TMSJ
TURKISH MEDICAL STUDENT JOURNAL

Turk Med Stud J 2026;13(1):13-20
DOI: 10.4274/tmsj.galenos.2026.2025-10-1

ORIGINAL ARTICLE

EXPLORING THE DIAGNOSTIC LIMITS OF CHATGPT: HOW FAR CAN A LARGE LANGUAGE MODEL GO IN HISTOPATHOLOGICAL IMAGE INTERPRETATION?

Aghajan Musali[1], Jamal Musayev[2,3]

[1]Yozgat Bozok University School of Medicine, Yozgat, TÜRKİYE
[2]Baku Pathology Center, Baku, AZERBAIJAN
[3]Karabakh University School of Medicine and Health Sciences, Khankendi, AZERBAIJAN

ABSTRACT

Aims: Artificial intelligence's integration into pathology has accelerated with the adoption of digital workflows. Large language models like ChatGPT offer unique opportunities but have yet to be systematically evaluated in diagnostic image interpretation.

Methods: In this comparative study, 24 histopathological images representing various tissue types and pathological entities were evaluated by ChatGPT-4o mini and 15 experienced pathologists. The model was prompted with a standard diagnostic query without access to clinical information. Pathologists independently assessed the same images. Responses were categorized as correct, false positive, false negative, low-impact error, or no interpretation. Standard diagnostic metrics were calculated, and group comparisons were conducted using McNemar's test and Fisher's exact test. Interobserver agreement among pathologists was analyzed using Fleiss' kappa.

Results: ChatGPT-4o mini achieved an accuracy of 71.4%, with a sensitivity of 60.0% and a specificity of 77.8%. The average accuracy of pathologists was 89.8%, with 97.7% sensitivity and 87.1% specificity. Low-impact errors were more frequent with ChatGPT-4o mini (33.3%) compared to pathologists (6.9%). McNemar's test revealed a statistically significant difference in favor of pathologists. The interobserver agreement among pathologists was in the lower range.

Conclusion: While ChatGPT-4o mini demonstrated partial diagnostic capabilities, it underperformed compared to experienced pathologists. The absence of a clinical context likely impacted the results. Future artificial intelligence models integrating image analysis and clinical data may enhance performance. Despite limitations, the potential ChatGPT holds as a supportive diagnostic tool in pathology is highlighted in this study.

Keywords: Artificial intelligence, generative artificial intelligence, microscopy, pathology

INTRODUCTION

ChatGPT is an artificial intelligence (AI) virtual assistant launched in November 2022, with its applications in medicine rapidly expanding. The first Food and Drug Administration-approved AI application in medicine emerged in 2017 with software designed to detect diabetic retinopathy (1). More recently, large language models (LLMs) such as ChatGPT have gained attention for their potential applications in pathology and other medical fields. The advancement and widespread adoption of digital pathology have further facilitated AI integration into diagnostic pathology (2).

Unlike AI-driven image analysis platforms such as PathAI or Google Health AI, ChatGPT does not perform direct image interpretation. Instead, it provides text-based insights that may complement pathology assessments when combined with expert knowledge. Thus, the number of studies evaluating its potential in pathology is steadily increasing, highlighting the need for a systematic assessment of its capabilities and limitations (3-6). While ChatGPT demonstrates proficiency

Address for Correspondence: Aghajan Musali, Yozgat Bozok University School of Medicine, Yozgat, TÜRKİYE
e-mail: agacan.musali@gmail.com
ORCID iD of the authors: AM: 0009-0008-1792-576X; JM: 0000-0002-9202-6990
Received: 04.10.2025 Accepted: 25.02.2026 Publication Date: 27.02.2026

Cite this article as: Musali A, Musayev J. Exploring the diagnostic limits of ChatGPT: how far can a large language model go in histopathological image interpretation? Turk Med Stud J. 2026;13(1):13-20.

in certain specialized fields, such as bone pathology, its performance relies primarily on theoretical knowledge rather than direct image analysis (7, 8). Consequently, its practical application in diagnostic pathology remains constrained.

Moreover, limitations such as the lack of deep critical thinking and reasoning, as well as the inability to verify source reliability, necessitate cautious use. Expert validation remains essential (9). Additionally, ChatGPT's diagnostic accuracy is significantly influenced by the availability of contextual information; in the absence of clinical data, its performance in microscopic image analysis is markedly reduced (10).

Despite these limitations, ChatGPT has unique attributes that explain why it was chosen for the present study. It is the most widely known and most accessible LLM in medical and educational settings, with many clinicians and trainees already experimenting with its image-upload function, despite its limitations as a general-purpose tool not specifically trained for histopathological image analysis. This accessibility and widespread informal use created the rationale for our study: to systematically evaluate ChatGPT's diagnostic performance against experienced pathologists under controlled conditions.

