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

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Dermatological classification algorithms developed without sufficiently diverse training data may generalize poorly across populations. While more intentional data collection and annotation is the best way to increase representation, new computational approaches for generating training data may also aid in reducing representation bias. In this paper, we show that DALL·E 2, a large text-to-image diffusion model, can generate synthetic and photorealistic skin disease images across skin types. Using the Fitzpatrick 17k dataset as a benchmark, we demonstrate that including DALL·E 2-generated synthetic images improves classification accuracy of skin disease models overall and particularly for underrepresented groups.

## 1 Introduction

Skin disease classification algorithms based on modern machine learning algorithms are now entering clinical and commercial use [3, 7, 13]. Although these algorithms have demonstrated results on par with those of board-certified dermatologists, there remain concerns about their use given the limited representation of darker skin tones in development datasets, which may negatively impact their performance across diverse populations [1].

In response, several groups have collected and released benchmark data repositories with data from more diverse populations as well as skin tone annotations that may be used to study and improve bias in classification models [5, 2]. However, even with intentional collection, it is difficult to achieve parity in representation or performance for underrepresented groups, particularly across the dozens or hundreds of skin conditions that may be included as possible prediction outputs.

Recent breakthroughs in large-scale diffusion models [11, 9, 12] have introduced the capability to generate photorealistic images based on text alone, or text and image inputs in combination. Such models may enable the generation of synthetic data to improve prediction [8], especially for underrepresented disease classes and skin tones. In this paper, we develop a DALL·E 2-based pipeline for generating synthetic photorealistic images. We show that targeted generation of synthetic images by DALL·E 2 can be used to improve the performance for dermatological classifiers on a benchmark dataset overall and particularly for underrepresented groups.

## 2 Methods

### 2.1 The Fitzpatrick 17k Dataset

Images of skin conditions with accompanying diagnostic and skin tone labels are derived from the Fitzpatrick 17k dataset [5]. These data sourced thousands of images from two online dermatology atlases—DermaAmin and Atlas Dermatalogico—and assigned Fitzpatrick skin type (FST) labels

Submitted to NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research. Do not distribute.

