Barw Medical Journal | 2025;3(3):6-12
Journal home page: https://barw.krd/index.php/BMJ

Original Article

Exploring Large Language Models Integration in the Histopathologic Diagnosis of Skin Diseases: A Comparative Study

Talar Sabir Ahmed[1], Rawa M. Ali[2,3], Ari M. Abdullah[2,4], Hadeel A. Yasseen[2,5], Ronak S. Ahmed[2,6], Ameer M. Salih[2,7], Dilan S. Hiwa[2*], Shvan H. Mohammed[8]

1. *Hiwa Cancer Hospital, Shorsh Street, Sulaymaniyah, Iraq*
2. *Scientific Affairs Department, Smart Health Tower, Madam Mitterrand Street, Sulaymaniyah, Iraq*
3. *Hospital for Treatment of Victims of Chemical Weapons, Mawlawy Street, Halabja, Iraq*
4. *Department of Pathology, Sulaymaniyah Teaching Hospital, Sulaymaniyah, Iraq*
5. *College of Medicine, University of Sulaimani, Madam Mitterrand Street, Sulaymaniyah, Iraq*
6. *Shahid Nabaz Dermatology Teaching Center for Treating Skin Diseases, Sulaymaniyah Directorate of Health, Sulaymaniyah, Iraq*
7. *Civil Engineering Department, College of Engineering, University of Sulaimani, Sulaymaniyah, Iraq*
8. *Xzmat polyclinic, Rizgari, Kalar, Sulaymaniyah, Iraq*

[*] ***Corresponding author***: dilan.sarmad.hiwa@gmail.com *(D.S. Hiwa). Ashty Street 30 -Zone 1 -house number 6, Zip code: 46001, Sulaimani, Iraq*

Keywords:
ChatGPT
Gemini
Large language models
Dermatology
Histopathology

Received: April 05, 2025
Revised: April 19, 2025
Accepted: April 24, 2025
First Published: April 30, 2025

Citation: Ahmed TS, Ali RM, Abdullah AM, Yasseen HA, Ahmed RS, Salih AM et al. Exploring Large Language Models Integration in the Histopathologic Diagnosis of Skin Diseases: A Comparative Study. Barw Medical Journal.2025;3(3):6-12.
https://doi.org/10.58742/bmj.v3i3.180

Abstract

Introduction

The exact manner in which large language models (LLMs) will be integrated into pathology is not yet fully comprehended. This study examines the accuracy, benefits, biases, and limitations of LLMs in diagnosing dermatologic conditions within pathology.

Methods

A pathologist compiled 60 real histopathology case scenarios of skin conditions from a hospital database. Two other pathologists reviewed each patient's demographics, clinical details, histopathology findings, and original diagnosis. These cases were presented to ChatGPT-3.5, Gemini, and an external pathologist. Each response was classified as complete agreement, partial agreement, or no agreement with the original pathologist's diagnosis.

Results

ChatGPT-3.5 had 29 (48.4%) complete agreements, 14 (23.3%) partial agreements, and 17 (28.3%) none agreements. Gemini showed 20 (33%), 9 (15%), and 31 (52%) complete agreement, partial agreement, and no agreement responses, respectively. Additionally, the external pathologist had 36(60%), 17(28%), and 7(12%) complete agreements, partial agreements, and no agreements responses, respectively, in relation to the pathologists' diagnosis. Significant differences in diagnostic agreement were found between the LLMs and the pathologist ($P < 0.001$).

Conclusion

In certain instances, ChatGPT-3.5 and Gemini may provide an accurate diagnosis of skin pathologies when presented with relevant patient history and descriptions of histopathological reports. However, their overall performance is insufficient for reliable use in real-life clinical settings.

1. Introduction

The healthcare sector is undergoing significant transformation with the emergence of large language models (LLMs), which have the potential to revolutionize patient care and outcomes. In November 2022, OpenAI introduced a natural language model called Chat Generative Pre-Trained Transformer (ChatGPT). It is renowned for its ability to generate responses that approximate human interaction in various tasks. Gemini, developed by Google, is a text-based AI conversational tool that utilizes machine learning and natural language understanding to address complex inquiries. These models generate new data by identifying structures and patterns from existing data, demonstrating their versatility in producing content across different domains. Generative LLMs rely on sophisticated deep learning methodologies and neural network architectures to scrutinize, comprehend, and produce content that closely resembles human-created outputs. Both ChatGPT and Gemini have gained global recognition for their unprecedented ability to emulate human conversation and cognitive abilities [1-3].