Recent works illustrate the growing interest in this area. Vaira et al. (11) compared ChatGPT with experts in the analysis of oral mucosal lesions, reporting lower accuracy with ChatGPT than human specialists. Mazzucchelli et al. (12) assessed ChatGPT's diagnostic ability in glioma histopathology, with similar findings. In cervical cytology, Laohawetwanit et al. (13) confirmed that LLMs can function as supportive tools but cannot replace expert evaluation. In addition to empirical evaluations, conceptual discussions have questioned whether ChatGPT should have a role in medicine at all, emphasizing both its potential utility and its inherent limitations (14, 15). Together, these studies highlight the research gap our work addresses: to benchmark ChatGPT against trained pathologists in a structured, case-based design.

In this study, we evaluated the ability of AI to interpret microscopic images of human tissues and compared its performance with that of experienced pathologists.

MATERIALS AND METHODS

Ethics Statement

This study did not involve identifiable patient data. All histopathological images were de-identified archival cases and were used solely for research and educational purposes. According to institutional and international ethical guidelines, a formal approval from an ethics committee was therefore not required. All participating pathologists took part voluntarily and provided informed consent to contribute their diagnostic assessments.

Tissue Sample Selection

This study was designed to explore and assess the performance of an AI rather than to establish definitive clinical conclusions. This study analyzed static microscopic images of 24 tissue samples, categorized into four groups:

- **Normal tissue samples (n=6):** Ureteral wall, thyroid, prostate, gastric mucosa, gallbladder wall, colonic mucosa (Figure 1).
- **Non-neoplastic lesions (n=6):** Helicobacter pylori gastritis, fibrous dysplasia of bone, cystitis cystica, ulcerative colitis, chronic lymphocytic thyroiditis, Sertoli cell-only syndrome (Figure 2 a-f).
- **Benign tumors (n=6):** Intradermal nevus, meningioma, osteochondroma, pleomorphic adenoma, intraductal papilloma of the breast, tubular adenoma of the colon (Figure 3).
- **Malignant tumors (n=6):** Papillary urothelial carcinoma, mucinous carcinoma, adenocarcinoma of the prostate, renal cell carcinoma, squamous cell carcinoma, papillary thyroid carcinoma (Figure 4).

This study was designed to explore and assess the performance of a LLM rather than to establish definitive clinical conclusions.

Image Acquisition and Processing

All images were captured using a Leica DM1000 microscope equipped with an ICC50 camera and proprietary software. The magnification levels were selected based on the histopathological characteristics of each tissue, ensuring the

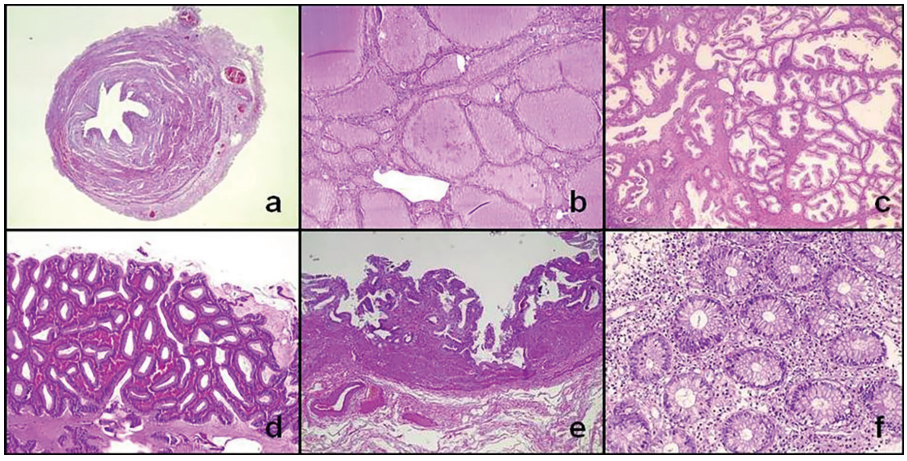

Figure 1: Normal tissue samples: A cross-section of the ureter showing the lumen and the ureteral wall (a), thyroid follicles (b), prostatic acini (c), gastric mucosa (d), mucosa and submucosa of the gallbladder (e), colonic mucosa (f).

