Lasse Cirkel, Fabian Lechner, Lukas Alexander Henk, Martin Krusche, Martin C. Hirsch, Michael Hertl, Sebastian Kuhn and Johannes Knitza*

Large language models for dermatological image interpretation – a comparative study

https://doi.org/10.1515/dx-2025-0014
Received January 27, 2025; accepted March 31, 2025;
published online May 23, 2025

Abstract

Objectives: Interpreting skin findings can be challenging for both laypersons and clinicians. Large language models (LLMs) offer accessible decision support, yet their diagnostic capabilities for dermatological images remain underexplored. This study evaluated the diagnostic performance of LLMs based on image interpretation of common dermatological diseases.

Methods: A total of 500 dermatological images, encompassing four prevalent skin conditions (psoriasis, vitiligo, erysipelas and rosacea), were used to compare seven multimodal LLMs (GPT-4o, GPT-4o mini, Gemini 1.5 Pro, Gemini 1.5 Flash, Claude 3.5 Sonnet, Llama3.2 90B and 11B). A standardized prompt was used to generate one top diagnosis.

Results: The highest overall accuracy was achieved by GPT-4o (67.8 %), followed by GPT-4o mini (63.8 %) and Llama3.2 11B (61.4 %). Accuracy varied considerably across conditions, with psoriasis with the highest mean LLM accuracy of 59.2 % and erysipelas demonstrating the lowest accuracy (33.4 %). 11.0 % of all images were misdiagnosed by all LLMs, whereas 11.6 % were correctly diagnosed by all models. Correct diagnoses by all LLMs were linked to clear, disease-specific features, such as sharply demarcated erythematous plaques in psoriasis. Llama3.2 90B was the only LLM to decline diagnosing images, particularly those involving intimate areas of the body.

Conclusions: LLM performance varied significantly, emphasizing the need for cautious usage. Notably, a free, locally hostable model correctly identified the top diagnosis for approximately two-thirds of all images, demonstrating the potential for safer, locally deployed LLMs. Advancements in model accuracy and the integration of clinical metadata could further enhance accessible and reliable clinical decision support systems.

Keywords: artificial intelligence; large language models; skin pathology; dermatology; diagnosis; ChatGPT

Introduction

Artificial Intelligence (AI) has become a transformative force in medicine, offering innovative solutions to enhance diagnostic accuracy, streamline workflows, and improve patient outcomes [1–3]. In contrast to traditional models, multimodal large language models (LLMs) can process and integrate a range of data types, including images, text, and structured information. The promise of multimodal LLMs is their ability to address a range of diverse diagnostic challenges across specialties [4, 5] using nearly any given data source.

Dermatology offers a unique testing ground for multimodal LLMs, as visual features are central to diagnosis. Even experienced dermatologists are experimenting to incorporate LLMs into their diagnostic processes [6]. Concurrently, the field of dermatology presents a challenging environment for AI due to the presence of ambiguous presentations, overlapping features, and diverse skin tones. The application of advanced machine learning algorithms and deep learning models has yielded

***Corresponding author: Johannes Knitza**, MD, MHBA, PhD, Institute for Digital Medicine, University Hospital Giessen-Marburg, Philipps University Marburg, Baldingerstr. 1, 35043 Marburg, Germany; and Université Grenoble Alpes, AGEIS, Grenoble, France, E-mail: johannes.knitza@uni-marburg.de. https://orcid.org/0000-0001-9695-0657
Lasse Cirkel and Fabian Lechner, Institute of Artificial Intelligence, University Hospital Gießen-Marburg, Philipps University, Marburg, Germany; and Institute for Digital Medicine, University Hospital Gießen-Marburg, Philipps University, Marburg, Germany
Lukas Alexander Henk, Institute for Digital Medicine, University Hospital Gießen-Marburg, Philipps University, Marburg, Germany; and Department of Dermatology and Allergology, University Hospital Gießen-Marburg, Philipps University, Marburg, Germany
Martin Krusche, Division of Rheumatology and Systemic Inflammatory Diseases, III. Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
Martin C. Hirsch, Institute of Artificial Intelligence, University Hospital Gießen-Marburg, Philipps University, Marburg, Germany
Michael Hertl, Department of Dermatology and Allergology, University Hospital Gießen-Marburg, Philipps University, Marburg, Germany
Sebastian Kuhn, Institute for Digital Medicine, University Hospital Gießen-Marburg, Philipps University, Marburg, Germany

impressive results in tasks such as melanoma detection and lesion classification [7–9]. Nevertheless, a first comparative LLM study analyzing dermoscopic images for melanoma detection yielded an accuracy of only roughly 50 % [10]. Despite the great potential and increasing usage, there is still a paucity of knowledge regarding the performance of general purpose multimodal LLMs in dermatology. This study aimed to compare the diagnostic performance of seven multimodal LLMs for four common dermatological conditions.

