Journal of Pakistan Association of Dermatologists. 2026;**36(3)**:413-418.

Original Article

Application of Generative Artificial Intelligence Models in diagnosis of localized skin diseases in a tertiary care teaching hospital in Pakistan

Samrat Habib*, Zafar Ullah Khan, Khawar Khurshid

Department of Dermatology, Rashid Latif Medical College, Lahore.

Abstract

Background Generative artificial neural networks like ChatGPT, Gemini, and Grok are rapidly being investigated for applications in medicine. Their application in dermatology, particularly for visual diagnostics, is still in its early phases.

Objective To assess the clinical usefulness and diagnostic precision of generative AI tools in identifying localized skin diseases from photographs, as well as the satisfaction of consultant dermatologists and suggestions for their application in clinical training.

Methods A total of 216 patients were enrolled with clinical images examined utilizing ChatGPT, Gemini,and Grok (72 patients each). Each AI tool made a diagnosis based only on the image input. Consultant dermatologists independently reviewed all patients and served as the gold standard. Consultant satisfaction ratings and recommendations for future use were also collected. SPSS was used to analyze the data, with Chi-square tests used to determine statistical significance ($P<.05$).

Results With a diagnostic accuracy of 61.1%, ChatGPT outperformed Gemini, (1.4%) and Grok (0%). Grok had the most missing or improper responses, whereas ChatGPT had the highest consultant satisfaction rate (54.2%). Chatgpt was recommended (79.1%) by consultant dermatologist in comparison with other AI tools. Variations were statistically significant ($P<.01$) in all domains: recommendation, accuracy, and satisfaction.

Conclusion In terms of clinical recommendation, consultant satisfaction, and diagnostic accuracy, ChatGPT performed better than Gemini, and Grok. Although more research involving clinical data is necessary, its image-based diagnostic capability shows potential for future usage in dermatology teaching and decision assistance.

Keywords Artificial Intelligence(AI); Skin Diagnosis; ChatGPT; Gemini; Grok.

Article Received on 01.04.2026

Revised on 18.09.2026

Accepted on 25.09.2026

Published on 30.09.2026

Citation: Habib S, Khan ZU, Khurshid K. Application of generative artificial intelligence models in diagnosis of localized skin diseases in a tertiary care teaching hospital in Pakistan. *J Pak Assoc Dermatol*. 2026;**36(3)**:413-418.
Doi: https://doi.org/10.66344/jpad.v36i3.3387

Introduction

In recent years AI has considerably improved disease diagnosis by expanding the integration of data and diagnostic capacities.[1] The technology has enhanced data processing and pattern recognition capabilities, leading to greater precision in medical assessment.[2] It has evolved from simple neural networks to the complex generative models, which aids diagnostic imaging in health sector. This innovation promises huge potential in different applications in healthcare by providing new insights and tools for healthcare professionals as well as

Address for correspondence
Dr. Samrat Habib, Post Graduate Resident,
Department of Dermatology,
Rashid Latif Medical College, Lahore,
Pakistan.
Email: samrathabib1@gmail.com

academics.[3] With all of this progress, practical implementation of these types of models remains questionable, especially under difficult clinical situations.[2]

AI is quickly changing the field of dermatology through increased image recognition and predictive algorithms.[4] The practice in clinical dermatology is being transformed by complex AI models that use large-scale self-supervised learning to reduce health disparities, improve screening, generate images, and support clinical tasks.[5] Beyond diagnostics, generative AI has a significant impact on patient education by providing relevant information about symptoms, therapies, drugs, and lifestyle changes.[6]