ChatGPT offers a notable advantage in medical decision-making due to its proficiency in analyzing complex medical data. It is a valuable resource for healthcare professionals, providing quick insights derived from patient records, medical research, and clinical guidelines [1,4]. Moreover, ChatGPT can play a crucial role in the differential diagnostic process by synthesizing information from symptoms, medical history, and risk factors, and comprehensively processing this data to present a range of potential medical diagnoses, thereby assisting medical practitioners in their assessments. This has the potential to improve diagnostic accuracy and reduce instances of misdiagnosis or delays [4].

The integration of ChatGPT and Gemini into the medical decision-making landscape has generated interest from various medical specialties. Multiple disciplines have published articles highlighting the significance and potential applications of ChatGPT and Gemini in their respective fields [2,5]. Despite the growing number of these models used in diagnostics, patient management, preventive medicine, and genomic analysis across medicine, the integration of LLMs in dermatology remains limited. This study emphasizes the exploration of large language models, highlighting their less common yet promising role in advancing dermatologic diagnostics and patient care [6]

This study aims to explore the role of LLMs and its decision-making capabilities in the field of pathology, specifically in dermatologic conditions. It focuses on ChatGPT 3.5 and Gemini and compares their accuracy and concordance with the diagnoses of human pathologists. The study also investigates the potential advantages, biases, and constraints of integrating LLM tools into pathology decision-making processes.

2. Methods

2.1. Case Selection

A pathologist selected 60 real case scenarios, with half being neoplastic conditions and the other half non-neoplastic, from a hospital's medical database. The cases involved patients who had undergone biopsy and histopathological examination for skin conditions. The records included information on age, sex, and the chief complaint of the patients, in addition to a detailed description of the histopathology reports (clinical and microscopic description without the diagnosis).

2.2. Consensus Diagnosis

Two additional board-certified pathologists reviewed each case, reaching a collaborative consensus diagnosis through a meticulous review of clinical and microscopic descriptions. This process ensured diagnostic accuracy and reliability while minimizing individual biases.

2.3. Eligibility Criteria

The study included cases that had complete and relevant histopathological reports and comprehensive patient demographic information. Specifically, cases were included if they provided a definitive diagnosis in the histopathological report and contained detailed patient data such as age, gender, and clinical history. Cases were excluded if the histopathological report was incomplete, lacked critical patient information, or if the diagnosis could not be definitively made based solely on the textual description.

2.4. Sampling Method

The selection process involved a systematic review of available cases from the hospital's medical database to ensure a representative sample of different dermatologic diagnoses. A random sampling method was employed to minimize selection bias and to ensure the sample was representative of the broader population of dermatologic conditions within the database. The selected cases span a range of common and less common dermatologic conditions, enhancing the generalizability of the study's findings.

2.5. Evaluation by LLMs and External Pathologist

In March 2023, these cases were evaluated using two LLM systems, namely ChatGPT-3.5 and Gemini. In addition, an external board-certified pathologist was tested similarly to the AI systems, receiving only the necessary histopathology report descriptions (without histopathological images) to ensure a fair comparison between the LLM systems and the external pathologist.

2.5. Pathologists' Experience

The Pathologists involved in the study had a minimum of eight years of experience in their respective specialties, handling an average of 30 cases per month. This level of experience ensured a deep familiarity with a wide range of case scenarios. Crucially, the pathologists conducted their assessments were fully informed of the study design, including the comparative analysis with AI systems. Their expertise and understanding were vital in upholding the integrity and reliability of the diagnostic evaluations throughout the study

2.6. LLMs Prompting Strategy

The LLM systems were initially greeted with a prompt saying "Hello," followed by standardized inquiries presented as: "Please provide the most accurate diagnoses from the texts that will be given below." Each case was individually presented by copy-pasting it from a Word document and requesting each system to provide a diagnosis of the case scenario based on the information presented. The first response of each system to the inquiry was documented. If no diagnosis was given, the prompt was repeated as such: "Please, based on the histopathological report information given above, provide the most likely disease that causes it." Until a diagnosis was obtained. In some cases, after a diagnosis was provided, an additional question was asked to specify the histologic subtype of the condition (e.g., if the diagnosis was "seborrheic keratosis", the system was asked to specify the histologic subtype). Furthermore, the board-certified external pathologist was tested with the same questions, and the correct diagnosis was inquired.

