HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Fairness Matters: Evaluating and Mitigating Bias in Skin Cancer Classification

Amy Zhang[1], Ana Zhao[2] and Andrew Han[3]

[1]Cupertino High School, USA
[2]LASA High School, USA
[3]Westlake High School, USA

ABSTRACT

Almost 1 in 5 Americans develop skin cancer by the age of 70, and it is the most commonly diagnosed cancer in the United States, with most cases being preventable. Artificial intelligence (AI) has shown great potential to accelerate the diagnosis of skin cancer for early treatment. However, the fairness of using AI for skin cancer detection has raised concerns due to the lower accuracy of "darker" skin tone detection. This paper conducts a comprehensive study on the bias problem using the Fitzpatrick dataset, while analyzing the effects of different methods towards mitigating this bias. Throughout our experiments, we found that not only was the darkest skin type biased, one of the specific light skin types also had a relatively low accuracy. Five AI models were evaluated and compared in terms of bias, and four augmentation techniques were applied to mitigate bias. Moreover, we studied the impact of training parameters (e.g. batch size, data splitting) on bias.

Introduction

Skin cancer affects people of all skin types, light or dark. Regrettably, individuals with darker skin tones often receive diagnoses of skin cancer at later stages, posing greater challenges for treatment. This condition is primarily attributed to exposure to ultraviolet light, making it nearly unavoidable. Early and accurate detection is crucial for successful treatment of skin cancer. When melanoma is identified early, the 5-year survival rate stands at 99%. However, this rate decreases to 68% if the cancer spreads to the lymph nodes and further drops to 30% if it reaches distant organs [21]. There has been a widely recognized problem within AI Skin Cancer Identification that has caused racial bias. On numerous platforms, AI based facial recognition systems have shown issues recognizing people of color. The dataset used in this research, FitzPatrick 17k [16], has a massive under-representation of skin lesion images from darker skinned populations, which can lead to discrepancies in identification accuracy. Additionally, people of color are more likely to develop rare melanoma subtypes that aren't caused by the sun, and because of this bias and the lack of data from darker skin, the delayed identification or even lack thereof can lead to much lower survival rates [17].

There is a lot of past work that has studied skin cancer detection through machine learning. For example, Melarkode et al. outlined the significant impact of skin cancer on global healthcare, stressing the importance of early detection for successful treatment [18]. However, the reliance on skin cancer specialists for diagnosis creates accessibility issues, especially in developing countries, prompting the need for automated diagnosis systems. AI methods have emerged as promising solutions to aid in early skin cancer detection, potentially reducing mortality rates associated with the disease. They explored the application of these AI techniques in skin cancer diagnosis, drawing comparisons with established datasets and review papers, while also discussing insights gained and future directions for improving automated diagnosis systems in this critical area of healthcare. A review done by Brancaccio et al. explored the use of AI in settings with human expert supervision, making the general claim that this typically offers the best results [3]. However, a significant portion of studies

have assessed accuracy in a controlled environment, which does not entirely reflect performance in a real-world setting. Furthermore, they emphasize that the practical challenges and deployment strategies of AI within skin cancer detection are very understudied. Li et al. introduced a novel approach to identifying biases learned by AI algorithms, particularly focusing on discovering unknown biased attributes of classifiers predicting target attributes in input images [15]. By utilizing a hyperplane within the generative model's latent space and a novel total-variation loss function, the study effectively detects biases that may not be readily recognized by human experts. Extensive experiments demonstrate the method's efficacy in uncovering biases and improving disentanglement regarding target attributes across diverse image domains.

While AI diagnostic tools show promise in early skin cancer detection, many models lack assessment on images of diverse skin tones and uncommon diseases. To bridge this gap, Daneshjou et al. introduces the DDI dataset, a publicly available collection of pathologically confirmed images featuring diverse skin tones [6]. Dark skin tones and uncommon diseases are identified as key factors leading to these performance drop-offs. Unfortunately, most of the published work mainly focused on the bias towards dark skin. Our experimental findings indicated that despite the underrepresentation of darker skin images, the accuracy of light skin tones also appeared to be affected. The absence of quantitative analysis in prior research highlights the issue of potential biases that may manifest directly in accuracy comparisons.

To tackle the issue of underrepresentation in datasets featuring patients with darker skin tones, we employed various strategies. These included employing data augmentation techniques to broaden and diversify the dataset, specifically incorporating a greater number of images depicting darker skin lesions. Additionally, we experimented with multiple AI models to assess their performance and ensure consistency. Furthermore, we adjusted various training parameters to observe their impact on addressing bias concerns. Throughout our research, we discovered a discrepancy from previous scientific findings regarding the skin color predominantly depicted in images, which exhibited accuracy issues. Even with thousands of images featuring one of the lighter skin types, accuracy remains relatively low. While other studies emphasize the significance of data volume, our research demonstrates the equal importance of data quality.

The rest of this paper will expand on the related works, dataset used, methodology, experimental process, and experimental results.

Related Works

Hosney et al. addressed the challenge of accurately classifying skin lesions, particularly melanoma [12]. The similarity in color images between melanoma and nevus lesions complicates detection and diagnosis, emphasizing the need for a reliable automated classification system. The proposed method employed a pre-trained deep learning network and transfer learning techniques, including fine-tuning and data augmentation.

