Original Research Article

Diagnostic Precision of Chat GPT-5.2 in Analyzing Oral Histopathological Images

Syed Fareed Mohsin[1*], Abdullah Fahad Abdullah Al-Moshwih[2]

[1]Associate Professor, Oral Pathology, Department of Oral and Maxillofacial Diagnostic Sciences, College of Dentistry, Qassim University, Buraydah, Kingdom of Saudi Arabia
[2]Dental Intern, College of Dentistry, Qassim University, Buraydah, Kingdom of Saudi Arabia
[*]Corresponding author email: **s.syedabdulmohsin@qu.edu.sa**

International Archives of Integrated Medicine, Vol. 13, Issue 6, June, 2026.
Available online at http://iaimjournal.com/
ISSN: 2394-0026 (P) **ISSN: 2394-0034 (O)**
Received on: 5-6-2026 **Accepted on:** 14-6-2026
Source of support: Nil **Conflict of interest:** None declared.

DOI: **10.5281/zenodo.20986905**

How to cite this article: Syed Fareed Mohsin, Abdullah Fahad Abdullah Al-Moshwih. Diagnostic Precision of Chat GPT-5.2 in Analyzing Oral Histopathological Images. Int. Arch. Integr. Med., 2026; 13(6): 1-7.

Abstract

Background and purpose: Artificial intelligence (AI) has been increasingly popular in recent years as a diagnostic tool. It can assist in medical and dental workflow to support clinical decisions. A multimodal large language model (LLM) like ChatGPT-5.2, has advanced processing capabilities which can aid in interpreting histopathology images.

Objective: This study assessed the diagnostic precision of ChatGPT-5.2 in detecting oral histopathology images.

Materials and methods: The study included seventy histological images sourced from a standard reference textbook "Oral Histology and Oral Histopathology" Springer Nature, 2025. Three subjects experts in oral pathology independently assessed the images for the final diagnosis. Comparative analysis of ChatGPT-5.2 versus expert consensus, utilizes Cochran's Q test and Post hoc McNemar with Bonferroni's, whereas Light kappa, Fleiss Kappa and Gwet's AC1 were applied for chance-corrected agreement.

Results: ChatGPT-5.2 correctly diagnosed 65 oral histological images, achieving 92.9% accuracy. Agreement among three experts and ChatGPT was 64.3% however Fleiss kappa showed a significant agreement (0.159, p=0.001) and Gwet's AC1 demonstrated substantial agreement (80.7% <0.001). Cochran's Q revealed a significant difference (p value 0.003). Post hoc McNemar with Bonferroni's test found significant differences between expert 2 and expert 3 with ChatGPT 5.2. Overall ChatGPT5.2 attained closely comparable score when compared to the highest-scoring expert.

Conclusion: This study found that use of ChatGPT-5.2 in oral and maxillofacial pathology can act as an supplementary educational and decision support tool, while experts retains full responsibility for

final diagnosis. Further validation is required using large datasets, multiple AI model and real world histopathology cases are necessary before its integration into clinical decisions.

Key words

ChatGPT-5.2; Oral Histopathology; Diagnostic Accuracy; Artificial Intelligence in Dentistry; Multimodal Large Language Models.

Introduction

In recent years, there has been an immense increase in the use of artificial intelligence (AI), a set of technologies across multiple disciplines that aim to mimic human intelligence [1]. Machine learning (ML) is a branch of AI which learn pattern from the data and making accurate predictions overtime. Deep learning is a subtype of ML that uses multilayered neural network language input and plays a vital role in supporting clinical decisions [2, 3].

A shift in AI happened in 2017 by the introduction of transformer architecture; this innovation paved the way for the foundation of large language model (LLM). These developments have made a remarkable fluency in generating human language [4]. The first widely available LLM was ChatGPT-3.5. Since then, other more models have been developed, like DeepSeek (DeepSeek AI), Claude (Anthropic), Gemini (Google), Copilot (Microsoft), Mistral (Mistral AI), and Llama (Meta) [5].