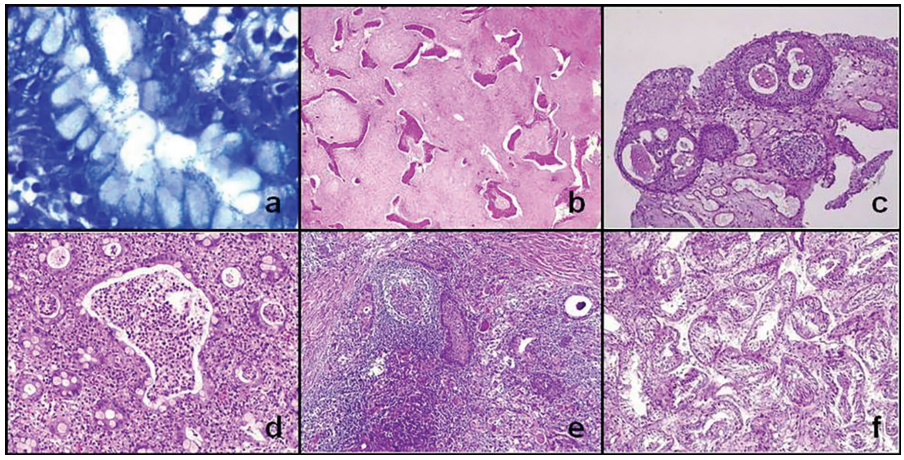

Figure 2: Non-neoplastic lesions: Helicobacter pylori gastritis (a), fibrous dysplasia of bone (b), cystitis cystica (c), ulcerative colitis (d), chronic lymphocytic thyroiditis (e), Sertoli cell-only syndrome (f).

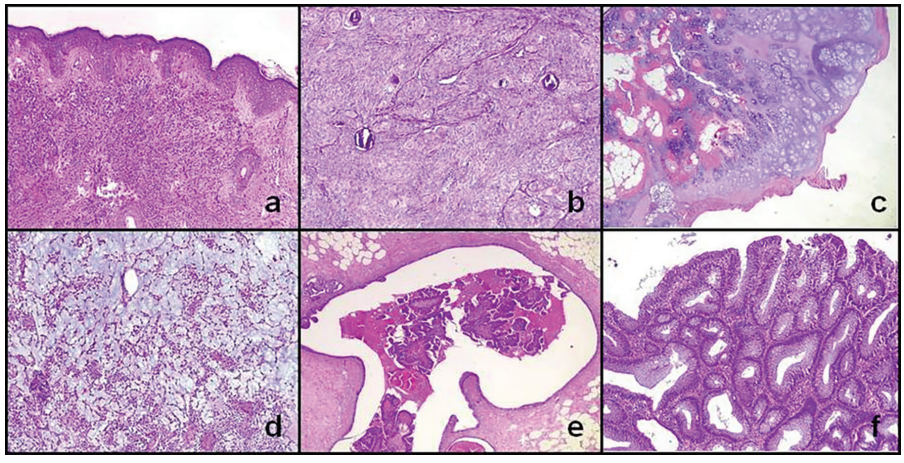

Figure 3: Benign tumors: Intradermal nevus (a), meningioma (b), osteochondroma (c), pleomorphic adenoma (d), intraductal papilloma of the breast (e), tubular adenoma of the colon (f).

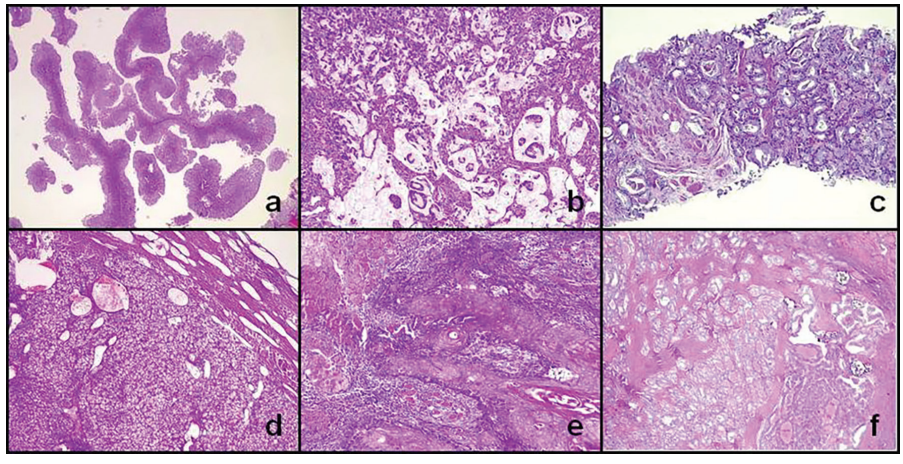

Figure 4: Malignant tumors: Papillary urothelial carcinoma (a), mucinous carcinoma (b), adenocarcinoma of the prostate (c), renal cell carcinoma (d), squamous cell carcinoma (e), papillary thyroid carcinoma (f).

most representative view. Images were saved in JPEG format, with a color depth of 24-bit and a resolution of 1600×1200 pixels at 96 dots per inch (both horizontal and vertical). No color correction or contrast adjustments were applied. All histopathological specimens originated from the same pathology department and were processed using identical tissue fixation, embedding, sectioning, and staining protocols to ensure methodological consistency across cases.

