Boosting GPT-4V's accuracy in dermoscopic classification with few-shot learning. Comment on "can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study"

To the Editor: We read with great interest the publication by Shifai et al,[1] which evaluates the baseline performance of ChatGPT-4 Vision (hereafter denoted as GPT-4V) in diagnosing melanoma from dermoscopic images. The reported low accuracies underscore the need for further refinement before GPT-4V can be effectively integrated into clinical settings.

The efficacy of responses from large language model chatbots like GPT-4V is heavily influenced by the design of the input instructions, a practice known as prompt engineering. A notable method within this scope is few-shot learning,[2] which involves incorporating examples of similar tasks directly into the prompt. This approach has significantly improved GPT-4V's accuracy in classification tasks involving histopathologic images.[3] However, it remains to be seen whether few-shot learning can similarly enhance GPT-4V's performance in analyzing dermoscopic images, which primarily feature macroscopic pigmentation patterns and color distributions within lesions.

To explore this, we revisited the same dermoscopic dataset used by Shifai et al,[1] comprising 100 histopathology-verified cases evenly divided between melanomas and benign nevi. For each query image to be classified, we implemented few-shot learning by randomly selecting k images ($k = 0$, 1, 2, 3, 4, and 5) from the remaining images in each category. We prompted GPT-4V through Application Programming Interface (version "GPT-4-Turbo-2024-04-09") to identify the most similar image among the selected examples and apply the corresponding label for classification (Fig 1, *A*). Initial results without few-shot examples indicated that GPT-4V's performance was marginally better than random chance (51.3 ± 1.3%) (Fig 1, *B*; dash line). In contrast, the integration of few-shot learning increased the accuracy to 71.1 ± 1.5% with just one example per category, and further to 75% to 77% with 2 or more examples (Fig 1, *B*; red bars). Extending to 100 additional testing images randomly chosen from the International Skin Imaging Collaboration archive (https://www.isic-archive.com) yielded comparable results. As a reference, 81% to 83% are the accuracies reported from a recently proposed handcrafted algorithm and a convolutional neural network model (ResNet-50) on the same topic.[4]

Our further exploration into a targeted approach using k-nearest neighbor to select examples from each category[3] did not improve the accuracy (Fig 1, *B*; blue bars). Additional analysis of GPT-4V's explanations revealed that GPT-4V frequently employs the ABCD rule (Asymmetry, Border, Color, Diameter) to aid classification. Although GPT-4V exhibits limitations in color perception,[5] its inclusion of color to make a prediction is evident as converting images to grayscale substantially decreased the accuracy (Fig 1, *B*; orange bars).

The primary limitations of our study include the binary classification approach and the small sample size inherited from Shifai et al,[1] which may affect the generalizability of our findings. Despite these limitations, our results suggest that few-shot learning can significantly enhance GPT-4V's capabilities in dermoscopic classification, even with minimal training examples. As GPT and prompt engineering continue to evolve, we are optimistic about their potential in advancing artificial intelligence-aided diagnosis. This is particularly pertinent for rare melanomas, which typically receive less representation in the training datasets of traditional deep learning models.

Supporting prompts and data are available through Mendeley (Supplementary Tables I and II, available via Mendeley at http://doi.org/10.17632/g6bkyvmn3s.3).

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the authors used ChatGPT-4 to improve readability, clarity, and professionalism in writing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Jinge Wang, PhD, and Gangqing Hu, PhD

