Vita Medica Scientica Journal

VMS J Vol.2, No.2, 80- 85. (2026)
doi: 10.64514/vmsjournal.2026.25
ORIGINAL RESEARCH

Comparative Evaluation of Widely Available Tools (ChatGPT, Grok, and Gemini) for Skin Lesion Diagnosis in Primary Care

Aslıhan Çelik Çoban[1], Yıldız Büyükdereli Atadağ[2], Hatice Tuba Akbayram[3], Tuba Kalınkara Seyhan[4]

[1] Gaziantep Şehitkamil District Health Directorate, Gaziantep, Türkiye (0009-0009-8808-5224)
[2] Gaziantep University Faculty of Medicine, Department of Family Medicine, Gaziantep, Türkiye (0000-0002-8516-6477)
[3] Gaziantep University Faculty of Medicine, Department of Family Medicine, Gaziantep, Türkiye (0000-0002-9777-9596)
[4] Gaziantep University Faculty of Medicine, Department of Family Medicine, Gaziantep, Türkiye (0009-0000-2612-1810)

Accepted 29 May 2026.

Abstract

Objective and Aim

The study aims to evaluate the accuracy and reliability levels of artificial intelligence (AI) models in diagnosing dermatological lesions and determining the prognosis of the disease using clinical case images.

Materials and Methods

The images of 34 dermatological cases in the book “Approach to Dermatological Diseases in Family Medicine” published by Türkiye Klinikleri Press were used with permission. In March 2025, the images were uploaded to ChatGPT-4, Grok-3, and Gemini-2.0 models. Their response and accuracy rates were analyzed comparatively.

Results

The images of 34 cases were presented to the AI models. ChatGPT responded to all cases, Grok to 28, and Gemini to 22. The accuracy rates were 41.2% for ChatGPT, 38.2% for Grok, and 29.4% for Gemini. When evaluating the 21 cases in common, the accuracy rates were found to be 43% for ChatGPT, 52.4% for Grok, and 47.6% for Gemini. There was no statistically significant difference between the correct answer scores of the AI models (p=0.829). The correct diagnosis was consistent across all models in eight cases (verruca vulgaris (n=2), melanoma (n=2), urticaria, onychomycosis, acne vulgaris, and vitiligo). All nine cases (molluscum, HSV, lichen planus, herpangina, orf, monkeypox, Behçet's disease, pityriasis alba, felon) were misdiagnosed. In cases with a correct diagnosis, the distinction between malignancy and benignity was clear.

Corresponding Author:
Aslıhan Çelik Çoban
Gaziantep Şehitkamil District Health Directorate, Gaziantep, Türkiye
Email: aslihancelikcoban@gmail.com

To cite:
Çoban AÇ, Atadağ YB, Akbayram HT, Seyhan TK. Comparative Evaluation of Widely Available Tools (ChatGPT, Grok, and Gemini) for Skin Lesion Diagnosis in Primary Care. VMS Journal 2026;2(2):80-85.
doi: 10.64514/vmsjournal.2026.25

Conclusion

This study demonstrated the limitations of using ChatGPT-4, Grok-3, and Gemini-2.0 models as independent diagnostic tools in dermatological cases. However, it was observed that they could potentially be beneficial as decision support tools for specific lesions. AI models can be developed to increase their future applications.

Keywords: Artificial Intelligence, Skin Lesions, Primary Care, Large Language Models.

1. Introduction

Artificial intelligence (AI) refers to computer-based systems that replicate the cognitive functions of the human mind (such as learning, reasoning, perception, problem-solving, and decision-making) through software and algorithms.[1] Today, with the advancement of technology, AI is increasingly being used in large language models (LLMs) in the field of medicine.[2] AI algorithms have developed image-based models using imaging methods (e.g., ECG, CT, and eye fundus imaging) for diagnosing and treating diseases. [1]

Accurate diagnosis of skin diseases depends on physicians' ability to identify lesions, establish a comprehensive differential diagnosis, and refine it with additional clinical inquiries or tests.[3] Physicians must also have extensive knowledge, clinical experience, and contextual analysis skills. The insufficient number of dermatologists and the increasing workload make accurately evaluating dermatological images difficult. This situation may lead to an increased likelihood of misdiagnosis.[4] Image-based artificial intelligence models, particularly in dermatology, can enable fast and highly accurate diagnosis by supporting the physician's assessment in the recognition and analysis of medical images.[5] A study found that the rate of artificial intelligence use among family physicians was 56.9%.[6] When used correctly, artificial intelligence can enhance the capabilities of family physicians.[7]

LLMs are pioneering innovations and transformations in many fields, including medicine, alongside advancing AI technologies. ChatGPT (OpenAI), Grok (xAI), and Gemini (Google DeepMind) are examples of LLMs with different infrastructures that are increasingly being used in healthcare for accessing information and generating recommendations.[8,9] Although the use of LLM is increasing, its accuracy and reliability in clinical decision support systems has not been fully proven.[10]

The aim of this study is to examine how well the ChatGPT-4, Grok-3, and Gemini-2.0 models can diagnose and determine the prognosis of diseases using images of dermatological lesions.

