Alex Baroody, Katie McLaughlin
COS 429

# Dermatology Datasets for Neural Network Classification:

# Assessing and Mitigating Skin-Tone Bias

## 1. Introduction.

The pervasive inequalities in healthcare that disproportionately impact people of color are well-documented. In dermatology textbooks and literature, black and brown skin is not represented proportionately. Skin cancer mortality rates are significantly higher for African Americans compared to white Americans (Gupta, Bharadwaj, & Mehrotra, 2016). Using Machine Learning (ML) algorithms to diagnose skin conditions holds great potential for increasing accessibility to early diagnoses and treatment. However, as we turn to ML to assist in identifying skin blemishes, it is a concerning possibility that this demographic gap carries over to training datasets and reinforces these disparities among darker-skinned patients (Lashbrook, 2018). We hypothesize that this underrepresentation in dermatology datasets leads to a worse performance by a neural network when classifying skin conditions on patients with darker skin. If this is realized, it would mean that careful consideration would need to be taken into account when crafting training datasets for dermatology studies. If it isn't, it is a promising conclusion for Convolutional Neural Networks (CNNs): it means that datasets with a disproportionate number of black and brown skin images can still be used to train effective classification models.

## 2. Previous Work.

The challenges in testing this possibility all stem from the dataset: the recent advancements in CNN's mean that accessible, modular, accurate models are available with minimal setup. Indeed, the pre-trained ResNet used in this project took mere hours to get fully functional. The datasets, however, are not so easy to work with. These challenges are rooted in the size of the available

datasets. Due to both the lack of robust research in this area and the nature of the images themselves, large, well-labeled datasets do not exist.[1] A deep neural network built by a research group at Stanford outperformed dermatologists in skin cancer classification (Esteva et al., 2017). The network was trained on a robust dataset of 129,450 clinical images. To analyze the accuracy of a neural network across skin tones, each image must be labeled in a way that quantifies skin tone. The Fitzpatrick scale is one such way in dermatology to numerically classify skin tone based on pigmentary phototype and response to ultraviolet light. At one end of the scale, skin that falls under Fitzpatrick type I is the palest shade, which always burns and never tans. At the other end, Type VI skin is deeply pigmented and never burns (Ward et al., 2017). The Stanford dataset was not annotated for Fitzpatrick type and was too large for us to annotate within the timeframe of the project.

Researchers from MIT curated a well-labeled dataset of clinical images in the Fitzpatrick 17k dataset, a collection of 16,577 images of 114 skin conditions from two dermatology atlases