Evaluation by ChatGPT and Pathologists

The images were presented to ChatGPT-4o mini without any accompanying clinical information, and the model was prompted with the question "What is this?":

Responses of both ChatGPT and pathologists were categorized into five distinct categories:

- **Correct:** Accurate diagnosis or a partially correct response that aligned with the correct interpretation [true negative (TN) and true positive (TP) results].
- **Low-impact error:** Errors that did not significantly impact the patient's management, prognosis, or treatment (labeling a chronic inflammatory lesion as "reactive changes" instead of providing the exact clinicopathological entity, or misclassifying a benign neoplasm subtype without altering its benign nature).
- **False negative:** Failure to identify a lesion or classify it as benign when it was malignant.
- **False positive:** Incorrect classification of benign lesions as malignant.
- **No interpretation:** Inability to provide any meaningful diagnosis or classification.

The reference standard (ground truth) for each case was defined according to the original finalized histopathological diagnosis issued in routine diagnostic practice at the originating institution. Although participants evaluated only a single static microphotograph per case, the reference diagnosis had been established through comprehensive assessment of the entire specimen, including full slide review and clinical correlation when applicable.

As a control, the same microphotographs were independently evaluated by 15 pathologists with experience ranging from four to 23 years (mean: 11.25 years), also without clinical context. The inclusion of 15 pathologists was intended to reflect inter-individual diagnostic variability in routine practice rather than to optimize interobserver agreement metrics. Pathologists were eligible for inclusion if they had a minimum of three years of independent diagnostic experience. All participating pathologists were general surgical pathologists and did not have formal subspecialty training. They were recruited from multiple institutions, both national and international, ensuring a heterogeneous expert pool. To avoid observer bias, pathologists affiliated with the laboratory from which the cases originated were excluded from participation, and none of the participants had prior exposure to the evaluated samples. The diagnostic accuracy and error categories for each pathologist were recorded and classified using the same criteria applied to ChatGPT.

Statistical Analysis

Specificity, sensitivity, negative predictive value (NPV), positive predictive value (PPV), and overall accuracy were the diagnostic metrics used to assess diagnostic performance. Only classifications labeled as TP, TN, false positive (FP), or false negative (FN) were included in the analysis. Cases with clinically insignificant errors or without any interpretation were excluded.

Diagnostic performance metrics for pathologists were calculated individually for each observer by comparing their classifications with the predefined reference standard. Subsequently, mean values and standard deviations were computed across the 15 observers to obtain group-level performance estimates.

For group comparison:

- Fisher's exact test was performed to assess overall differences in diagnostic accuracy between the model and the pathologists.
- McNemar's test was used to evaluate case-level discordance in paired classifications.
- The pathologists' performance was calculated as the average across 15 independent evaluations per case.

In addition, interobserver agreement among the pathologists was assessed using Fleiss' kappa statistic, which measures the degree of agreement beyond chance among multiple raters. Two analyses were conducted:

1. A full categorical model including correct diagnoses, incorrect diagnoses, and clinically insignificant errors.
2. A binary classification where both incorrect and minor errors were grouped as "incorrect".

A kappa value greater than 0.75 was considered to be excellent, a value between 0.40 and 0.75 was considered to be fair to good, and a value below 0.40 was considered to be poor agreement.

Statistical significance was determined by a p-value of less than 0.05. All analyses were conducted using Python (v3.11) and the Statsmodels library.

RESULTS

The correct response rate was 41.67% for ChatGPT-4o mini, compared with a significantly higher rate of 78.06% for pathologists. Among the pathologists, correct response rate ranged from 58.33% to 91.67%, indicating considerable variability in performance across evaluators. These calculations were based on case-specific diagnoses rather than a binary malignant-benign distinction (Table 1).

The error types made by ChatGPT-4o mini and pathologists were analyzed to assess their diagnostic tendencies. Low-impact diagnostic errors were observed in 33.3% of ChatGPT-4o mini's responses, in contrast to only 6.9% among the pathologists. Similarly, both FNs and FPs were observed in 8.3% of ChatGPT-4o mini's responses, while the corresponding rates for pathologists were 0.5% and 8.8%, respectively.

The frequency of no interpretation was comparable, with ChatGPT-4o mini providing no meaningful diagnosis in 8.3% of cases, compared to 5.5% for pathologists (Table 1).