2.7. Response Categorization

The responses from both systems and the external pathologist were categorized into three subtypes: complete agreement with the original diagnosis by the human pathologists, partial agreement, or none agreement. The criteria for categorizing agreement levels into "complete," "partial," and "none agreement" are based on the distinction between general and specific diagnostic classifications. For instance, when the original diagnosis provides a detailed type and subtype (e.g., "Seborrheic keratosis, irritated type"), a LLM's or external pathologist's response was classified as demonstrating "complete agreement" if it accurately identifies both the general diagnosis ("Seborrheic keratosis") and the specific subtype ("irritated type"). This classification acknowledges that accurate identification of both components reflects a thorough understanding and alignment with the original diagnosis. Conversely, an assessment was categorized as "partial agreement" if the response correctly identifies the general diagnosis but inaccurately specifies the subtype. Furthermore, a diagnosis was classified as demonstrating "no agreement" when both the general diagnosis and subtype provided by the AI tool or external pathologist are incorrect. These classification criteria draw upon established methodologies in diagnostic agreement studies, emphasizing the importance of distinguishing between different levels of agreement based on the precision and correctness of diagnostic outputs [7].

2.8. Data Processing and Statistical Analysis

The initial processing of the acquired data involved several steps before statistical analysis. First, the data were inputted into Microsoft Excel 2019. Subsequently, they were transferred to Statistical Package for the Social Sciences software (SPSS) 27.0 and the DATA tab for further analysis. Fleiss kappa was utilized to measure agreement among ChatGPT, the external pathologist, and Gemini. Additionally, Chi-square tests were applied to investigate associations between the two LLMs and the external pathologist. In this study, significance was defined as a p-value of < 0.05. A literature review was performed for the study, selectively considering papers from reputable journals while excluding those from predatory sources based on established criteria [8].

3. Results

ChatGPT-3.5 provided 29 (48.4%) complete agreement, 14 (23.3%) partial agreement, and 17 (28.3%) none agreement responses for the scenarios presented. In contrast, Gemini offered 20 (33%), 9(15%), and 31 (52%) complete agreement, partial agreement, and none agreement responses, respectively, for the same scenarios. Moreover, the external pathologist provided 36 (60%) complete agreement, 17 (28%) partial agreement, and 7 (12%) none agreement responses (Table 1). The complete details of the scenarios, including the diagnosis from the pathologists, ChatGPT's, Gemini's, and the external pathologist diagnoses are available in (Supplement 1).

Table 1. Distribution of the dermatopathology questions that the AI systems underwent.

Variables	Frequency/ percentage
Pathological classification	
Neoplastic	30 (50%)
Non-neoplastic	30 (50%)
Neoplastic	
Benign	19 (31.7%)
Malignant	11 (18.3%)
Non-neoplastic	
Dermatosis	9 (15%)
Infectious, pilosebaceous	2 (3.3%)
Connective tissue disease	2 (3.3%)
Infectious	2 (3.3%)
Granulomatous	2 (3.3%)
Vascular	2 (3.3%)
Epidermal maturation/keratinization disorder	2 (3.3%)
Dermatosis, pilosebaceous	2 (3.3%)
Pilosebaceous	2 (3.3%)
Panniculitis,	1 (3.3%)
Dermatosis, infectious	1 (1.7%)
Dermatosis, pigmentation disorder	1 (1.7%)
Granulomatous, panniculitis	1 (1.7%)
Bullous	1 (1.7%)
External Pathologist	
Complete agreement	36 (60%)
Partial agreement	17 (28%)
None agreement	7 (12%)
ChatGPT	
Complete agreement	29 (48.4%)
Partial agreement	14 (23.3%)
None agreement	17 (28.3%)
Gemini	
Complete agreement	20 (33%)
Partial agreement	9 (15%)
None agreement	31 (52%)

The agreement between ChatGPT, the external pathologist, and Gemini was assessed using Fleiss' kappa, which indicated a statistical significance at a level of <0.001, demonstrating slight to moderate agreement with respect to the original diagnosis made by the pathologists. Out of the 29 questions where ChatGPT agreed with the original diagnosis, only 12 (41.4%) instances also received complete agreement from both Gemini and the external pathologist (Table 2).

