPT's potential in diagnosing oral and maxillofacial pathologies: a study of 123 challenging cases. *BMC Oral Health.* 2025;25:1187. doi:10.1186/s12903-025-06444-x.
6. Pradhan P. Accuracy of ChatGPT 3.5, 4.0, 4o and Gemini in diagnosing oral potentially malignant lesions. *Med Oral Patol Oral Cir Bucal.* 2025;30(2):e224–e231. doi:10.4317/medoral.26824.
7. Schmidt HG, Rotgans JI, Mamede S. Bias sensitivity in diagnostic decision-making: comparing ChatGPT with residents. *J Gen Intern Med.* 2024;40(4):790–795. doi:10.1007/s11606-024-09177-9.
8. Hong D-R, Huang C-Y, Zhong H-H, Lyu G-R. ChatGPT-4 Vision: a promising tool for diagnosing thyroid nodules. *Front Med.* 2025;12:1634976. doi:10.3389/fmed.2025.1634976.
9. Durmazpinar PM, Ekmekci E. Comparing diagnostic skills in endodontic cases: dental students versus ChatGPT-4o. *BMC Oral Health.* 2025;25:457. doi:10.1186/s12903-025-05857-y.
10. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. *Cureus.* 2023;15(7):e42133. doi:10.7759/cureus.42133.
11. Khan MP, O'Sullivan ED. ChatGPT in medical education and clinical decision-making. *Front Artif Intell.* 2024;7:1379297. doi:10.3389/frai.2024.1379297.
12. Sun SH, Liu H, Ma Y, et al. Testing the ability and limitations of ChatGPT to generate differential diagnoses from transcribed clinical cases. *Radiology.* 2024;313(1):e232346. doi:10.1148/radiol.232346.
13. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *Nejm Ai.* 2024 Jan 1;1(1):AIp2300031.
14. Guimaraes GR, Silva CS, Contreras JC, Figueiredo RG, Uros-Grupo de Pesquisa, Tiraboschi RB, Gomes CM, de Bessa J. Diagnosis in Bytes: Comparing the Diagnostic Accuracy of Google and ChatGPT 3.5 as Diagnostic Support Tools. *medRxiv.* 2023 Nov 12:2023-11.
15. Choi J, Hong DR, Kim JH, et al. Application of ChatGPT for diagnostic support in oral pathology: potential and challenges. *Front Med.* 2025;12:1634976.
16. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine: current applications and future directions. *NEJM AI.* 2024;1(1):AIp2300031. doi:10.1056/AIp2300031.