

Mitigating Algorithmic Bias in Machine Learning Models for Skin Cancer Detection

Word Count: 4938

Introduction

Skin cancers are the most common cancers (Cassidy et al., 2022). If detected and treated early, skin cancer is unlikely to result in fatality (American Academy of Dermatology Association, 2022). Unfortunately, there is a global shortage of dermatologists to detect early signs of skin cancer, especially in remote areas and developing countries (Li et al., 2022). To fill this need, digital teledermatology solutions based on machine learning (ML) algorithms trained on massive public datasets (Li et al., 2022) have become increasingly popular. In recent years, teledermatology models have shown comparable or better diagnostic performance than professional dermatologists (Wen et al., 2022). These models are especially prevalent in the diagnosis stage, using algorithms that classify images of skin lesions as malignant or benign.

Some dermatologists expressed concerns about teledermatology, citing algorithmic bias stemming from the historical biases and underrepresentation of certain skin tones in popular dermatology datasets (Butt et al., 2021, Fathy & Lipoff, 2022). Approaches to mitigate this bias include initiatives for more representative datasets (Ganapathi et al., 2022) and directly modifying algorithms (Rezk et al., 2022a). Others advocate for unlocked algorithms that will be retrained as more representative datasets are released (Mittermaier et al., 2023).

Literature Review

History of Algorithmic Bias

According to Fathy & Lipoff (2022), underrepresentation in training materials for dermatologists and machine learning models deteriorates the standard of care for patients of color. White-centric practices outside of medicine, like the use of “Shirley Cards” to calibrate early color cameras (Butt et al., 2021), have contributed to underrepresentation and improper

capture of darker skin tones. Today, this “white lens phenomenon” persists in dermatology textbooks, where “merely 4% to 18% of the total number of images” (Rezk et al., 2022b) are of darker skin tones.

Machine Learning in Dermatology

The Convolutional Neural Network

Existing literature describes ML as a “formidable tool to potentially assist dermatologists in their diagnostic decisions” (Cullell-Dalmau et al., 2020). Amongst neural networks used in teledermatology, the Convolutional Neural Network (CNN) is the most popular because it is “well-suited for a variety of tasks in computer vision” (Amini, 2024). Besides dermatology, CNNs are used in pathology, radiology, and other medical disciplines (Li et al., 2022). In dermatology, CNNs detect skin cancer by analyzing “dermoscopic images only or in combination with regular photographic images” (Cullell-Dalmau et al., 2020).

General CNN models are retrained on smaller, more specific datasets (Google for Developers, 2023); researchers often make use of transfer learning to specialize past winners of the ImageNet Large Scale Visual Recognition Challenge (Stanford Vision Lab et al., 2020) for skin cancer diagnosis.

Diverse Datasets

Abubakar et al. (2020) analyzed bias against darker skin tones in CNNs for burn detection. They concluded that a “lack of racial or diverse ethnic inclusion during the training process of a machine learning algorithm tends to produce... unrealistic model[s],” which are less generalizable to real-world applications. Similarly, Rezk et al. (2022b) reported that “models trained on data with a certain skin color range could not be generalized when tested on data collected from a different population,” corroborating Abubakar et al. (2020).

Approaches to Debiasing ML

Wen et al. (2022) analyzed images in open-access datasets, finding a “lack of transparency in metadata reporting for clinically essential characteristics,” which limited the utility of the images and compounded the effects of skin tone underrepresentation.

Policy Initiatives

Numerous researchers have advocated for transparent, standardized data collection for future ML models. Government agencies, like the U.S. Food and Drug Administration, emphasized the importance of future regulation for AI systems (Mittermaier et al., 2023). The STANDING Together initiative addresses similar concerns, setting standard practices to develop criteria for evaluating large datasets used to train ML models (Ganapathi et al., 2022). Furthermore, the American Academy of Dermatology changed its residency curriculum and educational resources (Fathy & Lipoff, 2022).

Following these policy initiatives, “the first publicly available, deeply curated, and pathologically confirmed image dataset with diverse skin tones” was published by researchers at Stanford University (Daneshjou et al., 2022).

Adding Skin Tone Metadata

Other researchers used mathematical algorithms to add metadata to existing images. Kinyanjui et al. (2019) calculated individual topology angles (ITA) and converted them to one of the 6 FST categories. Groh et al. (2022) refined Kinyanjui’s calculations using ML algorithms. Both researchers’ ITA-FST conversion algorithms have been automated into the Python package *derm-ita* (*Derm-ita*, 2021). However, ITA-FST conversion for skin tone has drawn criticism because FST measures how skin tans or burns in sunlight, not skin tone (Groh et al., 2022).

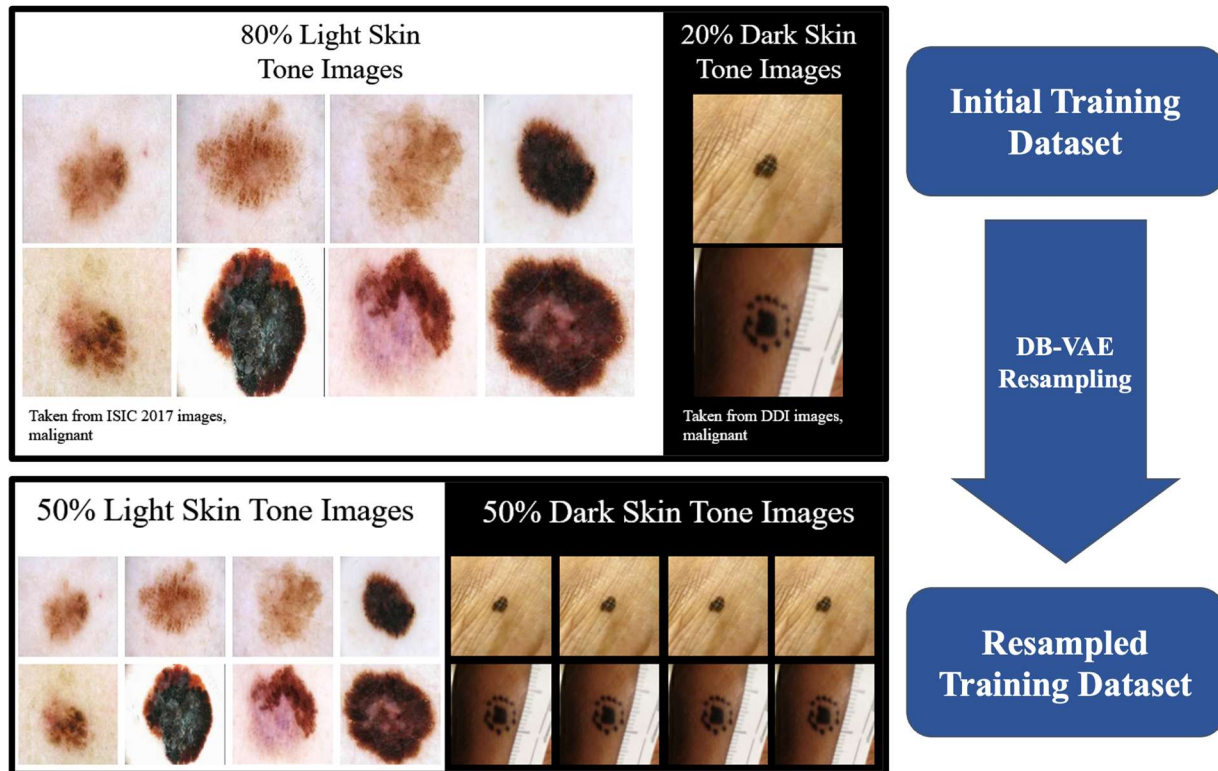
This has led to the development of alternative skin tone scales. Recently, Google released its Monk Skin Tone (MST) Scale (Skin Tone Research @ Google AI, n.d.), but other researchers have opted for their other categories. For example, Rezk et al. (2022b) used the categories “very light, light, intermediate, tan, brown, or black.”