Materials and methods

This study focused on four common dermatological diagnoses, namely, psoriasis, vitiligo, erysipelas and rosacea. Images from two publicly available verified dermatology datasets were used in this study. Atlas Dermatológico, is a comprehensive online database containing over 12,000 freely available images representing a wide range of skin diseases as of December 2024 [11]. The second dataset, DermIS.net, is described as the most comprehensive dermatology information service on the internet [12]. These datasets contain images labeled and categorized by dermatology experts and have been widely referenced in dermatological research, providing a diverse set of real-world skin disease images. While we did not conduct an independent review of image representativeness with board-certified dermatologists, we relied on the expert-curated nature of these databases to ensure data quality. All images associated with the respective skin diseases available on 01 December 2024 were used. No images were manually excluded or reviewed prior to analysis to ensure an unbiased sample. The final dataset contained a total of 500 images covering the four different skin diseases. The distribution of images per disease is shown in Table 1.

Disease selection

The selection of psoriasis, vitiligo, erysipelas, and rosacea was based on the availability of a large number of well-annotated, publicly accessible images in dermatological databases and the presence of distinct visual features that often allow for reliable diagnosis based on images. Many other dermatological conditions require additional patient history and clinical context for accurate diagnosis, which is beyond the scope of this image-based evaluation.

Table 1: Dataset with respective image sources and diseases.

Disease	Atlas dermatológico	DermIS	Total
Psoriasis	162	92	254
Vitiligo	70	38	108
Erysipelas	52	33	85
Rosacea	26	27	53
Total	310	190	500

Large language models

Seven state-of-the-art multimodal LLMs were compared in this study: OpenAI's GPT4o (gpt-4o-2024-11-20) and GPT-4o mini (gpt-4o-mini-2024-07-18), Google's Gemini 1.5 Pro (gemini-1.5-pro-002) and Gemini 1.5 Flash (gemini-1.5-flash-002), Anthropic's Claude 3.5 Sonnet (claude-3-5-sonnet-20241022), and Meta's Llama3.2 90B and Llama3.2 11B.

Five of these models (GPT-4o, GPT-4o mini, Gemini 1.5 Pro, Gemini 1.5 Flash and Claude 3.5 Sonnet) represent fully multimodal LLMs with comprehensive capabilities in processing various data types. Meta's Llama3.2 models are medium-sized vision LLMs with more focused capabilities, specifically designed for image understanding tasks. While they don't support video processing or image generation, these models offer distinct advantages, as they can be run locally, providing enhanced data privacy and independence from external API services.

Data preparation

All images were used at their original full resolution to preserve diagnostic detail. To ensure compatibility with the input requirements of each LLM's API interface, image formats were adjusted accordingly. For OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet, images were converted to Base64 encoding using the standard Python library base64. No conversion was required for Google's Gemini and Meta's Llama3.2 models, which accept.jpg images. All metadata embedded in the image files, such as EXIF data, was removed during processing using the Python Imaging Library (PIL/Pillow) [13]. This step ensures that no inadvertent information could bias the output of the models. Images were renamed with sequential numerical identifiers to further anonymize the data.

Experimental procedure

For five of the models (GPT-4o, GPT-4o mini, Gemini 1.5 Pro, Gemini 1.5 Flash and Claude 3.5 Sonnet), each image was sequentially input using their respective API interfaces. For

System Prompt

You are a helpful dermatological assistant who classifies skin conditions.

User Prompt

Please examine the provided dermatological image and identify the disease. Respond only with the name of the disease. Do not include any additional information or explanations.
`{image}`

Figure 1: Standardized system and user prompts.

Meta's Llama3.2 models, which were deployed locally, images were sequentially processed using Ollama (Version v0.5.1) as the interface. A new session was created for each image to ensure that the models did not retain information from previous interactions that could influence their responses. To ensure comparability and reduce bias the same standardized prompt was used for all models Figure 1. LLMs were prompted to generate the most likely diagnosis based on the image provided. Cases where models refused to provide a diagnosis or where API requests were blocked were rated as incorrect classifications. No API failures were noted during the study period. LLM requests were made between 07 December 2024 and 14 December 2024.

