# Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset

Matthew Groh
MIT Media Lab
Cambridge, MA
groh@mit.edu

Caleb Harris
MIT Media Lab
Cambridge, MA
harrisc@mit.edu

Luis Soenksen
MIT, Harvard University
Cambridge, MA
soenksen@mit.edu

Felix Lau
Scale
San Francisco, CA
felix.lau@scale.com

Rachel Han
Scale
San Francisco, CA
rachel.han@scale.com

Aerin Kim
Scale
San Francisco, CA
aerin.kim@scale.com

Arash Koochek
Banner Health
Phoenix, AZ
arash.koochek@bannerhealth.com

Omar Badri
Northeast Dermatology Associates
Beverly, MA
obadri@gmail.com

## Abstract

*How does the accuracy of deep neural network models trained to classify clinical images of skin conditions vary across skin color? While recent studies demonstrate computer vision models can serve as a useful decision support tool in healthcare and provide dermatologist-level classification on a number of specific tasks, darker skin is underrepresented in the data. Most publicly available data sets do not include Fitzpatrick skin type labels. We annotate 16,577 clinical images sourced from two dermatology atlases with Fitzpatrick skin type labels and open-source these annotations. Based on these labels, we find that there are significantly more images of light skin types than dark skin types in this dataset. We train a deep neural network model to classify 114 skin conditions and find that the model is most accurate on skin types similar to those it was trained on. In addition, we evaluate how an algorithmic approach to identifying skin tones, individual typology angle, compares with Fitzpatrick skin type labels annotated by a team of human labelers.*

## 1. Motivation

How does the accuracy of deep neural network models trained to classify clinical images of skin conditions vary across skin color? The emergence of deep neural network models that can accurately classify images of skin conditions presents an opportunity to improve dermatology and healthcare at large [23, 36, 51, 48, 12]. But, the data upon which these models are trained are mostly made up of images of people with light skin. In the United States, dark skin is underrepresented in dermatology residency programs [35], textbooks [5, 3], dermatology research [34], and dermatology diagnoses [43, 28]. With the exception of PAD-UFES-20 [42], none of the publicly available data sets identified by the Sixth ISIC Skin Image Analysis Workshop at CVPR 2021 (Derm7pt [30], Dermofit Image Library, ISIC 2018 [16, 52], ISIC 2019 [15, 52, 18], ISIC 2020[46, 17], MED-NODE [27], PH2 [37], SD-128 [49], SD-198, SD-260) include skin type or skin color labels or any other information related to race and ethnicity. The only dataset with such skin type labels, PAD-UFES-20, contains Fitzpatrick skin type labels for 579 out of 1,373 patients in the dataset. The lack of consideration of subgroups within a population has been shown to lead deep neural networks to produce large accuracy disparities across gender and skin color for facial recognition [11], across images with and without surgical markings in dermatology [56, 10], and across treated and untreated conditions in radiology [40]. These inaccuracies arise from dataset biases, and these underlying data biases can unexpectedly lead to systematic bias against groups of people [8, 1]. If these dataset biases are left unexamined in dermatology images, machine learning models have the potential to increase healthcare disparities in dermatology [2].

By creating transparency and explicitly identifying likely sources of bias, it is possible to develop machine learning models that are not only accurate but also serve as discrimination detectors [41, 32, 19]. By rigorously examining potentials for discrimination across the entire pipeline for machine learning model development in healthcare [14], we can identify opportunities to address discrimination such as collecting additional data from underrepresented groups [13] or disentangling the source of the disparities [44]. In this paper, we present the *Fitzpatrick 17k* dataset which is a collection of images from two online dermatology atlases annotated with Fitzpatrick skin types by a team of humans. We train a deep neural network to classify skin conditions solely from images, and we evaluate accuracy across skin types.

We also use the *Fitzpatrick 17k* dataset to compare Fitzpatrick skin type labels to a computational method for estimating skin tone: individual typology angle (ITA). ITA is promising because it can be computed directly from images, but its performance varies with lighting conditions and may not always be effective for accurately annotating clinical images with skin types [55, 33, 31].

## 2. Fitzpatrick 17k Dataset

The *Fitzpatrick 17k* dataset contains 16,577 clinical images with skin condition labels and skin type labels based on the Fitzpatrick scoring system [25]. The dataset is accessible at https://github.com/mattgroh/fitzpatrick17k.

The images are sourced from two online open-source dermatology atlases: 12,672 images from DermaAmin and 3,905 images from Atlas Dermatologico [4, 26]. These sources include images and their corresponding skin condition label. While these labels are not known to be confirmed by a biopsy, these images and their skin condition labels have been used and cited in dermatology and computer vision literature a number of times [23, 29, 9, 45, 6, 50, 53]. As a data quality check, we asked a board-certified dermatologist to evaluate the diagnostic accuracy of 3% of the dataset. Based on a random sample of 504 images, a board-certified dermatologist identified 69.0% of images as diagnostic of the labeled condition, 19.2% of images as potentially diagnostic (not clearly diagnostic but not necessarily mislabeled, further testing would be required), 6.3% as characteristic (resembling the appearance of such a condition but not clearly diagnostic), 3.4% are considered wrongly labeled, and 2.0% are labeled as other. A second board-certified dermatologist also examined this sample of images and confirmed the error rate. This error rate is consistent with the 3.4% average error rate in the most commonly used test datasets for computer vision, natural language processing, and audio processing [39].