Groh et al. introduced a novel neural network for skin cancer detection achieving a 62.4% accuracy on a three-category partition [11].In this case, random guessing theoretically yields an accuracy rate of 33.3%. Similarly, this model achieved an accuracy of 26.1% on a nine-category partition in which random guessing would yield approximately 11.1%. In their work, they utilized an original dataset, known as the FitzPatrick 17k. This dataset consists of approximately 17,000 images, spanning different categories based on skin tone.

Waweru et al. explored Deep Convolutional Neural Networks (DCNNs), as they offer improved diagnostic performance [24]. It assists by providing probability diagnoses for skin lesions. The study utilized DCNNs for automated melanoma region segmentation in dermoscopy images, ultimately developing a web tool for rapid diagnosis.

Janney et al. introduced a skin cancer identification system based on deep learning algorithms, minimizing human intervention [1]. The system utilizes past image records for training, enabling earlier treatment initiation. Evaluation metrics such as precision, sensitivity, and accuracy are employed to assess the system's efficiency, with comparative analysis conducted to gauge the effectiveness of the proposed approaches.

Jain et al. explored skin cancer diagnosis by leveraging the HAM1000 dataset [13]. They conducted a comparative analysis of six transfer learning models: VGG19, InceptionV3, InceptionResNetV2, ResNet50, Xception, and MobileNet. To address class imbalances, image replication techniques were employed for classes with low frequencies. The results highlight the efficacy of replication in improving classification accuracy and F-score while reducing false negatives.

Faghihi et al. explored skin lesion classification using DCNNs, showcasing a notable increase in classification accuracy achieved through a transfer learning framework applied to pre-trained neural networks [7]. Specifically, the fusion of VGG16 and VGG19 architectures with a modified AlexNet network, fine-tuned using a subject-specific dermatology image dataset, yielded a significant improvement in accuracy.

Salman et al. explored transfer learning, which involves adapting a pre-trained source model to a specific task, typically resulting in improved performance [20]. They uncovered a potential downside known as bias transfer, where biases from the source model persist even after adaptation to the target class. Through a series of synthetic and natural experiments, they found that bias transfer occurs in practical scenarios, such as when pre-training on standard datasets like ImageNet, and can persist even in explicitly de-biased target datasets.

Wen et al. highlighted a flaw in publicly available skin cancer image datasets, noticing that there was consistently a significant under-representation of darker skin tones in such datasets [5]. Their work also revealed that there is limited reporting on characteristics and metadata in these image datasets. Furthermore, this problem is even more prevalent in the cases of dark skin tones. They called for the implementation of more stringent quality standards for characteristics of skin image datasets, relating to our exploration of image augmentation to improve image quality.

Bissoto et al. presented a series of experiments that uncover both positive and negative biases within these datasets [2]. Their results reveal that machine-learning models can accurately classify skin lesion images even when lacking clinically-meaningful information. This indicates the presence of spurious correlations guiding the models' predictions. Moreover, attempts to introduce additional clinically-meaningful information did not significantly improve model performance, suggesting the destruction of cogent correlations.

Model biases have also been attributed to subject gender, particularly when analyzing images from the facial area. Research by Buolamwini et al. highlighted the fact that alongside skin tones, different genders have different characteristics, such as in facial shapes [4]. In their study on facial recognition models, they found that darker-skinned females are the most misclassified group while lighter-skinned males are the least. This drastic difference is the primary reason for their advocacy for reform in classification systems.

Rezk et al. addressed skin tone differences in skin cancer detection by developing a deep learning approach that generates realistic images depicting darker skin colors, enhancing the diversity of dermatology data for malignant and benign lesions [19]. Skin clinical images were sourced from DermNet NZ, the International Skin Imaging Collaboration, and Dermatology Atlas. Two deep learning methods, style transfer and deep blending, were employed to create images with darker skin tones from lighter skin images, which were then evaluated both quantitatively and qualitatively. They revealed that the style transfer method was superior in image quality, achieving a lower loss of realism score and higher disease presentation similarity score.

Dataset

In our research, we utilized the FitzPatrick 17k dataset, comprising 16,577 clinical images sourced from two dermatology atlases — DermaAmin and Atlas Dermatologico. The dataset includes Fitzpatrick skin type labels and was annotated by two data annotation services: Scale AI and Centaur Labs.

The Fitzpatrick labeling system is a six-point scale categorizing different skin colors, as shown in Figure 1, and it was originally developed for classifying sun reactivity of skin phenotype. The Fitzpatrick scale has been used in many computer vision applications to evaluate algorithmic fairness and model accuracy [4].

From the dataset given, we created our own dataset by selecting all of the images that were tumors. This new dataset contains annotated images consisting of 4320 skin tumor images: 2146 benign and 2174 malignant. It includes 105 images classified as category 6, 306 images for category 5, 668 images for category 4, 931 images for category 3, 1413 images for category 2, and 897 images for category 1. Using these images, we split the images into their 6 respective folders, and also created 3 additional folders: one with skin types 1 and 2, one with skin types 3 and 4, and one with skin types 5 and 6.

Figure 1. The Fitzpatrick scale consists of 6 levels: Types 1-6. Starting from Type 1, the lightest shade, the shade progresses to darker and darker shades as you go down the scale. The dataset we used was divided into 6 folders, each labeled with one of the 6 skin types. Shown above are some example skin images from each correlated folder. The Fitzpatrick scale image is from Emerge Tulsa [8].