Learning Latent Structure

Some researchers view metadata labeling as infeasible because it requires annotation of massive datasets and human validation of annotations by dermatologists. The debiasing variational autoencoder (DB-VAE, Figure 1) architecture circumvents this limitation, learning a dataset’s latent skin tone distribution to “increase the probability of selecting rarer data for training” (Amini et al., 2019).

Figure 1

Visual Representation of DB-VAE Resampling



Generating Artificial Images

Last, some researchers have deviated from looking at real images of underrepresented skin tones entirely. Rezk et al. (2022a) investigated artificial image generation to supplement datasets that underrepresent darker skin tones. Rezk et al. combined images of lesions on lighter skin tones with images of darker skin tones using style transfer, generating artificial yet realistic images of lesions on darker skin tones (Figure 2). When these artificial images were used in model training, Rezk et al. found that model performance increased, like how real images of underrepresented skin tones boosted model performance in Abubakar et al. (2020).

Figure 2

Visual Representation of Neural Style Transfer for Artificial Image Generation



The Gap

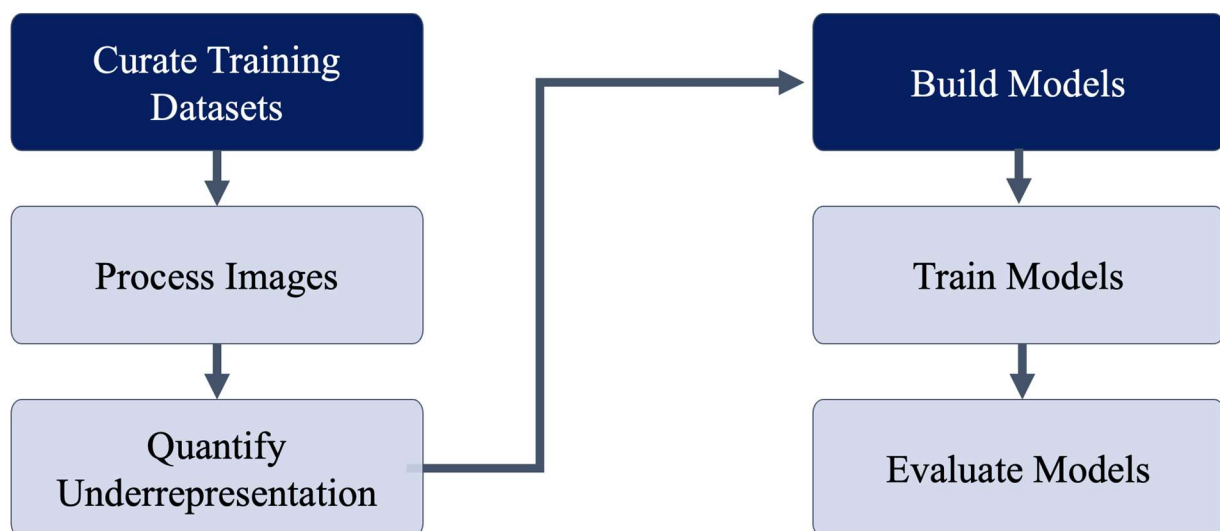
Much research has been done on the use of variational autoencoders (VAE) to generate artificial images, such as generating hairless images (Bardou et al., 2022) or complex pipelines that incorporate VAE into a series of different neural networks to “improve the quality of the samples produced” (Kebaili et al., 2023). However, previous researchers have not combined the DB-VAE architecture in a pipeline with artificial image generation. Furthermore, few researchers have implemented the Diverse Dermatology Images (DDI) dataset (Daneshjou et al., 2022) into model training.

Of the review articles that mention DDI and review various debiasing approaches, none compare the approaches on a uniform standard. For example, Petersen et al. (2023) analyzed future paths to help medical practitioners mitigate the different causes of bias in ML models. Other articles discussed risk factors beyond skin tone, including “mass index, education, insurance type, geography, and genetics” (Chen et al., 2023). Li et al. (2022) analyzed dermatology ML and future trends; none of these reviews ran ML models on a standardized architecture. Debelee (2023) addressed variational autoencoders based on “extracted features,” akin to Amini et al. (2019), and covered Rezk et al.’s artificial image generation, but still did not run any models. Rezk et al. (2023) cover artificial image generation and summarize different debiasing approaches, but mostly focus on summarizing common data challenges and various implementations without running any models.

Method

Figure 3

Summary of Method



The goal of this study is to answer the question: what is the best way to mitigate algorithmic bias in machine learning malignance classifiers used for dermatology? This research question is best answered with experimental research, holding certain variables constant and isolating the effect of each bias-mitigating approach. Isolating variables and holding others constant makes the study more robust and rigorous (ByPass Publishing, 2013). The research methodology is outlined above in Figure 3.

Dataset Curation

One consideration for datasets is size. After processing, each dataset has around 2000 images, within the precedent ranging from 200 to over 150,000 images (Li et al., 2022). To avoid memory allocation errors and hardware limitations, images used in this study were downsampled to a lower resolution of 64x64 pixels and saved into compressed .h5 files.

Diverse Dermatology Images Dataset (DiDI)

The Diverse Dermatology Images Dataset (Daneshjou et al., 2022) contains 656 images and corresponding metadata with identification codes, Fitzpatrick Skin Types, Boolean values for malignance, and clinical diagnoses. The metadata reduced the I-VI FST scale to a three-category scale: “12” for types I and II, “34” for types III and IV, and “56” for types V and VI.

Processing

Clinical diagnoses were irrelevant to the study because the classification networks differentiate between benign and malignant lesions without classifying the specific dermatological disorder. Boolean malignance values were converted into binary values: 1 for true/malignant and 0 for false/benign. After processing, each of the 656 DiDI images had a 64x64 resolution with metadata containing the image’s ID, FST value, and malignance label.

International Skin Imaging Collaboration Dataset (ISIC)

The next data source is the International Skin Imaging Collaboration (ISIC), which published 75.2% of all skin lesion images between 2013 and 2020 (Wen et al., 2022). ISIC Challenge datasets from 2016-2020 are publicly available to download; of the available challenge datasets, 2016 and 2017 were selected because both were often used in previous machine learning work (Cassidy et al., 2022, Li et al., 2022). The 2017 Challenge provided metadata about sex, age, and lesion diagnosis as melanoma, seborrheic keratosis, or neither. Melanoma is a malignant skin lesion disorder, while seborrheic keratosis is benign. The 2016 Challenge metadata merely labels each image as “malignant” or “benign.”

Processing

The 2017 Challenge also included super-pixel images like Figure 4 and binary mask images like Figure 5, which were not used in this study.