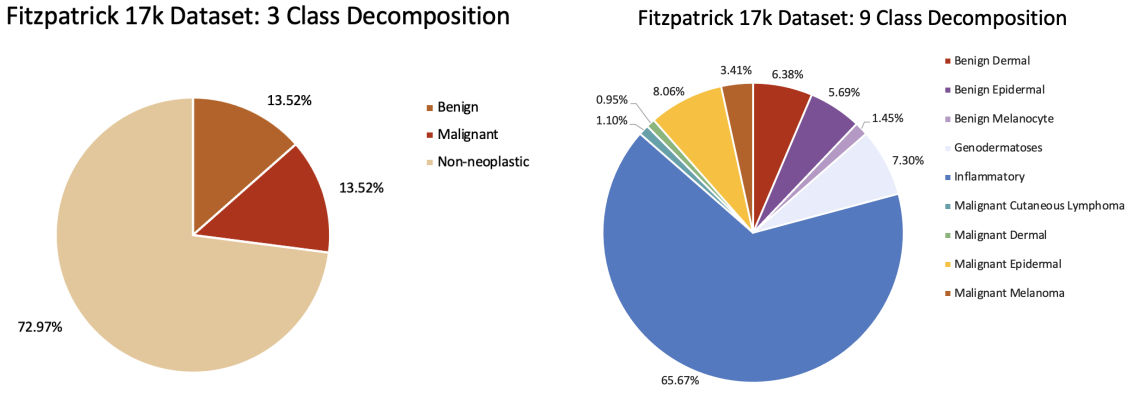


Figure 1. Breakdown of the Fitzpatrick 17k dataset along two of the classification systems. The chart on the left shows the breakdown of the broadest level, a three-partition classification between benign (2158), malignant (2158), and non-neoplastic (11651). The other classification system on the right is a nine-way split in which an image is classified as one of the following: benign dermal (1018), benign epidermal (909), benign melanocyte (231), genodermatoses (1165), inflammatory (10486), malignant cutaneous lymphoma (176), malignant dermal (151), malignant epidermal (1287), malignant melanoma (544).

[1] In this context, "well-labeled" means that the dataset has labels for both skin condition and Fitzpatrick skin type. Any combination of labels that does not include at least these two sets of labels is considered "poorly-labeled."

(Groh et al., 2021). Each image in the dataset is also labeled for two other aggregate classification levels, one with a three-way split and the other with nine categories.


Figure 2. Example images of similar skin conditions for people of different skin types. The top row is skin types I, II, and III (left, center, right), while the bottom row is skin types IV, V, and VI (left, center, right).

All images have a corresponding label on the Fitzpatrick scale, with a label of -1 if skin type is not detectable in the image (i.e. skin condition is on the gums or tongue). For the purposes of our research, we discarded all images with a Fitzpatrick label of -1, as we are most interested in bias issues of skin tone. This leaves us with 15,967 downloadable images from the original set of 16,577.

The paper authored by Groh et al. that was written alongside the open-source release of this dataset used the 114 skin condition labels. The group aimed to train a network that could achieve overall accuracies near those of larger datasets. However, approaching the dataset using the 114-class split meant that some classes had no images of darker Fitzpatrick labels. For example, 25 classes had no images of Fitzpatrick type VI. We know that a network will perform poorly with no data of a certain class to train on, especially given that skin conditions present differently based on the patient's skin color (Gupta, Bharadwaj, & Mehrotra, 2016). Thus, we

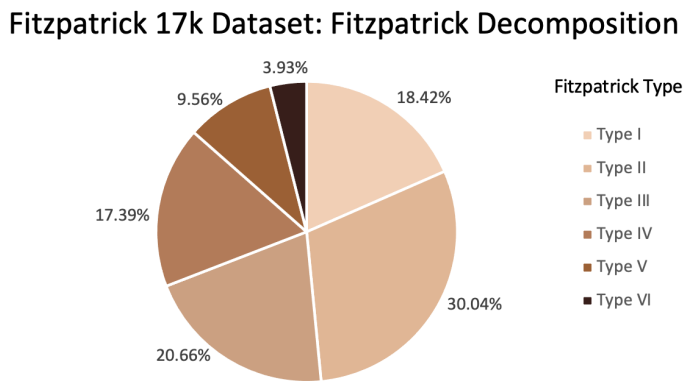


Figure 3. Distribution of 15,967 images in the Fitzpatrick 17k dataset over the 6 Fitzpatrick skin types: Type I (2941), Type II (4796), Type III (3299), Type IV (2776), Type V (1527), Type VI (628).

are more curious about how a network will perform with less data for one class than another. We chose to work with the 3-class and 9-class splits instead to ensure all labels had at least one image of each Fitzpatrick type. One of the primary focuses of the MIT research effort was also to annotate the dataset with Fitzpatrick labels to reveal the underrepresentation of dark skin images. While this underrepresentation can be observed and confirmed with their annotations, there still remains the question of if this disparity is a condemning sentence for lower accuracies on skin conditions in patients with darker skin. In other words, does a neural network trained on a biased dataset perform consistently worse on patients with darker skin? Answering this question requires that we assess and compare the accuracies of a 3-class and 9-class neural network for each of the six Fitzpatrick types.