Statistical methods

All descriptive statistical analyses were performed using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Ethical aspects

All images were publicly available on the respective databases. The Philipps-University Marburg Research Ethics Committee confirmed that no ethical approval was required (reference number: 23–300 ANZ) due to the anonymous and non-interventional nature of the study.

Results

Overall model accuracies

The evaluation revealed notable differences in model performance, see Table 2 and Figure 2. The highest overall accuracy was achieved by GPT-4o, with a score of 67.8 %, followed by GPT-4o mini (63.8 %), and Llama3.2 11B (61.4 %). The larger Llama3.2 90B model exhibited an inferior performance compared to its smaller counterpart, achieving only 50.8 %. This can be attributed, at least in part, to its tendency to frequently refuse to diagnose certain images (91, 18.2 %). Gemini 1.5 Flash demonstrated the lowest overall accuracy with 37.0 %.

Disease specific performance

Diagnostic accuracy varied across the four conditions Figure 2, with the highest mean LLM accuracy for psoriasis (59.2 %) and lowest for erysipelas (33.4 %). Performance of LLMs varied across diseases. For psoriasis, the GPT-4o mini model was identified as the most effective, with an accuracy of 80.3 %, followed by the Llama3.2 11B model (77.6 %). For vitiligo GPT-4o demonstrated the highest accuracy with 78.7 %. Erysipelas proved a significant challenge for all models, with accuracy rates ranging from 16.5 to 50.6 %. Llama3.2 90B achieved the second-best accuracy for this condition, with 44.7 %. This was despite its refusal to generate a diagnosis for

Table 2: Accuracy of LLMs in % across multiple diseases.

Disease, n	GPT-4o	GPT-4o Mini	Gemini 1.5 pro	Gemini 1.5 Flash	Claude 3.5 Sonnet	Llama 3.2 90B	Llama 3.2 11B
Psoriasis (254)	69.69	**80.31**	45.67	39.37	53.15	48.43	77.56
Vitiligo (108)	**78.70**	62.96	55.55	26.85	57.41	61.11	63.89
Erysipelas[a] (85)	**49.41**	12.94	31.76	36.47	22.35	44.71	36.47
Rosacea (53)	66.04	**67.92**	56.60	47.17	54.72	50.94	18.87
Overall accuracy[b]	**67.80**	63.80	46.60	37.00	49.00	50.80	61.40

Bold values indicate the highest accuracy achieved for each dermatological condition among the tested models. [a]Cellulitis was also rated as correct. [b]Weighted accuracy.