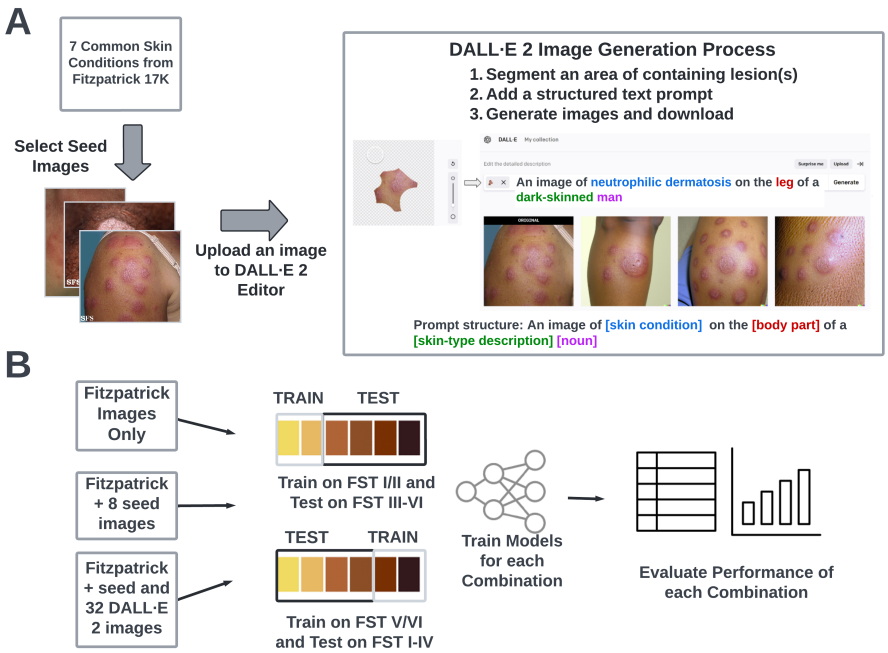


Figure 1: **A schematic overview of the study A)** We selected seven skin conditions from the Fitzpatrick 17k dataset [5]. For each skin condition, we randomly sampled eight images from the lightest and darkest Fitzpatrick skin types (16 total images for one condition) and used these images as seed images in the editor of OpenAI's DALL·E 2 model. From these groups of eight images, DALL·E 2 generated synthetic variations. We selected four DALL·E 2-synthetic images per seed image for inclusion in the subsequent analysis. **B)** We trained image classification models to predict skin condition labels using different train/test splits. Models were trained on skin images from lighter skin types (FST I-II) and tested on images with darker skin types (FST V-VI), and vice versa. We ran experiments where training data included either (1) Fitzpatrick 17k images only, (2) Fitzpatrick 17K + seed images, and (3) Fitzpatrick 17k + seed images + DALL·E 2 generated synthetic images.

to each image using a dynamic consensus process involving 2-5 annotators from Scale AI. The label error rate was estimated at 3.4% relative to board-certified dermatologists. An independent set of annotations were later assigned by a separate team from Centaur Labs [6]. This produced a final dataset comprising 16,577 clinical images. Together, these images represented 114 of the most common dermatology conditions, with a minimum of 53 images per condition.

### 2.2 Selecting skin conditions for retraining models with DALL·E 2-generated synthetic images

We selected a subset of seven disease labels from the 114 available in the Fitzpatrick 17k dataset: basal cell carcinoma, folliculitis, neutrophilic dermatoses, prurigo nodularis, squamous cell carcinoma, nematode infection, and psoriasis. These seven conditions were selected as follows: (1) first, we reproduced the analysis by Groh et al. [5] across all 114 conditions in the Fitzpatrick 17k dataset; (2) second, we identified skin conditions that had both the largest sample size at the extremes of FST (I-II or V-VI) and non-zero accuracy when tested on conditions unrepresented in model training. Table 1 lists the sample sizes of each skin condition across FST labels in the dataset.

### 2.3 DALL·E 2 photorealistic data augmentation pipeline

We generated photorealistic synthetic images from seed images in the Fitzpatrick 17k dataset using OpenAI's DALL·E 2 model with the following workflow (Figure 1A). For a given skin condition, we randomly sampled eight images from the lightest and darkest Fitzpatrick skin types (16 total images for one condition) to use as seed images. We cropped each seed image with square dimensions

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*** |

centered around the disease pathology. We then utilized DALL·E 2 editor's 'inpainting' function to isolate the primary dermatologic pathology and surrounding skin from background artifacts. The resulting image was used alongside a text prompt to generate synthetic images. Text prompts were produced using the following template: "An image of [skin condition] on the [body part] of a [skin type description] [noun]." All text prompts are listed in Supplementary Table 2. We created eight (on average) synthetic images from each image-text pair, from which we chose four to download and include in our DALL·E 2 training sets. The four images were chosen for photorealism and pathophysiologic consistency. Examples of seed and selected synthetic images are shown in Figure 2. Examples of unselected synthetic images are shown in Supplementary Figure 1.

### 2.4 Model training and evaluation

We trained image classification models to predict skin condition labels among the seven skin conditions using different train and test splits. Following Groh et al. [5], initial models were trained on images of the lightest skin types (FST I-II) and tested on images of darker skin types (FST III-IV and V-VI). The model was also trained on images of the darkest skin types (FST V-VI) and tested on images of lighter skin types (FST I-II and III-IV). These accuracy metrics were used as baselines for comparison against alternate training procedures (Figure 3).

For three skin conditions (squamous cell carcinoma, psoriasis, and neutrophilic dermatoses) we created training sets that included either FST I-II or FST V-VI images from Fitzpatrick 17k supplemented with either 8 seed images from the opposite FST group or both the 8 seed images and 32 DALL·E 2 generated images.

Seed images were not included in any test set. The model architecture and training pipeline, comprising a VGG16 network pre-trained on ImageNet, follows the same structure provided by Groh et al. [5]. Training was performed using Adam optimization and with a weighted random sampler to address class imbalance across skin conditions. Standard image transformations and normalization was applied at the time of training.