**3. Design and Implementation.**

To classify a skin condition present in an image, we employ the use of a Convolutional Neural Network. A Convolutional Neural Network takes an image as input and outputs one of $n$ classes. It uses convolutional filters to process the image and narrow it down from the input size, which

can be tens of thousands of pixels, to one of the *n* input classes. Specifically, the network repeatedly performs the convolution operation between each filter and the output of the previous layer, outputting a new "image." The filters work to expose different structural patterns that might be common to a specific class. Additionally, there are non-convolutional layers scattered throughout the network, like pooling or ReLU (Rectified Linear Unit) layers. A pooling layer makes its input image smaller by a factor of $m^2$, where *m* is the size of the pooling layer. This downscaling makes the network more robust to location variance within images. This can be done in several ways, including outputting the average of the $m^2$ pixels and outputting the maximum of the $m^2$ pixels. A ReLU layer simply sets the value of the input pixel equal to 0 if that value is less than 0, and leaves it alone otherwise. Finally, the network has at least one fully connected layer to transform its 2D input to a 1D output, along with one last activation layer, such that the network will finally output the correct number of outputs and those outputs will be in the correct ranges. For a multiclass output (that is, a network that outputs more than two classes), the way to evaluate the model is by measuring accuracy, or the total number of correctly predicted labels divided by the total number of images.

The CNN we used for this project is comprised of a softmax layer, a fully connected layer, and an average pooling layer postfixed onto the ResNet CNN (Figure 4). The code used to implement ResNet was adapted from a PyTorch tutorial on finetuning CNNs (Inkawhich, 2015). These layers go from the 2D output of ResNet to a 1-by-3 or a 1-by-9 output, depending on which set of labels we are classifying. Each of these values is between 0 and 1 and represents the confidence the model has that the input image has the specified label. This architecture allows us to leverage the full power of ResNet to classify images in our dataset. Further, we used a pre-trained version of ResNet (trained on the ImageNet dataset by Google), which

| layer name | output size | 18-layer | 34-layer | 50-layer | 101-layer | 152-layer |
|---|---|---|---|---|---|---|
| conv1 | 112×112 | 7×7, 64, stride 2 | | | | |
| conv2_x | 56×56 | 3×3 max pool, stride 2 | | | | |
| | | $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$ | $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$ |
| conv3_x | 28×28 | $\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$ | $\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$ |
| conv4_x | 14×14 | $\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$ | $\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$ |
| conv5_x | 7×7 | $\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$ | $\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$ |
| | 1×1 | average pool, 1000-d fc, softmax | | | | |
| FLOPs | | $1.8\times10^9$ | $3.6\times10^9$ | $3.8\times10^9$ | $7.6\times10^9$ | $11.3\times10^9$ |

Figure 4. Possible ResNet Architectures. For this project, we used ResNet18 because it was the easiest to implement and we didn't need to use more layers to test our hypothesis. If the goal was the highest possible accuracy, a larger net likely would have been a better choice.

means that our training process only meant fine-tuning the network. Finally, we trained the network in two different ways: the first was only training our new layers (the FC and Average Pooling layers) and the second was training the entire network. The latter had the potential for higher accuracy, but also for adverse results like overfitting, especially with a small dataset.

To accurately test the dataset and avoid overfitting, we split the dataset into two uneven pieces: a random 25% of the dataset became our test data, while the remaining 75% of the data was our training data. This meant that the 3,981 images in the test dataset were never seen by the model during training, and the model was never given the labels for any of these images.

To thoroughly test our hypothesis, we ran three sets of experiments on images divided according to the two sets of labels. The two sets of labels used for each of the three experiments were used to test the model's accuracy at differing levels of granularity. The first set consisted of 3 labels: benign-neoplastic, malignant-neoplastic, and non-neoplastic; the second set consisted of 9 labels: 3 different types of benign-neoplastic conditions, 4 different types of

malignant-neoplastic conditions, and 2 different types of non-neoplastic conditions. Both sets of labels provide meaningful results and each has enough images to properly train our model. The first set of experiments was on the unchanged dataset, while the second and third sets were performed on augmented versions of the dataset. The second and third experiments were performed on a balanced and augmented dataset, respectively. The balanced data set was a subset of the original dataset with images removed until each label had the same number of training images: 475, the number of Type VI-labeled images in the unbalanced dataset. The augmented dataset copied and randomly rotated images until each label had as many images as the label that had the most images in the original dataset.