When assessing the agreement between ChatGPT, the external pathologist, and Gemini, using the external pathologist as the reference, the external pathologist showed complete agreement

Table 2. Measurement of agreement between ChatGPT, external pathologist, and Gemini

			ChatGPT		Measurement of Agreement (Fleiss)	Significance level
		Complete agreement	Partial agreement	None agreement		
Gemini	Complete agreement	12 (41.4%)	1(7.1%)	2 (11.8%)	0.25	<0.001
	Partial agreement	3 (10.4%)	1(7.1%)	0 (0.0%)		
	None agreement	1(3.4%)	0(0.0%)	0 (0.0%)		
External Pathologist	Complete agreement	2 (7%)	3(21.4%)	0 (0.0%)		
	Partial agreement	1(3.4%)	3(21.4%)	0 (0.0%)		
	Complete agreement	5(17.2%)	2(14.4%)	9 (53%)		
	Partial agreement	4(13.8%)	4(28.6%)	1(5.8%)		
	None agreement	1(3.4%)	0(0.0%)	5 (29.4%)		
Total		29	14	17		

with the original diagnosis in 36 cases. Among these, ChatGPT achieved complete agreement in 19 cases (52.7%), while Gemini achieved complete agreement in 15 cases (41.7%). Additionally, the external pathologist showed none agreement with the original diagnosis in only 7 cases. Among these, ChatGPT achieved none agreement in 5 cases (71.4%), while Gemini achieved none agreement in 6 cases (85.7%). Statistical analysis indicated significant differences in agreement levels between AI tools (ChatGPT and Gemini) and the external pathologist, with a P-value of <0.001 (Table 3).

In addition, the agreement between the external pathologist, ChatGPT, and Gemini was assessed for both neoplastic and non-neoplastic cases. Statistical analysis revealed significant differences in the agreement levels between the LLMs and the external pathologist, with a P-value of <0.001, highlighting the statistically significant disparity in agreement rates between the LLMs and the external pathologist (Table 4 and 5).

4. Discussion

Despite being in existence for over five decades, LLM has recently garnered substantial attention in the public sphere. The increased focus on LLMs in the medical field has led to speculation about the potential replacement of doctors by these systems. However, LLMs are more likely to serve as a complementary tool, aiding clinicians in efficiently processing data and making clinical decisions. This is substantiated by the fact that LLMs can "learn" from extensive collections of medical data. Modern systems are also noted for their self-correcting capabilities. As electronic medical records become more prevalent, there is a growing reservoir of stored patient data. While having access to more data is undoubtedly advantageous, scanning through patient charts can be challenging. Algorithms have been developed to sift through patient notes and detect individuals with specific risk factors, diagnoses, or outcomes. This capability is particularly valuable because, in theory, a LLM system could be developed to review and extract data from medical charts, including pathology reports, and promptly identify patients at highest risk for conditions that could cause significant morbidity or mortality if missed by the physician [6,9].

The field of pathology is no exception to the adaptation of LLMs and the utilization of these technological advancements. Various in recent years have assessed LLM's accuracy, potential use, and associated limitations. For instance, a study by Vaidyanathaiyer et al., evaluated ChatGPT's proficiency in pathology through thirty clinical case scenarios. These cases were evenly distributed across three primary subcategories: hematology, histopathology, and clinical pathology, with ten cases from each category. The researchers reported that ChatGPT received high grade of "A" on nearly three-quarters of the questions; in the remaining questions, and "B" grades on remaining questions. They found that ChatGPT demonstrated moderate proficiency in these subcategories, excelling in rapid data analysis and providing fundamental insights, though it had limitations in generating thorough and elaborate information [10]. Furthermore, Passby et al. demonstrated capacity of ChatGPT to address multiple-choice inquiries in the Specialty Certificate Examination of dermatology, with ChatGPT-4 outperforming ChatGPT-3.5, scoring 90% versus 63%, respectively, compared to an approximate passing score of 70% [11]. In an investigation by Delsoz et al., twenty corneal pathologies with their respective case descriptions were provided to ChatGPT-3.5 and ChatGPT-4. ChatGPT-4 performed better, correctly answering 85% of the questions, whereas ChatGPT-3.5 answered only 60% correctly [12]. The current study found that ChatGPT-3.5 performed similarly in the percentage of correct responses. However, this study further evaluated the LLM responses and found that nearly