We selected images to annotate based on the most common dermatology conditions across these two data sources excluding the following 22 categories of skin conditions: (1) viral diseases, HPV, herpes, molluscum, exanthems, and others (2) fungal infections, (3) bacterial infections, (4) acquired autoimmune bullous disease, (5) mycobacterial infection (6) benign vascular lesions (7) scarring alopecia, (8) non-scarring alopecia (9) keratoderma (10) ichthyosis, (11) vasculitis, (12) pellagra like eruption (13) reiters disease (14) epidermolysis bullosa pruriginosa (15) amyloidosis, (16) pernio and mimics (17) skin metastases of tumours of internal organs (18) erythrokeratodermia progressive symmetric, (19) epidermolytic hyperkeratosis, (20) infections, (21) generalized eruptive histiocytoma, (21) dry skin eczema. We excluded these categories because they were either too broad, the images were of poor quality, or the categories represented a rare genodermatosis. The final sample includes 114 conditions with at least 53 images (and a maximum of 653 images) per skin condition.

This dataset also includes two additional aggregated levels of skin condition classification based on the skin lesion taxonomy developed by Esteva et al. 2017, which can be helpful to improve the explainability of a deep learning system in dermatology [23, 7]. At the highest level, skin conditions are split into three categories: 2,234 benign lesions, 2,263 malignant lesions, and 12,080 non-neoplastic lesions. At a slightly more granular level, images of skin conditions are split into nine categories: 10,886 images labeled inflammatory, 1,352 malignant epidermal, 1,194 genodermatoses, 1,067 benign dermal, 931 benign epidermal, 573 malignant melanoma, 236 benign melanocyte, 182 malignant cutaneous lymphoma, and 156 malignant dermal. At the most granular level, images are labeled by skin condition.

The images are annotated with Fitzpatrick skin type labels by a team of human annotators from Scale AI. The Fitzpatrick labeling system is a six-point scale originally developed for classifying sun reactivity of skin and adjusting clinical medicine according to skin phenotype [25]. Recently, the Fitzpatrick scale has been used in computer vision for evaluating algorithmic fairness and model accuracy across skin type [11, 36, 22]. Fitzpatrick labels allow us to begin assessing algorithmic fairness, but we note that the Fitzpatrick scale does not capture the full diversity of skin types [54]. Each image is annotated with a Fitzpatrick skin type label by two to five annotators based on Scale AI's dynamic consensus process. The number of annotators per image is determined by a minimal threshold for agreement, which takes into account an annotator's historical accuracy evaluated against a gold standard dataset, which consists of 312 images with Fitzpatrick skin type annotations provided by a board-certified dermatologist. This annotation process resulted in 72,277 annotations in total.

In the *Fitzpatrick 17k* dataset, there are significantly more images of light skin types than dark skin. There are

7,755 images of the lightest skin types (1 & 2), 6,089 images of the middle skin types (3 & 4), and 2,168 images of the darkest skin types (5 & 6). Table 1 presents the distribution of images by skin type for each of the three highest level categorizations of skin conditions. A small portion of the dataset (565 images) are labeled as unknown, which indicates that the team of annotators could not reasonably identify the skin type within the image.

The imbalance of skin types across images is paired with an imbalance of skin types across skin condition labels. The *Fitzpatrick 17k* dataset has at least one image of all 114 skin conditions for Fitzpatrick skin types 1 through 3. For the remaining Fitpatrick skin types, there are 113 skin conditions represented in type 4, 112 represented in type 5, and 89 represented in type 6. In other words, 25 of the 114 skin conditions in this dataset have no examples in Fitzparick type 6 skin. The mean Fitzpatrick skin types across these skin condition labels ranges from 1.77 for basal cell carcionma morpheaform to 4.25 for pityriasis rubra pilaris. Only 10 skin conditions have a mean Fitzpatrick skin type above 3.5, which is the expected mean for a balanced dataset across Fitzpatrick skin types. These 10 conditions include: pityriasis rubra pilaris, xeroderma pigmentosum, vitiligo, neurofibromatosis, lichen amyloidosis, confluent and reticulated papillomatosis, acanthosis nigricans, prurigo nodularis, lichen simplex, and erythema elevatum diutinum.

| | Non-Neoplastic | Benign | Malignant |
|---|---|---|---|
| # Images | 12,080 | 2,234 | 2,263 |
| Type 1 | 17.0% | 19.9% | 20.2% |
| Type 2 | 28.1% | 30.0% | 32.8% |
| Type 3 | 19.7% | 21.2% | 20.2% |
| Type 4 | 17.5% | 16.4% | 13.3% |
| Type 5 | 10.1% | 7.1% | 6.5% |
| Type 6 | 4.4% | 2.0% | 2.7% |
| Unknown | 3.2% | 3.3% | 4.6% |

Table 1. Distribution of skin conditions in *Fitzpatrick 17k* by Fitzpatrick skin type and high level skin condition categorization.

| | Accuracy | Accuracy (off-by-one) | # of Images |
|---|---|---|---|
| Type 1 | 49% | 79% | 10 |
| Type 2 | 38% | 84% | 100 |
| Type 3 | 25% | 71% | 98 |
| Type 4 | 26% | 71% | 47 |
| Type 5 | 34% | 85% | 44 |
| Type 6 | 59% | 83% | 13 |

Table 2. Accuracy of human annotators relative to the gold standard dataset of 312 Fitzpatrick skin type annotations provided by a board-certified dermatologist.