2. Materials and Method

This comparative and descriptive study was designed to evaluate the accuracy and reliability levels of various AI models in diagnosing and prognosing dermatological lesions. The study was carried out from March 6 to 14, 2025, at the Department of Family Medicine, Faculty of Medicine, Gaziantep University. The AI models evaluated in this study were ChatGPT-4 (OpenAI), Grok-3 (xAI), and Gemini 2.0 (Google DeepMind), all of which are large language models (LLMs) and were publicly available at the time of data collection (March 2025).

The images of dermatological lesions used in this study were obtained from the book Approach to Dermatological Diseases in Family Medicine, published by Türkiye Klinikleri Press.[11] Permission to use the images for research purposes was obtained from the publisher in February 2025, and permission for their publication in the manuscript was granted in October 2025. Images of 34 dermatological conditions commonly encountered in primary care were selected.

A message written in Turkish was sent to the ChatGPT-4, Grok-3, and Gemini-2.0 models,

along with the uploaded images: "I will now upload a skin lesion image. I want you to answer the following questions based solely on this image. Provide your responses in a short and clear manner, following the specified format. Do not share the data with third parties. 1.What is the most likely diagnosis for this skin lesion? Provide your answer with a single diagnosis. 2. Is this lesion malignant or benign? Answer only with 'malignant' or 'benign'." The diagnostic and prognostic responses obtained from AI models, along with their accuracy rates and instances of non-responsiveness, were recorded and comparatively analyzed.

This study did not involve human participants or animals. All data were obtained from publicly available artificial intelligence tools, and the dermatological images were used with permission from the publisher. Therefore, ethical committee approval was not required. All procedures were conducted in accordance with the principles of the Declaration of Helsinki.

Statistical Analysis

Statistical analyses were performed using SPSS version 26.0 (IBM Corp., USA). Descriptive statistics were presented as frequency and percentage for categorical variables, and as mean and standard deviation for numerical variables. The Chi-square test was used to evaluate differences in total scores among the models, and a p-value of < 0.05 was defined as statistically significant.

3. Results

A total of 34 dermatological case images were uploaded to the models. ChatGPT responded to all cases, whereas Grok provided answers for 28 cases (leaving 6 unanswered) and Gemini for 22 cases (leaving 12 unanswered). When comparing diagnostic accuracy across the 34 cases, ChatGPT correctly answered 14 cases (41.2%), Grok 13 cases (38.2%), and Gemini 10 cases (29.4%). (Table 1)

In an evaluation based on 21 common cases answered by all models, ChatGPT was found to have a 43% accuracy rate, Grok 52.4%, and Gemini 47.6%. There was no statistically significant difference in the correct answer rates among the three AI models (p=0.829). (Table 2)

The models correctly identified eight cases, including verruca vulgaris (n=2), melanoma (n=2), urticaria, onychomycosis, acne vulgaris, and vitiligo. In contrast, all models failed to provide correct diagnoses in nine cases: molluscum contagiosum, HSV infection, lichen planus, herpangina, orf, monkeypox, Behçet's disease, pityriasis alba, and felon. In the cases where all models correctly identified the diagnosis, they also demonstrated complete accuracy in determining whether the lesion was malignant or benign.

Discussion

General Performance

In this study, the diagnostic performance of the large language models ChatGPT-4, Grok-3, and Gemini 2.0 on dermatological lesions was comparatively evaluated. Among the models, only ChatGPT-4 was able to respond to all cases. In a previous study by Sozer et al., artificial intelligence models (ChatGPT-4o, Grok, and Gemini) were evaluated using magnetic resonance (MR) images to assess their radiological diagnostic capabilities. ChatGPT refused to respond to 28.02% of the images, whereas Grok and Gemini answered all cases.[8] Overall, when the findings of both our study and that of Sozer et al. are considered, no consistent conclusion can be drawn regarding the comparative performance of the models. This may primarily be due to the limited number of studies that evaluate large language models (LLMs) using medical images. This variability may also reflect differences in model architecture, training data diversity, or even prompt interpretation across platforms.