Skin malignancies notably melanomas, squamous cell carcinomas and basal cell carcinomas pose unique therapeutic and clinical concerns. Melanomas are estimated to impact 13.1 million individuals worldwide by 2030. The use of artificial intelligence in dermatology offers potential advances in early diagnosis and significantly improve patient outcomes.[7] Large language models (LLMs), such as ChatGPT, Grok, and Gemini, have shown potential in dermatological diagnosis and dermatopathology. Moreover, acne, which impacts 9.4% individuals worldwide, is an outstanding example of how artificial intelligence might change dermatological diagnostics. Its diagnosis has traditionally been subjective, and recent advances in AI have demonstrated potential for improving the recognition and classification of different variants of acne.[8] Unlike image-processing models such as convolutional neural networks (CNNs) and generative adversarial networks (GANs), LLMs such as ChatGPT, Grok, and Gemini are conversational AI systems capable of generating diagnostic responses based on the information provided to them. However, these models require greater validation because they can produce factually wrong predictions, particularly in complex situations such as cutaneous adnexal lesions.[9,10]

Despite these advancements, difficulties such as the accuracy of data, credibility, and ethical issues remain major barriers, which highlights the importance of ongoing research and development.[11] Ethical and regulatory frameworks along with involvement of all stakeholders is needed for the successful integration of AI in Dermatology.[4,12] Although generative AI in Pakistan has enormous potential in dermatology, it faces significant limitations such as inadequate digital record-keeping, lack of skilled professionals, unstable internet access, technical infrastructure and lack of awareness among healthcare professionals in Pakistan. In order to fully benefit from generative AI, these issues must be resolved.[13]

This study aims to assess how well the generative AI tools ChatGPT, GROK, and Gemini diagnose localized skin diseases using dermatological images alone, without additional patient history or clinical information. By comparing these AI-generated diagnoses with the clinical diagnosis provided by consultant dermatologists, the study seeks to evaluate the accuracy of these diagnoses. It also aims to comprehend how these tools might help with diagnostic decision-making in practical contexts. Given the rising prevalence of skin diseases and the demand for effective diagnostic assistance, this research is important. However, this study specifically evaluates the visual diagnostic accuracy of these AI tools and does not represent comprehensive clinical diagnosis. Only dermatological images showing lesions that could be reasonably identified on visual examination were included, while complex cases requiring detailed patient history, physical examination, laboratory investigations, or other clinical information were excluded. Therefore, the findings should be interpreted in the context of image-based diagnostic assessment and may not be directly applicable to complex dermatological cases requiring full clinical evaluation.

Methods

This cross-sectional, observational study was carried out in the Department of Dermatology, Rashid Latif Medical College, Lahore, using 216 clinical photographs of different localized skin diseases. This study was conducted between 01-01-2025 to 01-07-2025. Identical photographs were provided independently to all three AI tools- Grok (n = 72), Gemini (n=72), and ChatGPT (n=72)- to allow a direct head-to-head comparison of their diagnostic performance. The Epi Calculator was used to determine the sample size, using a 95% confidence level, an expected proportion of 50%, and an absolute precision of 7%. Informed consent from patients was taken. Standard protocol was followed to obtain high-quality clinical photographs of localised skin diseases. The AI tools were given specific instruction prompt with photograph of specific disease to produce a diagnosis. A consultant dermatologist also conducted an independent evaluation of each patient. The gold standard for comparison was the dermatologist's diagnosis. This study adhered to all relevant ethical requirements and approved by the Institutional Review Board (IRB) of RLKU medical and dental college (Ref no: RLKUMC/IRB/0035/24 dated 28.08.2024). Principles of research ethics and relevant legal requirements were fully complied with, and duly considered during the entire process. Patient information was kept private and safe. All the results obtained were evaluated using SPSS version 26.0. A bivariate analysis was conducted, Chi-square test was employed for this analysis. To allow for a direct head-to-head comparison of diagnostic performance, ChatGPT, Gemini, and Grok independently assessed the identical clinical photos. The model version and testing date were noted, and all three AI models were assessed using the same standardized prompt, picture quality, diagnostic job, and testing circumstances. The exact standardized prompt provided to each AI tool was: “Please identify the most likely diagnosis based solely on the attached clinical photograph. Provide one primary dermatological diagnosis only.” To evaluate AI performance across various illness groups, cases were divided into key diagnostic categories based on the final diagnosis. When clinically necessary, dermoscopy, microscopy, culture, histopathology, or other pertinent tests were used to support the reference diagnosis, which was made by consultant dermatologists. AI replies were categorized as accurate, somewhat correct, wrong, or non-response/ refusal; incorrect diagnoses and refusals were examined independently. Since all three AI models evaluated the identical instances, diagnostic performance was examined as paired data using the relevant statistical tests, and 95% confidence intervals were provided. Using predetermined, standardized criteria, consultant satisfaction and perceived clinical utility were evaluated and examined independently of diagnostic accuracy. Before being sent to AI platforms, all clinical pictures were anonymised, and the proper patient consent was acquired for their usage in research and processing by third-party AI platforms. The comparative performance of the AI models was not seen as proof of any platform's intrinsic superiority; rather, it was interpreted under the particular standardized research settings. The appropriate completed checklist was supplied as supplemental material, and the study was published in accordance with the EQUATOR guidelines for diagnostic-accuracy and clinical-AI research.