Methodology/Models

We assessed the accuracies of various transfer learning and deep learning models across different skin types, such as the InceptionResNetV2 and the EfficientNet series. Our evaluations were conducted locally and on diverse platforms, including Google Colab [10] and Kaggle [14]. The InceptionResNetV2 model is a deep Convolutional Neural Network (CNN) model that is an extension of the Inception and ResNet architectures. The inception modules are designed to capture multi-scale features by using a range of different-sized filters within a single layer. InceptionResNetV2 is an architecture introduced by Szegedy et al. which also includes residual connections that allow the gradients to flow more during training [22]. These connections mitigate the vanishing gradient problem and enable the training of very deep networks. EfficientNet is a family of convolutional neural network (CNN) models designed by Tan et al. that introduce the concept of compound scaling [23]. This scales the network in multiple dimensions simultaneously, including the network depth (number of layers), width (number of channels), and input image resolution. This primarily serves as a method of increasing performance without significantly increasing computational cost. EfficientNet models consist of a series of layers and operations that progressively capture more abstract features. Both the InceptionResNetV2 and EfficientNet models are pre-trained and mutable to detect a specific object. Both of these architectures are also well-suited for more intricate object detection tasks due to their robust feature extraction capabilities.

During model training, we utilized a standard mean squared error loss function for each architecture and the Adam optimizer with a learning rate of 0.0001. This helped to mitigate discrepancies between the built-in loss functions of each model. This loss function was applied to optimize objection detection accuracy and convergence, refining the model's performance on our specific dataset. Our training process involved twenty epochs and callbacks such as ReduceLROnPleateau, EarlyStopping, and ModelCheckpoint to enhance the efficiency of our training and monitor model performance. We also conducted hyperparameter tuning experiments to optimize the model's performance. This involved adjusting parameters such as batch size, learning rate, and the number of training epochs.

We performed a detailed subclass analysis based on the Fitzpatrick scale categories, ranging from 1 to 6, to evaluate the model's performance for different skin types. The model, at this stage, was trained on the entire Fitzpatrick dataset then applied to an individual skin type. Following the training process, we utilized the classification report import from sklearn. These calculated metrics helped us to assess the model's ability to accurately classify images. By obtaining results for the model's precision, we can analyze how well the model avoids misclassification. This ensures the reliability of our results. By considering metrics for both accuracy and precision, we gain a comprehensive understanding of the model's performance and how applicable it would be in a real-world context.

The model training process involved preprocessing the dataset, dividing it into training and validation sets, and then training the models. Despite the distinct architectures of the Inception ResNetV2 and Efficient models, there was no discernible variance in their training durations.

Experimental Process

In our experiments, we evaluated several models, such as Inception ResNet V2, Efficient V2S, DenseNet121, and Xception. Our findings revealed that directories containing images of light-skinned individuals exhibited lower accuracies compared to those with darker-skinned individuals, contrary to much of the prevailing research. This is surprising because the light-skinned image files comprised a significantly larger quantity. Based on these observations, we hypothesized that there might be some interference with the training process specifically related to the light-skinned files.

First, we investigated the impact of different image augmentation methods on the accuracy of the tests. Other studies have shown that adding variance to a dataset can be helpful, as it helps models recognize more diverse and unseen data [9]. After testing the effects of multiple augmentations on our images, we ended up choosing 4: Resize, Shift Scale Rotate, Transpose, and Horizontal Flip, as illustrated in Figure 2. These augmentations were selected because they effectively altered the images, maintaining their distinctiveness from the original while preserving recognizability.

Originally, we used these augmentations to generate images for Type 6 skin in order to create a larger data set, and to see if they helped increase the accuracy. Although our results from these experiments were helpful, we later, more importantly, used them to broaden the data set for our other experiments. Next, we thought that the training split could affect how the accuracies turned out. In order to test this, we first used an 80-20 data split for training, validation, and prediction – with 80% of the data used for training and 20% of the data used for validation and prediction. Next, we tested a 70-30 data split between the training and testing datasets.

Subsequently, we studied the six files by restricting the number of input images to the minimum, aiming to determine whether the problem lay solely with the dataset or stemmed from another source. First, we ran data using 105 images from each of the files (which is the number of images of Type 6 skin we have), which expectedly would reduce all of the accuracies. We created code to randomly select 105 images from each file in Google Colab, and created 6 new files. After running them separately, we compared their accuracy differences from the original tests. Then, we replicated this procedure with 306 images (representing Type 5 skin),

utilizing the augmentations employed in the prior experiment to generate additional images for Type 6 skin, which originally consisted of only 105 images. Finally, we created six folders, each containing 668 images (representing the quantity of Type 4 skin images). Likewise, we applied identical augmentations to generate additional sets for the Type 6 and Type 5 skin folders, ensuring their sizes matched the 668-image benchmark.

Lastly, we examined how varying batch sizes during model training affected overall accuracy. By comparing batch sizes of 64 and 32, we aimed to determine whether this factor influenced the accuracy gap between different skin tones.