Among the cases included in the study, lymphocytic thyroiditis (Figure 2e) proved to be the most challenging for the participating pathologists. Four pathologists (26%) did not provide an interpretation, nine pathologists (60%) classified the case as a malignant tumor, and only two pathologists (14%) reached the correct diagnosis. This difficulty can be attributed to the presence of squamous metaplasia in the image, a finding that is uncommonly observed in lymphocytic thyroiditis. These results indicate that, particularly when clinical information is limited, rare reactive histomorphological features may complicate interpretation and influence diagnostic outcomes.

Diagnostic Performance

The diagnostic capabilities of ChatGPT-4o mini and pathologists were evaluated across 24 cases, using standard performance metrics based on binary malignant-benign classification. Only valid classifications -TP, TN, FP, and FN- were included in the analysis; low-impact errors and uninterpretable cases were excluded.

ChatGPT-4o mini achieved a sensitivity of 60.0%, specificity of 77.8%, and overall accuracy of 71.4%. In contrast, the average performance of the 15 pathologists demonstrated a sensitivity of 97.7%, a specificity of 87.1%, and an accuracy of 89.8%. PPV and NPV followed a similar trend, with the pathologists outperforming ChatGPT across all metrics (Table 2).

These findings indicate a substantial performance gap, particularly in sensitivity and NPV, where pathologists showed perfect scores while ChatGPT-4o mini underperformed.

Statistical Comparison

Both case-agnostic and case-wise tests were conducted to assess if the differences between the two groups were statistically significant.

Fisher's exact test, comparing aggregated counts of correct and incorrect diagnoses, yielded a p-value of 0.191, suggesting no statistically significant difference when data were treated as independent observations. However, this independent-sample approach overlooks the paired nature of the diagnostic process, where the same cases were evaluated by both the model and the pathologists, thereby failing to account for case-specific correlations.

To account for this, McNemar's test was applied using case-level matched data, focusing on discordant classifications. This analysis revealed a difference between the two groups that is statistically significant ($p<0.001$), favoring the pathologists. Specifically, there were multiple instances where the model misclassified cases that the majority of pathologists diagnosed correctly. These findings underscore the importance of using paired statistical methods in diagnostic comparison studies and highlight the superior performance of the human experts in this setting. The results indicate that the model's misclassifications were not random but occurred systematically on specific cases that were correctly diagnosed by the pathologists.

Interobserver Agreement

The level of agreement among the 15 pathologists was evaluated using Fleiss' kappa statistic. When all three diagnostic outcomes were included - correct, incorrect, and low-impact diagnostic errors - the overall interobserver agreement was low, with a Fleiss' kappa of 0.13, indicating slight agreement. To assess whether diagnostic disagreement was driven by the presence of clinically insignificant errors, a binary classification was applied by grouping both incorrect and low-impact diagnostic errors as "incorrect". Even under this simplified dichotomy (correct vs. incorrect), Fleiss' kappa remained low at 0.14, suggesting that variability in pathologists' judgments persisted regardless of error type categorization. Importantly, diagnostic performance metrics were calculated by comparing each observer's classification against the predefined reference standard (original finalized diagnosis), rather than being derived from interobserver consensus.

Table 1: ChatGPT-4o mini's and pathologists' accuracy and error profiles.

Metric	ChatGPT-4o mini	Pathologists [mean ± standard deviation (range)]
Correct response rate (%)	41.67	78±10.9 (62.5-91.6)
False negative rate (%)	8.3	0.5±1.4 (0-4.1)
False positive rate (%)	8.3	8.8±6.2 (0-20.8)
Low-impact diagnostic errors (%)	33.3	6.9±4.9 (0-16.6)
No interpretation (%)	8.3	5.5±5.8 (0-16.6)

Table 2: Diagnostic performances of ChatGPT-4o mini and pathologists.

Metric	ChatGPT-4o mini	Pathologists [mean ± standard deviation (range)]
Sensitivity (%)	60.0	97.7±5.8 (83.3-100)
Specificity (%)	77.8	87.1±9.5 (68.7-100)
Positive predictive value (%)	60.0	75.1±14.7 (50-100)
Negative predictive value (%)	77.8	99.2±2.0 (94.1-100)
Accuracy (%)	71.4	89.8±6.9 (76.1-100)

DISCUSSION

Artificial intelligence has been rapidly transforming the field of pathology, particularly with the advent of digital pathology and whole slide imaging. AI-based systems have demonstrated their utility in detecting malignancies, identifying cellular atypia, and predicting prognostic factors. Studies have shown that convolutional neural networks can achieve high diagnostic accuracy in tasks such as identifying lymph node metastases in breast cancer and detecting prostate cancer in histopathological images (16, 17).

