Comparing the Performance of Claude-3 Opus and ChatGPT-4 in Dermoscopic images Analysis and Melanoma Diagnosis: Exploring the Application Potential of Different Large Language Models in Dermatology

Xu Liu, Chaoli Duan, Min-kyu Kim, Eunjin Jee, Beenu Maharjan, Dan Du, Yuwei Huang, Lu Zhang, Xian Jiang

Submitted to: JMIR Medical Informatics
on: April 08, 2024

Table of Contents

Comparing the Performance of Claude-3 Opus and ChatGPT-4 in Dermoscopic images Analysis and Melanoma Diagnosis: Exploring the Application Potential of Different Large Language Models in Dermatology

Xu Liu[1*] BS; Chaoli Duan[1*] MS; Min-kyu Kim[2] MS; Eunjin Jee[2] MS; Beenu Maharjan[3] MS; Dan Du[1] MS; Yuwei Huang[1] MD; Lu Zhang[1] MD; Xian Jiang[1*] MD

[1]Department of Dermatology, West China Hospital, Chengdu, People's Republic of China Sichuan University chengdu CN
[2]Department of Dermatology, West China Hospital, Chengdu, People's Republic of China Sichuan University chengdu KR
[3]Department of Dermatology, West China Hospital, Chengdu, People's Republic of China Sichuan University chengdu NP
[*]these authors contributed equally

Corresponding Author:
Xu Liu BS
Department of Dermatology, West China Hospital, Chengdu, People's Republic of China
Sichuan University
Chengdu 610041, China
chengdu
CN

Abstract

Background: Recent advancements in artificial intelligence (AI) and large language models (LLMs) have shown promising potential in various medical fields, including dermatology. LLMs, such as ChatGPT, have demonstrated their ability to generate human-like responses to text-based prompts and assist in clinical decision-making. With the introduction of image analysis capabilities in LLMs, such as ChatGPT Vision , the application of these models in dermatological diagnostics has garnered significant interest.However, the emergence of other LLMs, such as Claude 3 Opus, warrants investigation. Claude 3 Opus is an advanced conversational AI model that has shown promising performance in various natural language processing tasks. Its ability to engage in context-aware dialogues and provide coherent responses makes it a potential candidate for assisting in clinical decision-making, including dermatological diagnostics.

Objective: We compared the diagnostic performance of Claude 3 Opus and ChatGPT-4 to provide insights into their strengths and weaknesses and guide the selection and optimization of AI-assisted diagnostic tools in dermatology.

Methods: We randomly selected 100 histopathology-confirmed dermoscopic images (50 malignant, 50 benign) from the International Skin Imaging Collaboration (ISIC) Archive database. Each model was prompted to provide the top 3 differential diagnoses for each image, ranked by likelihood. The models' responses were recorded for further analysis. We assessed primary diagnosis accuracy, top 3 differential diagnoses accuracy, and malignancy discrimination ability.

Results: McNemar's test determined statistical significance (?=0.05). For primary diagnosis accuracy, Claude 3 Opus achieved 54.90% sensitivity, 57.14% specificity, and 56.00% accuracy, while GPT4-Vision demonstrated 56.86% sensitivity, 38.78% specificity, and 48.00% accuracy (p=0.170). For top 3 differential diagnoses accuracy, Claude 3 Opus and ChatGPT-4 included the correct diagnosis in 76.00% and 78.00% of cases, respectively (p=0.564). For malignancy discrimination, Claude 3 Opus outperformed ChatGPT-4 with 47.06% sensitivity, 81.63% specificity, and 64.00% accuracy compared to 45.10%, 42.86%, and 44.00%, respectively (p=0.001). Further quantifying the difference in malignancy discrimination ability, we calculated odds ratios (ORs) and 95% confidence intervals (CIs). Claude 3 Opus had an OR of 3.951 (95% CI: 1.685-9.263), indicating a stronger association between its predictions and actual malignancy compared to ChatGPT-4's OR of 0.616 (95% CI: 0.297-1.278) .

Conclusions: Our study highlights the potential of LLMs in assisting dermatologists but also reveals their limitations. Both models made errors in diagnosing melanoma and benign lesions. Claude 3 Opus misdiagnosed melanoma as benign lesions in several cases, while ChatGPT-4 made similar errors. Conversely, both models misclassified benign lesions as melanoma in some examples. These findings underscore the limitations of current AI models and emphasize that they may not replace clinical diagnosis and treatment. In the future, more research should focus on developing robust, transparent, and clinically validated

models through collaborative efforts between AI researchers, dermatologists, and other healthcare professionals. While AI can provide valuable insights, it is crucial to recognize that these models are not yet capable of replacing the expertise and judgment of trained clinicians in diagnosing and managing skin lesions. Clinical Trial: None

(JMIR Preprints 08/04/2024:59273)
DOI: https://doi.org/10.2196/preprints.59273

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Comparing the Performance of Claude-3 Opus and ChatGPT-4 in Dermoscopic images Analysis and Melanoma Diagnosis: Exploring the Application Potential of Different Large Language Models in Dermatology