Healthcare has been influenced by AI which helps in enhancing diagnostic precision, clinical decision and patient management. ML algorithms have been trained on large data sets which enable pathologist to accurately and quickly analyze tissue samples. These models can detect various diseases and infections by identifying their pattern and abnormalities. ChatGPT can understand the text in a way comparable to human abilities which enables it to be well known in dental and medical field due to its quick, organized clinical insight and decision making [2, 6].

Oral diagnosis remains a challenge in dentistry particularly when it needs histopathological image analyzing. The difference between benign and malignant leads to diagnostic uncertainty which results in underdiagnosis and ultimately delayed treatment. The use of LLM offers a fascinating tool for disease identification and can lead to enhance diagnostic rate [7].

In dentistry AI can help students and practitioners in guessing the differential diagnosis based on providing detailed information in the form of clinical scenario and laboratory investigations to the LLM. Various studies on ChatGPT diagnostic reliability have been tested, however depending on the model version the accuracy ranges from 57% to 80%. There is shortage of qualified pathologists around the globe and due to the burden of diseases the clinicians are overburdened which results in misdiagnosis. As the model has capability to scan slides with accuracy and speed its application will lead to reduce burden on pathologist to some extent [8-12].

It is essential to evaluate ChatGPT in general and specific pathology images, therefore this study was conducted to assess ChatGPT5.2 ability to diagnose histopathological images sourced from a standard textbook.

Materials and methods

This cross sectional observational study assessed the diagnostic performance of ChatGPT 5.2. The model was used to evaluate histopathology images. A standard reference textbook of Oral Histology and Oral Histopathology: A Practical Guide for Dental Students and a Companion to Pathologists (Springer Nature, 2025) [13] was utilized to acquire high quality well annotated histological images that is needed for diagnostic training and evaluation. The images were strictly

used for academic research purpose. Altogether 70 histological images were randomly selected from each chapter of book which includes 60 histopathological images depicting common and less common diseases, and 10 normal histological images.

The high and low power photomicrograph were selected to the region of interest without compromising diagnostic features and saved as JPEG format. Each image was accompanied by limited information about the disease. The chat history was deleted before a new session was commenced in order to eliminate contextual bias. The images of high resolution were included whereas blurred or poorly focused were excluded.

Data collection

A standardized prompt condition was developed and model was instructed to evaluate each query accompanied by image with description and provide the most probable diagnosis. The following prompt was used "What is your final diagnosis based on the information provided?". All responses given by ChatGPT have been documented verbatim.

The ChatGPT 5.2 was selected due to its multimodal capability to analyze both textual and image based request which is required for histopathological evaluation. It offers improved contextual understanding and structured response generation which makes it suitable for complex diagnostic queries. Same images were independently evaluated by three subject expert and their diagnosis served as a reference standard for correctness. Each response was given a score of 1 for correct answer and 0 for incorrect answer. The final diagnosis (gold standard) of each case was established using histopathological confirmation. All cases were independently assessed by three raters who were blinded to the reference standard.

This study does not require ethical approval, as it did not involve human participants or animal subjects and relied exclusively on publicly accessible software and fully anonymized text book images.

Statistical Analysis

SPSS Version 26 was used to analyze data. ChatGPT 5.2 diagnostic accuracy of final diagnosis was documented as frequency and percentage table, whereas the Cochran's Q test was used to assess the differences in the responses by all evaluators, Post hoc McNemar test with Bonferroni's method was applied and p-values were adjusted for multiple comparisons. Lightkappa, Fleiss Kappa and Gwet's AC1 were tabulated for chance-corrected agreement. Statistical significance was determined at $p < 0.05$.

Results

Cochran's Q test was used to assess the differences in responses by all four evaluators. Cochran's Q showed a significant difference (p value 0.003). Post hoc McNemar test was applied and p-values were adjusted for multiple comparisons using Bonferroni's correction. It was found, that there were significant differences between expert 2 and expert 3 with ChatGPT (**Table - 1**).