# 3. Classifying Skin Conditions with a Deep Neural Network

## 3.1. Methodology

We train a transfer learning model based on a VGG-16 deep neural network architecture [47] pre-trained on ImageNet [20]. We replace the last fully connected 1000 unit layer with the following sequence of layers: a fully connected 256 unit layer, a ReLU layer, dropout layer with a 40% change of dropping, a layer with the number of predicted categories, and finally a softmax layer. As a result, the model has 135,335,076 total parameters of which 1,074,532 are trainable. We train the model by using the Adam optimization algorithm to minimize negative log likelihood loss. We address class imbalance by using a weighted random sampler where the weights are determined by each skin condition's inverse frequency in the dataset. We perform a number of transformations to images before training the model which include: randomly resizing images to 256 pixels by 256 pixels, randomly rotating images 0 to 15 degrees, randomly altering the brightness, contrast, saturation, and hue of each image, randomly flipping the image horizontally or not, center cropping the image to be 224 pixels by 224 pixels, and normalizing the image arrays by the ImageNet means and standard deviations.

We evaluate the classifier's performance via 5 approaches: (1) testing on the subset of images labeled by a board-certified dermatologist as diagnostic of the labeled condition and training on the rest of the data (2) testing on a randomly selected 20% of the images where the random selection was stratified on skin conditions and training on the rest of the data (3) testing on images from Atlas Dermatologico and training on images from Derma Amin (4) testing on images from Derma Amin and training on images from Atlas Dermatologico (5) training on images labeled as Fitzpatrick skin types 1-2 (or 3-4 or 5-6) and testing on the rest of the data. The accuracy on the validation set begins to flatten after 10 to 20 epochs for each validation fold. We trained the same architecture on each fold and report accuracy scores for the epoch with the lowest loss on the validation set.

## 3.2. Results

We report results of training the model on all 114 skin conditions across 7 different selections of holdout sets in Table 3.

In the random holdout, the model produces a 20.2% overall accuracy on exactly identifying the labeled skin condition present in the image. The top-2 accuracy (the rate at which the first or second prediction of the model is the same as the image's label) is 29.0% and the top-3 accuracy is 35.4%. These numbers can be evaluated against random guessing, which would be 1/114 or 0.9% accuracy. Across

| Holdout Set | Verified | Random | Source A | Source B | Fitz 3-6 | Fitz 1-2 & 5-6 | Fitz 1-4 |
|---|---|---|---|---|---|---|---|
| # Train Images | 16,229 | 12,751 | 12,672 | 3,905 | 7,755 | 6,089 | 2,168 |
| # Test Images | 348 | 3,826 | 3,905 | 12,672 | 8,257 | 10,488 | 14,409 |
| Overall | 26.7% | 20.2% | 27.4% | 11.4% | 13.8% | 13.4% | 7.7% |
| Type 1 | 15.1% | 15.8% | 40.1% | 6.6% | - | 10.0% | 4.4% |
| Type 2 | 23.9% | 16.9% | 27.7% | 8.6% | - | 13.0% | 5.5% |
| Type 3 | 27.9% | 22.2% | 25.3% | 13.7% | 15.9% | - | 9.1% |
| Type 4 | 30.9% | 24.1% | 26.2% | 17.1% | 14.2% | - | 12.9% |
| Type 5 | 37.2% | 28.9% | 28.4% | 17.6% | 10.1% | 21.1% | - |
| Type 6 | 28.2% | 15.5% | 25.7% | 14.9% | 9.0% | 12.1% | - |

Table 3. Accuracy rates classifying 114 skin conditions across skin types on six selections of holdout sets. The verified holdout set is a subset of a randomly sampled set of images verified by a board-certified dermatologist as diagnostic of the labeled condition. The random holdout set is a randomly sampled set of images. The source A holdout set are all images from Atlas Dermatologico. The source B holdout set are all images from Derma Amin. The 3 Fitzpatrick holdout sets are selected according to Fitzpatrick labels. In all cases, the training data are the remaining non-held out images from the *Fitzpatrick 17k* dataset.

| | | **Predicted Class** | | |
|---|---|---|---|---|
| | | Benign | Malignant | Non-neoplastic |
| **Actual Class** | Benign | 275 | 52 | 54 |
| | Malignant | 106 | 487 | 109 |
| | Non-neoplastic | 788 | 448 | 1586 |

Table 4. Confusion matrix for deep neural network performance on predicting the high-level skin condition categories in the holdout set of images from Atlas Dermatologico.

the 114 skin conditions, the median accuracy is 20.0% and ranges from a minimum of 0% accuracy on 10 conditions (433 images in the random holdout) and a maximum of 93.3% accuracy on 1 condition (30 images).

