See related letters on pgs 2275 and 2277

# Automated Classification of Skin Lesions: From Pixels to Practice

Akhila Narla[1], Brett Kuprel[2], Kavita Sarin[3], Roberto Novoa[4,5] and Justin Ko[3,5]

**The letters "Interpretation of the Outputs of Deep Learning Model trained with Skin Cancer Dataset" and "Automated Dermatological Diagnosis: Hype or Reality?" highlight the opportunities, hurdles, and possible pitfalls with the development of tools that allow for automated skin lesion classification. The potential clinical impact of these advances relies on their scalability, accuracy, and generalizability across a range of diagnostic scenarios.**

*Journal of Investigative Dermatology* (2018) **138,** 2108–2110. doi:10.1016/j.jid.2018.06.175

As researchers and clinicians delve into the medical applications of artificial intelligence (AI) and develop deep learning-based tools, dermatology's visually oriented tasks stand out as ripe for innovation. Both providers and patients have ready access to the tissue of interest, and with their smartphones, they possess the imaging devices needed to collect data at scale. We have seen a number of recent advances, including the work of Han et al. (2018), on the automated skin lesion classification tool, "ModelDerm." The dermatological applications of AI hold both opportunities and pitfalls as we cross from "pixels to practice," deploying these tools across diverse patient populations.

## Contextual learning in lesion classification

A robust AI system of automated solitary lesion classification may be feasible for clinical integration and can augment clinical practice. However, the greatest utility would come from a one-system contextual learning model for multiple tasks that accommodates multimodal input, detection of changes in lesions, and comparison of lesions to others on the body. Without multilesion change detection and classification capability, consumer-facing technology runs the risk of reassuring a hypothetical patient about the lentigo on her arm, while missing the melanoma on her leg. Lesion classification can also benefit from multimodal inputs such as age, gender, race, location on the body, or examples of other lesions on the body.

A one-system model may be capable of answering a number of clinical questions across a breadth of dermatological diseases, beyond the binary classification of benign versus malignant (Esteva et al., 2017), whereas from a logistical and usability perspective, it may be suboptimal to have a different model for each skin type or clinical classification task. Multiple models may worsen the performance of the algorithm on "edge" cases, such as patients with intermediate skin types or background skin disease (i.e., a patient with extensive psoriasis and squamous cell carcinoma). Furthermore, a task-specific classifier may fall short if mistakes are made in selecting the initial task. An AI model for dermatology that generalizes between multiple tasks and transfers abilities to new tasks using prior experience with similar tasks would also be the most clinically functional and streamlined solution. Although there may be fewer data for some skin types, we hypothesize that the benefits of training a joint model with sufficient data may outweigh its limitations.

## Nonstandardized and standardized input in AI classification

Making an artificially intelligent system robust enough to handle the variation inherent in image input also poses a hurdle, yet supports the technology's potential for scalability. Dermatology images are the easiest to capture of all medical images, but also the least standardized. Standardization of images is difficult, even with dermoscopic images, as shown in Figure 1. Variability must be incorporated into training algorithms to create capacity to handle noisy data. This includes multiple camera angles, different orientations, blurry photos, multiple skin backgrounds, pen markings or rulers included in the photo, or variations in lighting. Otherwise, the algorithm will use features of nonstandardized photos to guide decision making. For instance, in our work, we noted that the algorithm appeared more likely to interpret images with rulers as malignant. Why? In our dataset, images with rulers were more likely to be malignant; thus the algorithm inadvertently "learned" that rulers are malignant. These biases in AI models are inherent unless specific attention is paid to address inputs with variability. An alternative approach is to incorporate stringent standards and/or hardware that allows for standardization of photos, at the cost of decreased potential for scalability.

Unanswered questions remain with "wide-open" image collection: (1) How can we define the safe limits for analysis given the input—for example, how do we know that a blurry image or dark image is too dark or too blurry? (2) Can we address the issue of "fidelity," whereby analysis of images of the same lesion, for example, rotated or oriented differently, or with different angle, zoom, or brightness, outputs a stable result? One example of a hybrid intermediate approach might be to incorporate an initial algorithm that determines

[1]*Stanford School of Medicine, Stanford University, Stanford, California, USA;* [2]*Department of Electrical Engineering, Stanford University, Stanford, California, USA;* [3]*Department of Dermatology, Stanford University, Stanford, California, USA; and* [4]*Department of Pathology, Stanford University, Stanford, California, USA*

[5] *The last two authors share senior authorship.*

*Correspondence: Justin Ko, Department of Dermatology, Stanford University, 440 Broadway Street Pavilion C 2F, Redwood City, CA USA 94063. E-mail: jmko@stanford.edu*

© 2018 The Authors. Published by Elsevier, Inc. on behalf of the Society for Investigative Dermatology.

## Clinical Implications

- Specific characteristics of automated skin lesion detection make technology valuable and scalable.
- Ideal clinical use would accommodate both standardized and nonstandardized photos.
- A one-system model decision support tool needs clinical validation and robust training.

the suitability of an image for analysis, much as modern banking smartphone apps detect the presence of a check in a picture before depositing it.

