– Independent Work Report Spring, 2022 –

Bias in Skin Lesion Classification

Tanushree Banerjee

Adviser: Prof. Olga Russakovsky

Abstract

The use of automated machine learning-based skin lesion diagnosis systems has a huge potential to not only speed up the diagnosis process for deadly skin conditions such melanoma, but also help bridge disparities in healthcare and dermatology in care for patients with darker skin [1]. However, the overrepresentation of images of lighter skin in skin lesion datasets [2] used to train such models could correspond to discrepancies in accuracy across different skin tones [3]. This paper trains a model to classify skin conditions based on images of skin lesions from the Fitzpatrick17k dataset [4]. A confidence interval on the accuracy of the trained model across the different skin types is analyzed using the bootstrap method [11] in order to determine the existence of bias in the trained model.

1. Introduction

In the United States, images of skin conditions on darker skin are underrepresented in dermatology residency programs [13], textbooks [14], dermatology research [15] and dermatology diagnoses [16]. As a result, physicians are mostly trained to recognise skin conditions on lighter skin [3]. Since skin conditions manifest themselves differently on lighter and darker skin (for instance, with a different shade of redness, pigmentation, discoloration etc.), skin conditions are often misdiagnosed on patients with darker skin [2]. Thus, the lack of training

physicians get in diagnosing skin conditions on darker skin harms care for patients of color. Therefore, the emergence of deep neural networks that can classify skin conditions such as cancer with human-level accuracy [1] presents a huge opportunity to improve healthcare and bridge diagnosis accuracy disparities in dermatology as well as in healthcare at large [4].

Moreover, automating the process of diagnosis also promises to speed up the process of diagnosis, allowing the early detection of deadly skin diseases such as melanoma early. Since early detection of skin conditions such as melanoma are crucial for good prognosis, the use of deep learning models to assist diagnosis can help develop early-detection techniques and save thousands of lives.

However, even the datasets used to train these neural networks are mostly made up of images of skin conditions on lighter skin. For example, the Fitzpatrick17k dataset [4] containing images of skin lesions corresponding to 114 different skin conditions contains almost twice the number of images displaying lighter skin than darker skin [4]. The lack of consideration of subgroups within a population was hypothesized as a possible cause for the large disparity across gender and skin color for facial recognition in Buolamwini and Gebru's work [5]. Thus, it is plausible that this overrepresentation of light skin in skin lesion datasets may lead to diagnosis accuracy disparities across skin color, thereby even further increasing healthcare disparities in dermatology rather than bridging them, should such biased skin models be deployed in the real world.

Thus, the goal of this project is to rigorously analyze the existence of accuracy discrepancies across skin color and their statistical significance in a neural network trained on one such biased dataset - Fitzpatrick17k [4]. First, a machine learning model is trained such that it takes as input an image of a skin lesion and outputs a label for the skin condition displayed in

the input image. Next, we examine the distribution of image labels in the dataset used in order to aid our analysis of bias. Finally, we examine the discrepancies in accuracy of the trained model on the testing data to assess the existence of bias. We use the bootstrapping method [11] to determine if any accuracy discrepancies are statistically significant.

A rigorous analysis of bias in a dataset and accuracy disparities across subpopulations in a model is crucial before deploying such models into practice. Such an analysis can help pave the way for the development of models that are more accurate for underrepresented subpopulations, that can also serve as discrimination detectors [4]. The analysis of model bias also helps identify opportunities to address this bias, such as collecting additional data from the underrepresented groups or disentangling the source of the disparities in accuracy [4].

Previous work conducted on skin lesion classification has focused on older datasets such as the International Skin Imaging Collaboration (ISIC) dataset, which do not contain any skin type or demographic labels (apart from the PAD-UFES-20 dataset with skin type labels on 579 out of 1,373 images [17]). In addition, previous analysis such as in [12] is mostly focussed on discrepancies due to spurious correlations rather than discrepancies due to underrepresentation of subgroups in a dataset.

Thus, the main challenge in conducting analysis on bias due to underrepresentation of darker skin in the dataset is a lack of large enough datasets containing skin type or demographic labels in order to be able to conduct any rigorous analyses on accuracy disparities. This challenge is compounded by the difficulty of the task of classifying skin lesions from images, owing to the vast variation in the appearance of skin lesions, as well as the subtlety of the cues that distinguish malignant and benign conditions.

With the release of the Fitzpatrick17k in 2021, it is now possible to conduct such an analysis on bias, as the dataset contains images with skin condition labels on three levels of granularity – 3 low-level, 9 mid-level and 114 high-level labels. The paper introducing Fitzpatrick17k [4] trains a model for the low level classification. Their analysis reveals the underrepresentation of darker skin in online dermatology atlases, and determine that the model is most accurate on skin types similar to those it was trained on [4].

However, no analysis has been done on models trained on the high or mid-level labels. Thus, in this project, we focus our analysis on the model trained to classify images by the mid level categories instead of the low level categories, for reasons discussed later in the paper. Moreover, the implementation of the model used in the original paper is changed in order to reduce the time required for training as well as remove the need for a GPU to train.

Feature vectors for each image are obtained by conducting one forward pass of the image and then storing the values obtained at the penultimate layer of AlexNet [9] pretrained on ImageNet [10] after this forward pass. A linear classifier is then trained independently on these image feature vectors and the ground truth labels. Finally, the accuracies across skin types are computed and analyzed.

The analysis conducted fails to conclude that a statistically significant discrepancy in accuracy exists between images of different skin types. However, this does not necessarily mean that no bias exists – only that the current analysis does not reveal it.