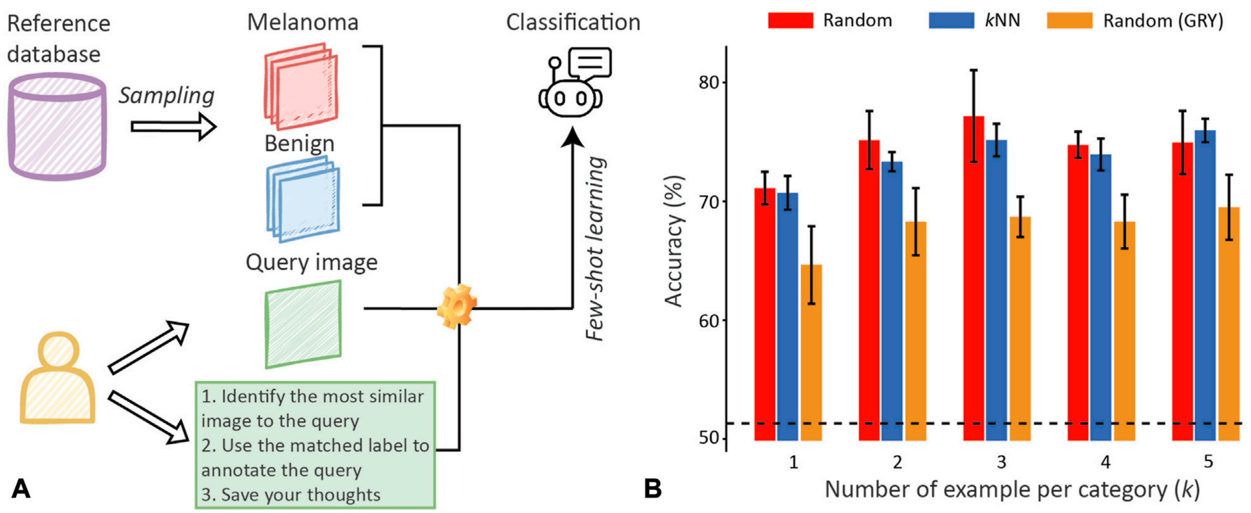
From the Department of Microbiology, Immunology & Cell Biology, West Virginia University, Morgantown, West Virginia.

Funding sources: This work was supported by NIH-NIGMS grants P20 GM103434 and U54 GM-104942 as well as NSF 2125872.

Patient consent: Not applicable.

IRB approval status: Not applicable.

Key words: ChatGPT; few-shot learning; GPT-4V; large language models; melanoma; nevi; skin cancer.



1. Identify the most similar image to the query
2. Use the matched label to annotate the query
3. Save your thoughts

Fig 1. Dermoscopic classification of melanomas versus benign nevi by GPT-4V using few-shot learning. **A,** Schematic representation of GPT-4V employing few-shot learning for dermoscopic classification. **B,** Prediction accuracy of GPT-4V as a function of the number of few-shot examples, calculated under 3 conditions: Random, where examples are randomly selected; *k*NN, where examples are chosen through *k*-nearest neighbor search; and Random (GRY), where all images are converted to grayscale including both query image and randomly chosen examples. Error bar: standard deviation. *Dashed line*: average accuracy for zero-shot (no example image). All experiments are in 5 replicates.

Correspondence to: Gangqing Hu, PhD, Department of Microbiology, Immunology & Cell Biology, West Virginia University, 64 Medical Center Dr, Morgantown, WV 26506-9177

E-mail: Michael.Hu@hsc.wvu.edu

X handle: @hugangqing

Conflicts of interest

None disclosed.

REFERENCES

1. Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. *J Am Acad Dermatol*. 2024;90(5):1057-1059. https://doi.org/10.1016/j.jaad.2023.12.062
2. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-1901.
3. Ferber D, Wölflein G, Wiest IC, et al. In-context learning enables multimodal large language models to classify cancer pathology images, 2024. Accessed March 25, 2024. https://ui.adsabs.harvard.edu/abs/2024arXiv240307407F
4. Monnier J, Gouabou Foahom AC, Serdi M, et al. Automated melanoma detection. An algorithm inspired from human intelligence characterizing disordered pattern of melanocytic lesions improving a convolutional neural network. *J Am Acad Dermatol*. 2024;91:350-352. https://doi.org/10.1016/j.jaad.2024.02.063
5. Wang J, Ye Q, Liu L, Guo NL, Hu G. Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception. *NPJ Precis Oncol*. 2024;8(1):84. https://doi.org/10.1038/s41698-024-00576-z

https://doi.org/10.1016/j.jaad.2024.06.098