Dermatology Practical & Conceptual

Artificial Intelligence-Based Image Analysis is Insufficient as a Stand-Alone Assessment of Skin Tumors in Real Clinical Practice

Aimilios Lallas,[1] Konstantinos Liopyris,[2] Zoe Apalla,[3] Elvira Moscarella,[4] Gabriella Brancaccio,[4] Alexander Stratigos,[2] Giuseppe Argenziano[4]

1 First Department of Dermatology, School of Medicine, Faculty of Health Sciences, Aristotle University, Thessaloniki, Greece
2 First Department of Dermatology, National and Kapodistrian University of Athens School of Medicine, Andreas Sygros Hospital, Athens, Greece
3 Second Department of Dermatology, School of Medicine, Faculty of Health Sciences, Aristotle University, Thessaloniki, Greece
4 Dermatology Unit, University of Campania L. Vanvitelli, Naples, Italy

Key words: Artificial Intelligence (AI), Skin tumors, Total body photography (TBP), Digital dermoscopy (DD), Diagnostic accuracy

Citation: Lallas A, Liopyris K, Apalla Z, et al. Artificial Intelligence-Based Image Analysis is Insufficient as a Stand-Alone Assessment of Skin Tumors in Real Clinical Practice. *Dermatol Pract Concept*. 2025;15(2):5353. DOI: https://doi.org/10.5826/dpc.1502a5353

Accepted: November 30, 2024; **Published:** April 2025

Funding: None.

Competing Interests: None.

Authorship: All authors have contributed significantly to this publication.

Corresponding Author: Aimilios Lallas, First Department of Dermatology, School of Medicine, Faculty of Health Sciences, Aristotle University, Thessaloniki, Greece.Address: Hospital for Skin and Venereal Diaeases, 124 Delfon Street, 54643, Thessaloniki, Greece. ORCID ID: 0000-0002-7193-0964. E-mail: alallas@auth.gr

Introduction

Multiple reader studies have shown that artificial intelligence (AI)- based analysis can classify images of benign and malignant skin tumors with an accuracy comparable, or superior to, human readers [1-3]. Although not tested in real-life clinical settings, automated classification has been incorporated in most available devices for total body photography (TBP) and digital dermoscopy (DD). This offers an opportunity for an initial assessment of its performance in a real-life clinical setting.

Case Presentation

We randomly examined the performance of automated AI classification of all photographed skin lesions in consecutive patients who underwent TBP and DD over a period of one week in September 2025. No patient was excluded from the analysis, regardless of the reason that justified the indication for TBP and DD. Similarly, no lesion was excluded from the analysis. We subsequently compared the AI assessment outcome with the management decision of the visiting clinicians, considering the latter as the gold standard for clinical diagnosis. The study was conducted in three skin cancer referral centers in Greece and Italy that use different devices for TBP and DD. All involved clinicians have an adequate level of expertise in skin cancer diagnosis. The clinicians were unaware of the AI assessment and the study aim at the time of the visit. Overall, 4792 images from 76 patients were included in the analysis. Of 76 patients, 49 had at least one previous TBP and DD visit, while for 27 this was the

Table 1. Diagnostic Accuracy of Artificial Intelligence (AI) algorithm.

	AI Assessment		Clinical Assessment	
	Not suspicious	Suspicious	Not suspicious	Suspicious
Number of patients	4	72	61	15
Total number of lesions	4620	172	4787	15
Correct AI diagnosis (lesion level)	4633/4792 = 96.7%			
AI Sensitivity (lesion level)	14/15 = 93.3%			
AI Specificity (lesion level)	4620/4787 = 96.5%			
AI PPV (lesion level)	14/172 = 8.1%			
AI NPV (lesion level)	4619/4620 = 99.9%			
AI Sensitivity (patient level)	15/15 = 100%			
AI Specificity (lesion level)	4/61 = 6.6%			
AI PPV (lesion level)	15/72 = 20.8%			
AI NPV (lesion level)	4/4 = 100%			

All the accuracy calculations were done using clinical assessment as referral (gold standard).

baseline documentation. The analytical results are shown in Table 1.

Discussion

Overall, of 4792 images, 172 were assessed by the AI algorithm as suspicious. In 72 of 76 patients, there was at least one lesion assessed as suspicious by the AI algorithm. According to the standard clinical assessment, 15 of 4792 lesions were deemed suspicious enough to be surgically excised. Of them, 12 were proven to be malignant (six basal cell carcinomas, three melanomas, and three squamous cell carcinomas) and three benign (atypical nevi). Of the 15 excised lesions, 14 had also been assessed as suspicious by the AI algorithm. The one lesion that escaped AI detection was proven to be a microinvasive melanoma. In terms of accuracy, the AI assessment was correct in 4633/4792 lesions (96.7%). The sensitivity of the AI algorithm for malignancy was 93.3%, and the specificity at the lesion level was 96.5%. The negative predictive value (NPV) was 99.9%, and the positive predictive value (PPV) was 8.1%. At the patient level, the specificity of the AI algorithm was 6.6%, with 57 of 61 healthy patients being falsely assessed as carrying one or more suspicious lesions. There was no significant difference in the PPV and NPV between the devices. The accuracy calculations were repeated only for those patients with available previous visit, with essentially identical results. These numbers suggest that the concept of using AI-based classification of skin tumors as a stand-alone method for diagnosis and patient management in clinical practice, without the involvement of a skilled clinician, is contraindicated, at least at this moment. The very low PPV indicates a wide margin of potential false positive predictions that would lead to numerous unnecessary excisions of benign moles, which today are not excised because of the much superior performance of skilled clinicians. Such a practice would induce unnecessary morbidity to patients and a tremendous burden on any health system.

Conclusion

What remains to be further elucidated is whether AI-based classification of skin tumors could serve as a supportive tool for clinicians by reducing the number of lesions to be evaluated. This hypothesis is supported by the high NPV, suggesting that it could function as a safe triage mechanism for the selection of lesions that require further evaluation by clinicians. In the present example, only 3.6% of the 4792 lesions were assessed as “suspicious” by AI. Sparing 4620 of 4792 lesions from evaluation would have an obvious time-saving effect and would cost the overlooking of one melanoma.

Whether this is a reasonable cost to pay for the time saved remains a complicated discussion. There is little doubt about the great potential of AI-based analysis of macroscopic images from TBP and/or DD. Whether this great potential will transform into practice-changing developments depends on the tasks that should be assigned to AI algorithms. The initial experience suggests that AI-based classification of skin tumors is inadequate to accomplish the task of establishing a final diagnosis in clinical practice. Instead, it looks potentially appropriate for the task of triaging lesions to be evaluated or to highlight changes in sequential photographs, supporting clinicians' decision-making process.

References

1. Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol.* 2019; 20(7): 938-947. DOI:10.1016/S1470-2045(19)30333-X.
2. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. *Nat Med*. 2020; 26(8): 1229-1234. DOI: 10.1038/s41591-020-0942-0.
3. Haggenmüller S, Maron RC, Hekler A, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *Eur J Cancer*. 2021; 156: 202-216. DOI: 10.1016/j.ejca.2021.06.049.