When we train the model on the 3 category partition of non-neoplastic, benign, and malignant, the model produces an accuracy of 62.4% on the random holdout (random guessing would produce 33.3% accuracy). Likewise, the model trained on the 9 category partition produces an accuracy of 36.1% on the random holdout (random guessing would produce 11.1% accuracy). Another benchmark for this 3 partition and 9 partition comes from Esteva et al. which trained a model on a dataset 7.5 times larger to produce 72.1% accuracy on the 3 category task and 55.4% accuracy on the 9 category task [23].

Depending on each holdout selection, the accuracy rates produced by the model vary across skin types. For the first four holdout selections in Table 3 – the verified selection, the random holdout, the source A holdout based on images from Atlas Dermatologico, and the source B holdout based on images from Derma Amin – we do not find a systematic pattern in accuracy scores across skin type. For the second three holdout selections where the model is trained on images from two Fitzpatrick types and evaluated on images in the other four Fitzpatrick types, we find the model is most accurate on the images with the closest Fitzpatrick skin types to the training images. Specifically, the model trained on images labeled as Fitzpatrick skin types 1 and 2 performed better on types 3 and 4 than types 5 and 6. Likewise, the model trained on types 3 and 4 performed better on types 2 and 5 than 1 and 6. Finally, the model trained on types 5 and 6 performed better on types 3 and 4 than types 1 and 2.

# 4. Evaluating Individual Typology Angle against Fitzpatrick Skin Type Labels

## 4.1. Methodology

An alternative approach to annotating images with Fitzpatrick labels is estimating skin tone via individual typology angle (ITA), which is calculated based on statistical features of image pixels and is negatively correlated with the melanin index [55]. Ideally, ITA is calculated over pixels in a segmented region highlighting only non-diseased skin [31]. But, segmentation masks are expensive to obtain, and instead of directly segmenting healthy skin, we apply the YCbCr algorithm to mask skin pixels [33]. We compare Fitzpatrick labels on the entire dataset with ITA calculated on the full images and the YCbCr masks.

The YCbCr algorithm takes as input an image in RGBA color space and applies the following masking thresholds.

$$R > 95 \tag{1}$$

$$R > G \tag{2}$$

$$R > B \tag{3}$$

$$G > 40 \tag{4}$$

$$B > 20 \tag{5}$$

$$|R - G| > 15 \tag{6}$$

$$A > 15 \tag{7}$$

Then, the image is converted from RGBA to YCbCr color space, and applies a further masking along the following thresholds:

$$Cr > 135 \tag{8}$$

$$Cr \geq (0.3448 \cdot Cb) + 76.2069 \tag{9}$$

$$Cr \geq (-4.5652 \cdot Cb) + 234.5652 \tag{10}$$

$$Cr \leq (-1.15 \cdot Cb) + 301.75 \tag{11}$$

$$Cr \leq (-2.2857 \cdot Cb) + 432.85 \tag{12}$$

where $R - G - B - A$ are the respective Red-Green-Blue-Alpha components of the input image, and $Y - Cb - Cr$ are the respective luminance and chrominance components of the color-converted image. As a result, the YCbCr algorithm attempts to segment healthy skin from the rest of an image.

We calculate the ITA of each full and YCbCr masked image by converting the input image to $CIE - LAB$ color space, which contains $L$: luminance and $B$: yellow, and applying the following formula [38]:

$$ITA = arctan(\frac{L^* - 50}{B^*}) \cdot \frac{180}{\pi} \tag{13}$$

where $L^*$ and $B^*$ are the mean of non-masked pixels with values within one standard deviation of the actual mean.

### 4.2. Results

In Table 5, we compare ITA calculations on both the full images and YCbCr masks with Fitzpatrick skin type labels. Furthermore, we compare two different methods for calculating Fitzpatrick type given ITA, as described in Equations 14 and 15. For each entry, we calculate the proportion of ITA scores in the range of plus or minus one of the annotated Fitzpatrick score.

$$Fitzpatrick(ITA) = \begin{cases} 1 & ITA > 55 \\ 2 & 55 \geq ITA > 41 \\ 3 & 41 \geq ITA > 28 \\ 4 & 28 \geq ITA > 19 \\ 5 & 19 \geq ITA > 10 \\ 6 & 10 \geq ITA \end{cases} \tag{14}$$

$$Fitzpatrick(ITA) = \begin{cases} 1 & ITA > 40 \\ 2 & 40 \geq ITA > 23 \\ 3 & 23 \geq ITA > 12 \\ 4 & 12 \geq ITA > 0 \\ 5 & 0 \geq ITA > -25 \\ 6 & -25 \geq ITA \end{cases} \tag{15}$$