In the remainder of this paper, we first discuss in depth the prior work done on the analysis of bias in skin lesion classification, and why it does not meet the goal of this paper. We also give more detail on the Fitzpatrick17k dataset and the labels on each image. Next, we introduce our key novel idea and how it drives our approach. We then describe the implementation of the model in further detail. Finally, we present and discuss in detail our analysis of accuracy disparities by skin type

2. Problem Background

2.1. The Fitzpatrick17k dataset

The Fitzpatrick17k dataset contains 16,577 images [4] of a skin lesion or several lesions. Each image is annotated with

1. Skin condition labels [4]
2. Skin type labels based on the Fitzpatrick scoring system [4]

The images are obtained from the following two online dermatology atlases, containing images with their corresponding skin condition labels:

1. DermaAmin: 12,672 images [4]
2. Atlas Dermatologico: 3,905 images [4]

Although the labels on these images are not known to be confirmed via a biopsy [4], these labels have been used and cited by computer vision literature [4]. Moreover, the paper by Groh et al. also conducts a quality check on the labels of the images on the dataset by asking board-certified dermatologists to evaluate the accuracy of 3% of the full dataset [4], consisting of a random sample of 504 images [4]. The quality check confirms that the error rate on the sampled dataset is consistent with the average error rate of humans of 3.4% in the most

commonly used test datasets in computer vision, natural language processing and audio processing [4].

From the images obtained from the aforementioned online dermatology atlases, a subset of images are chosen to be annotated with Fitzpatrick skin type labels. 22 categories of skin conditions that were either too broad, with images of poor quality or represented a rare hereditary skin disease were excluded from being in the dataset [4]. Thus, the final dataset contains images of skin lesions corresponding to 114 different skin conditions, with at least 53 images and at most 653 images of each of the 114 skin conditions represented [4].

In addition to the low-level labels for the 114 specific skin conditions displayed in the image, each image is also annotated with the two additional aggregated levels of skin condition classification labels. These aggregated levels of skin condition labels are based on the taxonomies developed in the paper by Esteva et al. [8] which were shown to be helpful in improving the explainability of deep learning models to classify skin lesions on the ISIC 2017 and 2018 dataset [8].

The highest level categories split the 114 labels into three broad categories. The three high level categories and number of images labeled with this category are given in Table 1 below.

High-level category label	Number of images
Benign lesions	2,234
Malignant lesions	2,263
Non-neoplastic lesions	12,080

Table 1: High-level categories and number of images in each category in the dataset []

At a more granular level, the 114 labels are split into 9 mid level categories. Each of the 9 mid-level categories and the number of images in each category is shown in Table 2 below.

Mid-level category label	Number of images
images labeled inflammatory	10,886
malignant epidermal	1,352
genodermatoses	1,194
benign dermal	1,067
benign epidermal	931
malignant melanoma	573
benign melanocyte	236
malignant cutaneous lymphoma	182
malignant dermal	156

Table 2: Mid level categories and number of images in each category in the dataset []

Each of the images (apart from the images in the 22 categories chosen to be excluded from the dataset) are annotated by labels based on the Fitzpatrick scoring system by a team of human annotators from Scale AI [4]. For a small subset of images in the dataset, the Fitzpatrick skin type could not be identified by the annotators, for whom the Fitzpatrick skin type is labeled 'unknown'. The Fitzpatrick labeling system is based on a six-point scale, originally created to classify sun reactivity of skin as well as adjusting dermatology medicine according to skin color [18]. However, more recently, labels based on the Fitzpatrick scale have been used to evaluate fairness and model accuracy across skin type in computer vision [5], as they are assumed to serve as a proxy for race and ethnicity. It is important to note, however, that Fitzpatrick labels do not always perfectly align with race and do not capture the full diversity of skin types [19]. Nevertheless, using Fitzpatrick labels allows us to at least begin to evaluate algorithmic fairness based on skin color.

2.2. Related Work

Most prior analysis on skin lesion classification models has been done on older datasets such as the International Skin Imaging Collaboration (ISIC) dataset. Apart from the PAD-UFES-20 [] dataset, none of the public skin lesion image datasets at the ISIC Skin Image Analysis workshop at CVPR 2021, such as Derm7pt [20], Dermofit Image Library [4], ISIC 2018 [21], ISIC 2019 [22], ISIC 2020[22], MED-NODE [4], PH2 [4], SD-128 [4], SD-198 [4] and SD-260 [4] contain any skin type or demographic labels, and even the PAD-UFES-2 [17] dataset only contains skin type labels for only 579 out of 1,373 images [17]. Thus, conducting any robust analysis of accuracy based on skin type was difficult until the release of the Fitzpatrick17k dataset.

Bissoto et al.'s paper, *(De)Constructing Bias on Skin Lesion Datasets*, proposes a set of experiments that reveal positive and negative biases in existing skin lesion datasets and models trained on them [12]. The paper's analysis leads to some concerning results – that the trained model seems to correctly classify skin lesion images even when all information about the lesion is removed from the image [12]. This strongly suggests that the model trained is learning some spurious correlations in order to make its prediction. They conduct their analysis on the Interactive Atlas of Dermoscopy (Atlas) dataset as well as the International Skin Imaging Collaboration (ISIC) Archive dataset. Neither of these datasets contain any labels indicating skin type or race, and the paper conducts no analysis of bias by skin color.

Thus, Groh et al. introduce the Fitzpatrick17k dataset with almost 17,000 images labeled with skin type labels based on the Fitzpatrick scale. The images are obtained from two online dermatology atlases: Atlas and DermAmin [4]. First, the paper analyzes the distribution of images by Fitzpatrick skin type in the dataset, which is summarized in Table 3 below. It is evident that there are significantly more images of lighter skin than darker skin in the dataset.

Skin type	Number of images
Lightest skin types: 1 and 2	7,755
Middle skin types: 3 and 4	6,089
Darkest skin types: 5 and 6	2,168

Table 3: Distribution of images by skin type

In addition to the imbalance in the distribution of images by skin type, there is an imbalance in skin types across skin condition labels. For example, there is at least one image for each of the 114 skin conditions for skin types 1, 2 and 3. However, skin type 6 is only represented in 89 skin conditions, i.e. 25 skin conditions have no examples on type 6 skin.