Table 3. Comparative analysis of ChatGPT, Gemini, and the external pathologist responses regarding all skin pathologies.

AI tools		External pathologist			P-value
		Complete agreement	Partial agreement	None agreement	
ChatGPT	Complete agreement	19(52.7%)	8(47.1%)	2(28.6%)	<0.001
	Partial agreement	6(16.7%)	8(47.1%)	0(0%)	
	None agreement	11(30.6%)	1(5.8%)	5(71.4%)	
Gemini	Complete agreement	15(41.7%)	4(23.5%)	1(14.3%)	<0.001
	Partial agreement	5(13.9%)	4(23.5%)	0(0%)	
	None agreement	16(44.4%)	9(53%)	6(85.7%)	
Total		36(100%)	17(100%)	7(100%)	

23.3% and 15% of ChatGPT and Gemini answers, respectively, were fair but still had inaccuracies. This highlight areas where these systems can improve, as they sometimes almost answer correctly but not fully. For instance, when a histopathology report of squamous cell carcinoma in situ was given to ChatGPT-3.5, it answered with squamous cell carcinoma. On further prompting, the system favored an invasive squamous cell carcinoma over an in-situ one, even when the suggestion was made to it whether an in-situ lesion was more appropriate for that scenario. similarly, in the case of guttate psoriasis, Gemini answered with only "psoriasis" did not specify the type, while ChatGPT-3.5 responded with "psoriasis vulgaris". In a study by Rahsepar et al. on pulmonary malignancies, Google Bard (the former name of Gemini) provided 9.2% partially correct answers, similar to Gemini's 15% partially correct responses in this study. However, ChatGPT-3.5 answered 17.5% of lung cancer questions incorrectly, whereas in the present study, ChatGPT-3.5's incorrect answers were nearly twice as frequent. This may be due to ChatGPT broader access to data and medical information on lung cancer compared to the dermatological conditions tested in this study, highlighting the limitations and risks of relying on these systems for rarer diseases [13]. Although existing language models have access to extensive medical data, they often lack a nuanced understanding of individual diseases or specific patient cases. They have not undergone specialized training for medical tasks, relying solely on the provided data and information. The unclear methodology behind the LLM's diagnostic process leads to skepticism regarding the reliability of LLM-generated diagnoses. Consequently, their ability to accurately diagnose complex or unique cases may be limited, as demonstrated in the current study on skin histopathology cases. Notably, in a few cases,

Table 4. The agreement status between external pathologist, ChatGPT, and Gemini regarding non- neoplastic cases.

AI tools		External pathologist			P-value
		Complete agreement	Partial agreement	None agreement	
ChatGPT	Complete agreement	11(61.1%)	2(40%)	4(57.1%)	<0.001
	Partial agreement	3(16.7%)	3(60%)	1(14.3%)	
	None agreement	4(22.2%)	0(0%)	2(28.6%)	
Gemini	Complete agreement	9(50%)	1(20%)	1(14.3%)	<0.001
	Partial agreement	7(38.9%)	4(80%)	4(57.1%)	
	None agreement	2(11.1%)	0(0%)	2(28.6%)	
Total non-neoplastic cases		18(100%)	5(100%)	7(100%)	

Table 5. The agreement status between external pathologist, ChatGPT, and Gemini regarding neoplastic cases.