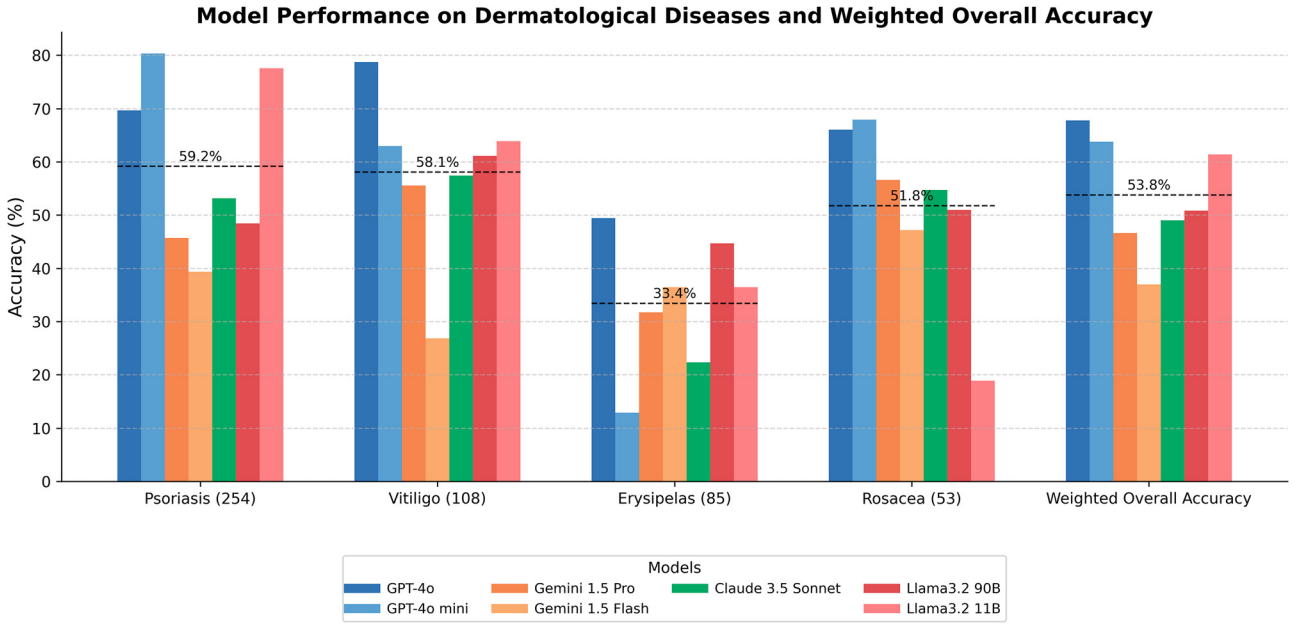

Figure 2: Diagnostic accuracy of respective large language models according to respective dermatological diseases.

23.5 % (20/85) of the erysipelas images. For rosacea GPT-4o mini performed best with an accuracy 67.9 %.

Patterns in misdiagnosed images

Analysis of the misdiagnosed images revealed notable patterns (Table 3). A total of 11.0 % (55/500) images were misclassified by all models, highlighting the challenges of interpreting ambiguous or visually overlapping features. In contrast, 11.6 % (58/500) images were correctly classified by all models, characterized by clear and distinctive disease-specific features, such as the sharply demarcated erythematous plaques in psoriasis. Visual examples of three vitiligo images correctly classified by all models and three images misclassified by all models are provided in Figure 3 to illustrate these patterns.

Table 3: Classification results by disease and overall.

Disease	Total images, n (%)	All LLMs are false, n (%)	≥5 LLMs are false, n (%)	All LLMs are right, n (%)	≥5 LLMs are right, n (%)
Psoriasis	254 (100.0)	15 (5.9)	66 (26.0)	39 (15.4)	123 (48.4)
Vitiligo	108 (100.0)	10 (9.3)	27 (25.0)	14 (13.0)	52 (48.1)
Erysipelas	85 (100.0)	20 (23.5)	48 (56.5)	3 (3.5)	14 (16.5)
Rosacea	53 (100.0)	10 (18.9)	23 (43.4)	2 (3.8)	13 (24.5)
Overall	500 (100.0)	55 (11.0)	164 (32.8)	58 (11.6)	202 (40.4)

Inability to generate diagnoses

All LLMs generated a diagnosis except for Meta's Llama3.2 90B model, which frequently refused to diagnose images, particularly those depicting intimate areas of the body. This behavior was most evident in psoriasis with 19.7 % (50/254) refusals, vitiligo 19.4 % (21/108) and erysipelas 23.5 % (20/85). The pattern of refusals suggests that the model's responses are influenced by the body regions depicted in the images. Interestingly, there were no refusals for rosacea, although the dataset consisted mainly of facial images. The smaller Llama3.2 11B model consistently provided diagnoses for all images, regardless of body region, contributing to its superior performance compared to its larger counterpart.

Discussion

This study assessed the performance of LLMs in interpreting images of common dermatological conditions. The results of this study demonstrate a relatively high diagnostic accuracy of certain models even without domain-specific further training. The strong performance of GPT-4o, which achieved the highest overall accuracy, highlights the potential of general-purpose multimodal models for visually distinct conditions such as vitiligo and psoriasis. The competitive performance of Llama3.2 11B, despite its smaller size, free availability, and local deployment capabilities, highlights the potential of medium-sized, privacy-preserving models for

Figure 3: Examples of correctly and incorrectly classified images. The top row shows three images (A–C) that were correctly classified as vitiligo by all seven LLMs, whereas the bottom row shows three images (D–F) that could not be correctly diagnosed as vitiligo by any of the LLMs. Image sources: www.dermis.net and www.atlasdermatologico.com.br.

diagnostic decision support. However, the limitations observed highlight the need for significant improvements before such systems can be reliably integrated into clinical workflows. SkinGPT-4 exemplifies a privacy-preserving LLM trained using a two-step strategy [14]. The model was rigorously tested on data from 150 real-world patients, achieving a diagnostic accuracy or relevance rate of 80.6 % as evaluated by board-certified dermatologists. The variability in accuracy between conditions highlights the inherent complexity of dermatological diagnosis. Conditions with high-quality images and clear and well-defined visual features, such as vitiligo, were generally well-recognized by most models. In contrast, conditions such as erysipelas, which are characterized by overlapping and more subtle features, posed a greater challenge.