**4. Results and Analysis.**

We hypothesized that the observed underrepresentation of patients with darker skin in dermatology datasets leads to poor performance – measured by accuracy – on images of skin conditions on darker skin. To test this hypothesis, we curated a balanced dataset by creating an equal number of images per Fitzpatrick skin type labels. We can look at the differences in accuracy on the test set for each skin type after training one model on the existing imbalanced dataset and another on the balanced dataset. In general, CNNs trained on more robust datasets tend to have better performance on unseen data. We would apply similar logic to anticipate lower accuracies on skin types with fewer images (specifically Types VI, Type V, Type IV) from a model trained on the unbalanced dataset. If the underrepresentation of darker toned images does cause disparities in accuracy, we would expect these disparities to be remedied by the balanced training dataset.

We were surprised to see our initial experiments produce accuracies that did not align with our initial intuitions. It was in fact Type VI – the Fitzpatrick type with the least number of images – that had the highest accuracy across the 3-way and 9-way split for models trained on the imbalanced and balanced dataset. The general distribution of accuracies across skin types was preserved between the models trained on the imbalanced and balanced datasets (see Figure 5). In other words, the difference between each Fitzpatrick accuracy and the mean of accuracies did not change significantly from imbalanced to balanced. For classifying images labeled with Fitzpatrick type II, the 9-class partition imbalanced model had 77.65% accuracy and the balanced model accurately predicted 71.66% accuracy, a 4.37% decrease and 3.47% decrease from the mean accuracy, respectively. While Type II is the skin type with the most images, this was the lowest accuracy across the six skin types. Also interesting was the high accuracy on Type VI images, where the 9-class partition imbalanced model performed 4.14% better than the average and the balanced model performed 3.97% better. Thus, we see that the difference in performance by skin type is not completely attributable to the size of the training set for that skin type.

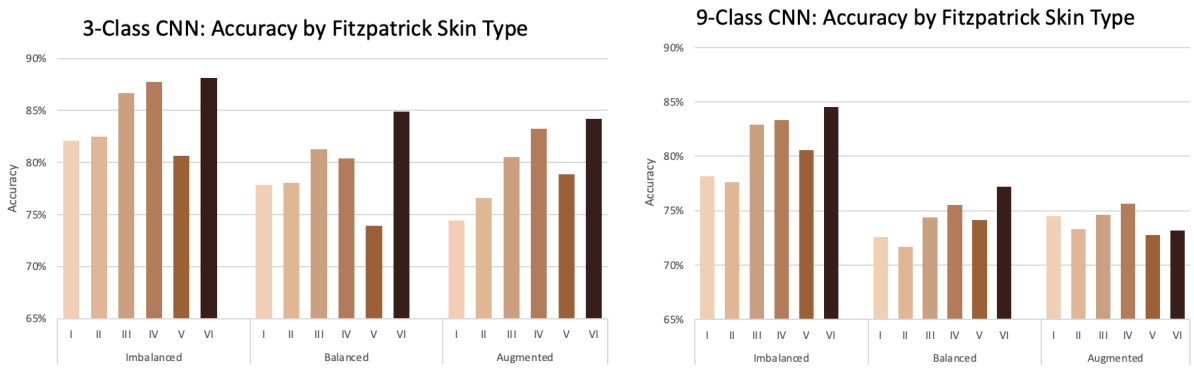


Figure 5. Accuracy by Fitzpatrick skin type for both the 3-class and 9-class split. Each graph shows accuracy results on the test set for a model trained on the imbalanced, balanced, and augmented dataset.

