Evaluating GPT-4o for Automated Classification of Skin Lesions Using the HAM10000 Dataset

Nitin Chetla, BA; Matthew Chen, BS; Aaron Smith, BS; Tamer R. Hage, BS; Joseph Chang, BS; Priyanka Kadam, BS; Manasi Ladrigan, MD

PRACTICE **POINTS**

- Even with a multiclass classification framework designed to assist GPT-4o, the model encountered notable challenges in accurately diagnosing skin lesions.
- In its current form, GPT-4o may provide inaccurate and misleading information to patients who use its interface to evaluate suspected skin lesions. Patients should continue to seek clinical consultation from health care professionals.

To the Editor:

The widespread availability and popularity of ChatGPT (OpenAI) have sparked interest in its potential applications within various fields, including medical diagnostics.[1] In dermatology, large language models (LLMs) already are being cited as a possible way to reliably respond to common patient queries and produce concise patient education materials.[2,3] That being said, there is skepticism regarding the technology's efficacy and reliability in producing accurate treatment plans, with variability among popular LLMs; for example, a recent study by Chau et al[4] demonstrated that ChatGPT was best at providing specific and accurate information regarding patient-facing responses to questions about 5 dermatologic diagnoses compared to Google Bard (now rebranded as Google Gemini) and Bing AI (now rebranded as Microsoft Copilot), which more often produced inaccurate or nonspecific responses. Google Bard also declined to answer one prompt.[4] Large language models also have been evaluated in diagnosing skin lesions. In 2024, SkinGPT-4 (a pretrained multimodel LLM developed by Zhou et al[5]) achieved just over 80% accuracy in interpreting images of skin lesions and was considered informative by 82.5% of board-certified dermatologists, demonstrating that LLMs may have the potential to become integrated into clinical practice.[5]

Our study aimed to evaluate the performance of GPT-4o (OpenAI)—a widely accessible, low-cost LLM—in diagnosing dermatologic conditions using the HAM10000 dataset, a well-curated collection of dermatoscopic images developed for training and benchmarking artificial intelligence (AI) algorithms.[6] HAM10000 comprises images representing 7 distinct skin conditions: actinic keratoses (ak), basal cell carcinoma (bcc), benign keratosis (bk), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular skin lesions (vsl), providing a robust platform for multiclass classification assessment. We evaluated GPT-4o using 100 dermatoscopic images per condition to assess diagnostic accuracy, potential biases, and limitations in skin lesion identification. The HAM10000 dataset was selected because it offers a large standardized reference set of dermatoscopic (rather than conventional clinical) images commonly used in dermatologic AI research. GPT-4o was chosen due to its patient-friendly interface, widespread use, and prior reports suggesting greater reliability in skin lesion assessment compared with other LLMs.

One hundred images from each of the 7 dermatologic categories were randomly selected for use in our analysis in 2024. The images were selected by our data scientist (J.C.) through random sampling from the dataset. Each image was separately presented to GPT-4o without any preprocessing or modification alongside 2 prompts designed to evaluate the diagnostic capabilities of GPT-4o. Both prompts included the same list of 7 dermatologic conditions for answer choices but differed in contextual information, where prompt 1 provided patient

Nitin Chetla and Aaron Smith are from the School of Medicine, University of Virginia, Charlottesville. Matthew Chen and Priyanka Kadam are from the Renaissance School of Medicine, Stony Brook University, New York. Tamer R. Hage is from the School of Medicone, Virginia Commonwealth University, Richmond. Joseph Chang is from the University of Passau, Germany. Dr. Ladrigan is from Comprehensive Dermatology of Rochester, New York.

The authors have no relevant financial disclosures to report.

The eTables and eFigures are available in the Appendix online at www.mdedge.com/cutis.

Correspondence: Tamer R. Hage, BS (tamerwh@gmail.com).

Cutis. 2026 March;117(3):98-100, E2-E4. doi:10.12788/cutis.1359

demographic information and localization of the dermatological condition but prompt 2 did not provide these details (Table). No follow-up questions were presented.