In Table 5, the columns labeled "Kinyananjui" compare Fitzpatrick skin type labels with ITA following Equation 14, a modified version of the ITA thresholds described by Kinyanjui et al. [31]. The columns labeled "Empirical" follow Equation 15, which we developed based on the empirical distribution of ITA scores minimizing overall error. Figure 1 plots the empirical distribution of ITA scores for each Fitzpatrick skin type label. The discrepancy between Fitzpatrick skin type labels and the ITA approach appears to be driven mostly by high variance in the ITA algorithm as Figure 3 reveals.

| | Full Image | | YCbCr Mask | |
|---|---|---|---|---|
| | Kinyanjui | Empirical | Kinyanjui | Empirical |
| Overall | 45.87% | 60.34% | 53.30% | 70.38% |
| Type 1 | 50.97% | 65.35% | 52.22% | 66.00% |
| Type 2 | 42.60% | 59.57% | 49.15% | 69.47% |
| Type 3 | 35.43% | 55.20% | 45.13% | 66.41% |
| Type 4 | 34.09% | 58.54% | 40.24% | 72.10% |
| Type 5 | 78.21% | 65.49% | 93.41% | 82.26% |
| Type 6 | 74.80% | 65.04% | 90.71% | 79.69% |

Table 5. Plus or minus one concordance of individual typology angle (ITA) with Fitzpatrick skin type labels. Each column shows the percent of ITA scores that are within plus or minus 1 point of the annotated Fitzpatrick labels after converting ITA to Fitzpatrick types via Equations 14 and 15.

## 5. Conclusion

We present the *Fitzpatrick 17k*, a new dataset consisting of 16,577 clinical images of 114 different skin conditions annotated with Fitzpatrick skin type labels. These images are sourced from Atlas Dermatologico and Derma Amin and contain 3.6 times more images of the two lightest Fitzpatrick skin types than the two darkest Fitzpatrick skin types. By annotating this dataset with Fitzpatrick skin type labels, we reveal both an underrepresentation of dark skin images in online dermatology atlases and accuracy disparities that arise from training a neural network on only a subset of skin types.

By training a deep neural network based on an adapted VGG-16 architecture pre-trained on ImageNet, we achieve accuracy results that approach the levels reported on a much larger dataset [23]. We find that the skin type in the images on which a model is trained affects the accuracy scores across Fitzpatrick skin types. Specifically, we find that models trained on data from only two Fitzpatrick skin types are most accurate on holdout images of the closest Fitzpatrick skin types to the training data. These relationships between the type of training data and holdout accuracy across skin types are consistent with what has been long known by dermatologists: skin conditions appear differently across skin types [3].

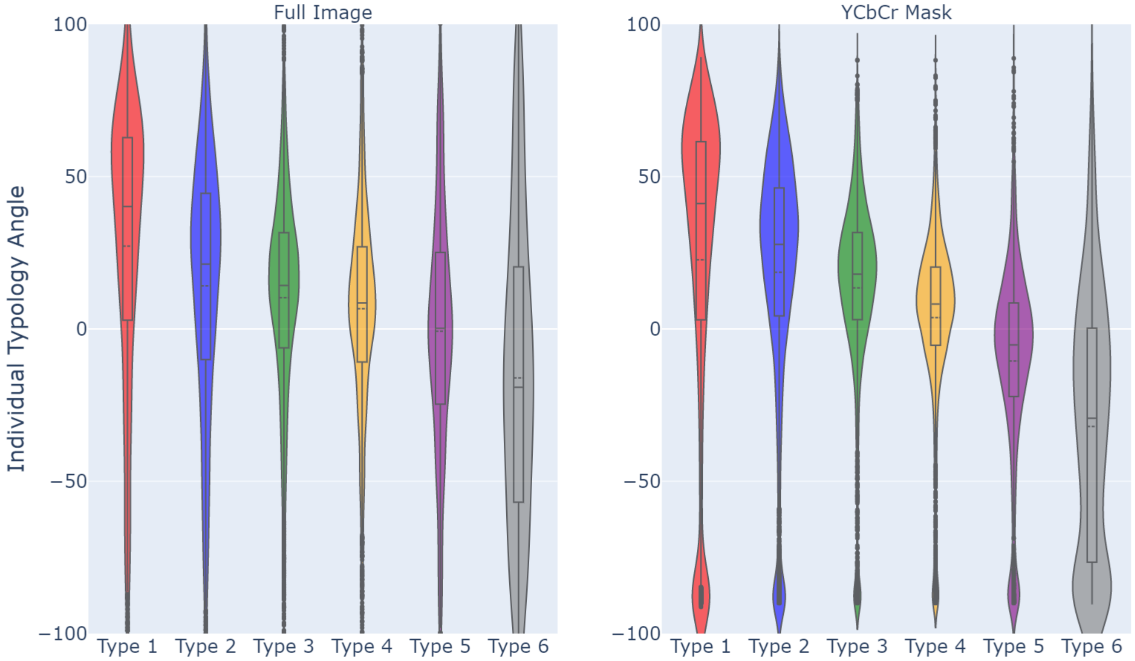

Figure 1. Observed distribution of individual typology angles by Fitzpatrick.