In order to understand why the distributions of the unbalanced and balanced models were similar, we computed confusion matrices for both the 3-class and 9-class split (see Figure 6). These tools revealed that the imbalance of the labels was also affecting the way the model trained. It is clear that the model predicts one label ("non-neoplastic" and "inflammatory" for the 3-split and the 9-split respectively) much more often than the others. This makes sense given the makeup of the dataset: there are many more non-neoplastic and inflammatory images than any other label. This imbalance is exacerbated for skin type VI, which has far fewer total images than any other skin type. The combination of imbalances led us to conclude that, rather than learning what skin conditions looked like, the model predicts non-neoplastic or inflammatory the vast majority of the time, which results in a high accuracy because the vast majority of the images are labeled as such. Indeed, when we tested the imbalance 3-split model on only the benign and malignant images in the test dataset, it returned only 37.50% accuracy, far lower than the 88.16% accuracy it reported on the entire test set. This shows why the balanced and unbalanced models had such similar distributions: they both were subject to this example of Simpson's Paradox.[2]

Armed with this knowledge, we modified our approach to test if the disparities among skin types arose from an unequal distribution of 3-class and 9-class labeled images. We created a uniformly balanced training dataset but this time along the 3-class and 9-class labels rather than Fitzpatrick labels. We designed this in order to train the model to recognize image-specific features rather than recognizing a probability distribution among the inputs. We created this augmented dataset according to the procedure described above in the implementation section. If

[2] "Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations" (Sprenger & Weinberger, 2021). In this case, the overwhelming number of non-neoplastic/inflammatory images mean that the accuracy for the entire dataset is high, even though the accuracies for all labels that aren't non-neoplastic/inflammatory are objectively low.

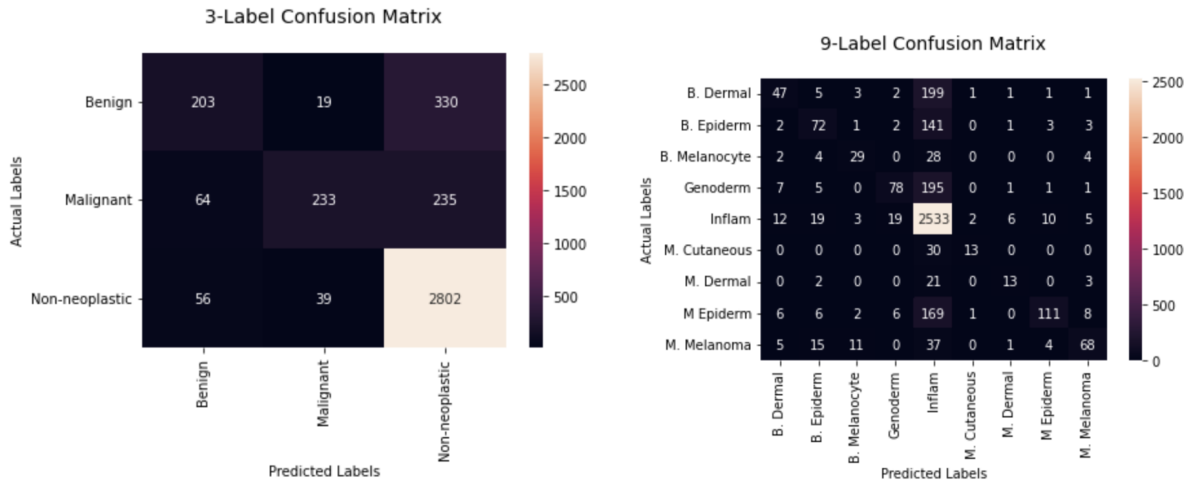


Figure 6. Confusion matrices for the 3-split and 9-split models. Note that the Non-neoplastic and Inflammatory columns are by far the most predicted.