Correct Answer Performance

The correct response rates of the AI models were 41.2% for ChatGPT-4, 38.2% for Grok-3, and 29.4% for Gemini 2.0. When evaluating the commonly answered case questions, the correct response rates were found to be 43% for ChatGPT-4, 52.4% for

Table 1. Comparison of Case Response Rates and Diagnostic Accuracy of Artificial Intelligence Models

Model	Cases Answered (Total of 34 Questions) (n, %)	Correct Answer (n, %)
ChatGPT-4	34 (100%)	14 (41.2%)
Grok-3	28 (82.35%)	13 (38.2%)
Gemini 2.0	22 (64.70%)	10 (29.4%)

Grok-3, and 47.6% for Gemini 2.0. The performance scores of the AI models on the commonly answered cases were similar.

In a study evaluating the image-based performance of LLMs (ChatGPT 3.5, 4.0, 4o, and Gemini) in diagnosing potentially malignant lesions in the mouth, ChatGPT-4o correctly answered 66.6% of cases, while Gemini correctly answered 45.2%.[12] Consistent with this, a recent meta-analysis examining the text-based and image-based performance of artificial intelligence models in dermatology board-style exams analyzed the results of a total of 14 articles. The analysis found that the average performance of all ChatGPT models in answering text-based questions correctly (65.24%) was significantly higher than their performance in answering image-based questions correctly (52.26%). When comparing the models, the ChatGPT-4 model was found to have an overall performance total score of 77.98% and a visual-based performance total score of 59.32%.[13] The lower diagnostic accuracy observed in our study may be due to its image-based design, in contrast to previous text-based studies. Interpreting medical images remains a challenge for current LLMs, which may explain the performance gap.

In a study assessing the dermatological knowledge and image analysis performance of various freely accessible LLMs (Claude-3.5 Sonnet, Copilot, Gemini, ChatGPT-4o, and Perplexity), the accuracy rates of ChatGPT and Gemini were found to be 90% and 75%, respectively. Statistically significant differences in performance were observed among the LLMs ($p = 0.023$).[14] Similarly, in a study comparing the diagnostic accuracy of ChatGPT and Gemini across different medical specialties, the models achieved accuracy rates of 90% and 80%, respectively, in the clinical diagnosis category.[15] In a study conducted in Hungary, the diagnostic performance of AI models was evaluated by uploading images of acne and rosacea cases. The correct diagnosis rates were found to be 93% for GPT-4o and 21% for Gemini Flash 2.0, indicating that GPT-4o performed significantly better than Gemini Flash 2.0 in diagnosing acne and rosacea.[16] In addition to individual studies, based on a meta-analysis, it was suggested that ChatGPT should continue to be considered not as an independent clinical decision-making tool, but rather as a complementary support system in dermatological diagnosis.[17] In a separate study, the correct pathological diagnosis prediction rates for AI models based on MR images were reported as 74.33% for ChatGPT-4o, 65.64% for Grok, and 74.54% for Gemini.[8] These differences in model accuracy largely depend on the type of data used and the working methods. In our study based solely on lesion images, diagnostic success was lower compared to previous multimodal studies. Our findings show that ChatGPT has a higher accuracy rate than Gemini, and this is consistent with other studies in the literature.

Table 2. Comparative Performance of Artificial Intelligence Models on Commonly Answered Cases

Model	Correct Answer (Total of 21 Questions) (n, %)	*P* value
ChatGPT-4	9 (43%)	0.829
Grok-3	11 (52.4%)	
Gemini 2.0	10 (47.6%)	

Limitations

In this study, only visual data was presented to the AI models; clinical history, physical examination, and laboratory findings, which play an important role in diagnosis, were not used. This situation may limit the diagnostic accuracy of the models. All tools were evaluated using the same standardized

prompt, which may not fully reflect variations in real-world clinical interactions. Furthermore, the tools were not allowed to request additional clinical information, which may have limited their diagnostic performance. Additionally, due to the abundance of valid treatment options and differences based on expert opinions, case-based treatment recommendations have not been systematically analyzed. Therefore, the study focused solely on diagnosis and distinguishing between malignant and benign conditions. The limited number of cases (n = 34) reduces the generalizability of the results obtained. Furthermore, the limited number of studies in the literature that evaluate all three artificial intelligence models (ChatGPT-4, Grok-3, Gemini 2.0) together has prevented the findings related to the Grok model from being compared with the literature. In addition, the tools were accessed between March 6 and 14, 2025, and their performance may vary over time due to ongoing updates.

Strengths

Our study is one of the first comparative analyses conducted in a Turkish-speaking context that directly compares three popular AI models (ChatGPT-4, Grok-3, Gemini-2.0) using dermatological images for diagnosis. Providing only lesion images to the AI models simulates the visual preliminary assessment (inspection) process at the initial contact in the clinic. All models were tested on the same cases and the same questions, and their accuracy rates and response capabilities were compared in an unbiased and standardized manner. It was observed that all models provided correct answers in distinguishing malignant from benign lesions when they reported a diagnosis, demonstrating their potential for clinical safety in this regard.