Result

Compared to Gemini (1.39%) and Grok (0%), ChatGPT had the highest diagnostic accuracy, with 61.11% of diagnoses being accurate. Although ChatGPT demonstrated higher diagnostic accuracy than other AI tools evaluated in this study, its overall accuracy of 61.1% indicates that substantial diagnostic errors remain. Therefore, these findings suggest that ChatGPT should not be relied upon as an independent diagnostic or training tool without appropriate clinical oversight and supervision. To evaluate the relationship between the AI model and diagnosis accuracy, a chi-square test was used. Due to a large percentage of incorrect and non-generated

responses, Grok and Gemini did poorly. Gemini and Grok refused to provide answers in a number of instances, citing that they were not licensed dermatologists, ethical concerns, or policy limitations. These safety-related refusals were recorded as a separate outcome category and were not classified as incorrect medical diagnoses. Diagnostic errors were analyzed independently from non-response or safety-related refusals to distinguish limitations in diagnostic performance from the models' refusal behavior.

Consultant satisfaction was highest with ChatGPT (54.17%), followed by Gemini (0.0%) and Grok (0.0%), having the highest proportion of unsatisfactory outcomes. In terms of future use of AI tools for training residents and general physicians, Chatgpt is preferred mostly by consultant dermatologists (79.2%), compared to 16.7% for Gemini and 13.9% for Grok.

Three major performance indicators-diagnostic accuracy, consultant satisfaction, and referral rate-were examined to determine whether there were statistically significant differences amongst the three AI models-ChatGPT, Grok, and Gemini-using the chi-square test. The *P*-value in each of these domains was 0.01, which is less than the standard cut off of 0.05. This suggests that the observed variations between the AI models are statistically significant and unlikely to have happened by coincidence. The reported performance benefits of ChatGPT are not the result of chance, but rather represent actual underlying differences in model capabilities and clinician acceptance, according to the statistically significant Chi-square results across all three domains. These results support the idea that ChatGPT is more appropriate than Grok and Gemini for clinical decision assistance in dermatology.

Discussion

This study intended to compare and assess the performance of three AI models- ChatGPT, Grok, and Gemini- in the diagnosis of localised skin diseases. The key areas were diagnostic accuracy of the AI models, consultant dermatologist acceptance, and the possibility of recommending these tools to younger doctors and general practitioners.

Among the three models, ChatGPT had the highest diagnostic accuracy, correctly diagnosing 61.1% of cases. Grok, on the other hand, provided no completely accurate diagnoses and had the highest rate of missing responses (44.4%), while Gemini demonstrated limited performance, with only one accurate diagnosis and a high percentage of partially correct (37.5%) and incorrect responses.

Table 1 Diagnostic accuracy by different AI models.

AI models	*Accurate diagnosis f (%)*	*Partially Accurate diagnosis f (%)*	*Inaccurate diagnosis f (%)*	*Non-generated responses f (%)*	*Total cases*	*P-value*
Chat gpt	44 (61.11%)	21 (29.17%)	7 (9.72%)	0 (0.00%)	72	.01
Grok	0 (0.00%)	19 (26.39%)	21 (29.17%)	32 (44.44%)	72	.01
Gemini	1 (1.39%)	27 (37.50%)	24 (33.33%)	20 (27.78%)	72	.01

Table 2 Consultant Dermatologists satisfaction of different AI models.