Overall percent agreement among three experts and ChatGPT was 64.3%. However, Light's Kappa was 0.201 (p-value 0.963) showed a fair but non-significant chance-corrected agreement. Fleiss Kappa also showed slight but significant agreement (0.159, p-value 0.001). Gwet's AC1 demonstrated substantial inter-rater agreement (AC1 = 0.75, 95% CI: 0.65–0.85), with an observed agreement of 80.7%.

This study included a wide spectrum of oral lesions with multiple diagnostic categories, many of which had low frequencies. In such multi-category settings with sparse observations per category, kappa statistics (including Fleiss' Kappa and Light's Kappa) are known to yield low values despite high accuracy. In this study, experts demonstrated high agreement with the gold standard (validity), but variability in

assigning specific lesion subtypes reduced inter-rater agreement (reliability). This reflects the inherent complexity of multi-class oral lesion diagnosis rather than inconsistency in clinical competence. In order to provide robust measure of percent agreement this study also applied Gwet's AC1 which was not affected by category prevalence (**Table - 2**).

Figure - 1 represents the proportion of correct interpretation of each slide by each of the evaluator. Expert 1 has the highest proportion of correct answers (94.3%) followed by ChatGPT (92.9%). Expert number 2 had the least score, 78.6%.

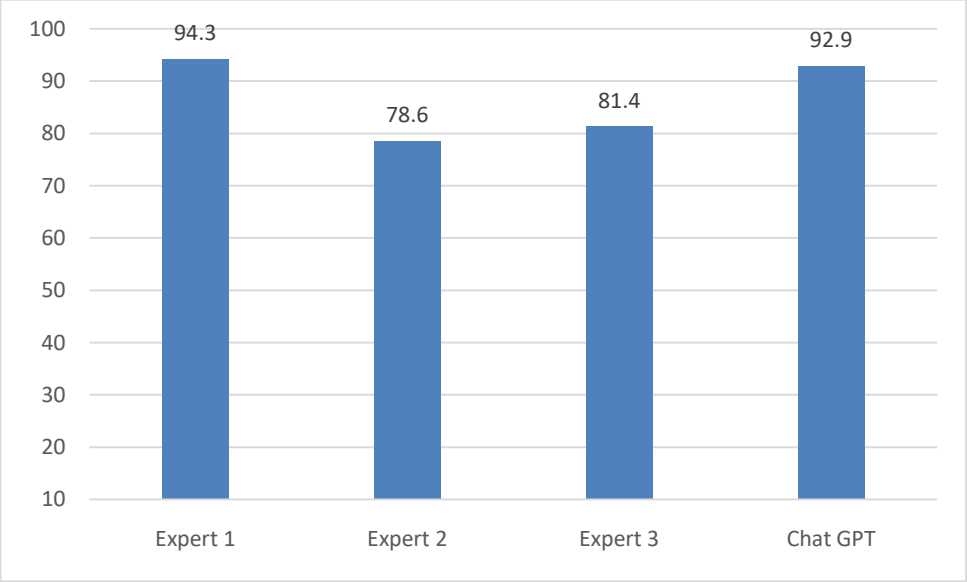
Table - 1: Comparison of Experts' diagnosis with Chat GPT.

Evaluators	**Correct % (n)**	**Incorrect % (n)**	**P-value£**
Expert 1	94.3 (66)	5.7 (4)	0.936
Expert 2	78.6 (55)	21.4 (15)	0.025
Expert 3	81.4 (57)	18.6 (13)	0.029
Chat GPT 5.2	92.9 (65)	7.1 (5)	NA
£ McNemar p-value in comparison with Chat GPT, NA; not applicable			

Table - 2: Inter-rater agreement.

Measure	**Value**	**P-value**
Simple agreement	64.3%	NA
Light's Kappa	0.201	0.963
Fleiss Kappa	0.159	0.001
Gwet's AC1	75%	<0.001

Figure - 1: Proportion of correct interpretation of slides.