Conclusion

This study demonstrated the limitations of using ChatGPT-4, Grok-3, and Gemini 2.0 models as independent diagnostic tools in the clinical evaluation of dermatological cases. However, AI models may serve as complementary decision-support systems, particularly in the assessment of specific skin lesions in primary care settings. Enhancing model performance through training on larger and more diverse datasets, along with improving alignment with current clinical guidelines, may strengthen the future applicability of AI in dermatology.

Acknowledgments

The authors would like to thank Türkiye Klinikleri Press for granting permission to use the dermatological case images from the book Approach to Dermatological Diseases in Family Medicine in this study.

Previous Publication Status

This study was presented as an oral presentation at the 24th International Eastern Mediterranean Family Medicine Congress held in Adana on May 09–12, 2025.

Funding

The authors received no specific funding for this work.

Conflict of Interest

The authors declare that they have no competing interests.

Artificial Intelligence Usage Statement

Artificial intelligence–based tools (ChatGPT-4, Grok-3, and Gemini 2.0) were used as part of the study to evaluate their diagnostic performance on dermatological images. In addition, AI-assisted tools were used for minor language editing and clarity improvement during manuscript preparation. All study design, data analysis, and interpretation of the results were performed by the authors.

Ethics Committee Approval

Ethics committee approval was not required as this study did not involve human or animal participants.

Informed Consent Statement

Not applicable.

References

1. Xu Y, Liu X, Cao X, et al. Artificial intelligence: A powerful paradigm for scientific research. Innov 2021;2:100179.

2. Sanli DET, Sanli AN, Buyukdereli Atadag Y, Kurt A, Esmerer E. GPT-4o and specialized AI in breast ultrasound imaging. J Ultrasound Med 2025;44:1993-2004.

3. Rimoin L, Altieri L, Craft N, Krasne S, Kellman PJ. Training pattern recognition of skin lesion morphology, configuration, and distribution. J Am Acad Dermatol 2015;72:489-495.

4. Zheng J, Xie K, Zhang D, Lv Z, Yu X. Stepwise self-knowledge distillation for skin lesion image classification. Sci Rep 2025;15:25238.

5. Li Z, Koban KC, Schenck TL, Giunta RE, Li Q, Sun Y. Artificial intelligence in dermatology image analysis: Current developments and future trends. J Clin Med 2022;11:6826.

6. Clebak KT, Partin MT, Newman R, et al. Artificial intelligence (AI) adoption, policies, and goals in family medicine: A survey of department chairs. J Am Board Fam Med 2025;38.

7. Liaw W, Kakadiaris IA. Artificial intelligence and family medicine: Better together. Fam Med 2020;52:8-10.

8. Sozer A, Sahin MC, Sozer B, et al. Do LLMs have “the eye” for MRI? Evaluating GPT-4o, Grok, and Gemini on brain MRI performance. Diagnostics (Basel) 2025;15.

9. Wu X, Cai G, Guo B, et al. A multi-dimensional performance evaluation of large language models in dental implantology: Comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios. BMC Oral Health 2025;25.

10. Freeman K, Dinnes J, Chuchu N, et al. Algorithm-based smartphone apps to assess risk of skin cancer in adults. BMJ 2020;368:m127.

11. Kurtoglu Y, editor. Approach to dermatological diseases in family medicine. Ankara: Türkiye Klinikleri Press; 2024.

12. Pradhan P. Accuracy of ChatGPT 3.5, 4.0, 4o and Gemini in diagnosing oral potentially malignant lesions based on clinical case reports and image recognition. Med Oral Patol Oral Cir Bucal 2025;30:e224-e231.

13. Chen R, Fettel KD, Nguyen DH, Nambudiri VE. Evaluating the performance of ChatGPT on dermatology board-style exams: A meta-analysis of text-based and image-based question accuracy. J Am Acad Dermatol 2025;93:520-522.

14. Fan KS, Fan KH. Dermatological knowledge and image analysis performance of large language models based on specialty certificate examination in dermatology. Dermato 2024;4:124-135.

15. Fattah FH, Salih AM, Salih AM, et al. Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: A scoping review. Front Digit Health 2025;7:1482712.

16. Boostani M, Banvolgyi A, Goldust M, et al. Diagnostic performance of GPT-4o and Gemini Flash 2.0 in acne and rosacea. Int J Dermatol 2025;64:1881-1882.

17. Chen R, Fettel KD, Nguyen DH, Nambudiri VE. Diagnostic accuracy of ChatGPT in dermatology: A meta-analysis of textual versus visual prompts. J Am Acad Dermatol 2025.