An open question for future research is in which skin conditions do accuracy disparities appear largest across skin types. Recent research shows that diagnoses by medical students and physicians appears to vary across skin types [24, 21]. Future research at the intersection of dermatology and computer vision should focus on specific groups of skin conditions where accuracy disparities are expected to arise because visual features of skin conditions (e.g. redness in inflammatory conditions) do not appear universally across skin types.

The large set of Fitzpatrick skin type labels enable an empirical evaluation of ITA as an automated tool for assessing skin tone. Our comparison reveals that ITA is prone to error on images that human labelers can easily agree upon. The most accurate ITA scores are off by more than one point on the Fitzpatrick scale in about one third of the dataset. One limitation of this comparison is that we calculated ITA based on either the entire image or an automatic segmentation mask. Future work should refine this comparison based on more precise segmentation masks.

We present this dataset and paper in the hopes that it inspires future research at the intersection of dermatology and computer vision to evaluate accuracy across subpopulations where classification accuracy is suspected to be heterogeneous.

# References

[1] Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E. J. Michels, Gerard Schouten, and Veronika Cheplygina. Risk of Training Diagnostic Algorithms on Data with Demographic Bias. *arXiv:2005.10050 [cs, stat]*, June 2020. arXiv: 2005.10050.

[2] Adewole S. Adamson and Avery Smith. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology*, 154(11):1247, Nov. 2018.

[3] Ademide Adelekun, Ginikanwa Onyekaba, and Jules B Lipoff. Skin color in dermatology textbooks: an updated evaluation and analysis. *Journal of the American Academy of Dermatology*, 84(1):194–196, 2021.

[4] Jehad Amin AlKattash. Dermaamin. *https://www.dermaamin.com/site/*.

[5] Savannah M. Alvarado and Hao Feng. Representation of dark skin images of common dermatologic conditions in educational resources: a cross-sectional analysis. *Journal of the American Academy of Dermatology*, page S0190962220311385, June 2020.

[6] M Shamsul Arifin, M Golam Kibria, Adnan Firoze, M Ashraful Amini, and Hong Yan. Dermatological disease diagnosis using color-skin images. In *2012 international conference on machine learning and cybernetics*, volume 5, pages 1675–1680. IEEE, 2012.

[7] Catarina Barata, M Emre Celebi, and Jorge S Marques. Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition*, 110:107413, 2021.