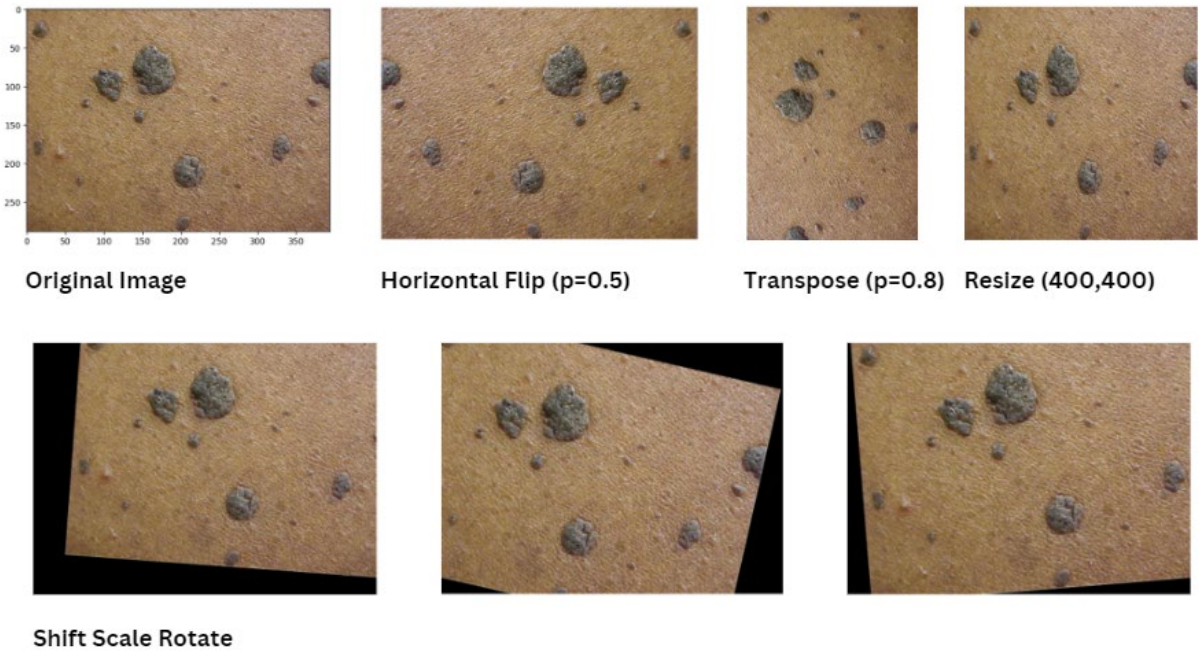

Figure 2. Using a sample Type 5 skin image, we performed the augmentations as labeled. As you can see, the "Horizontal Flip" function reflects the image across the y-axis, the "transpose" function flips the orientation of the image, the "Resize" function generates the image with a new length and width, and finally, the "Shift Scale Rotate" function keeps the same original size but slightly skews the image along that frame.

Experimental Results

The Evidence of Bias

We used the entire dataset to train the CNN Baseline model and test the model (80% for training and 20% for testing). The CNN Baseline model contains four Convolution and Pooling pairs, followed by a flatten layer, which is usually used as a connection between Convolution and the Dense layers. The dense layers are the ones that are mostly used for the output layers. The activation used was the "Softmax," which gives a probability for each class and they sum up to 1. The purpose of this experiment was to establish a baseline for comparison with our subsequent results. This allowed us to assess the extent of impact our subsequent experiments had. We repeated this process five times and depicted our findings in Figure 3. The bias is evident because the accuracies

vary greatly across different skin types, with Type 2 exhibiting the lowest accuracy and Type 5 showing the second lowest accuracy. Surprisingly, Type 6 demonstrates average accuracy, despite that it only has 105 training images. This indicates that the prediction accuracy was not only affected by the quantity of training data, but also the quality of training data.

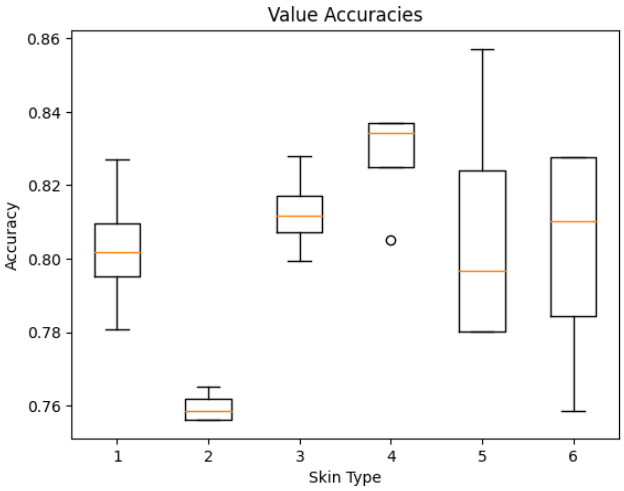

Figure 3. Box plot of accuracies for the entire dataset

The Impact of Model Type On Bias

To explore the influence of various models on skin cancer detection bias, we assessed one deep-learning-based model alongside four transfer learning models. Deep-learning-based systems for skin cancer classification have shown potential in enhancing detection accuracy. The model we employed for this purpose was the CNN-Baseline. Transfer learning models leverage knowledge gained from prior machine learning tasks and apply it to a new, but related, problem.