### 2.5 Ablation Studies

We measured model accuracy on all skin conditions in the dataset in order to see how prediction accuracies change for other skin conditions when synthetic images from one skin condition are added to the training data. Additionally, we re-ran the models that included synthetic images while changing the number of synthetic images in the training data. We compared the accuracies of models trained with a random sample of 2, 8, and 16 DALL·E 2 generated images of three skin conditions from the lightest and darkest skin types.

## 3 Results

### 3.1 Dermatological classifiers can generalize poorly to populations underrepresented in the training data

We observed that models trained using data from one end of the Fitzpatrick skin-type (FST) scale may exhibit worse performance on skin-types on the opposite end of the FST scale (Table 2). An example of this is seen for neutrophilic dermatoses, where a model trained on images with the lightest FST labels (I-II), exhibited worse performance for the darkest skin types (prediction accuracy for

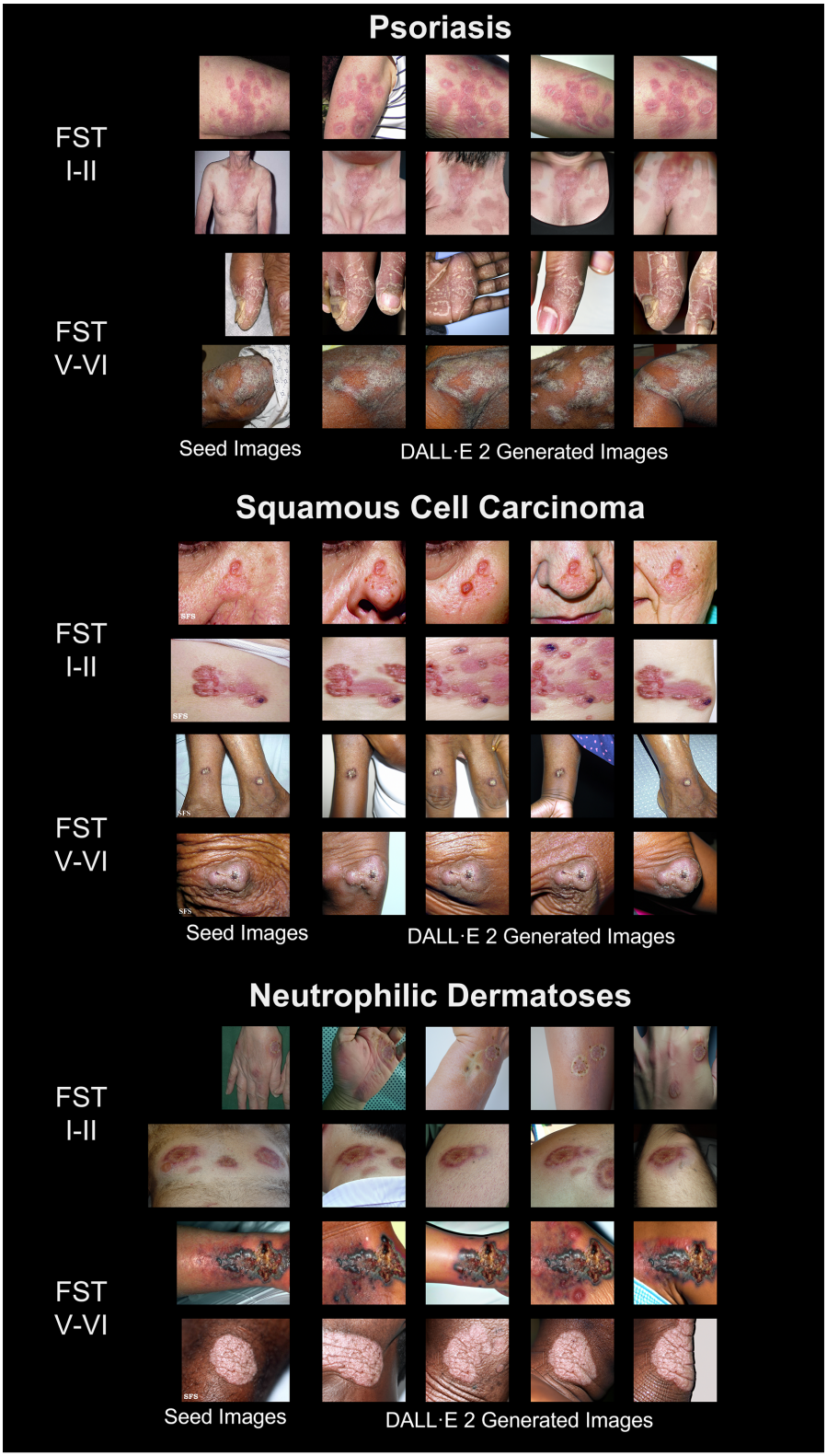


Figure 2: **Examples of DALL·E 2 generated synthetic images for three skin conditions** We used DALL·E 2 to generate synthetic images for three conditions: psoriasis, squamous cell carcinoma, and neutrophilic dermatoses. For each seed image (left), four synthetic images are shown (right). A full table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

Table 2: Model performance across different real and synthetic training datasets

| **Neutrophilic Dermatoses: Classification Accuracy (95% CI), N** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Trained on FST I-II* | | | *Trained on FST V-VI* | | | |