AI models	*Satisfied f (%)*	*Moderately Satisfied f (%)*	*Not satisfied f (%)*	*Non-generated responses f(%)*	*Total cases*	*P-value*
Chat gpt	39 (54.17%)	26 (36.11%)	7 (9.72%)	0 (0.00%)	72	<.01
Grok	0 (0.00%)	18 (25%)	22 (30.56%)	32 (44.44%)	72	<.01
Gemini	0 (0.00%)	32 (44.44%)	20 (27.78%)	20 (27.78%)	72	<.01

Table 3 Consultant Dermatologists recommendation of AI models.

AI models	*Recommended F (%)*	*Not recommended f (%)*	*Non-generated responses f (%)*	*Total cases*	*P-value*
Chat gpt	57(79.17%)	15(20.83%)	0(0.00%)	72	<.01
Grok	10(13.89%)	30(41.67%)	32(44.44%)	72	<.01
Gemini	12(16.67%)	39(54.17%)	20(27.78%)	72	<.01

The Chi-square test revealed a significant difference ($P<.01$) in diagnostic response categories across the AI models. However, this statistical difference reflects performance under the conditions of the present study and should not be interpreted as evidence that ChatGPT has inherently superior medical reasoning or that Grok and Gemini lack technical diagnostic capability. In particular, the high proportion of non-responses from Grok and Gemini may have been influenced by organisational safety policies and restrictions on diagnosing medical images rather than solely by limitations in their medical capabilities. Therefore, refusal to answer should be distinguished from an incorrect diagnosis when interpreting comparative AI performance.

Similarly, in another study, Hirosawa *et al*.[14] used ChatGPT to generate differential diagnoses and reported an accuracy rate of 53.3%. Our study demonstrated a comparable accuracy of 61.1%. In contrast to their study, which included systemic symptoms, we used dermatological images without clinical presentation to simulate a self-diagnosis. Johnson *et al*.[15] similarly performed a comprehensive evaluation in which ChatGPT was asked a variety of medical questions. They found that the median accuracy of responses was relatively high. This suggests that ChatGPT can provide accurate responses for a considerable proportion of medical questions, highlighting its potential as a supplementary clinical and educational tool.[15] Prior research indicates that, while ChatGPT has potential for self-diagnosis, its reliability and consistency are variable, particularly in the absence of symptom-specific clinical information.[16]

Consultant feedback revealed significant differences between models. ChatGPT received the highest satisfaction and recommendation rate (79.1%), indicating greater perceived usability and acceptance among participating consultants. However, this recommendation rate represents perceived usefulness and should not be interpreted as evidence that ChatGPT is sufficiently reliable for independent diagnosis or unsupervised resident training. It's approximately 39% combined rate of partially correct and incorrect responses demonstrates that substantial diagnostic limitations remain. Therefore, ChatGPT should never be recommended as a stand-alone training tool for medical residents; its output should only be used as supplementary information under strict supervision and verification by qualified clinicians. Grok and Gemini had lower recommendation rates and higher proportions of refusals. These refusals may have resulted from ethical and safety mechanisms designed to prevent automated diagnosis from medical photographs and should not necessarily be regarded as evidence of poor medical or technical ability. Such safety limitations are important when interpreting AI performance and considering integration into clinical practice. In addition, data quality should be improved, algorithms made more transparent, AI incorporated appropriately into clinical workflows, ethical concerns addressed, regulations followed, and relevant stakeholders involved to ensure the safe and sustainable application of AI in dermatology.[12]

Conclusion

ChatGPT showed significantly higher diagnostic precision, dermatologist satisfaction, and recommendation rates than Grok and Gemini. However, because the models were tested using isolated images without clinical history, these findings cannot be directly generalized to routine practice or medical training. Further clinical studies are needed before AI can be routinely integrated into dermatology

Limitations This study has a few limitations, including the clinical assessment of photographs by consultant dermatologists rather than histo-pathological confirmation.

Financial support and sponsorship None.

Conflict of interest No conflict of interest.