Xu Liu, BS,[a,b†] Chaoli Duan, MS,[a,b†] Min-kyu Kim, MS,[a,b†] Eunjin Jee, MS,[a,b] Beenu Maharjan, MS,[a,b] Yuwei Huang, MD,[a,b] Dan Du, MS,[a,b] Lu Zhang, MD,[a,b△] Xian Jiang, MD[a,b△]
[a]Department of Dermatology, West China Hospital, Sichuan University, Chengdu, People's Republic of China. [b] Laboratory of Dermatology, Clinical Institute of Inflammation and Immunology, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, People's Republic of China
† These authors contributed equally to the article and claim as co-1st authors

△Corresponding Author: Dr. Xian Jiang, Department of Dermatology, West China Hospital, Sichuan University, No. 37 Guoxue Xiang, Chengdu, China, 610041.
Dr. Lu Zhang, Department of Dermatology, West China Hospital, Sichuan University,
No. 37 Guoxue Xiang, Chengdu, China, 610041.
E-mail: jiangxian@scu.edu.cn and zhangluhx@scu.edu.cn

Abstract
We compared Claude 3 Opus and ChatGPT-4 in diagnosing dermoscopic images. We randomly selected 100 histopathology-confirmed dermoscopic images (50 malignant, 50 benign) from the International Skin Imaging Collaboration (ISIC) Archive database. We assessed primary diagnosis accuracy, top 3 differential diagnoses accuracy, and malignancy discrimination ability. In terms of malignancy discrimination, Claude 3 Opus outperformed ChatGPT-4 with 47.06% sensitivity, 81.63% specificity, and 64.00% accuracy compared to 45.10%, 42.86%, and 44.00%, respectively (p=0.001). Claude 3 Opus had an OR of 3.951 (95% CI: 1.685-9.263), indicating a stronger association between its predictions and actual malignancy compared to ChatGPT-4's OR of 0.616 (95% CI: 0.297-1.278).

KeywordsArtificial Intelligence, Large Language Models, Claude, ChatGPT, Dermatology

To the Editor: Recent advancements in artificial intelligence (AI) and large language models (LLMs) have shown promising potential in various medical fields, including dermatology. LLMs, such as ChatGPT, have demonstrated their ability to generate human-like responses to text-based prompts and assist in clinical decision-making [1]. With the introduction of image analysis capabilities in LLMs, such as ChatGPT Vision [2], the application of these models in dermatological diagnostics has garnered significant interest.However, the emergence of other LLMs, such as Claude 3 Opus, warrants investigation. Claude 3 Opus is an advanced conversational AI model that has shown promising performance in various natural language processing tasks [3]. Its ability to engage in context-aware dialogues and provide coherent responses makes it a potential candidate for assisting in clinical decision-making, including dermatological diagnostics. We compared the diagnostic performance of Claude 3 Opus and ChatGPT-4 to provide insights into their strengths and weaknesses and guide the selection and optimization of AI-assisted diagnostic tools in dermatology.
We randomly selected 100 histopathology-confirmed dermoscopic images (50 malignant, 50 benign) from the International Skin Imaging Collaboration (ISIC) Archive database [4]. Each model was prompted to provide the top 3 differential diagnoses for each image, ranked by likelihood. The models' responses were recorded for further analysis (Fig.1 a-b). We assessed primary diagnosis accuracy, top 3 differential diagnoses accuracy, and malignancy discrimination ability. McNemar's test determined statistical significance (α=0.05). For primary diagnosis accuracy, Claude 3 Opus achieved 54.90% sensitivity, 57.14% specificity, and 56.00% accuracy, while GPT4-Vision

demonstrated 56.86% sensitivity, 38.78% specificity, and 48.00% accuracy (p=0.170). For top 3 differential diagnoses accuracy, Claude 3 Opus and ChatGPT-4 included the correct diagnosis in 76.00% and 78.00% of cases, respectively (p=0.564). For malignancy discrimination, Claude 3 Opus outperformed ChatGPT-4 with 47.06% sensitivity, 81.63% specificity, and 64.00% accuracy compared to 45.10%, 42.86%, and 44.00%, respectively (p=0.001). Further quantifying the difference in malignancy discrimination ability, we calculated odds ratios (ORs) and 95% confidence intervals (CIs). Claude 3 Opus had an OR of 3.951 (95% CI: 1.685-9.263), indicating a stronger association between its predictions and actual malignancy compared to ChatGPT-4's OR of 0.616 (95% CI: 0.297-1.278) (Fig.1 c-d).