Moreover, AI models such as Google's DeepVariant and PathAI have exhibited remarkable accuracy in histological image analysis, surpassing or matching the performance of experienced pathologists in certain contexts (18, 19). However, these models are primarily trained on digital slide data, whereas LLMs such as ChatGPT-4o mini rely solely on textual input-output relationships. While ChatGPT lacks the capability to analyze microscopic images directly, its ability to synthesize information and provide differential diagnoses from textual data presents a unique opportunity for integration into pathology workflows.

Looking ahead, AI is expected to expand its role in pathology by enhancing real-time intraoperative consultations, automated grading of tumors, and even predicting therapeutic responses. However, for AI models like ChatGPT to reach their full potential in pathology, hybrid models that combine image analysis with natural language processing (NLP) will be necessary (20).

Our study compared the diagnostic accuracy of ChatGPT-4o mini with that of 15 experienced pathologists. The correct response rate demonstrated by ChatGPT-4o mini was 41.7%, which was significantly lower than the same rate of pathologists (78.06%). This performance gap is consistent with the results of previous studies that have evaluated AI models trained on non-image datasets, where the lack of image-based contextual understanding led to inferior diagnostic accuracy (21).

The error profile analysis revealed that low-impact diagnostic errors were observed in 33.3% of ChatGPT-4o mini's responses, whereas pathologists made similar errors at a rate of 6.9%. Additionally, ChatGPT's FN and FP rates were both 8.3%, compared to 0.5% and 8.8% for pathologists. These findings align with the studies highlighting that AI models often excel in detecting common patterns but may struggle with atypical or rare cases, leading to higher rates of misclassification (22).

Real-world practice always involves clinical, radiological, and demographic context. Adding such data would likely improve accuracy for both AI and humans. Therefore, future studies should aim to include multimodal data combining histopathological images with clinical context to more accurately reflect real-world diagnostic environments and to better assess the full potential of AI in pathology.

Our findings on ChatGPT-4o mini's diagnostic performance are largely consistent with recent studies conducted in different areas of pathology. In our series, the model achieved an accuracy of 41.7%, with a sensitivity of 60.0% and a specificity of 77.8%, whereas pathologists reached an average accuracy of 78.0%. Similarly, in the evaluation of oral mucosal lesions, ChatGPT's accuracy was reported as 52.5% compared to over 84% for experts; in glioma histopathology, accuracy was 43% for ChatGPT and 89% for neuropathologists; and in cervical cytology, ChatGPT achieved around 50% while pathologists exceeded 85% (11-13). Taken together, these results show that ChatGPT demonstrates similar limitations across organ systems and specimen types, particularly struggling with the recognition of complex morphological patterns, while human experts consistently achieve much higher levels of accuracy. At the same time, ChatGPT's most immediate value may not lie in primary diagnosis but rather in supportive roles such as simplifying pathology reports, education, and reducing workload (23). Our study reinforces these trends, confirming that ChatGPT cannot yet function as a standalone diagnostic tool but may, in the future, contribute to pathology practice through hybrid approaches.

The integration of AI into pathology workflows raises critical ethical considerations. The principle of "primum non nocere" (first, do no harm) is paramount in medical practice. To align with this principle, both ChatGPT-4o mini and human pathologists were given the option to refrain from making a diagnosis rather than providing a potentially misleading result. This aspect is particularly important in medical AI applications, where incorrect classifications may have serious implications for patient management and prognosis.

Moreover, the potential for AI-induced bias, over-reliance on technology, and the risk of automation bias require careful consideration. Regulatory frameworks and ethical guidelines must be continually updated to ensure that AI systems are deployed in a manner that minimizes harm and maximizes patient benefit (24, 25).

Our findings underscore the need for ongoing refinement of AI models in pathology, particularly through the integration of hybrid systems that combine histopathological image analysis with NLP capabilities. Such approaches could lead to the development of intelligent diagnostic assistants that not only flag suspicious findings but also facilitate diagnostic consensus among pathologists and contribute to reducing workload in high-volume settings.

To enhance the clinical utility of ChatGPT-like models, future efforts should focus on incorporating image interpretation algorithms capable of analyzing complex tissue patterns and morphological features. Equally important is the continued training of these models on large, diverse, and pathology-specific datasets, which would enable better handling of diagnostically challenging or rare cases. Moreover, implementing AI successfully in clinical practice will require rigorous validation in real-world settings to ensure safety, reliability, and generalizability across different institutions and diagnostic environments.