AI tools		External pathologist			P-value
		Complete agreement	Partial agreement	None agreement	
ChatGPT	Complete agreement	8(44.4%)	0(40%)	4(40%)	<0.001
	Partial agreement	8(44.4%)	2(1000%)	0(0%)	
	None agreement	2(11.1%)	0(0%)	6(60%)	
Gemini	Complete agreement	6(33.3%)	0(20%)	3(30%)	<0.001
	Partial agreement	9(50%)	2(100%)	5(50%)	
	None agreement	3(16.7%)	0(0%)	2(20%)	
Total neoplastic cases		18(100%)	2(100%)	10(100%)	

LLMs declined to provide a diagnosis on the initial prompt, citing concerns about giving medical advice, and only issued a diagnosis after repeated prompting with the same scenario. Despite their ability to offer insights based on existing knowledge, LLMs may lack a complete understanding of the intricate details and visual indicators crucial for pathologists' diagnosis. In the current study, the pathologist initially examined the histopathology slides and then provided the report to the AI systems. Another issue is that preserving the integrity of LLMs and safeguarding the confidentiality of associated data from unauthorized access is critical, particularly in scenarios involving sensitive patient information [14,15]. The case scenarios in this study did not include specific patient identifiers. Additionally, failure to evolve the LLM tools utilized in the pathological assessment alongside advancements in clinical practice and treatment poses the risk of stagnation and adherence to outdated methodologies. Although it is possible to manually update LLM algorithms to align with new protocols, their efficacy depends heavily on the availability of pertinent data, which might not be readily accessible during transitional periods. Such adaptations could introduce errors, particularly in pathology, through misclassifications of entities as classification and staging systems undergo revisions. Another concern is automation bias, which refers to the tendency of clinicians to regard LLM-based predictions as flawless or to adhere to them without questioning their validity. This bias often emerges soon after exposure to new technology and may stem from concerns about the legal consequences of disregarding an algorithm's output. Research across various fields has shown that automation bias can reduce clinician accuracy, affecting areas such as electrocardiogram interpretation and dermatologic diagnoses. Clinicians at all proficiency levels, including experts, are susceptible to this phenomenon [3,14-16].

The LLM has numerous applications in the medical field, with various technologies being developed at an unprecedented pace. For example, in the field of epilepsy, Empatica has created a wearable monitor called Embrace, which detects the onset of seizures in patients with epilepsy and notifies designated family members or trusted physicians. This innovation enhances safety and facilitates early management of such cases and received FDA approval six years ago [17]. Additionally, one of the earliest uses of LLM was for the detection of atrial fibrillation. AliveCor mobile application, which facilitates ECG monitoring and atrial fibrillation detection using a mobile phone, was FDA-approved. Recent findings from the REHEARSE-AF study indicated that traditional care methods are less effective at detecting atrial fibrillation in ambulatory individuals compared to remote ECG monitoring using Kardia [17,18]. Another example is the artificial immune recognition system, which has demonstrated remarkable accuracy in diagnosing tuberculosis by using support vector machine classifiers. These advanced systems significantly outperform traditional methods, making them a robust tool in identifying tuberculosis cases with high reliability. This underscores the potential of these models to enhance diagnostic processes in infectious diseases [19]. The advancements across various medical disciplines render the application of LLMs in histopathological diagnostics increasingly viable and anticipated for future clinical implementation. This progress motivates further research by scientists and numerous companies, as the focus has shifted from questioning whether LLM will be used in pathology or not to when and how these models will be utilized precisely.

One limitation of this study is that the aforementioned LLM systems were not evaluated for their ability and accuracy in directly reaching a diagnosis from histopathological images. Instead, the study relied on providing necessary information from the histopathological reports in text form, which imposes practical constraints and still requires an expert pathologist. Future studies focusing on both histopathological images and texts are necessary to further evaluate the comprehensive capabilities of LLM tools in this domain.

5. Conclusion

In certain instances, ChatGPT-3.5 and Gemini may provide an accurate diagnosis of skin conditions when provided with pertinent patient history and descriptions of histopathological reports. Specifically, Gemini showed higher accuracy in diagnosing non-neoplastic cases, while ChatGPT-3.5 demonstrated better performance in neoplastic cases. However, despite these strengths, the overall performance of both models is insufficient for reliable use in real-life clinical settings.

Declarations

Conflicts of interest: The authors have no conflicts of interest to disclose.

Ethical approval: Not applicable.

Patient consent (participation and publication)**:** Not applicable.

Funding: The present study received no financial support.

Acknowledgements: None to be declared.