Groh et al. continuen their analysis by training a transfer learning model based on a VGG-16 neural network architecture [23] pretrained on ImageNet [10]. They then replace the last fully connected layer in this architecture with the following sequence of layers [4]:

1. Fully connected 256-unit layer
2. ReLU layer
3. Dropout layer with 40% change of dropping
4. Linear layer with number of predicted categories
5. Softmax layer

Next, Groh et al. evaluate the model's performance using the following five experiments with different sets of training and testing data [4]. The experiments are summarized in Table 4 below.:

Experiment number	**Test set**	**Train set**
1	images labeled by a board-certified dermatologist as diagnostic of the labeled condition	Remaining data
2	randomly selected 20% of the images where the random selection was stratified on skin conditions	Remaining data
3	images from Atlas Dermatologico	images from Derma Amin
4	images from DermaAmin	images from Atlas Dermatologico
5	Remaining data	images labeled as Fitzpatrick skin types 1-2 (or 3-4 or 5-6)

Table 4: Experiments conducted by Groh et al. to evaluate accuracies across skin type

Through their analysis, they conclude that models trained on data from only two Fitzpatrick skin types are most accurate on images of the closest Fitzpatrick skin types to images they were trained on [4]. However, the paper does not conduct any analysis on whether these accuracy disparities are statistically significant given the smaller size of the dataset and the lack of training examples for the rarer skin conditions, especially for darker skin types.

3. Approach

Previous work by Groh et al. has only conducted a rigorous analysis on a model trained to classify images on the 114 low level labels representing the specific skin conditions. In this paper, we focus our analysis of bias on the nine mid-level category label classification. Since each of the 9 mid-level categories have more images per label, any analysis on accuracy discrepancies by skin type would be less noisy than the 114-way classification. Thus, any results obtained would be more robust, i.e. it is more likely that there would be sufficient data to fully evaluate accuracy disparities. Yet, the 9 mid-level categories give more fine-grained information than the broad 3 high level category labels.

In this project, the implementation of the model used to train the 9-way classifier is changed so as to reduce the time required for training. This reduces the computational cost and hence the need for a GPU as well. This allows for the model to potentially be deployed in applications where computational power such as a high performance GPU is not available or too expensive to include, as well as allow the GPU to be used for other tasks so as to not waste computational power.

4. Implementation

4.1. Transfer learning to train the classifier

Training a convolutional neural network (CNN) from scratch to be able to classify the images into the 9 mid-level categories would be computationally expensive and time consuming. Moreover, the dataset is of a relatively small size. Thus, instead of training a CNN from scratch, transfer learning is used to build the model. As described in [24], transfer learning involves pre-training a CNN on a very large dataset, and then using the CNN as an initialization or fixed feature extractor for the intended task.

In this project, the AlexNet [23] architecture is used as the CNN, and is pretrained on ImageNet [10], an image database containing 1.2 million images with 1000 categories. AlexNet is a CNN architecture that competed in the ImageNet Large Scale Visual Recognition Challenge [23] in 2012. This architecture was chosen since the AlexNet architecture has often been used in other computer vision papers, making comparisons easier. ImageNet was chosen as the dataset used for training since it is one of the largest and most widely used dataset for pretraining.

In particular, the pretrained CNN is used as a fixed feature extractor as follows. The last fully connected layer, whose outputs are the 1000 task scores for each of the 1000 categories in ImageNet, is removed [24]. Next, each image is forward-propagated once through this fixed feature extractor, and the 4096-dimensional vector obtained at the last layer of this model (i.e. the penultimate layer of the original CNN model) is used as the feature vector representing the image[24]. These feature vectors are called CNN codes. For best performance, it is crucial that these obtained CNN codes are ReLUd (i.e. thresholded at zero), since they were also thresholded during the training of the AlexNet on ImageNet during the pre-training [24].

Once the 4096-D codes for all images are extracted, a linear classifier is trained such that takes as input the CNN code of an image and outputs one of the 9 mid-level skin condition labels. The linear classifier used in this project is the Stochastic Gradient Descent classifier, since the SGD classifier can be trained much faster, and at a lower computational cost than other linear classifiers like support vector machines (SVMs) or Softmax classifiers.

4.2 System overview

The flowchart in Figure 1 below shows the key steps involved in the training and evaluation of the model. Each of the key steps are described in more detail below.

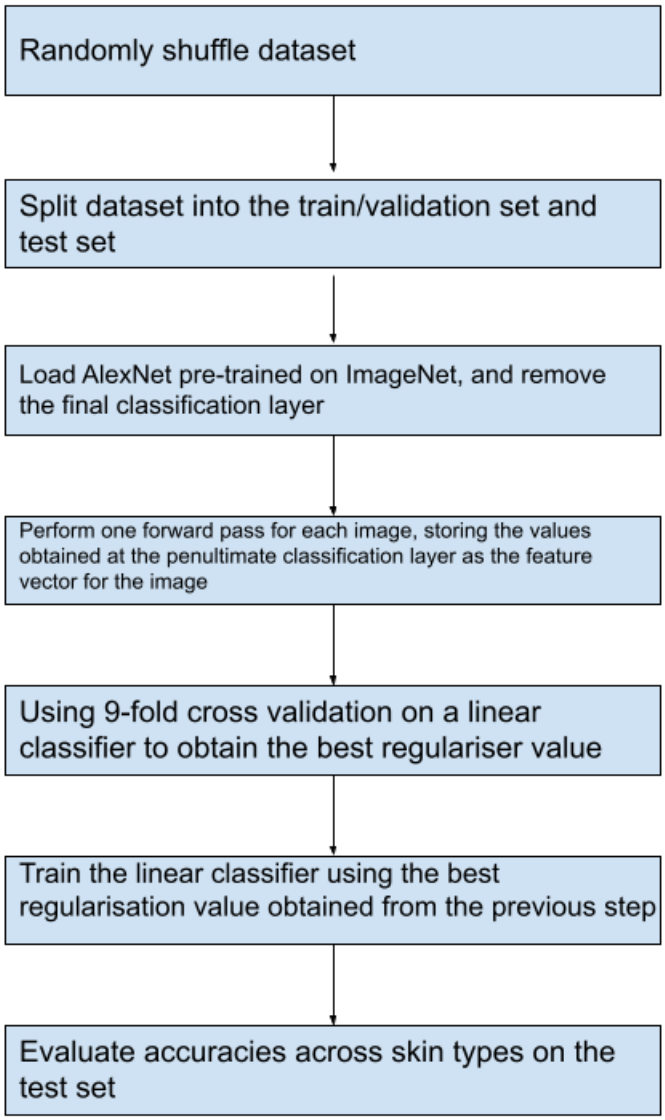

Figure 1: Flowchart showing the key steps involved in implementing the goal of the project

Step 1: Randomly shuffle the dataset.