For prompt 1, the confusion matrix showed a strong bias toward detecting mel and bcc, with high true positives (mel, 83%; bcc, 37%)(eFigure 1). This pattern possibly suggests a tendency to favor malignant labels (eg, mel, BCC) when uncertainty is present. Interestingly, df and vsl also had notable true positives (46% and 37%, respectively), which is unexpected for less critical conditions because the model's correct classifications were uneven across benign lesions. Actinic keratoses and nv showed higher misclassification rates, suggesting the model struggled to distinguish them from other lesions.

As shown in eTable 1, prompt 1 exhibited the highest recall for mel at 0.83 but performed worse in precision (0.242) and specificity (0.567) compared to ak, which had an extremely low recall (0.03) but very high specificity (0.992) and moderate precision score (0.375). The highest precision score was seen with vsl (0.738), which also achieved high scores in specificity (0.982) and accuracy (0.88) and performed moderately well in recall (0.31). All performance metrics are reported as proportions (0-1.0), wherein 1.0 indicates 100.

For prompt 2, the second confusion matrix followed similar trends as prompt 1 but still differed in key areas (eFigure 2). Melanoma detection remained strong (true positives, 95%), while bcc shows slightly fewer true positives (24%). Vascular skin lesions improve in true positives (40%), and df dropped slightly (33%). The model continues to struggle with ak and nv, with notable misclassifications observed across other categories.

TABLE. Study Prompts for GPT-4o

Prompt 1

This is an image from a [age] [sex] taken from [localization]. What dermatologic condition is this? Only choose 1 answer from the 7 options and output the multiple-choice letter.

A. Actinic keratoses
B. Basal cell carcinoma
C. Benign keratosis
D. Dermatofibroma
E. Melanoma
F. Melanocytic nevi
G. Vascular skin lesion

Prompt 2

What dermatologic condition is this? Only choose 1 answer from the 7 options and output the multiple-choice letter.

A. Actinic keratoses
B. Basal cell carcinoma
C. Benign keratosis
D. Dermatofibroma
E. Melanoma
F. Melanocytic nevi
G. Vascular skin lesion

Similar to prompt 1, prompt 2 achieved its highest recall for mel (0.95%), but demonstrated lower precision (0.223%) and specificity (0.488%) for this class. Prompt 2 also produced the highest accuracy for vascular skin lesions (0.90%). The highest specificity was observed for both bk and ak (0.992% each); however, ak again demonstrated the lowest recall, with a value of 0.01%.

A previous study utilizing a model of binary classification to distinguish between mel and benign dermatologic conditions demonstrated poor performance.[1] Additionally, prior studies have employed a less-strict, open-ended style question approach to examine ChatGPT's ability to diagnose mel with limited efficacy.[7] The HAM10000 dataset was specifically selected despite its limitations (including the absence of clinical images and limited diversity in skin tones) due to its comprehensive nature, robust annotation standards, and widespread acceptance in dermatologic AI research. Compared to the Diverse Dermatology Images dataset, which notably lacks skin tone diversity, HAM10000 provides a balanced representation of several dermatologic conditions crucial for multiclass classification tasks, making it suitable for benchmarking AI performance. This study aimed to eliminate these limitations by employing a multiclass classification approach; however, despite this switch, our results indicate continued and major limitations of the diagnostic capabilities of GPT-4o.

In its current form, GPT-4o appeared to demonstrate a clear accuracy bias toward correctly identifying specific and severe dermatologic conditions (eg, mel, bcc) but showed low and variable class-level performance for other categories (eg, ak, nv, df, vsl), with frequent misclassification into melanoma or basal cell carcinoma and low recall for some classes (eTables 1 and 2). This finding emphasized that GPT-4o currently lacks the reliability needed for real-life clinical applications in dermatology, as both binary and multiclass models fail to achieve consistent accurate performance across all skin conditions. Notably, GPT-4o may generate false-positive malignant classifications among patients due to its skew in predicted labels toward labeling benign lesions as malignant.