Author's contribution

SH: Substantial contribution to conception and design and manuscript writing.

ZUK: Substantial contributions to conception and design and manuscript writing.

KK: Substantial contributions to conception and design and critical review of the manuscript.

Every author has given final approval of the manuscript version to be published and agreed to be accountable for all aspects of the work.

References

1. Vallverdú J. Challenges and Controversies of Generative AI in Medical Diagnosis. *Euphyía*, 2023;**17(32)**:88-121. https://doi.org/10.33064/32euph4957
2. Akyon SH, Akyon FC, Camyar AS, Hızlı F, Sari T, Hızlı Ş. Evaluating the capabilities of generative AI tools in understanding medical papers: qualitative study. *JMIR Med Inform*. 2024;**12**:e59258. doi:10.2196/59258.
3. Shokrollahi Y, Yarmohammadtoosky S, Nikahd MM, Dong P, Li X, Gu L. A comprehensive review of generative AI in healthcare. arXiv [Preprint]. 2023 Oct 1 [cited 2026Sep24];arXiv:2310.00795. doi:10.48550/arXiv.2310.00795
4. Sengupta D. Artificial intelligence in diagnostic dermatology: challenges and the way forward. *Indian Dermatol Online J*. 2023;**14(6)**:782-7. doi:10.4103/idoj.idoj_462_23. PMID: 38099026; PMCID: PMC10718130.
5. Gui H, Omiye JA, Chang CT, Daneshjou R. The promises and perils of foundation models in dermatology. *J Invest Dermatol*. 2024;**144(7)**:1440-8. doi:10.1016/j.jid.2023.12.019. PMID:38441507.
6. Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: a pilot study. *Indian Dermatol Online J*. 2023;**14(4)**:482-6. doi:10.4103/idoj.idoj_72_23. PMID:37521213; PMCID:PMC10373821.
7. Saeed M, Naseer A, Masood H, Rehman SU, Gruhn V. The power of generative AI to augment for enhanced skin cancer classification: a deep learning approach. IEEE Access.2023;11:130330-130344. doi:10.1109/ACCESS.2023.3332628
8. Sankar A, Chaturvedi K, Nayan AA, Hesamian MH, Braytee A, Prasad M. RETRACTED: Utilizing generative adversarial networks for acne dataset generation in dermatology. *Bio Med Informatics*. 2024;**4(2)**:1059-70.
9. McCrary MR, Galambus J, Chen WS. Evaluating the diagnostic performance of a large language model-powered chatbot for providing immunohistochemistry recommendations in dermatopathology. *J Cutan Pathol*. 2024;**51(9)**:689-695. doi:10.1111/cup.14631. Epub 2024 May 14. PMID:38744501.
10. Pillai J, Li B. Generative artificial intelligence in dermatology: Recommendations for future studies evaluating the clinical knowledge of models. Vol. 30, Skin Research and Technology. John Wiley and Sons Inc; 2024.
11. Singh JP. The impacts and challenges of generative artificial intelligence in medical education, clinical diagnostics, administrative efficiency, and data generation. Int J Appl Health Care Anal. 2023;8(5):37-46.
12. Manoharan P, Surapaneni KM. Assessing the diagnostic capability of ChatGPT through clinical case scenarios in dermatology. Indian J Dermatol Venereol Leprol. 2025;91(Suppl 1):S27-S28. doi:10.25259/IJDVL_1267_2023.
13. Habib MM, Hoodbhoy Z, Siddiqui MR. Knowledge, Attitudes, and Perceptions of Healthcare Students and Professionals on the Use of Artificial Intelligence in Healthcare in Pakistan. *PLOS Digital Health*. 2024 May;**3(5)**:e0000443. doi: 10.1371/journal.pdig.0000443.
14. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health*. 2023;**20(4)**:3378.
15. Johnson D, Goodman R, Patrinely JR, Stone CA, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq [Preprint]. 2023 Feb 28:rs.3.rs-2566942. doi:10.21203/rs.3.rs-2566942/v1.
16. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, Fujita K. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. *J Med Internet Res*. 2023;**25**:e47621. doi:10.2196/47621