Accuracy for Different Model Types

Model Type	Accuracy for Skin Type 1-2	Accuracy for Skin Type 3-4	Accuracy for Skin Type 5-6	Notes
CNN Model (Baseline)	64.96%	68.32%	65.51%	Deep-learning
Inception ResNet V2	77.01%	82.92%	81.61%	Transfer Learning
Efficient V2S	77.46%	87.05%	82.76%	Transfer Learning
DenseNet121	74.78%	82.64%	73.56%	Transfer Learning

Xception	80.80%	87.05%	82.76%	Transfer Learning

From our results, we observed that the transfer learning models had a higher accuracy compared to the deep-learning model. Specifically, the Xception model increased the baseline model accuracy from around 65% up to an accuracy of around 85%. This shows that using transfer learning towards the data set does in fact create an increase in accuracy, and therefore, better results. Through this experiment, we also observed that the accuracy for skin type 3-4 is the highest while the accuracy for skin type 1-2 is the lowest.

The Impact of Skin Types in Dataset On Accuracy

Given the Xception model's high accuracy, we selected it for this experiment. To assess the impact of diversity within the training dataset, we initially trained the Xception model using all available images and evaluated it on the complete dataset. After that, we trained the Xception model exclusively on skin type 1 and 2 images, yet tested it on the entire dataset. We then compared the outcomes of these two approaches to discern any disparities.

The objective of this experiment was to ascertain whether the inclusion of dark skin images in the training dataset would enhance overall accuracy. This investigation aimed to demonstrate whether the distribution of skin type categories in the training set would affect accuracy.

Accuracy for Training Dataset with All Images Vs Training Datasets with Images from Skin Types 1 And 2

Training Dataset	All Images	Skin types 1-2	All Images	Skin type 1-2	All Images	Skin types 1-2
Testing dataset	Skin types 1-2	Skin Type 1-2	Skin types 3-4	Skin types 3-4	Skin types 5-6	Skin types 5-6
Images in test set	343	N/A	269	1599	62	411
Ben Precision	83%	N/A	85%	80%	91%	73%
Mal Precision	80%	N/A	86%	79%	79%	78%

Average	82%	N/A	85.5%	79.5%	85%	76%

The "N/A" labels denote instances where the model was initially trained on skin types 1 and 2, thus we refrained from retesting it with the same skin types. Our findings revealed that the accuracy consistently improves when testing data with the model trained on the entire dataset, compared to the model trained solely on skin types 1 and 2. This shows the significance of enhancing diversity in our training dataset, as it correlates with an overall increase in accuracy across all skin types.

The Impact of Equalizing the Distribution of Images

To thoroughly assess the correlation between data volume and accuracy, we conducted three distinct experiments. In the first experiment, we randomly sampled 400 images from each skin type group (1-2, 3-4, and 5-6). This selection was designed to equalize the image count across all skin type groups, given that skin types 5-6 originally contained 411 images. Both benign and malignant categories were represented in the samples to maintain a 1:1 ratio between the two categories. This resulted in a total of 1200 images. Utilizing the Inception ResNet V2 model, we employed an 80-20 split, allocating 80% of the images for training and 20% for testing.

We repeated the selection process 3 times, creating 3 random sets to use for training. Then, we used each of the random sets 3 times, and showed the average accuracies in the following table.

Experiment 1 Separated Average Accuracies

Random Set	Accuracy for Skin Types 1-2	Accuracy for Skin Types 3-4	Accuracy for Skin Types 5-6
1	76.19%	70.65%	75.50%
2	67.14%	74.64%	73.98%
3	76.32%	84.15%	80.84%

Although it is hard to determine whether the overall accuracy improved due to the randomness, we can still see that in most cases, the accuracy for skin types 5-6 was greater than the accuracy for skin types 1-2.

Similarly, because skin types 3-4 have around 1500 images, in our second experiment, we randomly selected 1500 images from skin types 1-2 and combined those images with all of the original images from skin types 3-4 and 5-6. We selected samples from both the benign category and also the malignant category to maintain a 1:1 ratio. In total, the dataset comprised around 3400 images, with 80% allocated for training and 20% for testing.

The purpose of this experiment was similar to the previous experiment. However, this time, we based the baseline number on the images in the skin types 3-4 category, while maintaining the number from skin types 5-6.

Additionally, we select only one random selection set, as choosing more would offer limited variety due to the large number of images required.

Experiment 2 Accuracies

Trial #	Accuracy for Skin Types 1-2	Accuracy for Skin Types 3-4	Accuracy for Skin Types 5-6
1	75%	83%	77%
2	78%	78%	77%
3	77%	82%	82%
Average	77%	81%	79%

Similar to our initial experiment, the accuracy could have varied significantly due to the randomness in the selection of 1500 images. As depicted in the above table, the accuracy for skin types 3-4 is notably higher. This discrepancy between skin types 3-4 and 5-6 is expected, given the considerably lower number of images from the latter category. However, the overall accuracy for skin types 5-6 appears to surpass that of skin types 1-2, reinforcing concerns regarding the dataset's integrity for skin types 1-2. This observation aligns with our findings from the first experiment, suggesting a potential bias in our skin cancer detection leaning either towards skin types 3-4 and 5-6 or against skin types 1-2.

The Impact of Training with Only One Group of Skin Type Images

To test what would happen if we only trained our dataset with one group of skin type images, like we did in the very first experiment, we decided to train the model using this strategy with each group of images. We also wanted to see if the accuracy of a specific skin type would be close to 100% if we trained the model using only images from that skin type and test if some of the training datasets were potentially harmful to the overall accuracy rather than helping it. Then we tested that model with all the skin type images.