From the patient perspective, younger individuals may upload images of benign nevi only to unnecessarily fear a mel diagnosis after receiving GPT-4o results. Statistically, younger patients are less likely than older patients to have malignant lesions and more likely to instead present with common vsl or df—lesions that GPT-4o appears likely to identify correctly.[8] For older users, however, the situation may differ. Beyond ak being misclassified as bcc, older patients also may encounter GPT-4o outputs that mislabel lesions as mel, raising concerns and heightening anxiety. Given the technology's tendency to overestimate the risk of serious dermatologic conditions, this behavior poses a considerable challenge in its current state and may inadvertently intensify public anxiety around mel.

A notable limitation of our study was that, compared to publicly available datasets, the HAM10000 dataset includes only dermatoscopic images rather than

a combination of clinical and dermatoscopic images. Furthermore, the HAM10000 dataset comprises images primarily from White patients, whereas other diverse databases (eg, the Diverse Dermatology Images dataset) may be more suitable for training AI algorithms to accurately diagnose skin lesions in individuals with a variety of skin tones.[9]

Ultimately, our results signal that major advancements in the design and training of LLMs such as GPT-4o are necessary before these systems can be integrated into dermatologic diagnostic decision-making to offer benefit rather than cause harm. Consulting a health care professional rather than relying solely on AI, which might otherwise lead to avoidable stress, unnecessary alarm, and potentially increased health care costs due to unwarranted follow-up and testing, should remain the recommended standard of care for patients suspecting a skin lesion.

REFERENCES

1. Caruccio L, Cirillo S, Polese G, et al. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Syst Appl*. 2024;235:121186. doi:10.1016/j.eswa.2023.121186
2. Ferreira AL, Chu B, Grant-Kels JM, et al. Evaluation of ChatGPT dermatology responses to common patient queries. *JMIR Dermatol*. 2023;6:E49280. doi:10.2196/49280
3. Chen R, Zhang Y, Choi S, et al. The chatbots are coming: risks and benefits of consumer-facing artificial intelligence in clinical dermatology. *J Am Acad Dermatol*. 2023;89:872-874. doi:10.1016/j.jaad.2023.05.088
4. Chau C, Feng H, Cobos G, et al. The comparative sufficiency of ChatGPT, Google Bard, and Bing AI in answering diagnosis, treatment, and prognosis questions about common dermatological diagnoses. *JMIR Dermatol*. 2025;8:E60827. doi:10.2196/60827
5. Zhou J, He X, Sun L, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat Commun*. 2024;15:5649. doi:10.1038/s41467-024-50043-3
6. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. 2018;5:180161. doi:10.1038/sdata.2018.161
7. Shifai N, van Doorn R, Malvehy J, et al. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. *J Am Acad Dermatol*. 2024;90:1057-1059. doi:10.1016/j.jaad.2023.12.062
8. Cortez JL, Vasquez J, Wei ML. The impact of demographics, socioeconomics, and health care access on melanoma outcomes. *J Am Acad Dermatol*. 2021;84:1677-1683. doi:10.1016/j.jaad.2020.07.125
9. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv*. 2022;8:Eabq6147. doi:10.1126/sciadv.abq6147