This is done to ensure that the labels in the dataset are randomly distributed, so that any conclusions drawn are not made due to the order of images in the dataset.

Step 2: Split the dataset into the train/validation and test set.

90% of the dataset is kept as the train/validation set, while the remaining 10 percent of the dataset is used as the test set. Thus, the number of images in the train/validation and test set are given as follows:

1. Train/validation set: 14919 images
2. Test set: 1658 images

Step 3: Load the Alexnet pretrained on ImageNet and remove the final classification layer

The original AlexNet architecture [23] is loaded pre-trained on ImageNet [10], i.e. with all the final model parameters after the model was trained on the 1000-way classification task on the ImageNet dataset [10]. The final fully connected 1000 unit layer is removed, such that the last layer is the penultimate layer of the original model, with 4096 densely connected neurons [23]. The architecture of the image feature extractor thus obtained is given in Figure 2 below.

```
AlexNet(
  (features): Sequential(
    (0): Conv2d(3, 64, kernel_size=(11, 11), stride=(4, 4), padding=(2, 2))
    (1): ReLU(inplace=True)
    (2): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (3): Conv2d(64, 192, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
    (4): ReLU(inplace=True)
    (5): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
    (6): Conv2d(192, 384, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (7): ReLU(inplace=True)
    (8): Conv2d(384, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (9): ReLU(inplace=True)
    (10): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (11): ReLU(inplace=True)
    (12): MaxPool2d(kernel_size=3, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (avgpool): AdaptiveAvgPool2d(output_size=(6, 6))
  (classifier): Sequential(
    (0): Dropout(p=0.5, inplace=False)
    (1): Linear(in_features=9216, out_features=4096, bias=True)
    (2): ReLU(inplace=True)
    (3): Dropout(p=0.5, inplace=False)
    (4): Linear(in_features=4096, out_features=4096, bias=True)
    (5): ReLU(inplace=True)
  )
```

Figure 2: Python representation of the architecture of the feature extractor obtained after modifying AlexNet

Step 4: For each image, perform one forward pass through the modified AlexNet architecture, and store the values at the last 4096 densely connected neurons as the CNN codes for the image

The final layer of the modified AlexNet has 4096 units [23], the CNN codes obtained are 4096 dimensional vectors. Thus, the CNN codes obtained are 4096-dimensional vectors.

Step 5: Use 9-fold cross validation to determine the best regulariser value for the linear classifier

K-fold cross validation is a statistical method used to compare and select the best hyperparameters such as regularization values [25]. The method involves the following key steps []:

1. Shuffle the train/validation set randomly.
2. Split the rain/validation set into k groups (in this project, k is chosen to be 9)

3. For each of the unique possible combinations of hyperparameter values (in this case, a given list of possible regularization factors):
    a. For each unique group of the k groups, i.e. ‘folds’ created:
        i. Take the group as the validation set
        ii. Take the remaining groups as the train set
        iii. Fit the linear classifier model on the train set
        iv. Evaluate the performance of the model on the validation set using a performance evaluation metric. In this paper, accuracy is the metric used to compare performance.
        v. Store the evaluation score and discard the obtained model.
    b. Calculate the average of the performance comparison metric (in this case, the average of the accuracies on all the k groups) on the validation set
4. Pick the combination of hyperparameters (in this case, a regularization factor value) for which the model had the highest average evaluation score (in this case, the highest average accuracy over each of the k ‘folds’) on the validation.

The we try regularization values between 10^{-15} and 10^{10}, increasing by an order of magnitude, i.e. [10^{-15}, 10^{-14}, 10^{-13}, 10^{-12}, 10^{-11}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}, 10^{7}, 10^{8}, 10^{9}, 10^{10}]. Using the procedure above, we plot the average accuracy over the 9 ‘folds’ for each of the regularization values listed above with the x-axis on a logarithmic scale. The plot is given in Figure 3 below. Thus, the best average accuracy over the 9 folds is obtained for a regularization factor of 0.1. Thus, this is the value of the regularization factor used to train the final linear classifier.

