Improving dermatology classifiers across populations using images generated by large diffusion models

Luke W. Sagers*
Department of Biomedical Informatics
Harvard Medical School
luke_sagers@hms.harvard.edu

James A. Diao*
Department of Biomedical Informatics
Harvard Medical School
james_diao@hms.harvard.edu

Matthew Groh
MIT Media Lab
Massachussetts Institute of Technology
groh@mit.edu

Pranav Rajpurkar
Department of Biomedical Informatics
Harvard Medical School
pranav_rajpurkar@hms.harvard.edu

Adewole S. Adamson
Department of Internal Medicine
UT Austin Dell Medical School
adewole.adamson@austin.utexas.edu

Arjun K. Manrai
Department of Biomedical Informatics
Harvard Medical School
arjun_manrai@hms.harvard.edu

1 Introduction

Skin classification algorithms developed without sufficiently diverse training data may generalize poorly across populations. While intentional data collection and annotation offer the best means for improving representation [1], new computational approaches for generating training data may also aid in mitigating sampling bias. In this study, we show that the text-to-image diffusion model DALL·E 2 [4] can produce photorealistic images of skin disease across skin types, and that using these synthetic images to supplement training data can improve classification of skin disease overall and especially for underrepresented groups.

2 Methods

Dermatologic images with accompanying diagnostic and Fitzpatrick skin type (FST) labels are derived from the Fitzpatrick 17k dataset , comprising 16,577 images (Table 1) [3]. For each condition, we sampled 8 images from the lightest and darkest skin types as seed images. We utilized DALL·E 2's 'inpainting' function to isolate the primary pathology and surrounding skin. This was paired with a structured text prompt to generate synthetic images. From each seed image, 4 synthetic images were selected for photorealism and pathophysiologic consistency (Figure 2).

We trained deep learning models using a VGG16 architecture pre-trained on ImageNet to predict skin condition labels among seven skin conditions. The model was trained using Adam optimization, weighted random sampling to address class imbalance, and data augmentation with standard image perturbations. Initial models were trained on images of the lightest skin types and tested on images of darker skin types, and vice versa. Training sets were augmented with images from the opposite FST group: either 8 seed images only, or 8 seed images and 32 synthetic images. We compared model developed using these separate training data and assessed the impact of successively increasing the number of synthetic images (from 2 to 32) to assess for a dose-response relationship.

*Equal contribution

Submitted to the 2023 Symposium on Artificial Intelligence for Learning Health Systems.

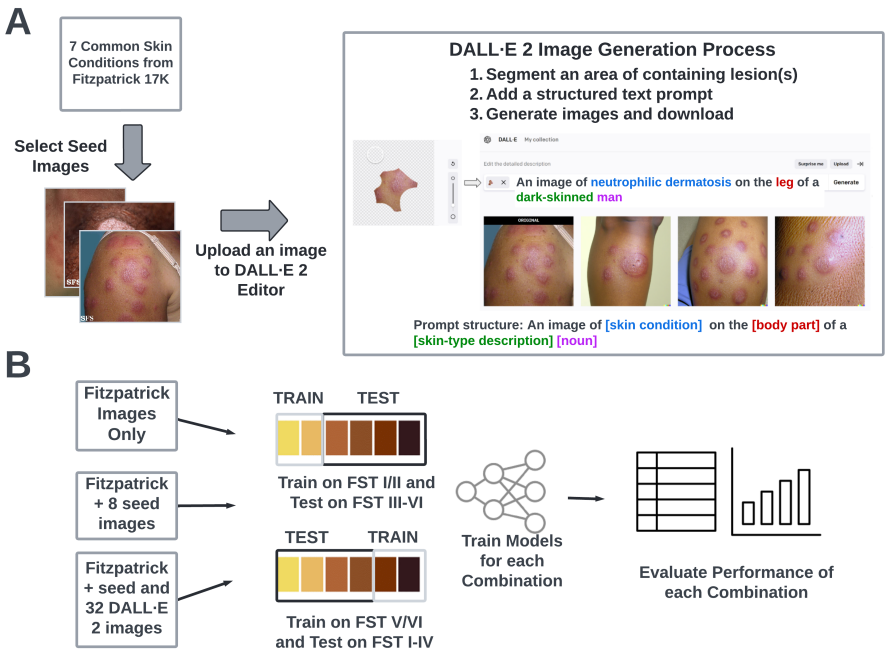

Figure 1: **A schematic overview of the study A)** For each skin condition, we randomly sampled 8 images each from the lightest and darkest skin types. These images were used as seed images for OpenAI's DALL·E 2 model to produce synthetic variations. **B)** Deep learning models for skin disease classification were trained using lighter skin types and tested using darker skin types, and vice versa. Experiments compared models trained using: (1) Fitzpatrick 17k images only, (2) Fitzpatrick 17k + seed images, and (3) Fitzpatrick 17k + seed images + DALL·E 2 generated synthetic images.