Results from Training Using Skin Types 5-6

	Accuracy for Skin Types 1-2	Accuracy for Skin Types 3-4	Accuracy for Skin Types 5-6
1	68.24%	69.84%	99.51%
2	65.72%	65.78%	99.51%

3	66.49%	66.86%	100%
Average	66.82%	67.49%	99.68%

Results from Training Using Skin Types 3-4

	Accuracy for Skin Types 1-2	Accuracy for Skin Types 3-4	Accuracy for Skin Types 5-6
1	76.28%	99.68%	80.79%
2	77.54%	99.81%	78.59%
3	76.60%	99.62%	79.56%
Average	76.81%	99.70%	79.64%

Results from Training Using Skin Types 1-2

	Accuracy for Skin Types 1-2	Accuracy for Skin Types 3-4	Accuracy for Skin Types 5-6
1	99.76%	79.78%	75.43%
2	99.72%	78.52%	74.70%
3	99.92%	78.83%	73.72%
Average	99.80%	79.05%	74.61%

As depicted in the tables, when the model was trained exclusively on a single skin type, the accuracy for that particular skin type approached 100% across all three training sessions. However, the lower accuracies observed

when training solely on skin types 5-6 cannot draw definitive conclusions due to the considerably smaller dataset size. Yet, by comparing the accuracies between training with skin types 3-4 and skin types 5-6, a trend emerges. It is evident that the accuracy for skin types 5-6 consistently improved when trained alongside skin types 3-4, as opposed to training solely on skin types 1-2. Despite the anticipation that training with a larger dataset (such as 1-2, which contains nearly 1000 more images than 5-6) would yield higher accuracy, these findings suggest that training with skin types 1-2 might detrimentally impact the accuracy of skin types 5-6, contrasting with the effects of training with skin types 3-4.

The Impact of Augmentation in Bias

Utilizing the augmentations delineated in Figure 2, we conducted an experiment to assess whether augmentation could mitigate bias. The experiment comprised two parts: one involving augmentation and the other without it.

Part I: Given that skin type 6 has 105 images, we randomly selected an equal number of images from each of the other skin type groups (1, 2, 3, 4, and 5) and amalgamated them. Employing a 70-30 split, we reserved 30% of the dataset for testing purposes. Subsequently, 80% of the remaining data was allocated for training, with 20% of that subset designated for validation. Consequently, each category comprised approximately 58 images for training, 15 images for validation, and 32 images for testing. We repeated this process 30 times, each time selecting a random set of 105 images from each skin type, and computed their average accuracies. Finally, we generated a boxplot representing the accuracies for all skin types, as illustrated in Figure 4.

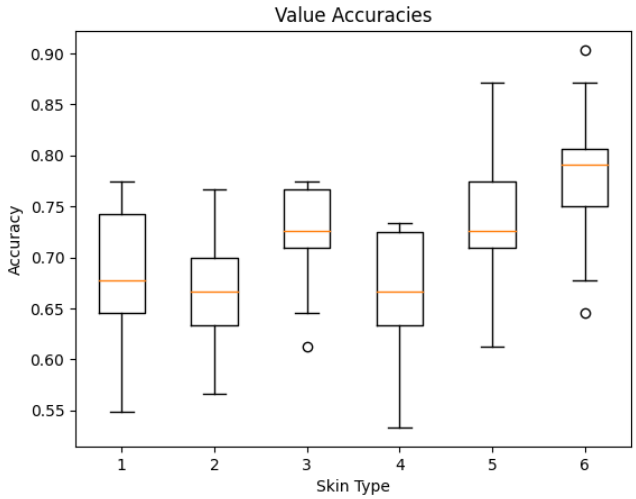

Figure 4. Box plot for training dataset consisting of 105 images per skin type

Part II: We did the same as the first part of the experiment, but instead of using 105 images, we chose 306 images because skin type 5 has 306 images. Since skin type 6 only had 105 images, we used augmentation to increase the number to 306. We chose a random set and ran it 40 times and Figure 5 shows all skin types' accuracies in a box plot.

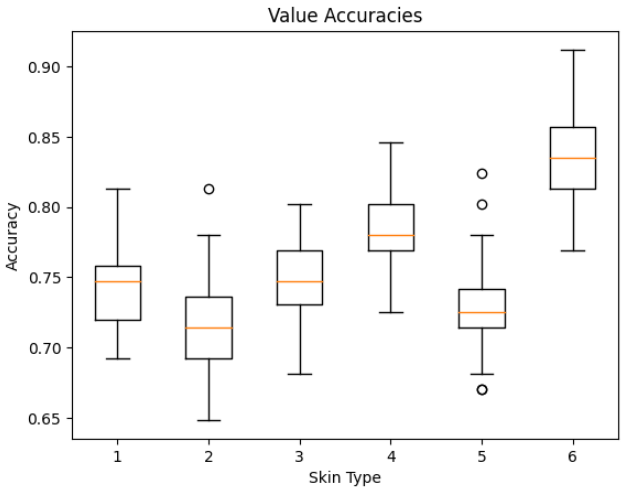

Figure 5. Box plot for training dataset consisting of 306 images per skin type

Part III: Leveraging the outcomes derived from parts I and II, we plotted them in Figure 6 as a line chart. Each point on the line signifies a distinct skin type. Notably, upon incorporating augmented images for skin type 6, there was a significant surge in accuracy. This experiment demonstrates the efficacy of augmenting images to balance the distribution across different skin types in our dataset, leading to heightened accuracy across most categories, with a particular improvement in the category targeted by the augmentations.