Discussion

This study compared the diagnostic accuracy of ChatGPT 5.2 in diagnosing histopathological images compared with three subject experts. The model performed well 92.9% and was consistent in structured diagnostic task when compared to subject expert 1 (94.3%), although the differences between evaluators were not statistically significant, indicating similar diagnostic performance, whereas high diagnostic

accuracy observed. Fleiss Kappa also showed slight but significant agreement this occurs because kappa is sensitive to category prevalence and marginal distributions, and treats all disagreements equally even between clinically similar categories therefore Gwet's AC1 was applied which was not affected by category prevalence and showed a significant inter rater agreement.

Study findings indicate that ChatGPT-5.2's overall diagnostic accuracy was 92.9%. Our findings are in line with recent study which evaluated 12 histopathological images revealed that Chat GPT 4 has strong capabilities in interpreting histopathology images [14]. Another study highlights 88% diagnostic accuracy of Chat GPT-4 in diagnosing cases with histopathological description [15].

The study revealed a noteworthy observation which was variation in diagnostic accuracy by subject experts, these phenomena has been observed in other studies conducted in clinical settings. Due to the diverse training experience and borderline interpretation of cases may impact the final diagnosis therefore careful consideration is required when using ChatGPT 5.2 to support physician decision-making [16].

Study findings indicate that AI can effectively assist in histopathological diagnosis. Our findings are in line with previous studies which include random selection of queries from the text book *Clinical Guide to Oral Disease* has 87.4% overall accuracy rate. In addition, other studies also demonstrated that AI models can achieve diagnostic precision in well-defined histopathological features. Based on these findings AI tools can be used as a supplementary aid in diagnostic workflow [17-19].

The findings of this study align with a study conducted on 100 pathological lesions from google images and 10 healthy mucosa reveals 91.7% of correct responses which show that ChatGPT has appropriate diagnostic accuracy and its application may reduce workload and human errors [20].

ChatGPT-5.2 fast processing ability supports health care professionals by providing evidence-based dentistry approach reducing human error and burden. However, it may influence result and pose challenges for less experienced practitioners. The current performance suggest that it can be used for generating differential diagnosis and educational feedback while expert OMF pathologist maintain full responsibility for final diagnosis. AI is still in development stage and it can be use as a supplementary educational or decision support tool, further research is necessary to determine its role in diagnosis and management in healthcare [21, 22].

Limitations and Future Directions

The study has small sample size which limits it generalizability and its dependence on histopathological images from text book only. Secondly the study evaluated only a single artificial intelligence model. Future studies should therefore include, large sample size, use of multiple AI models and real world histopathological cases.

Conclusion

This study found that use of ChatGPT-5.2 in oral and maxillofacial pathology can act as an supplementary educational and decision support tool, while experts retains full responsibility for final diagnosis. Further validation is required using large datasets, multiple AI model and real world histopathology cases are necessary before its integration into clinical decisions.

References

1. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res., 2020; 99(7): 769–74.
2. Zhou B, Yang G, Shi Z, Ma S. Natural Language processing for smart health-care. IEEE Rev Biomed Eng., 2024; 17: 4–18.