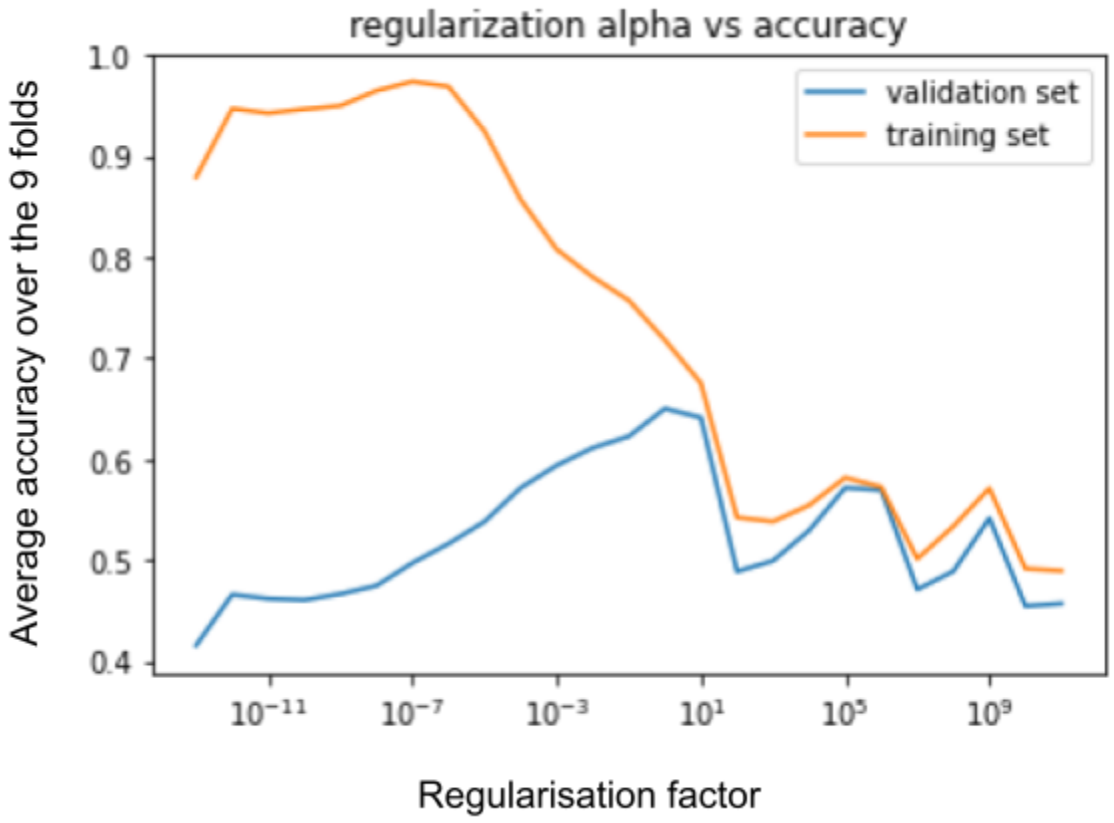

Figure 3: Plot of the average accuracy over the 9 folds vs the regularization factor

Step 6: Train the linear classifier using the best regularization value obtained in the previous step.

The best regularization factor value of 0.1 is used to train a stochastic gradient descent (SGD) classifier.

Step 7: Evaluate accuracies across skin type on the held out test set.

The details of the method used to evaluate accuracies and the results from this step are given and discussed in detail in the next Evaluation section.

5. Evaluation

For ease of reading, each of the 9 mid-level skin condition labels are referred to by their corresponding category code as given in Table 5 below. Note that images whose Fitzpatrick skin type is labeled ‘unknown’ are denoted by the label -1.

Mid-level skin condition label	Category code
benign dermal	0
benign epidermal	1
benign melanocyte	2
genodermatoses	3
inflammatory	4
malignant cutaneous lymphoma	5
malignant dermal	6
malignant epidermal	7
malignant melanoma	8

Table 5: Mid level skin condition categories and their corresponding category code

5.1. Distribution of images by skin type and mid-level skin condition label

Figure 4 below shows the number of images of each Fitzpatrick skin type in the full dataset, with -1 representing images labeled as having an ‘unknown’ skin type. Moreover, Figure 5 below shows the number of images of each mid-level skin condition category in the full dataset. Fitzpatrick skin type 2 is the most represented skin type in the dataset, while skin type 6 is the least represented. Moreover, the lightest three skin types make up 67% of the full dataset, while the darkest three types only make up less than 30% of the full dataset. Thus, a disproportionate proportion of the dataset, compared to the darkest three skin types.

Moreover, category code 4 (‘inflammatory’ category) makes up almost 66% of the full dataset, while only 0.94% of the dataset is labeled ‘malignant dermal’ (category code 6). Thus, a disproportionate number of images are labeled category 4, so the relative number of training examples for the other categories is relatively much smaller.