| **FST** | N | Fitzpatrick | +Seed | +DALLE & Seed | Fitzpatrick | +Seed | +DALLE & Seed | N |
| **I-II** | x | x | x | x | 0.175 (0.12-0.23) | 0.310 (0.24-0.38) | 0.61 (0.54-0.68) | 177 |
| **III-IV** | 119 | 0.311 (0.23-0.39) | 0.395 (0.31-0.48) | 0.411 (0.32-0.5) | 0.311 (0.23-0.39) | 0.378 (0.29-0.47) | 0.571 (0.48-0.66) | 119 |
| **V-VI** | 37 | 0.243 (0.1-0.38) | 0.594 (0.44-0.75) | 0.675 (0.52-0.83) | x | x | x | x |

| **Psoriasis: Classification Accuracy (95% CI), N** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Trained on FST I-II* | | | *Trained on FST V-VI* | | | |
| **FST** | N | Fitzpatrick | +Seed | +DALLE & Seed | Fitzpatrick | +Seed | +DALLE & Seed | N |
| **I-II** | x | x | x | x | 0.249 (0.20-0.29) | 0.359 (0.31-0.41) | 0.504 (0.45-0.55) | 337 |
| **III-IV** | 192 | 0.495 (0.42-0.57) | 0.557 (0.48-0.63) | 0.573 (0.50-0.64) | 0.255 (0.19-0.31) | 0.370 (0.30-0.44) | 0.463 (0.39-0.53) | 192 |
| **V-VI** | 77 | 0.753 (0.66-0.85) | 0.857 (0.78-0.94) | 0.857 (0.78-0.94) | x | x | x | x |

| **Squamous Cell Carcinoma: Classification Accuracy (95% CI), N** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | *Trained on FST I-II* | | | *Trained on FST V-VI* | | | |
| **FST** | N | Fitzpatrick | +Seed | +DALLE & Seed | Fitzpatrick | +Seed | +DALLE & Seed | N |
| **I-II** | x | x | x | x | 0.272 (0.22-0.32) | 0.349 (0.29-0.41) | 0.577 (0.52-0.64) | 272 |
| **III-IV** | 193 | 0.492 (0.42-0.56) | 0.487 (0.42-0.56) | 0.461 (0.39-0.53) | 0.389 (0.32-0.46) | 0.430 (0.36-0.5) | 0.461 (0.39-0.53) | 193 |
| **V-VI** | 55 | 0.545 (0.41-0.68) | 0.618 (0.49-0.75) | 0.691 (0.57-0.81) | x | x | x | x |

