Deep learning-aided decision support for diagnosis of skin disease across skin tones

Matt Groh[1,2], Omar Badri[3], Roxana Daneshjou[4], Arash Koochek[5], Caleb Harris[2], Luis Soenksen[6], P. Murali Doraiswamy[2,7], Rosalind Picard[2]

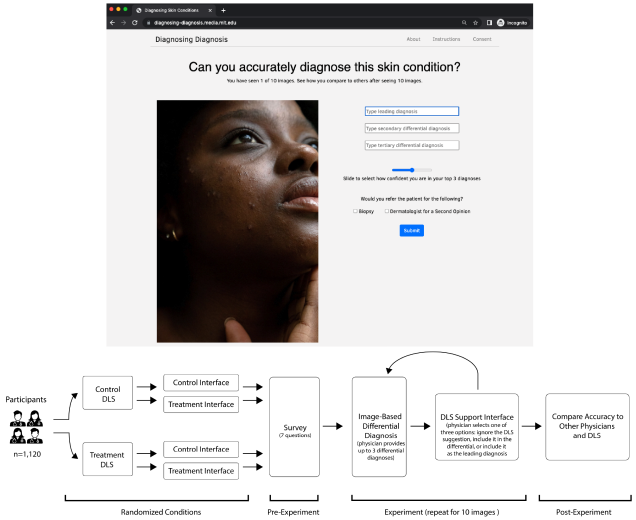
KELLOGG SCHOOL OF MANAGEMENT

Abstract

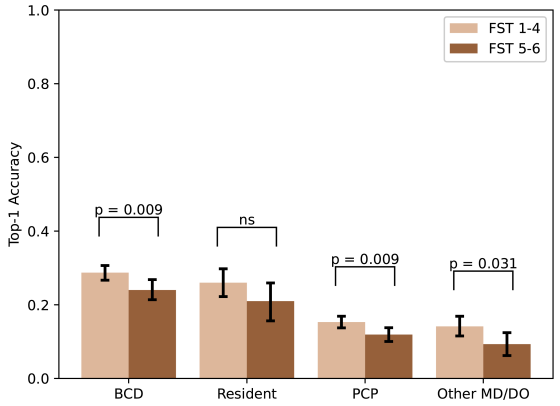
How accurately do physicians assisted by deep learning systems diagnose skin disease? How does accuracy vary across (1) patients' skin color, (2) physicians' expertise, (3) type of disease, (4) access to AI assistance, (5) quality of AI assistance, and (6) design of the decision support interface? How quickly do physicians adapt to different levels of AI assistance? Here, we design a custom, digital experiment to evaluate these questions focusing on physicians' diagnostic accuracy on images of inflammatory appearing skin diseases.

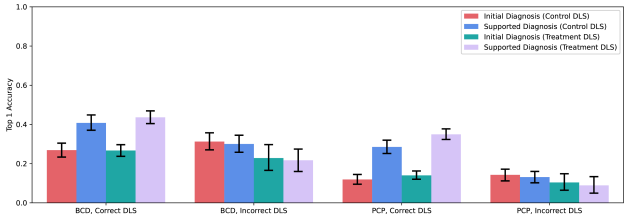
This image-based experimental setup mimics store-and-forward teledermatology and the kinds of patient images that physicians are sent through electronic health record messaging systems, which often have minimal clinical context. The experiment contains 364 images of 46 skin diseases where the vast majority (78%) are 8 main diseases. The images are nearly uniform in their distribution across patient skin color.

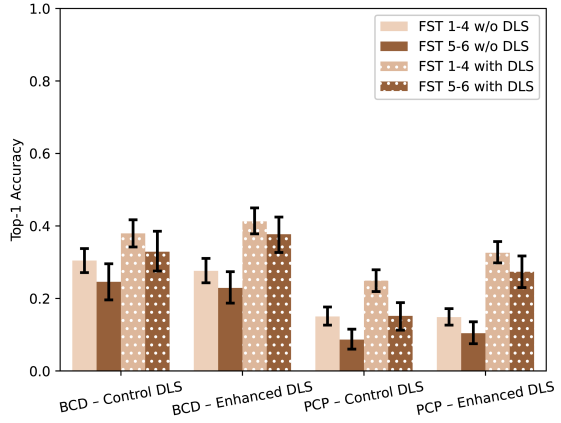
Experimental Design



Results







Discussion

- Diagnostic accuracy disparities appear across skin color

Board-certified dermatologists' (BCDs) diagnostic accuracy is 10% lower and primary care physicians' (PCPs) diagnostic accuracy is 22% lower on dark skin than light. For CTCL (a life threatening disease), we find both BCDs and PCPs report that they would refer patients for biopsy significantly more often in light skin than dark skin.

- AI assistance improves diagnostic accuracy

We find the DLS-based decision support increases top-1 diagnostic accuracy by 33% for BCDs and 69% for PCPs *($p<0.001$)* . This translates into improved sensitivity in diagnosing skin diseases with minimal effects on specificity; for example, we find specialists' sensitivity for diagnosing cutaneous t-cell lymphoma increases by 44% with control DLS assistance and 72% with treatment DLS assistance while specialists' specificity remains constant.

- AI assistance exacerbates accuracy disparities in generalists but not specialists

One potential explanation for the magnification of diagnostic accuracy disparities in generalists (despite overall improved accuracy) may be related to the nature of the DLS prompting physicians to consider alternatives that they cannot rule out and generalists' differential knowledge of what can and cannot be ruled out in dark skin.

- Physicians quickly adapt to quality of AI assistance

Physicians assigned to the treatment (more accurate) DLS system are 7 percentage points *($p<0.01$)* more likely to incorporate AI assistance into their diagnosis on the exact same image

- Images contain information but information is limited

We find the most common leading diagnosis for each image by BCDs is correct in 48% of observations. At least one BCD identified the reference label in their differential diagnosis in 77% of images. A single image contains significantly less information than an in-person interaction (or even a video call), which include adjustments in light and angle of view), a patient's symptoms, clinical history, behavioral information, and more.

Affiliations: 1 Northwestern, 2 MIT, 3 Northeast Dermatology Associates, 4 Stanford, 5 Banner Health, 6 Harvard, 7 Duke