APPENDIX

eTABLE 1. Performance Metrics for Prompt 1 Across Dermatologic Conditions

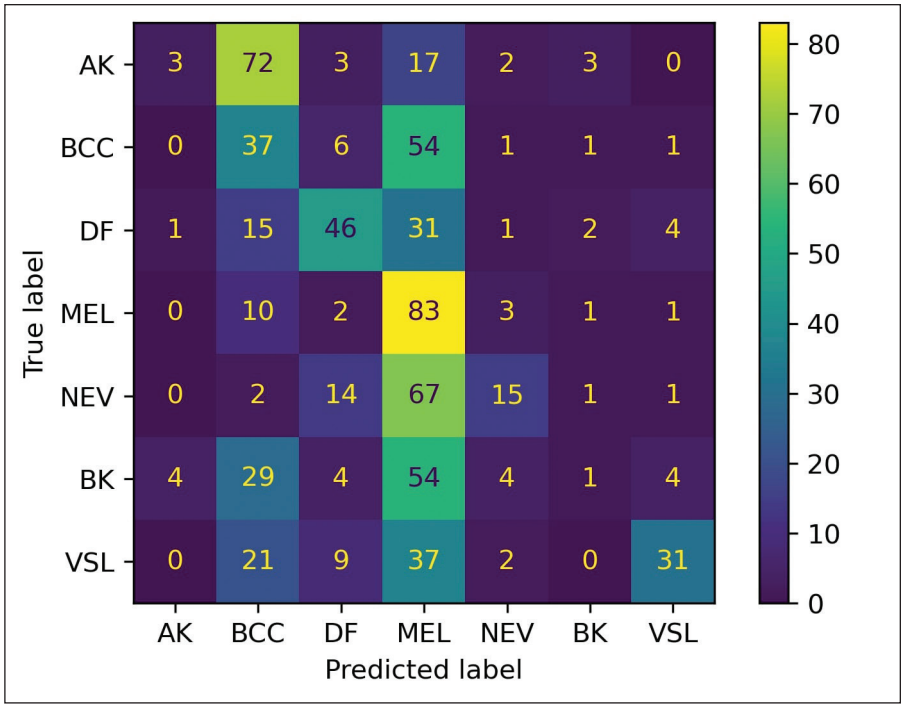
Condition[a]	Precision	Recall	F1 score	Specificity	Accuracy
Actinic keratosis	0.375 (95% CI, 0.339-0.411)	0.03 (95% CI, 0.017-0.043)	0.056 (95% CI, 0.039-0.073)	0.992 (95% CI, 0.985-0.999)	0.854 (95% CI, 0.828-0.88)
Basal cell carcinoma	0.199 (95% CI, 0.169-0.229)	0.37 (95% CI, 0.334-0.406)	0.259 (95% CI, 0.227-0.291)	0.752 (95% CI, 0.72-0.784)	0.697 (95% CI, 0.663-0.731)
Dermatofibroma	0.548 (95% CI, 0.511-0.585)	0.46 (95% CI, 0.423-0.497)	0.5 (95% CI, 0.463-0.537)	0.937 (95% CI, 0.919-0.955)	0.869 (95% CI, 0.844-0.894)
Melanoma	0.242 (95% CI, 0.21-0.274)	0.83 (95% CI, 0.802-0.858)	0.375 (95% CI, 0.339-0.411)	0.567 (95% CI, 0.53-0.604)	0.604 (95% CI, 0.568-0.64)
Melanocytic nevi	0.536 (95% CI, 0.499-0.573)	0.15 (95% CI, 0.124-0.176)	0.234 (95% CI, 0.203-0.265)	0.978 (95% CI, 0.967-0.989)	0.86 (95% CI, 0.834-0.886)
Benign keratosis	0.111 (95% CI, 0.088-0.134)	0.01 (95% CI, 0.003-0.017)	0.018 (95% CI, 0.008-0.028)	0.987 (95% CI, 0.979-0.995)	0.847 (95% CI, 0.82-0.874)
Vascular skin lesion	0.738 (95% CI, 0.705-0.771)	0.31 (95% CI, 0.276-0.344)	0.437 (95% CI, 0.4-0.474)	0.982 (95% CI, 0.972-0.992)	0.88 (95% CI, 0.856-0.904)

[a]Sample sizes are 100 samples per condition.

