

Research Letter

Assessing the Diagnostic Accuracy of ChatGPT-4 in Identifying Diverse Skin Lesions Against Squamous and Basal Cell Carcinoma

Nitin Chetla^{1*}, BS; Matthew Chen^{2*}, BS; Joseph Chang³, BS; Aaron Smith¹, BS; Tamer Rajai Hage⁴, BS; Romil Patel¹, BS; Alana Gardner⁵; Bridget Bryer⁶, MD

¹University of Virginia School of Medicine, University of Virginia, Charlottesville, VA, United States

²Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, United States

³University of Passau, Passau, Germany

⁴Virginia Tech, Blacksburg, VA, United States

⁵University at Albany, State University of New York, Albany, NY, United States

⁶Department of Dermatology, University of Virginia, Charlottesville, VA, United States

*these authors contributed equally

Corresponding Author:

Nitin Chetla, BS

University of Virginia School of Medicine

University of Virginia

828 Cabell Avenue

Charlottesville, VA, 22903

United States

Phone: 1 571-581-0562

Email: nc8qh@virginia.edu

Abstract

Our study evaluates the diagnostic accuracy of ChatGPT-4o in classifying various skin lesions, highlighting its limitations in distinguishing squamous cell carcinoma from basal cell carcinoma using dermatoscopic images.

JMIR Dermatol 2025;8:e67299; doi: [10.2196/67299](https://doi.org/10.2196/67299)

Keywords: chatbot; ChatGPT; ChatGPT-4; squamous cell carcinoma; basal cell carcinoma; skin cancer; skin cancer detection; dermatoscopic image analysis; skin lesion differentiation; dermatologist; machine learning; ML; artificial intelligence; AI; AI in dermatology; algorithm; model; analytics; diagnostic accuracy

Introduction

Squamous cell carcinoma (SCC) and basal cell carcinoma (BCC) are prevalent skin cancers that can cause significant local tissue damage and disfigurement as well as mortality in cases of aggressive SCCs [1,2]. With the rising incidence, early and accurate diagnosis is essential for appropriate treatment [3]. Differentiating SCC and BCC from other common skin lesions, such as actinic keratoses (AK), benign keratoses (BK), and melanocytic nevi, can be challenging [4]. As artificial intelligence (AI) becomes increasingly integrated into clinical practice, concerns arise about its ability to provide accurate diagnostic assessments, given AI's growing accessibility [5,6]. We assessed the ability of ChatGPT to distinguish images of SCC and BCC from other lesions.

Methods

OpenAI's application programming interface was used to query ChatGPT-4 Omni (ChatGPT-4O) for assessing the performance in classifying 200 dermatoscopic images each of SCC, BCC, BK, melanocytic nevi, and 150 images of AK from the HAM10K database [7]. Images were verified using histopathology (>50%), follow-up examination, expert consensus, or in-vivo confocal microscopy. Two standardized prompts were used:

Prompt 1

This is an image on the Step 1 examination, and the multiple-choice question is as follows: Based on the image, does the patient have (A) Nevus, (B) Actinic Keratosis (AK), (C)

Benign Keratosis (BK), or (D) BCC, or (E) SCC. Only output (A), (B), (C), (D) or (E).

Prompt 2

This is an image from a patient. Based on the image, does the patient have (A) Nevus, (B) AK, (C) BK, (D) BCC, or (E) SCC. Only output (A), (B), (C), or (D) or (E).

The key metrics calculated include accuracy, sensitivity, and specificity. Images that ChatGPT refused to answer were excluded from calculations. The exclusion criterion for this study was any dermatoscopic image that ChatGPT refused to classify. These images were not included in the calculations of accuracy, sensitivity, and specificity.

The study did not employ further prompt engineering to enhance ChatGPT's performance because the goal was to evaluate its diagnostic accuracy using straightforward, unrefined prompts that reflect real-world scenarios. This ensures that the findings are applicable to patient or clinician

usage. Additionally, the use of simple prompts highlights the model's sensitivity to language variations, underscoring the unpredictability and variability of these AI systems.

Results

For Prompt 1, ChatGPT classified nevi with an accuracy of 79.3% (95% CI 76.7%-81.9%), sensitivity of 0.844, and specificity of 0.758. The accuracy for classifying BCC was 77.8% (95% CI 75.2%-80.4%), with low sensitivity (0.081) and high specificity (0.959). The accuracy for classifying SCC was 66.1% (95% CI 52.8%-59.2%), with sensitivity of 0.477 and specificity of 0.711 (Table 1).

