

## Multimodal Large Language Models in Dermatologic Diagnosis: A Game Changer or a Clinical Risk?

Ümit AKPINAR<sup>1\*</sup>, Mehmet Semih ÇELİK<sup>2</sup>, Ekrem CIVAŞ<sup>1</sup>, Yıldız HAYRAN<sup>3</sup> and Başak YALÇIN<sup>1</sup>

<sup>1</sup>Private Dermatology Practice, Ankara, Turkey, <sup>2</sup>Gazi Yaşargil Education and Research Hospital, Diyarbakır, Turkey, and <sup>3</sup>Ankara City Hospital Department of Dermatology, Ankara, Turkey. Email: drumitakpinar@gmail.com

Submitted Apr 6, 2026. Accepted after revision Sept 15, 2026

Published Oct 6, 2026. DOI: 10.2340/actadv.v106.adv-2026-0586 Acta Derm Venereol 2026; 106: adv-2026-0586.

Dermatology is a visually driven specialty and is therefore particularly suited to artificial intelligence (AI)-based image analysis, with high diagnostic accuracy reported in selected settings (1). However, concerns remain regarding the real-world applicability, data quality, algorithmic bias, generalizability and clinical integration of AI systems (2–5). Multimodal large language models (MLLMs), which integrate visual and contextual information, are increasingly used for diagnostic queries, yet their reliability in real-world dermatologic settings remains uncertain. We therefore evaluated the diagnostic accuracy of ChatGPT and Gemini using histopathologically confirmed dermatologic lesions and assessed whether providing clinician-generated differential diagnoses improves their performance.

### MATERIALS AND METHODS

This retrospective image-based diagnostic accuracy study included 329 anonymized clinical photographs representing 138 diagnoses from the investigator's clinical archive. Only histopathologically confirmed lesions were included, with histopathology serving as the reference standard. Images were reviewed for adequate quality and absence of identifiable patient information. Diagnoses and case-level data are provided in Table SI.

Lesions were categorized into 6 predefined groups: (1) tumoral/tumour-like lesions (2), inflammatory dermatoses (3), infectious dermatoses (4), urticarial/reactive dermatoses (5), vesiculobullous and pustular dermatoses and (6) systemic/immune-related dermatoses.

ChatGPT and Gemini were evaluated under 2 conditions. In the image-only condition, each model received a clinical image and provided the 3 most likely diagnoses and the single most likely diagnosis. In the clinical-context condition, the same image was accompanied by 3 clinician-generated differential diagnoses, from which the model selected the most likely diagnosis.

All cases were entered into the models between 20 March and 15 April 2026. Evaluations were performed using paid subscription versions of ChatGPT (OpenAI; GPT-5.3 Instant) and Gemini (Google; Gemini 3 Flash)

through their respective consumer web interfaces. Standard inference modes were used, without manually selecting Thinking, Deep Think or other extended-reasoning modes. A new chat session was initiated for each case to prevent carryover effects and ensure independent model responses. No dedicated anonymity mode was used during the evaluation process. All submitted images were anonymized and contained no identifiable patient information.

Standardized English-language prompts were used consistently for both models according to the 2 evaluation conditions described above.

Performance was assessed using top-1 accuracy (primary diagnosis matching the reference diagnosis) and top-3 accuracy (reference diagnosis included among the 3 suggested diagnoses). For the clinical-context condition, only top-1 accuracy was calculated.

### Statistical analysis

All statistical analyses were performed using IBM SPSS Statistics version 26 (IBM Corp., Armonk, NY, USA). Descriptive data were expressed as frequencies and percentages. Comparisons between diagnostic accuracies were performed using McNemar and chi-square tests, as appropriate. A *p*-value <0.05 was considered statistically significant.

### RESULTS

A total of 329 dermatologic lesions were included. Their distribution across diagnostic groups is presented in **Table I** (see complete list of diagnoses in Table SI).

Overall diagnostic performance is summarized in **Table II**. For both models, top-3 accuracy was higher than top-1 accuracy. Providing clinician-generated differential diagnoses significantly improved diagnostic accuracy for both models (*p*<0.001), with a greater improvement observed for Gemini.