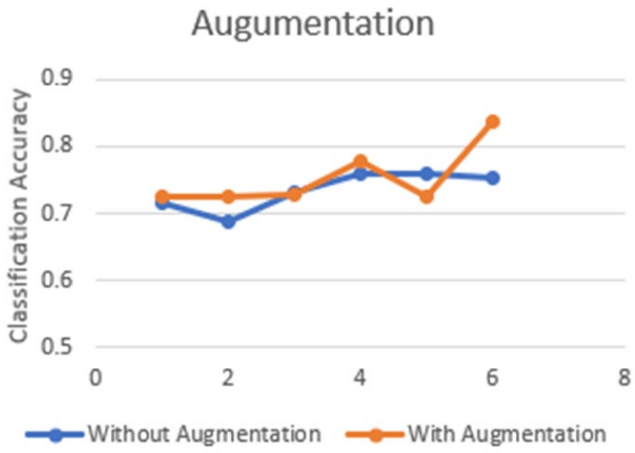

Figure 6. Graph comparing the results from parts I and II.

The Impact of Batch Size On Bias

To examine if altering the batch size in our testing would lead to a more equitable accuracy distribution among skin types, we compared the results obtained using batch sizes of 64 and 32. In our experiment, we randomly assembled 105 images from the other skin type groups and merged them. Employing a 70-30 data split, the outcomes are presented in Figures 7 and 8.

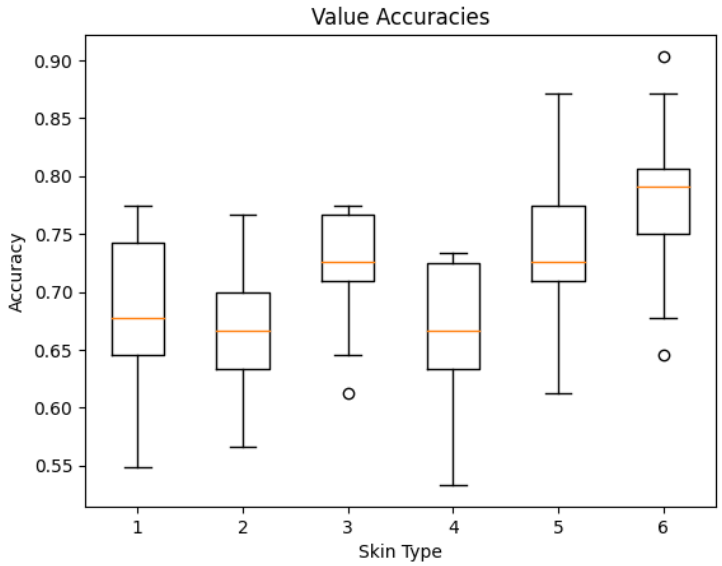

Figure 7. Box plot for batch size of 64

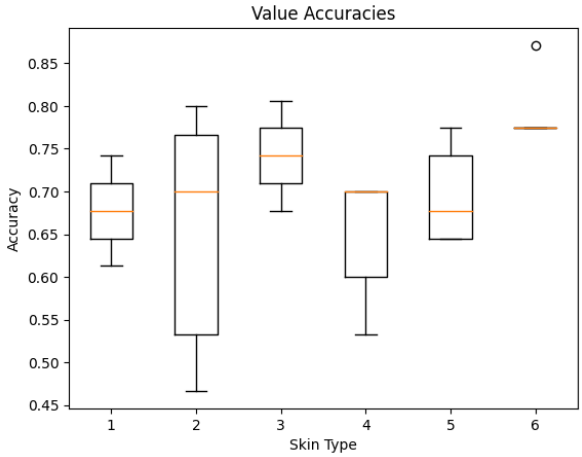

Figure 8. Box plot for batch size of 32

When comparing the means of each skin type, most exhibited minimal changes in accuracy between batch sizes of 64 and 32, except for skin type 5. The accuracy for skin type 5 declined from approximately 74% with a batch size of 64 to around 69% with a batch size of 32. From this experiment, we infer that while the correlation isn't robust, a batch size of 64 may be preferable over a batch size of 32.

Conclusion

This research paper aimed to examine how different skin types impact the accuracy of AI models in predicting whether a skin mark is a tumor. The training utilized a dataset sourced online, containing images of benign and malignant tumors across various skin types, ensuring applicability to a broad demographic. Analysis of results with similar data splits from each skin type revealed a bias against skin types 1-2.

The study's findings shed light on several factors influencing the overall accuracy of skin cancer detection. Transfer-learning models achieved a maximum accuracy of 87.05%, whereas deep-learning models reached a maximum of 68.32%. Increasing dataset diversity emerged as a significant factor in enhancing accuracy across all skin types. Generally speaking, training with larger and more diverse datasets yielded higher average accuracy. Moreover, our research underscores the importance of image type/quality alongside quantity, as evidenced by lower accuracy for skin types 5-6 when trained with skin types 1-2, despite the latter comprising over 2000 images compared to 3-4's approximately 1500 images.