This discrepancy highlights a key limitation of image-only diagnosis, where contextual clinical information often plays a critical role in accurate decision-making. Incorporating multimodal data streams, including demographic variables, medical history and laboratory findings, could enable these models to align visual patterns with contextual cues, moving beyond pattern recognition towards more nuanced and clinically relevant analysis. Future research should prioritize the integration of such multimodal approaches to bridge the observed performance gaps. In addition, direct comparisons with board-certified dermatologists could provide valuable context for assessing LLM performance. Although beyond the scope of the current study, expert benchmarks would help determine whether models truly add diagnostic value or simply reflect patterns already present in publicly available datasets. Establishing such comparisons in future research could clarify the strengths and limitations of LLM-based dermatological diagnosis.

The interpretability of LLM-based diagnostics remains a pressing concern. While models such as GPT-4o and Llama3.2 11B have demonstrated substantial diagnostic accuracy, their decision-making processes remain inherently opaque. This lack of transparency may hinder clinical uptake and buy-in, as understanding the rationale behind a diagnosis is crucial for trust and validation in medical practice. The provision of interpretable output, such as visual explanations highlighting pathologies could address this issue [15]. Improved interpretability would not only promote confidence among clinicians but also support regulatory approval processes by ensuring that diagnostic decisions are explainable, reproducible and ethically sound. Data-related limitations also require careful consideration [16]. The use of publicly available datasets carries the risk of geographical bias and under-representation of different skin colors and conditions. In addition, the lack of histopathological confirmation and unclear labelling criteria may

compromise the validity of the ground truth, making it difficult to distinguish between model error and dataset imperfections.

The integration of LLMs into diagnostic workflows not only poses significant regulatory challenges but also necessitates rigorous clinical validation. Under the European Union's Medical Device Regulation, AI systems intended for diagnostic purposes are classified as medical devices, requiring stringent validation to ensure safety and efficacy. Additionally, the recently passed EU Artificial Intelligence Act categorizes healthcare AI systems as high-risk, mandating compliance with strict performance, transparency, and ethical standards [17]. Beyond regulatory requirements, these systems must undergo robust clinical evaluation through preclinical and early clinical validation studies, including clinical simulation settings and/or Early Feasibility Studies [18, 19]. These studies provide critical insights into the safety, effectiveness, and real-world applicability of LLMs, identifying potential risks and limitations early in the development process. Combining regulatory compliance with phased clinical validation will be essential for achieving both trust and widespread adoption in clinical practice.

A main limitation of this study is that tested LLM models likely had access to parts of the publicly available image datasets. This familiarity could artificially inflate diagnostic accuracy, as the models may have been trained or fine-tuned on these datasets or closely related ones. Consequently, the results may not fully reflect the models' true capabilities when applied to unseen, real-world clinical data. To address this, future research should prioritize the use of private, independently curated datasets that are not accessible during model training. Incorporating real-world clinical images from diverse patient populations would enhance the robustness and generalizability of findings. This approach would also mitigate concerns of data leakage, ensuring that model performance more accurately reflects its diagnostic potential in practice.

Another limitation of this study is its focus on only four dermatological conditions. While these conditions represent a spectrum of diagnostic complexity, they do not encompass the full range of dermatological presentations. Consequently, the findings provide limited insight into the models' capabilities across the broader spectrum of skin diseases, including rarer or more nuanced conditions. Future studies should aim to include a wider variety of dermatological diagnoses to ensure a more comprehensive evaluation of LLM performance. This broader analysis would help establish whether the observed diagnostic trends hold across diverse and less well-defined conditions.

Nevertheless, we believe comparative benchmarking studies are crucial and can guide decisions on which models to choose for which purposes. Given the rapid iteration of large language models, newer versions with potentially improved capabilities are frequently released. While this study assessed models available at the time of testing, future research should continue to apply systematic benchmarking to newer models to track progress and evaluate whether performance improvements extend to dermatological image interpretation. To ensure meaningful comparisons over time, it is equally important that these evaluations are conducted on high-quality, diverse datasets. Therefore, future studies should incorporate private, independently curated datasets that include a variety of skin tones, conditions, and clinical contexts to ensure robust and generalizable model evaluations. Additionally, it seems crucial to evaluate the actual clinical impact of model usage.