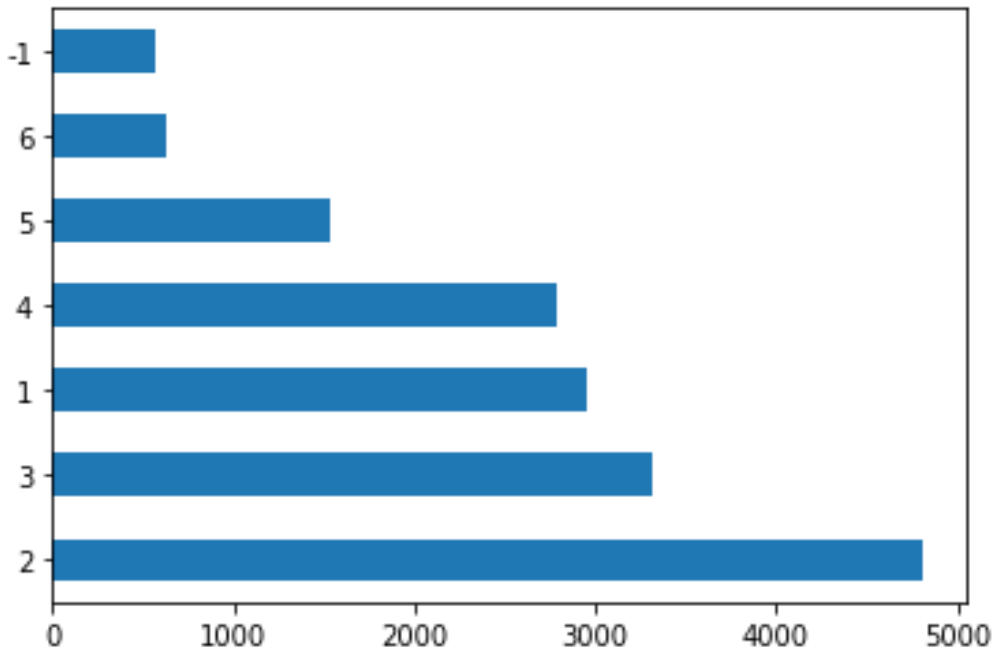

Figure 4: number of images by Fitzpatrick skin type label in the full dataset

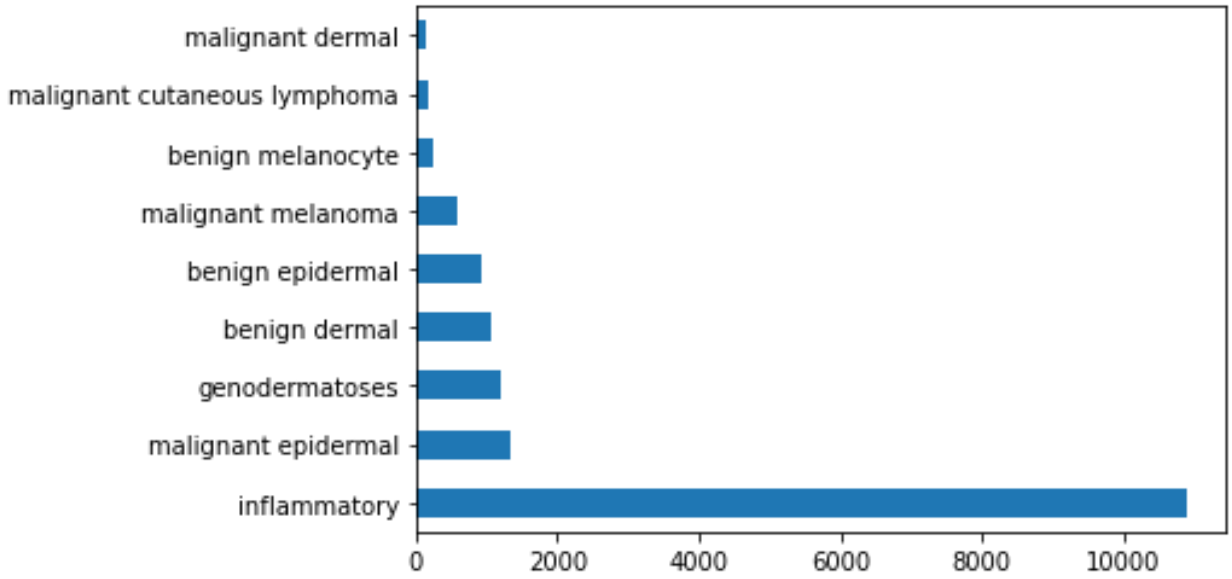

Figure 5: number of images of each mid level label in the full dataset

5.2. Confusion matrix

The confusion matrix summarizes the number of correct and incorrect predictions by giving the count values of the predicted labels broken down by each true label. The confusion matrix of the final trained model using the procedure explained above is given below in Table 6, normalized such that the sum of the values in each row is 1. The accuracies of each category on the test set are also given in Figure 7 below, alongside the percentage of images in the full dataset that are labeled as that category.

		Predicted label								
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
True label	**0**	11.46%	3.13%	0.00%	0.00%	81.25%	0.00%	0.00%	4.17%	0.00%
	1	1.12%	10.11%	0.00%	0.00%	84.27%	0.00%	0.00%	4.49%	0.00%
	2	4.55%	0.00%	0.00%	0.00%	90.91%	0.00%	0.00%	0.00%	4.55%
	3	0.00%	0.93%	0.00%	16.67%	82.41%	0.00%	0.00%	0.00%	0.00%
	4	0.44%	0.26%	0.09%	0.44%	97.89%	0.00%	0.00%	0.61%	0.26%
	5	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	6	0.00%	0.00%	0.00%	0.00%	94.44%	0.00%	5.56%	0.00%	0.00%
	7	0.87%	2.61%	0.00%	0.00%	63.48%	0.00%	0.00%	31.30%	1.74%
	8	0.00%	4.65%	0.00%	0.00%	60.47%	0.00%	0.00%	9.30%	25.58%

Table 6: Confusion matrix of the final trained classifier on the test set

Category code	Accuracy on test set	% of full dataset set
0	11.50%	6.44%
1	10.10%	5.62%
2	4.60%	1.42%
3	16.70%	7.20%
4	97.90%	65.67%
5	0%	1.10%
6	5.60%	0.94%
7	31.30%	8.16%
8	25.60%	3.46%

Table 7: Accuracies on test set and percentage of full dataset labeled as each category code

Discussion

Most images are labeled category 4 (inflammatory), and this category has the highest accuracy rate amongst all other categories (97.9%). However, most other classes are misclassified more often than they are correctly classified. Infact, whenever they are misclassified (>60% of the time for all classes except category 4), they are usually wrongly classified as category 4. This is expected since most of the images in the dataset (and hence also in the training set) are of category 4. Thus, whenever the model is unsure, perhaps it learns to simply output 4 when the model is unsure. Thus, it seems that the overall high accuracy of the model may be misleading - the model has a significantly lower accuracy on all categories apart from category 4 (under 30% for all categories).

Moreover, from Table 7, it seems that higher the proportion of a label in the full dataset, the higher the accuracy for that category on the training set. This suggests that a greater number of images of a specific category in the dataset correspond to a higher accuracy for that category on the test set.

5.2. Error bounds on accuracies using the bootstrap method

In order to analyze discrepancies in accuracy across different Fitzpatrick skin types, we use the bootstrapping method suggested in [11]. This method gives us error bounds over the accuracies by skin type. The method is summarized as follows [11].

Let there be n images in the test set. n images are sampled with replacement from the test set, to obtain B bootstrapped samples. Accuracies by skin type are then computed for each of the B bootstrapped samples, and stored in an array such that there is one array for each skin type label. Next, each array is sorted, and the $\alpha / 2$ and $1 - (\alpha / 2)$ quantiles for each array are found and stored, where $\alpha = 0.05$, to get the 95% confidence interval, with the $\alpha / 2$ quartile value

being the lower bound and 1 – (α / 2) quantile being the upper bound of the confidence interval [11].

The accuracies are then plotted with error bars corresponding to the 95% confidence interval found in the previous step. Figure 4 below shows the bar graph with the 95% confidence interval over the accuracies displayed.