3 Results

Models trained on light skin performed poorly on dark skin and vice versa (Figure 3). For example, a model trained on images of neutrophilic dermatosis in light skin exhibited lower performance for the darkest skin types than for intermediate skin types: 24.3% vs. 31.1%. Performance generally improved when training was augmented with seed images from unrepresented skin types and improved further when additionally augmented with synthetic images, although not all conditions were powered to detect significant differences. Successive addition of synthetic images was associated with successive performance improvements, suggesting a dose-response effect (Table 2).

4 Discussion

We present a proof-of-concept that data augmentation using photorealistic synthetic images of skin pathologies may improve performance across diverse populations. The results extend upon prior work leveraging deep generative adversarial networks (GANs), style transfer, deep blending, and other generative methods [2, 5]. Limitations include the limited number of assessed skin conditions and manual involvement in image generation. Follow-up work may compare diffusion models with previous approaches, directly quantify photorealism, or investigate uses in other underrepresented domains. While collection of diverse real-world data remains the most important step for improving skin classification models, we believe that the concomitant use of synthetic data may act as a force-multiplier to continually improve classification models for skin pathology.

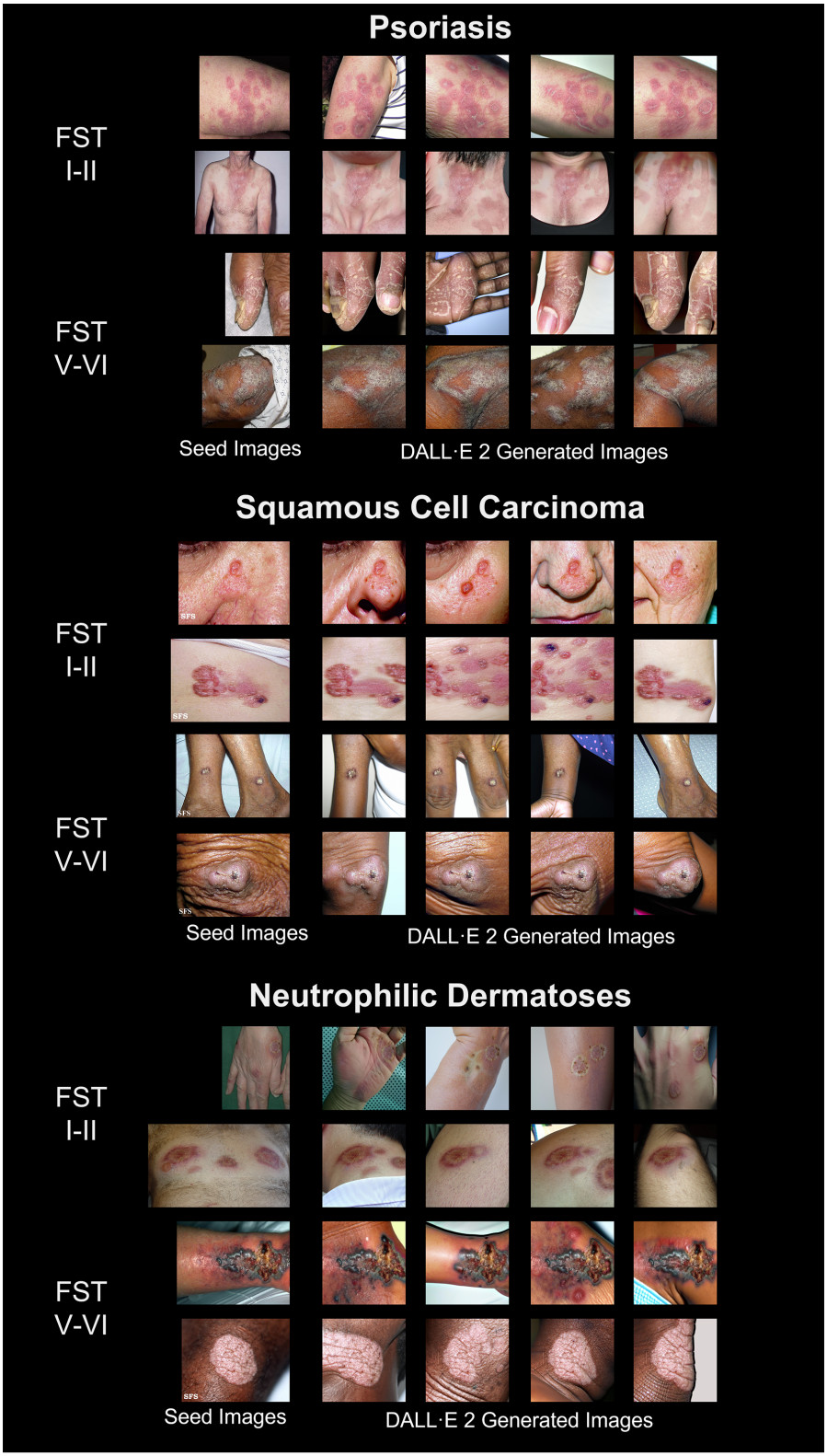

Figure 2: **Examples of DALL·E 2-generated synthetic images** We generated synthetic images for 3 conditions: psoriasis, squamous cell carcinoma, and neutrophilic dermatoses. For each seed image (left), 4 synthetic images are shown (right). A full table of text prompts used in these image generations can be found in Supplementary Table 2.