Our study highlights the potential of LLMs in assisting dermatologists but also reveals their limitations. Both models made errors in diagnosing melanoma and benign lesions (Fig.1 e-h). Claude 3 Opus misdiagnosed melanoma as benign lesions in several cases, while ChatGPT-4 made similar errors. Conversely, both models misclassified benign lesions as melanoma in some examples. These findings underscore the limitations of current AI models and emphasize that they may not replace clinical diagnosis and treatment. In the future, more research should focus on developing robust, transparent, and clinically validated models through collaborative efforts between AI researchers, dermatologists, and other healthcare professionals. While AI can provide valuable insights, it is crucial to recognize that these models are not yet capable of replacing the expertise and judgment of trained clinicians in diagnosing and managing skin lesions.

Conflicts of Interest

All authors disclosed no relevant relationships.

References

[1] Rundle CW, Szeto MD, Presley CL, Shahwan KT, Carr DR. Analysis of ChatGPT generated differential diagnoses in response to physical exam findings for benign and malignant cutaneous neoplasms. J Am Acad Dermatol. 2024;90(3):615-616. doi:10.1016/j.jaad.2023.10.040

[2] Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. J Am Acad Dermatol. Published online January 19, 2024. doi:10.1016/j.jaad.2023.12.062

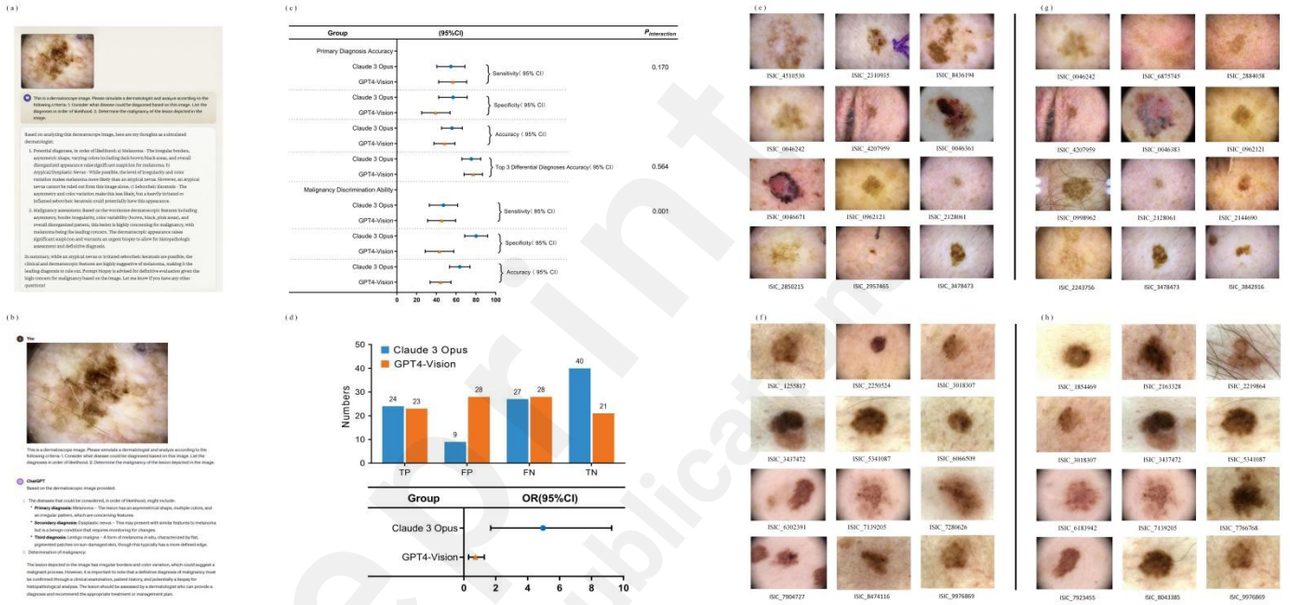
[3] Anthropic. Introducing Claude. Accessed October 20, 2023. https://www.anthropic.com/index/introducing-claude

[4] The International Skin Imaging Collaboration (ISIC) archive. Accessed October 19, 2023. https://www.isic-archive.com

Funding sources: The National Natural Science Foundation of China (82273559，82304052 and 82073473).

Figure legends

Fig. 1. Performance Comparison of Claude 3 Opus and GPT4-Vision in Skin Dermoscopy Image Analysis and Melanoma Diagnosis: Application scenario, Data Comparison Results, Statistical Analysis, and Misdiagnosis Examples.



[a.]Application scenario of Claude 3 Opus in the analysis process of dermoscopic images. [b.]Application scenario of GPT4-Vision in the analysis process of dermoscopic images. [c.]Claude 3 Opus and GPT4-Vision in the analysis of dermoscopic images and the diagnosis of melanoma based on data comparison. [d.]The OR values and their 95% confidence intervals for Claude 3 Opus and GPT4-Vision in terms of their ability to distinguish between benign and malignant conditions. [e.]Examples of Claude 3 Opus misdiagnosing melanoma as benign lesions. [f.]Examples of Claude 3 Opus misdiagnosing benign lesions as melanoma. [g.]Examples of GPT4-Vision misdiagnosing melanoma as benign lesions. [h.]Examples

of GPT4-Vision misdiagnosing benign lesions as melanoma.