ASSESSING THE DIAGNOSTIC ACCURACY OF CHATGPT-4 IN IDENTIFYING DIVERSE SKIN LESIONS AGAINST SQUAMOUS AND BASAL CELL CARCINOMA

Authors: Nitin Chetla BS[1], Matthew Chen BS[1], Joseph Chang BS[1], Aaron D. Smith BS1, Dr. Bridget Bryer MD[1]

[1]University of Virginia Dermatology, 1340 Jefferson Park Ave, Charlottesville, VA 22903

INTRODUCTION

- Squamous cell carcinoma (SCC) and basal cell carcinoma (BCC) are prevalent skin cancers that can cause significant local tissue damage and disfigurement if not properly treated.
- With rising incidence, early and accurate diagnosis is essential for appropriate treatment.[1] Differentiating SCC and BCC from other common skin lesions, such as actinic keratoses (AK), benign keratoses (BK), and melanocytic nevi, can be challenging. As artificial intelligence (AI) becomes increasingly integrated into clinical practice, concerns arise about its ability to provide accurate diagnostic assessments.[2]
- These concerns are particularly relevant given AI's growing accessibility, which may lead to suboptimal care in the absence of validated diagnoses. We assess the ability of ChatGPT to distinguish images of SCC and BCC from other lesions.

METHODS

OpenAI's application programming interface (API) was used to query ChatGPT-4 Omni (ChatGPT-4O) to assess its performance in classifying 200 dermatoscopic images each of SCC, BCC, BK, and melanocytic nevi, and 150 images of AK from the HAM10K database.[3]

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, the multiple-choice question is as follows. Based on the image, does the patient have A) Nevus, B) Actinic Keratoses (AKs), C) Benign Keratosis (BKs), 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) AKs, C) BKs, D) BCC, or E) SCC. Only output A), B), C), or D) or E)

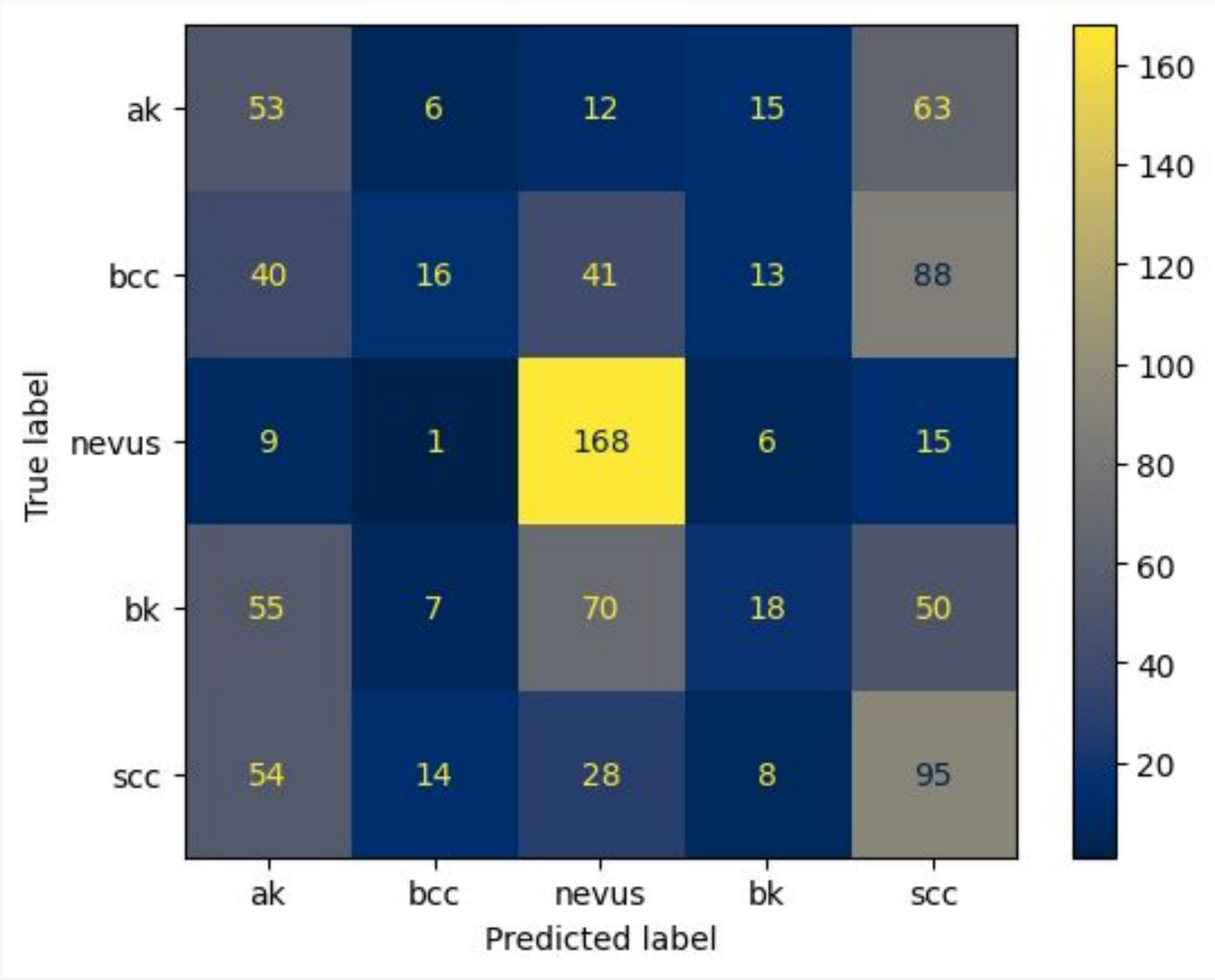
Key metrics calculated include accuracy, sensitivity, and specificity. Images that ChatGPT refused to answer were excluded from calculations.