3. Chen JH, Dhaliwal G, Yang D. Decoding artificial intelligence to achieve diagnostic excellence: learning from experts, examples, and experience. JAMA, 2022; 328(8): 709–10.
4. Takita H., Kabata D., Walston S.L., et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit. Med., 2025; 8: 175. https://doi.org/10.1038/s41746-025-01543-z
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al. Attention is all you need. Adv Neural Inf Process Syst., 2017; 30: 1–15.
6. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent., 2023; 35(7): 1098–102.
7. M.B. Stephens, J.P. Wiedemer, G.M. Kushner. Dental problems in primary care Am Fam Physician, 2018; 98: 654-660
8. Farhadi Nia M, Ahmadi M, Irankhah E. Transforming dental diagnostics with artificial intelligence: advanced integration of ChatGPT and large language models for patient care. Front Dent Med., 2025; 5: 1456208.
9. Tomo S, Lechien JR, Bueno HS, Cantieri-Debortoli DF, Simonato LE. Accuracy and consistency of ChatGPT-3.5 and 4 in providing differential diagnoses in oral and maxillofacial diseases: a comparative diagnostic performance analysis. Clin Oral Investig., 2024; 28(10): 5445–53.
10. Shafi S, Parwani AV. Artificial intelligence in diagnostic pathology. DiagnPathol., 2023; 18: 109.
11. Van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: The path to the clinic. Nat Med., 2021; 27: 775–84. doi: 10.1038/s41591-021-01343-4
12. Kim I, Kang K, Song Y, Kim TJ. Application of artificial intelligence in pathology: Trends and challenges. Diagnostics (Basel)., 2022; 12: 2794.
13. Balaji, S. M.. Oral Histology and Oral Histopathology – A Practical Guide for Dental Students and a Companion to Pathologists. Indian Journal of Dental Research, 2025; 36(2): 249-250. DOI: 10.4103/ijdr.ijdr_589_25
14. Gumilar KE, Ariani G, Wiratama PA, Rimbun, Yuliawati TH, Chen H, Ibrahim IH, Lin CH, Hung TY, Anggrahini D, Rajanagara AS, Omran KE, Yu ZY, Hsu YC, Dachlan EG, Yang JY, Liao LN, Tan M. Assessing the capabilities of AI-based large language models (AI-LLMs) in interpreting histopathological slides and scientific figures: Performance evaluation study. Biomedicine (Taipei)., 2026 Mar 1; 16(1): 41–52. doi: 10.37796/2211-8039.1698. PMCID: PMC12962759.
15. Mazzucchelli M, Salzano S, Caltabiano R, Magro G, Certo F, Barbagallo G, Broggi G. Diagnostic Performance of ChatGPT-4.0 in Histopathological Analysis of Gliomas: A Single Institution Experience. Neuropathology, 2025 Aug; 45(4): e70023. doi: 10.1111/neup.70023. PMID: 40726356; PMCID: PMC12305399.
16. Nori H, et al. Capabilities of GPT-4 for clinical decision support. JAMA Network Open, 2023; 6(10): e2336483.
17. Hajibagheri P, Sani SK, Samami M, Tabari-Khomeiran R, Azadpeyma K, Sani MK. ChatGpt's accuracy in the diagnosis of oral lesions. BMC Oral Health, 2025 Jul 21; 25(1): 1229. doi: 10.1186/s12903-025-06582-2. PMID: 40691556; PMCID: PMC12281746.
18. Islam A, Banerjee A, Wati SM, Banerjee S, Shrivastava D, Srivastava KC. Utilizing Artificial Intelligence Application for Diagnosis of Oral Lesions and Assisting Young Oral Histopathologist in Deriving Diagnosis

from Provided Features - A Pilot study. J Pharm Bioallied Sci., 2024 Apr; 16(Suppl 2): S1136-S1139. doi: 10.4103/jpbs.jpbs_1287_23. Epub 2024 Apr 16. PMID: 38882904; PMCID: PMC11174333.

19. Uranbey Ö, et al. Assessing ChatGPT's diagnostic accuracy and therapeutic strategies in oral pathologies: a cross-sectional study. Cureus, 2024; 16(4): e58607.
20. Vaira L.A., Lechien J.R., Maniaci A., De Vito A., Mayo-Yáñez M., Troise S., Consorti G., Chiesa-Estomba C.M., Cammaroto G., Radulesco T., et al. Diagnostic Performance of ChatGPT-4o in Analyzing Oral Mucosal Lesions: A Comparative Study with Experts. Medicina, 2025; 611379. https://doi.org/10.3390/medicina61081379
21. Kung TH, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education. PLOS Digital Health, 2023; 2(2): e0000198
22. Nguyen VA, Nguyen VH, Vuong TQT, Truong QT, Nguyen TT. Comparative study of advanced reasoning versus baseline large-language models for histopathological diagnosis in oral and maxillofacial pathology. PLoS One, 2025 Dec 31; 20(12): e0340220. doi: 10.1371/journal.pone.0340220. PMID: 41474818; PMCID: PMC12755752.