it was indeed the distribution of 3-class and 9-class labels that caused the difference in accuracies among Fitzpatrick skin types, we expected this dataset to yield a more uniformly distributed accuracy plot. While the augmented dataset for the 3-class model produced a slightly exaggerated distribution of the balanced dataset (more pronounced disparities), the 9-class model showed drastic improvement in reducing disparities between the accuracies of the Fitzpatrick types (see Figure 5). We were surprised to see this modification work for the 9-class model but not for the 3-class model. One reason this could be the case is the difference in dataset size. Data augmentations to create the 3-class augmented dataset increased the training set by about 10k images (26,262 total train), whereas the 9-class augmented training dataset held a total of 70,893 images. This decrease in disparity in a larger dataset suggests simply more data could be beneficial to creating a more uniform distribution of accuracies among skin types. Another potential reason for this observed performance is that the 9-class augmented dataset creates a more balanced distribution of labels within each Fitzpatrick type than that of the 3-class dataset. Upon examining the composition of labels per Fitzpatrick type in the imbalanced dataset, we see

that Type VI contains a notably higher percentage of the most popular label in its composition (non-neoplastic for 3-class and inflammatory for 9-class). If this trend carries over to 3-class augmentation but is mitigated in the 9-class augmentation, that would explain the difference in results between the two models' accuracy distributions.

**5. Discussion.**

Our system was effective at diagnosing skin conditions for people with skin types I-V. This is not a new discovery, but with more improvements, the accuracy of our model could be high enough to warrant use in some type of self-diagnosis tool. Of course, it should certainly not have the final say on your diagnosis, but a tool like this could have the potential to serve as an at-home, rapid test to determine if you need to make an appointment with your dermatologist.

Our system was not reliable at diagnosing skin conditions for people with skin type VI. Although it presented the highest test set accuracy for many of our experiments, those results were skewed by Simpson's paradox. In order to improve in this area, we could take multiple different steps, either in series or separately: we could augment the data in order to simulate a larger dataset (see above for a small foray into this step), we could collect and label more data so that we *actually* have a larger dataset, or we could consider different metrics to assess the quality of the model. If we were to augment the data, we could add more affine transformations on top of just rotating, cropping, and flipping. For example, we could skew images so that the skin condition has a different shape. This would make the model more robust against different angles and lens curvatures of an image. Considering different metrics would allow us to more comprehensively analyze this dataset. Three metrics come to mind when considering more comprehensive analysis: precision, recall, and accuracies on different test sets. Although this

problem is not a binary classification task — and thus precision and recall are not immediately applicable — we can group the classes in a binary manner and calculate precision and recall for a set of binary problems. For instance, we could consider the 3-class split and choose one label as our positive class and designate the other two labels as our negative class. This would leave us with three sets of precision and recall, telling us how well the machine performed on each label. Building a robust test set would also be very helpful in optimizing this model. The test set we used was randomly selected from the dataset as a whole, which means it has the same distribution as the entire dataset. This is a good strategy for a balanced dataset, but it means that the poor training of the model is not caught out in the testing phase. If we were to use a balanced test set — both in terms of skin types and skin conditions — we would be able to see how the model performs on an ideal data set and we would be able to catch any training anomalies more quickly.

Throughout the process of training and evaluating our model we have learned a lot about computer vision. For starters, we learned that unlike other types of coding, programming machine learning models should not be an iterative process. While I might test code for a website hundreds of times during a coding session, that is impossible to do with training a model. Because of this, testing pieces of code that all come together to create the final product is vital because training a model takes hours, even when running on an optimized machine. This problem is only exacerbated when the inputs are images because they are relatively large. We also learned how difficult it is to find a good dataset: our dataset was well-labeled, but smaller in magnitude than we would have liked. However, all larger possible datasets were poorly labeled or completely unlabeled. Finding a perfect dataset for dermatological projects is especially difficult because a dermatologist or biopsy is needed to give ground truth labels to images.