Figures 4-5

Examples of images removed from the dataset

Figure 4
ISIC_0012086_superpixels.png
(Codella et al., 2017)

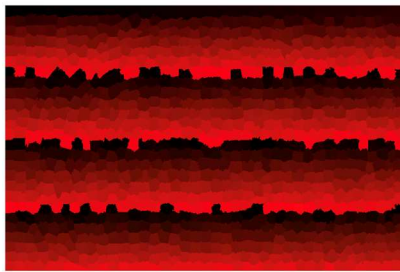
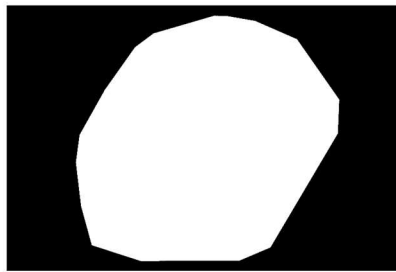


Figure 5
ISIC_0012086_segmentation.png
(Codella et al., 2017)



Duplicate images from the 2016 and 2017 Challenges were removed. To match the metadata for DiDI, metadata from the 2016 challenge was used to assign a malignance value: 1 for malignant and 0 for benign. For images that were only included in the 2017 Challenge, images labeled as melanoma were assigned as malignance value 1, sebhorrheic_keratosis as 0,

and neither as 2. All images with malignance value 2 were removed because the classifier would not have labels for those images.

After pruning the dataset, 1976 ISIC images were included in the study. Skin tone values were assigned using the `get_cropped_center_ita` function (*Derm-ita*, 2021), which calculates an individual topology angle by sampling pixels from the image—excluding the center because “many dermoscopic images have skin lesions in the center of the image which could throw off the ITA result” (*Derm-ita*, 2021). This ITA value was converted to a Fitzpatrick Skin Type using ranges established by Groh et al. (2021), then converted to a DDI skin tone category: “12,” “34,” or “56.” After processing, each of the 1976 ISIC images had 64x64 resolutions with metadata containing the image’s ID, FST value, and malignance label.

Artificially Generated Images Dataset (ArGI)

Sample code for neural style transfer (TensorFlow, n.d.) was adapted for the artificial generation of lesion images on underrepresented skin tones. To match the size of DiDI, 656 images were artificially generated. Images of various skin tones were used to stylize content images containing lesions.

Content images were selected by using Python to generate 656 random integers from 0-1975. These integers corresponded to the indices of ISIC images to be used as content images.

Stylizing the content images had several steps:

1. The malignance metadata from the content image was copied for the final stylized image’s metadata.
2. Then, using another random integer generator from 0-43, one of the 44 style images taken from AlexKaiLe (n.d.) was sampled as the style image.

3. Neural style transfer was used to stylize the content image (224x224 pixel resolution) with the style image (256x256 pixel resolution). Higher resolution style images preserved the detail of the content image after style transfer,

Last, each stylized image was processed using the same ITA-FST labeling system, adding an FST value of “12,” “34,” or “56.” After processing, each of the 656 artificially generated images had a resolution of 64x64 images, with metadata containing the image’s FST value and malignance label.

Quantify Underrepresentation

The underrepresentation of darker skin tones will be analyzed for each of the datasets in the study using chi-square goodness of fit tests, which test if samples “came from a population with a specific distribution” (NIST, n.d.). In this study, the specific distribution is the US national average FST values (Keiser et al., 2012). This test, using a significance value of $p < 0.05$, is modeled after Reilley-Luther et al. (2020), which conducted a chi-square analysis to quantify skin tone underrepresentation in dermatology textbooks.

Model Building

According to Li et al. (2022), “GoogleNet [AKA Inception-V1, (Szegedy et al., 2014)] Inception-V3, V4, ResNet, Inception-ResNet V2, and Dense Net” are the “most implemented CNN architectures in the field of dermatology.” Using Keras API applications (Keras Team, n.d.), each network architecture was trained on a subset of 100 ISIC images.

This study was conducted on a PC with 32 GB of RAM and an NVIDIA GeForce 1660 Ti GPU. Similar hardware configurations have been used in ML research (Cai et al., 2021), but many network architectures are not trainable on the 1660 Ti since it only has 6 gigabytes of video random-access memory. No Inception network completed training due to memory

allocation errors. A similarly sized DenseNet201 model also failed to complete training. ResNet50v2 completed training without any errors, so it was the standard CNN architecture and DB-VAE’s encoder architecture. DB-VAE’s decoder uses a sequence of convolutional layers to reconstruct images from the encoder’s output. Further considerations for architecture selection are in Appendix A.

Model Training

For ease of reference, each of the 6 models was assigned a Model ID. These are outlined in Table 1 below.

Table 1

Summary of Models, based on Architecture and Training Dataset

Model ID	Training Datasets	Model Architecture
1	ISIC	Standard Classifier
2	ISIC+ DiDI	
3	ISIC+ArGI	
4	ISIC	DB-VAE
5	ISIC+DiDI	
6	ISIC+ArGI	

Hyperparameters

Abubakar et al. (2020) used 200 epochs, Amini et al. (2019) used 50 epochs, and Rezk et al. (2022b) only used 30 epochs. Abubakar et al. noted that “the performance stabilized even before epoch 200,” implying that 200 epochs were excessive to reach stable accuracies. Thus, all models were trained for 50 epochs.

Keras Team (n.d.) used a batch size of 64, but this caused memory issues and performance problems on the training computer. All models completed training when the batch size was halved to 32.

Model Evaluation and Significance Testing

Each model was tested on subsets of ISIC and ISIC+DiDI to represent what classification networks might encounter in clinical settings. ArGI was not used for model evaluation because its images were not directly taken from dermatology clinics.

Each model test consisted of 30 evaluation runs; each run contained 256 images. This sample size is sufficiently large to assume a normal distribution according to the Central Limit Theorem (Kwak & Kim, 2017), satisfying the Normal condition for t-tests (Kim & Park, 2019).

Two-sample t-tests compare two sample means (NIST, n.d.), and paired tests compare dependent samples (Xu et al., 2017). In this study, paired two-sample t-tests compare differences in mean accuracy between models on the same evaluation dataset. One-tailed tests measure if modifications improved model performance. Two-tailed tests measure which modification was more effective.

Ethics Statement

Images are sourced from public databases (Codella et al., 2017, Daneshjou et al., 2022, Gutman et al., 2016) and public GitHub repositories (AlexKaiLe, n.d.). This research method received approval from an Institutional Review Board.

Results

Quantify Underrepresentation

Figure 6 shows that no dataset in this study had a skin tone distribution that reflected the distribution of Fitzpatrick Skin Types in the US population. Furthermore, the results of the chi-square goodness of fit tests (Appendix B) were all statistically significant, suggesting that no dataset in this study matched the distribution of Fitzpatrick Skin Types in the US population.

However, the chi-square test results do indicate that ISIC+ArGI and ISIC+DiDI represent darker skin tones more fairly than ISIC alone.

Figure 6

Skin tone distributions. US Average from Keiser et al. (2012).

