

The Role of Dermatologists in Developing AI Tools for Diagnosis and Classification of Skin Disease

Sara D. Ragi, MD; Amar D. Desai, MD; Rachel C. Hill, MD; Shari R. Lipner, MD, PhD

PRACTICE POINTS

- Artificial intelligence (AI) technology is emerging as a valuable tool in diagnosing and classifying dermatologic conditions.
- Despite advances in AI for dermatologic image analysis, the creation of these models often has been directed by nondermatologists.

Artificial intelligence (AI) technology has proven to be a valuable tool in the diagnosis and classification of dermatologic conditions. To our knowledge, no prior studies have investigated the role of dermatologists in the development of AI programs for classification of dermatologic conditions other than skin cancers. In this study, we aimed to analyze AI tools used for diagnosing and classifying skin disease and evaluate the role of dermatologists in the creation of AI technology. Additionally, we investigated the number of clinical images used in datasets to train AI programs and compared tools that were created with dermatologist input to those created without dermatologist/clinician involvement.

Use of artificial intelligence (AI) in dermatology has increased over the past decade, likely driven by advances in deep learning algorithms, computing hardware, and machine learning.¹ Studies comparing the performance of AI algorithms to dermatologists in classifying skin disorders have shown conflicting results.^{2,3} In this study, we aimed to analyze AI tools used for diagnosing and classifying skin disease and evaluate the role of dermatologists in the creation of AI technology. We also investigated the number of clinical images used

in datasets to train AI programs and compared tools that were created with dermatologist input to those created without dermatologist/clinician involvement.

Methods

A search of PubMed articles indexed for MEDLINE using the terms *machine learning*, *artificial intelligence*, and *dermatology* was conducted on September 18, 2022. Articles were included if they described full-length trials; used machine learning for diagnosis of or screening for dermatologic conditions; and used dermoscopic or gross image datasets of the skin, hair, or nails. Articles were categorized into 4 groups based on the conditions covered: chronic wounds, inflammatory skin diseases, mixed conditions, and pigmented skin lesions. Algorithms were sorted into 4 categories: convolutional/convoluted neural network, deep learning model/deep neural network, AI/artificial neural network, and other. Details regarding Fitzpatrick skin type and skin of color (SoC) inclusion in the articles or AI algorithm datasets were recorded. Univariate and multivariate analyses were performed using Microsoft Excel and SAS Studio 3.8. Sensitivity and specificity were calculated for all included AI technology. Sensitivity, specificity, and the number of clinical images were compared among the included articles using analysis of variance and *t* tests ($\alpha=0.05$; $P<.05$ indicated statistical significance).

Results

Our search yielded 1016 articles, 58 of which met the inclusion criteria. Overall, 25.9% (15/58) of the articles utilized AI to diagnose or classify mixed skin diseases; 22.4% (13/58) for pigmented skin lesions; 19.0% (11/58)

Dr. Ragi is from the Warren Alpert Medical School of Brown University, Providence, Rhode Island. Dr. Desai is from Rutgers New Jersey Medical School, Newark. Drs. Hill and Lipner are from Weill Cornell Medical College, New York, New York. Dr. Lipner is from the Department of Dermatology. The authors have no relevant financial disclosures to report.

The eTable is available in the Appendix online at www.mdedge.com/cutis.

Correspondence: Shari R. Lipner, MD, PhD, Associate Professor of Clinical Dermatology, Weill Cornell Medicine, 1305 York Ave, 9th Floor, New York, NY 10021 (shl9032@med.cornell.edu).

Cutis. 2025 November;116(5):184-185, E4. doi:10.12788/cutis.1295

for wounds; 17.2% (10/58) for inflammatory skin diseases; and 5.2% (3/58) each for acne, psoriasis, and onychomycosis. Overall, 24.0% (14/58) of articles provided information about Fitzpatrick skin type, and 58.7% (34/58) included clinical images depicting SoC. Furthermore, we found that only 20.7% (12/58) of articles on deep learning models included descriptions of patient ethnicity or race in at least 1 dataset, and only 10.3% (6/58) of studies included any information about skin tone in the dataset. Studies with a dermatologist as the last author (most likely to be supervising the project) were more likely to include clinical images depicting SoC than those without (82.6% [19/23] and 16.7% [3/18], respectively [$P=.0411$]).

The mean (SD) number of clinical images in the study articles was 28,422 (84,050). Thirty-seven (63.8%) of the study articles included gross images, 17 (29.3%) used dermoscopic images, and 4 (6.9%) used both. Twenty-seven (46.6%) articles used convolutional/convoluted neural networks, 15 (25.9%) used deep learning model/deep neural networks, 8 (13.8%) used other algorithms, 6 (10.3%) used AI/artificial neural network, and 2 (3.4%) used fuzzy algorithms. Most studies were conducted in China (29.3% [17/58]), Germany (12.1% [7/58]), India (10.3% [6/58]), multiple nations (10.3% [6/58]), and the United States (10.3% [6/58]). Overall, 82.8% (48/58) of articles included at least 1 dermatologist coauthor. Sensitivity of the AI models was 0.85, and specificity was 0.85. The average percentage of images in the dataset correctly identified by a physician was 76.87% vs 81.62% of images correctly identified by AI. Average agreement between AI and physician assessment was 77.98%, defined as AI and physician both having the same diagnosis.