In Prompt 2, SCC accuracy increased to 72.8% (95% CI: 70.0%-75.6%) but sensitivity dropped to 0.245. Nevi accuracy slightly declined to 72.8%, while SCC specificity improved to 0.857 (Table 2).

Table 1. Accuracy, sensitivity, and specificity of ChatGPT for lesion differentiation using Prompt 1.

Class	Sample size	Accuracy (95% CI)	Sensitivity	Specificity	F1 score
Actinic keratosis	149	73.0% (70.2-75.8)	0.356	0.802	0.294
Basal cell carcinoma	198	77.8% (75.2-80.4)	0.081	0.959	0.132
Nevus	199	79.3% (76.7-81.9)	0.844	0.758	0.649
Benign keratosis	200	74.4% (71.6-77.2)	0.090	0.939	0.138
Squamous cell carcinoma	199	66.1% (52.8-59.2)	0.477	0.711	0.373

Table 2. Accuracy, sensitivity, and specificity of ChatGPT for lesion differentiation using Prompt 2.

Class	Sample size	Accuracy (95% CI)	Sensitivity	Specificity	F1 score
Actinic keratosis	149	72.9% (70.1-75.7)	0.423	0.774	0.329
Basal cell carcinoma	200	79.5% (76.9-82.1)	0.07	0.987	0.125
Nevus	200	72.8% (70.0-75.6)	0.89	0.664	0.58
Benign keratosis	200	73.7% (70.9-76.5)	0.18	0.885	0.223
Squamous cell carcinoma	200	72.8% (70.0-75.6)	0.245	0.857	0.275

Discussion

ChatGPT-4o struggled to differentiate between SCC and BCC. Nevus classification was the most accurate, with high F1 scores and minimal false-positive results, demonstrating proficiency in identifying less ambiguous lesions. The model showed significant bias in SCC classification, frequently misclassifying SCC as BCC with a high rate of false-positive results. This aligns with previous research that observed SCC is often mistaken for BCC, particularly when features like pigmentation or rolled borders overlap [8]. ChatGPT's performance worsened in Prompt 2, where SCC was frequently misclassified as AK. Previous authors noted that AI performs comparably to dermatologists in binary choices, but our study further highlights the struggle AI faces in multiclass differentiation [9].

Prompt 1 was designed to emulate a standardized examination scenario, leveraging ChatGPT's ability to respond to structured, multiple-choice questions within a

controlled academic framework. This approach was necessary as ChatGPT restricts responses to direct health-related inquiries, necessitating creative prompt construction to elicit diagnostic outputs. In contrast, Prompt 2 adopted a more generic phrasing reflective of a patient inquiry to evaluate how conversational language might influence diagnostic accuracy. This design choice was informed by the observation that variations in prompt language can significantly impact AI-generated outputs.

Limitations include using a single dataset, which may not represent the diversity of skin lesions in clinical settings and not consider variations in image quality. Future improvements should focus on expanding training data diversity and improving image scenario handling to enhance diagnostic accuracy. We concur with Labkoff et al that precautions such as training clinicians on the limitations of AI systems and implementing standardized protocols to validate AI-generated diagnoses before acting on them would help ensure safe and effective integration into clinical workflows [10].

Conflicts of Interest

None declared.