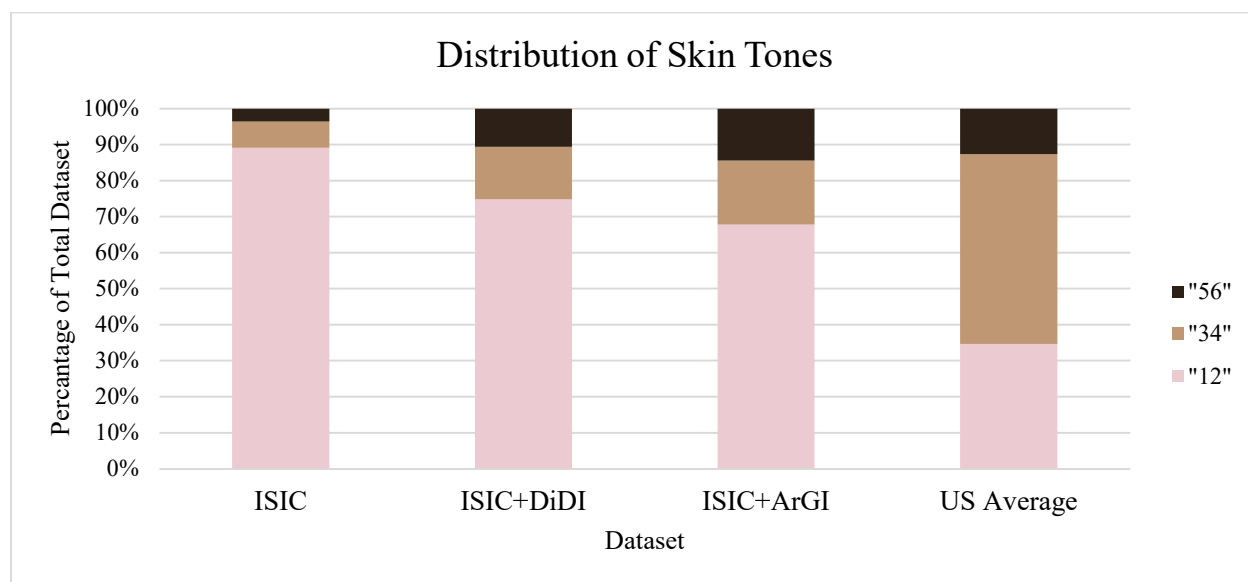
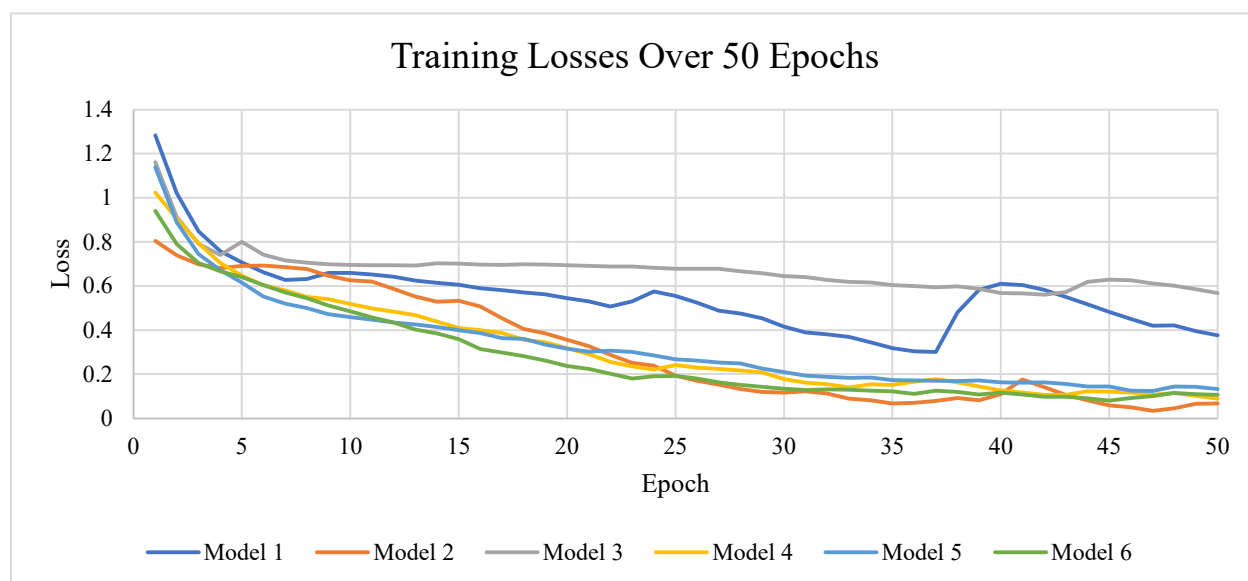


Figure 7

Loss values during each training epoch

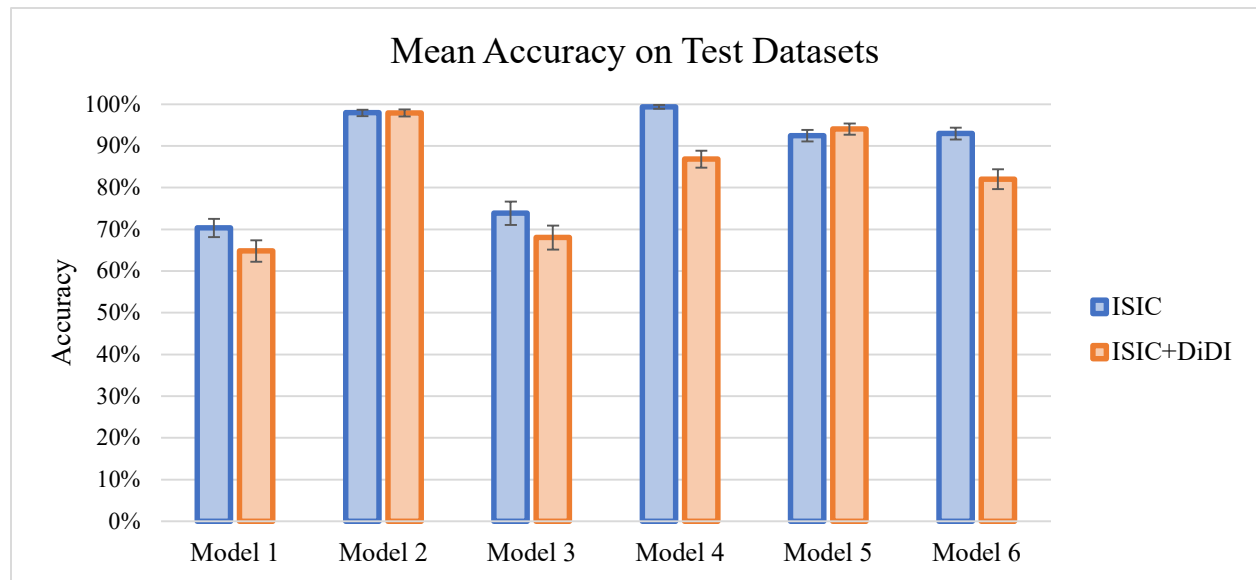


In Figure 7, all models showed decreasing losses. Standard classifiers showed spikes during this decreasing trend, while DB-VAEs showed consistently decreasing loss values, suggesting DB-VAE modification may increase stability during model training.

All DB-VAE models reached low final loss values, but the lowest was Model 2. Models 1 and 3 had the highest final loss values. This suggests that DB-VAE modification and training with more representative datasets improves losses. In contrast, training on datasets with poor representation for certain skin tones leads to higher loss values. Also, compared to real images, artificially generated images have a smaller effect on improving training loss.

Figure 8

Mean accuracy achieved by models on test datasets



Note. Error bars are 1x standard deviation. Models 1-3 are standard classifiers and 4-6 are DB-VAE. Models performed better when their test dataset matched their training dataset.

Significance Testing

Each group of tests in Table 2 answers different questions to help answer the broader research question.