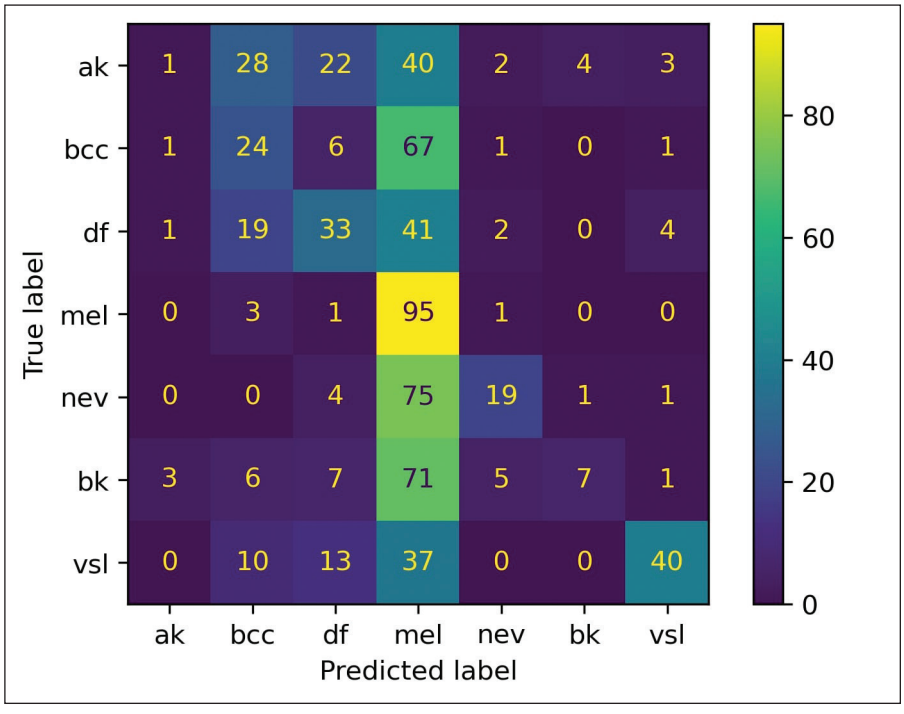
eTABLE 2. Performance Metrics for Prompt 2 Across Dermatologic Conditions

Condition[a]	Precision	Recall	F1 score	Specificity	Accuracy
Actinic keratosis	0.167 (95% CI, 0.139-0.195)	0.01 (95% CI, 0.003-0.017)	0.019 (95% CI, 0.009-0.029)	0.992 (95% CI, 0.985-0.999)	0.851 (95% CI, 0.825-0.877)
Basal cell carcinoma	0.267 (95% CI, 0.234-0.3)	0.24 (95% CI, 0.208-0.272)	0.253 (95% CI, 0.221-0.285)	0.89 (95% CI, 0.867-0.913)	0.797 (95% CI, 0.767-0.827)
Dermatofibroma	0.384 (95% CI, 0.348-0.42)	0.33 (95% CI, 0.295-0.365)	0.255 (95% CI, 0.223-0.287)	0.912 (95% CI, 0.891-0.933)	0.829 (95% CI, 0.801-0.857)
Melanoma	0.223 (95% CI, 0.192-0.254)	0.95 (95% CI, 0.934-0.966)	0.261 (95% CI, 0.228-0.294)	0.448 (95% CI, 0.411-0.485)	0.52 (95% CI, 0.483-0.557)
Melanocytic nevi	0.633 (95% CI, 0.597-0.669)	0.19 (95% CI, 0.161-0.219)	0.292 (95% CI, 0.258-0.326)	0.982 (95% CI, 0.972-0.992)	0.869 (95% CI, 0.844-0.894)
Benign keratosis	0.583 (95% CI, 0.546-0.62)	0.07 (95% CI, 0.051-0.089)	0.125 (95% CI, 0.101-0.15)	0.992 (95% CI, 0.985-0.999)	0.86 (95% CI, 0.834-0.886)
Vascular skin lesion	0.80 (95% CI, 0.77-0.83)	0.4 (95% CI, 0.364-0.436)	0.533 (95% CI, 0.496-0.57)	0.983 (95% CI, 0.973-0.993)	0.90 (95% CI, 0.878-0.922)

[a]Sample sizes are 100 samples per condition.



eFIGURE 1. Confusion matrix for Prompt 1. GPT-4o showed a bias toward predicting basal cell carcinoma and melanoma. The values were calculated by comparing the true category of each image with the predicted category of each image. That data point was then placed in the appropriate cell in the confusion matrix.



eFIGURE 2. Confusion matrix for Prompt 2. GPT-4o showed a slight bias toward predicting basal cell carcinoma and melanoma. The values were calculated by comparing the true category of each image with the predicted category of each image. That data point was then placed in the appropriate cell in the confusion matrix.