Table 1: Sample sizes of seven skin conditions analyzed in this study, by Fitzpatrick skin type

FST	Basal Cell Carcinoma	Folliculitis	Nematode Infection	Neutrophilic Dermatoses	Prurigo Nodularis	Psoriasis	Squamous Cell Carcinoma	*Total*
I	85	30	15	70	7	113	100	*420*
II	156	97	56	115	28	232	180	*864*
III	112	99	79	68	39	101	122	*620*
IV	76	51	60	51	56	91	71	*456*
V	24	31	32	31	29	64	40	*251*
VI	7	9	12	15	9	21	23	*96*
Total	*460*	*317*	*254*	*350*	*168*	*622*	*536*	***2707***

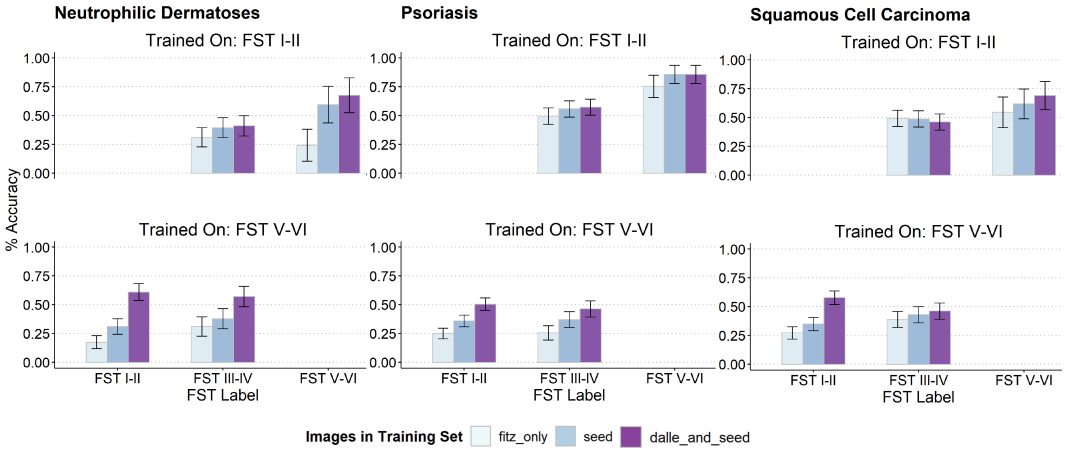

Figure 3: **Model accuracy for different real and synthetic training datasets** Models were trained on a subset of skin types (e.g. I-II) and tested on the remainder (e.g. III-IV & V-VI). Color labels represent which images were included in training. "fitz_only" includes only original images from the Fitzpatrick 17K dataset. "seed" includes the original images plus 8 seed images that were removed from the test set and used in the image generation process. "dalle_and_seed" includes the original images plus both seed and synthetic images.

Table 2: Classification accuracy by skin type with successive addition of synthetic training images

Classification accuracy by number of added synthetic training images					
FST	**+2 images**	**+8 images**	**+16 images**	**+32 images**	**N**
		Neutrophilic Dermatoses			
III-IV	0.37	0.34	0.39	0.45	119
V-VI	0.29	0.29	0.58	0.68	37
		Psoriasis			
III-IV	0.49	0.56	0.56	0.51	192
V-VI	0.78	0.82	0.83	0.86	77
		Squamous Cell Carcinoma			
III-IV	0.44	0.43	0.50	0.46	193
V-VI	0.56	0.53	0.60	0.69	55

References

[1] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong, R. Sahasrabudhe, J. A. C. Allerup, U. Okata-Karigane, J. Zou, and A. S. Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv*, 8(32):eabq6147, Aug. 2022.

[2] A. Ghorbani, V. Natarajan, D. Coz, and Y. Liu. DermGAN: Synthetic generation of clinical skin images with pathology. Nov. 2019.

[3] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. Apr. 2021.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical Text-Conditional image generation with CLIP latents. Apr. 2022.

[5] E. Rezk, M. Eltorki, and W. El-Dakhakhni. Improving skin color diversity in cancer detection: Deep learning approach. *JMIR Dermatology*, 5(3):e39143, Aug. 2022.