Table 2*Results of t-tests and questions answered.*

Question	Test #	Models Compared	Alternative Hypothesis (H_a)	$t(58)$
1. Does modifying model architecture with DB-VAE mitigate algorithmic bias?	1a	1 and 4	$\mu_1 < \mu_4$	-71.2502***
				-36.8960***
	1b	2 and 5	$\mu_2 < \mu_5$	19.1215***
				13.1042***
	1c	3 and 6	$\mu_3 < \mu_6$	-33.4347***
				-20.5095***
2. Does supplementing model training with real images of underrepresented skin tones mitigate algorithmic bias?	2a	1 and 2	$\mu_1 < \mu_2$	-65.4559***
				-67.3888***
	2b	4 and 5	$\mu_4 < \mu_5$	26.5285***
				-15.9942***
3. Does supplementing model training with artificial images of underrepresented skin tones mitigate algorithmic bias?	3a	1 and 3	$\mu_1 < \mu_3$	-5.3988***
				-4.5707***
	3b	4 and 6	$\mu_4 < \mu_6$	23.4723***
				8.4395***
4. Which of DiDI and ArGI was most impactful when used to supplement model training?	4a	2 and 3	$\mu_2 \neq \mu_3$	45.6489***
				54.4869***
	4b	5 and 6	$\mu_5 \neq \mu_6$	-1.4107
				24.0140***

*** $p < .001$

For each test, blue cells correspond to evaluation results on ISIC alone; orange corresponds to evaluation results on ISIC+DiDI. For each model evaluation, the sample size is $N = 30$, degrees of freedom $df = 58$, and the null hypothesis between samples A and B is $H_0: \mu_A = \mu_B$.

Tests 1a-1c

Test 1 yielded mixed results depending on the training dataset. For tests 1a and 1c, the t -values were negative and statistically significant, indicating that the DB-VAE outperformed the

standard classifier and suggesting that converting from the standard classifier to DB-VAE improves model accuracy on both evaluation datasets.

In contrast, in test 1b, the t -values were positive and statistically significant, indicating that the standard classifier outperformed the DB-VAE and suggesting that converting the standard classifier to DB-VAE decreases model performance.

Tests 2a-2b

Test 2 indicated that training models with more real images of underrepresented skin tones mostly mitigate algorithmic bias. For test 2a, the t -values were negative and statistically significant, indicating that the model trained on ISIC+DiDI outperformed the model trained on ISIC alone and suggesting that training on more real images of underrepresented skin tones improves model accuracy.

In test 2b, similar results were seen when testing Models 4 and 5 on ISIC+DiDI; however, the positive t -value for the comparison between Models 4 and 5 on ISIC indicated that training on more real images of underrepresented skin tones improves model accuracy on ISIC+DiDI, but not on ISIC alone.

Tests 3a-3b

Test 3 indicated that training models with artificial images of underrepresented skin tones sometimes mitigates algorithmic bias. Test 3a indicated that, when using a standard classifier (Models 1 and 3), training with additional, artificially generated images of underrepresented skin tones improves model performance, as the t -values were negative and statistically significant.

In contrast, test 3b indicated that, when using a DB-VAE (Models 4 and 6), training with additional, artificially generated images of underrepresented skin tones does not improve model performance, as the t -values were positive and statistically significant.

Tests 4a-4b

Test 4 indicated that training on ISIC+DiDI leads to a more significant impact on model performance than training on ISIC+ArGI. 3 of the 4 subtests yielded positive, statistically significant t -values, indicating that models trained on ISIC+DiDI outperformed models trained on ISIC+ArGI.

There was an exception: when comparing DB-VAE models (Models 5 and 6) evaluated on ISIC, the t -value was negative and statistically insignificant ($p = 0.1637$). This indicated that Model 6 outperformed Model 5 when evaluated on ISIC, but this difference was negligible.

Discussion

The gap addressed by this research has three components. First, there is not much research that uses the Diverse Dermatology Images dataset. Model evaluation on ISIC+DiDI and training Models 2 and 5 help close this gap.

Next, no study compares different debiasing approaches on standard architectures to answer the question: what is the best way to mitigate algorithmic bias in machine learning malignance classifiers used for dermatology? This study compares real image supplementation, artificial image generation, DB-VAE, and several combination modifications for dermatology classification networks, all on ResNet50v2.

The last component is the lack of combination approaches, which is addressed by the combination of DB-VAE and artificial image generation in Model 6.

Dataset Representation

Although the chi-square tests (Appendix B) show that no dataset fairly represents darker skin tones when compared to the national average, Figure 6 shows that ISIC+ArGI better

represents darker skin tones compared to ISIC+DiDI. However, ArGI was generated with content images from ISIC, which does not include any lesion images on darker skin tones. Thus, despite ISIC+ArGI quantitatively representing darker skin tones more fairly, ISIC+DiDI may represent the qualitative clinical presentations of lesions on darker skin tones more fairly.

Does the Diverse Dermatology Images Dataset help mitigate algorithmic bias? (Question 2)

Models trained on ISIC+DiDI outperformed models trained on ISIC alone, except when Model 4 outperformed Model 5 when evaluated on ISIC. This could be attributed to how DB-VAE models, like dermatologists, use experience to guide their diagnoses (Fathy & Lipoff, 2022). The DB-VAE encoder-decoder architecture is analogous to a human dermatologist, reconstructing past images to aid its diagnoses.

Model 4 performs exceptionally well on ISIC because its reconstructions are specialized during training to be very accurate on ISIC images. Model 5 performs worse because the network's reconstructions had to account for different clinical presentations in DiDI.

However, since ML classifiers are used on all skin tones, not just the lighter tones overrepresented in ISIC, thus, supplementing model training with real images of underrepresented skin tones still strongly mitigates algorithmic bias.

This aligns with the scholarly conversation, which asserts that training CNNs on more diverse datasets made algorithms “more robust and less biased when deployed to a new environment” (Abubakar et al., 2020).

Does combining DB-VAE with artificial image generation mitigate algorithmic bias?

(Question 3)

Improvements in ResNet50v2 performance with artificial image supplementation provide additional support for Rezk et al.'s findings that generated images “contributed to improvement

in the classification performance when used to augment the training of ResNet-50” (Rezk et al., 2022b).

However, test 3b shows that combining artificial image generation with DB-VAE was worse than DB-VAE trained on ISIC alone. One possible explanation is that artificial images reconstructed by DB-VAE are doubly artificial, deviating from real clinical presentations of lesions. Nonetheless, Figure 8 shows that DB-VAE combined with artificial image generation (Model 6) outperforms the base standard classifier trained on ISIC (Model 1).

What is the best way to mitigate algorithmic bias in machine learning malignance classifiers used for dermatology?

Supplementing model training with real images of underrepresented skin tones was most effective in debiasing ML classification models for skin cancer detection. DB-VAE modification was also very effective, performing better and more consistently than standard classifiers across most training datasets (Figure 8). Artificial image generation was effective, but not in combination with DB-VAE. All modifications showed improvement over Model 1.

DB-VAE

DB-VAE’s creators described the architecture as “an additional tool to promote systematic, algorithmic fairness of modern AI systems” (Amini et al., 2019), not as a replacement for previous debiasing approaches.

The results agree with that perspective: when both models were trained on ISIC and ISIC+ArGI, DB-VAE outperformed standard classifiers. However, standard classifiers outperformed DB-VAE when both models were trained on ISIC+DiDI, showing that DB-VAE is a beneficial short-term modification that improves model accuracy on more biased datasets. DB-

VAE’s benefit may decline as researchers add more clinical images of underrepresented skin tones to the publicly available dermoscopic image databases.

DiDI vs. ArGI

The results show that DiDI was more impactful; Tests 4a-4b showed that, for model training, DiDI significantly outperformed ArGI. When ArGI outperformed DiDI, the result was statistically insignificant.