Authors' contributions: RMA and AMA were significant contributors to the conception of the study and the literature search for related studies. DSH and SHM involved in the literature review, study design, and manuscript writing. TSA, HAY, RSA, and AMS were involved in the literature review, the study's design, the critical revision of the manuscript, and data collection. RMA and DSH confirm the authenticity of all the raw data. All authors approved the final version of the manuscript.

Use of AI: ChatGPT-3.5 was used to assist in language editing and improving the clarity of the introduction section. All content was reviewed and verified by the authors. Authors are fully responsible for the entire content of their manuscript.

Data availability statement: Not applicable.

References

1. Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019;8(7):2328-31. doi:10.4103/jfmpc.jfmpc_440_19
2. Aydın Ö. Google Bard Generated Literature Review: Metaverse. Journal of AI. 2023;7(1): 1-14. *doi:N/A*
3. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023. doi:10.1016/j.iotcps.2023.04.003
4. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Medical Education. 2023;9:e45312. doi:10.2196/preprints.50336
5. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, et al. Evaluating Artificial Intelligence Responses to Public Health Questions. JAMA Network Open. 2023;6(6):e2317517. doi:10.1001/jamanetworkopen.2023.17517
6. Rundle CW, Hollingsworth P, Dellavalle RP. Artificial intelligence in dermatology. Clinics in Dermatology. 2021;39(4):657-66. doi:10.1016/j.clindermatol.2021.03.011
7. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276-82. doi:10.11613/BM.2012.031
8. Kakamad FH, Abdalla BA, Abdullah HO, Omar SS, Mohammed SH, Ahmed SM et al. Lists of predatory journals and publishers: a review for future refinement. European Science Editing. 2024;50:e118119. doi:10.3897/ese.2024.e118119
9. MahmoodYM, MohammedRO, HabibullahIJ, RahimHM, SalihAM. Comparing ChatGPT and Google Bard: Assessing AI-Powered Information Retrieval in Nursing. Barw Medical Journal.2024;2(1):12-20. doi:10.58742/hsn32c73
10. Vaidyanathaiyer R, Thanigaimani GD, Arumugam P, Einstien D, Ganesan S, Surapaneni KM. Navigating the path to precision: ChatGPT as a tool in pathology. Pathology-Research and Practice. 2024;254:155141. doi:10.1016/j.prp.2024.155141
11. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. Clinical and experimental dermatology. 2023:llad197. doi:10.1093/ced/llad197
12. Delsoz M, Madadi Y, Raja H, Munir WM, Tamm B, Mehravaran S, Soleimani M, Djalilian A, Yousefi S. Performance of ChatGPT in diagnosis of corneal eye diseases. Cornea. 2022; 13:10-97. doi:10.1097/ico.0000000000003492
13. Rahsepar AA, Tavakoli N, Kim GH, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT versus Google Bard. Radiology. 2023;307(5):e230922. doi:10.1148/radiol.230922
14. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. Diagnostic Pathology. 2024;19(1):1-9. doi:10.1186/s13000-024-01464-7
15. Nakagawa K, Moukheiber L, Celi LA, Patel M, Mahmood F, Gondim D, Hogarth M, Levenson R. AI in pathology: what could possibly go wrong?. InSeminars in Diagnostic Pathology 2023 (Vol. 40, No. 2, pp. 100-108). WB Saunders. doi:10.1053/j.semdp.2023.02.006
16. Evans H, Snead D. Why do errors arise in artificial intelligence diagnostic tools in histopathology and how can we minimize them?. Histopathology. 2024;84(2):279-87. doi:10.1111/his.15071
17. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Frontiers in medicine. 2020;7:509744. doi:10.3389/fmed.2020.00027
18. Halcox JP, Wareham K, Cardew A, Gilmore M, Barry JP, Phillips C, et al. Assessment of remote heart rhythm sampling using the AliveCor heart monitor to screen for atrial fibrillation: the REHEARSE-AF study. Circulation. 2017;136(19):1784-94. doi:10.1161/circulationaha.117.030583
19. Agrebi S, Larbi A. Use of artificial intelligence in infectious diseases. InArtificial intelligence in precision health 2020 (pp. 415-438). Academic Press. doi:10.1016/B978-0-12-817133-2.00018-5