**Table I. Distribution of dermatologic lesions**

| Diagnostic category                        | <i>n</i> (%) |
|--------------------------------------------|--------------|
| Tumoral / Tumour-like lesions              | 78 (23.7)    |
| Inflammatory dermatoses                    | 93 (28.3)    |
| Infectious dermatoses                      | 24 (7.3)     |
| Urticarial / Reactive dermatoses           | 33 (10.0)    |
| Vesiculobullous and pustular dermatoses    | 34 (10.3)    |
| Other systemic / immune-related dermatoses | 67 (20.4)    |

Table II. Overall diagnostic performance of AI mode

| Model   | Top-3 accuracy (%) | Top-1 accuracy (%) | With clinical differential diagnosis (%) |
|---------|--------------------|--------------------|------------------------------------------|
| ChatGPT | 45.3               | 27.7               | 48.0                                     |
| Gemini  | 48.0               | 28.3               | 61.7                                     |

In image-only evaluation, no significant difference was observed between ChatGPT and Gemini in either top-1 or top-3 accuracy. When clinician-generated differential diagnoses were provided, Gemini achieved significantly higher accuracy than ChatGPT ( $p<0.001$ ).

In image-only evaluation, both models performed better in infectious and tumoral/tumour-like lesions, whereas vesiculobullous and pustular dermatoses were the most challenging (Table III). Clinician-generated differential diagnoses improved accuracy across categories, particularly in urticarial/reactive and systemic/immune-related dermatoses. With clinical context, Gemini achieved significantly higher accuracy in systemic/immune-related dermatoses ( $p=0.002$ ).

In malignant lesions, image-only performance was comparable between the models. Providing clinician-generated differential diagnoses significantly improved diagnostic accuracy, particularly for Gemini ( $p<0.001$ ).

DISCUSSION

Previous studies have demonstrated dermatologist-level AI performance in selected image-classification tasks (6, 7), although reliance on curated datasets has raised concerns regarding real-world generalizability (8).

Both ChatGPT and Gemini demonstrated moderate diagnostic accuracy, which improved significantly with clinician-generated differential diagnoses, supporting MLLMs as adjunctive rather than standalone diagnostic tools. This is consistent with evidence that clinical context can improve AI performance (9).

The models performed similarly in image-only evaluation, whereas Gemini achieved higher accuracy with clinician-generated differential diagnoses, suggesting greater benefit from contextual information. The mechanism underlying this difference cannot be determined from the present study.

The increasing use of MLLMs by patients and clinicians raises important safety concerns. In our study, top-1 accuracy in image-only evaluation was approximately 20% for both models. Although accuracy increased to 48% for ChatGPT and 61% for Gemini

when clinician-generated differential diagnoses were provided, these levels remain insufficient to support independent diagnostic use.

Malignant lesions represent a particular safety concern because diagnostic delay may adversely affect clinical outcomes. In this subgroup, image-only top-1 accuracy was only 25% for ChatGPT and 35% for Gemini, highlighting the limitations of independent MLLM-based assessment of potentially malignant lesions.

Previous studies reported improved diagnostic performance with combined visual and clinical information and variable accuracy across LLMs (10, 11). Real-world dermatology applications have also shown moderate accuracy and skin tone-related bias (12), further highlighting the importance of diverse and well-annotated datasets (13).

Several limitations should be considered. First, no additional clinical information beyond images was provided during image-only evaluation. Although this was a deliberate design choice to isolate visual diagnostic performance, it may have limited overall accuracy. Second, the retrospective design may introduce selection bias. Third, model outputs may be influenced by prompt structure and interaction variability, which could affect reproducibility. As the study was conducted in Türkiye, the geographic and demographic characteristics of the study population may limit generalizability to populations from other regions. Fitzpatrick skin phototype was not systematically recorded; therefore, the distribution of skin phototypes within the dataset could not be reliably determined.

Despite these limitations, the use of histopathologically confirmed diagnoses as the reference standard and the inclusion of a broad range of dermatologic conditions strengthen the validity of the diagnostic comparisons.

From a clinical perspective, these findings suggest that MLLMs may be most effectively used as adjunctive tools within structured diagnostic workflows,

Table III. Diagnostic performance across categories

| Diagnostic category                        | ChatGPT Top-3 (%) | ChatGPT Top-1 (%) | ChatGPT+Clinical differential (%) | Gemini Top-3 (%) | Gemini Top-1 (%) | Gemini+Clinical differential (%) |
|--------------------------------------------|-------------------|-------------------|-----------------------------------|------------------|------------------|----------------------------------|
| Tumoral / Tumour-like                      | 52.6              | 32.1              | 51.3                              | 62.8             | 37.2             | 64.1                             |
| Inflammatory                               | 48.4              | 30.1              | 50.5                              | 47.3             | 26.9             | 60.2                             |
| Infectious                                 | 66.7              | 45.8              | 45.8                              | 54.2             | 25.0             | 54.2                             |
| Urticarial / Reactive                      | 36.4              | 21.2              | 63.6                              | 45.5             | 30.3             | 60.6                             |
| Vesiculobullous / Pustular                 | 32.4              | 17.6              | 38.2                              | 17.6             | 11.8             | 55.9                             |
| Other systemic / immune-related dermatoses | 35.8              | 20.9              | 43.3                              | 46.3             | 28.4             | 67.2                             |

particularly when guided by clinician-generated differentials.