Rezk et al. (2022b) acknowledged the limitations of artificial image generation, stating that “images with real pathology in people of color are required to improve model training and validation.” This study reiterates this limitation but also acknowledges that adding more images with real clinical presentations of lesions is a long-term solution driven by policy solutions like the STANDING Together Initiative (Ganapathi et al., 2022).

Although supplementation with real images is better, artificial image generation mitigates algorithmic bias and can be used when images for lesions on darker skin tones are scarce.

Limitations

Model Training

ISIC had 656 fewer training images than the other training datasets. This may have negatively impacted performance in Models 1 and 4, as Thian et al. (2022) found that performance improved “rapidly as training volume increased up to 20,000 samples.” Thian et al. (2022) also investigated ResNet-50 and used a batch size of 32, though their models were used for pneumothorax classifiers, not dermatology malignance classifiers.

However, this limitation is not very significant because dataset sizes will increase as policy initiatives lead to increasingly representative datasets. Adding new images does not come at the expense of removing images of lighter skin tones.

Image Preprocessing

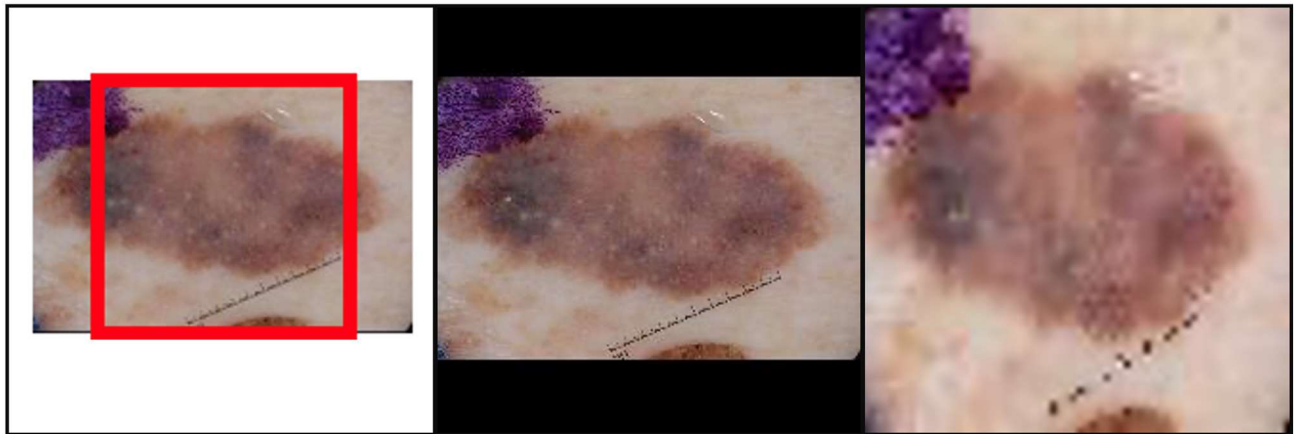
Figures 9-11

Comparing various image processing techniques

Figure 9
*ISIC_012773's original
image, overlayed with a 1:1
aspect ratio square*

Figure 10
*ISIC_012773, resized with
padding*

Figure 11
*ISIC_012773 after
processing*



One common sign of melanoma is asymmetry (The Skin Cancer Foundation, 2021); the asymmetry of malignant lesions may have been distorted during image processing because aspect ratios were not preserved when resizing images to a 1:1 aspect ratio. Comparing Figures 10 and 11 show this distortion; the shape of the lesion is much less elliptical and more circular in Figure 11.


Abubakar et al. (2020) and Amini et al. (2019) cropped their images to fit a 1:1 square aspect ratio during preprocessing, but the same approach could not be taken in this study because several images had lesions that were too large to confine to a square aspect ratio (Figure 9).

Instead of cropping to a 1:1 aspect ratio, another option was to resize the image and add black padding to the sides. However, this would lead to issues with ITA calculations, as addressed in the next section.

ITA Calculations

Figure 12

Fitzpatrick Skin Tone Scale (Ward et al., 2017)

Type I	Type II	Type III	Type IV	Type V	Type VI
White skin. Always burns, never tans.	Fair skin. Always burns, tans with difficulty.	Average skin color. Sometimes mild burn, tan about average.	Light-brown skin. Rarely burns. Tans easily.	Brown skin. Never burns. Tans very easily.	Black skin. Heavily pigmented. Never burns, tans very easily.
					

Figures 13-14

Images with misclassified skin tones

Figure 13
ISIC_000004 after processing

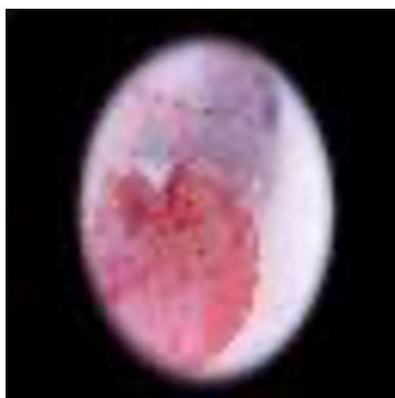
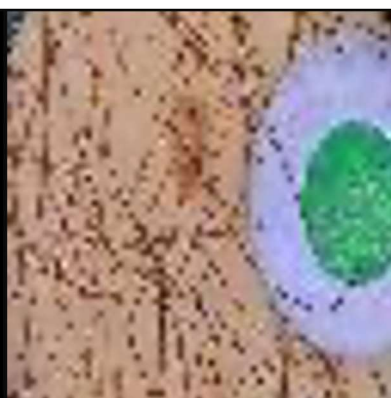


Figure 14
ISIC_003056 after processing



Figures 13 and 14 exemplify how visual artifacts, like microscope surroundings or stickers, could skew ITA-FST calculations. Figure 13 is a dermoscopic image taken through a microscope; the surrounding black pixels are not skin. However, Figure 13 had a calculated skin tone of “56” because the ITA-FST algorithm sampled pixels from the surrounding black area, skewing the ITA calculation toward a darker skin tone. This also explains why resizing with padding (Figure 10) was not viable for this study: additional black pixels skew the ITA calculations, leading to the misrepresentation of skin tone distributions.

Other artifacts like hair or stickers may have also affected the ITA calculations. Figure 14 was assigned an FST value of “12,” when it appears to be Type III or IV when compared to Figure 12. This could be attributed to the bright colors of the sticker skewing the ITA-FST calculations toward lighter skin tones. Overall, the IST-FTA labeling was not perfect, but provides strong estimates for the representation of skin tones in each dataset.

Conclusion

The study’s goal was to find the best ways to mitigate algorithmic bias in models for skin cancer detection. This study investigated several leading approaches on a uniform CNN architecture, comparing standard classifiers with debiasing variational autoencoders and real clinical images with artificially generated ones. Ultimately, the best solution is to train models with more images of darker skin tones, addressing algorithmic bias by directly bridging the gap in representation between lighter and darker skin tones, confirming results by Abubakar et al. (2020). However, DB-VAE and artificial image generation using style transfer are both powerful short-term solutions that can improve model performance, even on existing biased datasets.

This study fills the gap in scholarly conversation, showing that combining DB-VAE with artificial image generation does mitigate algorithmic bias. However, the two modifications may work against each other, leading to an undesirable result. Future studies can investigate why this occurs and experiment with combinations of other debiasing approaches.

Also, this study shows how DDI contains much useful metadata, making it easy to process. These results imply that the policy initiatives for improved data collection practices are important and future researchers should continue to increase representation for darker skin tones and create well-labeled datasets like the DDI Dataset (Daneshjou et al., 2022).