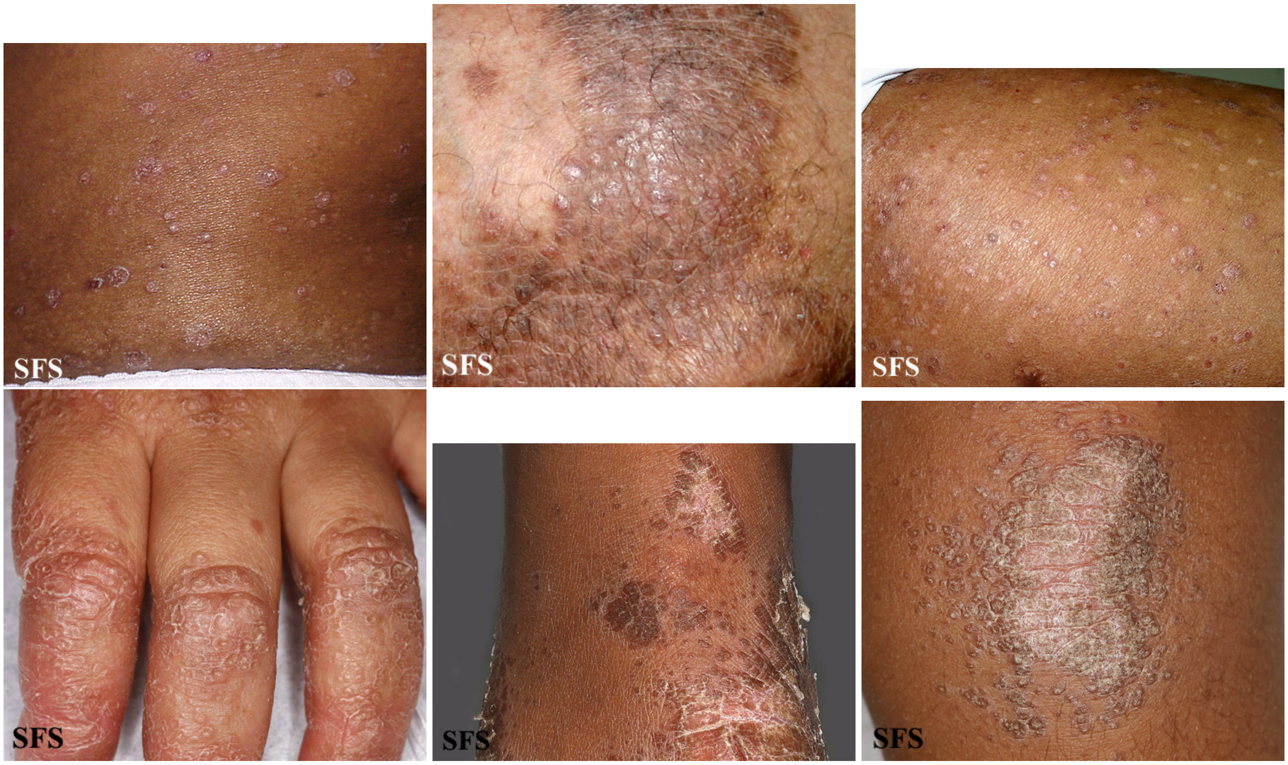

Figure 2. Example images of pityriasis rubra pilaris from Atlas Dermatologico that were accurately classified by the neural network trained on DermaAmin images. On the 174 images from Atlas Dermatologico labeled pityriasis rubra pilaris, 24% are accurately identified, 35% are accurately identified in the top 2 most likely predictions, and 45% are accurately identified in the top 3 most likely predictions.

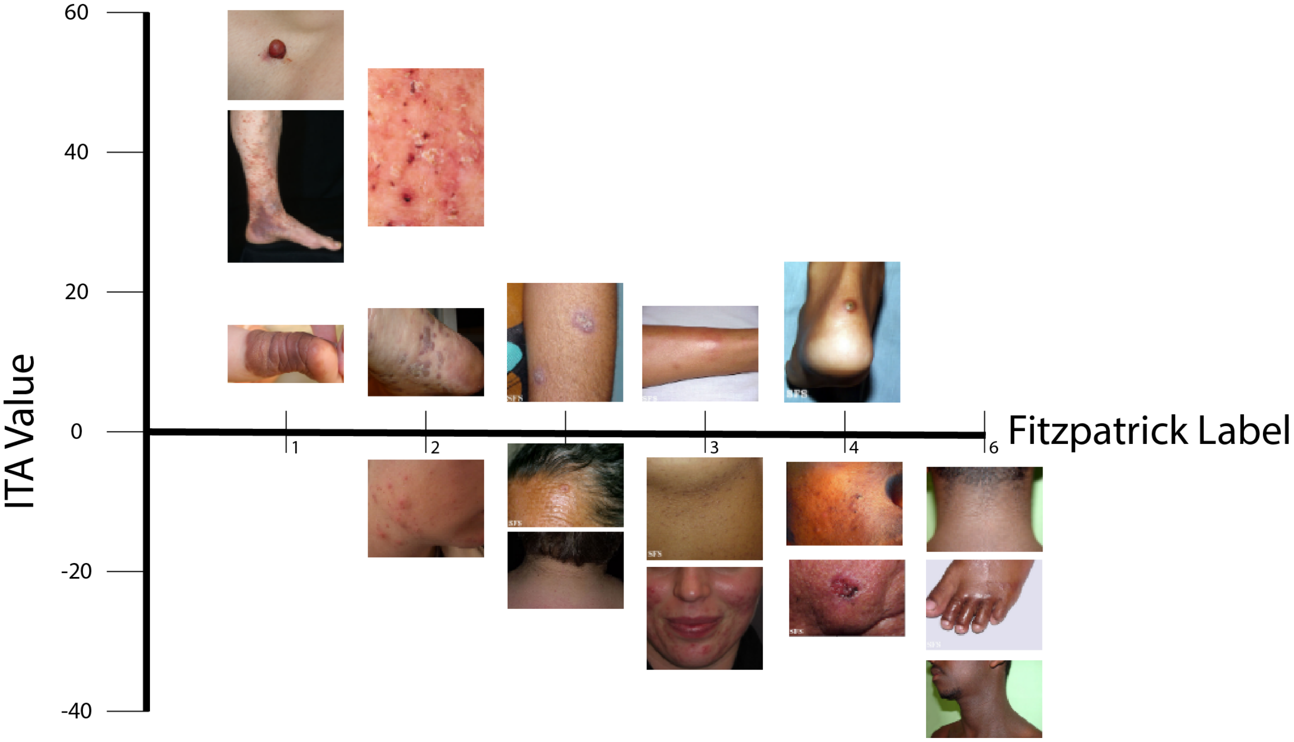

Figure 3. 18 images plot arranged based on ITA values and Fitzpatrick labels.

[8] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[9] Alceu Bissoto, Fábio Perez, Vinícius Ribeiro, Michel Fornaciali, Sandra Avila, and Eduardo Valle. Deep-learning ensembles for skin-lesion segmentation, analysis, classification: Recod titans at isic challenge 2018. *arXiv preprint arXiv:1808.08480*, 2018.

[10] Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 740–741, 2020.

[11] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[12] M Emre Celebi, Noel Codella, and Allan Halpern. Dermoscopy image analysis: overview and future directions. *IEEE journal of biomedical and health informatics*, 23(2):474–478, 2019.

[13] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018.

[14] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in health. *arXiv preprint arXiv:2009.10576*, 2020.

[15] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[16] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[17] International Skin Imaging Collaboration et al. Siim-isic 2020 challenge dataset. *International Skin Imaging Collaboration*, 2020.