In conclusion, MLLMs demonstrated limited diagnostic accuracy when used independently but performed substantially better when guided by clinician-generated differential diagnoses. Their low accuracy for malignant lesions remains an important clinical limitation. These systems should therefore be regarded as adjuncts to, rather than substitutes for, dermatologist assessment.

## ACKNOWLEDGEMENTS

*Data availability statement:* The data supporting the findings of this study are available from the corresponding author upon reasonable request. The full case-level dataset is also provided as Table SI.

*Ethics committee:* The study protocol was approved by the Clinical Research Ethics Committee of Gazi Yaşargil Training and Research Hospital, Diyarbakır Provincial Health Directorate (approval no. 116; March 18, 2026). The study was conducted in accordance with the principles of the Declaration of Helsinki. This study was conducted using fully anonymized dermatologic images. No identifiable patient information was present in the images; therefore, separate informed consent was not required.

*The authors have no conflicts of interest to declare.*

## REFERENCES

- Ohaya C, Ogbaudu E, Choi RE, Ko J. Artificial intelligence and deep learning for skin image analysis. *Dermatol Clin* 2025; 43: 541–552. <https://doi.org/10.1016/j.det.2025.05.004>
- Branaccio G, Balato A, Malvey J, Puig S, Argenziano G, Kittler H. Artificial intelligence in skin cancer diagnosis: A reality check. *J Invest Dermatol* 2024; 144: 492–499. <https://doi.org/10.1016/j.jid.2023.10.004>
- Gordon ER, Trager MH, Kontos D, Weng C, Geskin LJ, Dugdale LS, et al. Ethical considerations for artificial intelligence in dermatology: a scoping review. *Br J Dermatol* 2024; 190: 789–797. <https://doi.org/10.1093/bjd/ljae040>
- Sengupta D. Artificial intelligence in diagnostic dermatology: challenges and the way forward. *Indian Dermatol Online J* 2023; 14: 782–787. [https://doi.org/10.4103/idoj.idoj\\_462\\_23](https://doi.org/10.4103/idoj.idoj_462_23)
- Marri SS, Albadri W, Hyder MS, Janagond AB, Inamadar AC. Efficacy of an Artificial Intelligence App (Aysa) in dermatological diagnosis: Cross-sectional analysis. *JMIR Dermatol* 2024; 7: e48811. <https://doi.org/10.2196/48811>
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542: 115–118. <https://doi.org/10.1038/nature21056>
- Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr SMR, Jafari MH, Ward K, et al. Melanoma detection by analysis of clinical images using convolutional neural network. *Annu Int Conf IEEE Eng Med Biol Soc* 2016; 2016: 1373–1376. <https://doi.org/10.1109/EMBC.2016.7590963>
- De A, Sarda A, Gupta S, Das S. Use of artificial intelligence in dermatology. *Indian J Dermatol* 2020; 65: 352–357. [https://doi.org/10.4103/ijd.IJD\\_418\\_20](https://doi.org/10.4103/ijd.IJD_418_20)
- Gustafson E, Pacheco J, Wehbe F, Silverberg J, Thompson W. A machine learning algorithm for identifying atopic dermatitis in adults from electronic health records. *Proc (IEEE Int Conf Healthc Inform)* 2017; 2017: 83–90. <https://doi.org/10.1109/ICHI.2017.31>
- Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat Commun* 2024; 15: 5649. <https://doi.org/10.1038/s41467-024-50043-3>
- Cirkel L, Lechner F, Henk LA, Krusche M, Hirsch MC, Hertl M, et al. Large language models for dermatological image interpretation - a comparative study. *Diagnosis (Berl)* 2026; 13: 75–81. <https://doi.org/10.1515/dx-2025-0014>
- Police PR, Danda S, Karri SR. Diagnostic accuracy of artificial intelligence dermatology apps compared to clinical evaluation in Indian patients with common skin conditions. *Int J Res Dermatol* 2025; 11: 284–290. <https://doi.org/10.18203/issn.2455-4529.IntJResDermatol20252064>
- Biswas S, Achar U, Hakim B, Achar A. Artificial intelligence in dermatology: A systematized review. *Int J Dermatol Venereol* 2025; 8: 33–39. <https://doi.org/10.1097/JD9.0000000000000404>