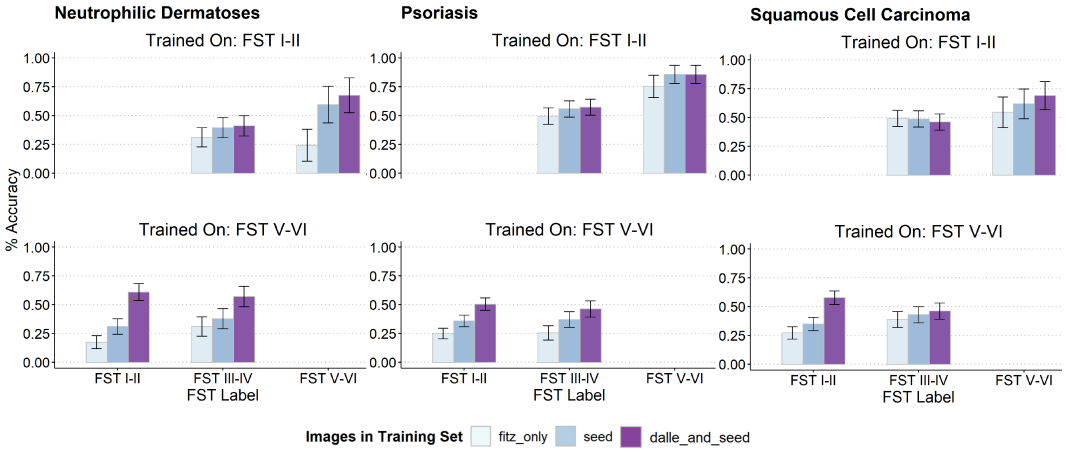


Figure 3: **Model accuracy across different subsets of train and test images** Results are shown for three disease labels across models. Each bar represents the performance of a model trained on a subset of Fitzpatrick skin types (e.g. I-II) and tested on the rest of the skin types (e.g. III-IV & V-VI). The 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 the 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 synthetic and seed images.

FST V-VI: 24.3%) than for the intermediate skin types (prediction accuracy for FST III-IV: 31.1%). This trend was observed for both models trained on light and tested on dark skin types as well as for models trained on dark and tested on light skin types. In squamous cell carcinoma, for instance, models trained on the darkest skin types (FST V-VI) performed worse on the lightest skin types (prediction accuracy for FST I-II: 27.2%) than on the intermediate skin types (prediction accuracy for FST III-IV: 38.9%). In other instances, the trend was reversed, as in psoriasis, where model performance was better in the darkest FST labels (prediction accuracy for FST V-VI: 75.3%) than in the intermediate labels (prediction accuracy for FST III-IV: 49.5%) when the model was trained on images from the lightest FST groups (FST I-II).

### 3.2 Improved performance for skin disease classification using DALL·E 2 generated synthetic images

Model performance generally improved when training was supplemented by seed images from unrepresented FST labels and improved further when additionally supplemented by synthetic images generated by DALL·E 2 (Figure 3), although substantial imprecision was observed for conditions

Table 3: Classification accuracy across subgroups when different numbers of synthetic images are added to the training data.

| **Neutrophilic Dermatoses: Classification Accuracy** | | | | | |
|---|---|---|---|---|---|
| | *Number of DALL·E 2 Images* | | | | |
| | +2 images | +8 images | +16 images | +32 images | **N** |
| **FST III-IV** | 0.37 | 0.34 | 0.39 | 0.45 | **119** |
| **FST V-VI** | 0.29 | 0.29 | 0.58 | 0.68 | **37** |
| **Psoriasis: Classification Accuracy** | | | | | |
| | +2 images | +8 images | +16 images | +32 images | **N** |
| **FST III-IV** | 0.49 | 0.56 | 0.56 | 0.51 | **192** |
| **FST V-VI** | 0.78 | 0.82 | 0.83 | 0.86 | **77** |
| **Squamous Cell Carcinoma: Classification Accuracy** | | | | | |
| | +2 images | +8 images | +16 images | +32 images | **N** |
| **FST III-IV** | 0.44 | 0.43 | 0.50 | 0.46 | **193** |
| **FST V-VI** | 0.56 | 0.53 | 0.60 | 0.69 | **55** |

with limited test data. The most substantial improvements occurred for images of skin tones least like those on which the model was trained. An example of this can be seen in squamous cell carcinoma for models trained on FST V-VI (bottom right Figure 3). The difference between the performance of the model trained on Fitzpatrick 17k images only and the model trained with Fitzpatrick 17k images plus synthetic and seed images is 7.2% for the FST III-IV group, whereas the performance difference is 30.5% in the FST I-II group (the group furthest from the FST V-VI images).