[18] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[19] Bo Cowgill and Catherine E Tucker. Algorithmic fairness and economics. *The Journal of Economic Perspectives*, 2020.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[21] James A Diao and Adewole S Adamson. Representation and misdiagnosis of dark skin in a large-scale visual diagnostic challenge. *Journal of the American Academy of Dermatology*, 2021.

[22] Brittany Dulmage, Kyle Tegtmeyer, Michael Z. Zhang, Maria Colavincenzo, and Shuai Xu. A Point-of-Care, Real-Time Artificial Intelligence System to Support Clinician Diagnosis of a Wide Range of Skin Diseases. *Journal of Investigative Dermatology*, page S0022202X20321679, Oct. 2020.

[23] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, Feb. 2017.

[24] Anne Fenton, Erika Elliott, Ashkan Shahbandi, Ekene Ezenwa, Chance Morris, Justin McLawhorn, James G Jackson, Pamela Allen, and Andrea Murina. Medical students' ability to diagnose common dermatologic conditions in skin of color. *Journal of the American Academy of Dermatology*, 83(3):957, 2020.

[25] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[26] Samuel Freire da Silva. Atlas dermatologico. *http://atlasdermatologico.com.br/*.

[27] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, 42(19):6578–6585, 2015.

[28] Alpana K Gupta, Mausumi Bharadwaj, and Ravi Mehrotra. Skin cancer concerns in people of color: risk factors and prevention. *Asian Pacific journal of cancer prevention: APJCP*, 17(12):5257, 2016.

[29] Seung Seog Han, Myoung Shin Kim, Woohyung Lim, Gyeong Hun Park, Ilwoo Park, and Sung Eun Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7):1529–1538, 2018.

[30] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

[31] Newton M. Kinyanjui, Timothy Odonga, Celia Cintas, Noel C. F. Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R. Varshney. Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets. *arXiv:1910.13268 [cs, stat]*, Oct. 2019. arXiv: 1910.13268.

[32] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences*, page 201912790, July 2020.

[33] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia. Human skin detection using rgb, hsv and ycbcr color models. *Proceedings of the International Conference on Communication and Signal Processing 2016 (ICCASP 2016)*, 2017.

[34] J.C. Lester, J.L. Jia, L. Zhang, G.A. Okoye, and E. Linos. Absence of images of skin of colour in publications of

COVID-19 skin manifestations. *British Journal of Dermatology*, 183(3):593–595, Sept. 2020.

[35] Jenna Lester and Kanade Shinkai. Diversity and inclusivity are essential to the future of dermatology. *Cutis*, 104(2):99–100, 2019.

[36] Yuan Liu, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, Vishakha Gupta, Nalini Singh, Vivek Natarajan, Rainer Hofmann-Wellenhof, Greg S. Corrado, Lily H. Peng, Dale R. Webster, Dennis Ai, Susan J. Huang, Yun Liu, R. Carter Dunn, and David Coz. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, June 2020.

[37] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

[38] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. Diversity in faces, 2019.

[39] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[40] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, Toronto Ontario Canada, Apr. 2020. ACM.

[41] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, Oct. 2019.

[42] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomão, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020.

[43] James R Palmieri. Missed Diagnosis and the Development of Acute and Late Lyme Disease in Dark Skinned Populations of Appalachia. *Biomedical Journal of Scientific & Technical Research*, 21(2), Sept. 2019.

[44] Emma Pierson, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.

[45] Maryam Ramezani, Alireza Karimian, and Payman Moallem. Automatic detection of malignant melanoma using macroscopic images. *Journal of medical signals and sensors*, 4(4):281, 2014.

[46] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8, 2021.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Luis R Soenksen, Timothy Kassis, Susan T Conover, Berta Marti-Fuster, Judith S Birkenfeld, Jason Tucker-Schwartz, Asif Naseem, Robert R Stavert, Caroline C Kim, Maryanne M Senna, et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, 13(581), 2021.

[49] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222. Springer, 2016.

[50] Christian C Swinney, Dennis P Han, and Peter A Karth. Incontinentia pigmenti: a comprehensive review and update. *Ophthalmic Surgery, Lasers and Imaging Retina*, 46(6):650–657, 2015.

[51] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, Aug. 2020.

[52] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[53] Lindi Van Zyl, Jeanetta Du Plessis, and Joe Viljoen. Cutaneous tuberculosis overview and current treatment regimens. *Tuberculosis*, 95(6):629–638, 2015.

[54] Olivia R Ware, Jessica E Dawson, Michi M Shinohara, and Susan C Taylor. Racial limitations of fitzpatrick skin type. *Cutis.*, 105(2):77–80, 2020.

[55] Marcus Wilkes, Caradee Y. Wright, Johan L. du Plessis, and Anthony Reeder. Fitzpatrick Skin Type, Individual Typology Angle, and Melanin Index in an African Population: Steps Toward Universally Applicable Skin Photosensitivity Assessments. *JAMA Dermatology*, 151(8):902–903, 08 2015.

[56] Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 155(10):1135, Oct. 2019.