RESULTS

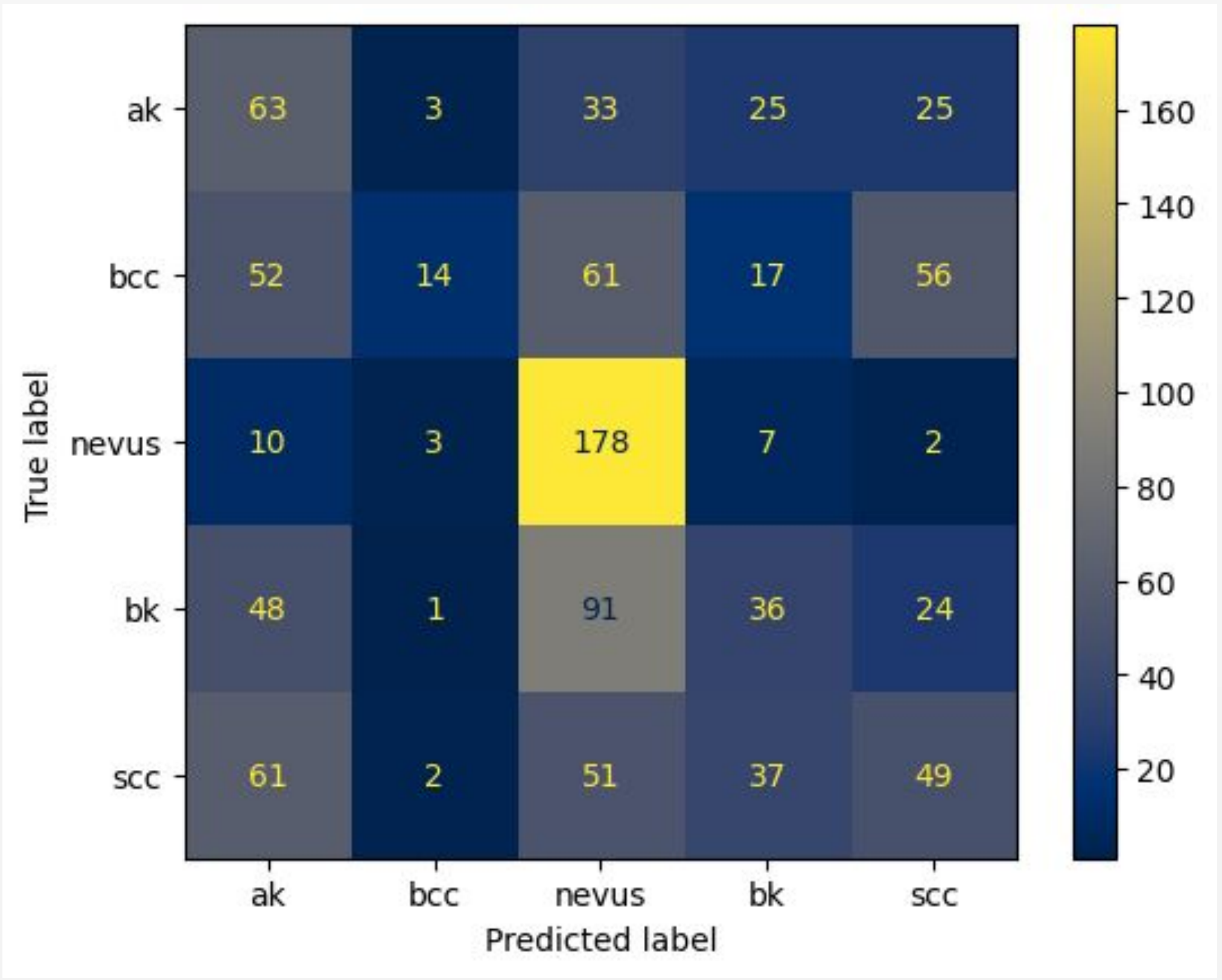
FOR PROMPT 1, ChatGPT classified nevi with 79.3% accuracy (95% CI: 76.7%-81.9%), sensitivity 0.844, and specificity 0.758. BCC had 77.8% accuracy (95% CI: 75.2%-80.4%), low sensitivity (0.081), and high specificity (0.959). SCC accuracy was 66.1% (95% CI: 52.8%-59.2%), with sensitivity 0.477 and specificity 0.711.