Articles authored by dermatologists contained more clinical images than those without dermatologists in key authorship roles ($P<.0001$) (eTable). Psoriasis-related algorithms had the fewest (mean [SD]: 3173 [4203]), and pigmented skin lesions had the most clinical images (mean [SD]: 53,191 [155,579]).

Comment

Our results indicated that AI studies with dermatologist authors had significantly more images in their datasets (ie, the set of clinical images of skin lesions used to train AI algorithms in diagnosing or classifying lesions) than those with nondermatologist authors ($P<.0001$) (eTable). Similarly, in a study of AI technology for skin cancer diagnosis, AI studies with dermatologist authors (ie, included in the development of the AI algorithm) had more images than studies without dermatologist authors.¹ Deep learning textbooks have suggested that 5000 clinical images or training input per output category are needed to produce acceptable algorithm performance, and more than 10 million are needed to produce results superior to human performance.⁴⁻¹⁰ Despite advances in AI for dermatologic image analysis, the creation of these models often has

been directed by nondermatologists¹; therefore, dermatologist involvement in AI development is necessary to facilitate collection of larger image datasets and optimal performance for image diagnosis/classification tasks.

We found that 20.7% of articles on deep learning models included descriptions of patient ethnicity or race, and only 10.3% of studies included any information about skin tone in the dataset. Furthermore, American investigators primarily trained models using clinical images of patients with lighter skin tones, whereas Chinese investigators exclusively included images depicting darker skin tones. Similarly, in a study of 52 cutaneous imaging deep learning articles, only 17.3% (9/52) reported race and/or Fitzpatrick skin type, and only 7.7% (4/52) of articles included both.^{2,6,8} Therefore, dermatologists are needed to contribute images representing diverse populations and collaborate in AI research studies, as their involvement is necessary to ensure the accuracy of AI models in classifying lesions or diagnosing skin lesions across all skin types.

Our search was limited to PubMed, and real-world applications could not be evaluated.

Conclusion

In summary, we found that AI studies with dermatologist authors used larger numbers of clinical images in their datasets and more images representing diverse skin types than studies without. Therefore, we advocate for greater involvement of dermatologists in AI research, which might result in better patient outcomes by improving diagnostic accuracy.

REFERENCES

1. Zakhem GA, Fakhoury JW, Motosko CC, et al. Characterizing the role of dermatologists in developing artificial intelligence for assessment of skin cancer. *J Am Acad Dermatol*. 2021;85:1544-1556.
2. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv*. 2022;8:eabq6147.
3. Wu E, Wu K, Daneshjou R, et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med*. 2021;27:582-584.
4. Murphree DH, Puri P, Shamim H, et al. Deep learning for dermatologists: part I. Fundamental concepts. *J Am Acad Dermatol*. 2022;87:1343-1351.
5. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. The MIT Press; 2016.
6. Kim YH, Kobic A, Vidal NY. Distribution of race and Fitzpatrick skin types in data sets for deep learning in dermatology: a systematic review. *J Am Acad Dermatol*. 2022;87:460-461.
7. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26:900-908.
8. Zhu CY, Wang YK, Chen HP, et al. A deep learning based framework for diagnosing multiple skin diseases in a clinical environment. *Front Med (Lausanne)*. 2021;8:626369.
9. Capurro N, Pastore VP, Touijer L, et al. A deep learning approach to direct immunofluorescence pattern recognition in autoimmune bullous diseases. *Br J Dermatol*. 2024;191:261-266.
10. Han SS, Park I, Eun Chang S, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J Invest Dermatol*. 2020;140:1753-1761.

APPENDIX

eTABLE. Comparison of Sensitivity, Specificity, and Number of Clinical Images Among Publications By Author And AI Characteristics ^a										
Characteristics	No. of articles (%)	Sensitivity			Specificity			No. of clinical images		
		Mean	SD	P value	Mean	SD	P value	Mean	SD	P value
First author dermatologist				.2116			.0838			<.0001
No	19 (47.5)	0.8126	0.1862		0.87	0.12		3149.8	3603.9	
Yes	21 (52.5)	0.8686	0.0743		0.81	0.11		25,221.9	56,128.7	
Last author dermatologist				.2999			.4092			<.0001
No	18 (43.9)	0.8678	0.1042		0.87	0.11		1377.6	2000.8	
Yes	23 (70.7)	0.8222	0.1592		0.84	0.12		24,978.6	2419.5	
First and last author dermatologist				.9275			.0384			<.0001
No	24 (60.0)	0.8383	0.1659		0.88	0.12		2746.0	3262.3	
Yes	16 (40.0)	0.8425	0.09		0.80	0.11		32,272.4	62,468.1	
Corresponding author dermatologist				.6532			.6889			<.0001
No	19 (46.3)	0.83	0.19		0.84	0.13		1908.2	3123.3	
Yes	22 (51.2)	0.85	0.08		0.86	0.11		22,857.0	51,746.1	
Algorithm used				.0993			.1258			.1417
AI/neural network	7 (15.2)	0.88	0.06		0.75	0.10		1515.5	2314.7	
Deep learning model/deep neural network	20 (43.5)	0.90	0.16		0.89	0.12		30,580.8	57,322.5	
Fuzzy	8 (17.4)	0.77	0.06		0.79	0.10		2728.8	2770.2	
Other	11 (23.9)	0.89	0.06		0.90	0.12		1332.4	3087.0	
Abbreviation: AI, artificial intelligence.										