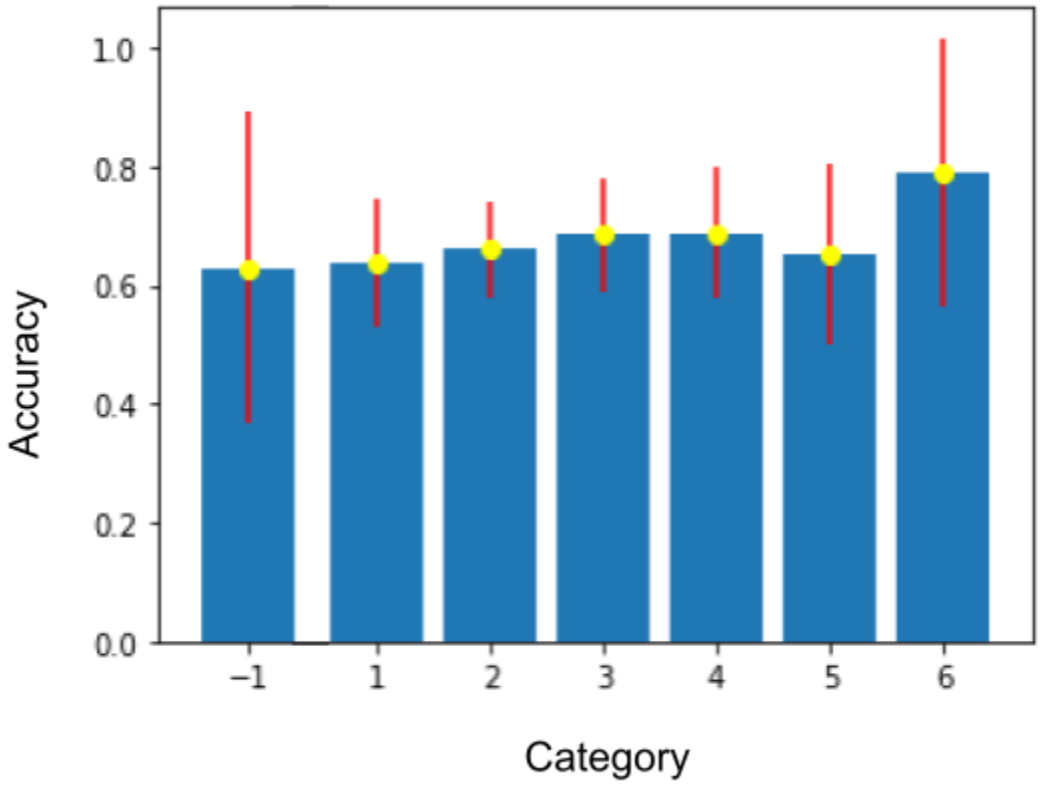

Figure 6: Classification accuracy by skin type, with 95% confidence interval obtained using the bootstrap method (-1 denotes unknown Fitzpatrick skin type)

Discussion

There is a relatively large uncertainty on the accuracy on the test set across all skin types. This may be due to the relatively small size of the test set, with under 2000 images, and hence under 200 images per mid-level category on average. The largest uncertainty in accuracy is for the darkest skin type, type 6, which is the least represented skin type in the full dataset, while the

smallest uncertainty in accuracy is for the skin type 2, which is the most represented skin type in the dataset. This suggests that a higher the number of training examples of the skin type in the full dataset corresponds to a lower the uncertainty on the accuracy for that skin type on the test set. Thus, this suggests that more images of the skin type in a test set corresponds to a smaller confidence interval on the accuracy for that skin type on the test set.

In order to see if there exists a statistically significant difference in the accuracies across skin type, we plot in Figure x below the difference in accuracy of each skin type on the test set from that of skin type 2 on the test set.

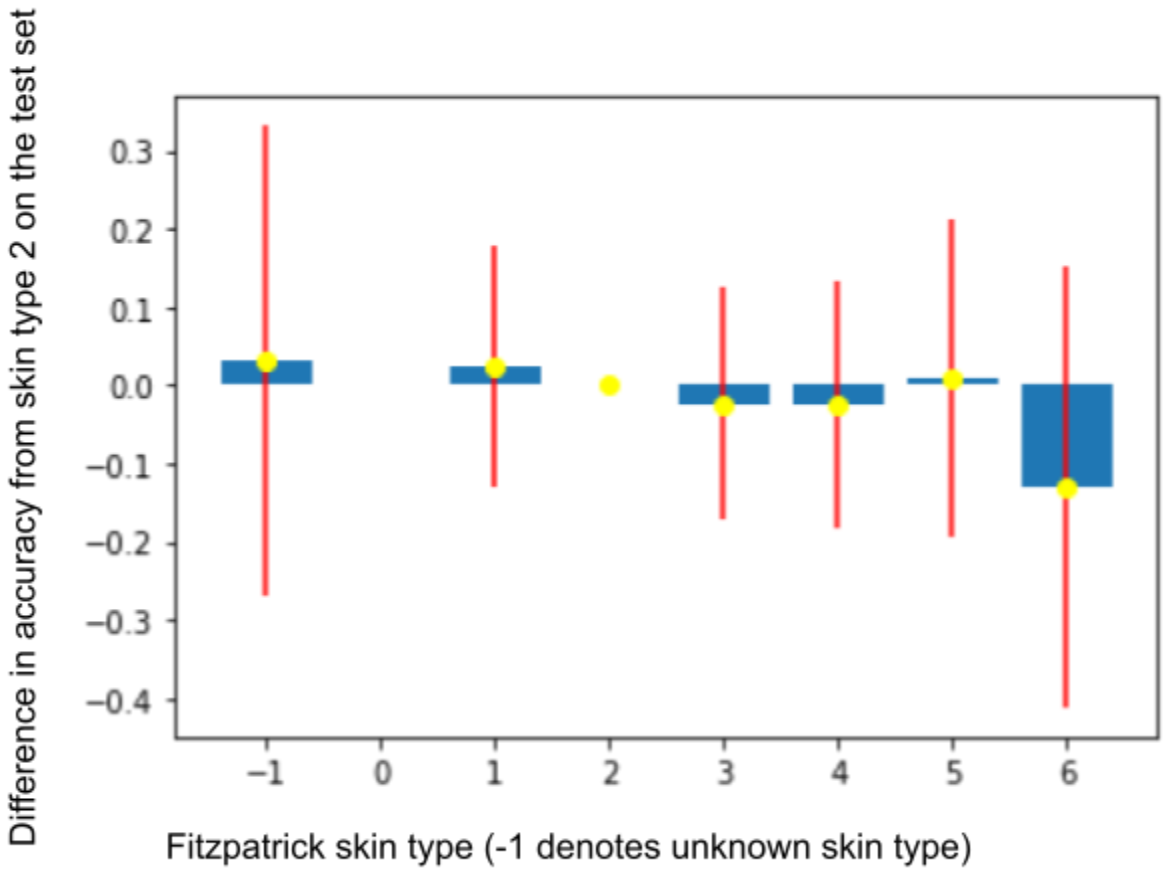

Figure 7: difference in accuracy of each skin type on the test set from that of skin type 2 on the test set

Discussion

From the figure above, it seems that the value 0 is within the confidence interval for the difference in accuracy of each skin type from that of skin type 2 on the test set. Thus, there is not enough evidence at the 95% confidence level to conclude that there is a difference in accuracy of each skin type from that of skin type 2 on the test set.