Conclusions

This study demonstrates that multimodal LLMs can effectively identify key diagnostic features in dermatological images, even in the absence of domain-specific training. Their ability to recognize conditions with distinctive visual patterns, such as vitiligo and psoriasis, highlights their potential for wider clinical applications. In particular, the comparable performance of locally deployed models such as Llama3.2 11B to large, cloud-based solutions highlights important implications for privacy and scalability in medical settings. Further studies are warranted investigating how to best integrate multimodal LLMs into clinical workflows.

Research ethics: The local Institutional Review Board deemed the study exempt from review (reference number: 23–300 ANZ).
Informed consent: Not applicable.
Author contributions: LC and JK contributed to the conception and design of the study, data collection, analysis, and interpretation. All authors were involved in drafting the manuscript, critically revising it for important intellectual content. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: ChatGPT-4o was used to improve the language of manuscript sections.
Conflict of interest: JK declares research support from Abbvie, Vila Health, honoraria and consulting fees from Abbvie, AstraZeneca, BMS, Boehringer Ingelheim, Chugai, GAIA, Galapagos, GSK, Janssen, Lilly, Medac, Novartis, Pfizer, Sobi, Rheumaakademie, UCB, Vila Health and Werfen. MK declares research support from Abbvie, Sobi and Sanofi;

honoraria and consulting fees from Abbvie, BMS, Boehringer Ingelheim, AlfaSigma, GSK, Janssen, Lilly, Medac, Novartis, Pfizer, Sobi, UCB. The remaining authors declare no competing interests.

Research funding: None declared.

Data availability: The raw data analysed during the current study are available from the corresponding author upon reasonable request.

References

1. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23:689.
2. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021;139:4–15.
3. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med 2023;3:141.
4. Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of gemini models in medicine. Available from: https://arxiv.org/abs/2404.18416.
5. Strotzer QD, Nieberle F, Kupke LS, Napodano G, Muertz AK, Meiler S, et al. Toward foundation models in radiology? Quantitative assessment of GPT-4V's multimodal and multianatomic region capabilities. Radiology 2024;313:e240955.
6. Gui H, Rezaei SJ, Schlessinger D, Weed J, Lester J, Wongvibulsin S, et al. Dermatologists' perspectives and usage of Large Language Models in practice: an exploratory survey. J Invest Dermatol 2024;144:2298–301.
7. Brancaccio G, Balato A, Malvehy J, Puig S, Argenziano G, Kittler H. Artificial intelligence in skin cancer diagnosis: a reality check. J Invest Dermatol 2024;144:492–9.
8. Escalé-Besa A, Yélamos O, Vidal-Alaball J, Fuster-Casanovas A, Miró Catalina Q, Börve A, et al. Exploring the potential of artificial intelligence in improving skin lesion diagnosis in primary care. Sci Rep 2023;13:4293.
9. Sanchez K, Kamal K, Manjaly P, Ly S, Mostaghimi A. Clinical application of artificial intelligence for non-melanoma skin cancer. Curr Treat Options Oncol 2023;24:373–9.
10. Liu X, Duan C, Kim M, Zhang L, Jee E, Maharjan B, et al. Claude 3 opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. JMIR Med Inform 2024;12:e59273.
11. da Silva SF: Atlas dermatológico. https://atlasdermatologico.com.br/index.jsf [Accessed 20 Dec 2024].
12. A cooperation between the dept of clinical social medicine (Univ of Heidelberg) and the dept of dermatology (Univ of Erlangen). DermIS - Dermatology Information System. https://www.dermis.net/dermisroot/en/home/index.htm [Accessed 20 Dec 2024].
13. Clark, JA and contributors. Pillow documentation. https://pillow.readthedocs.io/en/stable/index.html [Accessed 20 Dec 2024].
14. Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nat Commun 2024;15:5649.
15. Bhandari A. Revolutionizing radiology with artificial intelligence. Cureus 2024;16:e72646.
16. Navigli R, Conia S, Ross B. Biases in Large Language Models: origins, inventory, and discussion. J Data Inf Qual 2023;15. https://doi.org/10.1145/3597307.
17. Gilbert S. The EU passes the AI Act and its implications for digital medicine are unclear. NPJ Digit Med 2024;7:135.
18. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28:924–33.
19. Lau K, Halligan J, Fontana G, Guo C, O'Driscoll FK, Prime M, et al. Evolution of the clinical simulation approach to assess digital health technologies. Future Healthc J 2023;10:173–5.