Last, this research used the Fitzpatrick Skin Type Scale to measure skin tone. Future research can investigate these same debiasing combinations on novel skin tone scales like MST (Skin Tone Research @ Google AI, n.d.). Additionally, future researchers can compare new combinations as other researchers find new approaches to mitigating algorithmic bias. This study created 6 total models, but there are many possible combinations. The implications of these combinations are the potential avenues for increasing algorithmic fairness in clinical settings and improving diagnosis accuracy of the most common cancer.

References

- Abubakar, A., Ugail, H., & Bukar, A. M. (2020). Assessment of Human Skin Burns: A Deep transfer Learning approach. *Journal of Medical and Biological Engineering*, 40(3), 321–333. <https://doi.org/10.1007/s40846-020-00520-z>
- AlexKaiLe. (n.d.). *GitHub - AlexKaiLe/Debiasing-Melanoma-Images: Developed a deep learning computer vision program that is able to generate melanoma images for people with darker skin tones*. GitHub. <https://github.com/AlexKaiLe/Debiasing-Melanoma-Images/>
- American Academy of Dermatology Association. (2022). *Skin cancer*. <https://www.aad.org/media/stats-skin-cancer>
- Amini, A. (2024). *introtodeeplearning/lab2/Part2_Debiasing.ipynb at master · aamini/introtodeeplearning*. GitHub. https://github.com/aamini/introtodeeplearning/blob/master/lab2/Part2_Debiasing.ipynb
- Amini, A., Soleimany, A., Schwarting, W., Bhatia, S. & Rus, D. (2019). *Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure* [Conference Proceedings]. AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA. <https://doi.org/10.1145/3306618.3314243>
- Bardou, D., Bouaziz, H., Lv, L., & Zhang, T. (2022). Hair removal in dermoscopy images using variational autoencoders. *Skin Research and Technology*, 28(3), 445–454. <https://doi.org/10.1111/srt.13145>
- Butt, S., Butt, H., & Gnanappiragasam, D. (2021). Unintentional consequences of artificial intelligence in dermatology for patients with skin of colour. *Clinical and Experimental Dermatology*, 46(7), 1333–1334. <https://doi.org/10.1111/ced.14726>

- ByPass Publishing. (2013). *Research Methods: Experimental Design* [Video]. YouTube.
<https://www.youtube.com/watch?v=qtLnBz6lbRQ>
- Cai, C., Nishimura, T., Hwang, J., Hu, X., & Kuroda, A. (2021). Asbestos Detection with Fluorescence Microscopy Images and Deep Learning. *Sensors*, 21(13), 4582.
<https://doi.org/10.3390/s21134582>
- Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75, 102305. <https://doi.org/10.1016/j.media.2021.102305>
- Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T., Lipková, J., Lu, M., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6), 719–742.
<https://doi.org/10.1038/s41551-023-01056-8>
- Codella, N., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N. K., Kittler, H., & Halpern, A. C. (2017). 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). *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1710.05006>
- Cullell-Dalmau, M., Otero-Viñas, M., & Manzo, C. (2020). Research techniques made simple: deep learning for the classification of dermatological images. *Journal of Investigative Dermatology*, 140(3), 507-514.e1. <https://doi.org/10.1016/j.jid.2019.12.029>
- Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M. J., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. a. C., Okata-Karigane, U., Zou, J., & Chiou, A. S.

- (2022a). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32). <https://doi.org/10.1126/sciadv.abq6147>
- Debelee, T. G. (2023). Skin lesion Classification and Detection Using Machine Learning Techniques: A Systematic review. *Diagnostics*, 13(19), 3147. <https://doi.org/10.3390/diagnostics13193147>
- Derm-ita*. (2021). PyPI. <https://pypi.org/project/derm-ita/>
- Fathy, R., & Lipoff, J. B. (2022). Lack of skin of color in Google image searches may reflect under-representation in all educational resources. *Journal of the American Academy of Dermatology*, 86(3), e113–e114. <https://doi.org/10.1016/j.jaad.2021.04.097>
- Ganapathi, S., Palmer, J., Alderman, J., Calvert, M., Espinoza, C., Gath, J., Ghassemi, M., Heller, K., McKay, F., Karthikesalingam, A., Kuku, S., Mackintosh, M., Manohar, S., Mateen, B. A., Matin, R., McCradden, M., Oakden-Rayner, L., Ordish, J., Pearson, R., Pfohl, S.R., Rostamzadeh, N., Sapey, E., Sebire, N., Sounderajah, V., Summers, C., Treanor, D., Denniston, A. K., & Liu, X. (2022). Tackling bias in AI health datasets through the STANDING Together initiative. *Nature Medicine*, 28(11), 2232–2233. <https://doi.org/10.1038/s41591-022-01987-w>
- Google for Developers. (2023). *Machine Learning Glossary*. <https://developers.google.com/machine-learning/glossary>
- Groh, M., Harris, C., Daneshjou, R., Badri, O., & Koochek, A. (2022). Towards Transparency in Dermatology Image Datasets with Skin Tone Annotations by Experts, Crowds, and an Algorithm. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW2), 1–26. <https://doi.org/10.1145/3555634>

- Groh, M., Harris, C., Soenksen, L. R., Lau, F. D., Han, R., Kim, A., Koochek, A., & Badri, O. (2021). Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. *arXiv (Cornell University)*.
<https://doi.org/10.48550/arxiv.2104.09957>
- Gutman, D., Codella, N., Celebi, E., Helba, B., Marchetti, M. A., Mishra, N. K., & Halpern, A. C. (2016). Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv (Cornell University)*.
<https://doi.org/10.48550/arxiv.1605.01397>
- Kebaili, A., Lapuyade-Lahorgue, J., & Ruan, S. (2023). Deep Learning Approaches for data augmentation in Medical Imaging: A review. *Journal of Imaging*, 9(4), 81.
<https://doi.org/10.3390/jimaging9040081>
- Keiser, E., Linos, E., Kanzler, M. H., Lee, W., Sainani, K. L., & Tang, J. Y. (2012). Reliability and prevalence of digital image skin types in the United States: Results from National Health and Nutrition Examination Survey 2003-2004. *Journal of the American Academy of Dermatology*, 66(1), 163–165. <https://doi.org/10.1016/j.jaad.2011.02.044>
- Keras Team. (n.d.). *Keras Applications*. Keras Documentation. <https://keras.io/api/applications/>
- Kinyanjui, N., Odonga, T., Cintas, C., Codella, N., Panda, R., Sattigeri, P., & Varshney, K. (2019, Dec. 14). *Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets* [Workshop Presentation]. NeurIPS Fair ML for Health Workshop, Vancouver, BC, Canada. <https://doi.org/10.48550/arXiv.1910.13268>