However, the error bounds on the accuracy for all skin types, and especially skin type 6, are very big. Thus, even though the current analysis does not conclusively reveal a bias, it does not rule out the possibility that a bias does indeed exist.

6. Summary

6.1. Conclusions and Limitations

The Fitzpatrick17k dataset contains a disproportionate number of lighter skinned images compared to darker skinned images, as well as a vast difference in the number of images per mid-level skin condition category. Thus, the overall higher accuracy of the final trained model may thus be misleading due to this class imbalance, since although the model performs relatively well on the classes represented most in the dataset, it performs poorly on the less represented classes. Moreover, the analysis of accuracy discrepancies using the bootstrapping method [11] fails to conclusively determine a discrepancy in accuracy based on skin type in the final model.

However, this negative bias result does not mean that a bias does not exist in the model since the confidence intervals on the accuracy discrepancies are large. Our analysis also suggests that a greater number of images of a certain class in the full dataset corresponds to a greater accuracy in the test set. Moreover, the smaller confidence intervals on the accuracies for each Fitzpatrick skin type seem to correspond with a greater representation of the skin type in the dataset. This suggests that greater number of images of each skin type in the dataset would allow

us to obtain more precise values for the accuracies that would allow for a more robust evaluation of accuracy discrepancies by skin type.

6.2 Future work

Bevan and Atapour-Abarghouei's paper [26] proposes a method of quantifying and approximating skin tone using a metric called the individual typology angle (ITA). This could potentially allow larger skin lesion datasets without skin type labels to be used for analysis of accuracy disparities across different skin colors by estimating the skin tone using the ITA metric [26]. However, the ITA metric is very sensitive to lighting conditions, making it an unreliable metric to estimate skin tone and hence not a reliable proxy for race or ethnicity. Nevertheless, perhaps using the larger datasets with more examples of images of each skin type and skin condition would help overcome the problem of lack of images of certain skin types and conditions in the dataset, which would help us obtain smaller confidence intervals on the accuracies using the bootstrap method, and hence help conduct a more meaningful analysis of bias based on skin tone in skin lesion classification models. The paper also introduces and evaluates two potential debiasing techniques [26], which could be implemented. The debiased model could then be evaluated using a similar method as used in this report.

Moreover, further analysis could be conducted by changing the distribution of images by FItzpatrick skin type used in the training set versus the test set to further determine the extent to which the number of images of a given skin type affects the overall accuracy of the model by skin type on the test set.

More long term, such a robust evaluation of accuracy disparities in ML models across subpopulations where classification accuracy is suspected to be heterogeneous could perhaps pave the way for models that are more generalizable and fair towards those from

underrepresented subgroups. This would allow AI models deployed in the real world to bridge disparities due to race, ethnicity, gender, etc, rather than further worsen them.

References

[1] 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.

[2] "Dermatology Has a Problem With Skin Color." The New York Times - Breaking News, US News, World News and Videos. Last modified August 31, 2020. https://www.nytimes.com/2020/08/30/health/skin-diseases-black-hispanic.html.

[3] "Lack of Darker Skin in Textbooks, Journals Harms Patients of Color." STAT. Last modified July 20, 2020. https://www.statnews.com/2020/07/21/dermatology-faces-reckoning-lack-of-darker-skin-in-textbooks-journals-harms-patients-of-color/.

[4] Groh, Matthew, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badr. "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset." arXiv:2104.09957 [cs.CV]. Last modified April 20, 2021. https://arxiv.org/pdf/2104.09957.pdf

[5] 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.

[6] 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.

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

[8] Esteva et al 2017 Dermatologist-level classification of skin cancer

[9] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Last modified December 2012. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[10] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. "ImageNet Large Scale Visual Recognition Challenge." International Journal of Computer Vision 115, no. 3 (2015), 211-252. doi:10.1007/s11263-015-0816-y.

[11] Everingham, Mark, S. M. Ali Eslami, Luc Van Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. "Assessing the Significance of Performance Differences on the PASCAL VOC Challenges via Bootstrapping." The PASCAL Visual Object Classes. Last modified October 18, 2013. https://host.robots.ox.ac.uk/pascal/VOC/pubs/bootstrap_note.pdf.

[12] Bissoto, Alceu, Michel Fornaciali, Eduardo Valle, and Sandra Avila. "(De)Constructing Bias on Skin Lesion Datasets." ArXiv.org. Accessed April 19, 2022. https://arxiv.org/abs/1904.08818.

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

[14] 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.

[15] 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

[16] 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.

[17] Pacheco AGC, Lima GR, Salomão AS, Krohling B, Biral IP, de Angelo GG, Alves FCR Jr, Esgario JGM, Simora AC, Castro PBC, Rodrigues FB, Frasson PHL, Krohling RA, Knidel H, Santos MCS, do Espírito Santo RB, Macedo TLSG, Canuto TRP, de Barros LFS. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data Brief. 2020 Aug 25;32:106221. doi: 10.1016/j.dib.2020.106221. PMID: 32939378; PMCID: PMC7479321.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. https://doi.org/10.48550/arXiv.1409.1556.

[24] Stanford University. "CS231n Convolutional Neural Networks for Visual Recognition." CS231n Convolutional Neural Networks for Visual Recognition. Accessed April 20, 2022. https://cs231n.github.io/transfer-learning/.

[25] "A Gentle Introduction to K-fold Cross-Validation." Machine Learning Mastery. Last modified August 2, 2020. https://machinelearningmastery.com/k-fold-cross-validation/.

[26] Bevan, Peter J., and Amir Atapour-Abarghouei. "Detecting Melanoma Fairly: Skin Tone Detection and Debiasing for Skin Lesion Classification." Accessed April 20, 2022. https://doi.org/10.48550/arXiv.2202.02832.