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.

SUPPLEMENTARY MATERIAL I



Supplementary Figure I: Confusion Matrix of Prompt 1



Supplementary Figure II: Confusion Matrix of Prompt 2

DISCUSSION

- **Nevus Classification:** ChatGPT-4 excelled in accurately identifying nevi, demonstrating a high level of precision and minimal false positive rates.
- **SCC Classification:** The model encountered difficulties in distinguishing between squamous cell carcinoma (SCC) and basal cell carcinoma (BCC), particularly when presented with overlapping features like pigmentation or rolled borders. This aligns with previous research by Ryu et al. (2018), who observed similar challenges in AI-based skin cancer diagnosis.
- **Prompt 2:** The model's performance further declined in Prompt 2, frequently misclassifying SCC as actinic keratosis (AK). This finding is consistent with Escalé-Besa et al. (2024), who noted that AI models may struggle with multi-class differentiation tasks.
- **Limitations and Future Directions:** The limitations of this study include the use of a single dataset, which may not fully represent the diverse range of skin lesions observed in clinical practice. To improve the model's accuracy, future research should focus on expanding the training dataset to include a wider variety of images and exploring techniques to address variations in image quality and lighting conditions.

Class	Sample Size	Accuracy (95% CI)	Sensitivity	Specificity	F1 Score
AK	149	73.0% (70.2-75.8)	0.356	0.802	0.294
BCC	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
BK	200	74.4% (71.6-77.2)	0.090	0.939	0.138
SCC	199	66.1% (52.8-59.2)	0.477	0.711	0.373

Table 1: Accuracy, Sensitivity, and Specificity of ChatGPT lesion differentiation, Prompt 1

Class	Sample Size	Accuracy (95% CI)	Sensitivity	Specificity	F1 Score
AK	149	72.9% (70.1-75.7)	0.423	0.774	0.329
BCC	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
BK	200	73.7% (70.9-76.5)	0.18	0.885	0.223
SCC	200	72.8% (70.0-75.6)	0.245	0.857	0.275

Table 2: Accuracy, Sensitivity, and Specificity of ChatGPT lesion differentiation, Prompt 2

REFERENCES

1. 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. 2021 Jan 4;2:98-108.
2. O'Hern K, Yang E, Vidal NY. ChatGPT underperforms in triaging appropriate use of Mohs surgery for cutaneous neoplasms. JAAD Int. 2023 Jun 9;12:168-170.
3. HAM10k images dataset, Human Against Machine with 10,000 Training Images. Accessed via Kaggle. https://www.kaggle.com/datasets/drscarlat/melanoma
4. 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. 2018 Feb;30(1):64-70.
5. Escalé-Besa A, Vidal-Alaball J, Miró Catalina Q, Gracia VHG, Marin-Gomez FX, Fuster-Casanovas A. The Use of Artificial Intelligence for Skin Disease Diagnosis in Primary Care Settings: A Systematic Review. *Healthcare*. 2024; 12(12):1192.