### 3.3 Ablation studies

In our first ablation study, when we incrementally added 2, 8, and 16 synthetic images to the models, we generally saw increases in model performance in correspondence to the number of synthetic images added (Table 3).

In a separate analysis, we generally observed modest changes in model accuracy on other skin conditions when synthetic images from one skin condition (generated using images from the held-out FST) were added to the training data. The largest difference in performance occurred in basal cell carcinoma after light-skinned (FST I-II) synthetic images of squamous cell carcinoma were included in the training set that contained only images from FST V-VI. Classification accuracy for basal cell carcinoma in FST I-II decreased from 66.4% (N = 241) to 30.7% and from 68.1% to 50.5% (N = 188) in FST III-IV (Supplementary Table 1).

## 4 Discussion

Development datasets for machine learning models are limited by class imbalances that may limit generalizability, especially to underrepresented groups. Our study presents a proof-of-concept that using DALL·E 2-generated photorealistic synthetic images of dermatologic pathologies may improve performance across diverse populations. It also demonstrates performance improvement in several skin conditions, which had not yet been achieved by prior work leveraging generative adversarial networks [4], style transfer [10], deep blending, and other methods for synthetic data generation.

Limitations include use of a coarse photosensitivity scale to represent skin tone, possible inclusion of Fitzpatrick 17k test data in the DALL·E 2 training data, potential for data leakage from highly similar images in Fitzpatrick 17k, and significant manual involvement in the image generation process. Future work should develop more systematic pipelines for prompting, generating, and selecting synthetic images, which may enable the creation of large, expert-vetted repositories of synthetic data. Future work should also examine the generalizability of synthetic images as training aids for dermatology classifiers by evaluating on additional skin conditions and benchmarks. While collecting diverse real-world data remains the most direct and effective path for improving model training and evaluation, we believe that concomitant use of synthetic data, along with traditional methods of re-weighting and upsampling, may act as a force-multiplier to continually improve classification models for skin pathology.

# References


[1] A. S. Adamson and A. Smith. Machine learning and health care disparities in dermatology. *JAMA Dermatol.*, 154(11):1247–1248, Nov. 2018.

[2] 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.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542 (7639):115–118, Feb. 2017.

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

[5] 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.

[6] M. Groh, C. Harris, R. Daneshjou, O. Badri, and A. Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. July 2022.

[7] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, V. Gupta, N. Singh, V. Natarajan, R. Hofmann-Wellenhof, G. S. Corrado, L. H. Peng, D. R. Webster, D. Ai, S. J. Huang, Y. Liu, R. C. Dunn, and D. Coz. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.*, 26(6):900–908, June 2020.

[8] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. Jorge Cardoso. Brain imaging generation with latent diffusion models. Sept. 2022.

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

[10] 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.

[11] Rombach, Blattmann, Lorenz, and others. High-resolution image synthesis with latent diffusion models. *Proc. Estonian Acad. Sci. Biol. Ecol.*

[12] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. Sara Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic Text-to-Image diffusion models with deep language understanding. May 2022.

[13] L. R. Soenksen, T. Kassis, S. T. Conover, B. Marti-Fuster, J. S. Birkenfeld, J. Tucker-Schwartz, A. Naseem, R. R. Stavert, C. C. Kim, M. M. Senna, J. Avilés-Izquierdo, J. J. Collins, R. Barzilay, and M. L. Gray. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Sci. Transl. Med.*, 13(581), Feb. 2021.