**Models with robust training and clinical validation: When do we make it public?**

When incorporating AI tools in the clinic, they should aim to eliminate biases with inclusion of training images and data from a breadth of practice and patients. Subsequent prospective clinical validation demonstrating the robustness of a tool to handle the variability inherent in data input and diversity of patient population is necessary before public adoption for a role in medical decision augmentation. Without this stringency, there exists the potential for unintended effects and inherent bias to exist within these tools, as demonstrated in recent work highlighting racial and gender bias inherent in facial recognition systems (Buolamwini and Gebru, 2018). When assessing the clinical relevance and validity of an algorithm, area under the curve is ideal for validating clinical performance of binary classifiers (Ling et al., 2003). Area under the curve reports across every threshold and every prevalence. In our model (Esteva et al., 2017), we tailored the AI system to answer any number of clinical questions because we looked across the breadth of dermatological disease rather than just training on a certain number of diagnoses, so multiple areas under the curve were able to tell us how the tool would perform if a user were to freely input a skin lesion.

One of the complexities of AI research in a connected world is balancing the risks and benefits of making publicly available an algorithm that has not been validated prospectively. A publicly

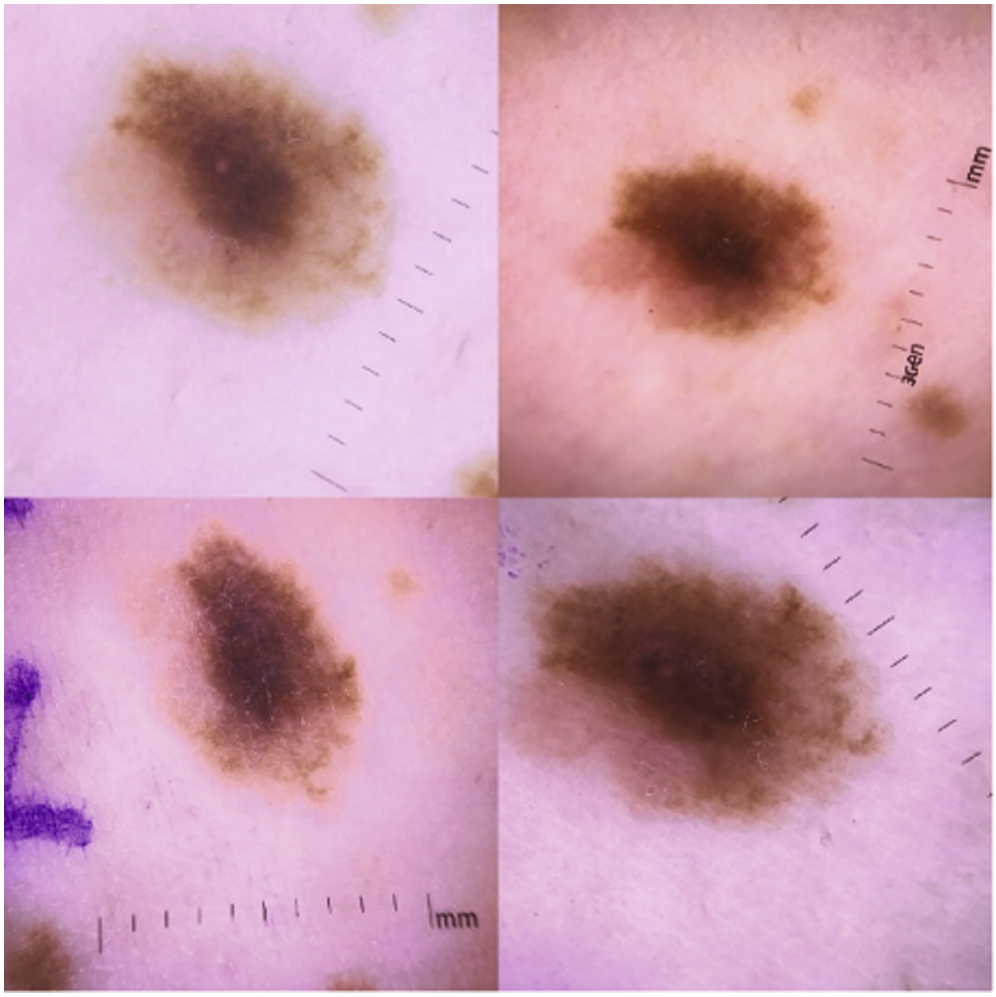


**Figure 1.** Variation even within standardized dermatoscope photos tracking the lesion of a patient over time in the Stanford prospective clinical trial "Change in Cutaneous Lesion Detection using Image Capture and Machine Learning."

available algorithm engages a broader community, allowing for rapid assessment and evaluation of both its capacity to handle varied data and its biases. The posited risk lies in creating the potential for an AI tool to reach or be inappropriately wielded by or for an individual who may be falsely reassured, alarmed, or otherwise use the tool in an unintended manner. Disclaimers regarding the investigative nature of a tool certainly are one approach and may satisfy some as sufficient. Another option is granting exclusive access to researchers and clinicians for a certain period before a tool is released to the general public. Such an approach may help balance the rich and fruitful discussion that comes from open-source engagement with such a tool with the safety of the public.

## Conclusion

Use of deep neural networks in the classification of skin lesions has made rapid progress in recent years while simultaneously generating significant "hype." As we continue improving on these models and building high-quality databases with increased data input modalities, we must continue addressing the many nuances involved in bringing these new technologies safely to the bedside.

## CONFLICT OF INTEREST

The authors state no conflict of interest.

## REFERENCES

Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness. Accountability and Transparency 21 January 2018. p. 77–91.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115.

Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018;138:1529–38.

Ling CX, Huang J, Zhang H. AUC: a better measure than accuracy in comparing learning algorithms. In: Conference of the Canadian Society for Computational Studies of Intelligence. Berlin, Heidelberg: Springer; 11 June 2003. p. 329–41.