By addressing these areas, future iterations of AI systems may evolve into robust and trustworthy tools that complement the expertise of human pathologists, supporting but not replacing their critical role in the diagnostic process.

Study Limitations

While our study contributes valuable insights into the comparative diagnostic performance of ChatGPT-4o mini and pathologists, several limitations must be acknowledged. First, the number of images was deliberately limited to 24 to ensure feasibility for 15 busy pathologists, while still representing a balanced spectrum of normal tissues, benign lesions, and malignancies. In addition, the clinical application of AI in medicine remains an advancing field. Our results should be interpreted as exploratory rather than definitive. This study should be viewed as providing preliminary insights into the potential role of AI. We recognize that a larger and more diverse dataset would increase generalizability and have highlighted this point more explicitly. Second, ChatGPT-4o mini is not designed as an image-analysis tool. Our rationale for using it nonetheless lies in its accessibility, widespread popularity, and the fact that clinicians already experiment with its image-upload functionality. In this sense, the study does not claim to evaluate a dedicated histopathology AI system but rather assesses how a general-purpose, widely used tool performs under diagnostic conditions. Third, although interobserver agreement among the 15 pathologists was assessed using Fleiss' kappa (κ=0.13), this low level of agreement may reflect differences in diagnostic thresholds, interpretive experience, and uncertainty in borderline cases. This variability may have been further amplified by the absence of clinical and radiological information and by the use of static microphotographs rather than whole-slide images, conditions that are known to increase subjectivity in histopathological interpretation. The absence of clinical information likely affected both human and AI performance, particularly in borderline entities where morphology alone is insufficient for confident classification. In daily practice, even experienced pathologists rely on clinical and radiological context to resolve such ambiguity. Therefore, the observed error patterns should not be interpreted as isolated failures of either the LLM or the human observers, but rather as a consequence of intentionally decontextualized image-based assessment. It is important to note that the observed variability does not necessarily undermine the overall high accuracy observed at the group level.

Another unique aspect of our study is that neither ChatGPT-4o mini nor the pathologists were provided with any clinical information. Only static microscopic images were evaluated. Using 24 static microscopic images allowed us to standardize image presentation and better evaluate the results. However, this is a serious limitation. In routine clinical practice, pathologists assess entire slides, evaluating different criteria across different magnifications. The use of static images may limit the interpretation of diagnostic cues. This stands in contrast to routine diagnostic workflows, where clinical history, radiologic findings, and demographic information play a crucial role in guiding interpretation. Previous research has demonstrated that access to relevant clinical data significantly enhances diagnostic accuracy among pathologists, and similar improvements have been observed in AI systems when clinical context is incorporated. Olawade et al. (26) emphasized that the integration of multimodal information, including clinical data, is critical to improving the accuracy and reliability of AI systems in healthcare delivery, warning that the absence of such context can limit their diagnostic utility. Similarly, Obuchowicz et al. (27), in their review of AI applications in medical imaging, highlighted that combining clinical, radiological, and histopathological data substantially increases diagnostic accuracy compared to image-only analysis.

CONCLUSION

In conclusion, while ChatGPT's diagnostic accuracy was inferior compared to that of the pathologist's, its ability to identify partially correct responses and error profiles suggests its potential as a diagnostic assistant. The absence of clinical information in this study likely influenced diagnostic accuracy, highlighting the importance of integrating clinical context in future AI-assisted diagnostics. Future AI models that integrate image analysis and NLP may further enhance diagnostic accuracy and improve pathology workflows. Ethical considerations, including the principle of "primum non nocere" (first, do no harm), must guide the deployment of AI models in clinical settings to minimize harm and ensure patient safety.

Ethics

Ethics Committee Approval: *This study did not involve identifiable patient data. All histopathological images were de-identified archival cases and were used solely for research and educational purposes. According to institutional and international ethical guidelines, a formal approval from an ethics committee was therefore not required.*

Informed Consent: *All participating pathologists took part voluntarily and provided informed consent to contribute their diagnostic assessments.*

Footnotes

Conflict of Interest: *Preliminary results of this study were presented orally at the 3rd National Medical Student Symposium with the main theme "Foundations of Modern Medicine" held between November 30-December 1, 2024. The abstract was published in the symposium proceedings, and the presentation was awarded first prize. The authors declared no conflict of interest.*

Author Contributions: *Surgical and Medical Practices: J.M., Concept: J.M., Design: A.M., Data Collection or Processing: A.M., J.M., Analysis and/or Interpretation: A.M., J.M., Literature Search: A.M., J.M., Writing: A.M., J.M.*

Financial Disclosure: *The authors declared that this study received no financial support.*