- Kim, T. K., & Park, J. H. (2019). More about the basic assumptions of t-test: normality and sample size. *Korean Journal of Anesthesiology*, 72(4), 331–335.
<https://doi.org/10.4097/kja.d.18.00292>
- Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology*, 70(2), 144. <https://doi.org/10.4097/kjae.2017.70.2.144>
- Li, Z., Koban, K. C., Schenck, T. L., Giunta, R. E., Li, Q., & Sun, Y. (2022). Artificial intelligence in Dermatology Image Analysis: Current developments and future trends. *Journal of Clinical Medicine*, 11(22), 6826. <https://doi.org/10.3390/jcm11226826>
- Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: challenges and mitigation strategies. *Npj Digital Medicine*, 6(1).
<https://doi.org/10.1038/s41746-023-00858-z>
- NIST. (n.d.). 1.3.5. *Quantitative Techniques*. National Institute of Standards and Technology Engineering Statistics Handbook.
<https://www.itl.nist.gov/div898/handbook/eda/section3/eda35.htm>
- Petersen, E., Holm, S., Ganz, M., & Feragen, A. (2023). The path toward equal performance in medical machine learning. *Patterns*, 4(7), 100790.
<https://doi.org/10.1016/j.patter.2023.100790>
- Reilley-Luther, J., Cline, A., Zimmerly, A., Azing, S., & Moy, J. (2020). Representation of Fitzpatrick skin type in dermatology textbooks compared with national percentiles. *Dermatology Online Journal*, 26(12). <https://doi.org/10.5070/d32612051349>
- Rezk, E., Eltorki, M., & El-Dakhkhni, W. (2022a). Leveraging artificial intelligence to improve the diversity of dermatological skin color pathology: Protocol for an algorithm

- development and validation study. *JMIR Research Protocols*, 11(3), e34896.
<https://doi.org/10.2196/34896>
- Rezk, E., Eltorki, M., & El-Dakhakhni, W. (2022b). Improving skin color diversity in Cancer Detection: Deep Learning approach. *JMIR Dermatology*, 5(3), e39143.
<https://doi.org/10.2196/39143>
- Rezk, E., Haggag, M., Eltorki, M., & El-Dakhakhni, W. (2023). A comprehensive review of artificial intelligence methods and applications in skin cancer diagnosis and treatment: Emerging trends and challenges. *Healthcare Analytics*, 4, 100259.
<https://doi.org/10.1016/j.health.2023.100259>
- Skin Tone Research @ Google AI. (n.d.). *Developing the Monk Skin Tone Scale*.
<https://skintone.google/the-scale>
- Stanford Vision Lab, Stanford University, & Princeton University. (2020). *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*. <https://image-net.org/challenges/LSVRC/>
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv (Cornell University)*.
<https://doi.org/10.48550/arxiv.1409.4842>
- TensorFlow. (n.d.). *Tutorials*. Tutorials | Tensorflow Core. <https://www.tensorflow.org/tutorials/>
- The Skin Cancer Foundation. (2021). *Melanoma Warning Signs and Images - The Skin Cancer Foundation*. <https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/>
- Thian, Y. L., Ng, D. W., Hallinan, J. T. P. D., Jagmohan, P., Sia, D. S. Y., Mohamed, J. S. A., Quek, S. T., & Feng, M. (2022). Effect of training data volume on performance of

- convolutional neural network pneumothorax classifiers. *Journal of Digital Imaging*, 35(4), 881–892. <https://doi.org/10.1007/s10278-022-00594-y>
- Ward, W. H., Lambreton, F., Goel, N., Yu, J., & Farma, J. M. (2017). Clinical presentation and staging of melanoma. In *Codon Publications eBooks* (pp. 79–89). <https://doi.org/10.15586/codon.cutaneousmelanoma.2017.ch6>
- Wen, D., Khan, S., Ji-Xu, A., Ibrahim, H., Smith, L. A., Caballero, J. A., Zepeda, L., De Blas Perez, C., Denniston, A. K., Liu, X., & Martin, R. (2022). Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1), e64–e74. [https://doi.org/10.1016/s2589-7500\(21\)00252-1](https://doi.org/10.1016/s2589-7500(21)00252-1)
- Xu, M., Fralick, D., Zheng, J. Z., Wang, B., Tu, X., & Feng, C. (2017). The differences and similarities between Two-Sample T-Test and paired T-Test. *PubMed*, 29(3), 184–188. <https://doi.org/10.11919/j.issn.1002-0829.217070>

Appendix A

Justification for Model Architecture Selection

All models were implemented using the Keras API on the Tensorflow framework. The Keras API includes applications (Keras Team, *n.d.*) for popular CNN architectures like “GoogleNet, Inception-V3, V4, ResNet, Inception-ResNet V2 and Dense Net” (Li et al., 2022).

Inception-style networks were tested first. Neither Inception V4 nor GoogleNet was available through the Keras API, but Inception-V3 and Inception-ResNetV2 were. Inception-V3, which is 92 megabytes large and contains 23.9 million parameters (Keras Team, *n.d.*) was tested first on a subset of 100 ISIC images. It failed to complete training, causing memory allocation errors after a few epochs because the model attempted to use more than 6 GB of memory, which surpassed the amount of memory available on the NVIDIA GeForce 1660 Ti GPU used for this study. Inception-ResNetv2, which is 215 megabytes large and contains 55.9 million parameters, was also tested. Similar memory allocation errors occurred, so it was concluded that Inception-style networks required too much memory to be trained on the study computer.

ResNet models were tested next. Considering Inception-V3 started model training, but failed during the process, a ResNet model with a similar size to Inception-V3 was selected. ResNet50 was a suitable choice, with a size of 98 megabytes and 25.6 million parameters compared to Inception-V3’s 92 megabytes and 23.9 million parameters (Keras Team, *n.d.*). Instead of testing ResNet50, the upgraded ResNet50v2 was tested because it had the same size and number of parameters but had a faster time per inference step on GPU and greater accuracy metrics on ImageNet (Stanford Vision Lab et al., 2020) compared to ResNet50 (Keras Team, *n.d.*). ResNet50v2 completed model training, making it a potential candidate for the study’s standardized CNN architecture.

DenseNet was the last style of CNN to be tested. A DenseNet with a similar size to Inception-V3 and ResNet50v2 was selected. DenseNet201, 80 megabytes large with 20.2 million parameters, was most similar to Inception-V3 and ResNet50v2. However, like Inception-V3, DenseNet201 also failed to complete model training.

These outcomes may be attributed to the “Time (ms) per inference step (GPU)” (Keras Team, n.d.) values for each model. ResNet50v2 takes only 4.4 milliseconds per inference step; in comparison, Inception-V3 needs 6.9 ms, InceptionResNetV2 needs 10.0 ms, and DenseNet201 needs 6.7 ms. It was concluded that ResNet50v2 was the best choice for the study because it was trainable on the study computer and contained comparable depth and accuracy metrics compared to other popular CNN architectures.

Appendix B

Chi-Square Analysis for Quantifying Underrepresentation

Table B1

National averages for FST categories provided by Keiser et al. (2012).

Fitzpatrick Skin Type	Percentage of US Population	Corresponding DDI-style skin tone	Percentage of US Population
I	1.60 %	12	34.70 %
II	33.10 %		
III	47.80 %	34	52.70 %
IV	4.90 %		
V	3.60 %	56	12.60 %
VI	9.00 %		

Table B2

Raw data for the chi-square goodness-of-fit tests

Skin Type	ISIC		ISIC+DiDI		ISIC+ArGI	
	Observed	Expected	Observed	Expected	Observed	Expected
"12"	1762	685.672	1970	913.304	1784	913.304
"34"	143	1041.352	384	1387.064	469	1387.064
"56"	71	248.976	278	331.632	379	331.632
Total	1976	1976	2632	2632	2632	2632

Table B3

Results of chi-square goodness-of-fit tests to determine if datasets fit national FST distribution.

Dataset	X^2	Degrees of freedom	Sample size	p-value
ISIC	2591.7690	2	1976	0*
ISIC+DiDI	1956.6465	2	2632	0*
ISIC+ArGI	1444.4858	2	2632	0*

* $p < .05$