References

1. Peris K, Fagnoli MC, Garbe C, et al. Diagnosis and treatment of basal cell carcinoma: European consensus-based interdisciplinary guidelines. *Eur J Cancer*. Sep 2019;118:10-34. [doi: [10.1016/j.ejca.2019.06.003](https://doi.org/10.1016/j.ejca.2019.06.003)] [Medline: [31288208](https://pubmed.ncbi.nlm.nih.gov/31288208/)]
2. Schmults CD, Blitzblau R, Aasi SZ, et al. NCCN Guidelines® insights: squamous cell skin cancer, version 1.2022. *J Natl Compr Canc Netw*. Dec 2021;19(12):1382-1394. [doi: [10.6004/jnccn.2021.0059](https://doi.org/10.6004/jnccn.2021.0059)] [Medline: [34902824](https://pubmed.ncbi.nlm.nih.gov/34902824/)]
3. Urban K, Mehrmal S, Uppal P, Giesey RL, Delost GR. The global burden of skin cancer: a longitudinal analysis from the Global Burden of Disease Study, 1990-2017. *JAAD Int*. Mar 2021;2:98-108. [doi: [10.1016/j.jdin.2020.10.013](https://doi.org/10.1016/j.jdin.2020.10.013)] [Medline: [34409358](https://pubmed.ncbi.nlm.nih.gov/34409358/)]
4. Ahnlide I. Aspects of skin cancer diagnosis in clinical practice. Lund University; 2015. URL: <https://lucris.lub.lu.se/ws/portalfiles/portal/3030914/8167764.pdf> [Accessed 2025-01-07] [Medline: [26854159](https://pubmed.ncbi.nlm.nih.gov/26854159/)]
5. O'Hern K, Yang E, Vidal NY. ChatGPT underperforms in triaging appropriate use of Mohs surgery for cutaneous neoplasms. *JAAD Int*. Sep 2023;12:168-170. [doi: [10.1016/j.jdin.2023.06.002](https://doi.org/10.1016/j.jdin.2023.06.002)] [Medline: [37404248](https://pubmed.ncbi.nlm.nih.gov/37404248/)]
6. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv*. Aug 12, 2022;8(32):eabq6147. [doi: [10.1126/sciadv.abq6147](https://doi.org/10.1126/sciadv.abq6147)] [Medline: [35960806](https://pubmed.ncbi.nlm.nih.gov/35960806/)]
7. Scarlet A. Melanoma dataset. URL: <https://www.kaggle.com/datasets/drscarlat/melanoma> [Accessed 2025-01-07]
8. Ryu TH, Kye H, Choi JE, Ahn HH, Kye YC, Seo SH. Features causing confusion between basal cell carcinoma and squamous cell carcinoma in clinical diagnosis. *Ann Dermatol*. Feb 2018;30(1):64-70. [doi: [10.5021/ad.2018.30.1.64](https://doi.org/10.5021/ad.2018.30.1.64)] [Medline: [29386834](https://pubmed.ncbi.nlm.nih.gov/29386834/)]
9. Escalé-Besa A, Vidal-Alaball J, Miró Catalina Q, Gracia VH, Marin-Gomez FX, Fuster-Casanovas A. The use of artificial intelligence for skin disease diagnosis in primary care settings: a systematic review. *Healthcare (Basel)*. Jun 13, 2024;12(12):1192. [doi: [10.3390/healthcare12121192](https://doi.org/10.3390/healthcare12121192)] [Medline: [38921305](https://pubmed.ncbi.nlm.nih.gov/38921305/)]
10. Labkoff S, Oladimeji B, Kannry J, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. *J Am Med Inform Assoc*. Nov 1, 2024;31(11):2730-2739. [doi: [10.1093/jamia/ocae209](https://doi.org/10.1093/jamia/ocae209)] [Medline: [39325508](https://pubmed.ncbi.nlm.nih.gov/39325508/)]

Abbreviations

AI: artificial intelligence
AK: actinic keratoses
BCC: basal cell carcinoma
BK: benign keratoses
SCC: squamous cell carcinoma

Edited by Robert Dellavalle; peer-reviewed by Chenxu Wang, Shaniko Kaleci; submitted 10.10.2024; final revised version received 15.01.2025; accepted 16.01.2025; published 21.03.2025

Please cite as:

Chetla N, Chen M, Chang J, Smith A, Hage TR, Patel R, Gardner A, Bryer B
Assessing the Diagnostic Accuracy of ChatGPT-4 in Identifying Diverse Skin Lesions Against Squamous and Basal Cell Carcinoma
JMIR Dermatol 2025;8:e67299
URL: <https://derma.jmir.org/2025/1/e67299>
doi: [10.2196/67299](https://doi.org/10.2196/67299)

© Nitin Chetla, Matthew Chen, Joseph Chang, Aaron Smith, Tamer Rajai Hage, Romil Patel, Alana Gardner, Bridget Bryer. Originally published in *JMIR Dermatology* (<http://derma.jmir.org>), 21.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Dermatology*, is properly cited. The complete bibliographic information, a link to the original publication on <http://derma.jmir.org>, as well as this copyright and license information must be included.