REFERENCES

1. Ratner, M. FDA backs clinician-free AI imaging diagnostic tools. Nat Biotechnol. 2018;36:673-4. [Crossref]
2. Tan GC, Wong YP. ChatGPT in pathology. Malays J Pathol. 2024;46(2):231-2. [Crossref]
3. Apornvirat S, Thinpanja W, Damrongkiet K, Benjakul N, Laohawetwanit T. ChatGPT for histopathologic diagnosis. Ann Diagn Pathol. 2024;73:152365. [Crossref]
4. Koga S, Du W, Ono D. Performance and limitations of customized ChatGPT in histopathologic diagnosis. Ann Diagn Pathol. 2024;73:152362. [Crossref]
5. Apornvirat S, Thinpanja W, Damrongkiet K, Benjakul N, Laohawetwanit T. Comparing customized ChatGPT and pathology residents in histopathologic

description and diagnosis of common diseases. Ann Diagn Pathol. 2024;73:152359. [Crossref]
6. Sun SH, Huynh K, Cortes G, Hill R, Tran J, Yeh L et al. Testing the ability and limitations of ChatGPT to generate differential diagnoses from transcribed radiologic findings. Radiology. 2024;313(1):e232346. [Crossref]
7. Huang L, Hu J, Cai Q, Ye A, Chen Y, Yang Xiao-Zhi Z et al. Preliminary discrimination and evaluation of clinical application value of ChatGPT4o in bone tumors. J Bone Oncol. 2024;48:100632. [Crossref]
8. Omar M, Ullanat V, Loda M, Marchionni L, Umeton R. ChatGPT for digital pathology research. Lancet Digit Health. 2024;6(8):e595-600. [Crossref]
9. Dwivedi YK, Kshetri N, Hughes L, Hughes L, Slade EL, Jeyaraj A et al. Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manag. 2023;71:102642. [Crossref]
10. Apornvirat S, Namboonlue C, Laohawetwanit T. Comparative analysis of ChatGPT and Bard in answering pathology examination questions requiring image interpretation. Am J Clin Pathol. 2024;162(3):252-60. [Crossref]
11. Vaira LA, Lechien JR, Maniaci A, De Vito A, Mayo-Yáñez M, Troise S et al. Diagnostic performance of ChatGPT-4o in analyzing oral mucosal lesions: a comparative study with experts. Medicina (Kaunas). 2025;61(8):1379. [Crossref]
12. Mazzucchelli M, Salzano S, Caltabiano R, Magro G, Certo F, Barbagallo G et al. Diagnostic performance of ChatGPT-4.0 in histopathological analysis of gliomas: a single institution experience. Neuropathology. 2025;45(4):e70023. [Crossref]
13. Laohawetwanit T, Apornvirat S, Asaturova A, Li H, Lami K, Bychkov A. Evaluation of general-purpose large language models as diagnostic support tools in cervical cytology. Pathol Res Pract. 2025;274:156159. [Crossref]
14. Aliyeva A. "Bot or not": turing problem in otolaryngology. Cureus. 2023;15(11):e48170. [Crossref]
15. Aliyeva A, Sari E. Be or not to be with ChatGPT? Cureus. 2023;15(11):e48366. [Crossref]
16. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88. [Crossref]
17. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301-09. [Crossref]
18. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C et al. Automated gleason grading of prostate cancer tissue microarrays via deep learning. Sci Rep. 2018;8(1):12054. [Crossref]
19. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017;318(22):2199-210. [Crossref]
20. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J. 2018;16:34-42. [Crossref]
21. Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol. 2019;20(7):938-47. [Crossref]
22. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B et al. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 2020;21(2):233-41. [Crossref]
23. Bheemireddy S, Leslie SE, Durden JA, Burnet G, Aryanpour Z, Fong A et al. The Use of ChatGPT-4.0 to simplify breast pathology reports: a study on readability and accuracy. Ann Surg Oncol. 2025;32(11):8400-8. [Crossref]
24. Rigby MJ. Ethical dimensions of using artificial intelligence in health care. AMA J Ethics. 2019;21(2):121-4. [Crossref]
25. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M et al. The ethics of AI in health care:a mapping review. Soc Sci Med. 2020;260:113172. [Crossref]
26. Olawade DB, David-Olawade AC, Wada OZ, Asaolu AJ, Adereni T, Ling J et al. Artificial intelligence in healthcare delivery: prospects and pitfalls. J Med Surg Public Health. 2024;3:100108. [Crossref]
27. Obuchowicz R, Strzelecki M, Piórkowski A. Clinical applications of artificial intelligence in medical imaging and image processing-a review. Cancers (Basel). 2024;16(10):1870. [Crossref]