**6. Conclusion.**

This project shows promise for building CNNs that are robust against disproportionality of skin types in training datasets. We hypothesized that underrepresentation of darker skin tones in dermatology training datasets would lead to poor accuracy when the model evaluates skin conditions on darker-skinned patients. Although the high accuracies of Type III through VI skin initially led us to reject our hypothesis, further research and analysis found that the spirit of our hypothesis holds true: the model is not as good at predicting skin conditions for people with underrepresented skin types. The reported accuracies were inflated due to the high composition of the most common label within darker skin types in the dataset. By balancing the dataset according to label, we observed that the model was simply learning the probability distribution of the labels and guessing the common label most frequently rather than learning the class-specific features. Future research could certainly include a more in-depth analysis of the precision and recall of these results (as described above), a more thorough data augmentation system (as described above), and potentially using a two-tiered system. Such a system would consist of seven classifiers, rather than one: one classifier would classify the given image into one of the six Fitzpatrick skin types, for which there would be one trained classifier. This classifier would only be trained on images of a specific skin type, so it would theoretically be able to learn very well. This would also remove the aforementioned problem of certain skin conditions presenting differently in people of different skin types. However, this type of model would need an incredible amount of well-labeled data, given that we need to be able to train six skin-condition classifiers rather than one.

**7. Honor Code Pledge**

This paper represents our own work in accordance with University regulations.

- Alex Baroody & Katie McLaughlin

**8. Acknowledgements.**

We would like to thank Dr. Olga Russakovsky and Sunnie Kim for their help with teaching us the concepts of computer vision throughout the semester, as well as their help refining our project idea, pointing us in the direction of helpful literature, and supporting us from start to finish. We would also like to thank Nathan Inkawhich for his *excellent* tutorial on using ResNet18 through the PyTorch library (see below for access). Lastly, we would like to thank Justin Curl ‘22 for drawing our attention to the appearance of Simpson’s Paradox in our results.

**9. Works Cited**

Esteva, A., Kuprel, B., Novoa, R. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118 (2017). https://doi.org/10.1038/nature21056

Gupta, A.K., Bharadwaj, M., & Mehrotra, R. (2016). Skin Cancer Concerns in People of Color: Risk Factors and Prevention. Asian Pacific journal of cancer prevention: APJCP, 17(12), 5257–5264. https://doi.org/10.22034/APJCP.2016.17.12.5257

Groh, M. *et al.*, "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1820-1828, DOI: 10.1109/CVPRW53098.2021.00201.

Groh, M. (n.d.). Fitzpatrick17k. Retrieved December 14, 2021, from https://github.com/mattgroh/fitzpatrick17k

Inkawhich, N. (2017). *Finetuning Torchvision Models—PyTorch Tutorials 1.2.0 documentation*. https://pytorch.org/tutorials/beginner/finetuning_torchvision_models_tutorial.html

Kaufman, Bridget P. & Alexis, Andrew F. *Skin Cancer Mortality in Patients With Skin of Color*. May 2017. Retrieved December 12, 2021, from https://www.mdedge.com/dermatology/article/137341/nonmelanoma-skin-cancer/skin-cancer-mortality-patients-skin-color

Lashbrook, A. (2018, August 16). AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind. Retrieved December 11, 2021, from https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

Sprenger, J., & Weinberger, N. (2021). Simpson's Paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/paradox-simpson/

Ward, W.H., Lambreton, F., Goel N., *et al.* Clinical Presentation and Staging of Melanoma. In: Ward WH, Farma JM, editors. Cutaneous Melanoma: Etiology and Therapy [Internet]. Brisbane (AU): Codon Publications; 2017 Dec 21. TABLE 1, Fitzpatrick Classification of Skin Types I through VI. Available from: https://www.ncbi.nlm.nih.gov/books/NBK481857/table/chapter6.t1/ DOI: 10.15586/codon.cutaneousmelanoma.2017.ch6