Exploration of alternative data preprocessing techniques, such as image augmentation, may further enhance performance. Additionally, varying batch sizes showed slight accuracy fluctuations, with a batch size of 64 exhibiting higher overall accuracy compared to 32. Additional experimentation may be necessary to further validate this disparity.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Bethanney Janney, J., Krishnamoorthy, N. R., Divakaran, S., Sudhakar, T., Krishnakumar, S., & Akshya., V. (2021). Diagnosis of skin malignancy using deep learning approaches. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA). https://doi.org/10.1109/icaeca52838.2021.9675722
[2] Bissoto, A., Fornaciali, M., Valle, E., & Avila, S. (2019). (DE) constructing bias on skin lesion datasets. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/cvprw.2019.00335
[3] Brancaccio, G., Balato, A., Malvehy, J., Puig, S., Argenziano, G., & Kittler, H. (2024). Artificial Intelligence in skin cancer diagnosis: A reality check. Journal of Investigative Dermatology, 144(3), 492–499. https://doi.org/10.1016/j.jid.2023.10.004
[4] Buolamwini, J., & Gebru, T. (2018, January 21). Gender shades: Intersectional accuracy disparities in commercial gender classification. PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html
[5] Characteristics of publicly available skin cancer image datasets: A systematic review - The Lancet Digital Health. (n.d.). https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00252-1/fulltext
[6] Daneshjou, R., Vodrahalli, K., Liang, W., Novoa, R. A., Jenkins, M., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Zou, J., & Chiou, A. (2021, November 15). Disparities in dermatology AI: Assessments using diverse clinical images. arXiv.org. https://arxiv.org/abs/2111.08006
[7] Faghihi, A., Fathollahi, M., & Rajabi, R. (2024, April 1). Diagnosis of skin cancer using VGG16 and VGG19 based transfer learning models. arXiv.org. https://arxiv.org/abs/2404.01160
[8] Fitzpatrick scale at Emerge in Tulsa | Wellness, med spa, salon. (2023, March 28). Emerge Medical. https://emergetulsa.com/fitzpatrick/
[9] Galdran, A., ∗, Alvarez-Gila, A., Meyer, M. I., Saratxaga, C. L., Ara´Ujo, T., Garrote, E., Aresta, G., Costa, P., Mendonc¸¸A, A. M., & Campilho, A. (2017). Data-Driven Color augmentation Techniques for Deep skin image analysis [Journal-article]. https://arxiv.org/pdf/1703.03702.pdf
[10] Google Colab. (n.d.). Colab.google. colab.google. https://colab.google/

[11] Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021, April 20). Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17K dataset. arXiv.org. https://arxiv.org/abs/2104.09957
[12] Hosny, K. M., Kassem, M. A., & Foaud, M. M. (2018). Skin cancer classification using Deep Learning and Transfer Learning. 2018 9th Cairo International Biomedical Engineering Conference (CIBEC). https://doi.org/10.1109/cibec.2018.8641762
[13] Jain, S., Singhania, U., Tripathy, B., Nasr, E. A., Aboudaif, M. K., & Kamrani, A. K. (2021). Deep learning-based transfer learning for classification of Skin cancer. Sensors, 21(23), 8142. https://doi.org/10.3390/s21238142
[14] Kaggle: your machine learning and data science community. (n.d.). https://www.kaggle.com/
[15] Li, Z., & Xu, C. (2021). Discover the unknown biased attribute of an image classifier. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv48922.2021.01470
[16] Mattgroh. (n.d.). GitHub - mattgroh/fitzpatrick17k. GitHub. https://github.com/mattgroh/fitzpatrick17k
[17] Melanoma & Skin of Color - Melanoma Research Alliance. (n.d.). Melanoma Research Alliance. https://www.curemelanoma.org/about-melanoma/people-of-color
[18] Melarkode, N., Srinivasan, K., Qaisar, S. M., & Plawiak, P. (2023). Ai-powered diagnosis of Skin cancer: A contemporary review, open challenges and future research directions. Cancers, 15(4), 1183. https://doi.org/10.3390/cancers15041183
[19] Rezk, E., Eltorki, M., & El-Dakhakhni, W. (2022). Improving skin color diversity in cancer detection: Deep Learning Approach. JMIR Dermatology, 5(3). https://doi.org/10.2196/39143
[20] Salman, H., Jain, S., Ilyas, A., Engstrom, L., Wong, E., & Madry, A. (2022, July 6). When does bias transfer in transfer learning?. arXiv.org. https://arxiv.org/abs/2207.02842
[21] Skin cancer. (2022, April 22). https://www.aad.org/media/stats-skin-cancer#:~:text=The%20five%2Dyear%20survival%20rate,the%20lymph%20nodes%20is%2099%25.&text=The%20five%2Dyear%20survival%20rate%20for%20melanoma%20that%20spreads%20to,and%20other%20organs%20is%2030%25.
[22] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016, August 23). Inception-V4, inception-resnet and the impact of residual connections on learning. arXiv.org. https://arxiv.org/abs/1602.07261
[23] Tan, M., & Le, Q. V. (2020, September 11). EfficientNet: Rethinking model scaling for Convolutional Neural Networks. arXiv.org. https://arxiv.org/abs/1905.11946
[24] Waweru, A. K., Ahmed, K., Miao, Y., & Kawan, P. (2020). Deep learning in skin lesion analysis towards cancer detection. 2020 24th International Conference Information Visualisation (IV). https://doi.org/10.1109/iv51561.2020.00130